INEQUALITY CONSTRAINED QUANTILE REGRESSION

ROGER KOENKER AND PIN NG

Abstract. An algorithm for computing parametric linear quantile regression estimates subject to linear inequality constraints is described. The algorithm is a variant of the interior point algorithm described in Koenker and Portnoy (1997) for unconstrained quantile regression and is consequently quite efficient even for large problems, particularly when the inherent sparsity of the resulting linear algebra is exploited. Applications to qualitatively constrained nonparametric regression are described in the penultimate section. Implementations of the algorithm are available in MATLAB and R.

1. Introduction

An early application of median regression in economics can be found in Arrow and Hoffenberg (1959). Their objective was to estimate input-output coefficients in a regression setting, but it was obviously desirable to impose the restriction that the coefficients were all positive. This was a relatively simple task given linear programming technology of the day provided that the conventional squared error fitting criterion was replaced by the absolute error criterion.

There are many other applications in which linear inequality constraints play a vital role in quantile regression. We will briefly describe some applications to nonparametric quantile regression in Section 5. First, we will introduce the quantile regression computational problem and briefly describe some basic duality theory and its relevance. We will then describe a modification of the Frisch-Newton algorithm introduced in Koenker and Portnoy (1997) that accommodates linear inequality constraints. Some details of the implementation of the algorithm and an evaluation of performance is provided in Section 4.

2. Quantile Regression as a Linear Program

The quantile regression problem

$$\min_{b \in \mathbb{R}^p} \sum_{i=1}^{n} \rho_\tau (y_i - x_i^\top b) \tag{1}$$

Version: August 24, 2004. This research was partially supported by NSF grant SES-02-40781. The authors would like to express their appreciation to Xuming He and Steve Portnoy for numerous discussions about this work. The authors would also like to thank Richard Blundell for providing the U.K. FES data used in Section 6.

where $\rho_\tau(u) = u(\tau - I(u < 0))$ is easily seen to be a linear program.[1] Let e denote an n-vector of ones and rewrite (1) as

$$\min_{(u,v,b)} \{\tau e^\top u + (1-\tau) e^\top v | Xb + u - v = y, \quad (u^\top, v^\top, b^\top) \in \mathbb{R}^{2n}_+ \times \mathbb{R}^p\} \tag{2}$$

In this formulation we seek the minimum of a linear function of the $2n + p$ variables $(u^\top, v^\top, b^\top)$, subject to n linear equality constraints and $2n$ linear inequality constraints. It turns out to be convenient to reformulate this primal version of the problem in the following way

$$\max_d \{y^\top d | X^\top d = (1-\tau) X^\top e, \quad d \in [0,1]^n\} \tag{3}$$

Here, $[0,1]^n$ denotes the n-field Cartesian product of the unit interval, and d may be interpreted as a vector of Lagrange multipliers associated with the linear equality constraints of the primal problem

To understand the transition from the primal problem (2) to the dual problem (3) it is helpful to recall a somewhat more general version of the duality theory of linear programming. Using the conventional notation of linear programming [2] and following Berman (1973), consider the primal problem

$$\min_x \{c^\top x | Ax - b \in T, \quad x \in S\} \tag{4}$$

where the sets $T = \{y \in \mathbb{R}^n\}$ and $S = \{y \in \mathbb{R}^{2n} \times \mathbb{R}^p\}$ can be arbitrary closed convex cones. This canonical problem has dual

$$\max_y \{b^\top y | c - A^\top y \in S^*, \quad y \in T^*\} \tag{5}$$

where $S^* = \{y \in \mathbb{R}^{2n} \times \mathbb{R}^p | x^\top y \geq 0 \text{ if } x \in S\}$ is the dual of S and $T^* = \{y \in \mathbb{R}^n\}$. For our purposes it suffices to consider the following special case: $T = \{O_n\}$, $S = \{\mathbb{R}^{2n}_+ \times \mathbb{R}^p\}$, $S^* = \{\mathbb{R}^{2n}_+ \times O_p\}$, and $T^* = \{\mathbb{R}^n\}$.

In our primal problem (2), $\left(\tau e^\top, (1-\tau) e^\top, O_p^\top\right)^\top$ and $\left(u^\top, v^\top, b^\top\right)^\top$ correspond to, respectively, c and x in (4), and the relation $Ax - b \in T$ in (4) becomes

$$[I \vdots - I \vdots X] \begin{bmatrix} u \\ v \\ b \end{bmatrix} - y \in \{O_n\} \tag{6}$$

[1] This observation for median regression, i.e., $\tau = 1/2$, can be traced to Charnes, Cooper, and Ferguson (1955) and Wagner (1959)

[2] An inherent difficulty in describing numerical algorithms for statistical procedures is that we are faced with two well established, but mutually incompatible notational schemes; one arising in statistics, the other in numerical analysis. In Section 2 we will introduce the quantile regression problem in its familiar statistical garb and then make the connection to linear programming. For the serious business of describing the algorithm in detail in Section 3 we will revert to the well established notational conventions of numerical analysis.

while $c - A^\top y \in S^*$ becomes

$$\begin{pmatrix} \tau e \\ (1-\tau)e \\ O_p \end{pmatrix} - \begin{bmatrix} I \\ -I \\ X^\top \end{bmatrix} \lambda \in \{\mathbb{R}_+^{2n} \times O_p\}$$

where λ denotes the n-vector of dual variables (Lagrange multipliers) associated with the equality constraints of the primal problem in (1). The requirement $y \in T^*$ in (5) may be translated as $\lambda \in \mathbb{R}^n$ so the dual problem in (5) can be expressed more concisely using the quantile regression notations as,

$$\max\{y^\top \lambda | X^\top \lambda = 0, \quad \lambda \in [\tau - 1, \tau]^n\}.$$

But this is equivalent to (3) after the transformation of variables, $d = 1 - \tau + \lambda$.

