Quantile Regression via an MM Algorithm

David R. Hunter[1]
Kenneth Lange[2]

Department of Statistics[1]
Penn State University
University Park, PA 16802-2111

Departments of Biomathematics and Human Genetics[2]
UCLA School of Medicine
Los Angeles, CA 90095-1766

e-mail: dhunter@stat.psu.edu[1]
phone: (814) 863-0979[1]
fax: (814) 863-7114[1]

Research supported in part by USPHS grant GM53275[2].

Submitted to J of Computational and Graphical Statistics
December 23, 1998

Resubmitted September 30, 1999

Abstract

Quantile regression is an increasingly popular method for estimating the quantiles of a distribution conditional on the values of covariates. Regression quantiles are robust against the influence of outliers, and taken several at a time, they give a more complete picture of the conditional distribution than a single estimate of the center. The current paper first presents an iterative algorithm for finding sample quantiles without sorting and then explores a generalization of the algorithm to nonlinear quantile regression. Our quantile regression algorithm is termed an MM, or Majorize-Minimize, algorithm because it entails majorizing the objective function by a quadratic function followed by minimizing that quadratic. The algorithm is conceptually simple and easy to code, and our numerical tests suggest that it is computationally competitive with a recent interior point algorithm for most problems.

Key words and phrases: L_1 regression, majorization, EM algorithm, Gauss-Newton method.

1 Introduction

The fact that the median $\mu_{1/2}$ of a random variable Y minimizes the expectation $f(\mu) = \mathrm{E}\,(|Y-\mu|)$ is well known among theoretical statisticians (Casella and Berger, 1990). Perhaps less well known is that this characterization of the median forms the basis of an iteratively reweighted least squares algorithm for finding the sample median of n numbers $y_1, \ldots, y_n$ without sorting (Borg and Groenen, 1997; Heiser, 1995; Mosteller and Tukey, 1977; Press et al., 1986). If $\mu_{1/2}^k$ is the kth iterate of this classical algorithm, then the next iterate is

$$\mu_{1/2}^{k+1} \quad = \quad \frac{\sum_{i=1}^n w_i^k y_i}{\sum_{i=1}^n w_i^k}, \tag{1}$$

where $w_i^k = |y_i - \mu_{1/2}^k|^{-1}$.

To generalize this algorithm to an arbitrary sample quantile μ_q, consider the following heuristic argument. Assuming that $w_i = |y_i - \mu_q|^{-1}$ is well defined for each i, note that $w_i(y_i - \mu_q) = \pm 1$ depending on whether $y_i > \mu_q$ or $y_i < \mu_q$. Since

nq of the y_i should be less than the q quantile μ_q, it follows that

$$\sum_{i=1}^{n}(y_i - \mu_q)w_i \quad = \quad -nq + n(1-q).$$

Rearranging, we obtain the algorithm

$$\mu_q^{k+1} \quad = \quad \frac{n(2q-1) + \sum_{i=1}^{n} w_i^k y_i}{\sum_{i=1}^{n} w_i^k} \tag{2}$$

analogous to equation (1).

We will not attempt to make this illustrative argument rigorous. A flaw of the algorithm is that the weight w_i^k is undefined whenever $y_i = \mu_q^k$. In mending this flaw, we will modify and generalize the algorithm so that it applies to the broader problem of quantile regression as defined by Koenker and Bassett (1978). The general algorithm we propose depends on a technique called Majorization-Minimization, or MM, in the rejoinder to the paper of Lange et al. (2000). Section 2 of the current paper presents a brief overview of the MM principle. In Section 3, we define regression quantiles and discuss the application of an MM algorithm to the problem of computing them. Section 4 sketches theory governing the convergence of the algorithm, and Section 5 describes numerical tests of the algorithm. The main body of the paper ends with a discussion of the comparative merits of the algorithm. All mathematical proofs appear in the appendix.

2 MM Algorithms

The technique of using majorizing functions to perform minimization is at least thirty years old (Ortega and Rheinboldt, 1970, p. 253) and has surfaced from time to time in the statistical literature. The best-known example of an MM algorithm is the EM algorithm for maximum likelihood estimation in the presence of missing

data (Dempster et al., 1977). In the case of EM though, MM stands for Minorize-Maximize instead of Majorize-Minimize because the goal is maximization, not minimization. Examples of the MM algorithm not involving missing data can be found, for example, in de Leeuw (1994), Heiser (1995), Becker et al. (1997), and Lange et al. (2000). In essence, the MM algorithm replaces a difficult optimization problem by a sequence of easier optimization problems. In most cases, the solutions of the substitute problems converge to a solution of the original problem.

To expose the nature of the MM algorithm, suppose we want to minimize the objective function $L(\theta) : R^p \to R$. If θ^k denotes the current iterate in finding the minimum point, then as the name suggests, the Majorize-Minimize algorithm proceeds in two steps. First, we create a surrogate function $Q(\theta \mid \theta^k) : R^p \times R^p \to R$ satisfying

$$\begin{aligned} Q(\theta^k \mid \theta^k) &= L(\theta^k) && (3) \\ Q(\theta \mid \theta^k) &\geq L(\theta) \quad \text{for all } \theta. && (4) \end{aligned}$$

The function $Q(\theta \mid \theta^k)$ is said to majorize $L(\theta)$ at θ^k. In what follows, we always understand the current iterate θ^k to be a constant; thus, through a slight abuse of notation, $Q(\theta \mid \theta^k)$ denotes the function $\theta \mapsto Q(\theta \mid \theta^k)$ on R^p. In the second step in the MM algorithm, we choose θ^{k+1} to minimize $Q(\theta \mid \theta^k)$. In general, it is a challenge to construct a good surrogate function which simultaneously majorizes $L(\theta)$ at θ^k and is itself easy to minimize. This is exactly what is accomplished by the E step of a well-conceived EM algorithm, though of course the object of an EM algorithm is maximization, not minimization.

Defining the next iterate θ^{k+1} to minimize $Q(\theta \mid \theta^k)$ implies in particular that

$$Q(\theta^{k+1} \mid \theta^k) \leq Q(\theta^k \mid \theta^k). \quad (5)$$

This inequality plus conditions (3) and (4) entail the descent property

$$L(\theta^{k+1}) \quad \leq \quad L(\theta^k). \tag{6}$$

Equality can hold in inequality (6) only if equality holds in inequality (5). The descent property lends MM algorithms their remarkable numerical stability.

As an illustration of the MM principle, we return to the quantile-finding algorithm (2). The q quantile of an integrable random variable Y minimizes the function $\mathrm{E}\,[\rho_q(Y-\mu)]$, where

$$\rho_q(r) \quad = \quad |r|\left[q1_{\{r\geq 0\}} + (1-q)1_{\{r<0\}}\right] \quad = \quad qr - r1_{\{r<0\}}. \tag{7}$$

Although long known, this fact has "languished in the status of curiosum—appearing for example as an exercise in Ferguson (1967, p. 51)," as Koenker and Bassett (1978) put it. Because the statistical literature seems to lack a rigorous proof of this principle that does not impose unnecessary distributional assumptions on Y such as the existence of a density, we prove it as Proposition 1 in the appendix. The empirical version of the principle says that a sample q quantile μ_q of n numbers $y_1, \ldots, y_n$ is a minimizer of the function $L(\mu) = \sum_{i=1}^n \rho_q(y_i - \mu)$.

