Pontíficia Universidade Católica do Rio de Janeiro
Departamento de Engenharia Elétrica

Quantile Regression

Marcelo Castiel Ruas*, Henrique Helfer Hoeltgebaum†, Alexandre Street‡,
Cristiano Fernandes§

January 16, 2017

*Aluno de doutorado do Departamento de Engenharia Elétrica da PUC-RIO.
†Aluno de doutorado do Departamento de Engenharia Elétrica da PUC-RIO.
‡Professor do Departamento de Engenharia Elétrica da PUC-RIO.
§Professor do Departamento de Engenharia Elétrica da PUC-RIO.

1 Introduction

Wind Firm Energy Certificate (FEC) [4] estimation imposes several challenges. First, it is a quantile function of an aleatory quantity, named here on wind capacity factor (WP). Due to its non-dispachable profile, accurate scenario generation models could reproduce a fairly dependence structure in order to the estimation of FEC. Second, as it is a quantile functions, the more close to the extremes of the interval, the more sensitive to sampling error.

In this work, we apply a few different techniques to forecast the quantile function a few steps ahead. The main frameworks we investigate are parametric linear models and a non-parametric regression. In all approaches we use the time series lags as the regression covariates. To study our methods performance, we use the mean power monthly data of Icaraizinho, a wind farm located in the Brazilian northeast.

The Icaraizinho dataset consists of monthly observations from 1981 to 2011 of mean power measured in Megawatts. The full Icaraizinho serie can be found on the appendices from this article. As is common in renewable energy generation, there is a strong seasonality component. Figures 1.1 and 1.2 illustrate this seasonality, where we can observe low amounts of power generation for the time span between February and May, and a yearly peek between August and November. Figure 1.3 shows four scatter plots relating y_t with some of its lags. We choose to present here the four lags that were selected for the quantile regression in the experiment of section **??**, which are the 1$^{\text{st}}$, 4$^{\text{th}}$, 11$^{\text{th}}$ and 12$^{\text{th}}$. They are most likely the four main lags to use for these analysis.

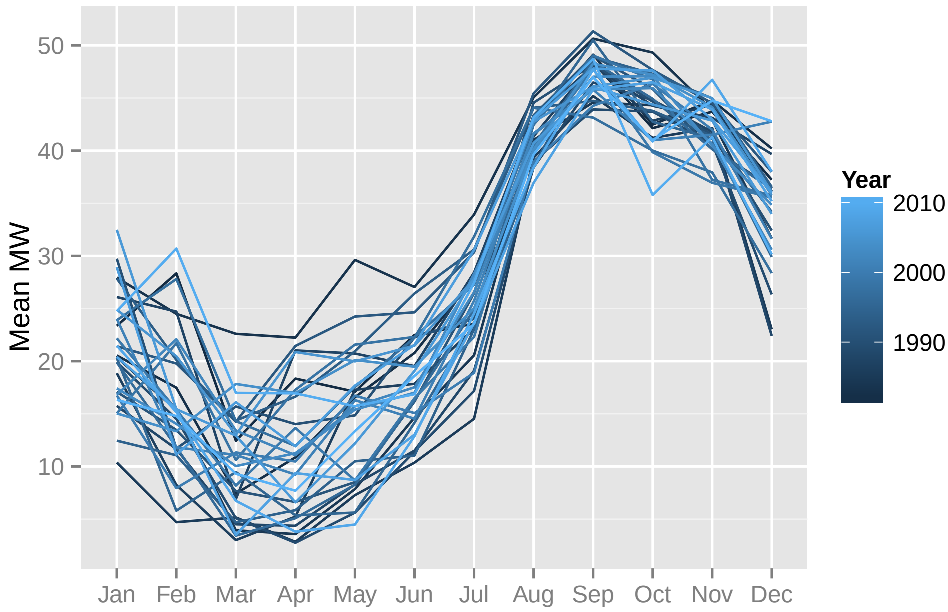


Figure 1.1: Icaraizinho yearly data. Each serie consists of monthly observations for each year.

