An efficient algorithm for structured sparse quantile regression

Vahid Nassiri[a] and Ignace Loris[b]
[a]Department of Mathematics, Vrije Universiteit Brussel, Belgium,
[b]Department of Mathematics, Université Libre de Bruxelles, Belgium

June 24, 2018

Abstract

Quantile regression is studied in combination with a penalty which promotes structured (or group) sparsity. A mixed $\ell_{1,\infty}$-norm on the parameter vector is used to impose structured sparsity on the traditional quantile regression problem. An algorithm is derived to calculate the piece-wise linear solution path of the corresponding minimization problem. A Matlab implementation of the proposed algorithm is provided and some applications of the methods are also studied.

Keywords: quantile regression; structured sparsity; variable selection; convex optimization

1 Introduction

As [1] have remarked, a good statistical model has seven key properties: 1. Parsimony, 2. Tractability, 3. Conceptual insightfulness, 4. Generalizability, 5. Falsifiability, 6. Empirical consistency, and 7. Predictive precision. In this paper a structured sparse quantile regression model is studied and an efficient algorithm is proposed to solve the corresponding minimization problem. As an illustration, two applications are discussed where such a model is preferable (in the sense of several of the above properties) to others, and where the proposed algorithm is more useful than others.

The quantile regression model of [2] allows for studying the effect of explanatory variables on the entire conditional distribution of the response variable, and not only on its center. In this sense quantile regression provides a deeper *conceptual insightfulness* into data than least squares models. Since *parsimony* is a key property of a good model, variable selection techniques have attracted a lot of attention in recent statistical literature, [3]. One reason that sparse models have become popular is the availability of very large data sets; see e.g. [4]. A promising approach to achieve a sparse model is to penalize models with a sparsity promoting penalty. Among different penalties, the ℓ_1-norm is the most popular one. It has e.g. been used in combination with least squares in Lasso regression of [5], and an effective algorithm was proposed by [6] and [7]. Authors such as [8] and [9] have proposed similar algorithms for the ℓ_1-norm penalized quantile regression problem.

Although the ℓ_1-norm penalty does a good job in selecting individual variables, there is no control on the groups of variables it selects. There are cases where one is interested in selecting a group of variables instead of individual variables. In some cases it is inevitable, see [10]. Consider e.g. race as a predictor with three levels: black, white, and other. The standard approach is making two dummy variables (X_1, X_2) out of it: $(0,0)$ for other, $(1,0)$ for black, and $(0,1)$ for white. Obviously in such a case one factor is represented by two variables.

Therefore, one may select the pair (X_1, X_2) together, or select none of them. The usual ℓ_1-norm penalty cannot guarantee such behavior. Finding an alternative to the ℓ_1-norm seems therefore to be necessary in the cases where groups of variables should be selected instead of individual variables. As an individual can be considered as a singleton group, the grouping approach should be a *generalization* of the usual ℓ_1-norm penalty.

An appropriate penalty for the grouping case would be a mixed norm which applies ℓ_1-norm penalty to groups, in order to promote sparsity of groups of variables. Within each group the ℓ_∞-norm (also known as the max-norm) can be used to ensure that, once a single member of a group is chosen (is nonzero), all other members of the same group can grow to the same size without penalty. Other choices are possible too, but we restrict ourselves to this mixed $\ell_1 - \ell_\infty$-norm (or $\ell_{1,\infty}$-norm for short) choice. [11] have introduced such a group sparsity penalty, and have solved the problem for the least squares model and have applied it to a low birth weight data set.

In this paper we formulate an algorithm for $\ell_{1,\infty}$-norm penalized quantile regression instead of least squares regression. As *tractability* is a key property of a good model, we choose the mixed norm penalty function because the corresponding models have a simple dependence on the penalty parameter. Choosing the ℓ_∞-norm within groups guarantees that the model is a piece-wise linear function both of the penalty parameter and of the $\ell_{1,\infty}$-norm of the model. Other choices (such as e.g. $\ell_{1,2}$) would lead to less tractable optimization problems. The main contribution of the present paper is the description of an efficient computational algorithm for solving this optimization problem. This algorithm applies only to the case of non-overlapping groups. A Matlab implementation of the algorithm is provided on the authors' webpage [12].

In order to illustrate the performance of the proposed penalized model and the corresponding optimization algorithm, we study two main applications: Firstly the use of qualitative explanatory variables with more than two levels, and secondly, simultaneous variable selection for a vector of response variables.

For the first application, we analyze a data set pertaining to 'low birth weight' (LBW). Low birth weight is a common subject in quantile regression literature (see e.g. [13]). If one is interested in the effect of different variables on the lower tail of the conditional distribution of the infant's weight, least squares models are not appropriate, because they analyze the effect of the explanatory variables on the conditional mean of the response variable. On the other hand, using quantile regression, one can study lower conditional quantiles of the response. Due to presence of a qualitative variable with more than two levels in the model, the sparsity promoting ℓ_1-norm penalty is not appropriate either. Therefore, a penalized model with a mixed norm penalty seems to be more appropriate for selecting effective variables and estimating their effect.

