Scenario generation for nongaussian time series via Quantile Regression

Marcelo Ruas and Alexandre Street, *Member, IEEE*

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, due to its unpredictability. To overcome this lack of certainty, one has to work with many different possibilities of outcome.

New statistical models capable of handling such difficulties are an emerging field in power systems literature [2]–[8] The main objective in such literature is to propose new models capable of generating scenarios of renewable generation (RG) which are demanded in (i) energy trading, (ii) unit commitment, (iii) grid expansion planning, and (iv) investment decisions (see ([9]–[12]) and references therein). In stochastic optimization, problems such as Unit Commitment, Economic Dispatch, Transmission Expansion Planning all use scenarios as input. Such scenarios are used to characterize the probability distribution within the optimization under uncertainty framework. When working with robust optimization, bounds for probable ranges of coefficients are needed.

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 it is important to consider probability forecasting. [2] 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 RG series by assuming the distribution *a priori*. On the other hand, **nonparametric models** (i) don't require a distribution to be specified, (ii) needs more 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, RG time series such as wind and solar are reported as non-Gaussian [3], [13]–[15]. To circumvent this problem, the usage of nonparametric methods - which doesn't rely on assuming any previously assumed distribution - is adequate. Quantile Regression (QR) is a tool for constructing a methodology for non-gaussian 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. This issue - also known as crossing quantiles - can be adressed by constraining the sequence of quantiles to be in an increasing order. Other possibility is making a transformation afterwards, as shown in [16].

The seminal work [17] 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 [18], 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 [4]–[8], QR is employed to model the conditional distribution of Wind Power Time Series. An updating quantile regression model is presented by [5]. The authors present a modified version of the simplex algorithm to incorporate new observations without restarting the optimization procedure. In [6], the authors build a quantile model from already existent independent Wind Power forecasts. The approach by [4] 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 online learning technique, where the model is easily updated after each new observation. In [8], 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.

We propose using Quantile Autoregression (QAR) to create a methodology capable of estimating and simulating a non-gaussian time series, such as RG. By estimating a regularized QAR we model the conditional quantile function. 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, with the objective of scenario generation in a parsimonious model on both covariates and quantiles. 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 crossing quantile issue is solved by introducing a constraint on the optimization problems that forces the quantile function monotonicity. Furthermore, in the quantile regression literature for wind forecasting, a sequence of quantiles is provided as output. In our work, we propose to estimate the conditional distribution as a whole.

The objective of this paper is, then, to propose a new methodology to address nonparametric time-series model focused on RG. This may be seen as a multiple quantile regression that specifies a time series model based on the empirical conditional distribution. The main contributions are:

- A nonparametric methodology to model the conditional distribution of RG time series to produce scenarios.
- We propose a methodology that selects the global optimal solution with parsimony both on the selection of covariates as on the quantiles. Regularization methods are based on two techniques: Best Subset Selection (MILP) and LASSO (Linear Programming)
- Regularization techniques applied to an ensemble of quantile functions to estimate the conditional distribution, solving the issue of non-crossing quantiles. On regularizing quantiles, we propose a smoothness on the coefficient value across the sequence of quantiles.

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 α-conditional quantile function of Y for a given value x of the d-dimensional random variable X, i.e., $Q_{Y|X} : [0,1] \times \mathbb{R}^d \to \mathbb{R}$, can be defined as

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

Let a dataset be composed from $\{y_t, x_t\}_{t\in T}$ and let ρ be the check function

$$\rho_\alpha(x) = \begin{cases} \alpha x & \text{if } x \geq 0 \\ (1-\alpha)x & \text{if } x < 0 \end{cases}. \quad (2)$$

The sample quantile function for a given probability α is then based on a finite number of observations and is the solution to minimizing the loss function $L(\cdot)$:

$$\hat{Q}_{Y|X}(\alpha, \cdot) \quad \in \quad \arg\min_{q\in\mathcal{Q}} L_\alpha(q) = \sum_{t\in T} \rho_\alpha(y_t - q(x_t)), \quad (3)$$

