Nongaussian time series model via Quantile Regression

Marcelo Ruas and Alexandre Street, *Member, IEEE*

Abstract

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

Keywords

Quantile Regression, Model Identification, Non-gaussian time series model

TODO LIST

I. INTRODUCTION

Renewable energy power is an emergent topic which is demanding attention from the academic community. The installed capacity of renewable energy plants has been increasing in a fast pace and projections point out that wind power alone will account to 18% of global power by 2050 [1]. In spite of its virtues, several new challenges are inherent when dealing with such power source. New statistical models capable to handle with such difficulties are an emerging field in power systems literature . The main objective in such literature is to propose new models capable of generating scenarios of renewable energy source which are demanded in (i) energy trading, (ii) unit commitment, (iii) grid expansion planning, and (iv) investment decisions (see ([2]–[5]) and references therein). To provide good scenarios from an array of potential influential factors, one has to properly select which features are relevant and create a good model for the conditional distribution. Notwithstanding, a little attention is devoted to addressing both at the same time.

Conventional statistical models are often focused on estimating the conditional mean of a given random variable. By reducing the outcome to a single statistic, we loose important informations about the series random behavior. In order to account for the process inherent variability we focus our work on probability forecasting. [6] reviews the commonly used methodologies regarding probabilistic forecasting models, splitting them in parametric and nonparametric classes. Main characteristics of **parametric models** are (i) assuming a distribution shape and (ii) low computational costs. ARIMA-GARCH, for example, model the renewable series by assuming the distribution *a priori*. On the other hand, **nonparametric models** (i) don't require a distribution to be specified, (ii) needs mode data to produce a good approximation and (iii) have a higher computational cost. Popular methods are Quantile Regression (QR), Kernel Density Estimation, Artificial Intelligence or a mix of them.

Most time series methods rely on the assumption of Gaussian errors. However, renewable series such as wind and solar are reported as non-Gaussian [7]–[10]. To circumvent this problem, the usage of nonparametric methods, which doesn't rely on assuming any previously assumed distribution of errors, is adequate. We choose to use Quantile Regression (QR) as a tool for constructing a methodology for non-Gaussian error time series, because of its facility to implement on commercial solvers and to extend the original model. However, when estimating a distribution function, as each quantile is estimated independently, the monotonicity of the distribution function may be violated. To get around this issue (known as crossing-quantiles) we propose to either add a constraint on the optimization model (which is more computationally intensive) or making a transformation as in [11], which can then be estimated independently.

The seminal work [12] defines QR as we use today. By this formulation, the conditional quantile is the solution of an optimization problem where we minimize the sum of the check function (defined formally in the next session). Instead of using the classical regression to estimate the conditional mean, the QR determines any quantile from the conditional distribution. Applications are enormous, ranging from risk measuring at financial funds (the Value-at-Risk) to a central measure robust to outliers. By estimating many quantiles on a thin grid of probabilities, one can have as many points as desired from the estimated conditional distribution function. In [13], the application of QR is extended to time series, when the covariates are lagged values of y_t. In our work, beyond autoregressive terms, it is also considered other exogenous variables as covariates.

In [14]–[18], QR is employed to model the conditional distribution of Wind Power Time Series. An updating quantile regression model is presented by [15]. The authors present a modified version of the simplex algorithm to incorporate new observations without restarting the optimization procedure. In [16], the authors build a quantile model from already existent independent Wind Power forecasts. The approach by [14] is to use QR with a nonparametric methodology. The authors add a penalty term based on the Reproducing Kernel Hilbert Space, which allows a nonlinear relationship between the explanatory variables and the output. This paper also develops an on-line learning technique, where the model is easily updated after each new observation. In [18], wind power probabilistic forecasts are made by using QR with a special type of Neural Network (NN) with one hidden layer, called extreme learning machine. In this setup, each quantile is a different linear combination of the features of the hidden layer. The authors of [19] use the weighted Nadaraya-Watson to estimate the conditional function in the time series.

Regularization is a topic already explored in previous QR papers. The work by [20] defines the proprieties and convergence rates for QR when adding a penalty proportional to the ℓ_1-norm to perform variable selection, using the same idea as the LASSO [21]. The ADALASSO equivalent to QR is proposed by [22]. In this variant, the penalty for each variable has a different weight, and this modification ensures that the oracle propriety is being respected.