For example, if $n = 5$, $q = 1/4$, and the sample consists of the points 1, 3, 4, 8, and 10, then the objective function $L(\mu)$ is pictured in Figure 1. Given the value of the current iterate $\mu^k \notin \{1, 3, 4, 8, 10\}$, one can construct a majorizing function by specifying for each i the unique quadratic curve tangent to the graph of $\rho_q(y_i - \mu)$ at the points $\mu = \pm\mu^k$, namely

$$\zeta_q(y_i - \mu \mid y_i - \mu^k) \quad = \quad \frac{1}{4}\left[\frac{(y_i-\mu)^2}{|y_i-\mu^k|} + (4q-2)(y_i-\mu) + |y_i-\mu^k|\right]. \tag{8}$$

Summing equation (8) over i gives a surrogate function $Q(\mu \mid \mu^k)$, which when minimized yields the update (2). As Figure 1 depicts, $Q(\mu \mid \mu^k)$ majorizes $L(\mu)$ at

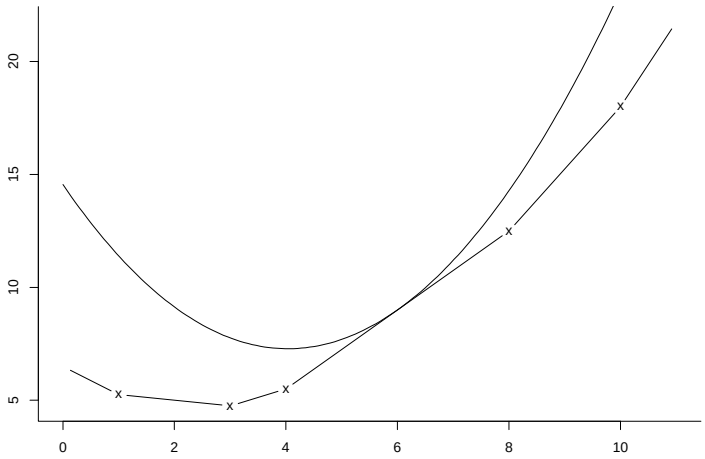

Figure 1: The lower, piecewise linear function is the objective function for the 1/4 quantile of the points $\{1, 3, 4, 8, 10\}$. The upper curve is the majorizing surrogate function $Q(\mu \mid \mu^k)$ corresponding to $\mu^k = 6$.

the point $\mu = \mu^k$. Verboon (1994) discusses MM algorithms for robust regression, among them the algorithm using the majorizer (8) in the case $q = 1/2$. Heiser (1995) and Borg and Groenen (1997) also depict this algorithm for $q = 1/2$.

In closing this section, we reiterate that the algorithm (2) highlighted in the preceding example suffers from the fact that $\rho_q(y_i - \mu^k) = 0$ when $y_i = \mu^k$. The importance of this shortcoming is underscored by the fact that for nq not an integer, $\{\mu^k\}$ should converge to one of the y_i. In the case $q = 1/2$, several authors (Mosteller and Tukey, 1977; Lange and Sinsheimer, 1993; Heiser, 1995) suggest bounding the terms $\rho_q(y_i - \mu^k)$ away from zero, but of course this changes the weights and the algorithm, and care must be taken to ensure that the desirable descent property of the MM algorithm is not destroyed. The vexing problem of zero residuals persists in the more general setting of quantile regression.

3 Quantile Regression

Koenker and Bassett (1978) define a regression quantile as any vector $\hat{\theta} \in R^p$ minimizing the sum

$$L(\theta) \quad = \quad \sum_{i=1}^{n} \rho_q \left[y_i - f_i(\theta) \right].$$

To simplify notation, we define the ith residual $r_i(\theta) = y_i - f_i(\theta)$. Often, we will simply write r_i, omitting the explicit dependence on θ. In the case of linear quantile regression, x_i will denote the ith row of the $n \times p$ matrix X of covariates; in this notation, $f_i(\theta) = x_i\theta$. In general, we will assume that each $f_i(\theta) : R^p \to R$ is continuously differentiable with differential $df_i(\theta)$. Since $L(\theta)$ is nonnegative, it must have an infimum; however, there is no guarantee that the infimum is actually attained. To ensure that a regression quantile actually exists, we impose the technical condition

$$\lim_{\|\theta\| \to \infty} \sum_{i=1}^{n} f_i(\theta)^2 \quad = \quad \infty. \tag{9}$$

For $q = 1/2$, quantile regression is least absolute deviation, or L_1, regression. Many authors have studied L_1 regression, particularly in the linear case (Bassett and Koenker, 1978; Rousseeuw and Leroy, 1987; Schlossmacher, 1973). For arbitrary q, a regression quantile $\hat{\theta}$ provides an intuitively appealing estimate of the q quantile of Y through the functions $f_i(\theta)$. For example, if $f_i(\theta) = f(x_i, \theta)$ with the x_i as predictors, then $\hat{Y} = f(x, \hat{\theta})$ is an estimate of the conditional quantile of Y given data x.

Alternative methods for estimating conditional quantiles exist (Cole, 1988; Efron, 1991; He, 1997), but the sole focus of this paper is quantile regression. Regression

quantiles are attractive not only because they are robust against non-Gaussian errors in a way that least squares estimates are not — a well-known feature of L_1 regression — but also because several quantiles convey a more complete picture of the conditional distribution of the dependent variable than the single mean derived from a traditional least squares approach. Applications of quantile regression, particularly in econometrics, (Buchinsky, 1995; Eide and Showalter, 1998; Taylor and Bunn, 1998) have advanced hand in hand with theory (Koenker and Bassett, 1982; Powell, 1986; Portnoy and Koenker, 1989). These advances, along with the obvious popularity of quantile regression as a tool for analyzing large data sets, motivate the search for improved methods of quantile regression (Portnoy and Koenker, 1997).

The objective function $L(\theta)$ in (9) is difficult to minimize because it can admit multiple minima and because the underlying function $\rho_q(r)$ is nondifferentiable at $r = 0$. Our approach to this minimization problem is first to construct a function that approximates $L(\theta)$ very closely and then to use an MM algorithm to minimize the approximating function. The first stage is hardly new; several authors have proposed approximations to the objective function $L(\theta)$ when $q = 1/2$ (Merle and Späth, 1974; El-Attar et al., 1979; and Madsen and Nielsen, 1990). However, even twice-differentiable approximations tend to be hard to minimize by standard methods. The novelty of our approach consists of combining a good approximation with an MM algorithm.

Starting with the function $\rho_q(r)$ which underlies $L(\theta)$, we define for $\epsilon > 0$ the perturbation

$$\rho_q^\epsilon(r) \quad = \quad \rho_q(r) - \frac{\epsilon}{2} \ln\left(\epsilon + |r|\right). \tag{10}$$

Then the sum

$$L_\epsilon(\theta) = \sum_{i=1}^n \rho_q^\epsilon(r_i) \tag{11}$$

approximates $L(\theta)$. It turns out that for a given residual value $r^k = r(\theta^k)$ at iteration k, $\rho_q^\epsilon(r)$ is majorized at r^k by the quadratic function

$$\zeta_q^\epsilon(r \mid r^k) = \frac{1}{4}\left[\frac{(r)^2}{\epsilon + |r^k|} + (4q-2)r + c\right], \tag{12}$$

where c is a constant chosen so that $\zeta_q^\epsilon(r^k \mid r^k) = \rho_q^\epsilon(r^k)$. We prove this claim as Proposition 2 in the appendix. The MM algorithm operates by minimizing the majorizer

$$Q_\epsilon(\theta \mid \theta^k) = \sum_{i=1}^n \zeta_q^\epsilon(r_i \mid r_i^k) \tag{13}$$

with respect to θ. The minimizer becomes the next iterate θ^{k+1}. Figure 2 continues the example depicted in Figure 1 and displays the objective function $L(\mu)$, the approximant $L_\epsilon(\mu)$, and the quadratic majorizing function $Q_\epsilon(\mu \mid \mu^k)$ for the relatively large choice $\epsilon = 1/5$.

