Revista Colombiana de Estadística
Special Issue: Colombian and Brazil Graduate Programs
December 2025, volume 48, issue 3, pp. 479–507
http://doi.org/10.15446/rce.v48n3.122651

Global Variable Selection for Quantile Regression

Selección global de variables para regresión cuantílica

Tais Bellini[1,a], Gabriela Cybis[2,b], Eduardo Horta[2,c]

[1]SAP Labs Latin America, São Leopoldo, Brazil

[2]Department of Statistics, Institute of Mathematics and Statistics, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil

Abstract

Quantile regression provides a parsimonious model for the conditional quantile function of the response variable Y given the vector of covariates X, and describes the whole conditional distribution of the response, yielding estimators that are more robust to the presence of outliers. Quantile regression models specify, for each quantile level τ, the functional form for the conditional τ-th quantile of the response, which brings complexity to perform variable selection using regularization techniques, such as LASSO or adaptive LASSO (adaLASSO), as one might obtain a different set of selected variables for each quantile level. In this work, we propose a method for global variable selection and coefficient estimation in the linear quantile regression framework, imposing few restrictions on the functional form of $\beta(\cdot)$, and applying group adaLASSO penalization for variable selection. We set up a Monte Carlo study comparing six different proposed estimators based on LASSO, adaLASSO and group LASSO in six scenarios that diversify sample and quantile levels grid sizes. The findings demonstrate that the selection of the tuning parameter λ for penalization is critical for model selection and coefficient estimation. It was observed that the methods using traditional LASSO are more prone to include the true model as compared to adaLASSO, but renouncing model shrinkage and not removing irrelevant covariates, while the grouped approaches are more effective in zeroing coefficients that are less relevant.

Keywords: Chebyshev polynomials; Group adaLASSO.

[a]Master's Degree. E-mail: tais38@gmail.com
[b]Ph.D. E-mail: gcybis@yahoo.com.br
[c]Ph.D. E-mail: eduardo.horta@ufrgs.br

Resumen

La regresión cuantílica proporciona un modelo parsimonioso para la función de cuantiles condicionales de la variable de respuesta Y dado el vector de covariables X, y describe toda la distribución condicional de la respuesta, produciendo estimadores más robustos ante la presencia de valores atípicos. Los modelos de regresión cuantílica especifican, para cada nivel de cuantil τ, la forma funcional del τ-ésimo cuantil condicional de la respuesta, lo que introduce complejidad a la hora de realizar selección de variables mediante técnicas de regularización, como LASSO o LASSO adaptativo (adaLASSO), ya que podríamos obtener un conjunto diferente de variables seleccionadas para cada nivel de cuantil. En este trabajo proponemos un método de selección global de variables y estimación de coeficientes en el marco de la regresión cuantílica lineal, imponiendo pocas restricciones sobre la forma funcional de $\beta(\cdot)$ y aplicando una penalización adaLASSO por grupos para la selección de variables. Realizamos un estudio de Monte Carlo comparando seis estimadores propuestos basados en LASSO, adaLASSO y LASSO grupal en seis escenarios que diversifican los tamaños de muestra y de la rejilla de niveles de cuantil. Los resultados demuestran que la selección del parámetro de ajuste λ para la penalización es crítica para la selección del modelo y la estimación de coeficientes. Se observó que los métodos que utilizan LASSO tradicional son más propensos a incluir el modelo verdadero en comparación con adaLASSO, pero a costa de renunciar a la contracción del modelo y de no eliminar covariables irrelevantes, mientras que los enfoques agrupados son más eficaces para anular (llevar a cero) los coeficientes menos relevantes.

Palabras clave: Group adaLASSO; Polinomios de Chebyshev.

1. Introduction

Quantile regression, brought to light in its modern guise by Koenker & Bassett (1978), provides a parsimonious model for the conditional quantile function of the response variable Y given the vector of covariates X, and, *a fortiori*, describes the whole conditional distribution of the response. Importantly, quantile regression also yields more robust estimators to the presence of outliers, as opposed to classical linear regression methods that only evaluate the conditional mean at a specific location (Davino et al., 2014). In Buchinsky (1998), practical uses of quantile regression exemplify an empirical analysis of change in education returns at various points of the log wage distribution. Koenker (2000), Koenker & Hallock (2001) and Koenker (2005) apply quantile regression to consolidated econometric examples. Recent studies use the quantile regression concepts to measure the social and economic impacts of COVID-19 in the society, as seen in Lu et al. (2020), Azimli (2020) and Bonaccorsi et al. (2020).

Variable selection techniques have been proposed for the regression context aiming to select a subset of predictors in the model, especially in cases where the number of studied covariates is large, thus bringing interpretability and tractability to the estimated model. Tibshirani (1996) introduced the Least Absolute Shrinkage and Selection Operator (hereafter, LASSO) regression, a regularization technique that applies an ℓ^1 penalty to the ordinary least squares (OLS) estimation,

thus forcing corner solutions that result in some estimated coefficients that are exactly zero. Further, Zou (2006) introduced the adaptive LASSO (adaLASSO), where coefficients are penalized with distinct, adaptive weights in the penalty factor, an approach that attains the so called *oracle property* that is lacking in the standard LASSO except under strong assumptions. LASSO and adaLASSO aim to select individual variables in the model, whereas the group LASSO of Yuan & Lin (2006) targets variables in a grouped manner by applying a penalty that is intermediate between the ℓ^1 and ℓ^2 penalties. Wang & Leng (2008) extend the group LASSO to the adaptive group LASSO, demonstrating consistency and oracle efficiency.

Regularization techniques for variable selection have been widely applied in quantile regression models. Koenker (2004) applies ℓ^1 regularization methods in longitudinal data to shrink the estimation of random effects, Li & Zhu (2008) propose an efficient algorithm to compute the solution path of the ℓ^1-norm quantile regression, and Belloni & Chernozhukov (2011) apply this regularization in high-dimensional sparse models. More recently, Man et al. (2022) propose fitting a penalized convolution smoothed quantile regression with several convex penalties. Furthermore, Wu & Liu (2009) explore adaptive LASSO penalization in linear quantile regression. In addition, Li et al. (2010) proposed regularized quantile regression with group LASSO from a Bayesian perspective and Hashem et al. (2016) apply the grouped approach for classification.

Quantile regression models are customarily presented by specifying, for each quantile level τ, the functional form for the conditional τ-th quantile of the response, seen as a function of the covariates. Therefore, for each desired quantile level, there corresponds one regression equation and, with regards to estimation, one optimization problem. This brings complexity to certain operations since there are, say, M different optimization procedures, where $M \geqslant 1$ is the cardinality of the set $\mathscr{T}$ of quantile levels we wish to evaluate. One scenario where this may raise inconsistent models occurs when we desire to perform variable selection using regularization techniques, such as LASSO or adaptive LASSO, as we might obtain a different set of selected variables for each quantile level.

Frumento & Bottai (2016) propose modeling the regression functional coefficient $\beta(\cdot)$ as a parametric function of the quantile level in a way that the functional space in the minimization problem is finite dimensional. Further applications of this proposal are explored in Frumento et al. (2021), Frumento & Salvati (2021), and Sottile & Frumento (2022). Adding to this approach, it is possible to tackle global selection of covariates: for instance, Sottile et al. (2020) study global estimation and variable selection using the LASSO, demonstrating its ability to efficiently approximate the true model with a high probability, although the (ada)LASSO, *per se*, does not properly tackle selection of covariates (see discussion below). Das & Ghosal (2018) and Park & He (2017), in turn, study *approximating* the function $\beta(\cdot)$ using B-splines, and Yoshida (2021) further employs the adaptive group LASSO (Wang & Leng, 2008) for variable selection in this connection. Ruas et al. (2022) propose an estimation for all quantile regression models in a single mathematical optimization in a time series context using a Lipschitz regularization.

In this work, we posit a method for *global* variable selection and coefficient estimation in the linear quantile regression framework. Selecting variables *globally* is of great practical importance in applied quantile regression, providing a crucial step in dimension reduction, determination of causal relationships, as well as aiding in providing interpretable post-fitting analyses. In spite of this relevance, traditional approaches typically only yield local (i.e., τ-dependent) selection, and the problem of global selection in QR has only recently gained momentum in the literature. Our proposal is similar to (and partially inspired by) the ideas put forth by Sottile et al. (2020), combined with the group adaLASSO penalty of Yoshida (2021), but we consider Chebyshev interpolation in contrast to the more flexible—albeit at the price of restricting the functional parameter to lie on a finite dimensional space—approach of Sottile et al. (2020) or the B-splines strategy of Yoshida (2021). In terms of theoretical assumptions, our method has the advantage of imposing little restrictions on the functional form of $\beta(\cdot)$, only requiring a condition that is slightly weaker than the continuous differentiability of its coefficients. A Monte Carlo study was performed to assess and compare the quality of the proposed (class of) estimators. We use a single data generating process to set up a study of 200 replications comparing six different proposed estimators in six scenarios that diversify sample and τ-grid sizes.

The paper is organized as follows: Section 2 describes the main concepts used in this work and the proposed estimators; Section 3 explains the data generation process used in the study, how the simulation procedure was set up, what are the evaluation criteria, the results of the simulation and comparison among methods; finally, Section 4 provides a final discussion enlightening future work.

2. Methodology

This heading gives an account of the theoretical framework used for this study. We describe the (linear) quantile regression model, as well as our proposed method for global estimation and variable selection.

2.1. Global Quantile Regression and Variable Selection

For a scalar random variable Y and a D-dimensional random vector X, the conditional τ-th quantile of Y given $X = x$ is

$$Q_{Y|X}(\tau|x) := \inf\{y \in \mathbb{R}\colon\ \mathbf{P}(Y \leqslant y \,|\, X = x) \geqslant \tau\},$$

for $0 < \tau < 1$ and $x \in \text{support}(X)$. The mapping $\tau \mapsto Q(\tau|x)$ is called the conditional quantile function of Y given $X = x$. The most studied specification is the linear one, presented by Koenker & Bassett (1978), which considers that there is some functional parameter $\beta\colon (0,1) \to \mathbb{R}^D$ such that the conditional quantile function admits the representation

$$Q_{Y|X}(\tau|x) = x^\intercal \beta(\tau), \tag{1}$$

for all $\tau \in (0,1)$ and $x \in \text{support}(X)$. Under this *globally concerned* linear quantile regression specification (the terminology was coined in Zheng et al., 2015) and a convexity assumption, it holds that, for any $\tau \in (0,1)$ and integrable Y,

$$\beta(\tau) = \arg\min_{b \in \mathbb{R}^D} \mathbf{E}\rho_\tau\big(Y - X^\intercal b\big),$$

where $\rho_\tau(\cdot)$ is the check function, $\rho_\tau(u) = u(\tau - \mathbb{I}_{[u<0]})$ (see Hunter & Lange (2000), for example). Along these lines, for a suitable grid composed of M quantile levels, say $\mathscr{T} = \{\tau_1, \ldots, \tau_M\}$, and letting $\boldsymbol{\beta}$ denote the $D \times M$ matrix whose component (d,m) is $\beta_d(\tau_m)$, it also holds that

$$\boldsymbol{\beta} = \arg\min_{\boldsymbol{b}} \sum_{m=1}^{M} \mathbf{E}\rho_{\tau_m}\big(Y - X^\intercal \boldsymbol{b}_{:,m}\big), \tag{2}$$

with the minimum running through all $D \times M$ matrices $\boldsymbol{b}$ having columns $\boldsymbol{b}_{:,m}$.