For the best of the authors knowledge, no other work has developed a methodology where regularization and estimation of the conditional distribution using QR is carried on at the same time. We propose to attack both problems simultaneously by using either Mixed Integer Linear Programming (MILP) or a LASSO penalization. On the LASSO formulation, regularization is performed for an individual quantile as described in [20], with the difference that all quantiles are estimated at the same time. In [23], the best subset with size K is selected by solving a MILP problem to minimize the sum of squared errors. The idea is straightforward: integer variables are used to count whether a variable is included or not in the model; a total number of K variables is allowed. Model selection for QR is performed using this same approach. The advantage we highlight on using the latter methodology is that the solution provided is optimal in the sense of minimizing the check function for a given number K of variables.

The objective of this paper is, then, to propose a new methodology to address nonparametric time-series focused on renewable energy. In our analysis, we develop both nonlinear and linear models for QR. The main contributions are:

- A nonparametric methodology to model the conditional distribution of time series.
- On the linear case, we propose a parsimonious methodology that selects the global optimal solution.
- Regularization techniques applied to an ensemble of quantile functions to estimate the conditional distribution.

The remaining of the paper is organized as follows. In section II, we present both the linear parametric and the nonlinear QR based time series models. In section III, we discuss the estimation procedures for them. The regularization strategies are also presented on this section. Finally, in section IV, a case study using real data from both solar and wind power is presented in order to test our methodology. Section V will conclude this article.

II. QUANTILE REGRESSION BASED TIME SERIES MODEL

Let the conditional quantile function $Q_{Y|X=x} : [0,1] \times \mathbb{R}^d \to \mathbb{R}$ (in short, from now on, $Q_{Y|X}(\cdot,\cdot)$) be defined as

$$Q_{Y|X}(\alpha, x) = F^{-1}_{Y|X}(\alpha, x) = \inf\{y : F_{Y|X}(y, x) \geq \alpha\}, \tag{1}$$

where Y is a random variable and X a random vector of dimension d. Let a dataset be composed of n observations of $\{y_t, x_t\}_{t=1}^{n}$. The sample quantile function is based on a finite number of observations and is the solution to the following optimization problem:

$$\hat{Q}_{Y|X}(\alpha, \cdot) \quad \in \quad \underset{q_\alpha(\cdot)}{\arg\min} \quad \sum_{t \in T} \rho_\alpha(y_t - q_\alpha(\cdot)), \tag{2}$$

$$q_\alpha \in \mathcal{Q}. \tag{3}$$

where ρ is the check function, defined as

$$\rho_\alpha(x) = \begin{cases} \alpha x & \text{if } x \geq 0 \\ (1-\alpha)x & \text{if } x < 0 \end{cases}. \tag{4}$$

Quantile q_α belongs to a function space $\mathcal{Q}$. We might have different assumptions for space $\mathcal{Q}$, depending on the type of function we want to find for q_α. A few properties, however, must be achieved by our choice of space, such as being continuous and having limited first derivative. In this paper, we consider the case where $\mathcal{Q}$ is a linear function's space.

The problem (2)-(3) can be rewritten as a Linear Programming problem as in (5)-(9), thus being able to use a modern solver to fit our model. Variables ε_t^+ e ε_t^- represent the quantities $|y - q(\cdot)|^+$ and $|y - q(\cdot)|^-$, respectively. A is the set containing a sequence of probabilities α_i such that $0 < \alpha_1 < \alpha_2 < \cdots < \alpha_Q < 1$. This set represents a finite discretization of the interval $[0,1]$.

$$\min_{\beta_{0\alpha}, \beta_\alpha, \varepsilon^+_{t\alpha}, \varepsilon^-_{t\alpha}} \sum_{\alpha \in A} \sum_{t \in T} \left(\alpha \varepsilon^+_{t\alpha} + (1-\alpha) \varepsilon^-_{t\alpha} \right) \tag{5}$$

$$\text{s.t.} \tag{6}$$

$$\varepsilon^+_{t\alpha} - \varepsilon^-_{t\alpha} = y_t - \beta_{0\alpha} - \beta_\alpha^T x_t, \qquad \forall t \in T, \forall \alpha \in A, \tag{7}$$

$$\varepsilon^+_{t\alpha}, \varepsilon^-_{t\alpha} \geq 0, \qquad \forall t \in T, \forall \alpha \in A, \tag{8}$$

$$\beta_{0\alpha} + \beta_\alpha^T x_t \leq \beta_{0\alpha'} + \beta_{\alpha'}^T x_t, \qquad \forall t \in T, \forall (\alpha, \alpha') \in A \times A, \alpha < \alpha', \tag{9}$$

After solving the problem, the sequence $\{q_\alpha\}_{\alpha \in A}$ is fully defined by the optimum values $\beta^*_{0\alpha}$ and β^*_α, for every α.