2.1. **Inequality Constraints.** Now consider augmenting the constraints of the primal problem in (2) with the new constraints, $Rb \geq r$. This is easily accommodated into the $Ax - b \in T$ constraint, (6) becomes,

$$(7) \qquad \begin{bmatrix} I & -I & X \\ 0 & 0 & R \end{bmatrix} \begin{bmatrix} u \\ v \\ b \end{bmatrix} - \begin{bmatrix} y \\ r \end{bmatrix} \in \{O_n\} \times \mathbb{R}_+^m$$

where $m \leq p$ denotes the row dimension of R. Now we have the dual variables $\lambda = (\lambda_1^\top \ \lambda_2^\top)^\top$ where λ_1 is associated with equality constraints and λ_2 is associated with the inequality constraints. The dual constraint $c - A^\top y \in S^*$ becomes,

$$(8) \qquad \begin{pmatrix} \tau e \\ (1-\tau)e \\ O_p \end{pmatrix} - \begin{bmatrix} I & 0 \\ -I & 0 \\ X^\top & R^\top \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix} \in \{\mathbb{R}_+^{2n} \times O_p\}$$

so the dual problem is

$$\max_\lambda\{y^\top \lambda_1 + r^\top \lambda_2 | X^\top \lambda_1 + R^\top \lambda_2 = 0, \quad \lambda_1 \in [\tau - 1, \tau]^n, \lambda_2 \geq 0\}.$$

Again, transforming variables $d_1 = 1 - \tau + \lambda_1, d_2 = \lambda_2$ we have

$$(9) \qquad \max_d\{y^\top d_1 + r^\top d_2 | X^\top d_1 + R^\top d_2 = (1-\tau)X^\top e, \quad d_1 \in [0,1]^n, d_2 \geq 0\}.$$

It is this form of the inequality constrained problem for which we will describe a solution algorithm.

3. A Frisch-Newton Algorithm

Until the mid 1980's the method of choice for solving linear programs of the sort described above was the simplex method. However, the work of Karmarker (1984) brought to fruition the idea of interior point methods. Rather than travel along the outer edges of the constraint set looking at each vertex for the direction of steepest descent, one might burrow from the center toward the boundary. This paradigm shift in thinking about linear programming has had a profound impact throughout the optimization literature. An influential observation by Gill, Murray, Saunders, Tomlin, and Wright (1986) connected Karmarker's interior point methods to earlier log-barrier methods elaborated by Fiacco and McCormick (1968) and others.

A pioneering early advocate of log barrier methods was Ragnar Frisch. In a series of Oslo technical reports, Frisch discovered interior point methods 30 years *avant la lettre.* Frisch (1956) described it in the following vivid terms for a talk in Paris,

> My method is altogether different than simplex. In this method we work systematically from the interior of the admissible region and employ a logarithmic potential as a guide – a sort of radar – in order to avoid crossing the boundary.

Despite considerable numerical experience with these methods, Frisch was unable to establish convergence and the resolution of many practical apects of the implementation of the methods had to wait for the intensive research effort that occurred only in the late 1980's.

3.1. **The Log-Barrier Formulation.** The algorithm that we will describe in this section is a variant of the log-barrier algorithm described in Koenker and Portnoy (1997) for unconstrained quantile regression problems. Following prior usage, we will refer to this as a Frisch-Newton method since the log-barrier formulation of Frisch, by replacing the sharply demarcated boundary of the inequality constraints with an objective function that smoothly tends to infinity as one approaches the boundary, enables us to take Newton steps toward a boundary solution. The strategy used to adjust the barrier parameter is based on the well-established Mehrotra (1992) predictor-corrector approach.

We adhere in this section to the notational conventions of Lustig, Marsden, and Shanno (1994). We will consider the following pair of primal and dual problems:

$$\min_{(x_1,x_2)} \{c_1^\top x_1 + c_2^\top x_2 | A_1 x_1 + A_2 x_2 = b,\ 0 \le x_1 \le u,\ 0 \le x_2\}$$

$$\max_{(y,w)} \{b^\top y - u^\top w | A_1^\top y + z_1 - w = c_1,\ A_2^\top y + z_2 = c_2,\ (z_1, z_2, w) \ge 0\}.$$

Note that we have reversed the roles of primal and dual so our new primal problem corresponds to the dual problem we derived in the previous section and vice-versa. Note also that we have generalized the problem slightly to allow u to be an arbitrary vector of (positive) upper bounds.

The classical Karush-Kuhn-Tucker (KKT) conditions for an optimum are:

$$\begin{aligned} A_1^\top y + z_1 - w &= c_1 \\ A_2^\top y + z_2 &= c_2 \\ Ax &= b \\ x_1 + s &= u \\ XZe &= 0 \\ SWe &= 0 \end{aligned}$$

Here we employ the convention that upper case letters corresponding to vectors in lower case are diagonal matrices with the elements of the vector along the diagonal, so for example $X = \mathrm{diag}(x)$. An easy way to see the KKT conditions is to write the primal problem above as the Lagrangian expression,

$$L = c^\top x - y^\top(Ax - b) - w^\top(u - x_1 - s) - \mu(\sum \log x_{1i} + \sum \log x_{2i} + \sum \log s_i).$$

This log-barrier formulation replaces the inequality constraints with a barrier function that penalizes feasible solutions as they approach the boundary of the constraint set. The parameter μ controls the severity of this penalization, and the strategy will be to gradually reduce μ. As μ tends to zero we approach the solution on the boundary of the constraint set. Differentiating with respect to x_1 and x_2 yields

$$\begin{aligned} A_1^\top y - w + \mu X_1^{-1} e &= c_1 \\ A_2^\top y + \mu X_2^{-1} e &= c_2. \end{aligned}$$

Writing $z = \mu X^{-1} e$, we have the first two equations of the KKT system. Differentiating with respect to y and w yields the next two. For fixed μ we have, from the definitions of z,

$$XZe = \mu e$$

and, from differentiating with respect to s, we obtain,

$$w = \mu S^{-1} e.$$

Substituting $\mu = 0$ yields the last two KKT conditions. The motivation for setting $\mu = 0$ stems from the question: 'What is the best value for μ at each iteration toward the optimal solution on the boundary of the constraint set?' The affine-scaling step described in Section 3.2 suggests computing the primal-dual step with $\mu = 0$. When the step thus obtained is feasible, we take it and continue the iteration. If it takes us outside the feasible region determined by the inequality constraints, i.e., the barrier function dominates the Lagrangian, we compute a Mehrotra predictor-corrector step described in Section 3.3 to modify the affine-scaling step to bring us back into the interior of the feasible region. The iterations stop when the duality gap is smaller than a specified tolerance, which is the requirement for optimality implied by the complementary slackness.

