Advanced algorithms for penalized quantile and composite quantile regression

Matthew Pietrosanu, Jueyu Gao, Linglong Kong, Bei Jiang & Di Niu

Computational Statistics

ISSN 0943-4062

Comput Stat
DOI 10.1007/s00180-020-01010-1

Springer

Computational Statistics
https://doi.org/10.1007/s00180-020-01010-1

ORIGINAL PAPER

Advanced algorithms for penalized quantile and composite quantile regression

Matthew Pietrosanu[1] · Jueyu Gao[1] · Linglong Kong[1] · Bei Jiang[1] · Di Niu[2]

Received: 7 March 2019 / Accepted: 25 June 2020

Abstract

In this paper, we discuss a family of robust, high-dimensional regression models for quantile and composite quantile regression, both with and without an adaptive lasso penalty for variable selection. We reformulate these quantile regression problems and obtain estimators by applying the alternating direction method of multipliers (ADMM), majorize-minimization (MM), and coordinate descent (CD) algorithms. Our new approaches address the lack of publicly available methods for (composite) quantile regression, especially for high-dimensional data, both with and without regularization. Through simulation studies, we demonstrate the need for different algorithms applicable to a variety of data settings, which we implement in the `cqrReg` package for R. For comparison, we also introduce the widely used interior point (IP) formulation and test our methods against the IP algorithms in the existing `quantreg` package. Our simulation studies show that each of our methods, particularly MM and CD, excel in different settings such as with large or high-dimensional data sets, respectively, and outperform the methods currently implemented in `quantreg`. The ADMM approach offers specific promise for future developments in its amenability to parallelization and scalability.

Keywords Adaptive lasso · Alternating direction method of multipliers · Coordinate descent · Interior point · Majorize minimization

1 Introduction

With recent rising interest in sparse regression for high-dimensional data, least squares regression with regularization—often via lasso penalty (Tibshirani 1996)—

Drs. Linglong Kong, Bei Jiang, and Di Niu are supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC).

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00180-020-01010-1) contains supplementary material, which is available to authorized users.

Extended author information available on the last page of the article

Published online: 12 July 2020

has become a focal point of computing scientists and statisticians in model selection procedures (He et al. 2016; Vidaurre et al. 2013). Furthermore, quantile regression has emerged as an alternative to traditional ordinary least squares methods with numerous advantages, including but not limited to higher efficiency with heavy-tailed error distributions, robustness against outlying data, and more informative insights into the distribution of the response under study (Koenker 2005).

Oracle model selection theory, introduced by Fan and Li (2001), illustrates optimal behaviour during model selection but is limited to the case where error variance is finite. In response, Zou and Yuan (2008) established composite quantile regression—a method to simultaneously model multiple quantile levels—that maintains desirable oracle properties even in the case of non-finite error variance. Beyond oracle model selection and the simultaneous modelling of multiple quantile levels, composite quantile regression also achieves a lower variance on estimated effects relative to quantile regression. These properties of composite quantile regression have proven attractive to many researchers who have widely applied this technique to improve the processing capabilities of artificial neural networks (Xu et al. 2017), provide an alternative to local polynomial regression (Kai et al. 2010), and smooth Harris chain stochastic processes (Li and Li 2016).

Applying existing optimization algorithms to (composite) quantile regression requires a non-trivial reformulation of the problem due to the non-linearity and non-differentiability of the loss and regularization terms of the objective function. The well-known `quantreg` package for R (Koenker 2017) uses an interior point (IP) approach for quantile and composite quantile regression, with native support for l_1 (lasso) regularization in only the former. Advanced IP algorithms in `quantreg`, e.g., using prediction-correction (Mehrotra 1992) for non-regularized quantile regression, have greatly improved upon earlier simplex methods. However, the time spent on matrix inversion in IP approaches (Chen and Wei 2005) motivates us to seek faster algorithms for quantile and composite quantile regression, particularly for high-dimensional data where regularization is required. Zou (2006), following the conjectures of Fan and Li (2001), showed lasso variable selection—currently the most commonly implemented penalty for quantile regression—to be inconsistent in certain situations and presented adaptive lasso regularization as a solution. Our work in the present paper is thus motivated by both a search for faster quantile regression algorithms as well as the lack of publicly available methods for adaptive lasso regularized quantile and composite quantile regression, particularly for high-dimensional data.

Our work in this paper is novel in its approach to quantile regression, composite quantile regression, and corresponding versions regularized by an adaptive lasso penalty using three different algorithms. First, we present an alternating direction method of multipliers (ADMM) approach that breaks up the model estimation problem into simpler convex optimization problems that can be solved in parallel (Boyd et al. 2011). Second, we give a majorize-minimization (MM) approach that iteratively minimizes a majorization, a particular differentiable approximation of the objective function containing both the quantile loss and penalty terms (Hunter and Lange 2000). Third, we detail a coordinate descent (CD) method that uses observations in a greedy algorithm to iteratively select and update individual model parameters while holding others constant (Wu and Lange 2008). For the sake of comparison, we also discuss

an IP formulation of the problem that seeks to minimize both loss and regularization functions after starting within rather than on the boundary of the feasible set (Koenker 2005). In numerical simulations, we compare our approaches to the advanced IP methods present in the `quantreg` package. We implement the proposed methods using the publicly available `cqrReg` package for R (Gao and Kong 2015), which performs computations in C++ and links back to R via the `Rcpp` (Eddelbuettel and François 2011) and `RcppArmadillo` (Eddelbuettel and Sanderson 2014) packages for increased computational efficiency. The results of these simulations suggest that our approaches generally improve upon `quantreg`'s computation time with roughly the same level of estimation error for the range of quantile regression problems considered. We find that the MM approach to non-regularized composite quantile regression greatly outperforms the other three methods in terms of computation time and that the CD method excels in regularized (composite) quantile regression with high-dimensional data. Our ADMM approach was at least comparable (in terms of computation time and estimate error) in most simulations performed but holds the promise of further improvement and scalability with distributed computing and parallelization. Indeed, ADMM has recently been explored in the context of penalized quantile regression for big data as well as in sparse settings (Yu and Lin 2017; Gu et al. 2018). Our new implementations provide users with new algorithms for quantile and composite quantile regression with competitive runtime in different data settings, all with comparable estimation error.