Regularization methods aimed to reduce the number of covariates in the estimated model, such as LASSO or adaLASSO, can be applied to the quantile regression context by incorporating a penalizing factor. In light of 2, for a sample of size N, denoting respectively by X_n and Y_n the covariates and response variable for the nth observation ($1 \leqslant n \leqslant N$), it is natural in this setting to estimate the parameter $\boldsymbol{\beta}$ by solving the following optimization problem:

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{b}} \sum_{n=1}^{N}\sum_{m=1}^{M} \left\{\rho_{\tau_m}(Y_n - X_n^\intercal \boldsymbol{b}_{:,m}) + \tilde{P}(\boldsymbol{b}_{:,m})\right\}, \tag{3}$$

where $\tilde{P}(\cdot)$ is a penalizing factor. We call this estimation procedure the "direct approach" and use it as a baseline in our simulation study.

Simply estimating $\beta(\cdot)$ from a finite set of quantile levels can be misleading since such an estimator may fail to provide a global picture of this functional parameter. For instance, there is no assurance that the values of β outside said grid would be close to its values at the grid. Thus, if β is "too irregular", it will not be correctly selected when the grid is poorly chosen. To give an example, if some of the β_d's, say $\beta_2(\tau)$, are defined as $\mathbb{I}[\tau > 0.9]$, then any grid $\mathscr{T} \subseteq (0, 0.9]$ will lead to problems in identifying X_2 as a relevant covariate. As the results below illustrate, such problems do not occur provided β is "sufficiently smooth". The proofs can be found in appendix.

Theorem 1. *Assume that, for each $0 < \delta < 1/2$, the coordinate functions $\beta_1, \ldots, \beta_D$ are absolutely continuous on $[\delta, 1-\delta]$, and that moreover the condition holds that*

$$\sum_{d=1}^{D} \int_{-1}^{1} \left\{\partial\beta_d\left(\tfrac{1}{2} + \tfrac{1-2\delta}{2}x\right)\right\}^2 \frac{\mathrm{d}x}{\sqrt{1-x^2}} < \infty. \tag{4}$$

Then, for each $M \geqslant 2$ and $\delta \in (0, 1/2)$, there exist a set of grid points $\mathscr{T} = \{\tau_1, \ldots, \tau_M\}$ with $1 - \delta = \tau_1 > \tau_2 > \cdots > \tau_M = \delta$, real coefficients $\boldsymbol{\alpha}_{dm}$ (with $1 \leqslant d \leqslant D$

and $1 \leqslant m \leqslant M$*), linearly independent polynomials* $\varphi_1(\cdot), \ldots, \varphi_M(\cdot)$*, and a positive constant* $C(\beta, \delta)$*, which does not depend on* M*, such that*

$$\sup_{\delta \leqslant \tau \leqslant 1-\delta} \left| \beta_d(\tau) - \sum\nolimits_{\ell=1}^{M} \boldsymbol{\alpha}_{d\ell} \varphi_\ell(\tau) \right| \leqslant \frac{C(\beta, \delta)}{\sqrt{M-1}}, \qquad 1 \leqslant d \leqslant D, \tag{5}$$

with the equality $\beta_d(\tau) = \sum_{\ell=1}^{M} \boldsymbol{\alpha}_{d\ell} \varphi_\ell(\tau)$ *holding whenever* $\tau \in \mathscr{T}$*.*

Remark 1. Recall that a real valued function ψ defined on the closed interval $[\delta, 1-\delta]$, where $0 < \delta < 1/2$, is said to be *absolutely continuous* if and only if (i) ψ is Lebesgue-almost everywhere differentiable on $[\delta, 1-\delta]$, and; (ii) its derivative $\partial\psi\colon [\delta, 1-\delta] \to \mathbb{R}$ is Lebesgue integrable on $[\delta, 1-\delta]$, and the representation $\psi(\tau) = \psi(\delta) + \int_\delta^\tau \partial\psi(u)\,\mathrm{d}u$ holds, for all $\tau \in [\delta, 1-\delta]$.

Remark 2. In the conditions of Theorem 1, denote by $\boldsymbol{\alpha}$ the $D \times M$ matrix whose component (d, m) is $\boldsymbol{\alpha}_{dm}$, by $\boldsymbol{\beta}$ the $D \times M$ matrix as defined above and by $\boldsymbol{\varphi}$ the $M \times M$ matrix whose component (ℓ, m) is $\varphi_\ell(\tau_m)$. Write moreover $\varphi(\cdot) = [\varphi_1(\cdot) \quad \cdots \quad \varphi_M(\cdot)]^\intercal$. Then (5) can be recast as

$$\sup_{\delta \leqslant \tau \leqslant 1-\delta} \|\beta(\tau) - \boldsymbol{\alpha}\varphi(\tau)\| \leqslant \frac{C(\beta, \delta)}{\sqrt{M-1}},$$

and in particular it holds that $\boldsymbol{\beta} = \boldsymbol{\alpha}\boldsymbol{\varphi}$. With this notation, we have the following direct consequence of Theorem 1.

Corollary 1. *The "basis matrix"* $\boldsymbol{\varphi}$ *is invertible, with* $\boldsymbol{\alpha} = \boldsymbol{\beta}\boldsymbol{\varphi}^{-1}$*. Additionally, letting*

$$R(\boldsymbol{b}) \coloneqq \sum_{m=1}^{M} \mathbf{E}\rho_{\tau_m}\big(Y - X^\intercal \boldsymbol{b}_{:,m}\big), \tag{6}$$

it holds that $\boldsymbol{\alpha}$ *is a minimizer of the mapping* $\boldsymbol{a} \mapsto R(\boldsymbol{a}\boldsymbol{\varphi})$*.*

Remark 3. It is important to notice that the constants δ, M, and even the functional parameter β, can be allowed to depend on the sample size N (if β depends on the sample size, then the data generating process should be indexed by N as well: we would have, e.g., observations $\big(Y_n^N, X_n^N\big)$ for $N \geqslant 1$ and $1 \leqslant n \leqslant N$, etc). Permitting δ and M to depend on N is of interest as this allows one to estimate β at a set of grid points that can get both finer and "wider", with the obvious benefits that such a grid provides. In turn, allowing β to depend on the sample size is a way to accommodate scenarios where more covariates are added to the model when N gets larger, for example. In this setting, by implication the bounding "constant" $C(\beta, \delta)$ will depend on the sample size too, although it can be difficult, especially when β varies with N, to explicitly derive conditions under which $C(\beta, \delta)/\sqrt{M-1} \to 0$ as $N \to \infty$. If this is the case, then we can state the following result.

Proposition 1. *Let* $(\gamma_N^i)_{N \geqslant 1}$*,* $i = 1, 2$*, be two sequences of non-negative real numbers. With the notation of Theorem 1, let* $\widehat{\boldsymbol{\alpha}}$ *be an estimator satisfying* $\|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\| = O_{\mathbb{P}}(\gamma_N^1)$ *and assume* $C(\beta, \delta)/\sqrt{M-1} = O(\gamma_N^2)$*. Define moreover*

$$\widehat{\beta}(\tau) \coloneqq \widehat{\boldsymbol{\alpha}}\varphi(\tau), \qquad \delta \leqslant \tau \leqslant 1-\delta, \tag{7}$$

where $\varphi(\cdot) = [\varphi_1(\cdot) \quad \cdots \quad \varphi_M(\cdot)]^\intercal$. *Then*

$$\sup_{\delta \leqslant \tau \leqslant 1-\delta} \left\| \widehat{\beta}(\tau) - \beta(\tau) \right\| = O_{\mathbb{P}}\big(\max\{\gamma_N^1 \sqrt{M}\,, \gamma_N^2\} \big). \tag{8}$$

In particular, $\widehat{\beta}(\cdot)$ *is uniformly consistent for* β *if and only if* $\max\{\gamma_N^1 \sqrt{M}\,, \gamma_N^2\} \to 0$ *as* $N \to \infty$.

We conclude this remark by noticing that, in a typical setting, one cannot be "too greedy" in expanding the grid $\mathscr{T}$ as the sample size grows. Indeed, in order to nearly preserve the convergence rate γ_N^1, a sensible choice is to set $M = O\big(1/\log(\gamma_N^1)\big)$.

In view of Corollary 1, a natural estimator for $\boldsymbol{\alpha}$ is given by

$$\widehat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{a}} \sum_{n=1}^{N} \sum_{m=1}^{M} \rho_{\tau_m}(Y_n - X_n^\intercal \boldsymbol{a} \boldsymbol{\varphi}_{:,m}) \;+\; P(\boldsymbol{a}), \tag{9}$$

where the minimization runs through all $D \times M$ matrices $\boldsymbol{a}$, and where $P(\cdot)$ is a (possibly random) penalty factor. Notice that this estimator is equivalent to the one put forth by Frumento & Bottai (2016), whenever $\beta(\cdot)$ is comprised of polynomials—and, when the parameter does not fall in this polynomial class, both estimators are still asymptotically equivalent. As a matter of fact, we claim a weaker assumption, stating that the representation $\beta(\tau) = \boldsymbol{\alpha}\varphi(\tau)$ holds for a grid of quantile levels $\tau \in \mathscr{T}$, instead of being valid for the whole unit interval as required in Frumento & Bottai (2016). We conjecture that, under mild ergodicity and convexity assumptions, together with a properly chosen penalty $P(\cdot)$, the global estimator (7) converges uniformly to $\beta(\cdot)$, and that the estimated active set $\{d : \widehat{\beta}_d(\cdot) \neq 0\}$ asymptotically identifies the relevant covariates.

As seen in the proof of Theorem 1, we choose the basis functions $\varphi\colon (0,1) \to \mathbb{R}^M$ and the grid of quantile levels $\mathscr{T}$ from the shifted Chebyshev polynomials to guarantee that the matrix $\boldsymbol{\beta}$ provides a fair picture of the whole $\beta(\cdot)$. Our estimator $\widehat{\boldsymbol{\alpha}}$, defined in (9), depends crucially on the choice of the penalization term $P(\cdot)$. Below, we describe the penalty functions that will be considered in this work.