3.2. **The Affine Scaling Step.** We are looking for a solution to the KKT equations, say $g(\xi) = 0$. Suppose we have an initial point $\xi_0 = (y_0, z_0, x_0, s_0, w_0)$, for which x_0 is feasible for our primal problem. Denote $g(\xi_0) = g_0$. We differentiate to get

$$g(\xi) \approx \nabla_\xi g(\xi_0) d\xi + g_0$$

and we choose a direction by setting this equal zero, i.e.,

$$d\xi = -[\nabla_\xi g(\xi_0)]^{-1} g_0.$$

This is just Newton's Method. The linear system we obtain looks like

$$\begin{bmatrix} A_1^\top & I & 0 & 0 & 0 & 0 & -I \\ A_2^\top & 0 & I & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & A_1 & A_2 & 0 & 0 \\ 0 & 0 & 0 & I & 0 & I & 0 \\ 0 & X_1 & 0 & Z_1 & 0 & 0 & 0 \\ 0 & 0 & X_2 & 0 & Z_2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & W & S \end{bmatrix} \begin{bmatrix} dy \\ dz_1 \\ dz_2 \\ dx_1 \\ dx_2 \\ ds \\ dw \end{bmatrix} = - \begin{bmatrix} A_1^\top y + z_1 - w - c_1 \\ A_2^\top y + z_2 - c_2 \\ Ax - b \\ x_1 + s - u \\ X_1 Z_1 e \\ X_2 Z_2 e \\ SWe \end{bmatrix} \equiv \begin{bmatrix} r_1 \\ r_2 \\ r_3 \\ r_4 \\ r_5 \\ r_6 \\ r_7 \end{bmatrix}$$

To solve, substitute out dw and dz to get

$$\begin{aligned} & Wds + Sdw = -SWe \\ & \qquad dw = -S^{-1}(SWe + Wds) = -We - S^{-1}Wds \\ & Xdz + Zdx = -XZe \\ & \qquad dz = -X^{-1}(XZe + Zdx) = -Ze - X^{-1}Zdx. \end{aligned}$$

This reduces the system to,

$$\begin{aligned} A_1^\top dy - X_1^{-1} Z_1 dx_1 + S^{-1}Wds &= r_1 - We + Z_1 e \\ A_2^\top dy - X_2^{-1} Z_2 dx_2 &= r_2 + Z_2 e \\ Adx &= r_3 \\ dx_1 + ds &= r_4, \end{aligned}$$

which we can assemble as,

$$\begin{bmatrix} A_1^\top & -X_1^{-1}Z_1 & 0 & S^{-1}W \\ A_2^\top & 0 & -X_2^{-1}Z_2 & 0 \\ 0 & A_1 & A_2 & 0 \\ 0 & I & 0 & I \end{bmatrix} \begin{bmatrix} dy \\ dx_1 \\ dx_2 \\ ds \end{bmatrix} = \begin{bmatrix} c_1 + w - z_1 - A_1^\top y - w + z_1 \\ c_2 - z_2 - A_2^\top y + z_2 \\ b - Ax \\ 0 \end{bmatrix}.$$

Note that we have assumed that we start with $u = x_1 + s$. Now, substitute out $ds = -dx_1$ to reduce further the system, and set $Q_1 = X_1^{-1}Z_1 + S^{-1}W$, and $Q_2 = X_2^{-1}Z_2$. We now have,

$$\begin{bmatrix} A_1^\top & -Q_1 & 0 \\ A_2^\top & 0 & -Q_2 \\ 0 & A_1 & A_2 \end{bmatrix} \begin{bmatrix} dy \\ dx_1 \\ dx_2 \end{bmatrix} = \begin{bmatrix} c_1 - A_1^\top y \\ c_2 - A_2^\top y \\ b - Ax \end{bmatrix} \equiv \begin{bmatrix} \tilde{r}_1 \\ \tilde{r}_2 \\ \tilde{r}_3 \end{bmatrix}.$$

Now solve for dx_1 and dx_2 in terms of dy in the first two equations:

$$\begin{aligned} A_1^\top dy - Q_1 dx_1 = \tilde{r}_1 \quad &\Rightarrow \quad dx_1 = Q_1^{-1}(A_1^\top dy - \tilde{r}_1) \\ A_2^\top dy - Q_2 dx_2 = \tilde{r}_2 \quad &\Rightarrow \quad dx_2 = Q_2^{-1}(A_2^\top dy - \tilde{r}_2) \end{aligned}$$

Substitution gives us one equation[3] in dy

$$(A_1 Q_1^{-1} A_1^\top + A_2 Q_2^{-1} A_2^\top) dy = \tilde{r}_3 + A_1 Q_1^{-1} \tilde{r}_1 + A_2 Q_2^{-1} \tilde{r}_2$$

We can now write the solution to the linear system for the "affine step" as:

$$\begin{aligned} dy &= (AQ^{-1}A^\top)^{-1}[\tilde{r}_3 + A_1 Q_1^{-1} \tilde{r}_1 + A_2 Q_2^{-1} \tilde{r}_2] \\ dx_1 &= Q_1^{-1}(A_1^\top dy - \tilde{r}_1) \\ dx_2 &= Q_2^{-1}(A_2^\top dy - \tilde{r}_2) \\ ds &= -dx_1 \\ dz &= -z - X^{-1} Z dx = -Z\left(e + X^{-1} dx\right) \\ dw &= -w - S^{-1} W ds = -W\left(e + S^{-1} ds\right). \end{aligned}$$

The real effort at each iteration involves the Cholesky decomposition of the matrix $AQ^{-1}A^\top$ where Q is the diagonal matrix with Q_1 and Q_2 on the diagonal. And in this respect the algorithm is essentially the same as the case without the inequality constraints.

3.3. **The Mehrotra Predictor – Corrector Step.** The tentative affine scaling step length is given by,

$$\phi_p = \min\{1, \sigma \min_{i:\, dx_i < 0}\{x_i/dx_i\}, \sigma \min_{i:\, ds_i < 0}\{s_i/ds_i\}\} \tag{10}$$

$$\phi_d = \min\{1, \sigma \min_{i:\, dz_i < 0}\{z_i/dz_i\}, \sigma \min_{i:\, dw_i < 0}\{w_i/dw_i\}\} \tag{11}$$

where the scaling factor σ determines how close the step is allowed to come to the boundary of the constraint set. In accordance with Lustig, Marsden and Shanno (1992, 1994) we take $\sigma = .99995$.

