THE GAUSSIAN HARE AND THE LAPLACIAN TORTOISE:
COMPUTABILITY OF SQUARED-ERROR VS. ABSOLUTE-ERROR ESTIMATORS

STEPHEN PORTNOY AND ROGER KOENKER

ABSTRACT. Since Gauss(1821), it has been generally accepted that ℓ_2 methods of combining observations by minimizing sums of squared errors have significant computational advantages over earlier ℓ_1 methods based on minimization of absolute errors advocated by Boscovich, Laplace and others. However, ℓ_1 methods are known to have significant robustness advantages over ℓ_2 methods in many applications, and related quantile regression methods provide a useful, complementary approach to classical least-squares estimation of statistical models. Combining recent advances in interior point methods for solving linear programs with a new statistical preprocessing approach for ℓ_1-type problems, we obtain a 10 to 100 fold improvement in computational speeds over current (simplex-based) ℓ_1 algorithms in large problems, demonstrating that ℓ_1 methods can be made competitive with ℓ_2 methods in terms of computational speed throughout the entire range of problem sizes. Formal complexity results suggest that ℓ_1 regression can be made faster than least squares regression for n sufficiently large, and p modest.

1. INTRODUCTION

Although ℓ_1 methods of estimation, which minimize sums of *absolute* residuals, have a long history in statistical applications, there is still some reluctance to adopt them for the analysis of large datasets because they are regarded as computationally highly demanding. In particular, the simplex algorithm of linear programming that is the mainstay of modern ℓ_1 computation has acquired a reputation as unwieldy in large problems. This reputation may be partially attributed to theoretical results on worst-case performance of the simplex algorithm, which establish that for certain pathological problems the number of simplex iterations required for a solution can increase exponentially with problem size over the range of problem dimensions typically encountered in statistical practice. However, *in practice* the simplex algorithm performs extremely well for problems of moderate size. Up to a few hundred observations ℓ_1-regression via the simplex algorithm is actually faster, for example, than

Version: May 30, 1997. This research has been partially supported by NSF Grant SBR 93-20555. Most of the computing carried out in the course of this research was conducted on the Sparc 20 `ragnar.econ.uiuc.edu` in the Econometrics Lab at the University of Illinois and primarily supported by NSF Grant SBR 95-12440. The authors would like to thank Kevin Hallock for providing the census data used in Section 7, and Marc Meketon, Mike Osborne, Gib Bassett and two referees for useful comments in the course of this research. We are, of course, solely responsible for any errors.

conventional ℓ_2 regression in the standard implementations provided by Splus. However, for problems exceeding a few thousand observations current implementations of simplex begin to live up to their slothful theoretical reputation.

Nevertheless, interest in the application of ℓ_1 estimation methods, and quantile regression more generally, in large-scale data analysis has grown steadily in recent years. Applications of quantile regression, Koenker and Bassett(1978) and Powell(1986), in economics to problems with sample sizes in the range 10,000 - 100,000 are now almost routine. See for example Buchinsky(1994,1995), and Manning, *et al* (1996). Interest in bootstrapped inference in such applications makes the need for efficient computational methods acute. Nonparametric quantile regression using local polynomials, Welsh (1996) and Fan and Gijbels (1996); or splines Koenker, Ng and Portnoy(1996), Green and Silverman(1994), has also stimulated the demand for more efficient ℓ_1 computation. Chen and Donoho (1995) and Tibshirani (1996) have recently proposed application of ℓ_1 penalties as model selection devices for a broad range of applications including image processing. Finally, there has been considerable interest in multivariate analysis in problems like the Oja (1983) median, which can be formulated as ℓ_1 regression problems with pseudo observations constructed as U-statistics from an initial sample. Taken together, these developments strongly motivate the search for improved methods of computing ℓ_1-type estimators when n is large.

Following Karmarkar (1984), there has also been intense interest among numerical analysts in alternative "interior point" methods for solving linear programs. Rather than moving from vertex to vertex around the outer surface of the constraint set as dictated by simplex, the interior point approach solves a sequence of quadratic problems in which the relevant interior of the constraint set is approximated by an ellipsoid. This approach, as shown by Karmarkar and subsequent authors, provides demonstrably better worst-case performance than the simplex algorithm and has also demonstrated impressive practical performance on a broad range of large-scale linear programs arising in commerce as well as in extensive numerical trials.

After a brief historical introduction to ℓ_1 computation, we compare recent interior point methods to existing simplex-based methods. We find, quite in accordance with recent literature on more general LPs, that the interior point approach is competitive with simplex in moderate sized problems, say n up to 1000, and exhibits a rapidly increasing advantage over simplex for larger problems. We then propose a new form of statistical preprocessing for general quantile regression problems that also has the effect of dramatically reducing the computation burden. This preprocessing step is somewhat reminiscent of the subsampling approach in earlier $O(n)$ univariate quantile algorithms like that of Floyd and Rivest(1975). Taken together, the preprocessing step and a careful choice of interior point vs. simplex yields an algorithm that is 10 to 100 times faster than current simplex methods for a variety of test problems with sample sizes in the range 10,000 - 200,000. In practice, as we will see, combining the preprocessing step and interior point methods yields ℓ_1 computations which rival the speeds achievable with current ℓ_2 methods over the entire range of problem dimensions. In theory, the results of Section 6 imply that ℓ_1 computations can be made strictly faster than their ℓ_2 counterparts for problems with n and p sufficiently large.