The fact that $Q_\epsilon(\theta \mid \theta^k)$ is quadratic in the residuals r_i does not imply that it is quadratic in θ unless the $f_i(\theta)$ are linear. In the linear case, we can explicitly solve for θ^{k+1}; otherwise, we usually settle for driving $Q_\epsilon(\theta \mid \theta^k)$ downhill rather than finding its minimum. This compromise preserves the descent property (6). Note that we still use the term MM to describe an algorithm in which the second step consists of merely decreasing the majorizer instead of actually minimizing it. Although one step of Newton's method is a natural choice for driving $Q_\epsilon(\theta \mid \theta^k)$ downhill, it suffers two drawbacks. First, if the Hessian $d^2Q_\epsilon(\theta \mid \theta^k)$ fails to be positive definite at θ^k, then one step of Newton's method may actually increase

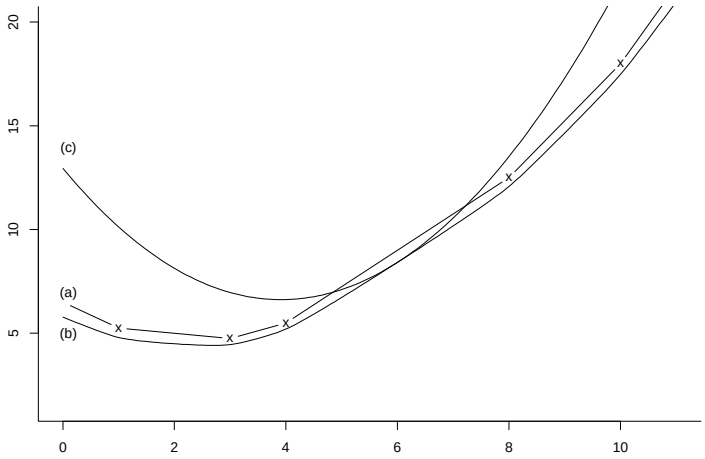

Figure 2: Continuing the example depicted in Figure 1, $L(\mu)$ is a piecewise linear objective function (a), $L_\epsilon(\mu)$ is the approximant (b) to the objective function, and $Q_\epsilon(\mu \mid \mu^k)$ is the surrogate (c) majorizing the approximant at $\mu^k = 6$. Here, $\epsilon = 1/5$.

$Q_\epsilon(\theta \mid \theta^k)$ as a function of θ. Second, calculation of $d^2 Q_\epsilon(\theta \mid \theta^k)$ requires the second differentials $d^2 f_i(\theta)$ of the various regression functions. These second differentials may be cumbersome and time consuming to evaluate.

Instead, we take a Gauss-Newton approach (Kennedy and Gentle, 1980) and approximate $d^2 Q_\epsilon(\theta \mid \theta^k)$ by

$$\begin{aligned} d^2 Q_\epsilon(\theta \mid \theta^k) &= \frac{1}{2} \sum_{i=1}^n \left[\frac{1}{\epsilon + |r_i^k|} dr_i(\theta)^t dr_i(\theta) + \left(\frac{r_i}{\epsilon + |r_i^k|} + 2q - 1 \right) d^2 r_i(\theta) \right] \\ &\approx \frac{1}{2} \sum_{i=1}^n \frac{1}{\epsilon + |r_i^k|} dr_i(\theta)^t dr_i(\theta). \end{aligned} \tag{14}$$

Since $dr_i(\theta) = -df_i(\theta)$, this approximation is exact when all $f_i(\theta)$ are linear. We may write (14) succinctly as

$$d^2 Q_\epsilon(\theta \mid \theta^k) \approx \frac{1}{2} df(\theta)^t W_\epsilon(\theta^k) df(\theta),$$

where $df(\theta)$ is the $n \times p$ matrix with entry $\frac{\partial}{\partial \theta_j} f_i(\theta)$ in row i and column j, and $W_\epsilon^k(\theta)$ is an $n \times n$ diagonal matrix with ith diagonal entry $[\epsilon + |r_i^k(\theta)|]^{-1}$. Given the first differential

$$\begin{aligned} dQ_\epsilon(\theta \mid \theta^k) &= \frac{1}{2} \sum_{i=1}^n \left[\frac{r_i}{\epsilon + |r_i^k|} + 2q - 1 \right] dr_i(\theta) \\ &= \frac{1}{2} v_\epsilon(\theta) df(\theta), \end{aligned} \tag{15}$$

where

$$v_\epsilon(\theta) = \left(1 - 2q - \frac{r_1(\theta)}{\epsilon + |r_1(\theta)|} \ , \ \cdots \ , \ 1 - 2q - \frac{r_n(\theta)}{\epsilon + |r_n(\theta)|} \right),$$

the Gauss-Newton step direction is

$$\Delta_\epsilon^k = -[df(\theta^k)^t W_\epsilon(\theta^k) df(\theta^k)]^{-1} df(\theta^k)^t v_\epsilon(\theta^k)^t. \tag{16}$$

In the special case of linear regression functions $f_i(\theta) = x_i\theta$, the matrix $df(\theta) = X$, and $\theta^{k+1} = \theta^k + \Delta_\epsilon^k$ exactly solves the equation $dQ_\epsilon(\theta \mid \theta^k) = \mathbf{0}$.

Although the matrix $df(\theta^k)^t W_\epsilon(\theta^k) df(\theta^k)$ seen in (16) is positive definite whenever the differential $df(\theta^k)$ has full rank, there is no guarantee that $\theta^k + \Delta_\epsilon^k$ will reduce the value of the surrogate function $Q_\epsilon(\theta \mid \theta^k)$ as required. However, if we take an appropriate fractional step size $\alpha^k \in (0, 1]$, then the iterate

$$\theta^{k+1} = \theta^k + \alpha^k \Delta_\epsilon^k \tag{17}$$

is guaranteed to decrease the value of the surrogate function. For instance, the tactic of step halving dictates that

$$\alpha^k = \max \{2^{-\nu} : Q_\epsilon(\theta^k + 2^{-\nu}\Delta_\epsilon^k \mid \theta^k) < Q_\epsilon(\theta^k \mid \theta^k), \ \nu \in N\}, \tag{18}$$

where N denotes the set of nonnegative integers. When $\Delta_\epsilon^k = \mathbf{0}$, the algorithm is at a stationary point, and we interpret $\alpha^k = 1$.

Actual implementation of the MM algorithm, then, involves the following steps:

1. Select a starting value θ^0 and small constant ϵ. Set $k = 0$.

2. Define θ^{k+1} as in equation (17), where α^k and Δ_ϵ^k are given in equations (18) and (16).

3. Replace k by $k + 1$; if convergence criterion is not met, return to step 2.

We defer further discussion of the selection of θ^0, ϵ, and the convergence criterion until Section 5.