The rest of this article is structured as follows. Section 2 presents quantile regression, starting with relevant notation in Sect. 2.1, followed by the description of our approaches to quantile regression using the ADMM, MM, and CD algorithms in Sect. 2.2 through 2.4. Sect. 3 continues with composite quantile regression, including relevant notation and commentary on the extension from quantile to composite quantile regression for our ADMM, MM, and CD methods. Numerical simulation results are presented in Sect. 4 and discussed in Sect. 5.

2 Quantile regression

In this section, we present the proposed ADMM, MM, and CD methods for quantile regression with adaptive lasso regularization. We refer interested readers to the online supplementary appendix for implementations of the non-regularized problems and further details (omitted for brevity) on our proposed methods. For completeness in the upcoming simulations, a basic IP formulation is also given in the online appendix.

2.1 Background and notation

We first introduce the necessary background and notation to be used throughout this paper regarding quantile regression, both with and without adaptive lasso regularization (Wu and Liu 2009; Zou 2006). We are concerned with the linear model

$$y = b_0 + \boldsymbol{x}^T\boldsymbol{\beta} + \varepsilon,$$

where we wish to estimate the level τ (for some $\tau \in (0, 1)$) conditional quantile of $y \in \mathbb{R}$ given $\boldsymbol{x} \in \mathbb{R}^p$, given by $b_0 + \boldsymbol{x}^T\boldsymbol{\beta} + b_\tau^\varepsilon$, where b_τ^ε is the (assumed unique) level τ quantile of the error distribution of ε, independent of $\boldsymbol{x}$ (Zou and Yuan 2008).

For a fixed quantile level $\tau \in (0, 1)$, define the quantile loss function, for any $t \in \mathbb{R}$, by $\rho_\tau(t) = \tau t_+ + (1 - \tau)t_-$, where $t_+ = \max\{t, 0\}$ and $t_- = \max\{-t, 0\}$. Given a design matrix $\boldsymbol{X} = [\boldsymbol{x}_1|\ldots|\boldsymbol{x}_n]^T \in \mathbb{R}^{n\times p}$ and response variable vector $\boldsymbol{Y} = (y_1, \ldots, y_n)^T \in \mathbb{R}^n$, adaptive lasso regularized quantile regression estimates are obtained as

$$(\hat{b}_0, \hat{\boldsymbol{\beta}}) = \underset{b_0\in\mathbb{R},\, \boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \sum_{i=1}^{n} \rho_\tau(y_i - b_0 - \boldsymbol{x}_i^T\boldsymbol{\beta}) + p_\lambda(|\boldsymbol{\beta}|),$$

where $\lambda > 0$ is a regularization parameter, $p_\lambda(|\boldsymbol{\beta}|) = \lambda \sum_{j=1}^{p} {}^{|\beta_j|}/_{|\beta_j^{\mathrm{QR}}|^2}$ is the adaptive lasso penalty, and $\boldsymbol{\beta}^{\mathrm{QR}} = (\beta_1^{\mathrm{QR}}, \ldots, \beta_p^{\mathrm{QR}})^T \in \mathbb{R}^p$ is the estimator (without intercept) obtained from non-regularized quantile regression (Koenker and Bassett 1978; Koenker 2005)—that is, the estimator in the problem with $\lambda = 0$.

Define the residuals for quantile regression by $r_i = r_i(b_0, \boldsymbol{\beta}) = y_i - b_0 - \boldsymbol{x}_i^T\boldsymbol{\beta}$, for $i = 1, \ldots, n$. For the ease of notation throughout this section, we sometimes assume that a design matrix $\boldsymbol{X}$ has an appropriate column for the intercept term of the model. Where intercepts are accounted for in the design matrix, the parameter vector $\boldsymbol{\beta}$ will be taken to include the corresponding intercept terms such that $\boldsymbol{\beta} = (b_0, \beta_1, \ldots, \beta_p)^T \in \mathbb{R}^{p+1}$. This will be made clear by the dimension of $\boldsymbol{\beta}$. Throughout this paper, p will always refer to the number of covariate parameters and β_j, for $j = 1, \ldots, p$, will always refer to a covariate effect and never an intercept term.

2.2 Alternating direction method of multipliers algorithm

Although developed in the 1960s and 1970s (Hestenes 1969; Gabay and Mercier 1976), interest in the ADMM algorithm was renewed with the findings of Boyd et al. (2011) and Lin et al. (2010). These studies demonstrate the ADMM algorithm's relative efficiency in solving optimization problems with large data sets, particularly when non-smooth terms are present in the objective function. This method has found notable use in quantile regression where the quantile loss and regularization term (if present) are not differentiable (Boyd et al. 2011; Kong et al. 2015; Zhang et al. 2017). For brevity, a general formulation of the ADMM algorithm is available in the online supplementary appendix. We apply the ADMM algorithm (Boyd et al. 2011) by reformulating regularized quantile regression as the convex optimization problem

$$\begin{aligned} \min_{\boldsymbol{\beta}\in\mathbb{R}^{p+1}} \quad & \sum_{i=1}^{n} \rho_\tau(r_i) + p_\lambda(|\boldsymbol{\beta}|) \\ \text{subject to} \quad & \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{r} = \boldsymbol{Y}, \end{aligned}$$

where $\boldsymbol{r}$ is a vector of residuals and where the intercept term is accounted for in both $\boldsymbol{\beta}$ and $\boldsymbol{X}$. We solve this problem using the ADMM iteration scheme (Boyd et al. 2011)