2. Introduction to ℓ_1 Computation

In 1760, the Croatian Jesuit Roger Boscovich, while on a visit to London, posed the following problem to Thomas Simpson:

> Let there be any number of quantities a, b, c, d, e, all given and let it be required to find corrections to be applied to them under these conditions
> 1. That their differences may be in a given ratio.
> 2. That the sum of the positive corrections may be equal to the sum of the negative.
> 3. That the sum of the positive and sum of the negative corrections may be a minimum.

In modern notation the problem may be formulated as follows: find $\hat{\alpha}, \hat{\beta}$ such that

$$y_i = \hat{\alpha} + \hat{\beta} x_i + \hat{u}_i$$

for given observations (y_i, x_i) $i = 1, \ldots, n$, the "corrections" $\hat{u}_i$ satisfy

$$\sum \hat{u}_i = 0$$

and

$$\sum |\hat{u}_i| = \min!$$

Clearly the differences in the corrected observations

$$\hat{y}_i - \hat{y}_j = \hat{\beta}(x_i - x_j)$$

would then satisfy the first requirement that they were in given ratios, determined solely by the x's.

Stigler (1984) in his lively commentary on this exchange, concludes that Simpson made some progress on the problem, recognizing, for example, that the solution should pass through the point $(\bar{x}, \bar{y})$ and one data point. Boscovich provided a geometric solution, but the problem does not seem to have been fully resolved in print until the publication of Laplace (1789), who recognized, see Stigler (1986), that one could obtain the slope estimate, $\hat{\beta}$, by computing a weighted median of the candidate slopes $s_i = (y_i - \bar{y})/(x_i - \bar{x})$ $i = 1, 2, \ldots, n$. More explicitly, let $s_{(i)}$ denote the ordered s_i's and $w_{(i)}$ be the associated weights, $w_i = |x_i - \bar{x}|$, ordered according to the s_i's, then $\hat{\beta} = s_{(m)}$ where

$$m = \min\{j | \sum_{i=1}^{j} |w_{(i)}| \geq \sum_{i=1}^{n} |w_{(i)}|/2\}.$$

With the advent of least squares at the end of the 18th century, Boscovich's prototype ℓ_1-estimator faded into obscurity. It was revived more than a century later by Edgeworth (1887) who, like Laplace earlier, argued that it could deliver better estimates when the required "corrections" did not happen to follow the Gaussian law. Subsequently, Edgeworth (1888) discarded Boscovich's second constraint that the residuals sum to zero, and proposed to minimize the sum of absolute residuals with respect to both intercept and slope parameters, calling this his "double median" method. He noted that this approach could be extended, in principle, to a "plural median" method. A geometric algorithm was given for the bivariate case, and a discussion of conditions under which such median methods were preferable to least-squares methods was also provided. Unfortunately, the geometric approach to computing Edgeworth's new median regression estimator was rather awkward requiring, as he admitted, "the attention of a mathematician, and in the case of many unknowns some power of hypergeometrical conception," (Edgeworth(1888), p. 190).

Only with the emergence of the simplex algorithm for linear programming in the late 1940's did ℓ_1 methods become practical on a large scale. Papers by Charnes, Cooper and Ferguson (1955), Wagner (1959) and others provided a foundation for modern implementations, such as Barrodale and Roberts (1974) and Bartels and Conn (1980). See Bloomfield and Steiger (1983) for an extensive discussion of the algorithmic development of ℓ_1 methods, including some very interesting empirical comparisons of the performance of several competing algorithms.

The simplex approach to solving the general ℓ_1 regression problem,

$$\min_{b \in \Re^p} \sum_{i=1}^{n} |y_i - x_i' b| \tag{2.1}$$

relies on the reformulation as the linear program,

$$\min\{\ e'u + e'v \ \mid \ y = Xb + u - v, (u, v) \in \Re_+^{2n}\}. \tag{2.2}$$

This problem has the dual formulation

$$\max\{y'd \ \mid \ X'd = 0, \quad d \in [-1, 1]^n\}, \tag{2.3}$$

or equivalently, setting $a = d + \frac{1}{2}e$,

$$\max\{y'a \ \mid \ X'a = \tfrac{1}{2}X'e, \ \ a \in [0, 1]^n\}. \tag{2.4}$$

A p-element subset of $\mathcal{N} = \{1, 2, ..., n\}$ will be denoted by h, and $X(h), y(h)$ will denote the submatrix and subvector of X, y with the corresponding rows and elements identified by h. Recognizing that solutions to (2.1) may be characterized as planes which pass through precisely $p = \dim(b)$ observations, or as convex combinations of such "basic" solutions, we can begin with any such solution, which we may write as,

$$b(h) = X(h)^{-1} y(h). \tag{2.5}$$

We may regard any such "basic" primal solution as an extreme point of the polyhedral convex constraint set. A natural algorithmic strategy is then to move to the adjacent vertex of the constraint set in the direction of steepest descent. This transition involves two stages: the first chooses a descent direction by considering the removal of each of the current basic observations and computing the gradient in the resulting direction, then having selected the direction of steepest descent and thus an observation to be removed from the currently active "basic" set, we must find the maximal step length in the chosen direction by searching over the remaining $n-p$ available observations for a new element to introduce into the "basic" set. Each of these transitions involves an elementary "simplex pivot" matrix operation to update the current basis. The iteration continues in this manner until no direction is found at which point the current $b(h)$ can be declared optimal.

The simplex algorithm offers an extremely efficient approach to computing ℓ_1-type estimators for many applications, yielding, as we shall see below, timings that are quite competitive with least squares on problems of moderate size. But the performance of simplex on large problems is somewhat less satisfactory. Problems of sample size 50,000 may require as much as 50 times the computational effort of least squares to compute a median regression, ℓ_1, estimate. Recent implementations of simplex, notably those of Bixby and collaborators, see e.g. Bixby discussion of Lustig, Marsden and Shanno (1994), primarily address efficient treatment of sparsity, and preprocessing to eliminate strictly dominated constraints, and therefore do not seem to be promising from the point of view of statistical applications. We begin our search for more efficient methods for large problems by looking in the most obvious place: the literature on interior point algorithms for linear programming, which have dramatically improved upon simplex for a broad class of problems.