For the second application, the '93CARS' data set of [14] is studied. It consists of 14 explanatory variables and 5 response variables which are highly correlated. [15] have suggested that a simultaneous variable selection approach is more interesting for modeling a vector of correlated response variables with a common set of explanatory variables. Using the proposed mixed norm as penalty one can put the coefficients of each explanatory variable for all the response variables in a group and perform a simultaneous variable selection. While [15] worked it out in the least-squares framework, we use quantile regression which is more robust and makes it possible to study the whole conditional distribution and not just its center. [16] have studied the latter problem too, using a standard linear programme solver. In contrast, our proposed algorithm produces the whole piece-wise linear solution path (for different values of the penalty parameter) which makes the variable selection computationally more efficient.

Section 2 states the structured sparse quantile regression problem and introduces notation. The main tools for describing the corresponding minimization problem are discussed in Sec-

tion 3. The algorithm to solve $\ell_{1,\infty}$-norm penalized quantile regression problem is introduced and discussed in Section 4. Section 5 illustrates the use of structured quantile regression for the analysis of a low birth weight data set. A simultaneous variable selection problem for a vector of response variables data set is also examined. The paper is concluded in Section 6.

2 Problem statement

In this paper, $X_{n \times m} = (X_1, \ldots, X_m)$ is the design matrix containing the explanatory variables. The linear model which we study is:

$$y = X\beta + \epsilon \tag{1}$$

The loss function which quantile regression (see [2]) tries to minimize is given by:

$$\sum_{i=1}^{n} \varrho_\tau \left((y - X\beta)_i \right) \tag{2}$$

where the function ϱ_τ is defined as:

$$\varrho_\tau(t) = \begin{cases} 2\tau\, t & \text{if } t \geq 0 \\ -2(1-\tau)\, t & \text{if } t \leq 0, \end{cases} \tag{3}$$

for $t \in \mathbb{R}$ and $0 < \tau < 1$. For $\tau = 1/2$, one recovers $\varrho_\tau(t) = |t|$. This loss function is more robust to outliers than the usual quadratic one: the minimizer of the least squares loss is the mean, while the minimizer of the least absolute deviations loss (expression (2) with $\tau = 1/2$) is the median. The optimization problem encountered in quantile regression thus is:

$$\hat{\beta}_{\text{QR}} = \arg\min_\beta \sum_{i=1}^{n} \varrho_\tau \left((y - X\beta)_i \right). \tag{4}$$

The so-called ℓ_1-norm $\|\beta\|_1 = \sum_{j=1}^{m} |\beta_j|$ is well known to promote sparsity [5] when used as a penalty (added to a loss function) or as a constraint (imposed on the minimizer of a loss function). Two major examples are the Lasso of [5] (which combines a quadratic loss with an ℓ_1-norm penalty or constraint) and sparse quantile regression of e.g. [8] (which combines the loss (2) with an ℓ_1-norm penalty. See also [9]). The advantage of using the ℓ_1-norm instead of the number of nonzero β_j is computational: the former leads to a convex minimization problem while the latter gives rise to a combinatorial minimization problem.

In this paper our goal is to impose structure between the m explanatory variables $\beta_1, \ldots \beta_m$ by dividing them into g non-overlapping groups ($g \leq m$). $G_1, \ldots, G_g$ represent the indices in each group. At the same time, we desire to select a small number of active groups. We are therefore interested in imposing a penalty or constraint of ℓ_1-norm type on the different groups. Within a given group, when a single member is active (nonzero), we allow the other members of that group to reach the same magnitude without penalty. In order to impose such a behavior within groups, the max-norm is appropriate. Indeed, the value of $\max_{j \in G_k}\{|\beta_j|\}$ does not increase as long as the largest of the $|\beta_j|$ (with $j \in G_k$) does not increase, regardless of the actual size of the smaller ones.

The mixed norm $\|\beta\|_{1,\infty}$ is defined as

$$\|\beta\|_{1,\infty} = \sum_{k=1}^{g} \|\beta_{G_k}\|_\infty, \tag{5}$$

where β_{G_k} are the components of β in group k, and $\|\beta_{G_k}\|_\infty = \max_{j \in G_k} |\beta_j|$. This mixed norm behaves as described above: Within a group, it behaves as a max-norm and between groups it imposes an ℓ_1-norm (sum of max-norms). The structured sparse $\hat{\beta}$ for the quantile regression loss function in (3) is now defined as the minimizer:

$$\hat{\beta} = \arg\min_\beta \sum_{i=1}^n \varrho_\tau \left((y - X\beta)_i\right) + \lambda \|\beta\|_{1,\infty} \tag{6}$$

where $\lambda \geq 0$ is the penalization parameter or as the minimizer:

$$\hat{\beta} = \arg\min_{\|\beta\|_{1,\infty} \leq R} \sum_{i=1}^n \varrho_\tau \left((y - X\beta)_i\right), \tag{7}$$

where R is a nonnegative parameter. By suitably choosing the parameters λ and R, the minimizers of these problems are identical. We therefore use the same symbol $\hat{\beta}$, both for the minimizer of the penalized problem (6) and for the minimizer of constrained problem (7). We do not indicate explicitly their dependence on λ or R.

The mixed norm penalty $\|\beta\|_{1,\infty}$ has been studied before in the framework of least squares loss functions by [11]. A special case of that least squares problem was already introduced by [15]. Here, as in [16], we combine its structure-imposing properties with the robustness properties of quantile regression. The aim of this paper is to present an efficient algorithm for the solution of the minimization problems (6) and (7), for various values of λ and R, and to illustrate its use with some applications.