2.2. Penalty functions

2.2.1. LASSO and adaLASSO

To perform variable selection using the adaLASSO penalization (Zou, 2006), the term $P(\cdot)$ in (9) is a weighted ℓ^1-norm penalizing factor,

$$P(\boldsymbol{a}) = \lambda \sum_{d=1}^{D} \sum_{\ell=1}^{L} \mathrm{w}_{d\ell} |\boldsymbol{a}_{d\ell}|, \tag{10}$$

where $\lambda > 0$ is a tuning parameter, and $\mathrm{w}_{d\ell} := (|\widehat{\boldsymbol{a}}_{d\ell}|)^{-p}$ is a weight based on a first-step estimator $\widehat{\boldsymbol{a}}$, with $p > 0$ fixed. For instance, traditional adaLASSO is achieved

by setting $\widehat{\boldsymbol{a}} := \widehat{\boldsymbol{\beta}}_{\text{qr}}\boldsymbol{\varphi}^{-1}$, where $\widehat{\boldsymbol{\beta}}_{\text{qr}}$ solves (3) with $\tilde{P}(\cdot)$ identically zero, whereas the standard LASSO selection (as implemented in Sottile et al., 2020) corresponds to $\text{w}_{d\ell} \equiv 1$. Under this choice of $P(\cdot)$, we are finding the matrix $\widehat{\boldsymbol{\alpha}}$ that solves the optimization problem (9), with the (ada)LASSO penalization setting to zero those components of $\widehat{\boldsymbol{\alpha}}$ that are not "relevant to the model". Notwithstanding, this approach zeroes elements of $\widehat{\boldsymbol{\alpha}}$ individually without any "pattern restriction", hence it does not coherently achieve variable selection since each covariate is represented by an entire row of $\widehat{\boldsymbol{\alpha}}$. The lack of pattern restrictions is illustrated by the following scheme:

$$\widehat{\boldsymbol{\alpha}} = \begin{bmatrix} \widehat{\boldsymbol{\alpha}}_{11} & \widehat{\boldsymbol{\alpha}}_{12} & \dots & 0 \\ \widehat{\boldsymbol{\alpha}}_{21} & 0 & \dots & \widehat{\boldsymbol{\alpha}}_{2L} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \widehat{\boldsymbol{\alpha}}_{(D-1)2} & \dots & \widehat{\boldsymbol{\alpha}}_{(D-1)L} \\ \widehat{\boldsymbol{\alpha}}_{D1} & \widehat{\boldsymbol{\alpha}}_{D2} & \dots & \widehat{\boldsymbol{\alpha}}_{DL} \end{bmatrix}.$$

As a consequence, although adaLASSO indeed shrinks coefficients (of $\widehat{\boldsymbol{\alpha}}$), no *de facto* variable selection is achieved.

2.2.2. Group adaLASSO

In view of the drawbacks of adaLASSO in achieving proper variable selection in our global framework (9), we propose introducing the group adaLASSO penalty (Yuan & Lin, 2006; Wang & Leng, 2008), which is an approach that applies an ℓ^s-norm penalization to groups of coefficients, thus zeroing coefficients in a grouped manner:

$$P(\boldsymbol{a}) = \lambda \sum_{d=1}^{D} \text{w}_d \|\boldsymbol{a}_{d,:}\|, \tag{11}$$

where $\|\boldsymbol{a}_{d,:}\| := \sqrt[s]{\sum_{\ell=1}^{L} |\boldsymbol{a}_{d\ell}|^s}$, with $s > 1$. Here, setting $\text{w}_d := \|\widehat{\boldsymbol{a}}_{d,:}\|^{-p}$ for some $p > 0$ yields the group adaLASSO procedure, whereas $\text{w}_d \equiv 1$ corresponds to the standard group LASSO. The penalty function (11) will consider an entire row as active or not, thus yielding a *bona fide* variable selection procedure, as the following scheme illustrates:

$$\widehat{\boldsymbol{\alpha}} = \begin{bmatrix} \widehat{\boldsymbol{\alpha}}_{11} & \widehat{\boldsymbol{\alpha}}_{12} & \dots & \widehat{\boldsymbol{\alpha}}_{1L} \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \widehat{\boldsymbol{\alpha}}_{(D-1)1} & \widehat{\boldsymbol{\alpha}}_{(D-1)2} & \dots & \widehat{\boldsymbol{\alpha}}_{(D-1)L} \\ \widehat{\boldsymbol{\alpha}}_{D1} & \widehat{\boldsymbol{\alpha}}_{D2} & \dots & \widehat{\boldsymbol{\alpha}}_{DL} \end{bmatrix}.$$

A similar selection strategy is adopted by Yoshida (2021), but the author uses B-splines instead of polynomial interpolation to approximate $\beta(\cdot)$.

3. Monte Carlo Simulation Study

To evaluate the proposed method for global estimation and variable selection, we executed a Monte Carlo simulation study employing, in (9), the four penalty procedures described in 2.2. We also applied the group LASSO and group adaLASSO penalties to what we call the "direct approach", namely the solution $\widehat{\boldsymbol{\beta}}$ to (3). This yields six different selection procedures: LASSO, adaLASSO, group LASSO, group adaLASSO, direct group LASSO, direct group adaLASSO.

3.1. Data Generating Process

We consider the linear quantile regression Model (1), where $X \in \mathbb{R}^D$, and with the functional parameter $\beta(\cdot)$ of polynomial type,

$$\beta_d(\tau) = \theta_d \cdot \tau^{d-1}, \qquad 0 \leqslant \tau \leqslant 1 \text{ and } 1 \leqslant d \leqslant D, \tag{12}$$

where θ is a vector that determines which regressors are relevant/active (that is, those for which $\theta_d \neq 0$), and at the same time the magnitude of non-zero coefficients. It carries the following pattern: $\theta_d > 0$ for $1 \leqslant d \leqslant D^*$ and $\theta_d = 0$ for $D^* < d \leqslant D$, where D^* is the number of relevant covariates; $\theta_d = {}^{2}/_{D^*}$ for $1 \leqslant d \leqslant D_{\text{strong}}$, where D_{strong} is the number of coefficients with a "strong signal"; and where the D_{weak} remaining positions in the vector correspond to coefficients with a "weak signal", determined by $\theta_d = {}^{0.1}/_{D^*}$ for $(D_{\text{strong}} + 1) \leqslant d \leqslant D^*$.

A random sample from (Y, X) can be generated using the fundamental theorem of simulation, via

$$Y_n := Q_{Y|X}(U_n|X_n), \quad n \in \{1, \ldots, N\}, \tag{13}$$

where (X_n, U_n) are i.i.d. draws from (X, U), with X multivariate uniform on the D-dimensional unit cube, except for the first coordinate which is identically one, and U is a standard uniform random variable on the unit interval, independent from X. In practice we fixed $D = 30$, $D^* = 20$, $D_{\text{strong}} = 9$; thus, $D_{\text{weak}} = 11$.

It is worth mentioning that, for this particular data generating process, the representation $\beta(\tau) = \boldsymbol{\alpha}\boldsymbol{\varphi}(\tau)$ is valid for any τ (not restricted to $\tau \in \mathscr{T}$), thus falling inside the framework of Frumento & Bottai (2016). Hence, the LASSO method evaluated in this work is tantamount to the one proposed by Sottile et al. (2020).

3.2. Simulation Procedure

The optimization algorithm to compute the estimator $\widehat{\boldsymbol{\alpha}}$ described in (9) was carried out through the package CVXR (Fu et al., 2020) from the statistical environment R. To determine $\widehat{\boldsymbol{a}}$ for the weights in (10) and (11) we used the quantreg package (Koenker, 2021) to obtain the canonical estimator $\widehat{\boldsymbol{\beta}}_{\text{qr}}$ and derive $\widehat{\boldsymbol{a}} = \widehat{\boldsymbol{\beta}}_{\text{qr}}\boldsymbol{\varphi}^{-1}$. The parameter p was fixed to 1, and $s \equiv 2$. The grid of quantile levels, $\mathscr{T}$, and the matrix $\boldsymbol{\varphi}$ were generated using the shifted Chebyshev polynomials as described

in the proof of Theorem 1. The initial grid of λ values was generated by setting $\Lambda_0 = \{10^i : i \in \texttt{seq}\}$, where `seq` is a vector of 50 values equally spaced in the interval $[-3, 3]$, resulting in 50 λ values ranging in the interval $[0.001, 1000]$, rounded to four decimal places. Notwithstanding, for the present data generating process, we found out in preliminary simulations that values of λ in points of the grid beyond the 30th value (3.5565) are either prone to numerical instability or effectively large enough so that every coefficient was zeroed, except for the intercept.[1] In view of this, we decided to restrict the values of λ to lie in the interval $[0.001, 3.5565]$, resulting in the grid $\Lambda = \{\lambda \in \Lambda_0 : \lambda \leqslant 3.5565\}$ with 30 points. In our preliminary simulations we also found out that, in a handful of scenarios, the numerical optimization algorithms ended up returning an error flag. In this connection, we set out with a dataset of $\texttt{nrep.tot} := 10000$ replications, each consisting of a sample $N_{\max} = 1000$ independent realizations from (X, Y), generated via the method described above. For any given `seed()`, this generation is reproducible, always yielding the same dataset for the same pair $(\texttt{nrep.tot}, N_{\max})$. The dataset contains the random data for the Monte Carlo study, hence it is possible to run the replications independently, spreading the execution across multiple platforms. The code is available from the authors upon request.

In each replication, we performed the optimization procedure corresponding to each one of the six proposed methods (LASSO, adaLASSO, group LASSO, group adaLASSO, direct group LASSO, direct group adaLASSO), with varying sample size[2] $N \in \{100, 500, 1000\}$, number of quantile levels $M \in \{5, 10\}$, and $\lambda \in \Lambda$. Optimization was carried incrementally (across replications), and we discarded those replications that resulted in numerical errors until the effective number of replications $\texttt{nrep} = 200$ was reached.[3]

3.3. Evaluation Metrics

Following Medeiros & Mendes (2015) and Konzen & Ziegelmann (2016), we used a set of metrics to evaluate and compare variable selection performance between the studied methods:

- **FVCI**: Average fraction of variables correctly identified. To calculate this metric, for each replication, we sum the number of variables correctly included, and the number of variables correctly excluded and divide by the number of covariates in the model to obtain the fraction of correctly included and excluded covariates. Then, we take the average across the number of

[1] Notice that if the set of active covariates contains only the intercept for a certain $\lambda_1 > 0$, then this will also be the case for any $\lambda_2 \geqslant \lambda_1$. Thus, in our simulation algorithm, given the computational burden of the optimization procedure, in each replication, we only computed the estimators (incrementally on $\lambda \in \Lambda$) up to the point where all non-constant covariates were excluded.

[2] In each replication, the sample size $N = 100$ is obtained from the first 100 observations from $\big((X_1, Y_1), \ldots, (X_{N_{\max}}, Y_{N_{\max}})\big)$, and similarly for $N = 500$.

[3] Preliminarily we also implemented the "direct approach" with the ("non-grouped") LASSO and adaLASSO penalties, but the two additional methods ended up with execution errors in more than half of the replications and were discarded.

replications. A variable is *correctly included* if its population coefficient is non-zero and the optimization algorithm included it in the selection procedure. Similarly, it is *correctly excluded* if its population coefficient is zero and it was excluded by the procedure.

- **TMI**: True model included. For this metric, we count how many replications included all relevant covariates in the model. Subsequently, we divide this count by the number of replications to obtain the fraction.
- **FRVI**: Average fraction of relevant variables included. To make this average, for each replication, we sum the number of covariates correctly included in the model and divide by the number of relevant covariates. Then, we take the average of this fraction across all replications.
- **FIVE**: Average fraction of irrelevant variables excluded. Similarly to FRVI, for this metric we sum the number of covariates correctly excluded from the model and divide by the number of irrelevant covariates for each replication. Next, we take the average of this fraction across all replications.

Additionally, to assess and compare the quality of the studied estimators, we considered the following two criteria:

- $\text{MSE}(\widehat{\boldsymbol{\beta}}) = \sum_{d=1}^{D}\sum_{m=1}^{M}\texttt{nrep}^{-1}\sum_{r=1}^{\texttt{nrep}}(\widehat{\boldsymbol{\beta}}_{d,m}^{r} - \boldsymbol{\beta}_{d,m})^2$
- $\mathcal{L}(\widehat{\boldsymbol{\beta}}) = \sum_{m=1}^{M}\left(\texttt{nrep}^{-1}\sum_{r=1}^{\texttt{nrep}}\rho_{\tau_m}(Y_1^r - X_1^{r\intercal}\widehat{\boldsymbol{\beta}}_{:m}^{r})\right)$, drawing from the concept of *elicitability* as proposed by Gneiting (2011).