Here we denote as parametric linear model the well-known quantile regression model [3]. In contrast to the linear regression model through ordinary least squares (OLS), which provides only an estimation of the dependent variable conditional mean, quantile regression model yields a much more detailed information concerning the complex relationship about the dependent variable and its covariates. A Quantile Regression for the α-quantile is the solution of the following optimization problem:

$$\min_q \sum_{t=1}^{n} \alpha |y_t - q(x_t)|^+ + (1-\alpha)|y_t - q(x_t)|^-, \tag{1.1}$$

where $q(x_t)$ is the estimated quantile value at a given time t and $|x|^+ = \max\{0, x\}$ and $|x|^- = -\min\{0, x\}$. To model this problem as a Linear Programming problem, thus being able to use a

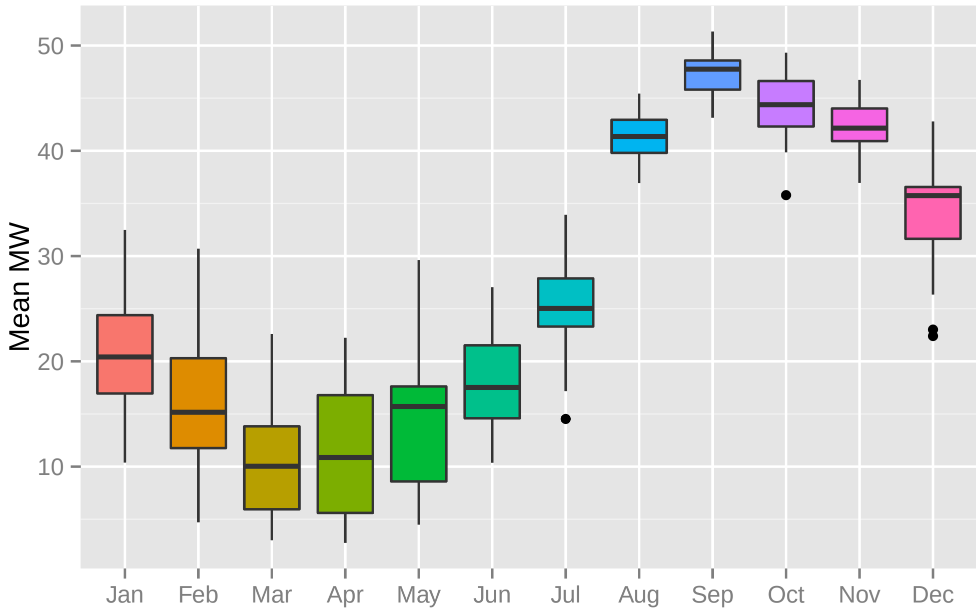


Figure 1.2: Boxplot for each month for the Icaraizinho dataset

modern solver to fit our model, we can create variables ε_t^+ e ε_t^- to represent $|y-q(x_t)|^+$ and $|y-q(x_t)|^-$, respectively. So we have:

$$\begin{aligned}
&\min_{q,\varepsilon_t^+,\varepsilon_t^-} \sum_{t=1}^{n} \left(\alpha\varepsilon_t^+ + (1-\alpha)\varepsilon_t^-\right) && \\
&\text{s.t. } \varepsilon_t^+ - \varepsilon_t^- = y_t - q(x_t), && \forall t \in \{1,\dots,n\}, \\
&\qquad \varepsilon_t^+, \varepsilon_t^- \geq 0, && \forall t \in \{1,\dots,n\}.
\end{aligned} \tag{1.2}$$

Section 2 is about linear models, so we investigate the quantile estimation when q is a linear function of the series past values, up to a maximum number of lags p:

$$q(y_t, \alpha; \beta) = \beta_0(\alpha) + \beta_1(\alpha)y_{t-1} + \beta_2(\alpha)y_{t-2} + \cdots + \beta_p(\alpha)y_{t-p}. \tag{1.3}$$

In section 3 we introduce a Nonparametric Quantile Autoregressive model with a ℓ_1-penalty term, in order to properly simulate FEC densities for several α-quantiles. In this nonparametric approach we don't assume any form for $q(x_t)$, but rather let the function adjust to the data. To prevent overfitting, the ℓ_1 penalty for the second derivative (approximated by the second difference of the ordered observations) is included in the objective function.