3. Interior Point Methods for Canonical LP's

Although prior work in the Soviet literature offered theoretical support for the idea that linear programs could be solved in polynomial time, Karmarkar(1984) constituted a watershed in thinking about linear programming both by making a more cogent theoretical argument and by offering direct evidence for the first time that interior point methods were demonstrably faster in specific, large, practical problems.

The close connection between the interior point approach of Karmarkar (1984) and earlier work on barrier methods for constrained optimization, notably Fiacco and McCormick (1968) was observed by Gill, et al. (1986) and others, and has led to what may be called without much fear of exaggeration a paradigm shift in the theory and practice of linear and nonlinear programming. Remarkably, some of the fundamental ideas required for this shift appeared already in the 1950's in a sequence of Oslo working papers by the economist Ragnar Frisch. This work is summarized in Frisch (1956). We will sketch the main outlines of the approach, with the understanding that further details may be found in the excellent expository papers of Wright (1992), Lustig, Marsden and Shanno (1994), and the references cited there.

Consider the canonical linear program

$$\min \ \{c'x \mid Ax = b, \ \ x \geq 0\}, \tag{3.1}$$

and associate with this problem the following logarithmic barrier (potential-function) reformulation,

$$\min \ \{B(x, \mu) \mid Ax = b\} \tag{3.2}$$

where

$$B(x, \mu) = c'x - \mu \sum \log x_k.$$

In effect, (3.2) replaces the inequality constraints in (3.1) by the penalty term of the log barrier. Solving (3.2) with a sequence of parameters μ such that $\mu \to 0$ we obtain in the limit a solution to the original problem (3.1). This approach was elaborated in Fiacco and McCormick (1968) for general constrained optimization, but was revived as a linear programming tool only after its close connection to the approach of Karmarkar (1984) was pointed out by Gill, et al. (1986). The use of the logarithmic potential function seems to have been introduced by Frisch (1956), who described it in the following vivid terms,

> 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.

Suppose that we have an initial feasible point, x_0, for (3.1), and consider solving (3.2) by the classical Newton method. Writing the gradient and Hessian of B with respect to x as

$$\begin{aligned} \nabla B &= c - \mu X^{-1} e \\ \nabla^2 B &= \mu X^{-2} \end{aligned}$$

where $X =$ diag (x) and e denotes an n-vector of ones, we have at each step the Newton problem

$$\min_p \{c'p - \mu p' X^{-1} e + \tfrac{1}{2} \mu p' X^{-2} p \mid Ap = 0\}. \tag{3.3}$$

Solving this problem, and moving from x_0 in the resulting direction p toward the boundary of the constraint set maintains feasibility and is easily seen to improve the objective function. The first order conditions for this problem may be written as

$$\mu X^{-2} p + c - \mu X^{-1} e = A'y \tag{3.4}$$

$$Ap = 0 \tag{3.5}$$

where y denotes an m-vector of Lagrange multipliers. Solving for y explicitly, by multiplying through in the first equation by AX^2 and using the constraint to eliminate p, we have

$$AX^2A'y = AX^2c - \mu AXe. \tag{3.6}$$

These normal equations may be recognized as generated from the linear least squares problem

$$\min_y \| XA'y - Xc - \mu e \|_2^2 \tag{3.7}$$

Solving for y, computing the Newton direction p from 3.4, taking a step in the Newton direction toward the boundary constitute the essential features of the primal log barrier method. A special case of this approach is the affine scaling algorithm in which we take $\mu = 0$ at each step in (3.6), an approach anticipated by Dikin (1967) and studied by Vanderbei, Meketon and Freedman (1986) and numerous subsequent authors.

Recognizing that similar methods may be applied to the primal and dual formulations simultaneously, recent theory and implementation of interior point methods for linear programming have focused on attacking both formulations. The dual problem corresponding to (3.1) may be written as

$$\max\{b'y \mid A'y + z = c, \;\; z \geq 0\}. \tag{3.8}$$

Optimality in the primal implies

$$c - \mu X^{-1}e = A'y \tag{3.9}$$

so setting $z = \mu X^{-1}e$ we have system

$$\begin{aligned} Ax &= b \quad x > 0 \\ A'y + z &= c \quad z > 0 \\ Xz &= \mu e. \end{aligned} \tag{3.10}$$

Solutions $(x(\mu), y(\mu), z(\mu))$ of these equations constitute the central path of solutions to the logarithmic barrier problem, which approach the classical complementary slackness condition $x'z = 0$, as $\mu \to 0$, while maintaining primal and dual feasibility along the path.

If we now apply Newton's method to this system of equations, we obtain

$$\begin{pmatrix} Z & 0 & X \\ A & 0 & 0 \\ O & A' & I \end{pmatrix} \begin{pmatrix} p_x \\ p_y \\ p_z \end{pmatrix} = \begin{pmatrix} \mu e - Xz \\ b - Ax \\ c - A'y - z \end{pmatrix} \tag{3.11}$$

which can be solved explicitly as,

$$
\begin{aligned}
p_y &= (AZ^{-1}XA')^{-1}[AZ^{-1}X(c - \mu X^{-1}e - A'y) + b - Ax] \\
p_x &= XZ^{-1}[A'p_y + \mu X^{-1}e - c + A'y] \\
p_z &= -A'p_y + c - A'y - z.
\end{aligned} \tag{3.12}
$$

Like the primal method, the real computational effort of computing this step is the Choleski factorization of the diagonally weighted matrix $AZ^{-1}XA'$. Note that the consequence of moving from a purely primal view of the problem to one that encompasses both the primal and dual is that $AX^{-2}A'$ has been replaced by $AZ^{-1}XA'$ and the right hand of the equation for the y-Newton step has altered somewhat. But the computational effort is essentially identical. To complete the description of the primal-dual algorithm we would need to specify how far to go in the Newton direction p, how to adjust μ as the iterations proceed, and how to stop.