In the above, $\widehat{\boldsymbol{\beta}}^r$ denotes the estimator computed in the rth replication, Y_1^r and X_1^r are the first observations of Y and X, respectively, in the rth replication. Regarding the elicitability criterion, notice that

$$\mathcal{L}(\boldsymbol{b}) \equiv \sum_{m=1}^{M}\left(\sum_{r=1}^{\texttt{nrep}}\frac{\rho_{\tau_m}(Y_1^r - X_1^{r\intercal}\boldsymbol{b}_{:m})}{\texttt{nrep}}\right) \approx \sum_{m=1}^{M}\mathbb{E}\rho_{\tau_m}(Y - X^{\intercal}\boldsymbol{b}_{:m}) =: \mathcal{L}^*(\boldsymbol{b}),$$

with $\mathcal{L}^*(\boldsymbol{\beta}) \leqslant \mathcal{L}^*(\boldsymbol{b})$ for any $\boldsymbol{b}$; thus, estimators that attain lower values of $\mathcal{L}$ can be regarded as better.

3.4. Results

In this section, we report the results and provide an account of patterns observed in the studied scenarios. We begin by describing the three criteria used to select the tuning parameter λ, as well as how each method behaves in terms of these λ-selection criteria. This observation enlightens the performance of the evaluated methods, as the λ parameter is determinant for the degree of shrinkage. Next, we compare the methods' performance according to the metrics described in Section 3.3. Afterward, we compare the estimated $\widehat{\boldsymbol{\beta}}$ coefficient with the real β function of four covariates: the intercept ($D = 1$), variables with strong ($D = 5$)

and weak ($D = 15$) coefficients, and an irrelevant covariate ($D = 25$). Finally, we exemplify how the increase in N and M positively impact the model fit. For every presented result, we start by outlining the observations in the scenario with the largest sample size ($N = 1000$) and grid ($M = 10$), followed by the scenarios and results that deviate from the standards identified in the highest sample size and grid.

3.4.1. Selection of Tuning Parameter

As mentioned above, the simulation procedure stores the estimated $\widehat{\boldsymbol{\beta}} \equiv \widehat{\boldsymbol{\beta}}_\lambda$ for each λ evaluated, with $\lambda \in \Lambda$, where Λ is described in Section 3.2. We analyze the results after selecting, at each replication, the "optimal" λ according to the Bayesian information criterion (BIC) and Akaike information criterion (AIC), as well as a fixed λ that gives the best average outcome for a given metric across all replications, which we call the *Omni* criterion. Notice that the latter is unfeasible in real world applications.

For this study, the BIC (Schwarz, 1978) and AIC (Akaike, 1973) criteria follow Equation 3.7 from Sottile et al. (2020), namely

$$\mathrm{BIC}_\lambda = \log \widehat{R}\big(\widehat{\boldsymbol{\beta}}_\lambda\big) + \frac{1}{N}\log(N)\mathrm{df}_\lambda \tag{14}$$

and

$$\mathrm{AIC}_\lambda = \log \widehat{R}\big(\widehat{\boldsymbol{\beta}}_\lambda\big) + \frac{1}{N}2\mathrm{df}_\lambda, \tag{15}$$

with $\widehat{R}(\boldsymbol{b}) := \sum_{n=1}^{N}\sum_{m=1}^{M}\rho_{\tau_m}(Y_n - X_n^\intercal \boldsymbol{b}_{:,m}) + P(\boldsymbol{b}\boldsymbol{\varphi}^{-1})$. Notice that $\widehat{R}(\cdot)$ implicitly depends on λ through $P(\cdot)$. Here, df is the number of non-zero coefficients in the model, that is, the number of non-zero rows in $\widehat{\boldsymbol{\beta}}_\lambda$. We consider that a variable was removed from the model if the entire row $\widehat{\boldsymbol{\beta}}_{d,:}$ contains absolute values below a given tolerance ($1\mathrm{e}^{-4}$).

The λ selected in each replication following the BIC criterion is the one that minimizes (14) and the one selected by the AIC criterion minimizes (15). The *Omni* criterion for each evaluation metric calculates the results for every λ tested, fixing to all replications the one that provided the best average outcome for that particular method. For example, the *Omni* criteria for MSE chooses the λ that provided the lowest MSE across replications, while the *Omni* criteria for FRVI chooses the λ that resulted in a higher average fraction of relevant variables included across replications.

Figures 1 and 2 present the histogram of the selected λ's, for each method, using the BIC and AIC criteria, respectively. The dotted vertical lines represent the λ selected by the *Omni* criterion for the MSE metric, while the dashed vertical lines indicate the *Omni* λ designated to minimize the elicitability loss criterion across replications. We observe in both the BIC and AIC histograms that the methods using the adaptive penalization (adaLASSO, group adaLASSO, direct group adaLASSO) opt for higher λ values than the traditional LASSO methods, being adaLASSO the one that picks lower λ values from this set, especially in the

AIC criterion. The BIC criterion has the attribute of applying a higher penalty to additional parameters (Bishop, 2006), in fact, we can detect that it selects higher λ's as compared to the AIC. Looking at the selection from the *Omni* criteria, the λ that optimizes the MSE across replications is smaller on methods using adaLASSO and bigger on traditional LASSO. When the *Omni* criterion is accounting for lower elicitability loss, it opts for lower λ's in all evaluated methods.

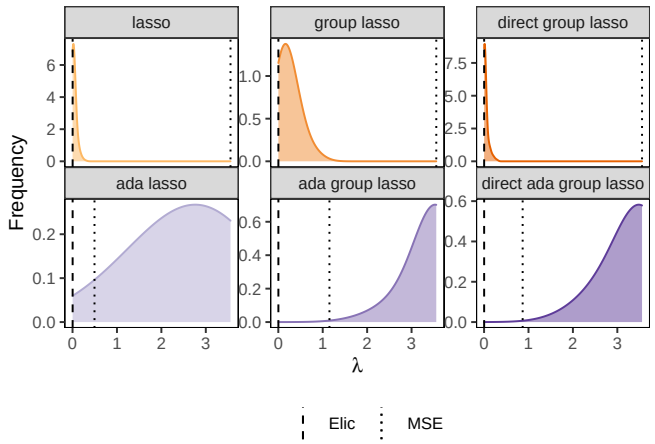

FIGURE 1: λ-selection pattern of BIC and *Omni* (MSE and Elicitability Loss) criteria for $N = 1000$ and $M = 10$. λ values chosen by BIC in each replication (histogram) and Omni choice across replications (vertical lines) for MSE (dotted line) and Elicitability Loss (dashed line) with $N = 1000$ and $M = 10$.

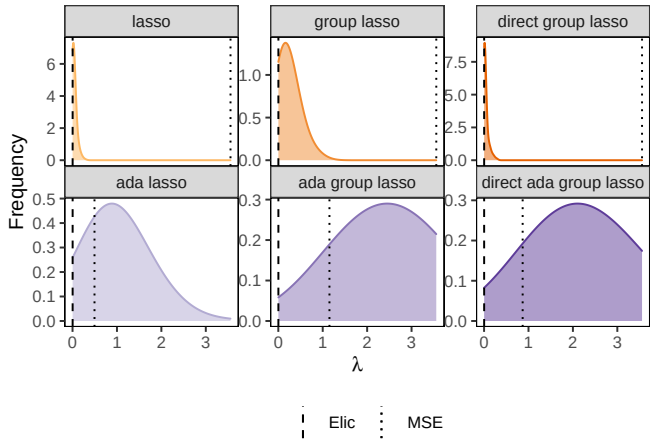

FIGURE 2: λ-selection pattern of AIC and *Omni* (MSE and Elicitability Loss) criteria for $N = 1000$ and $M = 10$. λ values chosen by AIC in each replication (histogram) and *Omni* choice across replications (vertical lines) for MSE (dotted line) and Elicitability Loss (dashed line) with $N = 1000$ and $M = 10$.

Similarly, we have the same histogram in Figures 3 and 4 comparing the BIC and AIC λ choices with the *Omni* criterion that favors the variable selection metrics. The solid vertical line with the down arrow represents the λ selected via the *Omni* criterion for the FIVE metric, the dashed line with a dot for FRVI, the dotted line with an X for FVCI, and the two dashed line with an up arrow for the TMI metric.

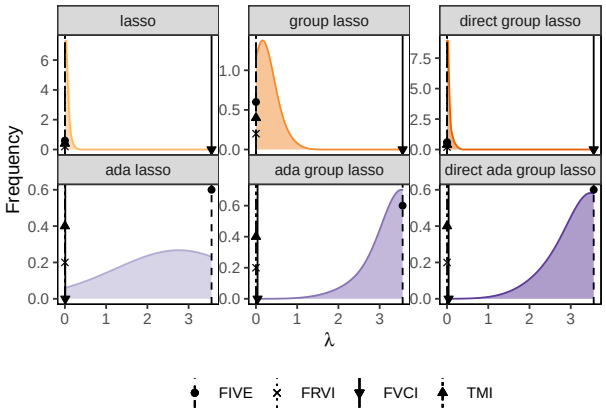

FIGURE 3: λ-selection pattern of BIC and *Omni* (FVCI, FRVI, TMI and FIVE) criteria for $N = 1000$ and $M = 10$. λ values chosen by BIC in each replication (histogram) and Omni choice across replications (vertical lines) for FVCI (dotted line), FRVI (dashed line), TMI (two dashed line) and FIVE (solid line) with $N = 1000$ and $M = 10$.

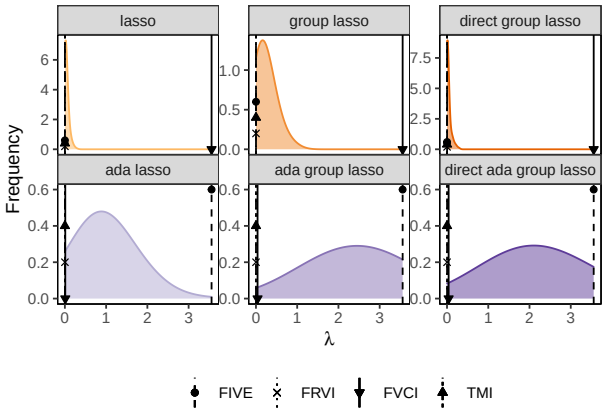

FIGURE 4: λ-selection pattern of AIC and *Omni* (FVCI, FRVI, TMI and FIVE) criteria for $N = 1000$ and $M = 10$. λ values chosen by AIC in each replication (histogram) and Omni choice across replications (vertical lines) for FVCI (dotted line), FRVI (dashed line), TMI (two dashed line) and FIVE (solid line) with $N = 1000$ and $M = 10$.