3 Solution of the minimization problem

Due to the presence of the non-smooth functions ϱ_τ and $\|\beta\|_{1,\infty}$, the optimization problems (6) and (7) are not differentiable. They are however convex minimization problems for which a general theory exists. We refer to e.g. [17] for an introduction to convex analysis. In particular, for any convex function f, the symbol ∂f denotes the subdifferential of f.

We express the condition for minimizing the cost function of expression (6) using subdifferentials instead of usual derivatives. Necessary and sufficient conditions for optimality of the optimization problem (6) are found by expressing that 0 belongs to the subdifferential of the functional (6). Therefore, we will have that β is the minimizer of (6) if and only if there exists a vector $w \in \partial \sum_{i=1}^n \varrho_\tau \left((y - X\beta)_i\right)$ and a vector $u \in \partial \|\beta\|_{1,\infty}$ such that:

$$-X^T w + \lambda u = 0. \tag{8}$$

Setting $r = y - X\beta$, the equations that we need to solve are:

$$\begin{cases} -X^T w + \lambda u = 0 \\ w_i \in \partial \varrho_\tau (r_i) \\ u \in \partial \|\beta\|_{1,\infty} \\ r = y - X\beta. \end{cases} \tag{9}$$

Solving the minimization problems (6) and (7) therefore requires detailed knowledge of the subdifferentials of ϱ_τ and of $\|\beta\|_{1,\infty}$.

In case of the function ϱ_τ the subdifferential is equal to:

$$\partial\varrho_\tau(r_i) = \begin{cases} 2\tau & \text{if} \quad r_i > 0, \\ [-2(1-\tau), 2\tau] & \text{if} \quad r_i = 0, \\ -2(1-\tau) & \text{if} \quad r_i < 0. \end{cases} \tag{10}$$

It is important to remark that, conversely, the knowledge of the value of the subgradient $w_i \in \partial\varrho_\tau(r_i)$ also gives a certain knowledge on r_i. In particular, if $w_i = 2\tau$, then r_i must be non-negative, if $w_i = -2(1-\tau)$, then r_i must be non-positive and if w_i belongs to the interval $]-2(1-\tau), 2\tau[$, then r_i must be zero. This will be important when describing the algorithm that solves the equations (9).

The subdifferential $\partial\|\beta\|_{1,\infty}$ for the penalty part $\|\beta\|_{1,\infty}$ is more difficult. As the groups are disjoint, each term in the sum (5) can be handled separately. Consider group k with $\|\beta_{G_k}\|_\infty = \max_{j \in G_k} |\beta_j|$, and $G_k = \{j_1, j_2, \ldots, j_{l_k}\}$. The subdifferential of the max function is the convex hull of the union of the subdifferentials of the 'maximal' arguments (see e.g. [18]). In our case, each argument of the max-function is a function of just a single variable β_j. The subdifferential (w.r.t. the single variable β_j) of the absolute value $|\beta_j|$ is:

$$\partial|\beta_j| = \begin{cases} 1 & \text{if} \quad \beta_j > 0, \\ [-1, 1] & \text{if} \quad \beta_j = 0, \\ -1 & \text{if} \quad \beta_j < 0, \end{cases} \tag{11}$$

and the subdifferential of $|\beta_{\tilde{j}}|$ w.r.t. β_j is zero ($j \neq \tilde{j}$).

In this way, when all components of β in group k are zero, we find that $\partial\|\beta_{G_k}\|_\infty$ consists of an ℓ_1-ball of radius 1. On the other hand, when all l_k coefficients in group G_k have the same (nonzero) absolute size, $\partial\|\beta_{G_k}\|_\infty$ consists of a single face of an ℓ_1-ball with radius 1 (the signs of the β_j determine which face). If the maximum is reached in $1, 2, \ldots l_k - 1$ nonzero β_j then the set $\partial\|\beta_{G_k}\|_\infty$ (with elements u) consists of that part of the same face with $u_j = 0$ for non-maximal components β_j (i.e. components that are smaller than the maximum of that group: $|\beta_j| < \|\beta_{G_k}\|_\infty$).

Here too, knowledge of $u_{G_k} \in \partial\|\beta_{G_k}\|_\infty$ yields partial information on β_{G_k}. E.g. if $\|u_{G_k}\|_1 < 1$ then all $|\beta_j|$ in this group are zero; if $\|u_{G_k}\| = 1$ and all $u_j \neq 0$ then all β_j in that group have the same absolute value. If $u_j = 0$ then β_j may not be maximal in its group.

This type of interplay between β, u, w and r will be used frequently in the algorithm which we describe in Section 4. The optimality conditions (9) are not analytically solvable for β. However, the problems (6) and (7) fall into the class of problems described by [19] which allow for a piece-wise linear solution path. This means that the minimizers of (6) and (7) are piecewise linear functions in terms of λ or R. Moreover the nodes that determine these piecewise linear functions can be calculated analytically (i.e. using linear algebra), provided one starts the calculation from $R = 0$ (which corresponds to $\beta = 0$ and λ large) and proceeds carefully for increasing values of R (decreasing values of λ). As explained in the next section, a new node of the minimizer $\hat\beta$ (as a function of R) appears e.g. when a new group becomes active (becomes nonzero) or when an equation among the $(y - X\beta)_i = 0$ is satisfied. There are several more events like this possible and they are explained in more detail in the next section.