When the full affine scaling step is infeasible, that is when $\min\{\phi_p, \phi_d\} < 1$, we attempt to modify the length and direction of the step. For fixed $\mu > 0$, the first

[3] Note that in prior implementations, Koenker and Portnoy (1997), we assumed that we had initial primal-dual feasibility, i.e., that the equality constraints were all satisfied and the starting value satisfied $0 \leq x \leq u$. This was easy since $x_0 = (1-\tau)e$ was a natural initial point. Here we have not assumed initial primal feasibility so the right hand side becomes a bit more complicated. But *crucially* we do not need to have $Ax = b$ at the start; this simplifies life in the inequality constrained case considerably.

order conditions corresponding to our Lagrangian expression are,

$$
\begin{aligned}
A_1^\top y + z_1 - w &= c_1 \\
A_2^\top y + z_2 &= c_2 \\
Ax &= b \\
x_1 + s &= u \\
XZe &= \mu e \\
SWe &= \mu e
\end{aligned}
\tag{12}
$$

Substituting $x \to x + dx, y \to y + dy$ etc., we obtain, assuming $x_1 + s = u$

$$
\begin{aligned}
A_1^\top dy + dz - dw &= c_1 - A_1^\top y - z_1 + w \\
A_2^\top dy + dz_2 &= c_2 - A_2^\top y - z_2 \\
Adx &= b - Ax \\
dx_1 + ds &= 0 \\
Xdz + Zdx &= \mu e - XZe - dXdZe \\
Sdw + Wds &= \mu e - SWe - dSdWe.
\end{aligned}
$$

Note that our provisional affine scaling step has been computed by solving almost the same system except that the bilinear terms $dXdZe$ and $dSdWe$ were ignored and μ was set to zero. The Mehrotra predictor-corrector step brings both of these aspects of the problem back into play by solving:

$$
\begin{aligned}
A_1^\top \delta y + \delta z_1 - \delta w &= 0 \\
A_2^\top \delta y + \delta z_2 &= 0 \\
A\delta x &= 0 \\
\delta x_1 + \delta s &= 0 \\
X\delta z + Z\delta x &= \mu e - dXdZe \\
S\delta w + W\delta s &= \mu e - dSdWe
\end{aligned}
$$

Solving, we proceed as before,

$$
\begin{aligned}
W\delta s + S\delta w &= \mu e - dSdWe \\
\delta w &= -S^{-1}W\delta s + S^{-1}(\mu e - dSdWe) \\
X\delta z + Z\delta x &= \mu e - dXdZe \\
\delta z &= -X^{-1}Z\delta x + X^{-1}(\mu e - dXdZe)
\end{aligned}
$$

substituting, and eliminating δs as before, we have

$$\begin{aligned} A_1^\top \delta y - Q_1 \delta x_1 &= S^{-1}(\mu e - dSdWe) - X_1^{-1}(\mu e - dX_1 dZ_1 e) \\ A_2^\top \delta y - Q_2 \delta x_2 &= -X_2^{-1}(\mu e - dX_2 dZ_2 e) \\ A\delta x &= 0, \end{aligned}$$

rewritten in matrix form,

$$\begin{bmatrix} A_1^\top & -Q_1 & 0 \\ A_2^\top & 0 & -Q_2 \\ 0 & A_1 & A_2 \end{bmatrix} \begin{bmatrix} \delta y \\ \delta x_1 \\ \delta x_2 \end{bmatrix} = \begin{bmatrix} \hat{r}_1 \\ \hat{r}_2 \\ 0 \end{bmatrix}$$

where

$$\begin{aligned} \hat{r}_1 &= S^{-1}(\mu e - dSdWe) - X_1^{-1}(\mu e - dX_1 dZ_1 e) \\ &= \mu(S^{-1} - X_1^{-1})e + X_1^{-1} dX_1 dZ_1 e - S^{-1} dSdWe \\ \hat{r}_2 &= -X_2^{-1}(\mu e - dX_2 dZ_2 e) \end{aligned}$$

Now solve again for δx in terms of δy, and substituting we have

$$\begin{aligned} \delta y &= (AQ^{-1}A^\top)[A_1 Q_1^{-1} \hat{r}_1 + A_2 Q_2^{-1} \hat{r}_2] \\ \delta x_1 &= Q_1^{-1}(A_1^\top \delta y - \hat{r}_1) \\ \delta x_2 &= Q_2^{-1}(A_2^\top \delta y - \hat{r}_2) \\ \delta s &= -\delta x_1 \\ \delta z &= -X^{-1} Z \delta x + X^{-1}(\mu e - dXdZe) \\ \delta w &= -S^{-1} W \delta s + S^{-1}(\mu e - dSdWe) \end{aligned}$$

We can interpret the solution of this system for the vector $(\delta y, \delta z, \delta x, \delta s, \delta w)$ as simply taking another Newton step, this time starting from the proposed affine scaling point. Since the left hand sides of the two linear systems are exactly the same, only the right hand side has been altered, a solution to the new system can be found by backsolving the triangular system using the Cholesky factorization of the affine step.

The crucial remaining question is: how does μ get updated? The duality gap is given by the expression,

$$\gamma = x^\top z + s^\top w.$$

Complementary slackness requires that the duality gap vanish at an optimum, so γ provides a direct measure of progress toward the solution. Iterations stop when γ is reduced below a prespecified tolerance. Solving the last two equations of the system (12) for μ we obtain,

$$\mu = \gamma/(2n_1 + n_2)$$

where n_1 is the dimension of the vectors: x_1, z_1, s, w, and n_2 is the dimension of the vectors: x_2, z_2. Were we to take the affine scaling step the duality gap would be

$$\hat{\gamma} = (x + \phi_p dx)^\top (z + \phi_d dz) + (s + \phi_p ds)^\top (w + \phi_d dw)$$

If $\hat{\gamma} << \gamma$ the step has made considerable progress toward the solution and it is reasonable to reduce μ considerably. On the other hand, if the duality gap is only slightly reduced, we should consider the affine step to be poor and conclude that μ should not be substantially reduced. Note that repeated Newton steps with a fixed value of μ bring the iterations toward a point on the "central path", that is a point that minimizes the Lagrangean for a fixed μ.

Good performance of any interior point algorithm must balance the objectives of trying to stay close to the central path while trying to rapidly reduce the barrier parameter μ and thus moving toward the boundary. See Gonzago (1992) for a detailed analysis. The heuristics described above are embodied in the updating rule,

$$\mu_{(k+1)} \rightarrow (\hat{\gamma}/\gamma)^3 \gamma/(2n_1 + n_2)$$

Substituting this new value of μ into the system and solving we obtain the modified step. The step length is again determined by the rules (10) and (11). The step is taken, and the iterations continue until the duality gap is reduced to satisfy the specified tolerance.

4. Implementation

The algorithm described above has been implemeted in four distinct versions.[4] Versions written in "R" and "Matlab" provide accessible and convenient tools for studying qualitative features of performance since they are written in higher-level, matrix-oriented languages. R, Ihaka and Gentleman (1996), is a open source dialect of the statistical language S developed by Chambers (1998). We have used the pure "R" version primarily as a debugging tool, but the Matlab version is reasonably efficient for problems of moderate size.[5]

Two distinct versions of the algorithm have also been written in Fortran and linked to "R". One employs standard (dense) linear algebra routines from LAPACK, the other uses more specialized sparse linear algebra to improve performance for problems having a high proportion of zeros in the matrix A. The latter formulation is particularly well-suited to the non-parametric regression problems we describe in the next section. We discuss in more detail the sparse matrix implementation aspects in Koenker and Ng (2004).