In fact, the most prominent examples of implementations of the primal-dual log barrier approach now employ a variant due to Mehrotra (1992), which resolves all of these issues. We will briefly describe this variant in the next section in the context of a slightly more general class of linear programs which encompasses the ℓ_1 problem as well as the general linear quantile regression problem.

4. Interior Point Methods for Quantile Regression

Quantile regression, as introduced in Koenker and Bassett (1978), places asymmetric weight on positive and negative residuals, and solves the slightly modified ℓ_1 problem,

$$
\min_{b \in \Re^p} \sum_{i=1}^{n} \rho_\tau(y_i - x_i'b) \tag{4.1}
$$

where $\rho_\tau(r) = r(\tau - I(r < 0))$ for $\tau \in (0,1)$. This yields the modified linear program,

$$
\min\{\tau e'u + (1-\tau)e'v \mid y = Xb + u - v, (u,v) \in \Re^{2n}_{+}\}, \tag{4.2}
$$

and dual formulations,

$$
\max\{y'd \mid X'd = 0, \;\; d \in [\tau - 1, \tau]^n\} \tag{4.3}
$$

or, setting $a = d + 1 - \tau$,

$$
\max\{y'a \mid X'a = (1-\tau)X'e, \;\; a \in [0,1]^n\}. \tag{4.4}
$$

The dual formulation of the quantile regression problem fits nicely into the standard formulations of interior point methods for linear programs with bounded variables. The function $a(\tau)$ that maps $[0,1]$ to $[0,1]^n$ plays a crucial role in connecting the statistical theory of quantile regression to the classical theory of rank tests as described in Gutenbrunner and Jurečková (1992) and Gutenbrunner, Jurečková , Koenker and Portnoy (1993). See Koenker and d'Orey (1987,1993) for a detailed description of modifications of the Barrodale and Roberts(1974) simplex algorithm for this problem.

Adding slack variables s, and the constraint $a+s=e$, we obtain the barrier function

$$B(a,s,\mu) = y'a + \mu \sum_{i=1}^{n} (\log a_i + \log s_i), \tag{4.5}$$

which should be maximized subject to the constraints, $X'a = (1-\tau)X'e$ and $a+s=e$. The Newton step, δ_a, solving

$$\max y'\delta_a + \mu\delta_a'(A^{-1} - S^{-1})e - \tfrac{1}{2}\mu\delta_a'(A^{-2} + S^{-2})\delta_a \tag{4.6}$$

subject to $X'\delta_a = 0$, satisfies

$$y + \mu(A^{-1} - S^{-1})e - \mu(A^{-2} + S^{-2})\delta_a = Xb \tag{4.7}$$

for some $b \in \Re^p$, and δ_a such that $X'\delta_a = 0$. As before, multiplying through by $X'(A^{-2}+S^{-2})^{-1}$ and using the constraint, we can solve explicitly for the vector b,

$$b = (X'WX)^{-1}X'W(y + \mu(A^{-1} - S^{-1})e) \tag{4.8}$$

where $W = (A^{-2}+S^{-2})^{-1}$. This is a form of the primal log barrier algorithm described above. Setting $\mu = 0$ in each step yields an affine scaling variant of the algorithm. We should stress again that the basic linear algebra of each iteration is essentially unchanged, only the form of the diagonal weighting matrix W has changed. We should also emphasize that there is nothing especially sacred about the explicit form of the barrier function used in (4.5). Indeed, one of the earliest proposed modifications of Karmarkar's original work was the affine scaling algorithm of Vanderbei, Meketon and Freedman (1986), which used, implicitly, $\mu \sum_{i=1}^{n} \log(\min(a_i, s_i))$ in lieu of the additive specification.

Again, it is natural to ask if a primal-dual form of the algorithm could improve performance. In the bounded variables formulation we have the Lagrangian,

$$L(a,s,b,u,\mu) = B(a,s,\mu) - b'(X'a - (1-\tau)X'e) - u'(a+s-e), \tag{4.9}$$

and setting $v = \mu A^{-1}$ we have the first order conditions, describing the central path, see Gonzaga(1992),

$$\begin{aligned} X'a &= (1-\tau)X'e \\ a + s &= e \\ Xb + u - v &= y \\ USe &= \mu e \\ AVe &= \mu e, \end{aligned} \tag{4.10}$$

yielding the Newton step,

$$\begin{aligned}
\delta_b &= (X'WX)^{-1}((1-\tau)X'e - X'a + X'W\xi(\mu)) \\
\delta_a &= W(X\delta_b + \xi(\mu)) \\
\delta_s &= -\delta_a \\
\delta_u &= \mu A^{-1}e - Ue - A^{-1}U\delta_a \\
\delta_v &= \mu S^{-1}e - Ve + S^{-1}V\delta_s
\end{aligned} \tag{4.11}$$