It is possible to notice that, for the non-adaptive methods (LASSO, group LASSO, direct group LASSO), the λ that optimizes the fraction of variables correctly included and correctly excluded at the same time is bigger than the λ's chosen by both BIC and AIC criteria, as well as the *Omni*-selected ones for the other variable selection metrics. However, when we look at the adaptive LASSO methods, the λ that performs better across all replications considering the fraction of irrelevant variables excluded (FIVE) has a higher value than the others. We can see that the histogram of the λ choice via BIC criterion often has a peak that coincides with the *Omni* selection of the FIVE metric, indicating that this criterion values the exclusion of irrelevant variables. As expected, when the metric accounted for is related to correctly including relevant variables (TMI and FRVI), the *Omni*-criterion always opts for the lowest λ option.

The λ-choice patterns are nearly unchanged when $N = 500$ and $M = 10$. However, the $N = 100$ scenario produces different outputs, as demonstrated in Figures 5 and 6. We see bigger λ values being chosen in the direct group LASSO and adaLASSO methods when using the BIC criterion to select the optimal tuning parameter. It is also noticeable that the *Omni* selection for MSE in adaptive penalization methods is higher as compared to the $N = 1000$, $M = 10$ scenario. The other scenarios can be reviewed in the On-line supplementary material.

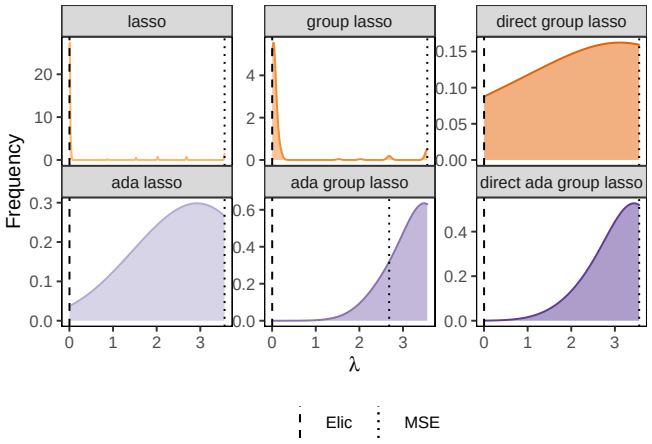

FIGURE 5: λ-selection pattern of BIC and *Omni* (MSE and Elicitability Loss) criteria for $N = 100$ and $M = 10$. λ values chosen by BIC in each replication (histogram) and *Omni* choice across replications (vertical lines) for MSE (dotted line) and Elicitability Loss (dashed line) with $N = 100$ and $M = 10$.

Figure 7 illustrates the impact of the λ-choice on variable selection by plotting, for each covariate on the vertical axis and each λ in the horizontal axis, how many replications have included that covariate in the model. The horizontal lines reflect the D_{strong} and D_{signal} values: variables below the dotted line have a strong coefficient, variables between them have a weak coefficient and variables above the dashed line are non-relevant. We can see that the methods without adaptive weights never remove the covariates for the λ values considered, while the adaptive

LASSO ones start excluding as λ increases. It is noticeable that variables with weak coefficients are often removed from the model in higher λ values together with the irrelevant ones. The same pattern is observed in other scenarios, however, the ones with sample size equal to 100 demonstrate a lighter blue color, indicating some replications have removed the variables across λ values (for other scenarios, see the On-line supplementary material.

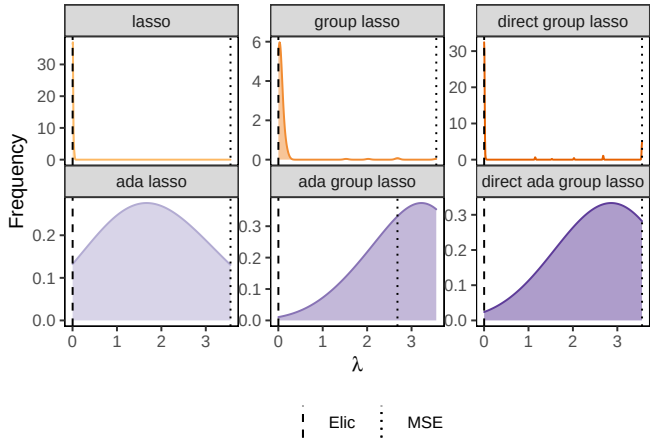

FIGURE 6: λ-selection pattern of AIC and *Omni* (MSE and Elicitability Loss) criteria for $N = 100$ and $M = 10$. λ values chosen by AIC in each replication (histogram) and *Omni* choice across replications (vertical lines) for MSE (dotted line) and Elicitability Loss (dashed line) with $N = 100$ and $M = 10$.

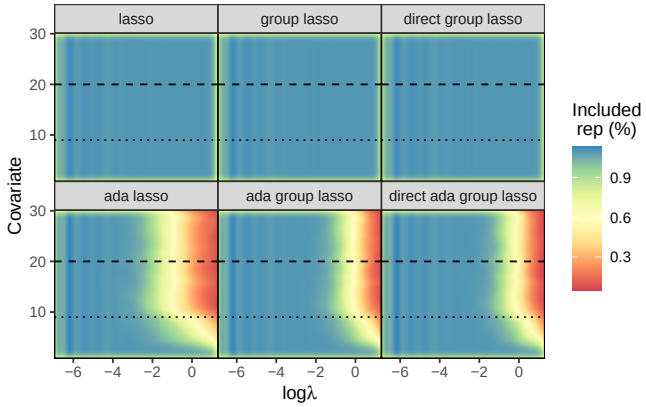

FIGURE 7: Number of times each variable was included across λ values for $N = 1000$ and $M = 10$. Horizontal axis: grid of evaluated λ's (log scale); vertical axis: model coefficients D; gradient from blue (highest - included) to red (lowest - excluded): proportion of replications coefficient was included in the model. Horizontal lines: D_{strong}(dotted) and D_{signal} (dashed) values.

3.4.2. Evaluation Results

Table 1 presents the results for all compared methods and evaluated metrics in the scenario with $N = 1000$ and $M = 10$. The method with adaLASSO penalization achieves the lowest MSE for all λ-selection criteria evaluated, however, it jeopardizes the variable selection metrics. As identified in Section 3.4.1, the adaLASSO penalization opts for lower λ values as compared to the other adaptive approaches but higher than the non-adaptive ones. The traditional LASSO penalization methods result in lower elicitability loss than the adaptive ones. When we look at the variable selection metrics, it is noticeable from the FRVI and FIVE metrics that the traditional LASSO methods never exclude irrelevant covariates, in other words they always include all of the 30 regressors in the model. On the other hand, the FRVI and TMI metrics indicate that none of the adaptive methods include the true model, being adaLASSO the one that includes the most in both BIC and AIC criteria. In the context of variable selection, the *Omni* criteria are not very meaningful, as they will maximize the λ when the metric evaluated is FIVE and minimize the λ when the metric is TMI or FRVI. Among the ada-penalized methods, adaLASSO outperforms in the FVCI metric, but excluding fewer variables as demonstrated in the FIVE metric, for both AIC and BIC. If we look at the count of metrics where the methods outperformed, group LASSO and direct group LASSO are the ones with better outcome in more metrics, but renouncing model shrinkage and not excluding irrelevant variables.

TABLE 1: Results for $N = 1000$ and $M = 10$. Results of all evaluated metrics for each method when $N = 1000$ and $M = 10$. In bold, the results with the best value. In grey, the results that numerically have the best result, but are not relevant in the context, since there weren't variables selected in that use case.

		MSE	Elic	FVCI	FRVI	TMI	FIVE
BIC	LASSO	0.1401	0.2867	0.6667	1	1	0
	adaLASSO	0.1071	0.2986	0.5098	0.3238	0.0100	0.8820
	gLASSO	0.1399	0.2866	0.6667	1	1	0
	gAdaLASSO	0.1195	0.2933	0.5043	0.2905	0	0.9320
	direct gLASSO	0.1408	0.2867	0.6667	1	1	0
	direct gAdaLASSO	0.1243	0.2935	0.4990	0.2775	0	0.9420
AIC	LASSO	0.1401	0.2867	0.6667	1	1	0
	adaLASSO	0.0826	0.2917	0.5695	0.5000	0.0600	0.7085
	gLASSO	0.1399	0.2866	0.6667	1	1	0
	gAdaLASSO	0.1019	0.2922	0.5273	0.3610	0	0.8600
	direct gLASSO	0.1408	0.2867	0.6667	1	1	0
	direct gAdaLASSO	0.1018	0.2913	0.5325	0.3790	0	0.83795
Omni	LASSO	0.0956	0.2867	0.6667	1	1	0
	adaLASSO	0.0637	0.2865	0.6667	1	1	0.9070
	gLASSO	0.1185	0.2867	0.6667	1	1	0
	gAdaLASSO	0.0776	0.2867	0.6667	1	1	0.9325
	direct gLASSO	0.0938	0.2867	0.6667	1	1	0
	direct gAdaLASSO	0.0797	0.2867	0.6667	1	1	0.9430

The pattern observed for $N = 1000$ and $M = 10$ is also observed when $N = 500$, for both tested τ-grid sizes. There is a variation in values in the third and fourth decimal places but the overall behavior pattern is maintained. The table of results for both scenarios is in the On-line supplementary material. When the sample size is smaller, $N = 100$, we observe some deviations from the standard. When $M = 10$, the direct group LASSO method performs very poorly in the TMI metric when using BIC criteria for λ-selection, as highlighted in Table 2. For this case, the average number of replications that includes the true model is less than 30%. It is worth recalling that this particular scenario also had a deviant pattern in the λ-choice, selecting higher values. The other metrics and methods follow the pattern observed in the scenarios with a higher sample size, but it is worth mentioning that the Elicitability Loss measurement has lower values overall and MSE values for non-weighted penalization methods are relatively higher. The $N = 100$ scenario with $M = 5$ presents the same pattern deviations as with $M = 10$, with the exception of the true model inclusion when using the BIC criteria, that has lower values for all evaluated methods as compared to the outcome of this metric in other scenarios, highlighted in Table 3. In this case, the direct group LASSO only includes the true model (TMI) in 35% of the replications when using AIC criteria.

Table 2: Results for $N = 100$ and $M = 10$. Results of all evaluated metrics for each method when $N = 100$ and $M = 10$. In bold, the results with the best value. In grey, the results that numerically have the best result, but are not relevant in the context, since there weren't variables selected in that use case. In red, the results that are significantly different from the $N = 1000$ and $M = 10$ scenario.

		MSE	Elic	FVCI	FRVI	TMI	FIVE
BIC	LASSO	1.1533	0.1948	0.6643	0.9942	0.9050	0.0045
	adaLASSO	0.2033	0.2992	0.4125	0.1735	0	0.8955
	gLASSO	1.1572	0.1954	0.6655	0.9965	0.9300	0.0040
	gAdaLASSO	0.2039	0.2815	0.4067	0.1455	0	0.9290
	direct gLASSO	0.5628	0.2219	0.6358	0.9040	0.2800	0.0095
	direct gAdaLASSO	0.2110	0.2804	0.4053	0.1428	0	0.9305
AIC	LASSO	1.2243	0.1935	0.6667	1	1	0
	adaLASSO	0.2288	0.2813	0.4395	0.2405	0.0050	0.8375
	gLASSO	1.2122	0.1953	0.6663	0.9990	0.9800	0.0010
	gAdaLASSO	0.2140	0.2762	0.4128	0.1638	0	0.9110
	direct gLASSO	1.1028	0.1998	0.6600	0.9740	0.8450	0.0320
	direct gAdaLASSO	0.2245	0.2745	0.4167	0.1728	0	0.9045
Omni	LASSO	0.435	0.1931	0.6667	1	1	0.0150
	adaLASSO	0.1975	0.1941	0.6667	1	1	0.8955
	gLASSO	0.5709	0.1933	0.6667	1	1	0.0040
	gAdaLASSO	0.2011	0.1937	0.6667	1	1	0.9290
	direct gLASSO	0.2917	0.1934	0.6667	1	1	0.1180
	direct gAdaLASSO	0.2083	0.1936	0.6667	1	1	0.9335

TABLE 3: Results for $N = 100$ and $M = 5$. Results of all evaluated metrics for each method when $N = 100$ and $M = 5$. In bold, the results with the best value. In red, the results that are significantly different from the $N = 1000$ and $M = 10$ scenario.