The α-quantile function 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. 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 quantile function is approximated by a sequence of $|J|$ (where J is an index set) quantiles $q_{\alpha_1} \leq q_{\alpha_2} \leq \cdots \leq q_{\alpha_{|J|}}$. We also define the closely related set $A = \{\alpha_j \mid j \in J\}$, whose elements all must range on $[0,1]$, for they are probability values such that $0 < \alpha_1 < \alpha_2 < \cdots < \alpha_{|J|} < 1$. The sequence $\{\alpha_j\}_{j\in J}$ provides a finite discretization of the interval $[0,1]$.

Problem (3) can be rewritten as a Linear Programming problem as in (4)-(7), thus being able to use a modern solver to fit our model. Variables ε_t^+ and ε_t^- represent the quantities $|y - q(\cdot)|^+$ and $|y - q(\cdot)|^-$, respectively. This new formulation estimates all quantiles at the same time, and we denote q as q_α to differentiate it from each α.

$$\min_{\beta_{0j},\beta_j,\varepsilon_{tj}^+,\varepsilon_{tj}^-} \sum_{j\in J}\sum_{t\in T} \left(\alpha_j \varepsilon_{tj}^+ + (1-\alpha_j)\varepsilon_{tj}^-\right) \quad (4)$$

subject to

$$\varepsilon_{tj}^+ - \varepsilon_{tj}^- = y_t - \beta_{0j} - \beta_j^T x_t, \forall t \in T, \forall j \in J, \quad (5)$$

$$\varepsilon_{tj}^+, \varepsilon_{tj}^- \geq 0, \qquad \forall t \in T, \forall j \in J, \quad (6)$$

$$\beta_{0j} + \beta_j^T x_t \leq \beta_{0,j+1} + \beta_{j+1}^T x_t, \qquad \forall t \in T, \forall j \in J_{(-1)}, \quad (7)$$

To estimate a conditional distribution based on quantile values, all quantiles $\{q_\alpha\}_{\alpha\in A}$ are estimated simultaneously. With the addition of constraint (7), we assure the monotonicity of the quantile function, solving the issue of crossing-quantiles. The output is the sequence $\{q_\alpha\}_{\alpha\in A}$, which is fully defined by the optimum values $\beta_{0\alpha}^*$ and β_α^*, for each α.

We apply QR to estimate the conditional distribution $\hat{Q}_{Y_{t+h}|X_{t+h},Y_t,Y_{t-1},\ldots}(\alpha, \cdot)$ for a h-step ahead forecast of time serie $\{y_t\}$, where X_{t+h} is a vector of exogenous variables at the time we want to forecast. Once the conditional distribution is estimated, we are able to simulate and generate scenarios.

In the next session, regularization techniques are presented, in order to choose parsimoniously which variables will be input for $\hat{Q}$.

III. Regularization on the covariates

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 $\beta_j = [\beta_{1j} \cdots \beta_{pj}]$ should not have all nonzero values. There are many ways of selecting a subset of variables among the available options. A classical approach for this problem is the Stepwise algorithm [24], [25], [21], which includes variables in sequence.

Two ways of selecting variables will be employed. In the first we use a Mixed Integer Linear Programming optimization problem (MILP) to find the best subset among all possible subsets 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 via MILP