4   Algorithm

The non-iterative algorithm we propose for finding the minimizers of (6) and (7) is described in this section. It is similar to the LARS algorithm of [7] for the Lasso and the algorithm of [8] and

[9] for sparse quantile regression (where the number of groups equals the number of explanatory variables). In our case, we are not restricted to singleton groups; however, similar to [11], only non-overlapping groups are allowed.

Following [9], we set $u = u^{(0)} + u^{(1)}/\lambda$ and $w = w^{(0)} + \lambda w^{(1)}$, where $u^{(0)}, u^{(1)}, w^{(0)}$ and $w^{(1)}$ do not depend on λ. In this way equation (8) is equivalent to:

$$-X^T w^{(0)} + u^{(1)} = 0 \qquad \text{and} \qquad -X^T w^{(1)} + u^{(0)} = 0. \tag{12}$$

As the minimizer $\hat{\beta}$ of the problem (7) is a piecewise linear function of $R = \|\hat{\beta}\|_{1,\infty}$, it is characterized by a set of (interpolation) nodes. If the value of $\hat{\beta}$ is known in these nodes, then $\hat{\beta}$ can also be found for other values of R by linear interpolation.

The proposed algorithm starts from $\hat{\beta} = 0$ for λ sufficiently large ($\lambda \geq \lambda_{\max}$). As $\hat{\beta} = 0$, one finds the value of $r = y - X\hat{\beta} = y$ and of w: $w_i^{(0)} = 2\tau$ if $r_i > 0$, $w_i^{(0)} = -2(1 - \tau)$ if $r_i < 0$ (assuming that all coefficients of r at this stage are nonzero) and $w^{(1)} = 0$. Using equations (12), one then calculates $u = u^{(0)} + u^{(1)}/\lambda$. The value of $\lambda_{\max}$ is now found by solving the equations $\|u_{G_k}\|_1 = 1$ (for all groups $k : 1 \ldots g$). This can be done on a computer as $\|u^{(0)} + u^{(1)}/\lambda\|_{1,\infty}$ is a piecewise linear function of λ^{-1}. The largest positive value among those numbers is the desired value of $\lambda_{\max}$. The group k for which $\|u_{G_k}\| = 1$ will be the first group to enter the set of active groups (in the next step) and become non-zero. Once this break point value of λ is known, it can be used to update u and w.

The algorithm then continues a loop with consists of several parts:

1. Express $\hat{\beta}$ as a linear function of R: $\hat{\beta} = \hat{\beta}^{(0)} + R\hat{\beta}^{(1)}$, where $R = \|\hat{\beta}\|_{1,\infty}$. This uses the knowledge of the active groups, of the maximal set within each group (from the knowledge of the subgradient u), of the relative signs of these components of β (also from the knowledge of u), and of the components of $y - X\hat{\beta}$ that are zero (from the knowledge of the subgradient w).

2. Determine the largest value of R for which the expression $\hat{\beta} = \hat{\beta}^{(0)} + R\hat{\beta}^{(1)}$ is valid. The expression may cease to be valid when:

   (a) an additional component of $r = y - X\hat{\beta}$ becomes zero,
   (b) an active group becomes non-active (all members are zero),
   (c) a non-maximal component of $\hat{\beta}$ (in some group) becomes equal in absolute value to the maximal value in that group.

3. Once the value of R is calculated, update the variables u and w (as a function of λ). Here equations (12) are used together with the knowledge of w_i for the nonzero r_i. The knowledge of the non-maximal β_j (in each group) is also used to set some u_j's to zero.

4. Calculate the smallest value of λ for which these expressions for u and w are valid. The expressions for u and w cease to be valid when:

   (a) One of the $\|u_{G_k}\|_1$ will reach 1 (a new active group will be added to the active set in the next step),
   (b) A coefficient of w equals 2τ or $-2(1 - \tau)$ (in this case, an equation $r_j = 0$ that is satisfied in the current step, will no longer be satisfied in the next step),
   (c) in an active group k, one of the coefficients of u_{G_k} becomes equal to 0. In this case, the corresponding component of $\hat{\beta}$ will be of smaller absolute value than the maximal value in that group in the next step.

6

5. Continue with step 1 or stop.

The algorithm may be stopped when the desired maximum number of active groups is reached, when $\lambda = 0$, or when some other suitable stopping criteria is satisfied.

In this algorithm, steps 1 and 3 require the solution of a linear system of equations. Steps 2 and 4 require the solution of simple linear equations to determine R or λ at break points. Here numerical round-off error may affect the accuracy of these calculations. Unfortunately, the decisions (groups entering or leaving the active set, coefficients becoming submaximal in a group, ...) depend on these numerical results. Round-off errors may therefore lead to the wrong decisions being taken by the algorithm. In that case the algorithm fails. This shortcoming is common to all the algorithms of this type [7, 8, 9].

When dealing with data X and y containing small integers, or when rows or columns of y and X repeat, it is possible that different events (2a–c or 4a–c) occur simultaneously. One could e.g. have two groups enter the active set at the same step. Another possibility is that a new group becomes active at the same step when a component of an active group becoming sub-maximal. Such possibilities are not accounted for in the current implementation of the algorithm. This "one-at-a-time condition" [7, p417] is also common to algorithms of this type (the work of [7, 8, 9] also does not handle such cases). [7, p438] have proposed to add a small amount of jitter to the variables to overcome the problem.