where $\xi(\mu) = y - Xb + \mu(S^{-1} - A^{-1})e$. The most successful implementations of this approach to date employ the predictor-corrector step of Mehrotra (1992) which is described in the context of bounded variables problems in Lustig, Marsden and Shanno (1992). A related earlier approach is described in Zhang(1992). In Mehrotra's approach we proceed somewhat differently. Rather than solving for the Newton step (4.11) directly, we substitute the step directly into (4.10), to obtain,

$$\begin{aligned}
X'(a+\delta_a) &= (1-\tau)X'e \\
(a+\delta_a) + (s+\delta_s) &= e \\
X(b+\delta_b) + (u+\delta_u) - (v+\delta_v) &= y \\
(U+\Delta_u)(S+\Delta_s) &= \mu e \\
(A+\Delta_a)(V+\Delta_v) &= \mu e,
\end{aligned} \tag{4.12}$$

where $\Delta_a, \Delta_v, \Delta_u, \Delta_s$ denote the diagonal matrices with diagonals, $\delta_a, \delta_v, \delta_u, \delta_s$ respectively. As noted by Lustig, Marsden and Shanno, the primary difference between solving this system and the prior Newton step is the presence of the nonlinear terms $\Delta_u\Delta_s$, $\Delta_a\Delta_v$ in the last two equations. To approximate a solution to these equations, Mehrotra (1992) suggests first solving for an affine primal-dual direction by setting $\mu = 0$ in (4.11). Given this preliminary direction, we may then compute the step length using the following ratio test,

$$\hat{\gamma}_P = \sigma \min\{\min_j\{-a_j/\delta_{a_j}, \delta_{a_j}\}, \min_j\{-s_j/\delta_{s_j}, \delta_{s_j}\}\} \tag{4.13}$$

$$\hat{\gamma}_D = \sigma \min\{\min_j\{-u_j/\delta_{u_j}, \delta_{u_j}\}, \min_j\{-v_j/\delta_{v_j}, \delta_{v_j}\}\}. \tag{4.14}$$

using scaling factor $\sigma = .99995$, as in Lustig, Marsden, and Shanno. Then defining the function,

$$\hat{g}(\hat{\gamma}_P, \hat{\gamma}_D) = (s + \hat{\gamma}_P\delta_s)'(u + \hat{\gamma}_D\delta_u) + (a + \hat{\gamma}_P\delta_a)'(v + \hat{\gamma}_D\delta_v) \tag{4.15}$$

the new μ is taken as,

$$\mu = \left(\frac{\hat{g}(\hat{\gamma}_P, \hat{\gamma}_D)}{\hat{g}(0,0)}\right)^3 \frac{\hat{g}(0,0)}{2n}. \tag{4.16}$$

To interpret (4.15) we may use the first three equations of (4.10) to write, for any primal-dual feasible point (u, v, s, a)

$$\tau e'u + (1-\tau)e'v - (a - (1-\tau)e)'y = u's + a'v. \tag{4.17}$$

So the quantity $u's + a'v$ is equal to the duality gap, the difference between the primal and dual objective function values at (u, v, s, a), and $\hat{g}(\hat{\gamma}_P, \hat{\gamma}_D)$ is the duality gap after the tentative affine scaling step. Note that the quantity $a - (1-\tau)e$ is simply the vector d appearing in the dual formulation (4.3). At a solution, classical duality theory implies that the duality gap vanishes, that is the values of the primal and dual objective functions are equal and the complementary slackness condition, $u's + a'v = 0$ holds. If, in addition to feasibility, (u, v, s, a) happened to lie on the central path, the last two equations of (4.10) would imply that,

$$u's + a'v = 2\mu n.$$

Thus, the function $\hat{g}$ in (4.15) may be seen as an attempt to adapt μ to the current iterate in such a way that for any given value of the duality gap, μ is chosen to correspond to the point on the central path with that gap. By definition, $\hat{g}(\hat{\gamma}_P, \hat{\gamma}_D)/\hat{g}(0, 0)$ is the ratio of the duality gap after the tentative affine-scaling step to the gap at the current iterate. If this ratio is small the proposed step is favorable and we should reduce μ further, anticipating that the recentering and nonlinearity adjustment of the modified step will yield further progress. If, on the other hand, $\hat{g}(\hat{\gamma}_P, \hat{\gamma}_D)$ isn't much different from $\hat{g}(0, 0)$, the affine scaling direction is unfavorable, and further reduction in μ is ill-advised. Since leaving μ fixed in the iteration brings us back to the central path, such unfavorable steps are intended to enable better progress in subsequent steps iy bringing the current iterate back to the vicinity of the central path. The rationale for the cubic adjustment in (4.16) which implements these heuristics, is based on the fact that the recentering of the Newton direction embodied in the terms $\mu A^{-1}e$ and $\mu S^{-1}e$ of (4.11) and (4.18) accomodates the $\mathcal{O}(\mu)$ term in the expansion of the duality gap function $\hat{g}$ while the nonlinearity adjustment described below accomodates the $\mathcal{O}(\mu^2)$ effect of the $\delta_s\delta_u$ and $\delta_a\delta_v$ terms.

We compute the following approximation to the solution of the system (4.12) with this μ and the nonlinear terms $\Delta_s\Delta_u$ and $\Delta_a\Delta_v$ taken from the preliminary primal-dual affine direction:

$$\begin{aligned}
\delta_b &= (X'WX)^{-1}((1-\tau)X'e - X'a + X'W\xi(\mu)) \\
\delta_a &= W(X\delta_b + \xi(\mu)) \\
\delta_s &= -\delta_a \\
\delta_u &= \mu A^{-1}e - Ue - A^{-1}U\delta_a + A^{-1}\Delta_s\Delta_u e \\
\delta_v &= \mu S^{-1}e - Ve + S^{-1}V\delta_s + S^{-1}\Delta_a\Delta_v e.
\end{aligned} \tag{4.18}$$

The iteration proceeds until the algorithm terminates when the duality gap $y'a - (1-\tau)e'Xb + e'v$ becomes smaller than a specified ϵ. Recall that the duality gap is zero at a solution, and thus, this criterion offers a more direct indication of convergence than is usually available in iterative algorithms.