We use MILP to select variables by including constraints which limits their number in K. Only K coefficients β_{pj} may have nonzero values, for each α. Binary variable z_{pj} indicates whether β_{pj} has a nonzero value. The optimization problem that incorporates this idea is described below:

$$\min_{\beta_{0j},\beta_j,z_{pj}\varepsilon^+_{tj},\varepsilon^-_{tj}} \sum_{j\in J}\sum_{t\in T}\left(\alpha_j\varepsilon^+_{tj} + (1-\alpha_j)\varepsilon^-_{tj}\right) \tag{8}$$

subject to

$$\varepsilon^+_{tj} - \varepsilon^-_{tj} = y_t - \beta_{0j} - \beta_j^T x_t, \quad \forall t \in T, \forall j \in J, \tag{9}$$

$$\varepsilon^+_{tj}, \varepsilon^-_{tj} \geq 0, \quad \forall t \in T, \forall j \in J, \tag{10}$$

$$-Mz_{pj} \leq \beta_{pj} \leq Mz_{pj}, \quad \forall j \in J, \forall p \in P, \tag{11}$$

$$\sum_{p\in P} z_{pj} \leq K, \quad \forall j \in J, \tag{12}$$

$$z_{pj} \in \{0,1\}, \quad \forall j \in J, \forall p \in P, \tag{13}$$

$$\beta_{0j} + \beta_j^T x_t \leq \beta_{0,j+1} + \beta_{j+1}^T x_t, \quad \forall t \in T, \forall j \in J_{(-1)}, \tag{14}$$

The objective function and constraints (9), (10) and (14) are the same from standard linear quantile regression. By constraint (11), variable z_{pj} is a binary that assumes 1 when coefficient β_{pj} is included, while (12) guarantees that at most K of them are nonzero. The value of M is chosen in order to guarantee that $M \geq \|\hat{\beta_h} j\|_\infty$. The solution given by β^*_{0j} and $\beta^*_j = [\beta^*_{1j} \cdots \beta^*_{pj}]$ will be the best linear α-quantile regression with K nonzero coefficients.

1) Defining groups for variables: Consider the optimization problem defined on (8)-(14). The choice of the best subset is independent for different values of probabilities α. This means that the best subset may include two completely different sets of regressors for two probabilities α_j and α_{j+1}. Take $K = 2$ for the example, selecting β_{1j} and β_{4j} for j while $\beta_{2,j+1}$ and $\beta_{5,j+1}$ is possible, but unlikely to be true.

To address this issue, we propose to divide all $j \in J$ 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 j in the same group g. To model these properties as constraints on problem (8)-(14), we substitute constraint (11) for the following equations:

$$z_{pjg} := 2 - (1 - z_{pg}) - I_{gj} \tag{15}$$

$$\sum_{g\in G} I_{gj} = 1, \quad \forall j \in J, \tag{16}$$

$$-Mz_{pjg} \leq \beta_{pj} \leq Mz_{pjg}, \quad \forall p \in P, \forall j \in J, \forall g \in G, \tag{17}$$

$$I_{gj}, z_{pg} \in \{0,1\}, \quad \forall p \in P, \forall g \in G, \tag{18}$$

on problem (8)-(14). 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_{gj} equals 1 iff the j^{th} quantile belongs to group g. Constraint (17) forces that

$$\text{if } z_{pg} = 0 \text{ and } I_{gj} = 1 \text{ then } \beta_{pj=0}.$$

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 j. Note that variable z_{pj} behaves differently than when we are not considering groups. This means that if the j^{th} quantile belongs to group g but variable p is not selected to be among the ones of group g, than β_{pj} is zero. Equation (15) defines z_{pj} to simplify writing.

B. Variable selection via LASSO

The second form of regularization we work here the LASSO technique. This method consists on the inclusion of the penalized coefficients ℓ_1-norm on the objective function of the QR problem. In [20], the reader can find properties and convergence rates when using the LASSO to select variables in a quantile regression setting. 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]. On the litterature, the LASSO QR regularization is applied for a single quantile only by the following optimization problem:

$$\min_{\beta_{0\alpha},\beta_\alpha} \sum_{t\in T} \alpha|y_t - q_\alpha(x_t)|^+ + \sum_{t\in T}(1-\alpha)|y_t - q_\alpha(x_t)|^- + \lambda\|\beta_\alpha\|_1, \tag{19}$$

$$q_\alpha(x_t) = \beta_0 - \beta^T x_t.$$

In our problem, however, we need to estimate multiple quantiles at once, in order to being able to circumvent the crossing quantiles issue.