In section 4 we investigate how to simulate S scenarios of y_t, considering a linear model and errors ε_t for which the distribution is unknown. To address this issue, we use quantile linear regression to calculate a thin grid of quantiles and fit a distribution function $\hat{F}_{y_t}$. This function will be used to simulate the innovations on the model.

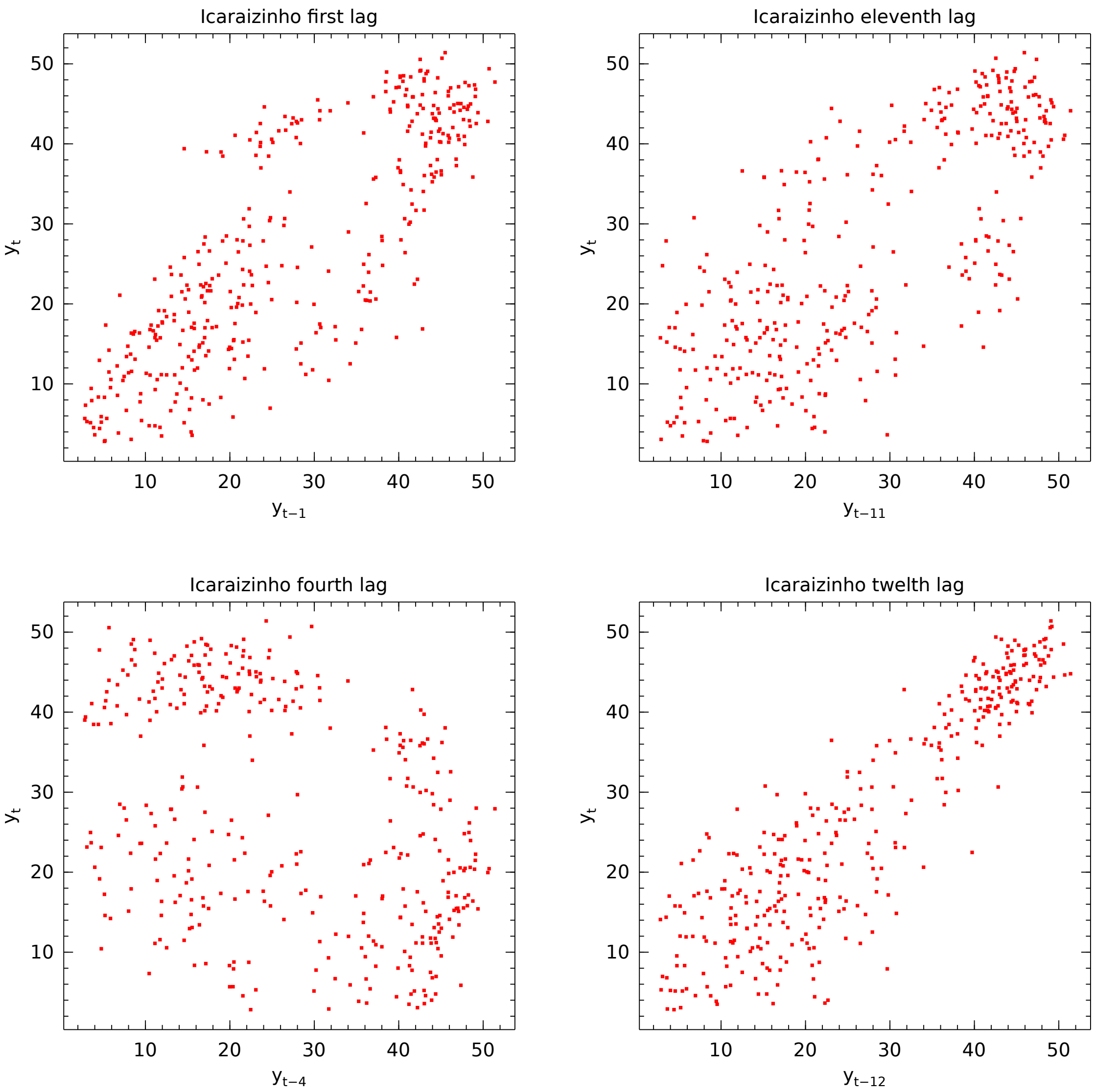


Figure 1.3: Relationship between y_t and some chosen lags.