Our expectations about satisfactory computational speed of regression estimators are inevitably strongly conditioned by our experience with least squares. In Figure 4.1 we illustrate the results of a small experiment to compare the computational speed of 3 ℓ_1 algorithms: the Barrodale and Roberts (1973) simplex algorithm which is employed in many contemporary statistical packages, Meketon's affine scaling algorithm, and our implementation of Mehrotra's (1992) predictor-corrector version of the primal-dual log barrier algorithm. The former is indicated in the figure as `mek` and the latter as `rqfn` for regression quantiles *via* Frisch-Newton. The two interior point algorithms were coded in Fortran employing Lapack, (Anderson, *et al* (1995)), subroutines for the requisite linear algebra. They were then incorporated as functions into Splus and timings are based on the Splus function `unix-time()`. The Barrodale and Roberts timings are based on the Splus implementation `l1fit(x,y)`. For comparison purposes we also illustrate timings for least squares estimation based on Splus function `lm(y ~ x)`.

Such comparisons are inevitably fraught with qualifications about programming style, system overhead, *etc.*. We have chosen to address the comparison within the Splus environment because a.) it is the computing environment in which we feel most comfortable, a view widely shared by the statistical research community, and b.) it offers a convenient means of incorporating new functions in lower level languages, like Fortran and C, providing a reasonably transparent and efficient interface with the rest of the language. We have considerable experience with the Barrodale and Roberts (1974) Fortran code as implemented in Splus for `l1fit`. This code also underlies the quantile regression routines described in Koenker and d'Orey (1987, 1993) and still represents the state-of-the-art after more than 20 years. The Splus function `l1fit` incurs a modest overhead getting problems into and out of BR's Fortran, but this overhead is quickly dwarfed by the time spent in the Fortran in large problems. Similarly, we have tried to write the interior point code to minimize the Splus overhead, although some improvements are still possible in this respect.

Least-squares timings are also potentially controversial. The Splus function `lm` as described by Chambers (1992) offers three method options: QR decomposition, Cholesky, and singular value decomposition. All of our comparisons are based on the default choice of the QR method. Again there is a modest overhead involved in getting the problem descriptions into and the solutions out of the lower level Lapack routines which underlie `lm`. We have run some very limited timing comparisons outside Splus directly in Fortran to evaluate these overhead effects and our conclusion from this is that any distortions in relative performance due to overhead effects are slight.

We would stress that the code underlying the least squares computations we report is the product of decades of refinement, while our interior point routines are still in their infancy. There is still considerable scope for improvement in the latter.

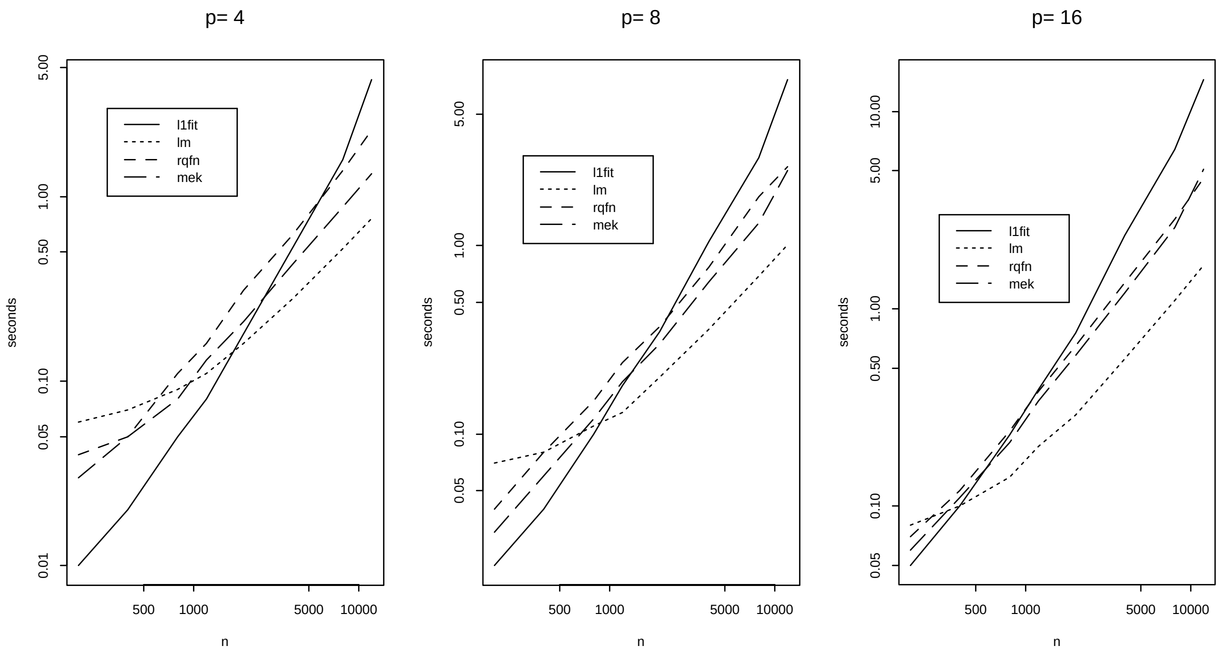

Figure 4.1. Timing comparison of three ℓ_1-algorithms for median regression: Times are in seconds for the median of five replications for iid Gaussian data. The parametric dimension of the models is $p+1$ with p indicated above each plot, p columns are generated randomly and an intercept parameter is appended to the resulting design. Timings were made at 8 design points in n: 200, 400, 800, 1200, 2000, 4000, 8000, 12000. The solid line represents the results for the simplex-based Barrodale and Roberts algorithm implemented in Splus as `l1fit`, the `rqfn` dashed line represents a primal-dual interior point algorithm, `mek` uses an affine scaling form of the interior point approach and the dotted line represents least squares timings based on $\texttt{lm}(\texttt{y} \sim \texttt{x})$ as a benchmark