		MSE	Elic	FVCI	FRVI	TMI	FIVE
BIC	LASSO	0.2759	0.1016	0.6342	0.8900	0.2400	0.1225
	adaLASSO	0.1216	0.1332	0.4198	0.1845	0.0050	0.8905
	gLASSO	0.3388	0.0949	0.6507	0.9388	0.3750	0.0745
	gAdaLASSO	0.1116	0.1334	0.3867	0.1013	0	0.9575
	direct gLASSO	0.1189	0.1074	0.5850	0.7220	0	0.3110
	direct gAdaLASSO	0.1114	0.1309	0.3902	0.1060	0	0.9585
AIC	LASSO	0.6075	0.0938	0.6627	0.9865	0.9100	0.0150
	adaLASSO	0.1290	0.1228	0.4480	0.2668	0.0050	0.8105
	gLASSO	0.5580	0.0918	0.6602	0.9782	0.8000	0.0240
	gAdaLASSO	0.1118	0.1244	0.4115	0.1622	0	0.9100
	direct gLASSO	0.3103	0.0994	0.6128	0.8100	0.3500	0.2185
	direct gAdaLASSO	0.1148	0.1226	0.4183	0.1808	0	0.8935
Omni	LASSO	0.1465	0.0908	0.6667	1	1	0.1405
	adaLASSO	0.1170	0.0912	0.6672	1	1	0.9165
	gLASSO	0.2015	0.0908	0.6667	1	1	0.0825
	gAdaLASSO	0.1040	0.0911	0.6667	1	1	0.9625
	direct gLASSO	0.1157	0.0909	0.6667	1	1	0.3105
	direct gAdaLASSO	0.1071	0.0913	0.6667	1	1	0.9650

Figure 8 illustrates how the results reported in Section 3.4.1 and Table 1 reflect estimation, by plotting the true $\boldsymbol{\beta}$ and the estimated $\widehat{\boldsymbol{\beta}}$ for the first 50 Monte Carlo replications, using the BIC criterion for λ-selection. A similar figure contemplating the AIC criterion can be found in the On-line supplementary material. The dashed line represents the average values of the 50 replications plotted and the solid line represents the true parameter $\beta(\tau)$. We observe that all methods have more difficulty estimating the coefficients for higher quantiles, due to the nature of how the β values are generated. When looking at the intercept, a constant coefficient of 0.1, the traditional LASSO penalization methods (which we observed in Section 3.4.1 select lower λ values) follow the true $\beta(\tau)$ pattern, while the adaptive methods appear to have a bias. This particular result differs from the asymptotic theory for large coefficients in Zou (2006), that states the adaptive LASSO results in unbiased estimates. The same is observed when $D = 5$, which is from the group of strong coefficients. However, in this case, we see that the adaptive LASSO methods already remove this coefficient in most evaluated replications. The average values across the replications follow the real $\boldsymbol{\beta}$ line, especially on non-adaptive methods. When $D = 15$, which is part of the D_{weak} set of coefficients, the ada penalization methods remove the variable in most of the replications, while the non-adaptive ones do not. However, the latter brings a lot of instability towards the higher quantile levels and the average does not match the original $\boldsymbol{\beta}$ curve in higher quantiles. When $D = 25$, which is an irrelevant coefficient, the methods with adaptive weights remove this coefficient in most of the replications, as opposed to the traditional LASSO ones, which was identified in Section 3.4.2.

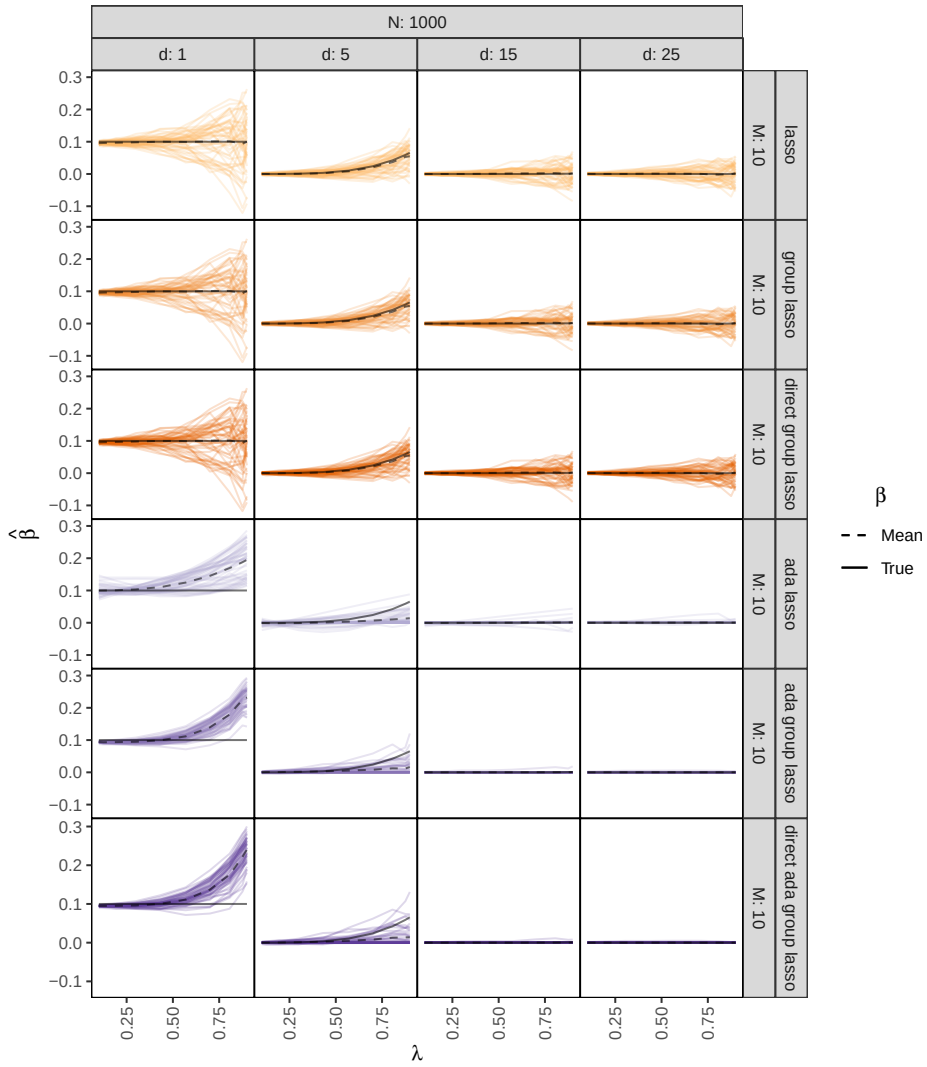

FIGURE 8: Plot of first 50 replications of β estimation for using BIC criterion for $N = 1000$ and $M = 10$. Columns: covariates - intercept ($d = 1$); variables with strong ($d = 5$) and weak ($d = 15$) coefficients; and an irrelevant covariate ($d = 25$). Rows: the six methods evaluated. Colorful lines: estimated functional coefficient $\tau \mapsto \widehat{\beta}_d(\tau)$ corresponding to the first 50 replications in our Monte Carlo study, according to the six considered methods. Dotted black line: average of the estimated functional coefficient $\tau \mapsto \widehat{\beta}_d(\tau)$ across replications. Solid black line: true parameter β_d.

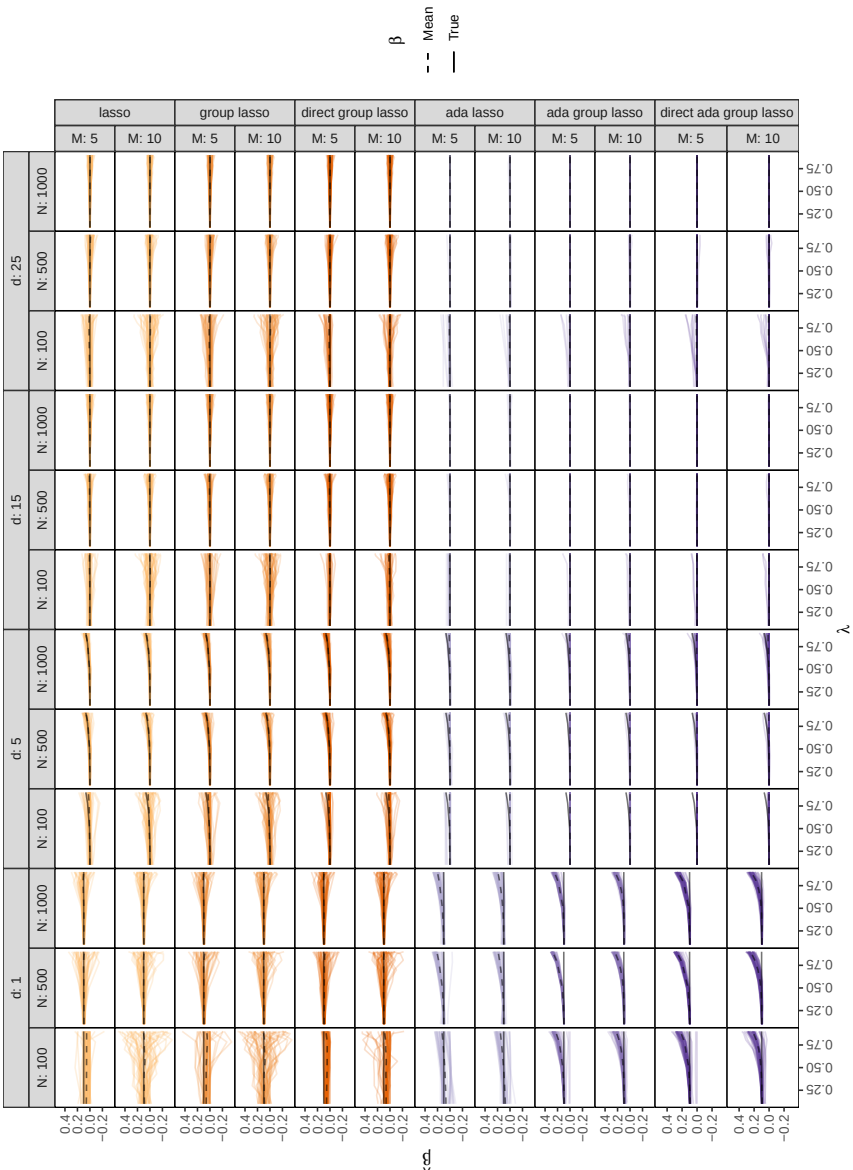

FIGURE 9: Plot of first 50 replications of β estimation for using BIC criteria. Columns: covariates - intercept ($d = 1$); variables with strong ($d = 5$) and weak ($d = 15$) coefficients; and an irrelevant covariate ($d = 25$). Rows: the six methods evaluated in each τ-grid size M. Colorful lines: estimated functional coefficient $\tau \mapsto \widehat{\beta}_d(\tau)$ corresponding to the first 50 replications in our Monte Carlo study, according to the six considered methods. Dotted black line: average of the estimated functional coefficient $\tau \mapsto \widehat{\beta}_d(\tau)$ across replications. Solid black line: true parameter β_d.