The process of estimation is done in two stages: (i) variable selection and (ii) coefficients estimation. At first, all

normalized covariates[1] are input on the following optimization problem:

$$\tilde{\beta}^*_\lambda = \underset{\beta_{0j},\beta_j,\varepsilon^+_{tj},\varepsilon^-_{tj}}{\arg\min} \sum_{j\in J}\left(\sum_{t\in T}(\alpha_j\varepsilon^+_{tj} + (1-\alpha_j)\varepsilon^-_{tj}) + \lambda\sum_{p\in P}(\xi^+_{pj}+\xi^-_{pj})\right) \quad (20)$$

subject to

$$\varepsilon^+_{tj} - \varepsilon^-_{tj} = y_t - \beta_{0j} - \beta_j^T x_t, \qquad \forall t\in T, \forall j\in J, \quad (21)$$

$$\xi^+_{pj} - \xi^-_{pj} = \beta_{pj}, \qquad \forall p\in P, \forall j\in J, \quad (22)$$

$$\beta_{0j} + \beta_j^T x_t \le \beta_{0,j+1} + \beta_{j+1}^T x_t, \forall t\in T, \forall j\in J_{(-1)}, \quad (23)$$

$$\varepsilon^+_{tj}, \varepsilon^-_{tj}, \xi^+_{tj}, \xi^-_{tj} \ge 0, \qquad \forall t\in T, \forall j\in J, \quad (24)$$

This model is built upon the standard linear programming model for the quantile regression (4)-(7). On the above formulation, the ℓ_1-norm of equation (19) is substituted by the sum $\xi^+_{pj}+\xi^-_{pj}$, for each j, which represents the absolute value of β_{pj}.

For low values of λ, the penalty over the size of coefficients is small, making the output of problem (20)-(24) be composed mainly of nonzero coefficients. On the other hand, when the penalty on $\|\beta_j\|_1$ is big, many covariates will have zero valued coefficients. When λ approaches infinity, one has a constant model. Note the linear coefficient β_{0j} is not penalized.

As the LASSO coefficients are shrunk towards zero they become biased. Our strategy will be to employ the LASSO as a variable selector, and estimate coefficients with regular QR on a second stage. The optimum vector of coefficients $\tilde{\beta}^*_\lambda$ on the first stage may be composed by both nonzero and zero coefficients, for a given λ . We then define S_λ as the set of indexes of selected variables given by

$$S_\lambda = \{p\in\{1,\dots,P\}|\ |\tilde{\beta}^*_{\lambda,p}| \neq 0\}.$$

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

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

On the second stage, the optimal coefficient vector $\tilde{\beta}^*_\lambda$ is estimated by the non-regularized QR, where only variables that belongs to S_λ are input:

$$(\mathcal{L}^*_\lambda, \beta^*_\lambda) \overset{(obj,var)}{\longleftarrow} \min_{\beta_{0j},\beta_j,z_{pj},\varepsilon^+_{tj},\varepsilon^-_{tj}} \sum_{j\in J}\sum_{t\in T}\left(\alpha_j\varepsilon^+_{tj} + (1-\alpha_j)\varepsilon^-_{tj}\right)$$

subject to

$$\varepsilon^+_{tj} - \varepsilon^-_{tj} = y_t - \beta_{0\alpha} - \sum_{p\in S_\lambda}\beta_p x_{t,p}, \qquad \forall t\in T, \forall j\in J,$$

$$\forall j\in J_{(-1)}, \forall p\in P,$$

$$\beta_{0j} + \beta_j^T x_t \le \beta_{0,j+1} + \beta_{j+1}^T x_t, \qquad \forall t\in T, \forall j\in J_{(-1)},$$

$$\varepsilon^+_{tj}, \varepsilon^-_{tj} \ge 0, \qquad \forall t\in T, \forall j\in J,$$

$$D2^+_{pj}, D2^-_{pj} \ge 0, \qquad \forall j\in J, \forall p\in P.$$

The variable $\mathcal{L}^*_\lambda$ receives the value of the objective function on its optimal solution. In summary, the optimization in equation 19 acts as a variable selection for the subsequent estimation, which is normally called the post-LASSO estimation [27].