[4] Code is available at: `http://www.econ.uiuc.edu∼/roger/rq/rq.html` for all four versions.

[5] Careful examination of the Matlab code reveals that some additional efficiency gain would be possible by reusing the Cholesky factorization in the computation of the modified step, as described in the preceding section.

5. Quantile Smoothing Splines

In Koenker, Ng, and Portnoy (1994) we proposed a variant of the classical cubic smoothing spline solving

$$\min_{g\in\mathcal{G}_2}\sum(y_i-g(x_i))^2+\lambda\int(g''(x))^2dx.$$

This quantile smoothing spline was constructed by solving,

$$\min_{g\in\mathcal{G}_1}\sum\rho_\tau(y_i-g(x_i))+\lambda\bigvee(g'). \tag{13}$$

Here $\bigvee(f)$ denotes the total variation of the function f. Recall, e.g., Natanson (1955) for absolutely continuous $f:\Re\to\Re$,

$$\bigvee(f)=\int|f'(x)|dx.$$

Thus, for sufficiently smooth g, we can interpret the roughness penalty in (13) as $\mathcal{L}_1$ norm of the *second* derivative,

$$\bigvee(g')=\int|g''(x)|dx.$$

However, solutions of the variational problem (13) turn out to take the form of piecewise linear functions, so the total variation interpretation of the penalty is preferable.[6]

The problem (13) has a simple linear programming formulation. Writing

$$g(x)=\alpha_i+\beta_i(x-x_i)\quad\text{ for } x\in[x_i,x_{i+1})$$

for the ordered, distinct values $x_1,\dots,x_n$, we have by the continuity of g, that

$$\beta_i=(\alpha_{i+1}-\alpha_i)/h_i\quad i=1,2,\dots,n-1$$

where $h_i=x_{i+1}-x_i$. So the penalty becomes

$$\bigvee(g')=\sum_{i=1}^{n-2}|\beta_{i+1}-\beta_i|=\sum_{i=1}^{n-2}|(\alpha_{i+2}-\alpha_{i+1})/h_{i+1}\quad-(\alpha_{i+1}-\alpha_i)/h_i|,$$

and the original problem may be written as,

$$\min_{\alpha\in\Re^n}\sum_{i=1}^{n}\rho_\tau(y_i-\alpha_i)+\lambda\sum_{j=1}^{n-2}|d_j^\top\alpha|$$

[6] The $\mathcal{L}_1$ interpretation can be extended to the piecewise linear case, but we then need to interpret the integral as a limiting form in the sense of (Schwartz) distributions. See Koenker and Mizera (2002) for details.

where $d_j^\top = (0, \ldots, 0, h_j^{-1}, (h_{j+1}^{-1} - h_j^{-1}), h_{j+1}^{-1}, 0, \ldots, 0)$. In the important median special case, $\tau = 1/2$, we can view this as simply a data-augmented ℓ_1 regression. We have the pseudo-design matrix,

$$X = \begin{bmatrix} I_n \\ D \end{bmatrix}$$

where $d_j^\top$ is the j^{th} row of D, and the pseudo response is $y^\top = (y_1, \ldots, y_n, 0, \ldots, 0) \in \Re^{2n-2}$. In the case that $\tau \neq 1/2$ the situation is almost the same, except that in the dual formulation of the problem we have equality constraints whose right hand side is $(1-\tau)e_n + 1/2D^\top e_{n-2}$ rather than $1/2e_n + 1/2D^\top e_{n-2}$.

The parameter λ in (13) controls the smoothness of the fitted function $\hat{g}$. The parametric dimension of $\hat{g}$ can be associated with the number of points interpolated exactly by $\hat{g}$, i.e., $\#\{i : y_i = \hat{g}(x_i)\}$. Koenker, Ng, and Portnoy (1994) discuss using this quantity in a Schwartz-type model selection criterion. They also suggest that further qualitative constraints on the fitted function such as monotonicity or convexity could be imposed by adding linear inequality constraints. This approach was implemented in He and Ng (1999b) using the Bartels and Conn (1980) projected gradient/Simplex algorithm.

5.1. **Monotonicity.** There is a vast literature on estimating non-parametric regression relationships subject to monotonicity constraints. The classical reference is Barlow, Bartholomew, Bremner, and Brunk (1972), recent developments are treated in Robertson, Wright, and Dykstra (1988). See also the survey of Matzkin (1994). Much of the early work focused on minimizing a squared error objective subject to a monotonicity constraint, but more recently there has been interest in adding a smoothing objective as well. Mammen (1991), for example, considers kernel smoothing followed by a pool-adjacent-violators step as well as a procedure that reverses the order of these operations. More in line with the approach suggested here is the work of Utreras (1985), Villalobos and Wahba (1987), Ramsay (1988), Mammen, Marron, Turlach, and Wand (2001), and most closely He and Ng (1999a) who all explore smoothing spline methods subject to linear inequality constraints as a way to impose monotonicity

Adding a monotonicity constraint to the quantile smoothing spline problem is quite straightforward given the algorithm described in the earlier sections. The function g is monotone increasing if

$$\beta_i = (\alpha_{i+1} - \alpha_i)/h_i \geq 0 \quad i = 1, 2, \ldots, n-1$$

so our constraint $Rb \geq r$ becomes,

$$\begin{bmatrix} -1 & 1 & 0 & \ldots & \ldots & 0 \\ 0 & -1 & 1 & \ldots & \ldots & 0 \\ \vdots & & & & & \\ 0 & \ldots & \ldots & \ldots & -1 & 1 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \vdots \\ \vdots \\ \alpha_n \end{bmatrix} \geq 0.$$

To illustrate the method, in Figure 1 we plot observations from the model

$$y_i = x_i + u_i$$

where the x_i are equally spaced on $[0, 5]$, and the u_i are Student t on 2 degrees of freedom. There are 100 observations. The pool adjacent violators curve appears in gray, and the monotone median smoothing spline with $\lambda = 0.1$ appears in black. Note that the outliers in the response tend to produce extended flat segments in the PAV fit. There are several advantages of the spline: it has a "knob" to control the smoothness of the fitted function, it has an inherent robustness that PAV fitting based on Gaussian fidelity does not, and there is also a knob to control the desired conditional quantile of the fit.