2 Linear Models for the Quantile Autoregression

Given a time series $\{y_t\}$, we investigate how to select which lags will be included in the Quantile Autoregression. We won't be choosing the full model because this normally leads to a bigger variance in our estimators, which is often linked with bad performance in forecasting applications. So our strategy will be to use some sort of regularization method in order to improve performance. We investigate two ways of accomplishing this goal. The first of them consists of selecting the best subset of variables through Mixed Integer Programming, given that K variables are included in the model. Using MIP to select the best subset of variables is investigated in [1]. The second way is including a ℓ_1 penalty on the linear quantile regression, as in [2], and let the model select which and how many variables will have nonzero coefficients. Both of them will be built over the standard Quantile Linear Regression model. In the end of the section, we discuss a information criteria to be used for quantile regression and verify how close are the solutions in the eyes of this criteria.

When we choose $q(x_t)$ to be a linear function, as on equation 1.1 (that we reproduce below for convenience):

$$\min_{f} \sum_{t=1}^{n} \alpha |y_t - q(x_t)|^{+} + (1-\alpha)|y_t - q(x_t)|^{-}, \tag{2.1}$$

we can substitute it on problem 1.2, getting the following LP problem:

$$\begin{aligned} \min_{\beta_0,\beta,\varepsilon_t^+,\varepsilon_t^-} \; & \sum_{t=1}^{n} \left(\alpha \varepsilon_t^+ + (1-\alpha)\varepsilon_t^- \right) \\ \text{s.t. } & \varepsilon_t^+ - \varepsilon_t^- = y_t - \beta_0 - \beta^T x_t, \qquad \forall t \in \{1,\dots,n\}, \\ & \varepsilon_t^+, \varepsilon_t^- \geq 0, \qquad \forall t \in \{1,\dots,n\}. \end{aligned} \tag{2.2}$$

In this work, we didn't explore the addition of terms other than the terms y_t past lags. For example, we could include functions of y_{t-p}, such as $log(y_{t-p})$ or $exp(y_{t-p})$. We leave such inclusion for further works.

3 Quantile Autoregression with a nonparametric approach

Fitting a linear estimator for the Quantile Auto Regression isn't appropriate when nonlinearity is present in the data. This nonlinearity may produce a linear estimator that underestimates the quantile for a chunk of data while overestimating for the other chunk (for example, scatter plot of y_t versus y_{t-1} that is seen on the upper left of figure 1.3). To prevent this issue from occurring we propose a modification which we let the prediction $\mathcal{Q}_{y_t|y_{t-1}}(\alpha)$ adjust freely to the data and its nonlinearities. To prevent overfitting and smoothen our predictor, we include a penalty on its roughness by including the ℓ_1 norm of its second derivative. For more information on the ℓ_1 norm acting as a filter, one can refer to [2].

Let $\{\tilde{y}_t\}_{t=1}^n$ be the sequence of observations in time t. Now, let $\tilde{x}_t$ be the $p-$lagged time series of $\tilde{y}_t$, such that $\tilde{x}_t = L^p(\tilde{y}_t)$, where L is the lag operator. Matching each observation $\tilde{y}_t$ with its $p-$lagged correspondent $\tilde{x}_t$ will produce $n-p$ pairs $\{(\tilde{y}_t, \tilde{x}_t)\}_{t=p+1}^n$ (note that the first p observations of y_t must be discarded). When we order the observation of x in such way that they are in growing order

$$\tilde{x}^{(p+1)} \leq \tilde{x}^{(p+2)} \leq \cdots \leq \tilde{x}^{(n)},$$

we can then define $\{x_i\}_{i=1}^{n-p} = \{\tilde{x}^{(t)}\}_{t=p+1}^n$ and $\{y_i\}_{i=1}^{n-p} = \{\tilde{y}^{(t)}\}_{t=p+1}^n$ and $T = \{2, \ldots, n-p-1\}$. As we need the second difference of q_i, I has to be shortened by two elements.