The most effective way of understanding the proposed algorithm is by going through a worked-out example step-by-step. Table 1 lists the complete solution path of a simple example with:

$$X = \begin{pmatrix} -4 & 3 & 5 \\ -4 & 5 & 1 \\ 4 & -3 & 0 \end{pmatrix}, \qquad y = \begin{pmatrix} 8 \\ 7 \\ -11 \end{pmatrix}, \tag{13}$$

groups $G_1 = \{1\}$ and $G_2 = \{2, 3\}$ and $\tau = 1/2$. The solution path is given as a function of R and λ, and the intermediate values of the subgradients u and w are also given. This example was chosen in such a way that every possibility in steps 2 and 4 of the algorithm occurs at least once.

As one can see in the example, there are certain values of λ (i.e. $\lambda = 17, 37/5, 20/3, \ldots$) for which the minimizer of the penalized problem (6) is not unique. One also sees that between these special values of λ the minimizer $\hat{\beta}$ of (6) is constant as a function of λ (the subgradients u and w do change). This behavior is easy to interpret by plotting the loss $\sum_{i=1}^{n} \varrho_\tau((y - X\hat{\beta})_i)$ as a function of $R = \|\hat{\beta}\|_{1,\infty}$, as was done for example (13) in Figure 1. We see that the graph is piecewise linear and that λ is locally equal to the slope of this trade-off curve. Therefore, between break points, λ is constant (several $\hat{\beta}$ correspond to the same value of λ) and at break points, λ takes on several values (in other words, for several values of λ, the solution $\hat{\beta}$ of (6) is constant).

A set of Matlab functions that implement the above algorithm was written by the authors, and is available on their web page [12].

5 Applications

The algorithm presented in Section 4 gives the entire solution path for different values of $R = \|\hat{\beta}\|_{1,\infty}$ in some interval $[0, R_{\max}]$. Selecting the appropriate value of R (and the corresponding coefficients $\hat{\beta}$) is an important issue for practical purposes. As [20] have proposed, the Bayesian information criterion (BIC) of [21] is a promising information criterion for model selection in

#	$\hat{\beta}$	r	R	λ	u	w	Step
0	$(0,0,0)$	$(8,7,-11)$	0	$[17,+\infty[$	$\lambda^{-1}(-12,11,6)$	$(1,1,-1)$	
	$(0,R,R)$	$\vdots$	$[0,1]$	17	$(\frac{-12}{17},\frac{11}{17},\frac{6}{17})$	$(1,1,-1)$	1
							$\downarrow 2a$
1	$(0,1,1)$	$(0,1,-8)$	1	$[\frac{37}{5},17]$	$\vdots$	$\vdots$	3
							$\downarrow 4c$
	$\vdots$	$\vdots$	$[1,\frac{27}{22}]$	$\frac{37}{5}$	$(\frac{-36}{37},1,0)$	$(\frac{-1}{5},1,-1)$	1
							$\downarrow 2a$
2	$(0,\frac{27}{22},\frac{19}{22})$	$(0,0,\frac{-161}{22})$	$\frac{27}{22}$	$[\frac{20}{3},\frac{37}{5}]$	$\vdots$	$\vdots$	3
							$\downarrow 4a$
	$\vdots$	$\vdots$	$[\frac{27}{22},\frac{3}{2}]$	$\frac{20}{3}$	$(-1,1,0)$	$(\frac{-1}{6},\frac{5}{6},-1)$	1
							$\downarrow 2c$
3	$(-1,\frac{1}{2},\frac{1}{2})$	$(0,0,\frac{-11}{2})$	$\frac{3}{2}$	$[3,\frac{20}{3}]$	$\vdots$	$\vdots$	3
							$\downarrow 4b$
	$\vdots$	$\vdots$	$[\frac{3}{2},2]$	3	$(-1,\frac{1}{12},\frac{11}{12})$	$(\frac{3}{4},-1,-1)$	1
							$\downarrow 2b$
4	$(-2,0,0)$	$(0,-1,-3)$	2	$[2,3]$	$\vdots$	$\vdots$	3
							$\downarrow 4a$
	$\vdots$	$\vdots$	$[2,\frac{25}{12}]$	2	$(-1,\frac{-1}{4},\frac{3}{4})$	$(\frac{1}{2},-1,-1)$	1
							$\downarrow 2a$
5	$(\frac{-23}{12},\frac{-1}{6},\frac{1}{6})$	$(0,0,\frac{-23}{6})$	$\frac{25}{12}$	$[\frac{20}{19},2]$	$\vdots$	$\vdots$	3
							$\downarrow 4c$
	$\vdots$	$\vdots$	$[\frac{25}{12},\frac{229}{40}]$	$\frac{20}{19}$	$(-1,-1,0)$	$(\frac{7}{38},\frac{-35}{38},-1)$	1
							$\downarrow 2a$
6	$(\frac{-161}{40},\frac{-17}{10},\frac{-3}{5})$	$(0,0,0)$	$\frac{229}{40}$	$[0,\frac{20}{19}]$	$(-1,-1,0)$	$\lambda(\frac{7}{40},\frac{-7}{8},\frac{-19}{20})$	3

Table 1: The various values of $\hat{\beta}, r, R, \lambda, u, w$ etc. that describe the minimizers (6) and (7) for the simple example (13). The last column indicates the step that is taken in the algorithm of Section 4. Values of $\hat{\beta}$, r, u or w that can be found through interpolation (w.r.t R for β and r, w.r.t λ for w and w.r.t. λ^{-1} for u) have been replaced by vertical dots for lack of space.