5.2. **Convexity.** There is also an extensive literature on estimating functions constrained to be convex, or concave. Such conditions are also easy to impose. Convexity is equivalent in our setting of linear splines to the monotonicity of the slope parameters, β_i, i.e. to the conditions,

$$\beta_{i+1} - \beta_i \geq 0 \quad i = 1, 2, ..., n-1.$$

So to impose convexity we need simply to add the constraint,

$$D\alpha \geq 0,$$

where D is the matrix defining the total variation roughness penalty introduced above. For concavity, D is replaced by $-D$. In Figure 2 we illustrate a simple application to fitting the quadratic model,

$$y_i = x_i + x_i^2 + u_i$$

where the u_i are iid $\mathcal{N}(0, 4)$. The plot illustrates two median smoothing spline fits, both with λ chosen to be 0.08, one with the convexity constraint, the other without the constraint. Clearly the convexity constraint acts as a powerful additional smoothing effect.

5.3. $\mathcal{L}_\infty$ **Roughness Penalty.** In earlier work, Koenker and Ng (1992), we have suggested that $\mathcal{L}_\infty$ penalties on the roughness of the fitted function,

$$\sup_x |g''(x)|$$

might serve as a useful alternative to the total variation penalty for some applications. For linear splines we may interpret this as,

$$\sup_i |\beta_{i+1} - \beta_i| \leq \nu.$$

Rather than assigning a Lagrange multiplier to determine the relative weight received by the roughness penalty we may choose a value for ν and vary the smoothness of the fitted function by adjusting the ν knob. It is well known that there is a one to one correspondance between the solutions determined by the Lagrange multiplier formulation and those indexed by the constraint parameter ν. Again the matrix D

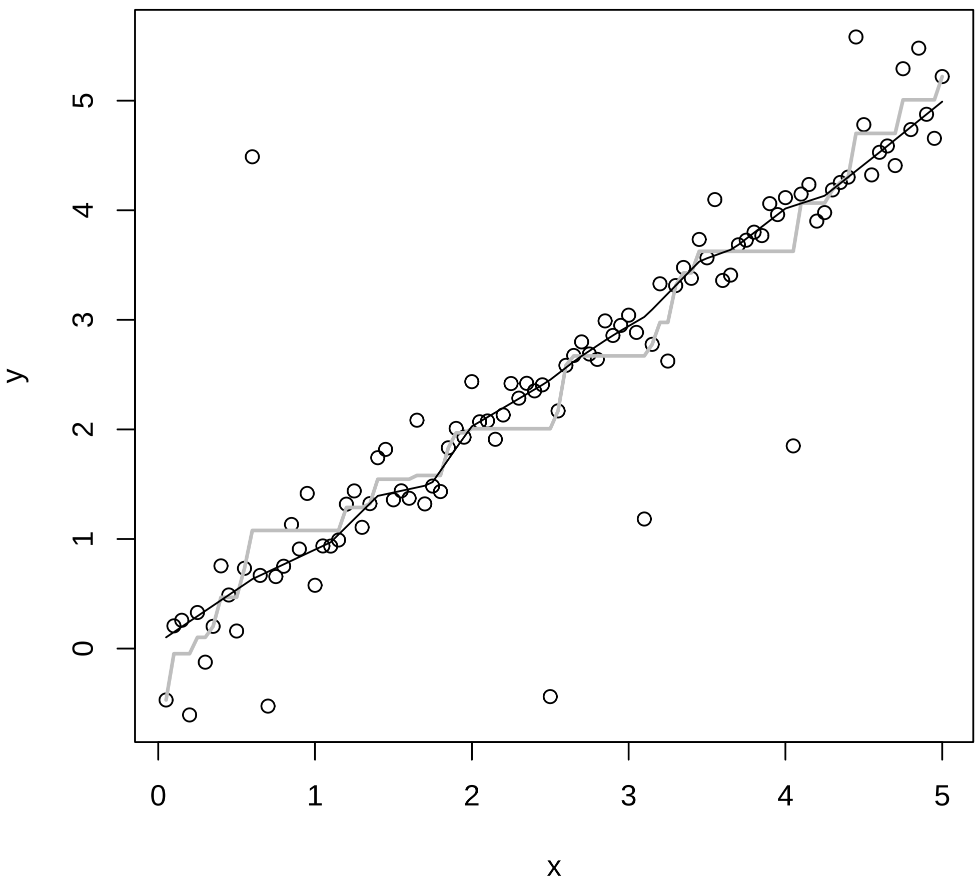

FIGURE 5.1. A comparison of the pool adjacent violators estimate (in gray) and a monotone median smoothing spline fit (in black). Note the sensitivity of the PAV solution to outliers in the response y, and the need for further smoothing.

plays a crucial role, and we may express the constraints as restricting both $D\alpha$ and $-D\alpha$ to exceed $-\nu$ times a vector of ones.

5.4. **Boundary Constraints.** In many smoothing problems there are natural constraints on the function being estimated near the boundary of the region of support. These might entail inequality constraints on the function itself, or on its derivitives,