Our optimization model to estimate the nonparametric quantile is as follows:

$$\begin{aligned}\mathcal{Q}^{\alpha}_{y_t|y_{t-1}}(t) = \arg\min_{q_t} \sum_{t\in T} \left(\alpha|y_t - q_t|^+ + (1-\alpha)|y_t - q_t|^-\right) \\ +\lambda \sum_{t\in T} |D^2_{x_t} q_t|,\end{aligned} \tag{3.1}$$

where $D^2 q_t$ is the second derivative of the q_t function, calculated as follows:

$$D^2_{x_t} q_t = \frac{\left(\frac{q_{t+1}-q_t}{x_{t+1}-x_t}\right) - \left(\frac{q_t - q_{t-1}}{x_t - x_{t-1}}\right)}{x_{t+1} - 2x_t + x_{t-1}}.$$

The first part on the objective function is the usual quantile regression condition for $\{q_t\}$. The second part is the ℓ_1-filter. The purpose of a filter is to control the amount of variation for our estimator q_t. When no penalty is employed we would always get $q_t = y_t$. On the other hand, when $\lambda \to \infty$, our estimator approaches the linear quantile regression.

The full model can be rewritten as a LP problem as bellow:

$$\min_{q_t} \quad \textstyle\sum_{t=1}^n \left(\alpha\delta_t^+ + (1-\alpha)\delta_t^-\right) + \lambda \sum_{t=1}^n \xi_t \tag{3.2}$$

$$s.t. \quad \delta_t^+ - \delta_t^- = y_t - q_t, \qquad \forall t \in \{3, \ldots, n-1\}, \tag{3.3}$$

$$D_t = \left(\frac{q_{t+1}-q_t}{x_{t+1}-x_t}\right) - \left(\frac{q_t-q_{t-1}}{x_t-x_{t-1}}\right) \qquad \forall t \in \{3, \ldots, n-1\}, \tag{3.4}$$

$$\xi_t \geq D_t, \qquad \forall t \in \{3, \ldots, n-1\}, \tag{3.5}$$

$$\xi_t \geq -D_t, \qquad \forall t \in \{3, \ldots, n-1\}, \tag{3.6}$$

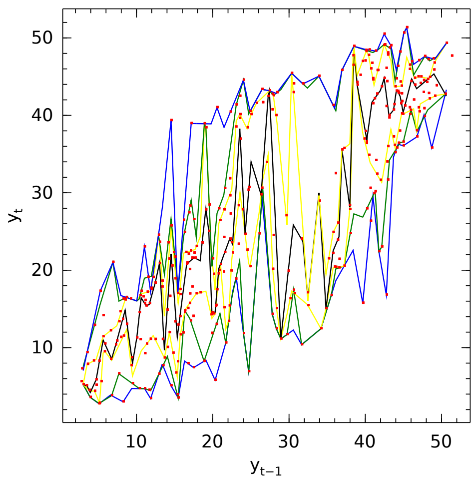
$$\delta_t^+, \delta_t^-, \xi_t \geq 0, \qquad \forall t \in \{3, \ldots, n-1\}. \tag{3.7}$$

The output of our optimization problem is a sequence of ordered points $\{(x_t, q_t)\}_{t\in T}$. The next step is to interpolate these points in order to provide an estimation for any other value of x. To address this issue, we propose using a B-splines interpolation, that will be developed in another study.

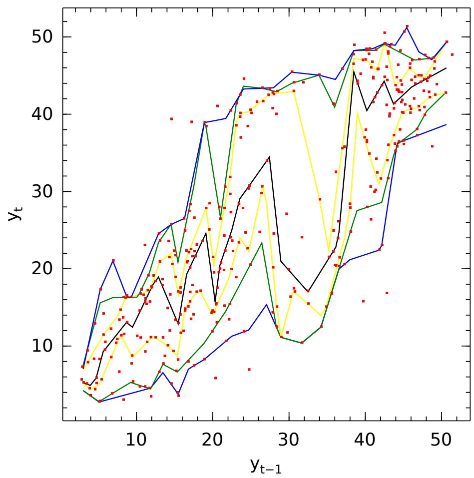
The quantile estimation is done for different values of λ. By using different levels of penalization on the second difference, the estimation can be more or less adaptive to the fluctuation. It is important to notice that the usage of the ℓ_1-norm as penalty leads to a piecewise linear solution q_t. Figure 3.1 shows the quantile estimation for a few different values of λ.