4 Convergence Results

For theoretical purposes, it is helpful to confine attention to a compact subset Ω of parameter space. The next proposition, proved in the appendix, shows how this goal can be achieved.

Proposition 3 *Let θ^0 be an arbitrary point of R^p. The compact set*

$$\Omega \;=\; \{\theta \in R^p : L_1(\theta) \leq L(\theta^0) + n\} \tag{19}$$

contains all θ satisfying $L(\theta) \leq L(\theta^0)$ or $L_\epsilon(\theta) \leq L_\epsilon(\theta^0)$ for any $\epsilon \in (0, 1]$. In particular, Ω contains all iterates of any MM algorithm beginning at θ^0 and all minimizers of $L(\theta)$ and $L_\epsilon(\theta)$ for all $\epsilon \in (0, 1]$.

In light of the proposition, we make the harmless assumption that ϵ is restricted to the interval $(0, 1]$.

We now consider the two theoretical issues of whether the algorithm (17) is guaranteed to minimize the function $L_\epsilon(\theta)$ and how close a minimum of $L_\epsilon(\theta)$ is to a minimum of the original objective function $L(\theta)$. The answer to the first question is, in general, no. However, for linear quantile regression the following proposition holds.

Proposition 4 *For linear quantile regression with a full-rank covariate matrix X, the algorithm (17) converges to the unique minimizer of $L_\epsilon(\theta)$.*

Several factors complicate the theoretical analysis of nonlinear quantile regression. First, the objective function $L_\epsilon(\theta)$ need not be strictly convex and consequently may not possess a unique stationary point. In practice, many different starting values θ^0 could be used to determine the global minimum. Second, step halving comes into play in the definition (18) of α^k. This spoils the continuity of the map $\theta^k \mapsto \theta^{k+1}$, which is used in a fundamental way in the proof of Proposition 4.

If we are willing to assume that $Q_\epsilon(\theta \mid \theta^k)$ is strictly convex in θ, then we may follow Lange (1995) and define

$$\alpha^k \;=\; \arg\min_{\alpha \in [0,1]} Q_\epsilon(\theta^k + \alpha \Delta_\epsilon^k \mid \theta^k).$$

This preserves continuity of the $\theta^k \mapsto \theta^{k+1}$ map, and it follows by the same argument used in the proof of Proposition 4 that all limit points of the modified algorithm are stationary points of $L_\epsilon(\theta)$. If $L_\epsilon(\theta)$ is also strictly convex, then it has at most a single stationary point, and this minimum point is the limit of the algorithm. Of course, in many nonlinear problems there is no reason to assume that either the surrogate function or the objective function is strictly convex. However, we have yet to see a practical example in which the algorithm fails to converge properly.

We now comment on the price paid for replacing the objective function by a differentiable approximation. The following proposition explicitly bounds the difference between the minimum of $L(\theta)$ and its value at a minimizer of $L_\epsilon(\theta)$.

Proposition 5 *Let* $d = 1 + \sup\{|r_i(\theta)| : \theta \in \Omega, 1 \leq i \leq n\}$. *If* $\epsilon d < 1$, *then*

$$|L(\theta) - L_\epsilon(\theta)| \quad \leq \quad -\frac{\epsilon n}{2} \ln \epsilon \tag{20}$$

for all $\theta \in \Omega$. *If* $\hat{\theta}$ *and* $\hat{\theta}_\epsilon$ *minimize* $L(\theta)$ *and* $L_\epsilon(\theta)$, *respectively, then*

$$L(\hat{\theta}_\epsilon) - L(\hat{\theta}) \quad \leq \quad -\epsilon n \ln \epsilon. \tag{21}$$

Putting a bound on $\|\hat{\theta}_\epsilon - \hat{\theta}\|$ proves more difficult because the objective function $L(\theta)$ may be nearly flat in the neighborhood of a minimum. However, we prove in the appendix the following proposition.

Proposition 6 *If* $\hat{\theta}_\epsilon$ *minimizes* $L_\epsilon(\theta)$, *then any limit point of* $\{\hat{\theta}_\epsilon\}$ *as* ϵ *tends to 0 minimizes* $L(\theta)$. *If* $L(\theta)$ *has a unique minimizer* $\hat{\theta}$, *then* $\lim_{\epsilon \to 0} \hat{\theta}_\epsilon = \hat{\theta}$.

5 Numerical Results

This section summarizes the results of several numerical tests of the MM algorithm. As a benchmark for comparison, we also test the interior point algorithm of Koenker and Park (1996). Their method is a primal-dual algorithm, with a dual loop nested within the main primal loop. As recommended by Koenker and Park, we take two dual iterations for each primal iteration and declare convergence whenever the change $|L(\theta^k) - L(\theta^{k+1})|$ in the value of the objective function between successive primal iterations is less than a specified tolerance τ. Additionally, each primal iteration calls for a line search along a specified direction. We use the MATLAB function `fmin` with a range of $(-1, 1)$ for this purpose. Koenker and Park do not specify how large a range to use; larger ranges seem to lead to more convergence failures of the type seen in the Rosenbrock problem of Table 1.

Problem $[p, n]$	Starting θ	Thousands of FLOPs (final value of $\hat{L}$)		
		$q = 0.05$	$q = 0.25$	$q = 0.5$
Bard	$(1, 1, 1)^t$	65.0 (0.035253)	26.4 (0.083237)	13.7 (0.062171)
$[3, 15]$		38.1 (0.035251)	38.5 (0.083095)	45.3 (0.062169)
Beale	$(1, 0)^t$	17.8 (4.4×10^{-6})	3.9 (3.4×10^{-7})	1.1 (0)
$[2, 3]$		9.6 (2.6×10^{-14})	9.6 (1.3×10^{-13})	9.7 (7.7×10^{-13})
Biggs (b)	$(1, 8, 2,$	272.6 (4.5×10^{-6})	56.6 (4.3×10^{-7})	39.6 (7.8×10^{-16})
$[6, 13]$	$2, 2, 2)^t$	114.6 (2.5×10^{-12})	99.9 (3.4×10^{-12})	115.8 (1.2×10^{-14})
Brown	$(25, 5,$	453.5 (45.1617)	263.8 (2.25809)	231.7 (451.617)
$[4, 20]$	$-5, -1)^t$	2169.4 (45.1972)	1795.8 (2.25818)	1276.6 (451.627)
El-Attar 1	$(1, 2)^t$	12.6 (0.050003)	2.3 (0.2500005)	3.3 (0.235212)
$[2, 3]$		7.0 (0.0500000)	6.9 (0.2500000)	8.8 (0.235212)
El-Attar 2	$(1, 1, 1)^t$	73.5 (0.447807)	14.5 (2.23900)	7.0 (4.17724)
$[3, 6]$		27.6 (0.447803)	12.0 (2.25773)	24.0 (3.96668)
Madsen	$(3, 1)^t$	33.3 (0.0500091)	9.2 (0.250001)	5.5 (0.500000)
$[2, 3]$		17.8 (0.0500002)	19.4 (0.250000)	19.7 (0.500000)
Osborne 1	$(.5, 1.5, -1,$	235.4 (0.0023885)	145.3 (0.010336)	445.1 (0.014696)
$[5, 33]$	$.01, .02)^t$	107.3 (0.0023876)	136.8 (0.010247)	167.2 (0.014696)
Osborne 2	$(1.3,.65,.65,.7,.6,$	1307.4 (0.104347)	1476.0 (0.402505)	922.4 (0.577625)
$[11, 65]$	$3,5,7,2,4.5,5.5)^t$	1165.9 (0.099585)	1281.3 (0.402503)	1464.3 (0.577624)
Powell	$(3, -1, 0, 1)^t$	61.0 (4.1×10^{-6})	14.2 (3.7×10^{-7})	6.6 (1.0×10^{-7})
$[4, 4]$		32.9 (1.6×10^{-7})	35.9 (2.0×10^{-7})	38.9 (1.0×10^{-7})
Rosenbrock	$(-1.2, 1)^t$	17.2 (4.1×10^{-6})	3.3 (4.9×10^{-7})	1.4 (0)
$[2, 2]$		Failed to converge	14.3 (6.7×10^{-15})	Failed to converge
Watson	$(1, 1, 1, 1)^t$	345.2 (0.286020)	172.1 (0.399931)	99.0 (0.300928)
$[4, 31]$		741.2 (0.286496)	735.1 (0.399962)	260.6 (0.301057)
Wood	$(0, 0, 0, 0)^t$	75.1 (4.1×10^{-6})	24.4 (4.4×10^{-7})	13.2 (0)
$[4, 6]$		16.1 (3.6×10^{-14})	25.1 (2.7×10^{-14})	33.1 (4.5×10^{-15})
Wormersley	$(0, 0)^t$	316.3 (0.602977)	62.9 (1.82720)	27.5 (1.51627)
$[2, 40]$		132.1 (0.598418)	151.3 (1.68123)	64.8 (1.51627)