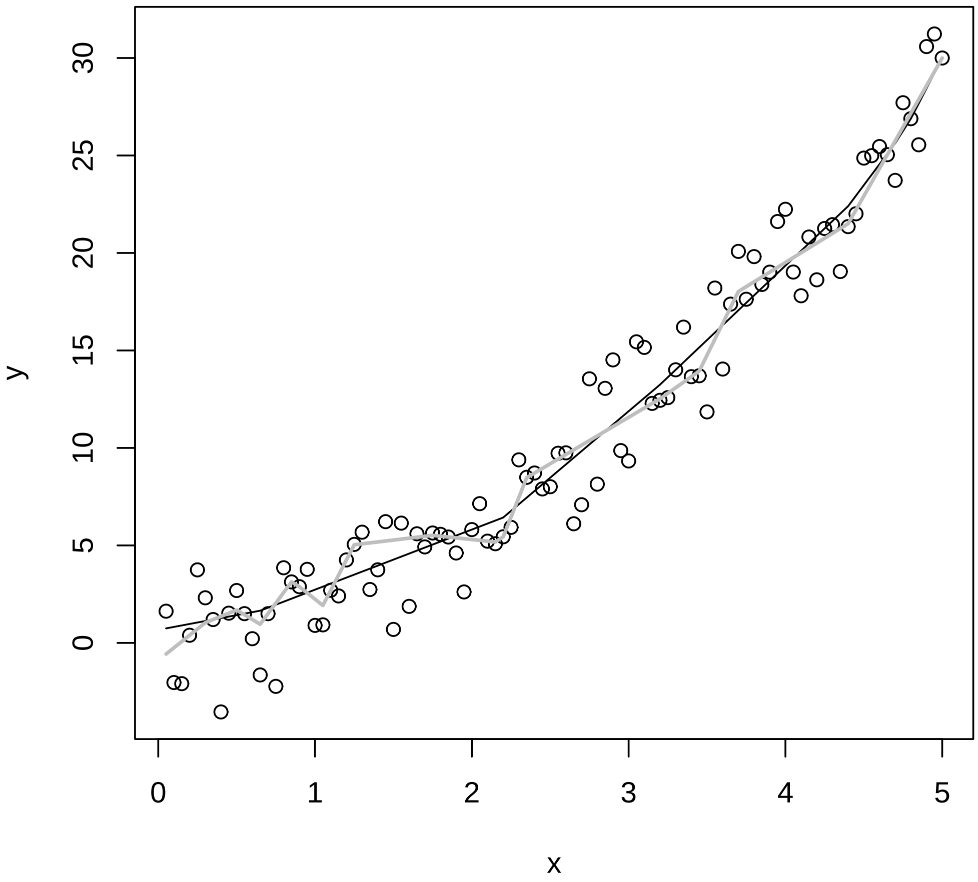

Figure 5.2. A comparison of the convex constrained median smoothing spline (in black) and the unconstrained median smoothing spline (in gray). Both of the fitted curves use the same smoothing parameter $\lambda = 0.08$.

or even equality constraints. Note that equality constraints, say $R\beta = r$ can easily be imposed by requiring both $R\beta \geq r$ and $R\beta \leq r$.

6. An Engel Curve Example

To illustrate the approach of the preceeding section we consider an application to the estimation of Engel curves. The data is taken from the U.K. Family Expenditure Survey for 1995. There are 3296 observations. The observations on household income

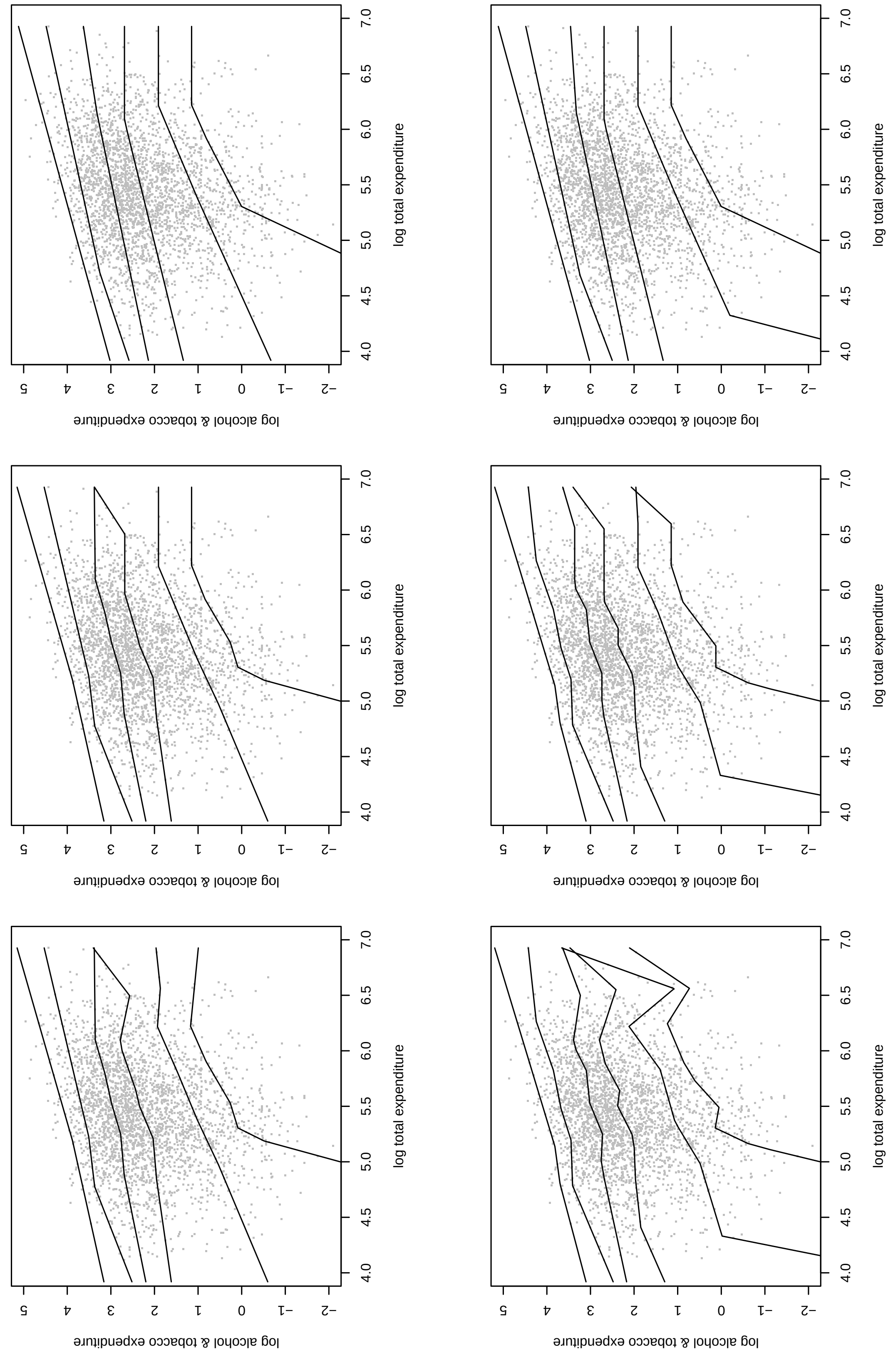

FIGURE 5.3. Families of Estimated Quantile Engel Curves.

were rounded to 4 significant digits; this yielded 1537 distinct values representing the parametric dimension of the the model prior to smoothing. Expenditure on alcohol and tobacco is modeled solely as a function of household total expenditure with both variables taken in natural logarithms, so the slope of the estimated curves can be taken as an estimate of the Engel elasticity. We estimate six distinct conditional quantile functions for $\tau \in \{.15, .25, .50, .75, .95, .99\}$. Quantiles below .15 can not be estimated since the proportion of reported zeros exceeds this level at lower household expenditure levels. Note that the treatment of the zero expenditure households when taking logarithms is not an issue for the estimation of these upper quantiles. Such observations can be notionally coded at $-\infty$. Only the the sign of their residuals influences the estimate of the model, so the actual coding only has to insure the they lie below the fitted function. In Figure 3 we depict six families of estimated conditional quantile Engel curves. All of the estimates are based on the total variation penalty method (13). We illustrate results for two different values of the smoothing parameter λ: the upper panels use $\lambda = 1.0$, while the lower panels use $\lambda = 0.5$. The left panels of the figure depict the unconstrained estimates, the middle panels illustrates the estimates constrained to be monotone increasing, and the right panels were constrained to be increasing and concave. It is apparent that the qualitative constraints are effective in imposing some additional discipline on the fitting and this is even more clear as one explores fitting with smaller values of the smoothing parameter, λ.