We apply QR to estimate the conditional distribution $\hat{Q}_{Y_{t+k}|X_t, Y_t, Y_{t-1}, \ldots}(\alpha, \cdot)$ for a k-step ahead forecast, where X_t is the

III. Regularization

When dealing with many candidates to use as covariates, one has to deal with the problem of selecting a subset of variables to use in constructing the model. This means that the vector of coefficients $\beta_\alpha = [\beta_{1\alpha} \cdots \beta_{P\alpha}]$ should not have all nonzero values. There are many ways of selecting a subset of variables among the available options. Classical approaches for this problem are the Stepwise algorithm [24], [25], [21], which includes variables in sequence.

Two approaches will be employed. At first, we use a Mixed Integer Linear Programming optimization problem (MILP) to find the best subset among all choices of covariates. The second way is by using a LASSO-type technique, which consists in penalizing the ℓ_1-norm of regressors, thus shrinking the size of estimated coefficients towards zero.

A. Best subset selection with MILP

In this part, we investigate the usage of MILP to select which variables are included in the model, by using a constraint which limits them to a number of K. Only K coefficients $\beta_{p\alpha}$ may have nonzero values, for each α. Binary variable $z_{p\alpha}$ indicates whether $\beta_{p\alpha}$ has a nonzero value. The optimization problem that incorporates this idea is described below:

$$
\begin{aligned}
\min_{\beta_{0\alpha},\beta_\alpha,z_{p\alpha},\varepsilon^+_{t\alpha},\varepsilon^-_{t\alpha}} \quad & \textstyle\sum_{\alpha\in A}\sum_{t\in T}\left(\alpha\varepsilon^+_{t\alpha} + (1-\alpha)\varepsilon^-_{t\alpha}\right) && && (10)\\
\text{s.t} \quad & \varepsilon^+_{t\alpha} - \varepsilon^-_{t\alpha} = y_t - \beta_{0\alpha} - \textstyle\sum_{p=1}^{P}\beta_{p\alpha}x_{t,p}, && \forall t\in T, \forall\alpha\in A, && (11)\\
& \varepsilon^+_{t\alpha}, \varepsilon^-_{t\alpha} \geq 0, && \forall t\in T, \forall\alpha\in A, && (12)\\
& -Mz_{p\alpha} \leq \beta_{p\alpha} \leq Mz_{p\alpha}, && \forall\alpha\in A, \forall p\in P, && (13)\\
& \textstyle\sum_{p=1}^{P} z_{p\alpha} \leq K, && \forall\alpha\in A, && (14)\\
& z_{p\alpha} \in \{0,1\}, && \forall\alpha\in A, \forall p\in P, && (15)\\
& \beta_{0\alpha} + \beta_\alpha^T x_t \leq \beta_{0\alpha'} + \beta_{\alpha'}^T x_t, && \forall t\in T, \forall(\alpha,\alpha')\in A\times A, \alpha<\alpha', && (16)
\end{aligned}
$$

ajeitar equação para duas colunas

The objective function and constraints (11), (12) and (16) are those from the standard linear quantile regression. By constraint (13), variable $z_{p\alpha}$ is a binary that assumes 1 when coefficient $\beta_{p\alpha}$ is included, while (14) guarantees that at most K of them are nonzero. The value of M is chosen in order to guarantee that $M \geq \|\hat{\beta}_\alpha\|_\infty$. The solution given by $\beta^*_{0\alpha}$ and $\beta^*_\alpha = [\beta^*_{1\alpha} \cdots \beta^*_{P\alpha}]$ will be the best linear α-quantile regression with K nonzero coefficients.