3.4.3. Variation Across Scenarios

We can see in Figure 9 how increasing the sample size improves $\widehat{\boldsymbol{\beta}}$ estimation under the BIC criterion, especially when we increase from $N = 100$ to $N = 500$. Similarly, we observe how the estimated $\widehat{\boldsymbol{\beta}}$ comes closer to the $\boldsymbol{\beta}$ when the grid size is bigger. A similar figure contemplating the AIC criterion can be found in the On-line supplementary material.

4. Final Discussion and Future Work

This work proposes a method for global variable selection and coefficient estimation in a (linear) quantile regression context through a single optimization procedure, applying the group adaLASSO regularization to achieve a meaningful selection of covariates. We waive the flexibility to choose the τ-grid in favor of a fair picture of the whole $\beta(\cdot)$ function, including function value disparities between grid points, by using Chebyshev interpolation. A remark about terminology is called for: although we name our method *global*, a more honest terminology would be to call it a *nearly global* approach for estimation and variable selection. This is because, in finite samples, it is always the case that the tails of the distribution (namely, the quantile levels $\tau < \delta$ and $\tau > 1 - \delta$) are left out of the estimation procedure. *De facto* global methods are only attainable (at least from a computational viewpoint) under strong parametric assumptions, as is the case of the generating processes considered by Frumento & Bottai (2016) and Sottile et al. (2020). We perform a Monte Carlo simulation study comparing six different optimization procedures, varying the objective function and penalization factors, in six different sample and quantile grid sizes scenarios.

Our findings demonstrate that each estimator studied displays different patterns for selecting the tuning parameter λ in the penalty factor, which is critical for the model selection and coefficient estimation. It was observed in the simulation study that the methods using adaptive LASSO penalization select larger λ values when compared to the ones using regular LASSO. This pattern is evidenced in the results: those methods without adaptive weights in the penalizing factor have a more conservative behavior in removing variables from the model. When comparing the grouped approaches with the traditional ones, we see that the grouped proposal is more effective in removing variables from the model. This is also observed when using the BIC criterion for λ-selection. For the studied data generation process, the direct approach was similar to the other methods. A word of caution is called for, however, before we jump to definitive conclusions: the scale of "reasonable" λ values may be widely distinct for different penalty factors, and we chose our effective grid Λ having in mind computational reasons (possibly at the expense of flexibility/specificity). Thus, the observed selection patterns are not granted to be comparable, which reiterates the necessity of deeper investigations regarding the λ choice. This sensitivity of the penalizing parameter to method specificities, and the consequent necessity of fine tuning, should be seen as a cautionary remark when comparing the different metrics as we did in Section

3.4.2. Indeed, when we rank the different methods using a "universal" grid of λ's, we may be too unfair to one of the methods, and too beneficial to another.

From a practical outlook, the inquiry of what distinguishes a "more suitable choice" for the regularization method is fundamentally tied to the specific aims of the researcher. For instance, the adaLASSO has shown to deliver a better balance between coefficient estimation, inclusion of true model, and exclusion of irrelevant covariates, especially when the tuning parameter is selected via the AIC criterion, due to picking intermediate values from the tested set of λ's. Indeed, this method appears to include only half of the relevant variables—thus, it may not be the best approach when including the correct model is paramount. On the other hand, if we wish to shrink the model as much as we can, then the BIC criterion, combined with an adaptive LASSO penalization approach, seems to be a good alternative. Tables 4 and 5 summarize the strengths of each approach.

TABLE 4: Comparing the strengths observed by each element of the methodology proposed according to evaluated metrics when $N \geqslant 500$. Adaptive penalization provides lower MSE and elicitability loss, and excludes the irrelevant covariates more often, using both AIC or BIC criteria to select the λ parameter. Grouped penalization provides lower elicitability loss and excludes the irrelevant covariates more often, for both λ criteria evaluated. Chebyshev interpolation provides lower MSE using both BIC and AIC, includes the true model regularly when using the BIC criterion and excludes the irrelevant covariates more often when the AIC criterion is used.

Recommendation summary for $N \geqslant 500$					
		MSE	Elicitability	Include model	Exclude irrelevant
Ada penalization	BIC	✓	✓		✓
	AIC	✓	✓		✓
Grouped penalization	BIC		✓		✓
	AIC		✓		✓
Interpolated	BIC	✓		✓	
	AIC	✓			✓

Throughout this research, the computational time to execute simulations, especially with regard to optimization, surfaced as a major challenge, even more so in settings with larger sample and grid sizes. This fact limited the amount of scenarios and variations to be evaluated, as well as further exploration based on preliminary findings. Opportunities to propose an algorithm to ameliorate execution time of simulations were preliminarily explored by the authors, inspired by the MM algorithm in Hunter & Lange (2000) and the so-called "η-trick" of Bach et al. (2012) and Mairal et al. (2014), but the results were not satisfying. It is noticeable that the optimization problems faced in the present framework are also connected to the computational problem of dealing with large matrices—thus, additional research on the statistical computing field would be valuable to unlock further exploration of scenarios and parametrization of the proposed methods. For future work, it is recommended to compare, in a more thorough manner, the $\beta(\cdot)$ approximation methodology developed here with the ones presented by Frumento & Bottai (2016) and Yoshida (2021), as well as different Data

TABLE 5: Comparing the strengths observed by each element of the methodology proposed according to evaluated metrics when $N = 100$. Adaptive penalization provides lower MSE and elicitability loss, and excludes the irrelevant covariates more often, using both AIC or BIC criteria to select the λ parameter. Grouped penalization provides lower MSE sing AIC criterion to select the λ parameter, as well as lower elicitability loss and more frequent irrelevant covariate exclusion for both BIC and AIC criteria. Chebyshev interpolation provides lower elicitability loss using the BIC criterion, includes the true model regularly when using both λ selection criteria and excludes the irrelevant covariates more often when the AIC criterion is used.

Recommendation summary for $N = 100$.					
		MSE	Elicitability	Include model	Exclude irrelevant
Ada penalization	BIC	✓	✓		✓
	AIC	✓	✓		✓
Grouped penalization	BIC		✓		✓
	AIC	✓	✓		✓
Interpolated	BIC		✓	✓	
	AIC			✓	✓

Generating Processes, including other functional forms, larger coefficients, etc., to assess the strengths of the proposed estimator in problematic scenarios as exemplified in Section 2. It is worth highlighting that the particular DGP studied in the present work provides values for the covariate vector on the same scale—in fact, aside from the constant regressor, the remaining ones are identically distributed. Hence, if one is to analyze data generated otherwise, covariate normalization is recommended. Likewise, we would want to explore scenarios where $D > N$, which is of interest in the variable selection literature. In particular, given the observed importance of tuning parameter selection, a wider range of criteria and deeper exploration on proposed metrics is desired. Applications to real world data would be interesting after an extensive evaluation of different β functional forms, to be more precise on the type of data this methodology can better contribute to. Last but not least, our method can potentially contribute to the literature on conditional density estimation (Fan et al., 1996; Spady & Stouli, 2020; Cattaneo et al., 2022) by exploring the well-known relation between the conditional probability density function of the response and the derivative of the corresponding conditional quantile function.

Acknowledgements

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

[Received: September 2025 — Accepted: November 2025]

References

Akaike, H. (1973), *Information Theory and an Extension of the Maximum Likelihood Principle*, Springer New York, New York, NY, pp. 199–213.

Azimli, A. (2020), 'The impact of covid-19 on the degree of dependence and structure of risk-return relationship: A quantile regression approach', *Finance Research Letters* **36**, 101648.

Bach, F., Jenatton, R., Mairal, J. & Obozinski, G. (2012), 'Optimization with sparsity-inducing penalties', *Foundations and Trends® in Machine Learning* **4**(1), 1–106.

Belloni, A. & Chernozhukov, V. (2011), 'l1-penalized quantile regression in high-dimensional sparse models', *The Annals of Statistics* **39**(1), 82 – 130.

Bishop, C. M. (2006), *Pattern Recognition and Machine Learning (Information Science and Statistics)*, Springer-Verlag, Berlin, Heidelberg, p. 217.

Bonaccorsi, G., Pierri, F., Cinelli, M., Flori, A., Galeazzi, A., Porcelli, F., Schmidt, A. L., Valensise, C. M., Scala, A., Quattrociocchi, W. & Pammolli, F. (2020), 'Economic and social consequences of human mobility restrictions under covid-19', *Proceedings of the National Academy of Sciences* **117**(27), 15530–15535.

Buchinsky, M. (1998), 'Recent advances in quantile regression models: A practical guideline for empirical research', *Journal of Human Resources* **33**(1), 88–126.

Cattaneo, M. D., Chandak, R., Jansson, M. & Ma, X. (2022), Boundary adaptive local polynomial conditional density estimators, arXiv preprint 2204.10359, arXiv.

Das, P. & Ghosal, S. (2018), 'Bayesian non-parametric simultaneous quantile regression for complete and grid data', *Computational Statistics & Data Analysis* **127**, 172–186.

Davino, C., Furno, M. & Vistocco, D. (2014), *Quantile Regression: Theory and Applications*, Wiley.

Fan, J., Yao, Q. & Tong, H. (1996), 'Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems', *Biometrika* **83**(1), 189–206.

Frumento, P. & Bottai, M. (2016), 'Parametric modeling of quantile regression coefficient functions', *Biometrics* **72**(1), 74–84.

Frumento, P., Bottai, M. & Fernández-Val, I. (2021), 'Parametric modeling of quantile regression coefficient functions with longitudinal data', *Journal of the American Statistical Association* **116**(534), 783–797.

Frumento, P. & Salvati, N. (2021), 'Parametric modeling of quantile regression coefficient functions with count data', *Statistical Methods & Applications* **30**(4), 1237–1258.

Fu, A., Narasimhan, B. & Boyd, S. (2020), 'CVXR: An R package for disciplined convex optimization', *Journal of Statistical Software* **94**(14), 1–34.

Gneiting, T. (2011), 'Making and evaluating point forecasts', *Journal of the American Statistical Association* **106**(494), 746–762.

Hashem, H., Vinciotti, V., Alhamzawi, R. & Yu, K. (2016), 'Quantile regression with group lasso for classification', *Advances in Data Analysis and Classification* **10**(3), 375–390.

Hunter, D. R. & Lange, K. (2000), 'Quantile regression via an MM algorithm', *Journal of Computational and Graphical Statistics* **9**(1), 60–77.

Koenker, R. (2000), 'Galton, Edgeworth, Frisch, and prospects for quantile regression in econometrics', *Journal of Econometrics* **95**(2), 347–374.

Koenker, R. (2004), 'Quantile regression for longitudinal data', *Journal of Multivariate Analysis* **91**(1), 74–89.

Koenker, R. (2005), *Quantile regression*, Econometric Society Monographs, Cambridge University Press.

