Pontíficia Universidade Católica do Rio de Janeiro
Departamento de Engenharia Elétrica

Quantile Regression

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

February 22, 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 Estimating distribution function from quantile regressions

In many applications where a time series model is employed, we often consider the innovations' distribution as known. Take, for example, the AR(p) model:

$$y_t = c + \varepsilon_t + \sum_{i=1}^{p} \phi_i y_{t-1}$$

In this model, errors ε_t are assumed to have normal distribution with zero mean.

When we are dealing with meteorological time series, however, we can't always assume normality. In these cases, one can either find a distribution that has a better fit to the data or have a nonparametric method to estimate the distribution directly from the available data.

In a time series framework, where a time series y_t is given by a linear model of its regressors x_t

$$y_t = \beta^T x_t + \varepsilon_t,$$

we propose to estimate the k-step ahead distribution of y_t with a nonparametric approach. Let an empirical α-quantile $\hat{q}_\alpha \in \mathcal{Q}$ be a functional belonging to a functional space. In any given t, we can estimate the sequence of quantiles $\{q_\alpha(x_t)\}_{\alpha \in A}$ by solving the problem defined on equations (2.5)-(2.8). After evaluating this sequence, by making equal

$$\hat{Q}_{y_t|X}(\alpha) = \hat{q}_\alpha(x_t), \qquad \forall \alpha \in A, \tag{1.1}$$

we have a set of size $|A|$ of values to define the discrete function over the first argument $\hat{Q}_{y_t|x_t}(\alpha, X = x_t) : A \times \mathbb{R}^d \to \mathbb{R}$. The goal of having function $\hat{Q}$ is to use it as base to construct the estimated quantile function $\hat{Q}'_{y_t|X=x_t}(\alpha, x_t) : [0,1] \times \mathbb{R}^d \to \mathbb{R}$.

A problem arises for the distribution extremities, because when $\alpha = 0$ or $\alpha = 1$, the optimization problem becomes unbounded. In order to find values for $\hat{Q}(\alpha, x_t)$ when $\alpha \in \{0,1\}$, we chose to linearly extrapolate its values. Note that as $A \subset [0,1]$, the domain of $\hat{Q}$ is also a subset of the domain of $\hat{Q}'$. The estimative of $\hat{Q}'$ is done by interpolating points of $\hat{Q}$ over the interval $[0,1]$. Thus, the distribution found for $\hat{y}_\tau$ is nonparametric, as no previous assumptions are made about its shape, and its form is fully recovered by the data we have.

We investigate two different approaches for Q_{y_t} by the functional structure of each individual $q_\alpha(x_t)$. In section 2.1, we explore the case where the individual quantiles $q_\alpha(x_t)$ are a linear function of its arguments:

$$\hat{q}_\alpha(x_t) = \beta_{0,\alpha} + \beta_\alpha^T x_t, \tag{1.2}$$

where β^α is a vector of coefficients for the explanatory variables.

In section 2.2 we introduce a Nonparametric Quantile Autoregressive model with a ℓ_1-penalty term, in order to properly simulate densities for several α-quantiles. In this nonparametric approach we don't assume any form for $q_\alpha(x_t)$, 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. The result of this optimization problem is that each $q_\alpha(x_t)$ will be a function with finite second derivative.

In order to find good estimates for $Q_{y_t}(\alpha)$ when α approaches 0 or 1, as well as performing interpolation on the values that were not directly estimated, we can either use a kernel smoothing function, splines, linear approximation, or any other method. **This will be developed later.**