When estimating quantiles for a few different values of α, however, sometimes we find them overlapping each other, which we call crossing quantiles. This effect can be seen in figure 3.1f, where the 95%-quantile crosses over the 90%-quantile. To prevent this, we can include a non-crossing constraint:

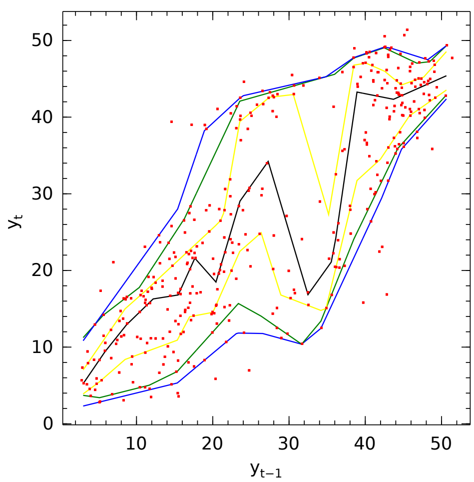
$$q_i^{\alpha} \leq q_i^{\alpha'}, \quad \forall i \in I, \alpha < \alpha'. \tag{3.8}$$



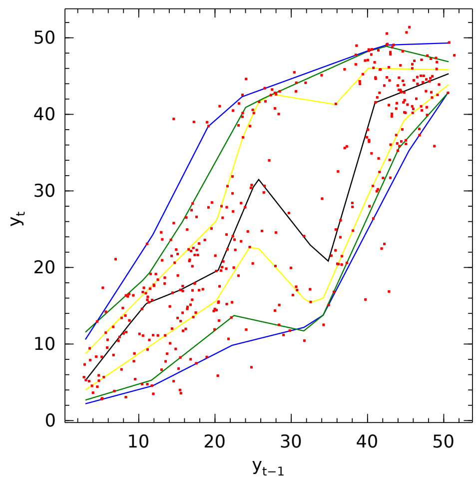
(a) $\lambda = 0.1$



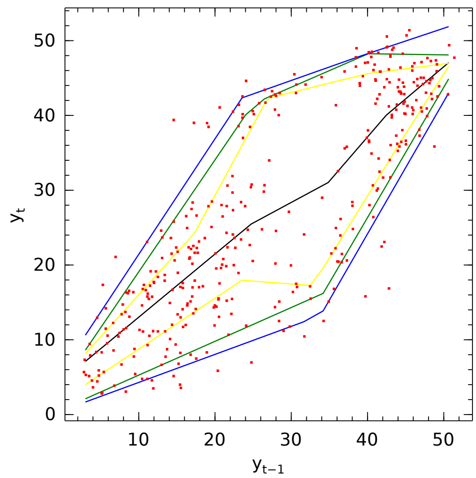
(b) $\lambda = 0.3$



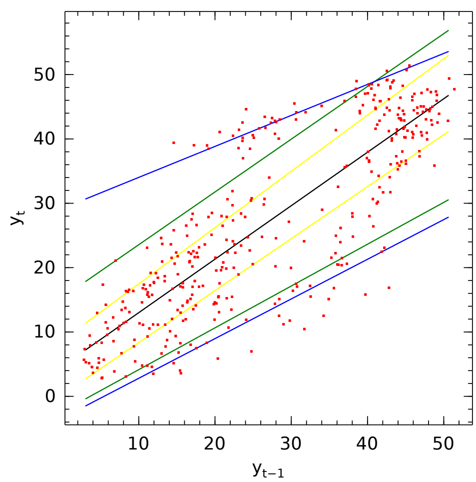
(c) $\lambda = 1$



(d) $\lambda = 3$



(e) $\lambda = 10$



(f) $\lambda = 150$

Figure 3.1: Quantile estimations for a few different values of λ. The quantiles represented here are $\alpha = (5\%, 10\%, 25\%, 50\%, 75\%, 90\%, 95\%)$. When $\lambda = 0.1$, on the upper left, we clarly see a overfitting on the estimations. The other extreme case is also shown, when $\lambda = 200$ the nonparametric estimator converges to the linear model.