$$\begin{aligned}
\boldsymbol{r}^{(t+1)} &= \underset{\boldsymbol{r}\in\mathbb{R}^n}{\arg\min} \sum_{i=1}^{n} \rho_\tau(r_i) + \frac{\rho}{2}||\boldsymbol{Y} - \boldsymbol{r} - \boldsymbol{X}\boldsymbol{\beta}^{(t)} + \boldsymbol{u}^{(t)}/\rho||_2^2 \\
\boldsymbol{\beta}^{(t+1)} &= \underset{\boldsymbol{\beta}\in\mathbb{R}^{p+1}}{\arg\min} \frac{\rho}{2}||\boldsymbol{Y} - \boldsymbol{r}^{(t+1)} - \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}^{(t)}/\rho||_2^2 + p_\lambda(|\boldsymbol{\beta}|) \\
\boldsymbol{u}^{(t+1)} &= \boldsymbol{u}^{(t)} + \rho(\boldsymbol{Y} - \boldsymbol{r}^{(t+1)} - \boldsymbol{X}\boldsymbol{\beta}^{(t+1)}),
\end{aligned}$$

where $\boldsymbol{u}$ is the rescaled Lagrange multiplier and $\rho > 0$ is a penalty parameter. For reference, ρ is chosen to be 1.2 by Boyd et al. (2011). The update for $\boldsymbol{r}$ can be written in a closed form as $S_{1/\rho}\big(\boldsymbol{c} - (2\boldsymbol{\tau}_{n\times 1} - \boldsymbol{1}_{n\times 1})/\rho\big)$ where $\boldsymbol{c} = \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}^{(t)} + \boldsymbol{u}^{(t)}/\rho$ and, for $a \in \mathbb{R}$, the soft thresholding operator $S_a : \mathbb{R}^m \to \mathbb{R}^m$ is defined component-wise via $(S_a(\boldsymbol{v}))_i = (v_i - a)_+ - (-v_i - a)_+$. Similarly, the update for $\boldsymbol{\beta}$ does not have a closed form but can be viewed as a least squares optimization problem with adaptive lasso penalty. We implement existing numerical methods to solve this problem and update $\boldsymbol{\beta}$.

Let $\boldsymbol{X}_*$ and $\boldsymbol{\beta}_*$ be $\boldsymbol{X}$ and $\boldsymbol{\beta}$ with the intercept term removed and $\boldsymbol{b}$ a vector of intercepts $(b_0)_{n\times 1}$. A generic stopping condition for the algorithm can be defined in terms of the primal and dual residuals $\boldsymbol{r}_{\text{primal}}^{(t+1)}$ and $\boldsymbol{r}_{\text{dual}}^{(t+1)}$, respectively, with the stopping conditions $||\boldsymbol{r}_{\text{primal}}^{(t+1)}||_2 \le \varepsilon_{\text{primal}}$ and $||\boldsymbol{r}_{\text{dual}}^{(t+1)}||_2 \le \varepsilon_{\text{dual}}$. In this regularized setting, we have (from the general ADMM algorithm) that

$$\begin{aligned}
\boldsymbol{r}_{\text{primal}}^{(t+1)} &= \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}^{(t+1)} - \boldsymbol{r}^{(t+1)} \\
\boldsymbol{r}_{\text{dual}}^{(t+1)} &= \rho\boldsymbol{X}_*^T(\boldsymbol{r}^{(t+1)} - \boldsymbol{r}^{(t)}) \\
\varepsilon_{\text{primal}} &= \sqrt{n}\varepsilon_{\text{abs}} + \varepsilon_{\text{rel}} \max\{||\boldsymbol{X}_*\boldsymbol{\beta}_*^{(t+1)}||_2^2, ||\boldsymbol{r}^{(t+1)}||_2^2, ||\boldsymbol{b} - \boldsymbol{Y}||_2^2\}, \\
\varepsilon_{\text{dual}} &= \sqrt{p}\varepsilon_{\text{abs}} + \varepsilon_{\text{rel}}||\boldsymbol{X}^T\boldsymbol{u}^{(t+1)}||_2^2,
\end{aligned}$$

with possible tolerance values $\varepsilon_{\text{abs}} = 10^{-4}$ and $\varepsilon_{\text{rel}} = 10^{-2}$, respectively (Boyd et al. 2011).

2.3 Majorize-minimization algorithm

The use of majorizing functions to solve minimization problems has been well-studied in the statistical literature for many years since Ortega and Rheinboldt (1970). It was not until a later time, however, that the general MM framework was put forward by Hunter and Lange (2000). In general, MM can refer to majorize-minimization or minorize-maximization, depending on whether the problem at hand is a minimization or maximization problem, respectively. MM algorithms operate iteratively by constructing an auxiliary function $g_t(\cdot|\boldsymbol{\beta}^{(t)})$ using a solution $\boldsymbol{\beta}^{(t)}$ for the current iteration that will simultaneously optimize the original objective function f. In the case of a minimization problem, this function is called a majorizer and must satisfy

$g_t(\boldsymbol{\beta}|\boldsymbol{\beta}^{(t)}) \geq f(\boldsymbol{\beta})$ for all $\boldsymbol{\beta}$ of interest and $g_t(\boldsymbol{\beta}^{(t)}|\boldsymbol{\beta}^{(t)}) = f(\boldsymbol{\beta}^{(t)})$. Arguably, the most well-known application of an MM method is in the expectation-maximization (EM) algorithm (Dempster et al. 1976) for maximum likelihood estimation. MM has also been applied in various areas of research, e.g., regression, survival analysis, discriminant analysis, and quantile regression (Hunter and Lange 2004). We use the MM algorithm developed by Hunter and Lange (2000) and Hunter and Li (2005) to solve the quantile regression problem with adaptive lasso regularization.