Table 1: Thousands of FLOPs required until convergence and final objective function values $\hat{L}$ for test problems of Koenker and Park (1996). In each cell, the MM results are on the first line and the interior point results are on the second line.

To test the performance of the MM algorithm on numerical data, several details must be clarified. First, how is convergence declared? Second, how is ϵ chosen? Third, how is the initial parameter vector θ^0 chosen? Once these issues are addressed, then the algorithm set forth at the end of Section 3 can proceed.

By analogy to the convergence criterion of the interior point algorithm given by Koenker and Park (1996), we declare convergence for the MM algorithm whenever

$$Q_\epsilon(\theta^k \mid \theta^k) - Q_\epsilon(\theta^{k+1} \mid \theta^k) < \tau. \tag{22}$$

Alternatively, one might adopt a scale-invariant convergence criterion which stops

the algorithm when

$$\left| \frac{Q_\epsilon(\theta^k \mid \theta^k) - Q_\epsilon(\theta^{k+1} \mid \theta^k)}{Q_\epsilon(\theta^k \mid \theta^k)} \right| \quad < \quad \tau.$$

Because several of our examples involve an objective function which converges in value to zero, we use criterion (22) for our examples. We also let τ guide the selection of ϵ; in light of Proposition 5, we choose ϵ to satisfy $\epsilon n |\ln \epsilon| = \tau$. Finally, we start the MM algorithm near an ordinary least squares solution unless otherwise specified. This is easy to do since a least squares solution may be approximated by replacing $W_\epsilon(\theta^k)$ by the identity matrix and $v_\epsilon(\theta)^t$ by the residual vector $y - f(\theta)$ in the update (16) for several iterations at the start of the algorithm. In the linear case, a single iteration suffices. We caution the reader that the least-squares starting value is essentially arbitrary, chosen for convenience only, and is not guaranteed to be close to the global minimum. If one suspects that there are multiple local minima, then one can start the algorithm from several different randomly chosen points in an attempt to find the global minimum.

Comparisons between the MM algorithm and the interior point algorithm cannot safely be made on the basis of iteration counts alone. MM algorithms tend to trade fewer arithmetic operations per iteration for more total iterations. Neither is CPU time an adequate measure of performance since different systems and even different compilers on the same system will run the same code at different speeds. Therefore, we opt for the number of FLOPs, or floating point operations, as the basis for comparison and code all examples in MATLAB, which counts FLOPs automatically. This is not a perfect solution; for example, MATLAB fails to count certain operations such as taking the maximum of n numbers, an operation used extensively by the interior point method. However, in our opinion, FLOPs provide the fairest basis for

comparison.

For both the MM algorithm and the interior point method, we use a value of $\tau = 10^{-6}$. Because most MM algorithms typically show very slow convergence in the neighborhood of an optimum point, this value represents a compromise between the competing desires for precise answers and quick convergence.

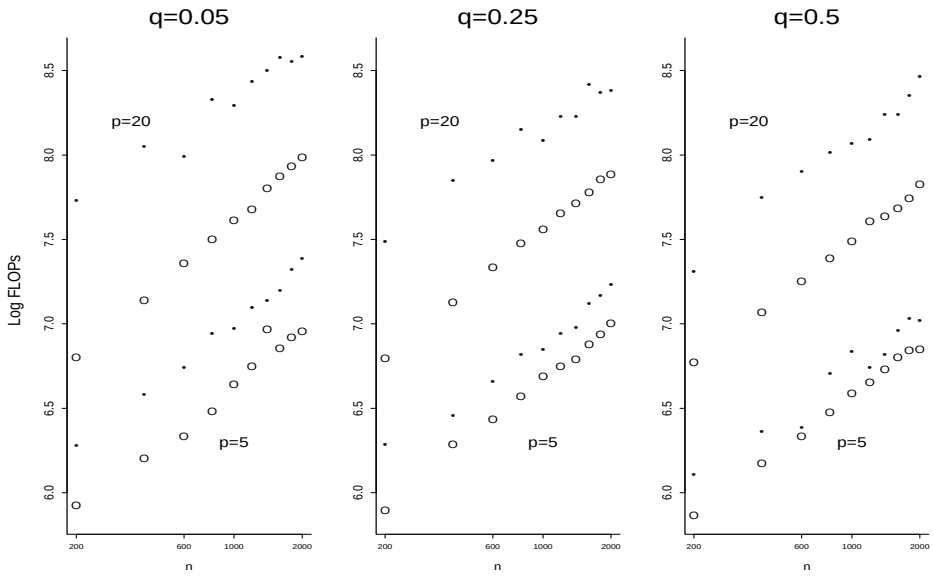

Figure 3: $\log_{10}$ of mean FLOP count for 10 runs of both methods on a linear problem for various values of q, p, and n. Solid dots represent MM runs; open circles represent interior point runs.

Table 1 displays the results of numerical tests of the two algorithms on the 14 nonlinear problems listed by Koenker and Park (1996). Most of these problems are fairly small, but together they represent a wide range of nonlinear objective functions. Overall, neither algorithm is clearly faster. Similarly, the quality of solutions seems comparable based on the values of the objective functions at convergence. Note that exact solutions are known whenever all residuals may be simultaneously zero, as is the case in the Beale, Biggs, Powell, Rosenbrock, and Wood problems. In

such cases, we consider solutions such as 4×10^{-6} correct because they are on the order of the choice $\tau = 10^{-6}$ used to declare convergence.

Although neither of the two algorithms consistently outpaces the other or produces more accurate solutions in the problems of Table 1, the MM algorithm is the more stable of the two. The interior point method fails to converge for the Rosenbrock problem for $q = 0.05$ and $q = 0.5$.