We ran this optimization on the Icaraizinho dataset for each value of $K \in \{0, 1, \ldots, 12\}$ and quantiles $\alpha \in \{0.05, 0.1, 0.5, 0.9, 0.95\}$. The full results table can be accessed on section **??**. For all tested α-quantiles the 12[th] lag was the one included when $K = 1$. When $K = 2$, the 1[st] lag was included for all values of α, sometimes with β_{12}, some others with β_4 and once with β_{11}. These 4 lags that were present until now are the only ones selected when $K = 3$. For $K = 4$, those same four lags were selected for three quantiles (0.05, 0.1 and 0.5), but for the others (0.9 and 0.95) we have β_6, β_7 and β_9 also as selected. From now on, the inclusion of more lags represent a lower increase in the fit of the quantile regression. The estimated coefficient values for all K's are available in the appendices section.

Defining groups for variables: Consider the optimization problem defined on (10)-(16). The choice of the best subset is independent for different values of α. This means that the best subset may include two completely different sets of regressors for two probabilities α and α' close to each other. Take $K = 2$ for the example, selecting $\beta_{1\alpha}$ and $\beta_{4\alpha}$ for α while $\beta_{2\alpha'}$ and $\beta_{5\alpha'}$ is possible, but unlikely to be true.

To address this issue, we propose to divide all $\alpha \in A$ in groups. The collection G of all groups g form a partition of A, and each α belongs to exactly one group g. The subset of selected covariates must be the same for all α in the same group g. To model these properties as constraints on problem (10)-(16), we substitute constraint (13) for the following equations:

ajeitar equação para duas colunas

$$z_{p\alpha g} := 2 - (1 - z_{pg}) - I_{g\alpha} \tag{17}$$

$$\sum_{g \in G} I_{g\alpha} = 1, \forall \alpha \in A, \tag{18}$$

$$-M z_{p\alpha g} \leq \beta_{p\alpha g} \leq M z_{p\alpha g}, \forall p \in P, \forall \alpha \in A, \forall g \in G, \tag{19}$$

$$I_{g\alpha}, z_{pg} \in \{0, 1\}, \qquad \forall p \in P, \forall g \in G, \tag{20}$$

on problem (10)-(16). where G is a set of group index and z_{pg} is a binary variable that equals 1 iff covariate p is included on group g and $I_{g\alpha}$ equals 1 iff probability α belongs to group g. Constraint (19) forces that

$$\text{if } z_{pg} = 0 \text{ and } I_{g\alpha} = 1 \text{ then } \beta_{p\alpha=0}.$$

Hence, if covariate p belongs to group g, this covariate is not among group's g subset of variables, than its coefficient must be equal to 0, for that α. Note that variable $z_{p\alpha}$ behaves differently than when we are not considering groups. This means that if probability α belongs to group g but variable p is not selected to be among the ones of group g, than $\beta_{p\alpha}$ is zero. Equation (17) defines $z_{p\alpha}$ to simplify writing.

B. Best subset selection with LASSO

Another way of doing regularization is including the ℓ_1-norm of the coefficients on the objective function. The advantage of this method is that coefficients are shrunk towards zero by changing a continuous parameter λ, which penalizes the size of the ℓ_1-norm. When the value of λ gets bigger, fewer variables are selected to be used. This is the same strategy of the LASSO methodology, and its usage for the quantile regression is discussed in [26]. The proposed optimization problem to be solved is:

$$\min_{\beta_{0\alpha},\beta_\alpha} \sum_{t\in T} \alpha|y_t - q_\alpha(x_t)|^+ + \sum_{t\in T}(1-\alpha)|y_t - q_\alpha(x_t)|^- + \lambda\|\beta_\alpha\|_1, \tag{21}$$

$$q_\alpha(x_t) = \beta_0 - \sum_{p=1}^{P} \beta_p x_{t,p}.$$

For such estimation to be coherent, however, each covariate must have the same relative weight in comparison with one another. So, before solving the optimization problem, we perform a linear transformation such that all variables have mean $\mu = 0$ and variance $\sigma^2 = 1$. We apply the transformation $\tilde{x}_{t,p} = (x_{t,p} - \bar{x}_{t,p})/\hat{\sigma}_{x_{t,p}}$, where $\bar{x}_{t,p}$ and $\hat{\sigma}_{x_{t,p}}$ are respectively the sample's unconditional mean and standard deviation. The $\tilde{y}_{t-p,i}$ series will be used to estimate the coefficients, as this series has the desired properties.