This means that when α' is a higher quantile than α, then the values from the α'-quantile must be bigger than those of the α-quantile for each and every point.

As a result of this nonparametric estimation, we are able to establish a relation between y_t and y_{t-p} in a way that the model adjusts itself automatically to the present nonlinearities. For this, we only have to supply a numeric value for λ. This approach, however, have yet some issues do be discussed.

The first issue is how to select an appropriate value for λ. A simple way is to do it by inspection, which means to test many different values and pick the one that suits best our needs by looking at them. The other alternative is to use a metric to which we can select the best tune. We can achieve this by using a cross-validation method, for example.

The other issue occurs when we try to add more than one lag to the analysis at the same time. This happens because the problem solution is a set of points that we need to interpolate. This multivariate interpolation, however, is not easily solved, in the sense that we can either choose using a very naive estimator such as the K-nearest neighbors or just find another method that is not yet adopted for a wide range of applications.

4 Simulation

In this section, we investigate how to simulate future paths of the time series y_t. Let T be the total number of observations of y_t. We produce S different paths with size K for each. We have T observations of y_t and we want to simulate Given a vector of explanatory variables x_t, let q_t^α be given by the following linear model:

$$q_t^\alpha = \beta_0^\alpha + x_t^T \beta^\alpha + \varepsilon_t, \tag{4.1}$$

where β^α is a vector of coefficients for the explanatory variables. The variables chosen to compose x_t can be either exogenous variables, autoregressive components of y_t or both. As the distribution of ε_t is unknown, we have to use a nonparametric approach in order to estimate its one-step ahead density.

The coefficients β_0^α and β^α are the solution of the minimization problem given in equation 2.2, reproduced here for convenience:

$$\begin{aligned} \min_{\beta_0,\beta,\varepsilon_t^+,\varepsilon_t^-} & \sum_{t=1}^{n} \left(\alpha\varepsilon_t^+ + (1-\alpha)\varepsilon_t^- \right) \\ \text{s.t. } & \varepsilon_t^+ - \varepsilon_t^- = y_t - \beta_0 - \beta^T x_t, \qquad \forall t \in \{1,\ldots,n\}, \\ & \varepsilon_t^+, \varepsilon_t^- \geq 0, \qquad \forall t \in \{1,\ldots,n\}. \end{aligned} \tag{4.2}$$

To produce S different paths of $\{\hat{y}_t\}_{t=T+1}^{T+K}$, we use the following procedure:

Procedure for simulating S scenarios of y_t

---

1. For every quantile $\alpha_i \in (0,1)$, we use equation 4.1 to produce a forecast of $\hat{q}_{T+1}^{\alpha_i}$, as x_{T+1} is supposed to be known at time $T+1$. In the presence of exogenous variables that are unknown, it is advisable to incorporate its uncertainty by considering different scenarios. In each scenario, though, x_{T+1} must be considered fully known.

2. In any given t, by choosing many different values of α_i, we can estimate a sequence of quantiles $q_t^{\alpha_1} \leq q_t^{\alpha_2} \leq \cdots \leq q_t^{\alpha_Q}$ with $0 < \alpha_1 < \alpha_2 < \cdots < \alpha_Q < 1$. Let $F_{y_{T+1}}$ be the estimated distribution function of y_{T+1}. The process of fitting $\hat{F}_{y_{T+1}}$ is by mapping every α_i with its estimated quantile $\hat{q}^{\alpha_i}$. A problem arises for the distribution extremities, because when $\alpha = 0$ or $\alpha = 1$, the problem 4.2 becomes unbounded. In order to find good estimates for y_{T+1} when $F_{y_{T+1}}$ approaches 0 or 1, we can either use a kernel smoothing function, splines, linear approximation, or any other method. **This will be developed later.** When this sequence of chosen α_i is thin enough, we can approximate well the distribution function of y_{T+1}, as is shown in Figure 4.1. Thus, the distribution found for $\hat{y}_{T+1}$ is nonparametric, as no previous assumptions are made about its shape, and its form is fully recovered by the data we have.