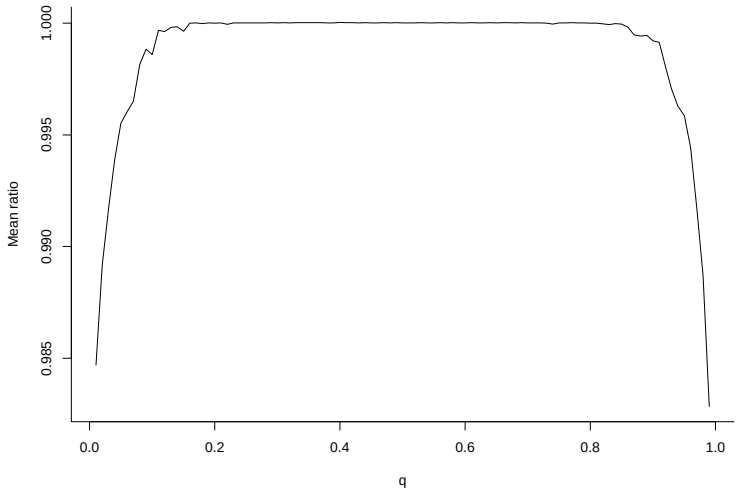

Figure 4: Ratio of $L(\hat{\theta})$ found by the MM algorithm to $L(\hat{\theta})$ found by the interior point method for various values of q, averaged over 100 repetitions. The problem used here is linear with $n = 100$ and $p = 5$.

Because most of the problems of Table 1 are quite small, we consider two additional problems which can be scaled up to any desired size. Let X be an $n \times p$ predictor matrix, each of whose entries is independently uniformly distributed in (0,1). The response vector Y consists of entries $y_i = f_i(x_i, \theta) + \epsilon_i$, where $\theta = (1, \ldots, 1)^t$ and the ϵ_i are independent and normally distributed with zero mean and variance 0.01. For each problem, we use the zero vector as the starting value of θ^0.

The first problem is linear; that is, $y_i = x_i\theta + \epsilon_i$. Figure 3 displays the logarithm base 10 of the mean FLOP counts for 10 repetitions of each problem for $q \in \{0.05, 0.25, 0.5\}$, $p \in \{5, 20\}$, and $n \in \{200, 400, 600, \ldots, 2000\}$. The graph shows a small but consistent gap indicating that the interior point method usually requires fewer FLOPs. However, Figure 4 suggests that the interior point solutions may be slightly inferior to the MM solutions for extreme values of q near 0 or 1 for the particular choices $p = 5$ and $n = 100$.

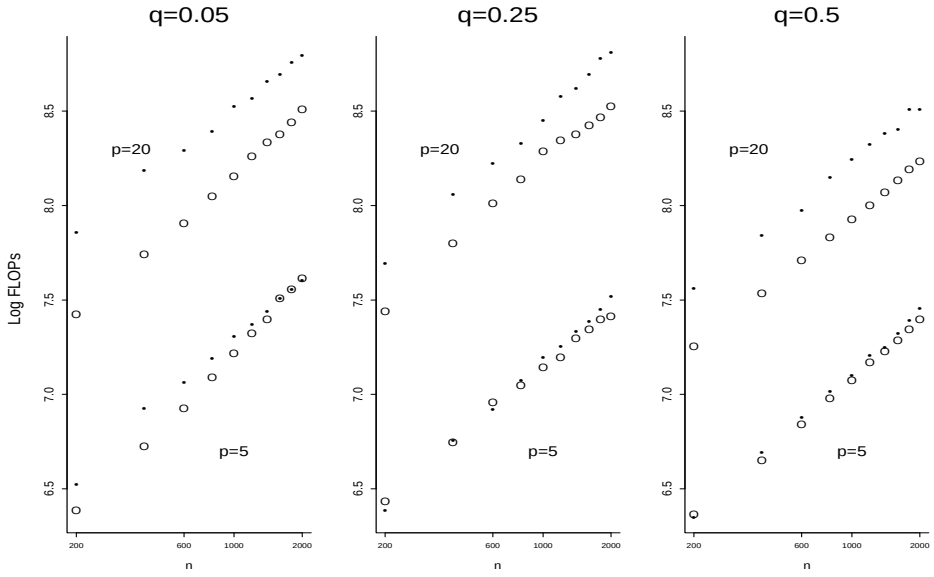

Figure 5: $\log_{10}$ of mean FLOP count for 10 runs of both methods on a nonlinear problem for various values of q, p, and n. Solid dots represent MM runs; open circles represent interior point runs.

The second problem we test is a nonlinear problem in which

$$Y_i \;=\; \sum_{j=1}^{p} \left[e^{-x_{ij}\theta_j^2} + x_{ij}\theta_{p-j+1} \right] + \epsilon_i.$$

Because these nonlinear problems may involve multiple local minima, we do not consider the ratio displayed in Figure 4. Figure 5 summarizes the comparison of

computational speed for the two methods for the same values of q, p, and n considered in Figure 3. The gap seen in Figure 3 is still evident but smaller. In several examples of this nonlinear problem not shown in the graph, the interior point method failed to converge at all. As in Figure 3, the gap between the two algorithms is larger for $p = 20$ than $p = 5$. This undoubtedly reflects the fact that each MM iteration enjoys less of a computational advantage compared to each interior point iteration as p grows because each algorithm inverts a $p \times p$ matrix at each iteration.

6 Discussion

The major strengths of the MM algorithm presented in this paper are its conceptual simplicity, ease of implementation, and numerical stability, qualities it shares with most other MM algorithms. Our numerical tests indicate that the MM algorithm is computationally competitive with Koenker and Park's (1996) interior point method, the current state of the art, on many nonlinear quantile regression problems. However, as the number of parameters increases, it appears that when the interior point method converges, it converges faster on average than the MM algorithm. We feel this will likely be the case for any MM algorithm which requires the inversion at each iteration of a $p \times p$ matrix, as ours does. However, given the success of other MM algorithms in circumventing the need for large matrix inversions (Lange et al., 2000; Erdoğan and Fessler, 1999), we are hopeful that the ideas in this paper will lead to a very fast algorithm for performing quantile regression in high-dimensional parameter space.

Portnoy and Koenker (1997) have recently shown how preprocessing the data points in quantile regression problems can significantly reduce the ensuing computational complexity, particularly when the number of observations is large. This

useful suggestion may be applied in the context of the MM algorithm as well.