After the process of normalization, we can rewrite problem 21 as a LP problem, as shown below:

$$\tilde{\beta}_\lambda^{*LASSO} = \underset{\beta_0,\beta,\varepsilon_{t\alpha}^+,\varepsilon_{t\alpha}^-}{\arg\min} \sum_{\alpha\in A}\sum_{t\in T}\left(\alpha\varepsilon_{t\alpha}^+ + (1-\alpha)\varepsilon_{t\alpha}^-\right) + \lambda\sum_{p=1}^{P}\xi_{p\alpha} \tag{22}$$

$$\text{s.t.} \tag{23}$$

$$\varepsilon_{t\alpha}^+ - \varepsilon_{t\alpha}^- = y_t - \beta_{0\alpha} - \sum_{p=1}^{P}\beta_{p\alpha}\tilde{x}_{t,p}, \qquad \forall t\in T, \forall\alpha\in A, \tag{24}$$

$$\varepsilon_{t\alpha}^+, \varepsilon_{t\alpha}^- \geq 0, \qquad \forall t\in T, \forall\alpha\in A, \tag{25}$$

$$\xi_{p\alpha} \geq \beta_{p\alpha}, \qquad \forall p\in P, \forall\alpha\in A, \tag{26}$$

$$\xi_{p\alpha} \geq -\beta_{p\alpha}, \qquad \forall p\in P, \forall\alpha\in A. \tag{27}$$

This model is built upon the standard linear programming model for the quantile regression (5)-(9). On the above formulation, the ℓ_1 norm of equation (21) is substituted by the sum of ξ_p, which represents the absolute value of $\beta_{p\alpha}$. The link between variables ξ_p and $\beta_{p\alpha}$ is made by constraints (26) and (27). Note that the linear coefficient $\beta_{0\alpha}$ is not included in the penalization, as the sum of penalties on the objective function 22.

For low values of λ, the penalty over the size of coefficients is small. Because of that, the output of problem (22)-(27) is a model where most coefficients have nonzero value. On the other hand, when the penalty on $\|\beta_\alpha\|_1$ is big, many covariates will have zero valued coefficients. When λ approaches infinity, one has a constant model. For instance, the penalty isn't applied to the linear coefficient $\beta_{0\alpha}$.

Even though we have coefficients that are estimated by this method, we don't use them directly. In fact, the LASSO coefficients are biased, so it is employed only as a variable selector. As so, the nonzero coefficient covariates will be the input of a unrestricted quantile regression problem, as in the linear programming problem (5)-(9). The set of selected indexes are given by

$$L_\lambda = \{p \in \{1, \ldots, P\} | \, |\beta^{*LASSO}_{\lambda,p}| \neq 0\}.$$

Hence, we have that, for each $p \in \{1, \ldots, P\}$,

$$\beta^{*LASSO}_{\lambda,p} = 0 \Longrightarrow \beta^{*}_{\lambda,p} = 0.$$

The post-lasso coefficients β^*_λ are the solution from the optimization problem given below:

$$\begin{aligned} (obj^*_\lambda, \beta^*_\lambda) \stackrel{(obj,var)}{\longleftarrow} & \min_{\beta_0,\beta,\varepsilon^+_t,\varepsilon^-_t} \sum_{t \in T} \left(\alpha \varepsilon^+_t + (1-\alpha)\varepsilon^-_t \right) \\ & \text{s.t. } \varepsilon^+_t - \varepsilon^-_t = y_t - \beta_0 - \sum_{p \in L_\lambda} \beta_p x_{t,p}, \quad \forall t \in T, \\ & \varepsilon^+_t, \varepsilon^-_t \geq 0, \qquad\qquad\qquad\qquad \forall t \in T. \end{aligned} \tag{28}$$

The variable obj^*_λ receives the value of the objective function on its optimal solution. In summary, the optimization in equation 21 acts as a variable selection for the subsequent estimation, which is normally called the post-LASSO estimation [27].

For the same quantile values, α, we experimented on section III-A ($\alpha \in \{0.05, 0.1, 0.5, 0.9, 0.95\}$), we estimate the post-LASSO (from now on, we call it just LASSO, for simplicity). Figure **??** shows the path of variables for each α-quantile. On the x-axis, we have the penalty λ in a log scale. On the y-axis we have the size of coefficients. One can see how increasing λ leads to a shrinking on the size of coefficients, up to a point where all coefficients are equal to 0.

C. Model selection distance

On sections III-A and III-B, we presented two ways of doing regularization. Nonetheless, regularization can be done with different levels of parsimony. For example, one can select a different number K of variables to be included in the best subset selection via MILP or choose different values of λ for the ℓ_1 penalty.