We first construct a function $\rho_\tau^\varepsilon(r)$ based on some perturbation parameter $\varepsilon > 0$ to approximate the fidelity portion $\sum_{i=1}^n \rho_\tau(r_i)$ of the objective function. For any $r \in \mathbb{R}$, define $\rho_\tau^\varepsilon(r) = \rho_\tau(r) - \frac{\varepsilon}{2}\ln(\varepsilon + |r|)$ so that the fidelity can be approximated by $\sum_{i=1}^n \rho_\tau^\varepsilon(r_i)$. At the t-th iteration, for each residual value $r_i^{(t)} = r_i^{(t)}(\boldsymbol{\beta}^{(t)})$, we have that $\rho_\tau^\varepsilon(r)$ is majorized by the quadratic function

$$\xi_\tau^\varepsilon(r|r_i^{(t)}) = \frac{1}{4}\left[\frac{r^2}{\varepsilon + |r_i^{(t)}|} + (4\tau - 2)r + c\right],$$

for some solvable constant c that satisfies the equation $\xi(r_i^{(t)}|r^{(t)}) = \rho_\tau^\varepsilon(r^{(t)})$. Given λ, $\boldsymbol{\beta}^{\mathrm{QR}}$, and an initial value $\boldsymbol{\beta}^{(0)} = (\beta_1^{(0)}, \ldots, \beta_p^{(0)})$ for $\boldsymbol{\beta}$, we can locally approximate the penalty $p_\lambda(|\boldsymbol{\beta}|)$ as a quadratic function. This yields a majorizer of the objective function (Hunter and Li 2005),

$$Q^\varepsilon(\boldsymbol{\beta}|\boldsymbol{\beta}^{(t)}) = \sum_{i=1}^n \xi_\tau^\varepsilon(r_i|r_i^{(t)}) + \lambda\sum_{j=1}^p \frac{1}{|\beta_j^{\mathrm{QR}}|^2}\left[|\beta_j^{(t)}| + \frac{\left(\beta_j^2 - (\beta_j^{(t)})^2\right)\mathrm{sgn}(\beta_j^{(t)})}{2|\beta_j^{(t)}| + \varepsilon|}\right].$$

For the t-th iteration of the algorithm, given an updated value $\boldsymbol{\beta}^{(t)}$ for $\boldsymbol{\beta}$, we minimize the quadratic function $Q^\varepsilon(\cdot|\boldsymbol{\beta}^{(t)})$ using a Newton-Raphson iterative method. The argument minimum is used to update $\boldsymbol{\beta}^{(t)}$ and can be used to decide when to terminate the algorithm. For our purposes, we use tolerance 10^{-3}.

2.4 Coordinate descent algorithm

Coordinate descent (CD) algorithms are iterative procedures that generally fix some components of the argument vector in an optimization problem and solve the resulting subproblem in terms of the unfixed components. CD methods have a long-standing history (Ortega and Rheinboldt 1970) and their convergence properties are well-documented (Luo and Tseng 1992; Tseng 2001). The simplest CD algorithms allow for exactly one unfixed variable per iteration and search for a subproblem solution along a line, while others will search along a hyperplane by allowing multiple unfixed components. Most implementations use the latter in a block coordinate descent method. CD methods have been developed extensively, particularly for non-differentiable, non-convex objective functions, permitting the use of regularization functions such as lasso (l_1) and ridge (l_2) penalties (Tseng 2001; Friedman et al. 2010).

To implement quantile regression with adaptive lasso regularization, we use an extended version of the greedy CD method put forward by Edgeworth and, more

recently, further developed by Wu and Lange (2008). This requires us to reformulate the quantile objective function. In each iteration, for fixed $\boldsymbol{\beta} \in \mathbb{R}^p$, replace b_0 by the level-τ sample quantile of the residuals $y_i - \boldsymbol{X}_i^T\boldsymbol{\beta}$ for $i = 1, \ldots, n$: this will necessarily drive the value of the objective function downwards. Define $\Theta_i = \rho_\tau(r_i)$ for $i = 1, \ldots, n$. For $m = 1, \ldots, p$, rewrite the loss function as

$$L(b_0, \boldsymbol{\beta}) = L_m(b_0, \boldsymbol{\beta}) = \sum_{i=1}^{n} |x_{im}| \left| \frac{y_i - b_0 - \sum_{j=1,\, j\neq m}^{p} x_{ij}\beta_j}{x_{im}} - \beta_m \right| \cdot \Theta_i + p_\lambda(|\boldsymbol{\beta}|)$$

and apply the CD algorithm. For each fixed m, define $z_i = \frac{1}{x_{im}}\big(y_i - b_0 - \sum_{j=1,\, j\neq m}^{p} x_{ij}\beta_j\big)$ if $r_i \geq 0$ and $z_i = 0$ if $r_i < 0$. We sort z_i, for $i = 1, \ldots, n$, and update β_m to the value of the i^*-th order statistic $z_{(i^*)}$ satisfying

$$\sum_{j=1}^{i^*-1} w_{(j)} < \frac{1}{2}\sum_{j=1}^{n} w_{(j)} \qquad \text{and} \qquad \sum_{j=1}^{i^*} w_{(j)} \geq \frac{1}{2}\sum_{j=1}^{n} w_{(j)},$$

where $w_i = |x_{im}| \cdot \theta_i$ if $r_i \geq 0$ and $w_i = {}^{\lambda}/_{|\beta_m^{\mathrm{QR}}|^2}$ if $r_i < 0$. In other words, using the weights w_i, the selected $z_{(i^*)}$ is the weighted median of all z_i (for the fixed value of m). At the end of each iteration, check for the convergence of $\boldsymbol{\beta}$ using the selected stopping criteria. Here, we use an absolute value difference threshold of 10^{-3}.

3 Composite quantile regression

In this section, we present an extension from quantile to composite quantile regression for the proposed ADMM, MM, and CD algorithms. We only show results for the case with adaptive lasso regularization. Readers interested in the non-regularized case are referred to the online supplementary appendix where more details and a similar extension for a basic IP formulation are given. With regards to the available `quantreg` package for R (Koenker 2017), we note that non-regularized composite quantile regression has only recently been implemented using an IP algorithm and that a regularized version is currently not natively available without further reformulation of the problem.