7 Appendix: Proofs

Proposition 1 *Any q quantile μ_q of an integrable random variable Y minimizes the expectation $\mathrm{E}\,[\rho_q(Y-\mu)]$, with $\rho_q(r)$ defined in equation (7).*

Proof: The inequalities $\Pr(Y \leq \mu_q) \geq q$ and $\Pr(Y \geq \mu_q) \geq 1-q$ define a q quantile μ_q of a random variable Y. Suppose that $\mu \leq \mu_q$. Taking expectations in the equality

$$\begin{aligned}\rho_q(Y-\mu) - \rho_q(Y-\mu_q) &= (1-q)(\mu-\mu_q)1_{\{Y<\mu_q\}} + q(\mu_q-\mu)1_{\{Y\geq\mu_q\}} \\ &\quad + (Y-\mu)1_{\{\mu<Y<\mu_q\}}\end{aligned}$$

yields

$$\begin{aligned}\mathrm{E}\,[\rho_q(Y-\mu)] - \mathrm{E}\,[\rho_q(Y-\mu_q)] &= (\mu_q-\mu)\,[q\Pr(Y\geq\mu_q) - (1-q)\Pr(Y<\mu_q)] \\ &\quad + \mathrm{E}\,[(Y-\mu)1_{\{\mu<Y<\mu_q\}}]. \qquad (23)\end{aligned}$$

In view of the fact that $q\Pr(Y \geq \mu_q) \geq (1-q)\Pr(Y < \mu_q)$, this representation makes it clear that $\mathrm{E}\,\{\rho_q(Y-\mu)\}$ is a decreasing function of μ on the interval $(-\infty, \mu_q]$. From the corresponding representation

$$\begin{aligned}\mathrm{E}\,[\rho_q(Y-\mu)] - \mathrm{E}\,[\rho_q(Y-\mu_q)] &= (\mu-\mu_q)\,[(1-q)\Pr(Y\leq\mu_q) - q\Pr(Y>\mu_q)] \\ &\quad + \mathrm{E}\,[(\mu-Y)1_{\{\mu_q<Y<\mu\}}]. \qquad (24)\end{aligned}$$

for $\mu \geq \mu_q$, we now conclude that $\mathrm{E}\,[\rho_q(Y-\mu)]$ attains its minimum at $\mu = \mu_q$.

Conversely, suppose that μ_q provides the minimum of $\mathrm{E}\,\{\rho_q(Y-\mu)\}$. Then inserting the limits

$$\lim_{\mu\to\mu_q-} \frac{1}{\mu_q-\mu}\,\mathrm{E}\left[(Y-\mu)1_{\{\mu<Y<\mu_q\}}\right] = 0$$

$$\lim_{\mu \to \mu_q +} \frac{1}{\mu - \mu_q} \mathrm{E}\left[(\mu - Y) 1_{\{\mu_q < Y < \mu\}}\right] = 0$$

derived from the bounded convergence theorem into equations (23) and (24) requires that

$$\begin{aligned} q \Pr(Y \geq \mu_q) - (1-q) \Pr(Y < \mu_q) &\geq 0 \\ (1-q) \Pr(Y \leq \mu_q) - q \Pr(Y > \mu_q) &\geq 0. \end{aligned}$$

These last two inequalities imply that μ_q is a q quantile of Y. ■

Proposition 2 *The function $\zeta_q^\epsilon(r \mid r^k)$ of equation (12) majorizes $\rho_q^\epsilon(r)$ of equation (10) at the point $r = r^k$.*

Proof: We must verify conditions (3) and (4). The first is satisfied by the definition of $\zeta_q^\epsilon(r \mid r^k)$, so it suffices to demonstrate that the difference $\zeta_q^\epsilon(r \mid r^{(k)}) - \rho_q^\epsilon(r)$ attains its minimum at $r = \pm r^{(k)}$. The straightforward calculation

$$\begin{aligned} \zeta_q^\epsilon(r \mid r^{(k)}) - \zeta_q^\epsilon(-r \mid r^{(k)}) &= \rho_q^\epsilon(r) - \rho_q^\epsilon(-r) \\ &= (2q-1)r \end{aligned}$$

shows that $\zeta_q^\epsilon(r \mid r^{(k)}) - \rho_q^\epsilon(r)$ is symmetric around 0. This fact allows us to restrict our attention to the interval $r \geq 0$. Next, note that the derivatives

$$\frac{d}{dr}\rho_q^\epsilon(r) = \begin{cases} q - \frac{\epsilon}{2(\epsilon + r)} & r \geq 0 \\ q - 1 + \frac{\epsilon}{2(\epsilon - r)} & r < 0 \end{cases} \tag{25}$$

and

$$\frac{d}{dr}\zeta_q^\epsilon(r \mid r^{(k)}) = \frac{r}{2(\epsilon + |r^{(k)}|)} + q - \frac{1}{2}$$

can be combined to give

$$\frac{d}{dr}\left[\zeta_q^\epsilon(r \mid r^{(k)}) - \rho_q^\epsilon(r)\right] = \frac{r(r - |r^{(k)}|)}{2(\epsilon + |r^{(k)}|)(\epsilon + |r|)}.$$

Therefore, for $r \geq 0$, the difference $\zeta_q^\epsilon(r \mid r^{(k)}) - \rho_q^\epsilon(r)$ is decreasing to the left of $r^{(k)}$, equals 0 at $r^{(k)}$, and is increasing to the right of $r^{(k)}$. This qualitative behavior validates condition (4). ■

Proof of Proposition 3: The set Ω must be compact because condition (9) implies that $\{\theta : L_\epsilon(\theta) \leq c\}$ is a compact subset of R^p. For any $\epsilon \in (0, 1]$ and $r \in R$, the elementary observation

$$-2 \;<\; \epsilon \ln(\epsilon + |r|) \;\leq\; \ln(1 + |r|) \tag{26}$$

along with definition (11) implies that

$$L_1(\theta) \;\leq\; L_\epsilon(\theta) \;\leq\; L(\theta) + n. \tag{27}$$

Defining $L_0(\theta) = L(\theta)$ for convenience of notation, the bounds (27) hold even for $\epsilon = 0$. Therefore, for any $\epsilon \in [0, 1]$ and any θ satisfying $L_\epsilon(\theta) \leq L_\epsilon(\theta^0)$,

$$L_1(\theta) \;\leq\; L_\epsilon(\theta) \;\leq\; L_\epsilon(\theta^0) \;\leq\; L(\theta^0) + n.$$

This implies that $\theta \in \Omega$. ■

Proof of Proposition 4: Because the step-size constant α^k must equal 1 for linear regression functions, we rewrite algorithm (17) as

$$\theta^{k+1} \;=\; \theta^k + \Delta_\epsilon^k.$$

Examination of equation (16) shows that Δ_ϵ^k is continuous in θ^k and therefore that the iteration map $M(\theta^k) : \theta^k \mapsto \theta^k + \Delta_\epsilon^k$ is continuous.

Differentiating equation (25) gives

$$\begin{aligned} d^2 L_\epsilon(\theta) &= \sum_{i=1}^n \frac{\epsilon}{2(\epsilon + |r_i|)^2} x_i^t x_i \\ &= \frac{\epsilon}{2} X^t W_\epsilon(\theta)^2 X \end{aligned}$$

and reveals that $L_\epsilon(\theta)$ is strictly convex when the covariate matrix X is of full rank. Since $Q_\epsilon(\theta \mid \theta^k)$ is tangent to the strictly convex function $L_\epsilon(\theta)$ at the point $\theta = \theta^k$, the condition $dQ_\epsilon(\theta^k \mid \theta^k) = dL_\epsilon(\theta^k)$ holds, and the MM algorithm has exactly one fixed point, namely the unique minimizer $\hat{\theta}_\epsilon$ of $L_\epsilon(\theta)$. Given a convergent subsequence $\{\theta^{k_n}\}_{n \geq 1}$ with limit θ^*, rewriting inequality (6) as

$$L_\epsilon[M(\theta^{k_n})] \;\leq\; L_\epsilon(\theta^{k_n}),$$

taking limits as n tends to ∞, and invoking the continuity of $M(\theta)$ and $L_\epsilon(\theta)$ demonstrate that

$$L_\epsilon[M(\theta^*)] \;=\; L_\epsilon(\theta^*). \tag{28}$$

If we can show that any θ^* satisfying equation (28) is a fixed point of the MM algorithm, then this forces $\theta^* = \theta_\epsilon$ and proves the proposition. Because equation (28) entails

$$Q_\epsilon[M(\theta^*) \mid \theta^*] \;=\; Q_\epsilon(\theta^* \mid \theta^*),$$

the desired equality $M(\theta^*) = \theta^*$ follows directly from the definition of $M(\theta^*)$ as the unique minimizer of the strictly convex function $Q_\epsilon(\theta \mid \theta^*)$. ■