1.1 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 $q_\alpha(x_t)$ to be a linear function

$$\hat{q}_\alpha(x_t) = \beta_{0\alpha} + \beta_\alpha^T x_t \tag{1.3}$$

we can substitute it on problem 1.6, getting the following LP problem:

$$\begin{aligned} \min_{\beta_0, \beta, \varepsilon_t^+, \varepsilon_t^-} \quad & \sum_{t \in T} \left(\alpha \varepsilon_t^+ + (1-\alpha) \varepsilon_t^- \right) \\ \text{s.t. } \quad & \varepsilon_t^+ - \varepsilon_t^- = y_t - \beta_0 - \beta^T x_t, \qquad \forall t \in \{1, \dots, n\}, \\ & \varepsilon_t^+, \varepsilon_t^- \geq 0, \qquad \forall t \in \{1, \dots, n\}. \end{aligned} \tag{1.4}$$

To solve the crossing quantile issue discussed earlier, we employ a minimization problem with all quantiles at the same time and adding the non crossing constraints. The new optimization problem is shown below:

$$\min_{\beta_{0\alpha}, \beta_\alpha, \varepsilon_{t\alpha}^+, \varepsilon_{t\alpha}^-} \quad \sum_{\alpha \in A} \sum_{t \in T} \left(\alpha \varepsilon_{t,\alpha}^+ + (1-\alpha) \varepsilon_{t,\alpha}^- \right) \tag{1.5}$$

$$\text{s.t.} \quad \varepsilon_{t,\alpha}^+ - \varepsilon_{t,\alpha}^- = y_t - \beta_{0\alpha} - \beta_\alpha^T x_t, \qquad \forall t \in T_\tau, \forall \alpha \in A, \tag{1.6}$$

$$\varepsilon_{t\alpha}^+, \varepsilon_{t\alpha}^- \geq 0, \qquad \forall t \in T_\tau, \forall \alpha \in A, \tag{1.7}$$

$$q_\alpha(x_t) \leq q_{\alpha'}(x_t), \qquad \forall t \in T_\tau, \forall (\alpha, \alpha') \in A \times A, \alpha < \alpha', \tag{1.8}$$

When solving problem (2.5)-(2.8), the sequence $\{q_\alpha\}_{\alpha \in A}$ is fully defined by the values of $\beta_{0\alpha}^*$ and β_α^*, for every α.

1.2 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. To prevent this issue from occurring we propose a modification which we let the prediction $q_\alpha(x_t)$ 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].

This time, as opposed to when employing linear models, we don't suppose any functional form for $q_\alpha(x_t)$. This forces us to build each q_α differently: instead of finding a set of parameters that fully defines the function, we find a value for $q_\alpha(x_t)$ at each instant t. On the optimization problem, we will find optimal values for a variable $q_{\alpha t} \in \mathbb{R}$, each consisting of a single point. The sequence of $\{q_{\alpha t}^*\}$ will provide a discretization for the full function $\hat{q}_\alpha(x_t)$, which can be found by interpolating these points.

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

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

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 $T' = \{2, \dots, n-p-1\}$. 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:

$$\begin{aligned}\hat{q}_\alpha(x_t) = \arg\min_{q_{\alpha t}} & \sum_{t \in T'} \left(\alpha |y_t - q_{\alpha t}|^+ + (1-\alpha)|y_t - q_{\alpha t}|^- \right) \\ & + \lambda_1 \sum_{t \in T'} |D^1_{x_t} q_{\alpha t}| + \lambda_2 \sum_{t \in T'} |D^2_{x_t} q_{\alpha t}|,\end{aligned} \tag{1.9}$$

where $D^1 q_t$ and $D^2 q_t$ are the first and second derivatives of the $q_\alpha(x_t)$ function, calculated as follows:

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

$$D^1_{t\alpha} = \frac{q_{\alpha t+1} - q_{\alpha t}}{x_{t+1} - x_t}.$$

The first part on the objective function is the usual quantile regression condition for $\{q_{t\alpha}\}_{\alpha \in A}$. The second part is the ℓ_1-filter. The purpose of a filter is to control the amount of variation for our estimator $q_\alpha(x_t)$. When no penalty is employed we would always get $q_{\alpha t} = y_t$, for any given α. 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:

$$\min_{q_{\alpha t}, \delta^+_t, \delta^-_t, \xi_t} \quad \textstyle\sum_{\alpha \in A} \sum_{t \in T'} \left(\alpha \delta^+_{t\alpha} + (1-\alpha)\delta^-_{t\alpha} \right) \tag{1.10}$$

$$+\lambda_1 \textstyle\sum_{t \in T'} \gamma_{t\alpha} + \lambda_2 \sum_{t \in T'} \xi_{t\alpha}$$

$$s.t. \quad \delta^+_t - \delta^-_{t\alpha} = y_t - q_{t\alpha}, \qquad \forall t \in T', \forall \alpha \in A, \tag{1.11}$$

$$D^1_{t\alpha} = \frac{q_{\alpha t+1} - q_{\alpha t}}{x_{t+1} - x_t}, \qquad \forall t \in T', \forall \alpha \in A, \tag{1.12}$$

$$D^2_{t\alpha} = \frac{\left(\frac{q_{\alpha t+1} - q_{\alpha t}}{x_{t+1} - x_t}\right) - \left(\frac{q_{\alpha t} - q_{\alpha t-1}}{x_t - x_{t-1}}\right)}{x_{t+1} - 2x_t + x_{t-1}}. \qquad \forall t \in T', \forall \alpha \in A, \tag{1.13}$$

$$\gamma_{t\alpha} \geq D^1_{t\alpha}, \qquad \forall t \in T', \forall \alpha \in A, \tag{1.14}$$

$$\gamma_{t\alpha} \geq -D^1_{t\alpha}, \qquad \forall t \in T', \forall \alpha \in A, \tag{1.15}$$

$$\xi_{t\alpha} \geq D^2_{t\alpha}, \qquad \forall t \in T', \forall \alpha \in A, \tag{1.16}$$

$$\xi_{t\alpha} \geq -D^2_{t\alpha}, \qquad \forall t \in T', \forall \alpha \in A, \tag{1.17}$$

$$\delta^+_{t\alpha}, \delta^-_{t\alpha}, \gamma_{t\alpha}, \xi_{t\alpha} \geq 0, \qquad \forall t \in T', \forall \alpha \in A, \tag{1.18}$$

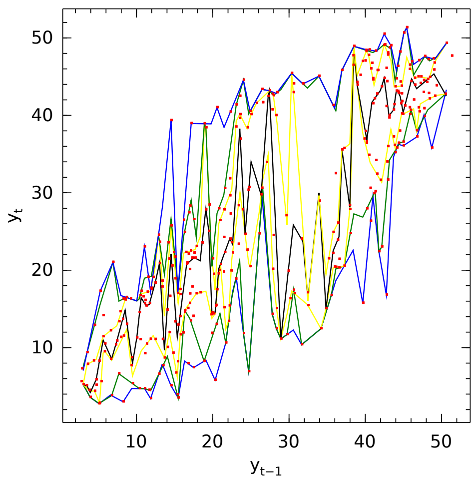
$$q_{t\alpha} \leq q_{t\alpha'}, \qquad \forall t \in T', \forall (\alpha, \alpha') \in A \times A, \alpha(1.19)$$

The output of our optimization problem is a sequence of ordered points $\{(x_t, q_{t\alpha})\}_{t \in T}$, for all $\alpha \in A$. The next step is to interpolate these points in order to provide an estimation for any other value of x_t. To address this issue, we propose using a linear interpolation, that will be developed in another study. Note that $q_{t\alpha}$ is a variable that represents only one point of the α-quantile function $q_\alpha(x_t)$.