Composite quantile regression (Zou and Yuan 2008) simultaneously estimates a sequence of K conditional quantiles of y given $\boldsymbol{X}$ at levels $0 < \tau_1 < \tau_2 < \cdots < \tau_K < 1$. Under the same linear model as before, these conditional quantiles are given by $b_0 + \boldsymbol{X}^T\boldsymbol{\beta} + b_k^{\varepsilon}$, where b_k^{ε} is the (assumed unique) level τ_k quantile of the error distribution of ε, again assumed independent to be independent of $\boldsymbol{X}$. Unlike K independent quantile regression models, the composite model assumes the same covariate effects across the K quantile levels. Adaptive lasso regularized composite quantile regression estimates are obtained as

$$(\hat{b}_1, \ldots, \hat{b}_K, \hat{\boldsymbol{\beta}}^{\mathrm{CQR}}) = \underset{b_1,\ldots,b_K\in\mathbb{R},\ \boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \sum_{k=1}^{K}\sum_{i=1}^{n} \rho_\tau(y_i - b_k - \boldsymbol{X}_i^T\boldsymbol{\beta}) + p_\lambda(|\boldsymbol{\beta}|),$$

where $\lambda > 0$ is a regularization parameter, $p_\lambda(|\boldsymbol{\beta}|) = \lambda \sum_{j=1}^{p} {}^{|\beta_j|}/_{|\beta_j^{\text{CQR}}|^2}$, and $\boldsymbol{\beta}^{\text{CQR}}$ is the solution (without intercepts) to the non-regularized composite quantile regression problem. To extend the residual notation defined before, let $r_{ik} = y_i - b_k - \boldsymbol{x}_i^T \boldsymbol{\beta}$, for $i = 1, \ldots, n$ and $k = 1, \ldots, K$. Zou and Yuan (2008) impose regularity conditions to ensure the asymptotic normality of the unregularized composite quantile estimates: the authors note these are essentially the same as those in standard quantile regression Koenker (2005).

The extension from quantile to composite quantile regression is relatively straightforward: we need only accommodate additional quantile levels and intercept terms. Since the composite quantile case only adds more intercept parameters, the penalty term remains unchanged. For explicit details on our methods for regularized composite quantile regression in the ADMM, MM, and CD approaches, refer to the online supplementary appendix.

To extend the ADMM method, we generate a new design matrix $\boldsymbol{X}^* \in \mathbb{R}^{nK \times (p+K)}$ by "stacking" the design matrices for each quantile level and adjusting all input accordingly. Written formally,

$$\boldsymbol{X}^*_{nK\times(p+K)} = \begin{bmatrix} [1\,0\,0\,\cdots\,0]\ \boldsymbol{X} \\ [0\,1\,0\,\cdots\,0]\ \boldsymbol{X} \\ [0\,0\,1\,\cdots\,0]\ \boldsymbol{X} \\ \vdots\ \vdots\ \vdots\ \ \vdots\ \ \vdots \\ [0\,0\,0\,\cdots\,1]\ \boldsymbol{X} \end{bmatrix}, \boldsymbol{Y}^*_{nK\times 1} = \begin{bmatrix} \boldsymbol{Y} \\ \boldsymbol{Y} \\ \boldsymbol{Y} \\ \vdots \\ \boldsymbol{Y} \end{bmatrix}, \boldsymbol{b}^* = \begin{bmatrix} (b_1)_{n\times 1} \\ (b_2)_{n\times 1} \\ (b_3)_{n\times 1} \\ \vdots \\ (b_K)_{n\times 1} \end{bmatrix}, \boldsymbol{\tau}^* = \begin{bmatrix} (\tau_1)_{n\times 1} \\ (\tau_2)_{n\times 1} \\ (\tau_3)_{n\times 1} \\ \vdots \\ (\tau_K)_{n\times 1} \end{bmatrix},$$

where, for example, $[\mathbf{1}\,\mathbf{0}\,\mathbf{0}\,\cdots\,\mathbf{0}]$ denotes the $n \times K$ matrix with rows $(1,\ 0,\ \cdots,\ 0)^T \in \mathbb{R}^K$. The methods presented in Sect. 2.2 for quantile regression then apply after replacing $\boldsymbol{X}$, $\boldsymbol{Y}$, $\boldsymbol{b}$, and τ with $\boldsymbol{X}^*$, $\boldsymbol{Y}^*$, $\boldsymbol{b}^*$, and $\boldsymbol{\tau}^*$, respectively. After replacement, the optimization problem becomes

$$\begin{aligned} \min_{\beta \in \mathbb{R}^{p+K}} \quad & \sum_{k=1}^{K} \sum_{i=1}^{n} \rho_{\tau_k}(r_{ik}) + p_\lambda(|\boldsymbol{\beta}|) \\ \text{subject to} \quad & \boldsymbol{X}^* \boldsymbol{\beta} + \boldsymbol{r} = \boldsymbol{Y}^*. \end{aligned}$$

With these changes, the explicit update scheme for ADMM is given by

$$\begin{aligned} \boldsymbol{r}^{(t+1)} &= \underset{\boldsymbol{r}\in\mathbb{R}^{nK}}{\arg\min} \ \sum_{k=1}^{K} \sum_{i=1}^{n} \rho_{\tau_k}(r_{ik}) + \frac{\rho}{2} ||\boldsymbol{Y}^* - \boldsymbol{r} - \boldsymbol{X}^* \boldsymbol{\beta}^{(t)} + \boldsymbol{u}^{(t)}/\rho||_2^2 \\ \boldsymbol{\beta}^{(t+1)} &= \underset{\boldsymbol{\beta}\in\mathbb{R}^{p+K}}{\arg\min} \ \frac{\rho}{2} ||\boldsymbol{Y}^* - \boldsymbol{r}^{(t+1)} - \boldsymbol{X}^* \boldsymbol{\beta} + \boldsymbol{u}^{(t)}/\rho||_2^2 + \lambda \sum_{j=1}^{p} {}^{|\beta_j|}/_{|\beta_j^{\text{CQR}}|^2} \\ \boldsymbol{u}^{(t+1)} &= \boldsymbol{u}^{(t)} + \rho(\boldsymbol{Y}^* - \boldsymbol{r}^{(t+1)} - \boldsymbol{X}^* \boldsymbol{\beta}^{(t+1)}), \end{aligned}$$