3. Once we have a distribution for y_{T+1}, we can generate S different simulated values, drawn from the distribution $\hat{F}_{y_{T+1}}$ found on step 2. Let X be a random variable with uniform distribution over the interval $[0,1]$. By using results from the Probability Integral Transform, we know that the random variable $F_{y_{T+1}}^{-1}(X)$ has the same distribution as y_{T+1}. So, by drawing a sample of size S from X and applying the inverse function of $F_{y_{T+1}}$, we have our sample of size K for y_{T+1}.

4. Each one of the S different values for y_{T+1} will be the starting point of a different path. Now, for each $t \in [T+2, T+K]$ and $s \in S$, we have to estimate the quantiles $q_{t,s}^{\alpha_i}$ and find a distribution function for $\hat{F}_{y_{t,s}}$ just like it was done on step 2. Note that when $t > T+2$, every estimate will be scenario dependent, hence there will be S distribution functions estimated for each period t. From now on, in each path just one new value will be drawn randomly from the one-step ahead distribution function - as opposed to what was carried on step 3, when S values were simulated. As there will be S distribution functions - one for each path, in each period t it will be produced exact S values for y_t, one for its own path. Repeating this step until all values of t and s are simulated will give us the full simulations that we are looking for.

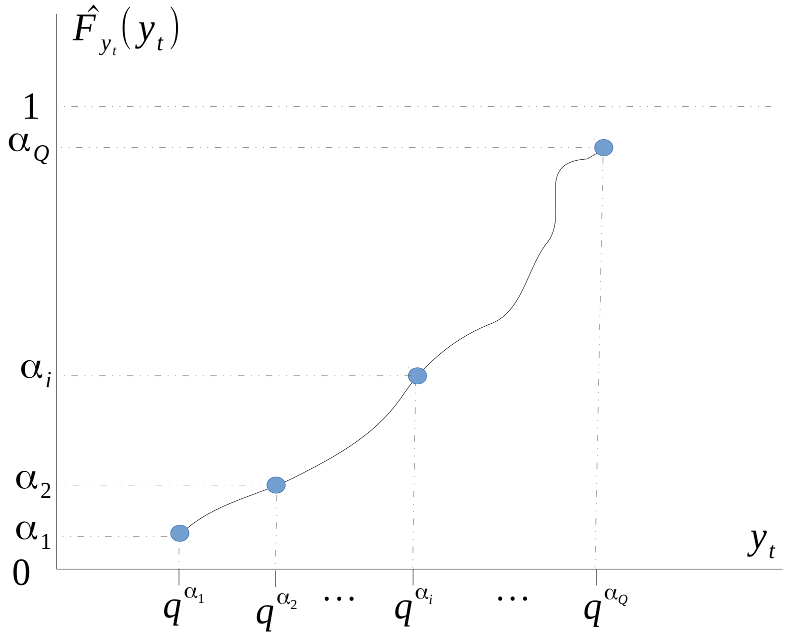


Figure 4.1: Fitting a distribution function from quantile estimations

References

[1] Dimitris Bertsimas, Angela King, and Rahul Mazumder. Best subset selection via a modern optimization lens. *arXiv preprint arXiv:1507.03133*, 2015.

[2] Seung-Jean Kim, Kwangmoo Koh, Stephen Boyd, and Dimitry Gorinevsky. ℓ_1 trend filtering. *SIAM review*, 51(2):339–360, 2009.

[3] Roger Koenker. *Quantile regression.* Number 38. Cambridge university press, 2005.

[4] Fernando Porrua, Bernardo Bezerra, Luiz Augusto Barroso, Priscila Lino, Francisco Ralston, and Mario Pereira. Wind power insertion through energy auctions in brazil. In *Power and Energy Society General Meeting, 2010 IEEE*, pages 1–8. IEEE, 2010.