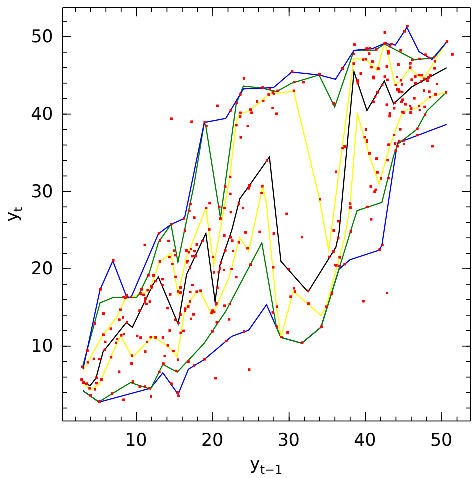
The quantile estimation is done for different values of λ. By using different levels of penalization on the second difference, the estimation can be more or less adaptive to the fluctuation. It is important to notice that the usage of the ℓ_1-norm as penalty leads to a piecewise linear solution $q_{t\alpha}$. Figure 2.1 shows the quantile estimation for a few different values of λ.

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.

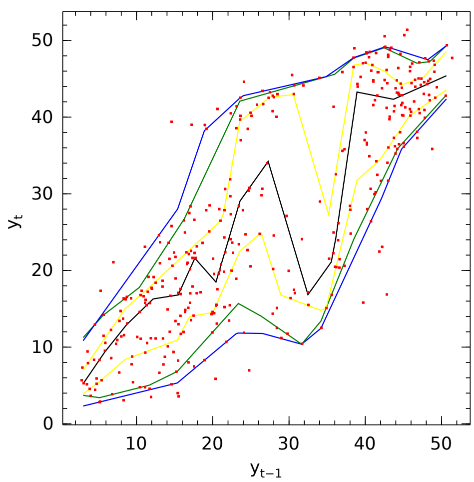
The other issue occurs when we try to add more than one lag to the analysis at the same time. This happens because the problem solution is a set of points that we need to interpolate. This multivariate interpolation, however, is not easily solved, in the sense that we can either choose using a very naive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



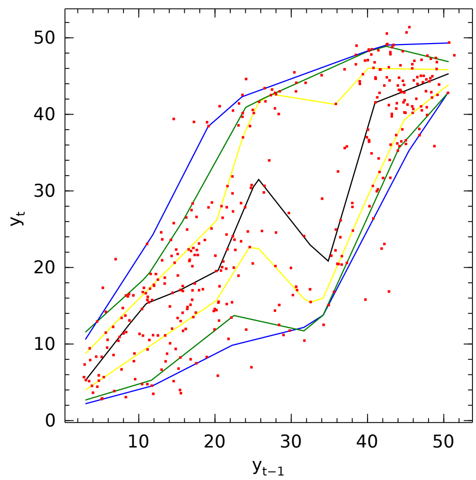
(a) $\lambda = 0.1$



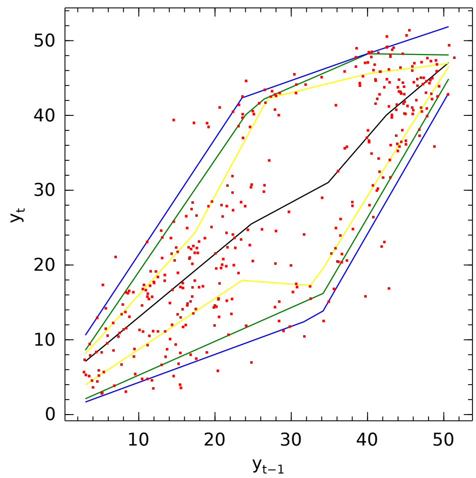
(b) $\lambda = 0.3$



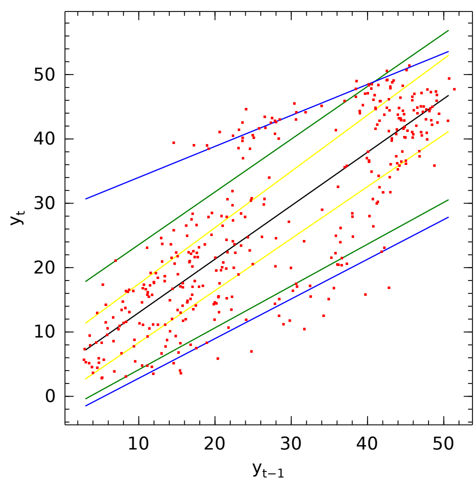
(c) $\lambda = 1$



(d) $\lambda = 3$



(e) $\lambda = 10$



(f) $\lambda = 150$

Figure 1.1: 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.