Here the fitting was carried out using the sparse versions of the algorithms described above in R, as a consequence it would be straightforward to add further complexity to the model in the form of parametric or nonparametric components. See the documentation for the function `rqss` in the `quantreg` package for R, Koenker (1991–), and the `SparseM` package for sparse linear algebra, Koenker and Ng (2003). The use of sparse algebra is quite essential since without it the size of underlying regression problems with column dimension 1537 and up to about 10,000 rows would be prohibitive on many machines; with it, required cpu time for fitting is about half a second on a Sun Ultra 2 for each quantile.

7. Prospects and Conclusions

Inequality constraints are relatively easy to impose in the context of quantile regression estimation and provide a flexible means of imposing qualitative restrictions in non-parametric quantile regression problems. Interior point methods based on Frisch's log-barrier approach offer an extremely efficient approach to the computation of such estimators. And sparse linear algebra leads to significant further gains in efficiency of computation.

There are several important open problems associated with inequality constrained quantile regression. There is an extensive literature on inference in the classical Gaussian regression setting subject to inequality restrictions, and there is also an extensive literature on tests of qualitative features in nonparametric mean regression. It would be useful to extend this inference apparatus to the present context.

References

Arrow, K., and M. Hoffenberg (1959): *A Time Series Analysis of Interindustry Demands.* North-Holland, Amsterdam.

Barlow, R. E., D. J. Bartholomew, J. M. Bremner, and H. D. Brunk (1972): *Statistical Inference Under Order Restrictions.* Wiley, New York.

Bartels, R., and A. Conn (1980): "Linearly constrained discrete ℓ_1 problems," *Transactions of the ACM on Mathematical Software*, 6, 594–608.

Berman, A. (1973): *Cones, Matrices and Mathematical Programming.* Springer-Verlag, Berlin.

Chambers, J. M. (1998): *Programming With Data: a Guide to the S Language.* Springer-Verlag.

Charnes, A., W. Cooper, and R. Ferguson (1955): "Optimal estimation of executive compensation by linear programming," *Management Science*, 1, 138–151.

Fiacco, A., and G. McCormick (1968): *Nonlinear Programming: Sequential Unconstrained Minimization Techniques.* Wiley: New York.

Frisch, R. (1956): "La Résolution des problèmes de programme lineaire par la mèthode du potential logarithmique," *Cahiers du Seminaire d'Econometrie*, 4, 7–20.

Gill, P., W. Murray, M. Saunders, T. Tomlin, and M. Wright (1986): "On projected Newton barrier methods for linear programming and an equivalence to Karmarker's projective method," *Mathematical Programming*, 36, 183–209.

Gonzago, C. (1992): "Path-following methods for linear programming," *SIAM Review*, 34, 167–224.

He, X., and P. Ng (1999a): "COBS: Qualitatively Constrained Smoothing Via Linear Programming," *Computational Statistics*, 14, 315–337.

He, X., and P. Ng (1999b): "Quantile splines with several covariates," *J. Statist. Planning Inference*, 75, 343–352.

Ihaka, R., and R. Gentleman (1996): "R: A Language for Data Analysis and Graphics," *Journal of Computational and Graphical Statistics*, 5(3), 299–314.

Karmarker, N. (1984): "A new polynomial time algorithm for linear programming," *Combinatorica*, 4, 373–395.

Koenker, R. (1991–): "Quantreg: A Quantile Regression Package for R," `http://cran.r-project.org`.

Koenker, R., and I. Mizera (2002): "Penalized triograms: total variation regularization for bivariate smoothing," Preprint.

Koenker, R., and P. Ng (1992): "Quantile Smoothing Splines," in *Nonparametric Statistics and Related Topics*, pp. 205–215. Elsevier/North-Holland (New York; Amsterdam).

Koenker, R., and P. Ng (2003): "SparseM: Sparse Linear Algebra for R," *Journal of Statistical Software*, 8.

——— (2004): "A Frisch-Newton Algorithm for Sparse Quantile Regression," preprint.

Koenker, R., P. Ng, and S. Portnoy (1994): "Quantile smoothing splines," *Biometrika*, 81, 673–680.

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

Lustig, I., R. Marsden, and D. Shanno (1992): "On implementing Mehrotra's predictor-corrector interior-point method for linear programming," *SIAM Journal of Optimization*, 2.

——— (1994): "Interior point methods for linear programming: computational state of the art with discussion," *ORSA J. on Computing*, 6, 1–36.

Mammen, E. (1991): "Nonparametric Regression Under Qualitative Smoothness Assumptions," *The Annals of Statistics*, 19, 741–759.

Mammen, E., J. S. Marron, B. A. Turlach, and M. P. Wand (2001): "A General Projection Framework for Constrained Smoothing," *Statistical Science*, 16(3), 232–248.

Matzkin, R. (1994): "Restrictions of economic theory in nonparametric methods," in *Handbook of Econometrics*, ed. by R. Engle, and D. McFadden. Elsevier, Amsterdam.

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

Natanson, I. P. (1955): *Theory of functions.* Ungar, New York.

Ramsay, J. O. (1988): "Monotone Regression Splines in Action," *Statistical Science*, 3, 425–441.

Robertson, T., F. T. Wright, and R. L. Dykstra (1988): *Order Restricted Statistical Inference.* Wiley.

Utreras, F. I. (1985): "Smoothing Noisy Data Under Monotonicity Constraints: Existence, Characterization and Convergence Rates," *Numerische Mathematik*, 47, 611–625.

Villalobos, M., and G. Wahba (1987): "Inequality-constrained Multivariate Smoothing Splines With Application to the Estimation of Posterior Probabilities," *Journal of the American Statistical Association*, 82, 239–248.

Wagner, H. (1959): "Linear programming techniques for regression analysis," *Journal of American Statistical Association*, 54, 206–212.

University of Illinois at Urbana-Champaign

University of Northern Arizona