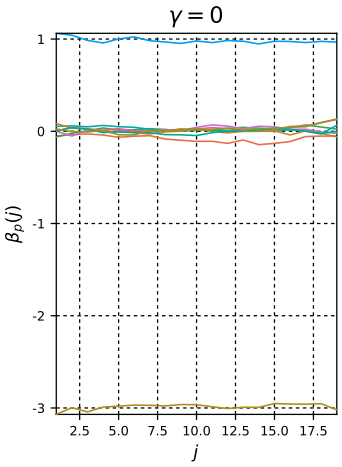

Fig. 1. $\mathcal{K}$-fold CV and $\mathcal{K}$-fold with non-dependent data. Observations in blue are used to estimation and in orange for evaluation. Note that non-dependent data doesn't use all dataset in each fold.

IV. Regularization on the Quantiles

The nonparametric approach we use in this paper consists of a different model for each α-quantile. From the assumption that the value of similar quantiles be produced by similar models, we need to prevent instabilities and big changes on the p coefficient $\beta_p(\alpha)$ when seen as a function of probability α.

As an example of the aforementioned issue, we simulate 100 *iid* observations of random variable $y = 1+\beta_1$, where X_1 are normal explanatory variables. Let $X_{reg} = [X_1\, X_2]$ be input on the problem and we have to figure out the real model. shown on Figure 1. To force the model to have parsimony on the quantiles, we propose to penalize the second derivative of $\beta_p(\alpha)$. On the implementation side, we must limit the second difference of the sequence of coefficients $\{\beta_{pj}\}_{j\in J}$ for every parameter p. This will bring smoothness to $\beta_p(\alpha)$.

$$\tilde{D}^2_{pj} := \frac{\left(\frac{\beta_{p,j+1}-\beta_{pj}}{\alpha_{j+1}-\alpha_j}\right) - \left(\frac{\beta_{p,j}-\beta_{p,j-1}}{\alpha_j-\alpha_{j-1}}\right)}{\alpha_{j+1}-2\alpha_j+\alpha_{j-1}}. \quad (25)$$

[1]For such estimation to be coherent each covariate must have the same relative weight in comparison with one another, i.e., they must be normalized. This normalization process is a linear transformation to each covariate such that all 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.

This idea is implemented on the optimization problem by adding a penalty on the objective function to penalize the absolute value $|D^2_{pj'}|$ by a tuning parameter γ, that controls how rough the sequence $\{\beta_{pj}\}_{j\in J}$ can be. The full optimization problem for the best subset selection via MILP which incorporates the derivative penalty is given below:

$$\min_{\beta_{0j},\beta_j,z_{pj}\varepsilon^+_{tj},\varepsilon^-_{tj}} \sum_{j\in J}\sum_{t\in T}\left(\alpha_j\varepsilon^+_{tj}+(1-\alpha_j)\varepsilon^-_{tj}\right) + \gamma\sum_{j\in J'}(D2^+_{pj}+D2^-_{pj}) \quad (26)$$

subject to

$$\varepsilon^+_{tj}-\varepsilon^-_{tj}=y_t-\beta_{0j}-\beta_j^T x_{t,p}, \qquad \forall t\in T,\forall j\in J, \quad (27)$$

$$-Mz_{p\alpha}\le\beta_{pj}\le Mz_{p\alpha}, \qquad \forall j\in J,\forall p\in P, \quad (28)$$