Solving a LP problem is many times faster than a similar-sized MILP problem. One of our goals is to test how much the LASSO approach gets close to the solution provided when solving the MILP problem. It would be interesting, then, if we could use a faster method that would provide a solution close to the best. To test how far are the solution given by both methods, we propose an experiment that is described as follows. Then, for each number K of total nonzero coefficients, there will be a penalty

λ_K^* which minimizes the errors from the quantile regression's objective function (given on equation (28)):

$$\lambda_K^* = \arg\min_{\lambda} \{ obj_\lambda^* \quad | \ \|\beta_\lambda^*\|_0 = K \}, \tag{29}$$

where the quantity $\|\beta_\lambda^*\|_0$ is the 0-norm, which gives the total of nonzero coefficients, for a given lambda of the LASSO estimations.

We, then, define the sets L_K^{LASSO} and L_K^{MILP}, which contains all nonzero indexes, for a given K, when using methods LASSO and MILP for regularization, respectively. Thus, we can compare the best LASSO fit where exactly K variables are selected with the best fit given by the MILP problem, also with K variables selected.

As the MILP solution is the exact solution for the problem, while the LASSO solution is an approximation, we use the former as a *benchmarking* for the quality of the latter solution. To help us view the difference of results between both methods, we define a similarity metric d between the subset of coefficients chosen by each one of them. It is desirable that the LASSO solution be as related with the MILP solution as possible. The similarity is calculated as the solution of the following optimization problem

$$d(\beta^*_{MILP(K)}, \beta^*_{\lambda_K^*}) = \min_{0 \leq \delta_{ij} \leq 1} \quad \sum_{i,j=1}^{K} \delta_{ij}(1 - |\rho_{ij}|) \tag{30}$$

$$\text{s.t.} \quad \sum_{j=1}^{K} \delta_{ij} = 1, \qquad i = 1, \dots, K, \tag{31}$$

$$\sum_{i=1}^{K} \delta_{ij} = 1, \qquad j = 1, \dots, K, \tag{32}$$

where ρ_{ij} is the correlation between the i-th and j-th independent variables in sets L_k^{MILP} and L_k^{LASSO}, respectively. The optimal value for the decision variables of this problem provides us with an assignment between selected covariates from both methods, namely, MILP and LASSO, that minimizes the overall "index of uncorrelation" between selected covariates. If $\delta^*_{ij} = 1$, the i-th selected variable in L_k^{MILP} is associated with the j-th variable in L_k^{LASSO}. For instance, if $d(\beta^*_{MILP(K)}, \beta^*_{\lambda_K^*}) = 0$, it means that there are K perfectly correlated pair of variables, despite not being the same.

As seen before, we have a best solution for each desired K. The question that arises now is how to select the ideal number of variables to use. One way of achieving this is by using an information criteria to guide our decision. An information criteria summarizes two aspects. One of them refers to how well the model fits the in-sample observations. The other part penalizes the quantity of covariates used in the model. By penalizing how big our model is, we prevent overfitting from happening. So, in order for a covariate to be included in the model, it must supply enough goodness of fit. In [28], it is presented a variation of the Schwarz criteria for M-estimators that includes quantile regression. The Schwarz Information Criteria (SIC), adapted to the quantile autoregression case, is presented below:

$$SIC(m) = n \log(\hat{\sigma}^*) + \frac{1}{2} K \log n, \tag{33}$$

where K is the model's dimension. This procedure leads to a consistent model selection if the model is well specified.

Figure **??** shows the results of these experiments for quantiles $\alpha \in \{0.05, 0.1, 0.5, 0.9, 0.95\}$. The results point us that for small values of K the distance between coefficients is bigger and where we observe the biggest differences between the SIC values. In this experiment, the minimum SIC value for the MILP problem is usually found between 4 and 6 variables in the model.

D. Time-series cross validation

Explain the necessity of CV; explain algorithm

Cross-validation (CV) is a technique used to have an estimative of the model's quality of prediction in an independent testing set. The best model that minimizes the CV error is the model which presumably will have the best performance on out of sample data.