Several features of the figures are immediately striking. For small problems all the ℓ_1 algorithms perform impressively. They are all faster than the QR implementation of least squares which is generally employed in `lm`. For small problems the simplex implementation of Barrodale and Roberts is the clear winner, but its roughly quadratic (in sample size) growth over the illustrated range quickly dissipates its initial advantage. The interior point algorithms do considerably better than simplex at larger sample sizes, exhibiting roughly linear growth, as does least-squares. Meketon's affine scaling algorithm performs slightly better than the primal- dual algorithm, which is somewhat surprising, but for larger p the difference is hardly noticeable.

Beyond the range of problem sizes illustrated here, the advantage of the interior point method over simplex grows exorbitant, fully justifying the initial enthusiasm with which Karmarkar (1984) was received. Nevertheless, there is still a significant gap between ℓ_1 and ℓ_2 performance in large samples. We explore this gap from the probabilistic viewpoint of computational complexity in the next section.

5. Computational Complexity

In this section we investigate the computational complexity of the interior point algorithms for quantile regression described above. We should stress at the outset, however, that the probabilistic approach to complexity analysis adopted here is rather different than that employed in the rest of the interior-point literature where the focus on worst case analysis has led to striking discrepancies between theoretical rates and observed computational experience. The probabilistic approach has the virtue that the derived rates are much sharper and consequently more consonant with observed performance. A similar gap between worst-case theory and average practice can be seen in the analysis of parametric linear programming via the simplex algorithm, where it is known that in certain problems with an n by p constraint matrix there can be as many as n^p distinct solutions. However, exploiting some special aspects of the quantile regression problem and employing a probabilistic approach, Portnoy (1991) was able to show that the number of distinct vertex solutions (in τ) is $\mathcal{O}_p(n \log n)$, a rate which provides excellent agreement with empirical experience.

For interior point methods the crux of the complexity argument rests on showing that at each iteration the algorithm reduces the duality gap by a proportion, say $\theta_n < 1$. Thus after K iterations, an initial duality gap of Δ_0 has been reduced to $\theta_n^K \Delta_0$. Once the gap is sufficiently small, say, less than ε, there is only one vertex of the constraint set at which the duality gap can be smaller. This follows obviously from the fact that the vertices are discrete. Thus, the vertex with the smaller duality gap must be the optimal one, and this vertex may be identified by taking p simplex-type steps. This process, called purification in Gonzaga (1992, Lemma 4.7), requires in our notation p steps involving $\mathcal{O}(np^2)$ operations, or $\mathcal{O}(np^3)$ operations. Hence, the number of iterations, K, required to make, $\theta_n^K \Delta_0 < \epsilon$ is,

$$K < \log(\Delta_0/\epsilon)/(-\log \theta_n).$$

In the worst-case analysis of the interior point literature, ε is taken to be 2^{-L} where L is the total number of binary bits required to encode the entire data of the problem. Thus, in our notation ε would be $\mathcal{O}(np)$. Further, the conventional worst-case analysis employs the bound $\theta_n < (1 - cn^{-1/2})$, and takes Δ_0 independent of n so the number of required iterations is $\mathcal{O}(\sqrt{n}L)$. Since each iteration requires a weighted least-squares solutions of $\mathcal{O}(np^2)$ operations, the complexity of the algorithm as a whole would be $\mathcal{O}(n^{5/2}p^3)$, apparently hopelessly disadvantageous relative to least squares. Fortunately, however, in the random problems for which quantile regression methods are designed, the ε bound on the duality gap at the second best vertex can be shown to be considerably larger, at least with probability tending to 1, than this worst case value of 2^{-L}. Lemma A.1 provides the bound $\log \epsilon = \mathcal{O}_p(p \log n)$ under mild conditions on the underlying regression model. This leads to a considerably more optimistic view of these methods for large problems.

Renegar (1988) and numerous subsequent authors have established the existence of a large class of interior point algorithms for solving linear programs which, starting from an initially feasible primal-dual point with duality gap Δ_0, can achieve convergence to a prescribed accuracy ϵ in $\mathcal{O}(\sqrt{n}\log(\Delta_0/\epsilon))$ iterations *in the worst case.* More recently, Sonnevend, Stoer, and Zhao (1991) have shown under somewhat stronger nondegeneracy conditions that this rate can be improved to $\mathcal{O}(n^a \log(\Delta_0/\epsilon))$ with $a < 1/2$. We will call an algorithm which achieves this rate an n^a-algorithm. They give explicit conditions, which hold with probablity one if the the y's have a continuous density, for the case $a = 1/4$. The following result then follows immediately from Lemma A.1.