$$\sum_{p\in P} z_{p\alpha}\le K, \qquad \forall j\in J, \quad (29)$$

$$D2^+_{pj}-D2^-_{pj}=\frac{\left(\frac{\beta_{p,j+1}-\beta_{pj}}{\alpha_{j+1}-\alpha_j}\right)-\left(\frac{\beta_{p,j}-\beta_{p,j-1}}{\alpha_j-\alpha_{j-1}}\right)}{\alpha_{j+1}-2\alpha_j+\alpha_{j-1}}, \qquad \forall j\in J_{(-1)},\forall p\in P, \quad (30)$$

$$\beta_{0j}+\beta_j^T x_t\le\beta_{0,j+1}+\beta_{j+1}^T x_t, \forall t\in T,\forall j\in J_{(-1)}, \quad (31)$$

$$z_{p\alpha}\in\{0,1\}, \qquad \forall j\in J,\forall p\in P, \quad (32)$$

$$\varepsilon^+_{tj},\varepsilon^-_{tj}\ge 0, \qquad \forall t\in T,\forall j\in J, \quad (33)$$

$$D2^+_{pj},D2^-_{pj}\ge 0, \qquad \forall j\in J,\forall p\in P. \quad (34)$$

where A' is the set formed by the same elements of A without the first and the last elements, that need to be taken out of the sum as their second derivative cannot be calculated. The sum $D2^+_{pj}+D2^-_{pj}$ represents the absolute value of the second derivative, and its penalty is tuned by parameter γ.

The same feature is incorporated in the LASSO estimation, which now incorporates penalization of both coefficients ℓ_1-norm as well as the second derivative of $\beta(\alpha)$, shown by the problem below: The modified LASSO problem to include the

$$\tilde{\beta}^{*LASSO}_\lambda = \arg\min_{\beta_0,\beta,\varepsilon^+_{tj},\varepsilon^-_{tj}} \sum_{j\in J}\sum_{t\in T}(\alpha_j\varepsilon^+_{tj}+(1-\alpha_j)\varepsilon^-_{tj}) + \lambda\sum_{p\in P}(\xi^+_{pj}+\xi^-_{pj})+\gamma\sum_{j\in J'}(D2^+_{pj}+D2^-_{pj}) \quad (35)$$

subject to

$$\varepsilon^+_{tj}-\varepsilon^-_{tj}=y_t-\beta_{0j}-\beta_j^T x_{t,p}, \qquad \forall t\in T,\forall j\in J, \quad (36)$$

$$\xi_{pj}\ge\beta_{pj}, \qquad \forall p\in P,\forall j\in J, \quad (37)$$

$$\xi_{pj}\ge-\beta_{pj}, \qquad \forall p\in P,\forall j\in J, \quad (38)$$

$$D2^+_{pj}-D2^-_{pj}=\frac{\left(\frac{\beta_{p,j+1}-\beta_{pj}}{\alpha_{j+1}-\alpha_j}\right)-\left(\frac{\beta_{p,j}-\beta_{p,j-1}}{\alpha_j-\alpha_{j-1}}\right)}{\alpha_{j+1}-2\alpha_j+\alpha_{j-1}}, \qquad \forall j\in J_{(-1)},\forall p\in P, \quad (39)$$

$$\beta_{0j}+\beta_j^T x_t\le\beta_{0,j+1}+\beta_{j+1}^T x_t, \ \forall t\in T,\forall j\in J_{(-1)}, \quad (40)$$

$$\varepsilon^+_{tj},\varepsilon^-_{tj}\ge 0, \qquad \forall t\in T,\forall j\in J, \quad (41)$$

$$D2^+_{pj},D2^-_{pj}\ge 0, \qquad \forall j\in J,\forall p\in P. \quad (42)$$

A. ADALASSO

A popular extension of the LASSO is the Adaptive LASSO (ADALASSO). In [22], the authors extended the ADALASSO for QR and shown properties of the ADALASSO for the QR