where $\boldsymbol{c} = \boldsymbol{Y}^* - \boldsymbol{X}^*\boldsymbol{\beta}^{(t)} + \boldsymbol{u}^{(t)}/\rho$, with residuals

$$\begin{aligned}
\boldsymbol{r}_{\text{primal}}^{(t+1)} &= \boldsymbol{Y}^* - \boldsymbol{X}^*\boldsymbol{\beta}^{(t+1)} - \boldsymbol{r}^{(t+1)} \\
\boldsymbol{r}_{\text{dual}}^{(t+1)} &= \rho \boldsymbol{X}_*^{*T}(\boldsymbol{r}^{(t+1)} - \boldsymbol{r}^{(t)}) \\
\varepsilon_{\text{primal}} &= \sqrt{n}\varepsilon_{\text{abs}} + \varepsilon_{\text{rel}} \max\{||\boldsymbol{X}_*^*\boldsymbol{\beta}_*^{(t+1)}||_2^2, ||\boldsymbol{r}^{(t+1)}||_2^2, ||\boldsymbol{b}^* - \boldsymbol{Y}||_2^2\}, \\
\varepsilon_{\text{dual}} &= \sqrt{p}\varepsilon_{\text{abs}} + \varepsilon_{\text{rel}}||\boldsymbol{X}^{*T}\boldsymbol{u}^{(t+1)}||_2^2.
\end{aligned}$$

The extension of the remaining two methods is similar, although requiring a slight change in the objective function. For the CD method, we modify our reformulation L_m of the objective function, for $m = 1, \ldots, p$, to include a second summation for the additional quantile levels as

$$L_m(b_1, \ldots, b_k, \boldsymbol{\beta}) = \sum_{k=1}^{K}\sum_{i=1}^{n}|x_{im}|\left|\frac{y_i - b_k - \sum_{j=1,\, j\neq m}^{p} x_{ij}\beta_j}{x_{im}} - \beta_m\right| \cdot \Theta_{ik} + p_\lambda(|\boldsymbol{\beta}|),$$

where $\Theta_{ik} = \rho_{\tau_k}(r_{ik})$ is analogous to Θ_i defined previously. The MM approach is similarly extended, yielding a final majorizer of the form

$$Q^\varepsilon(\boldsymbol{\beta}|\boldsymbol{\beta}^{(t)}) = \sum_{k=1}^{K}\sum_{i=1}^{n}\xi_{\tau_k}^{\varepsilon}(r_{ik}|r_{ik}^{(t)}) + \lambda\sum_{j=1}^{p}\frac{1}{|\beta_j^{\text{CQR}}|^2}\left[|\beta_j^{(t)}| + \frac{(\beta_j^2 - (\beta_j^{(t)})^2)\text{sgn}(\beta_j^{(t)})}{2|\beta_j^{(t)} + \varepsilon|}\right].$$

4 Numerical simulations

In this section, we evaluate the performance of the proposed ADMM, MM, and CD methods against that of the IP methods in `quantreg`. Because `quantreg` does not natively support regularized composite quantile regression, we do not make a comparison with IP approaches in that setting. Lasso regularization is used in place of adaptive lasso regularization for the IP method as the latter is not readily available in `quantreg`. Throughout this section, data is generated according to the model

$$y_i = b + \sum_{j=1}^{p} x_{ij}\beta_j + \varepsilon_i,$$

for $i = 1, \ldots, n$, where the ε_i are i.i.d. standard normal random variables. We use a convergence threshold of 10^{-4} to define our stopping criteria throughout.

We first focus on parameter estimation rather than variable selection and consider cases with $p = 5$ variables and $n = 200, 400, 600, 800, 1000, 2000$ observations in non-regularized quantile and composite quantile regression. In each simulation, the true value of each β_j is uniform randomly sampled from the interval $[-1, 1]$. In the

Table 1 Simulation results for quantile regression without regularization

(n,p)	IP		ADMM		MM		CD	
	Error	Time	Error	Time	Error	Time	Error	Time
(200,5)	0.08	0.002	0.063	0.002	0.060	**0.0002**	**0.036**	0.002
(400,5)	0.052	0.0022	0.055	0.0038	0.051	**0.0004**	**0.046**	0.003
(600,5)	0.043	0.0029	0.042	0.005	**0.033**	**0.0005**	0.043	0.0416
(800,5)	0.037	0.0048	0.035	0.006	**0.031**	**0.0005**	0.034	0.0046
(1000,5)	0.0336	0.0053	0.031	0.008	**0.026**	**0.0006**	0.031	0.0064
(2000,5)	0.0213	0.01	0.022	0.013	**0.018**	**0.001**	0.022	0.0096

Time measures the average computation time in seconds over 50 replications and Error measures the average absolute value difference between the estimated and true parameter values. The IP column displays the results from quantile regression using the IP method available in `quantreg`. The lowest Error and Time values for each (n,p) are noted in bold

Table 2 Simulation results for composite quantile regression without regularization

(n,p)	IP		ADMM		MM		CD	
	Error	Time	Error	Time	Error	Time	Error	Time
(200,5)	0.058	0.009	**0.057**	0.029	0.057	**0.0008**	0.058	0.008
(400,5)	0.043	0.021	0.043	0.057	0.047	**0.001**	**0.040**	0.011
(600,5)	0.035	0.03	**0.034**	0.088	0.034	**0.0012**	0.039	0.017
(800,5)	0.029	0.047	**0.029**	0.122	0.029	**0.0014**	0.031	0.018
(1000,5)	0.025	0.064	**0.024**	0.16	0.028	**0.0015**	0.024	0.025
(2000,5)	0.077	0.14	**0.017**	0.36	0.017	**0.0026**	0.018	0.044

Time measures the average computation time in seconds over 50 replications and Error measures the average absolute value difference between the estimated and true parameter values. The IP column displays results from composite quantile regression using the IP method available in `quantreg`. The lowest Error and Time values for each (n,p) are noted in bold

quantile regression case, we set $\tau = 0.3$ and in the composite quantile setting, we use quantile levels $0.1, 0.2, \ldots, 0.9$. Tables 1 and 2 present the performance of each method, averaged over 50 simulations.