The usage of CV is not straightforward when data is dependent, which is the case when working with time series. As the data is time dependent, one can be interested in using either all observations or to take the dependency away. The works [29], [30], we test K-fold CV and K-fold with non-dependent data. Both schemes are shown of Figure 1. This mimics real applications better even by dropping in a few times the number of observations. Will use a growing window in a 5-fold scheme, as described in figure

Criar novas figuras (R/grafico-cv.r) e citar todas as formas de CV

References

[1] International energy agency. [Online]. Available: https://www.iea.org/newsroom/news/2013/october/wind-power-seen-generating-up-to-18-of-global-power-by-2050.html

[2] A. Moreira, D. Pozo, A. Street, and E. Sauma, "Reliable renewable generation and transmission expansion planning: Co-optimizing system's resources for meeting renewable targets," *IEEE Transactions on Power Systems*, vol. PP, no. 99, pp. 1–1, 2016.

[3] R. Jabr, "Robust transmission network expansion planning with uncertain renewable generation and loads," *IEEE Transactions on Power Systems*, vol. 28, no. 4, pp. 4558–4567, 2013.

[4] C. Zhao and Y. Guan, "Data-driven stochastic unit commitment for integrating wind generation," *IEEE Transactions on Power Systems*, vol. 31, no. 4, pp. 2587–2596, July 2016.

[5] A. C. Passos, A. Street, and L. A. Barroso, "A dynamic real option-based investment model for renewable energy portfolios," *IEEE Transactions on Power Systems*, vol. 32, no. 2, pp. 883–895, March 2017.

[6] Y. Zhang, J. Wang, and X. Wang, "Review on probabilistic forecasting of wind power generation," *Renewable and Sustainable Energy Reviews*, vol. 32, pp. 255–270, Apr. 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1364032114000446

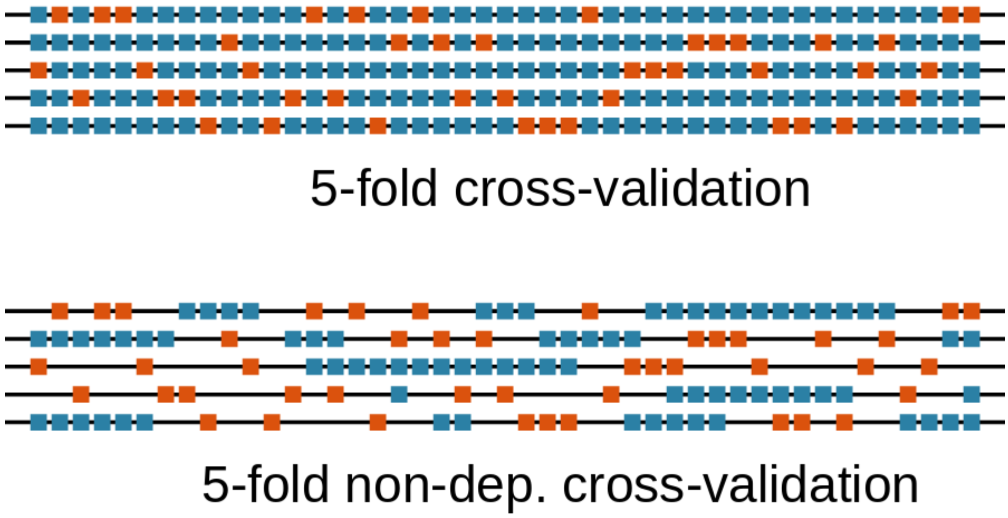

Fig. 1. K-fold CV and K-fold with non-dependent data. Observations in blue are used to estimation and in orange for evaluation. Note that non-dependent data doesn't use all dataset in each fold.

[7] R. J. Bessa, V. Miranda, A. Botterud, J. Wang, and M. Constantinescu, "Time adaptive conditional kernel density estimation for wind power forecasting," *Sustainable Energy, IEEE Transactions on*, vol. 3, no. 4, pp. 660–669, 2012.

[8] J. Jeon and J. W. Taylor, "Using conditional kernel density estimation for wind power density forecasting," *Journal of the American Statistical Association*, vol. 107, no. 497, pp. 66–79, 2012.

[9] J. W. Taylor and J. Jeon, "Forecasting wind power quantiles using conditional kernel estimation," *Renewable Energy*, vol. 80, pp. 370–379, 2015.

[10] C. Wan, J. Lin, J. Wang, Y. Song, and Z. Y. Dong, "Direct quantile regression for nonparametric probabilistic forecasting of wind power generation," *IEEE Transactions on Power Systems*, vol. 32, no. 4, pp. 2767–2778, July 2017.