When estimating the ADALASSO for quantile regression, we show a few adaptations and extensions of the original method. The full process consists of two steps, each consisting of a LASSO estimation:

- **First step:** First LASSO regularization, as in problem (35)-(42).
- **Second step:** The coefficients of the LASSO estimation are used to form the weights w_{pj} are:
  1) $w_{pj}=1/\beta_{pj}$.
  2) $w_{pj}=1/(\beta_{pj}\parallel\beta_j\parallel_1)$,

  where β_{pj} stands for the coefficients obtained on the first step and $\parallel\beta_j\parallel_1$ is the ℓ_1-norm of the j^{th} quantile coefficients. The weights w_j are input to a second-stage Lasso estimation:

$$\min_{\beta_{0j},\beta_j}\sum_{j\in J}\left(\sum_{t\in T}\rho_{\alpha_j}(y_t-(\beta_{0j}+\beta_j^T x_t))+\lambda\sum_{p\in P}w^\delta_{pj}\mid\beta_{pj}\mid\right) +\gamma\sum_{j\in J'}(D2^+_{pj}+D2^-_{pj}),$$

  where δ is an exponential parameter, normally set to 1.

V. Simulation

In this section, we investigate how to simulate future paths of the time series y_t. Let n be the total number of observations of y_t. We produce S different paths with size K for each. We have n observations of y_t and a vector of explanatory variables x_t. The variables chosen to compose x_t can be either exogenous variables, autoregressive components of y_t or both. We use a nonparametric approach which to estimate, at every t, the k-step ahead conditional density of y_t.

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

Procedure for simulating S scenarios of y_t

1) At first, let $\tau=n+1$.
2) In any given period τ, for every $\alpha\in A$, we use one of the methods presented in the last sections to estimate the value of each α-quantile. Note that x_τ is supposed to be known at time τ. 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_τ must be considered fully known.
3) Let $\hat{Q}_{y_\tau|X}(\alpha,x_\tau)$ be the estimated quantile function of y_τ. To estimate $\hat{Q}_{y_\tau}$, we first define a discrete quantile function $\tilde{Q}_{y_\tau}$. By mapping every $\alpha\in A$ with its estimated quantile $\hat{q}_\alpha$, we define function $\tilde{Q}_{y_\tau}$. When we interpolate

4) Once we have a distribution for y_{n+1}, we can generate S different simulated values, drawn from the distribution function $\hat{F}_{y_{n+1}} = \hat{Q}^{-1}_{y_\tau}$, derived from the quantile function found by doing steps 2 and 3. 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 $F^{-1}_{y_{n+1}}(X)$ has the same distribution as y_{n+1}. So, by drawing a sample of size S from X and applying the quantile function $Q_{y_{n+1}}(\alpha)$, we have our sample of size K for y_{n+1}.
5) Each one of the S different values for y_{n+1} will be the starting point of a different path. Now, for each $\tau \in [n+2, n+K]$ and $s \in S$, we have to estimate quantiles $q_{\alpha\tau,s}$ and find a quantile function for $\hat{Q}_{y_{\tau,s}}$ just like it was done on steps 2 and 3. Note that when $\tau > n+2$, every estimate will be scenario dependent, hence there will be S distribution functions estimated for each period τ. 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 τ it will be produced exact S values for y_τ, one for its own path. Repeating this step until all values of τ and s are simulated will give us the full simulations that we are looking for.

VI. Estimation and evaluation

In this section, we present the metric for which the fit can be evaluated. From this metric, we show two ways - information criteria and cross validation - to determine the best tuning parameter for the regularization techniques seen on section III.