Proof of Proposition 5: By the compactness of Ω and the continuity of the $f_i(\theta)$, the upper bound d is finite. Clearly $|\ln(\epsilon + |r|)| \leq -\ln \epsilon$ if $\epsilon + |r| \leq 1$. On the other hand, if $\epsilon + |r| > 1$, then setting $\epsilon < d^{-1}$ implies $|\ln(\epsilon + |r|)| < \ln d < -\ln \epsilon$ in view of the definition of d. Thus,

$$\begin{aligned} |L(\theta) - L_\epsilon(\theta)| &= \frac{\epsilon}{2} \left| \sum_{i=1}^n \ln(\epsilon + |r_i|) \right| \\ &\leq \frac{\epsilon}{2} \sum_{i=1}^n |\ln(\epsilon + |r_i|)| \\ &\leq -\frac{\epsilon n}{2} \ln \epsilon. \end{aligned}$$

Inequality (21) follows from

$$\begin{aligned} L(\hat{\theta}_\epsilon) - L(\hat{\theta}) &\leq L(\hat{\theta}_\epsilon) - L_\epsilon(\hat{\theta}_\epsilon) + L_\epsilon(\hat{\theta}) - L(\hat{\theta}) \\ &\leq \left|L(\hat{\theta}_\epsilon) - L_\epsilon(\hat{\theta}_\epsilon)\right| + \left|L_\epsilon(\hat{\theta}) - L(\hat{\theta})\right| \\ &\leq -\epsilon n \ln \epsilon. \end{aligned}$$

■

Proof of Proposition 6: Let ϵ_k be a sequence tending to 0 with $\lim_{k\to\infty} \hat{\theta}_{\epsilon_k} = \theta^*$. Definitions (10) and (11) imply that $\lim_k L_{\epsilon_k}(\theta) = L(\theta)$ for all θ. Therefore, taking limits in the inequality $L_{\epsilon_k}(\theta_{\epsilon_k}) \leq L_{\epsilon_k}(\theta)$ yields $L(\theta^*) \leq L(\theta)$. ■

References

Bassett, G., and Koenker, R. (1978), "Asymptotic theory of least absolute error regression," *Journal of the American Statistical Association*, **73**, 618–622.

Becker, M. P., Yang, I., and Lange, K. (1997), "EM algorithms without missing data," *Statistical Methods in Medical Research*, **6**, 37–53.

Borg, I., and Groenen, P. (1997), *Modern Multidimensional Scaling*, New York: Springer.

Buchinsky, M. (1995), "Quantile regression, Box-Cox transformation model, and the U.S. wage structure," *Journal of Econometrics*, **65**, 109–154.

Casella, G., and Berger, R. L. (1990), *Statistical Inference*, Pacific Grove, CA: Wadsworth & Brooks/Cole.

Cole, T. J. (1988), "Fitting smoothed centile curves to reference data," *Journal of the Royal Statistical Society, Series A*, **151**, 385–418.

de Leeuw, J. (1994), "Block-relaxation algorithms in statistics," in *Information Systems and Data Analysis*, eds. H. H. Bock et al., Berlin: Springer-Verlag, pp. 308–325.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society, Series B*, **39**, 1–38.

Efron, B. (1991), "Regression percentiles using asymmetric squared error loss," *Statistica Sinica*, **1**, 93–125.

Eide, E., and Showalter, M. H. (1998), "The effect of school quality on student performance: A quantile regression approach," *Economics Letters*, **58**, 345–350.

El-Attar, R. A., Vidyasagar, M., and Dutta, S. R. K. (1979), "An algorithm for L_1-norm minimization with application to nonlinear L_1 approximation," *SIAM Journal of Numerical Analysis*, **16**, 70–86.

Erdoğan, H. and Fessler, J. A. (1999), "Monotonic algorithms for transmission tomography," *IEEE Transactions on Medical Imaging*, **18**, 801–814.

Ferguson, T. S. (1967), *Mathematical Statistics: A Decision Theoretic Approach*, New York: Academic Press.

He, X. (1997), "Quantile curves without crossing," *The American Statistician*, **51**, 186–192.

Heiser, W. J. (1995), "Convergent computation by iterative majorization," in *Recent Advances in Descriptive Multivariate Analysis*, ed. W. J. Krzanowski, New York: Oxford University Press, pp. 157–189.

Kennedy, W. J. Jr., and Gentle, J. E. (1980), *Statistical Computing*, New York: Marcel Dekker.

Koenker, R., and Bassett, G. (1978), "Regression quantiles," *Econometrica*, **46**, 33–50.

Koenker, R., and Park, B. J. (1996), "An interior point algorithm for nonlinear quantile regression," *Journal of Econometrics*, **71**, 265–283.

Lange, K., and Sinsheimer, J. (1993), "Normal/Independent Distributions and Their Applications in Robust Regression," *Journal of Computational and Statistical Graphics*, **2**, 175–198.

Lange, K. (1995), "A gradient algorithm locally equivalent to the EM algorithm," *Journal of the Royal Statistical Society, Series B*, **57**, 425–437.

Lange, K., Hunter, D. R., and Yang, I. (2000), "Optimization transfer using surrogate objective functions," with discussion, *Journal of Computational and Graphical Statistics*, **9**, 1–59.

Madsen, K., and Nielsen, H. B. (1990), "Finite algorithms for robust linear regression," *Bit*, **30**, 682–699.

McLachlan, G. J., and Krishnan, T. (1996), *The EM Algorithm and Extensions*, Wiley: New York.

Merle, G., and Späth, H. (1974), "Computational Experiences with discrete l_p approximation," *Computing*, **12**, 315–321.

Mosteller, F., and Tukey, J. W. (1977), *Data Analysis and Regression: A Second Course in Statistics*, Reading, MA: Addison-Wesley.

Ortega, J. M., and Rheinboldt, W. C. (1970), *Iterative Solution of Nonlinear Equations in Several Variables*, Orlando: Academic Press.

Portnoy, S., and Koenker, R. (1997), "The Gaussian hare and the Laplacian tortoise: computability of squared-error versus absolute-error estimators," *Statistical Science*, **12**, 279–300.

Press, W., Flannery, B., Teukolsky, S., and Vetterling, W. (1986), *Numerical Recipes*, Cambridge: Cambridge University Press.

Rousseeuw, P. J. and Leroy, A. M. (1987), *Robust Regression and Outlier Detection*, New York: John Wiley & Sons.

Schlossmacher, E. J. (1973), "An iterative technique for absolute deviations curve fitting," *Journal of the American Statistical Association*, **68**, 857–859.

Taylor, J. W., and Bunn, D. W. (1998), "Combining forecast quantiles using quantile regression: Investigating the derived weights, estimator bias and imposing constraints," *Journal of Applied Statistics*, **25**, 193–206.

Verboon, P. (1994), *A Robust Approach to Nonlinear Multivariate Analysis*, Leiden, the Netherlands: DSWO Press.