Figure 1: Minimizers corresponding to the example X and y of equation (13). Left: parameters $\hat{\beta}$ as a function of $R = \|\hat{\beta}\|_{1,\infty}$. Right: the corresponding trade-off curve plotting the loss $\sum_{i=1}^{n} \varrho_\tau((y - X\hat{\beta})_i)$ as a function of the penalty $R = \|\hat{\beta}\|_{1,\infty}$. The slope of this curve is equal to $-\lambda$.

8

quantile regression. In view of Theorem 2 of [8], and the loss function in (2), the adapted BIC for quantile regression is as follows:

$$\text{BIC}(R) = \log\left(\frac{1}{n}\sum_{i=1}^{n} \varrho_\tau\left((y - X\hat{\beta})_i\right)\right) - \frac{\log(n)}{2n}\ n_R, \tag{14}$$

where $\hat{\beta}$ is a function of R and n_R is defined as the number of zeros in the residual vector $r = y - X\hat{\beta}$. The model with smaller BIC is more desirable (see e.g. [3]).

In this section two main applications of structured sparse quantile regression are studied using real data sets as illustrations.

5.1 Low birth weight data set

According to [22], *low birth weight* (LBW) is defined as a birth weight of a liveborn infant of less than 2500g regardless of gestational age. LBW has negative effect both on the infants and the parents, e.g mothers of LBW babies have a greater chance of having postpartum depression and they need more time before returning to work, infants who are born with LBW are at greater risk of having learning or vision difficulties. Also, LBW infants would impose large costs on society. The risks of LBW are discussed by many authors such as, [23], [24] and [25]. Therefore, determining the effective factors in LBW infants is very important. If one is interested in studying the effects of different factors on the lower tail of the conditional distribution of infants' weight, [13] have remarked that using the least squares regression methods (e.g. [11]) is not reasonable. Using quantile regression with τ-th quantile ($\tau \le 0.5$) would give the possibility to study lower tail of the infants weight given the explanatory variables.

As in [11], the data are taken from [26]. The data set contains the birth weight (expressed in grams) of 189 infants as the response variable and 8 explanatory variables: mother's age (in years), mother's weight (in pounds), mother's race $((1,0) = \text{black}, (0,1) = \text{white or } (0,0) = \text{other})$, smoking status during pregnancy (1 =yes or 0 =no), number of previous premature labours $(0,1,2,\ldots)$, history of hypertension ($1 = $ yes or $0 = $ no), presence of uterine irritability ($1 = $ yes or $0 = $ no), number of physician visits during the first trimester $(0,1,2,\ldots)$. The data were collected at Baystate Medical Center, Springfield, Massachusetts, in the year 1986.

As [13] and [11] have suggested, some non-linear effects of two of the quantitative predictors (mother's weight and age) may exist. Therefore, in accordance with [13] we consider a second-order polynomial for both of them. We put each corresponding pair of variables in one group. A variable which needs a non-singleton group is race (it is a nominal variable with more than two levels). Thus in order to study its effect one may create two dummy variables out of it (see also Section 1). As was already mentioned in the introduction, both or none of them should be included in the model. So we may put them both in one group. All other groups are singletons.

As it is mentioned in [13], and considering Tukey's dictum: *Never estimate intercepts, always estimate centercepts*, the quantitative variables in the model are centered and re-scaled by dividing by their standard deviations. Therefore, the estimated intercept may be interpreted as the weight of an infant born to a 23 year old mother, whose weight was 130 pounds, her race was other (neither black nor white), she was a non-smoker during her pregnancy, with on average 0.16 previous premature labours, no history of hypertension, no presence of irritability, and an average of 0.47 physician visits during the first trimester.

A small amount of jitter is added to the variables y and X, to guarantee that the one-at-a-time condition mentioned at the end of Section 4 is satisfied (see e.g. also [7, p438]). We have verified that this does not change the outcome of the numerical experiments.

Figure 2: Analysis of the LBW data set. The first row shows (part of) the solution path given by the algorithm in Section 4 for $\tau = 0.1, 0.5, 0.9$. The vertical dashed line indicates the model chosen by the BIC (for $R = R_{\mathrm{BIC}}$). The second row shows the corresponding coefficients $\hat{\beta}$ for the BIC model.

Figure 2 (top) presents (part of) the solution path of the model coefficients $\hat{\beta}$ for for $\tau = 0.1$, $\tau = 0.5$, and $\tau = 0.9$. A vertical dashed line is drawn at $R = R_{\mathrm{BIC}}$ to indicate the coefficients chosen by the BIC (14). The bottom row of this figure contains the three corresponding model coefficients (chosen by BIC). As one may see, the models chosen for various values of τ are different, i.e. the effective variables for the lower tail of the conditional distribution are different from the ones for the median and for the upper tail. Since we are interested in studying the effective variables on the LBW, using loss functions such as least-squares (with minimizer equal to the conditional mean) would only give partial information. This result is in accordance with [27] who remarked that the effects of the explanatory variables in LBW data set are not constant across the conditional distribution of the infant's weight. In other words, for different quantiles one may have different models.