A. Evaluation metrics

In order to evaluate our predictions, we need to define a metric for which we take as objective function to optimize. As conditional distribution is the focus in this paper, we use a performance measurement which emphasizes the correctness of each quantile. For each probability $\alpha \in A$, a loss function is defined by

$$L_\alpha(q) = \sum_{t \in T} \rho_\alpha(y_t - q(x_t)). \tag{43}$$

The loss score $\mathcal{L}$, which is the chosen evaluation metric to optimize, aggregates the score function over all elements of A:

$$\mathcal{L} = \frac{1}{|A|} \sum_{\alpha \in A} L_\alpha(q). \tag{44}$$

B. Time-series cross validation

Section III presented two different methods to estimate the conditional distribution in a parsimonious way. However, as presented, the aforementioned methods don't provide a unique solution, but a set of solutions for a range of tuning parameters. For instance, on the MILP method, the quantity $\mathcal{K}$ of nonzero coefficients is an input of the problem. Similarly, the LASSO needs a penalization parameter λ, that tunes how much penalty the ℓ_1-norm receives.

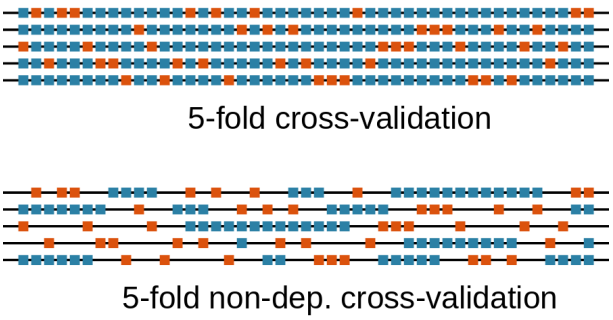

Fig. 2. $\mathcal{K}$-fold CV and $\mathcal{K}$-fold with non-dependent data. Observations in blue are used to estimation and in orange for evaluation. Note that non-dependent data doesn't use all dataset in each fold.

In statistics and machine learning, a popular technique is using Cross-validation (CV) to select the best model from this range of possibilities. It 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 [28] and [29] deals specifically with the usage of CV in a time series context. They provide tests with both $\mathcal{K}$-fold CV and $\mathcal{K}$-fold with non-dependent data. Both schemes are shown of Figure 2. In both settings, the training data is randomly split into a collection of sets J_k, forming a $\mathcal{K}$ size partition. Each of these J_k is used as test set, while the rest is used to estimate coefficients which will be used to predict values of J_k. As there are $\mathcal{K}$ folds, this procedure is done $\mathcal{K}$ times. So, for a given vector of tuning parameter θ (which can be either $[\gamma \quad \lambda]^T$ for the LASSO or $[\gamma \quad K]^T$ from the MILP problem), the CV score is given by the sum of the loss function for each fold. The optimum value of θ in this criteria is the one that minimizes the CV score:

$$\theta^* = \arg\min_\theta CV(\theta) = \sum_{k \in \mathcal{K}} \sum_{\alpha \in A} L_\alpha(q_\theta).$$

The optimization in VI-B is done by the Nelder-Mead method, which is an iterative method based on simplexes that is suitable for nonlinear problems. Given that this problem is only bidimensional, there is a high probability of the solution being close enough to the optimum θ^*.

C. Information Criteria for Quantile Regression

Sometimes, using CV can be computationally expensive, as the full estimation is done several times for each tuning

parameter - in this case, either K or λ. Other form of deciding the quantity of variables that provides a good equilibrium between in-sample prediction and parsimony is the Information Criteria.

Information criteria summarizes two aspects. One of them refers to how well the model fits the in-sample observations and 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 [30], 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:

$$SIC(m) = |T| \log(\mathcal{L}^*) + \frac{1}{2} K \log |T|, \tag{45}$$

where K is the model's dimension. This procedure leads to a consistent model selection if the model is well specified. By minimizing the SIC, the chosen model is the one with the best combination, according to this metric, of fit and parsimony among all models.

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] 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

[3] 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.

[4] 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.

[5] 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

[6] 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.

[7] 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

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

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

[15] 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.

[16] 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

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

[18] 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

[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] 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

[29] 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

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