Koenker, R. (2021), *quantreg: Quantile Regression*. R package version 5.86. https://CRAN.R-project.org/package=quantreg

Koenker, R. & Bassett, G. (1978), 'Regression Quantiles', *Econometrica* **46**(1), 33–50.

Koenker, R. & Hallock, K. F. (2001), 'Quantile regression', *Journal of Economic Perspectives* **15**(4), 143–156.

Konzen, E. & Ziegelmann, F. A. (2016), 'LASSO-Type Penalties for Covariate Selection and Forecasting in Time Series', *Journal of Forecasting* **35**(7), 592–612.

Li, Q., Lin, N. & Xi, R. (2010), 'Bayesian regularized quantile regression', *Bayesian Analysis* **5**(3), 533 – 556.

Li, Y. & Zhu, J. (2008), 'L1-norm quantile regression', *Journal of Computational and Graphical Statistics* **17**(1), 163–185.

Lu, H., Nie, P. & Qian, L. (2020), Do Quarantine Experiences and Attitudes Towards COVID-19 Affect the Distribution of Psychological Outcomes in China? A Quantile Regression Analysis, GLO Discussion Paper 512, Global Labor Organization, Essen.

Mairal, J., Bach, F. & Ponce, J. (2014), 'Sparse modeling for image and vision processing', *Foundations and Trends® in Computer Graphics and Vision* **8**(2-3), 85–283.

Man, R., Pan, X., Tan, K. M. & Zhou, W.-X. (2022), 'A unified algorithm for penalized convolution smoothed quantile regression'.

Medeiros, M. C. & Mendes, E. F. (2015), 'L_1-regularization of high-dimensional time-series models with flexible innovations', *Available at SSRN 2626507* .

Park, S. & He, X. (2017), 'Hypothesis testing for regional quantiles', *Journal of Statistical Planning and Inference* **191**, 13–24.

Quarteroni, A. & Valli, A. (1994), *Numerical Approximation of Partial Differential Equations*, Lecture Notes in Mathematics, Springer-Verlag.

Ruas, M., Street, A. & Fernandes, C. (2022), 'A multi-quantile regression time series model with interquantile Lipschitz regularization for wind power probabilistic forecasting', *Electric Power Systems Research* **209**, 107973.

Schwarz, G. (1978), 'Estimating the dimension of a model', *The Annals of Statistics* **6**(2), 461–464.

Sottile, G. & Frumento, P. (2022), 'Robust estimation and regression with parametric quantile functions', *Computational Statistics and Data Analysis* **171**, 107471.

Sottile, G., Frumento, P., Chiodi, M. & Bottai, M. (2020), 'A penalized approach to covariate selection through quantile regression coefficient models', *Statistical Modelling* **20**(4), 369–385.

Spady, R. & Stouli, S. (2020), 'Gaussian transforms modeling and the estimation of distributional regression functions'.

Tibshirani, R. (1996), 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society. Series B (Methodological)* **58**(1), 267–288.

Wang, H. & Leng, C. (2008), 'A note on adaptive group lasso', *Computational Statistics & Data Analysis* **52**(12), 5277–5286.

Wu, Y. & Liu, Y. (2009), 'Variable selection in quantile regression', *Statistica Sinica* **19**(2), 801–817.

Yoshida, T. (2021), 'Quantile function regression and variable selection for sparse models', *Canadian Journal of Statistics* **49**.

Yuan, M. & Lin, Y. (2006), 'Model selection and estimation in regression with grouped variables', *Journal of the Royal Statistical Society. Series B: Statistical Methodology* **68**(1), 49–67.

Zheng, Q., Peng, L. & He, X. (2015), 'Globally adaptive quantile regression with ultra-high dimensional data', *The Annals of Statistics* **43**(5), 2225 – 2258.

Zou, H. (2006), 'The Adaptive Lasso and Its Oracle Properties', *Journal of the American Statistical Association* **101**(476), 1418–1429.

Appendix. Proofs

Proof of Theorem 1. The proof is based on an inequality found in Quarteroni & Valli (1994). We adapt their notation in order to make the proof easier to follow. First, fix $\delta \in (0, 1/2)$ and $d \in \{1, \dots, D\}$, and set $K = M - 1$. Let f_δ denote the linear reparametrization $[-1, 1] \to [\delta, 1-\delta]$ defined through $2f_\delta(x) \coloneqq 1 + (1-2\delta)x$ for $x \in [-1, 1]$. Under the stated assumptions, the function $\psi \coloneqq \beta_d \circ f_\delta$ is an element of the Sobolev space[4] H^1_w with $w(x) \coloneqq (1-x^2)^{-1/2}$, $x \in [-1, 1]$. Indeed, ψ is continuous and bounded, hence square-integrable with respect to any finite measure on $[-1, 1]$, and its weak derivative $\psi^{(1)}$ coincides with the (almost everywhere defined, strong) derivative $\partial\psi = ((\partial\beta_d) \circ f_\delta) \cdot \partial f_\delta = ((\partial\beta_d) \circ f_\delta) \cdot (1 - 2\delta)/2$, and then Equation (4) ensures that $\partial\psi$ is square-integrable with respect to $w(x)\,\mathrm{d}x$. By inequality (4.3.42) in Quarteroni & Valli (1994), we have, for some constant $C_0 > 0$ that does not depend on ψ,

$$\sup_{-1 \leqslant x \leqslant 1} |\psi(x) - I_K\psi(x)| \leqslant \frac{C_0}{\sqrt{K}} \left(\int_{-1}^{1} \frac{\psi(x)^2}{\sqrt{1-x^2}}\,\mathrm{d}x + \int_{-1}^{1} \frac{\partial\psi(x)^2}{\sqrt{1-x^2}}\,\mathrm{d}x \right)^{\frac{1}{2}}, \quad \text{(A1)}$$

where $I_K\psi(\cdot) \coloneqq \sum_{k=0}^{K} \psi_k^* \cos\left(k \arccos(\cdot)\right)$ with

$$\psi_k^* \coloneqq \frac{2K^{-1}}{1 + \lfloor |\cos(\pi k K^{-1})| \rfloor} \sum_{j=0}^{K} \frac{\cos\left(\pi k j K^{-1}\right)}{1 + \lfloor |\cos(\pi j K^{-1})| \rfloor} \psi(x_j),$$

the x_j being defined through $x_j \coloneqq \cos(\pi j K^{-1})$ for $0 \leqslant j \leqslant K$. The remainder of the proof is just a matter of adjusting definitions: for $0 \leqslant k \leqslant K \equiv M - 1$, let $\varphi\colon [\delta, 1-\delta] \to \mathbb{R}^M$ have component functions

$$\varphi_{k+1}(\tau) \coloneqq \cos\left(k \arccos(f_\delta^{-1}(\tau))\right), \qquad \delta \leqslant \tau \leqslant 1 - \delta, \quad \text{(A2)}$$

put

$$\tau_{k+1} \coloneqq \frac{1}{2} + \frac{1-2\delta}{2} x_k, \quad \text{(A3)}$$

and let

$$\boldsymbol{\alpha}_{d,k+1} = \frac{2K^{-1}}{1 + \lfloor |\cos(\pi k K^{-1})| \rfloor} \sum_{j=0}^{K} \frac{\cos\left(\pi k j K^{-1}\right)}{1 + \lfloor |\cos(\pi j K^{-1})| \rfloor} \beta_d(\tau_{j+1}). \quad \text{(A4)}$$

Clearly $\psi_k^* = \boldsymbol{\alpha}_{d,k+1}$, as $\psi(x_j) \equiv \beta_d(\tau_{j+1})$, yielding the identity

$$(I_K\psi) \circ f_\delta^{-1}(\tau) = \sum_{k=0}^{K} \boldsymbol{\alpha}_{d,k+1} \varphi_{k+1}(\tau), \quad \delta \leqslant \tau \leqslant 1 - \delta.$$

Therefore, and noticing that $\psi \circ f_\delta^{-1} = \beta_d$, the equality

$$\sup_{-1 \leqslant x \leqslant 1} |\psi(x) - I_K\psi(x)| = \sup_{\delta \leqslant \tau \leqslant 1-\delta} |\psi \circ f_\delta^{-1}(\tau) - (I_K\psi) \circ f_\delta^{-1}(\tau)|$$

[4] H^1_w is the set of all real valued functions ψ on $[-1, 1]$ having a weak-derivative $\psi^{(1)}$ such that both ψ and $\psi^{(1)}$ are measurable and square-integrable with respect to the measure $w(x)\mathrm{d}x$.

yields the bound in (5), with $C(\beta, \delta)$ given implicitly in (A1) (sum along, or take the maximum with respect to d, if necessary, to get rid of the dependence of $C(\beta, \delta)$ on d). The fact that the functions $\varphi_1(\cdot), \ldots, \varphi_M(\cdot)$ are (linearly independent) polynomials is well known from the literature on Chebyshev interpolation. The validity of equality $\beta_d(\tau_m) = \sum_{\ell=1}^{M} \boldsymbol{\alpha}_{d\ell} \varphi_\ell(\tau_m)$ is a matter of direct verification. This completes the proof. ■

Proof of Corollary 1. Since the functions $\varphi_1, \ldots, \varphi_M$ are linearly independent polynomials, it follows that the M columns of $\boldsymbol{\varphi}$ are also linearly independent, each column being one of the former polynomials evaluated at the grid of points $\mathscr{T}$. Thus, from $\boldsymbol{\beta} = \boldsymbol{\alpha}\boldsymbol{\varphi}$, it follows that $\boldsymbol{\alpha} = \boldsymbol{\beta}\boldsymbol{\varphi}^{-1}$, as stated, and clearly $\boldsymbol{\alpha}$ is a minimizer of $\boldsymbol{a} \mapsto R(\boldsymbol{a}\boldsymbol{\varphi})$. Now, suppose $\boldsymbol{a}^*$ is another minimizer of the latter map. Since $R(\boldsymbol{\beta})$ is the unique minimum of R, it necessarily holds that $\boldsymbol{a}^*\boldsymbol{\varphi} = \boldsymbol{\beta}$. Since $\boldsymbol{\varphi}$ is the matrix of a one-to-one onto linear transformation from $\mathbb{R}^M \to \mathbb{R}^M$, the condition $(\boldsymbol{\alpha} - \boldsymbol{a}^*)\boldsymbol{\varphi} = \mathbf{0}$ forces $\boldsymbol{a}^* = \boldsymbol{\alpha}$. ■

Proof of Proposition 1. For $\tau \in [\delta, 1-\delta]$ we have, through Remark 2,

$$\begin{aligned} \left\|\widehat{\boldsymbol{\beta}}(\tau) - \boldsymbol{\beta}(\tau)\right\| &\leqslant \|\widehat{\boldsymbol{\alpha}}\varphi(\tau) - \boldsymbol{\alpha}\varphi(\tau)\| + \|\boldsymbol{\alpha}\varphi(\tau) - \boldsymbol{\beta}(\tau)\| \\ &\leqslant \|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\| \cdot |\varphi(\tau)| + \frac{C(\beta, \delta)}{\sqrt{M-1}}, \end{aligned}$$

with $|\varphi(\tau)|^2 = \sum_{\ell=1}^{M} \varphi_\ell(\tau)^2 \leqslant M$. This establishes (8), completing the proof. ■