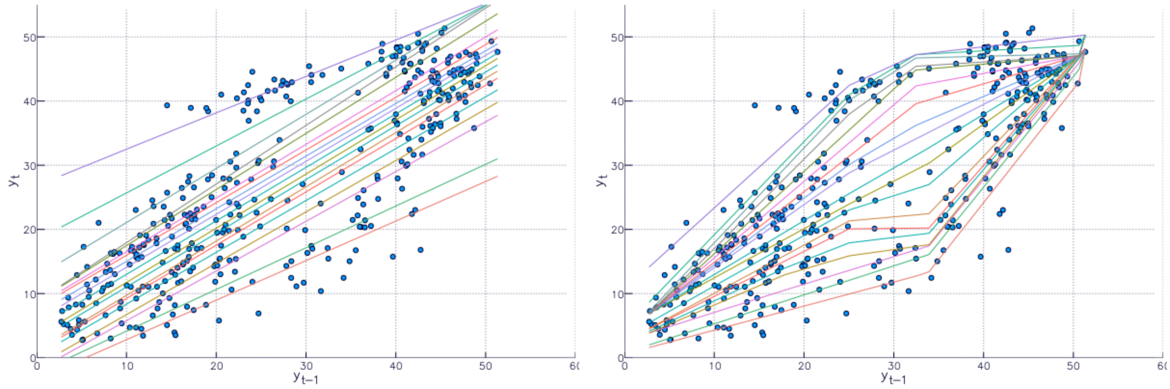


Figure 1.2: Estimated α-quantiles. On the left using a linear model and using a nonparametric approach on the right.

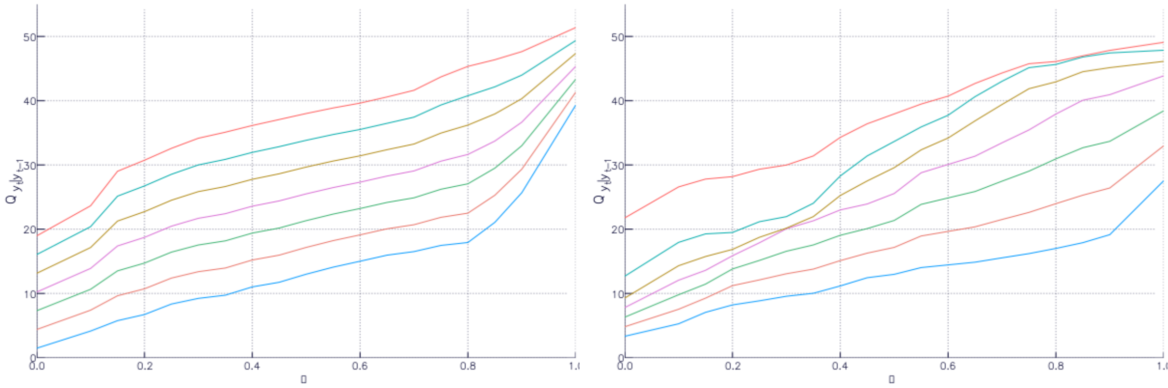


Figure 1.3: Estimated quantile functions, for different values of y_{t-1}. On the left using a linear model and using a nonparametric approach on the right.

1.3 A comparison between both approaches

The last two sections introduced two different strategies to arrive in a Quantile Function $Q_{y_t|X}$. But what are the differences between using one method or the other?

To provide a comparison between both approaches, we estimate a quantile function to predict the one-step ahead quantile function. We use as explanatory variable only the last observation y_{t-1}. Both

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.