5.2 Simultaneous variable selection for a vector of response variables

Consider $Y = (y_1, y_2, \ldots, y_p)$ a vector of possibly correlated response variables. One is interested in modelling these variables using a common set of explanatory variables $X = (x_1, x_2, \ldots, x_m)$. For variable selection purposes, one may consider p linear models $y_j = X\beta + \epsilon$, $(j = 1, \ldots, p)$ separately and find e.g. the penalized model solution for β for each model. But as [15] have suggested, it is sometimes interesting to select variables by considering all p response variables simultaneously, specially when these p variables are correlated.

Suppose we have n observations y_{ij} (with $i = 1, \ldots, n$ and $j = 1, \ldots, p$) for $Y = (y_1, y_2, \ldots, y_p)$

as response variables and $X = (x_1, x_2, \ldots, x_m)$ as explanatory variables. Let β_{kj}, $(k = 1, \ldots, m$ and $j = 1, \ldots, p)$ be the regression coefficient of x_k regressed on y_j. As [15] have observed, $\|(\beta_{k1}, \beta_{k2}, \ldots, \beta_{kp})\|_\infty = \max\{|\beta_{k1}|, |\beta_{k2}|, \ldots, |\beta_{kp}|\}$ is a reasonable measure of the explanatory power of the regressor x_k on all p response variables y_j simultaneously. If the least-squares loss of [15] is replaced by the more robust loss function (3), the following optimization problem:

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} \sum_{j=1}^{p} \varrho_\tau \left(y_{ij} - \sum_{k=1}^{m} x_{ik}\beta_{kj} \right) + \lambda \sum_{k=1}^{m} \max\{|\beta_{k1}|, |\beta_{k2}|, \ldots, |\beta_{kp}|\}. \tag{15}$$

should be solved to select the variables. Equivalently, one could also use the constrained formulation taking the form:

$$\hat{\beta} = \arg\min_{\sum_{k=1}^{m} \max\{|\beta_{k1}|, |\beta_{k1}|, \ldots, |\beta_{kp}|\} \leq R} \sum_{i=1}^{n} \sum_{j=1}^{p} \varrho_\tau \left(y_{ij} - \sum_{k=1}^{m} x_{ik}\beta_{kj} \right). \tag{16}$$

As the argument of ϱ_τ in these last two expressions is a linear function of the β_{kj}, problems (15) and (16) are special cases of problems (6) and (7), and can therefore be solved by the algorithm of Section 4. As [15] pointed out, this is an exploratory tool for identifying a suitable subset of regressor variables, not for actual parameter estimation. The problem (15) has already been proposed by [16] who solved it for a fixed value of R using a generic linear programming code. The algorithm of Section 4 finds the minimizer for a whole range of values of R. The latter algorithm is therefore more useful as the BIC criterion (14) (also used by [16]) requires a further minimization over many values of the parameter R.

As an example, we consider the '93CARS' data set which contains information on 93 new cars for the 1993 model year which is obtained from [14]. Table 2 presents the variables we have considered. Some of the observations have been omitted due to missing values, so in total 82 observations are used.

Figure 3 shows the pairwise scatter plots of the 5 response variables. The two main points which follow from this figure are: 1. the variables are correlated, and 2. there are some outliers in the data. Therefore, a simultaneous variables selection using least absolute deviation (QR with $\tau = 0.5$) seems reasonable here. Both response variables y and regression variables X are standardized, so one would be able to compare variables in different measures.

The algorithm of Section 4 is used with $\tau = 0.5$. In this application too, a small amount of jitter is added to the variables y, so as to guarantee that the one-at-a-time condition mentioned at the end of section 4 is satisfied (see e.g. also [7, p438]). We have again verified that this does not change the outcome of the numerical experiments. Figure 4 (left) shows the presence or absence of each group (as a functions of $R = \|\beta\|_{1,\infty}$).

The BIC (14) is used to select a model among all possible models calculated along the solution path. The corresponding value of $\|\hat{\beta}\|_{1,\infty}$ is called R_{BIC}, and is indicated on the first panel with a dotted line. The final model is given in Figure 4, (right). As one may see, the variables x_2, x_8, x_{11} and x_{12} were not selected.

6 Conclusions

A structured sparse solution (or group sparse solution) of a quantile regression model, based on penalizing or constraining the quantile regression loss function by a mixed $\ell_{1,\infty}$-norm of regression coefficients, was discussed.

Figure 3: Pairwise scatter plots of the 5 response variables in the 93CARS data set. The histogram of each response variable y is presented on the diagonal.

Figure 4: Structured sparse quantile regression of the 93CARS data set. Left: the ℓ_∞-norm of coefficients $\hat{\beta}$'s in each group (vertical) as a function of R (horizontal). Right: (top) presence or absence of each group along the solution path, (bottom) the ℓ_∞-norm of coefficients $\hat{\beta}$ in each group for $R = R_{\mathrm{BIC}}$ (selected using the BIC (14)).

Table 2: Variables in 93CARS data set.