We next consider variable selection for high-dimensional data using $n = 100, 200, 500$ and varying p from $1.5n$ to $5n$. The performance of each algorithm is summarized by the average number of false predictors selected, the average number of true predictors selected, and the average computation time in seconds over 25 replications. Simulation results in Table 3 are for regularized quantile regression with quantile level $\tau = 0.3$: here, the ADMM, MM, and CD methods use adaptive lasso regularization as described in previous sections, while the IP method uses the lasso regularization available in `quantreg`. Table 4 gives results based on composite quantile regression with adaptive lasso regularization using quantile levels $0.1, 0.2, \ldots, 0.9$: we do not make a comparison against an IP approach here, however, as a comparable method is not readily available in `quantreg`.

Table 3 Simulation results for regularized quantile regression: the ADMM, MM, and CD methods use adaptive lasso, while the IP method from `quantreg` uses lasso regularization

(n,p)	IP			ADMM			MM			CD		
	Time	N_T	N_F	Time	N_T	N_F	Time	N_T	N_F	Time	N_T	N_F
(100,200)	0.074	4	0	0.017	4	0	0.1	4	0.1	**0.014**	4	0
(100,300)	0.024	4	0	0.041	4	0	0.25	4	0	**0.02**	4	0
(100,500)	0.98	4	0	0.152	4	0	0.812	3.9	0	**0.035**	4	0
(200,400)	0.627	4	0	0.088	4	0	0.58	4	0	**0.048**	4	0
(200,600)	1.96	4	0	0.161	4	0	1.64	4	0	**0.054**	4	0
(200,1000)	8.85	4	0	0.791	4	0	6.23	4	0	**0.11**	4	0
(500,750)	5.1	4	0	0.522	4	0	4.09	4	0	**0.18**	4	0
(500,1000)	11	4	0	0.852	4	0	10.3	4	0	**0.24**	4	0
(500,1500)	38	4	0	2.41	4	0	24	4	0	**0.36**	4	0

Time measures the average computation time in seconds over 25 replications; N_T and N_F give the average number of true and false predictors selected, respectively. The lowest Time value for each (n,p) is noted in bold

Table 4 Simulation results for composite quantile regression with adaptive lasso regularization for the ADMM, MM, and CD algorithms

(n,p)	ADMM			MM			CD		
	Time	N_T	N_F	Time	N_T	N_F	Time	N_T	N_F
(100,200)	**0.043**	4	0	0.11	4	0.8	0.13	4	0
(100,300)	**0.089**	4	0	0.29	4	0.6	0.18	4	0
(100,500)	**0.21**	4	0	1.01	4	0.64	0.32	4	0
(200,400)	**0.22**	4	0	0.75	4	0.64	0.47	4	0
(200,600)	**0.452**	4	0	1.9	4	0.72	0.676	4	0
(200,1000)	1.41	4	0	7.4	4	0.25	**0.615**	4	0
(500,750)	**1.52**	4	0	5.4	4	0.8	2.4	4	0
(500,1000)	**2.43**	4	0	10.3	4	0.8	2.6	4	0
(500,1500)	5.86	4	0	28.5	4	0	**3.7**	4	0

An IP method from `quantreg` is not available in this setting. Time measures the average computation time in seconds over 25 replications; N_T and N_F give the average number of true and false predictors selected, respectively

5 Discussion and conclusions

In this paper we have presented three novel approaches to quantile and composite quantile regression and variable selection. Motivated by the lack of variety in algorithms for (composite) quantile regression, both with and without adaptive lasso regularization, and a desire to improve run times over the existing IP methods, we reformulated four types of quantile regression problems and presented estimators obtained using three algorithms. Using our existing implementation of these methods in the `cqrReg` pack-

age for R (Gao and Kong 2015), we used simulation studies to compare our methods to the IP algorithms available in the `quantreg` package (Koenker 2017).

In the non-regularized quantile regression setting, we do not observe substantial differences in the average estimation error between methods; the same is true of run time except for the MM approach, which performs considerably better than the other three methods in this setting. In non-regularized composite quantile regression, however, differences between the methods in terms of estimation error are more apparent, as the IP method has larger average estimation error than the ADMM, MM, and CD approaches, while MM and CD are faster and ADMM slower than the IP algorithm. Comparisons between IP and ADMM methods for non-regularized quantile regression already exist in the literature (Koenker et al. 2018, Chapter 5). The results so far suggest that the MM approach is the best suited for non-regularized (composite) quantile regression among the four methods tested, especially for data sets with p small relative to n. In regularized quantile regression, all of our approaches perform similarly in terms of variable selection, but CD and ADMM show clear superiority in run time, particularly relative to the IP and MM methods when p is large. In the case of regularized composite quantile regression, CD and ADMM have run time superior to MM. Furthermore, MM shows a tendency to select irrelevant variables, likely due to the algorithm's matrix inversion and selection of an approximating parameter. This second set of results suggests that our CD approach is best suited for regularized (composite) quantile regression among the three methods tested, although care should be taken with regards to its theoretical convergence properties, as noted by Tseng (2001). In particular, since the penalty is not continuously differentiable in $\boldsymbol{\beta}$ [so that the penalty is not separable as per Tseng (2001)], convergence results do not apply. This situation is similar to that noted by Friedman et al. (2007) in the context of fused lasso. In an example, CD is unable to achieve the global minimum of a strictly convex objective function. The authors show this problem stems from CD not allowing two particular components to be updated together, while no improvement to the value of the objective function is possible in one-component subproblem updates. With some specific modifications, however, Friedman et al. (2007) show that this CD approach can be modified for highly competitive performance for the fused lasso problem.