[11] V. Chernozhukov, I. Fernández-Val, and A. Galichon, "Quantile and Probability Curves Without Crossing," *Econometrica*, vol. 78, no. 3, pp. 1093–1125, May 2010. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.3982/ECTA7880/abstract

[12] R. Koenker and G. Bassett Jr, "Regression quantiles," *Econometrica: journal of the Econometric Society*, pp. 33–50, 1978.

[13] R. Koenker, Z. Xiao, J. Fan, Y. Fan, M. Knight, M. Hallin, B. J. M. Werker, C. M. Hafner, O. B. Linton, and P. M. Robinson, "Quantile Autoregression [with Comments, Rejoinder]," *Journal of the American Statistical Association*, vol. 101, no. 475, pp. 980–1006, 2006. [Online]. Available: http://www.jstor.org/stable/27590777

[14] C. Gallego-Castillo, R. Bessa, L. Cavalcante, and O. Lopez-Garcia, "On-line quantile regression in the rkhs (reproducing kernel hilbert space) for operational probabilistic forecasting of wind power," *Energy*, vol. 113, pp. 355–365, 2016.

[15] J. K. Møller, H. A. Nielsen, and H. Madsen, "Time-adaptive quantile regression," *Computational Statistics & Data Analysis*, vol. 52, no. 3, pp. 1292–1303, Jan. 2008. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167947307002502

[16] H. A. Nielsen, H. Madsen, and T. S. Nielsen, "Using quantile regression to extend an existing wind power forecasting system with probabilistic forecasts," *Wind Energy*, vol. 9, no. 1-2, pp. 95–108, 2006.

[17] J. B. Bremnes, "Probabilistic wind power forecasts using local quantile regression," *Wind Energy*, vol. 7,

no. 1, pp. 47–54, Jan. 2004. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1002/we.107/abstract

[18] C. Wan, J. Lin, J. Wang, Y. Song, and Z. Y. Dong, "Direct Quantile Regression for Nonparametric Probabilistic Forecasting of Wind Power Generation," *IEEE Transactions on Power Systems*, vol. 32, no. 4, pp. 2767–2778, Jul. 2017.

[19] Z. Cai, "Regression Quantiles for Time Series," *Econometric Theory*, vol. 18, no. 1, pp. 169–192, 2002. [Online]. Available: http://www.jstor.org/stable/3533031

[20] A. Belloni and V. Chernozhukov, "L1-Penalized Quantile Regression in High-Dimensional Sparse Models," *arXiv:0904.2931 [math, stat]*, Apr. 2009, arXiv: 0904.2931. [Online]. Available: http://arxiv.org/abs/0904.2931

[21] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[22] G. Ciuperca, "Adaptive LASSO model selection in a multiphase quantile regression," *Statistics*, vol. 50, no. 5, pp. 1100–1131, Sep. 2016. [Online]. Available: http://dx.doi.org/10.1080/02331888.2016.1151427

[23] D. Bertsimas, A. King, and R. Mazumder, "Best Subset Selection via a Modern Optimization Lens," *arXiv:1507.03133 [math, stat]*, Jul. 2015, arXiv: 1507.03133. [Online]. Available: http://arxiv.org/abs/1507.03133

[24] M. Efroymson, "Multiple regression analysis," *Mathematical methods for digital computers*, vol. 1, pp. 191–203, 1960.

[25] R. R. Hocking and R. N. Leslie, "Selection of the Best Subset in Regression Analysis," *Technometrics*, vol. 9, no. 4, pp. 531–540, Nov. 1967. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/00401706.1967.10490502

[26] Y. Li and J. Zhu, "L1-norm quantile regression," *Journal of Computational and Graphical Statistics*, 2012.

[27] A. Belloni and V. Chernozhukov, "Least squares after model selection in high-dimensional sparse models," 2009.

[28] J. A. Machado, "Robust model selection and m-estimation," *Econometric Theory*, vol. 9, pp. 478–493, 1993.

[29] C. Bergmeir, R. J. Hyndman, and B. Koo, "A Note on the Validity of Cross-Validation for Evaluating Time Series Prediction," Monash University, Department of Econometrics and Business Statistics, Tech. Rep. 10/15, 2017. [Online]. Available: https://ideas.repec.org/p/msh/ebswps/2015-10.html

[30] C. Bergmeir and J. M. Benítez, "On the use of cross-validation for time series predictor evaluation," *Information Sciences*, vol. 191, pp. 192–213, May 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0020025511006773