	Variable	Description
y_1:	Minimum price (in \$1,000)	Price for basic version of this model
y_2:	Midrange price (in \$1,000)	Average of Min and Max prices
y_3:	Maximum price (in \$1,000)	Price for a premium version
y_4:	City MPG	miles per gallon by EPA rating
y_5:	Highway MPG	-
x_1:	Number of cylinders	-
x_2:	Engine size	in liters
x_3:	Horsepower	maximum
x_4:	RPM	revs per minute at maximum horsepower
x_5:	Engine revolutions per mile	(in highest gear)
x_6:	Fuel tank capacity	in gallons
x_7:	Passenger capacity	in persons
x_8:	Length	in inches
x_9:	Wheelbase	in inches
x_{10}:	Width	in inches
x_{11}:	U-turn space	in feet
x_{12}:	Rear seat room	in inches
x_{13}:	Luggage capacity	in cubic feet
x_{14}:	Weight	in pounds

An algorithm to compute the solution of the corresponding minimization problem was presented. This algorithm computes the minimizer of the penalized or constrained loss function for all values of $R = \|\hat{\beta}\|_{1,\infty}$ within a range $[0, R_{\max}]$ instead of just for a single value of R. This is a strong point when using the BIC criterion (which needs a further minimization over R) for model selection.

In a first application, the effective variables for the lower and upper quantiles of the conditional distribution of the birth weight of infants in the LBW data set were identified, subject to a group sparsity constraint. As a second application we studied the problem of simultaneous variable selection in a quantile regression model for robustly modeling a vector of possibly correlated response variables using a common set of explanatory variables.

The implementation of the algorithm presented in Section 4 is not straightforward. Therefore, such an implementation in Matlab is provided on the authors' webpage [12], together with the scripts for processing the LBW data set and the 93CARS data set. The necessary functions for interpolating the solution $\hat{\beta}$ between the nodes and choosing the best model using the proposed BIC (14) are also provided.

The current short article dealt with the non-overlapping group case. An potential extension of the algorithm would consist of including the overlapping group case as well.

Acknowledgements

References

[1] X. Gabaix and D. Laibson. The seven properties of good models. *The Methodologies of Modern Economics: Foundations of Positive and Normative Economics*, Oxford University Press, 2008.

[2] R. Koenker and G. Bassett. Regression quantiles. *Econometrics*, 46:33–50, 1978.

[3] G. Claeskens and N. L. Hjort. *Model Selection and Model Averaging*. Cambridge University Press, Cambridge, 2008.

[4] Jianqing Fan and Runze Li. Statistical challenges with high dimensionality: feature selection in knowledge discovery. In *International Congress of Mathematicians. Vol. III*, pages 595–622. Eur. Math. Soc., Zürich, 2006.

[5] Robert Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996.

[6] M. R. Osborne, B. Presnell, and B. A. Turlach. A new approach to variable selection in least squares problems. *IMA J. Numer. Anal.*, 20(3):389–403, July 2000.

[7] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Ann. Statist.*, 32(2):407–499, 2004.

[8] Y. Li and J. Zhu. ℓ_1-norm quantile regression. *Journal of Computational and Graphical Statistics*, 17:1–23, 2008.

[9] J. J. Fuchs. Fast implementation of a $\ell_1 - \ell_1$ regularized representations algorithm. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 3329–3332, 2009.

[10] A. Cohen. Dummy variables in stepwise regression. *The American Statistician*, 45(3):226–228, 1991.

[11] M. Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 68(1):49–67, 2006.

[12] Ignace Loris and Vahid Nassiri. Group quantile regression Matlab code, 2013.

[13] R. Koenker and K. F. Hallock. Quantile regression. *Journal of Economic Perspectives*, 15:143–156, 2001.

[14] Robin H. Lock. 1993 new car data. *Journal of Statistics Education*, 1(1), 1993.

[15] Berwin A. Turlach, William N. Venables, and Stephen J. Wright. Simultaneous variable selection. *Technometrics*, 47(3):349–363, 2005.

[16] Hui Zou and Ming Yuan. Regularized simultaneous model selection in multiple quantiles regression. *Computational Statistics and Data Analysis*, 52:5296–5304, 2008.

[17] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

[18] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge University Press, Cambridge, 2004.

[19] Saharon Rosset and Ji Zhu. Piecewise linear regularized solution paths. *The Annals of Statistics*, 35(3):1012–1030, 2007.

[20] Roger Koenker, Pin Ng, and Stephen Portnoy. Quantile smoothing splines. *Biometrika*, 81(4):673–680, 1994.

[21] Gideon Schwarz. Estimating the dimension of a model. *Ann. Statist.*, 6(2):461–464, 1978.

[22] World Health Organization. International statistical classification of diseases and related health problems (ICD-10). `http://www.who.int/classifications/icd/en/`, 1992. 10th Revision.

[23] D. A. Boardman, J. D.and Powers, Y. C. Padilla, and R. A Hummer. Low birth weight, social factors, and developmental outcomes among children in the united states. *Demography*, 39:353–368, 2002.

[24] G. K. Auslander, D. Netzer, and I. Arad. Parental anxiety following discharge from hospital of their very low birth weight infants. *Family Relations*, 52:12–21, 2003.

[25] D. Almond, K. Y. Chay, and D. S. Lee. The costs of low birth weight. *The Quarterly Journal of Economics*, 120:1031–1083, 2005.

[26] D. W. Hosmer, , and S. Lameshow. *Applied Logistic Regression*. Wiley, 1989.

[27] Roger Koenker and José A. F. Machado. Goodness of fit and related inference processes for quantile regression. *J. Amer. Statist. Assoc.*, 94(448):1296–1310, 1999.