Overall, our methods provide reliable and efficient algorithms to estimate solutions to quantile and composite quantile regression problems, including those regularized by an adaptive lasso penalty. Our methods, already implemented in the `cqrReg` package for R, widen the variety of algorithms available for quantile and composite quantile regression and greatly improve upon the run time of the existing advanced IP methods, particularly for large or high-dimensional data sets. Our ADMM method was competitive and is further amenable to parallelization, naturally lending itself to distributed computing to handle data that is both high-dimensional and extremely large in volume. ADMM may have future application in training deep neural networks through gains in estimation error and computation time. This is explored in greater depth by Yu and Lin (2017) and Gu et al. (2018) for big data and in sparse, high-dimensional settings.

Acknowledgements Jueyu Gao acknowledges the supervision of Drs. Linglong Kong and Edit Gombay during his graduate studies. The authors have no declarations of interest to declare.

References

Boyd S, Parikh N, Chu E, Eckstein J (2011) Distributed optimization and statistical learning via the alternating direction method of multipliers. Found Trends Mach Learn 3(1):1–122

Chen C, Wei Y (2005) Computational issues for quantile regression. Sankhyā. Indian J Stat 67(2):399–417

Dempster A, Laird N, Rubin D (1976) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc: Ser B (Methodol) 39(1):1–38

Eddelbuettel D, François R (2011) Rcpp: seamless R and C++ integration. J Stat Softw 40(8):1–18

Eddelbuettel D, Sanderson C (2014) RcppArmadillo: accelerating R with high-performance C++ linear algebra. Comput Stat Data Anal 71:1054–1063

Fan J, Li R (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. J Am Stat Assoc 96(1):1348–1360

Friedman J, Hastie T, Höfling H, Tibshirani R (2007) Pathwise coordinate optimization. Ann Appl Stat 1(2):302–332

Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. J Stat Softw 31(1):1–22

Gabay D, Mercier B (1976) A dual algorithm for the solution of nonlinear variational problems via finite element approximation. Comp Math Appl 2(1):17–40

Gao J, Kong L (2015) cqrReg: Quantile, composite quantile regression and regularized versions. https://CRAN.R-project.org/package=cqrReg, R package version 1.2. Accessed 2017

Gu Y, Fan J, Kong L, Ma S, Zou H (2018) ADMM for high-dimensional sparse penalized quantile regression. Technometrics 60(3):319–331

He Q, Kong L, Wang Y, Wang S, Chan T, Holland E (2016) Regularized quantile regression under heterogeneous sparsity with application to quantitative genetic traits. Comput Stat Data Anal 95:222–239

Hestenes M (1969) Multiplier and gradient methods. J Optim Theory Appl 4(5):303–320

Hunter D, Lange K (2000) Quantile regression via an MM algorithm. J Comput Gr Stat 9(1):60–77

Hunter D, Lange K (2004) A tutorial on MM algorithms. Am Stat 58(1):30–37

Hunter D, Li R (2005) Variable selection using MM algorithms. Ann Stat 33(4):1617–1642

Kai B, Li R, Zou H (2010) Local composite quantile regression smoothing: an efficient and safe alternative to local polynomial regression. J R Stat Soc: Ser B (Stat Methodol) 72(1):49–69

Koenker R (2005) Quantile regression. Cambridge University Press, Cambridge

Koenker R (2017) quantreg: Quantile regression. https://CRAN.R-project.org/package=quantreg, R package version 5.33. Accessed 2017

Koenker R, Bassett G (1978) Regression quantiles. Econometrica 46(1):33–50

Koenker R, Chernozhukov V, He X, Peng L (2018) Handbook of quantile regression. CRC Press, Boca Raton

Kong L, Shu H, Heo G, He QC (2015) Estimation for bivariate quantile varying coefficient model. arXiv:1511.02552

Li D, Li R (2016) Local composite quantile regression smoothing for Harris recurrent Markov processes. J Econom 194(1):44–56

Lin Z, Chen M, Ma Y (2010) The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices. arXiv:1109.0367

Luo ZQ, Tseng P (1992) On the convergence of the coordinate descent method for convex differentiable minimization. J Optim Theory Appl 72(1):7–35

Mehrotra S (1992) On the implementation of a primal-dual interior point method. SIAM J Optim 2(4):575–601

Ortega J, Rheinboldt W (1970) Iterative solution of nonlinear equations in several variables. Academic Press, New York and London

Tibshirani R (1996) Regression shrinkage and selection via the lasso. J R Stat Soc: Ser B (Methodol) 58(1):267–288

Tseng P (2001) Convergence of a block coordinate descent method for nondifferentiable minimization. J Optim Theory Appl 109(3):475–494

Vidaurre D, Bielza C, Larrañaga P (2013) A survey of L_1 regression. Int Stat Rev 81(3):361–387

Wu T, Lange K (2008) Coordinate descent algorithms for lasso penalized regression. Ann Appl Stat 2(1):224–244

Wu Y, Liu Y (2009) Variable selection in quantile regression. Stat Sin 19(2):801–817

Xu Q, Deng K, Jiang C, Sun F, Huang X (2017) Composite quantile regression neural network with applications. Expert Syst Appl 76:129–139
Yu L, Lin N (2017) ADMM for penalized quantile regression in big data. Int Stat Rev 85(3):494–518
Zhang L, Yu D, Mizera I, Jiang B, Kong L (2017) Sparse wavelet estimation in quantile regression with multiple functional predictors. arXiv:1706.02353
Zou H (2006) The adaptive lasso and its oracle properties. J Am Stat Assoc 101:1418–1429
Zou H, Yuan M (2008) Composite quantile regression and the oracle model selection theory. Ann Stat 36(3):1108–1126

Affiliations

Matthew Pietrosanu[1] · Jueyu Gao[1] · Linglong Kong[1] · Bei Jiang[1] · Di Niu[2]

✉ Linglong Kong
lkong@ualberta.ca

Matthew Pietrosanu
pietrosa@ualberta.ca

Jueyu Gao
jueyu@ualberta.ca

Bei Jiang
bei1@ualberta.ca

Di Niu
dniu@ualberta.ca

[1] Department of Mathematical & Statistical Sciences, University of Alberta, Edmonton, AB, Canada

[2] Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB, Canada