Theorem 5.1. *Under the conditions of Lemma A.1, an n^a-algorithm for median regression converges in $\mathcal{O}_p(n^a p \log n)$ iterations. And with $\mathcal{O}(np^2)$ operations required per iteration and $\mathcal{O}(np^3)$ operations required for the final "purification" process such an algorithm has complexity, $\mathcal{O}_p(n^{1+a}p^3 \log n)$.*

Mizuno, Todd and Ye (1993) provide an alternative probabilistic approach to the existence of an n^a-algorithm, with $a < 1/2$ and provide a heuristic argument for $a = 1/4$. They also conjecture that n^a might be improvable to $\log n$, by a more refined probabilistic approach. This would improve the overall complexity in the above Theorem to $\mathcal{O}_p(np^3 \log^2 n)$ and seems quite plausible in light of the empirical evidence reported below, and elsewhere in the interior point literature. In either case we are still faced with a theoretical gap between ℓ_1 and ℓ_2 performance that substantiates the empirical experience reported in the previous section. We now introduce a new form of preprocessing for ℓ_1 problems that has been successful in further narrowing this gap.

6. Preprocessing for Quantile Regression

Many modern linear programming algorithms include an initial phase of preprocessing which seeks to reduce problem dimensions by identifying redundant variables and dominated constraints. See, for example, the discussion in Section 8.2 of Lustig, Marsden, and Shanno(1993) and the remarks of the discussants. Bixby, in this discussion, reports reductions of 20-30% in the row and column dimensions of a sample of standard commercial test problems due to "aggressive implementation" of preprocessing. Standard preprocessing strategies for LP's are not, however, particularly well-suited to the statistical applications which underlie quantile regression. In this section we describe some new preprocessing ideas designed explicitly for quantile regression, which can be used to reduce dramatically the effective sample sizes for these problems.

The basic idea underlying our preprocessing step rests on the following elementary observation. Consider the directional derivative of the median regression, ℓ_1, problem

$$\min_b \sum_{i=1}^n |y_i - x_i'b|$$

which may be written in direction w as

$$g(b, w) = \sum_{i=1}^n x_i'w \ \mathrm{sgn}^*(y_i - x_i'b, x_i'w),$$

where

$$\mathrm{sgn}^*(u, v) = \begin{cases} \mathrm{sgn}(u) & \text{if } u \neq 0, \\ \mathrm{sgn}(v) & \text{if } u = 0. \end{cases}$$

Optimality may be characterized as a b^* such that $g(b^*, w) \geq 0$ for all $w \in \Re^p$. Suppose for the moment that we "knew" that a certain subset J_H of the observations $N = \{1, \dots n\}$ would fall above the optimal median plane and another subset J_L would fall below. Then consider the revised problem

$$\min_{b \in \Re^p} \sum_{i \in N \setminus J_L \cup J_H} |y_i - x_i'b| + |y_L - x_L'b| + |y_H - x_H'b|$$

where $x_K = \sum_{i \in J_K} x_i$, for $K \in \{H, L\}$ and y_L, y_H can be chosen as arbitrarily small and large enough, respectively, to ensure that the corresponding residuals remain negative and positive. We will refer in what follows to these combined pseudo-observations as "globs". The new problem, under our provisional hypothesis, has exactly the same gradient condition as the original one, and therefore the same solutions, but the revision has reduced effective sample size by $\#\{J_L, J_H\} - 2$, i.e. by the number of observations in the globs.

How might we know J_L, J_H? Consider computing a preliminary estimate $\hat{\beta}$ based on a subsample of m observations. Compute a simultaneous confidence band for $x_i'\beta$ based on this estimate for each $i \in N$. Under plausible sampling assumptions we will see that the length of each interval is proportional to $p/\sqrt{m}$, so if M denotes the number of y_i falling inside the band, $M = \mathcal{O}_p(np/\sqrt{m})$. Take J_L, J_H to be composed of the indices of the observations falling outside the band. So we may now create the "globbed" observations $(y_K, x_K), K \in \{L, H\}$ and reestimate based on $M + 2$ observations. Finally, we must check to verify that all the observations in J_H, J_L have the anticipated residual signs; if so, we are done, if not, we must repeat the process. If the coverage probability of the bands is P, presumably near 1, then the expected number of repetitions of this process is the expectation of a geometric random variable, Z, with expectation P^{-1}. We will call each repetition a cycle.

6.1. **Implementation.** In this subsection we will sketch some further details of the preprocessing strategy. We should emphasize that there are many aspects of the approach that deserve further research and refinement. In an effort to encourage others to contribute to this process we have made all of the code described below available at the website `http://www.econ.uiuc.edu/research/rqn/rqn.html`. We will refer in what follows to the Frisch-Newton quantile regression algorithm *with preprocessing* as `prqfn`.

The basic structure of the current `prqfn` algorithm looks like this:

```
k ← 0
l ← 0
m ← [2n^{2/3}]
while(k is small){
    k = k + 1
    solve for initial rq using first m observations
    compute confidence interval for this solution
    reorder globbed sample as first M observations
    while(l is small){
        l = l + 1
        solve for new rq using the globbed sample
        check residual signs of globbed observations
        if no bad signs: return optimal solution
        if only few bad: adjust globs, reorder sample, update M, continue
        if too many bad: increase m and break to outer loop
        }
    }
```

The algorithm presumes that the data has undergone some initial randomization so the first m observations may be considered representative of the sample as a whole. In all of the experiments reported below we use the Mehrotra-Lustig-Marsden-Shanno primal-dual algorithm to compute the subsample solutions. For some "intermediately large" problems it would be preferable to use the simplex approach, but we postpone this refinement. Although the affine scaling algorithm of Meketon(1986) exhibited excellent performance on certain subsets of our early test problems, like those represented in Figure 4.1, we found its performance inconsistent in other tests. It was consequently abandoned in favor of the more reliable primal-dual formulation. This choice is quite consistent with the general development of the broader literature on interior point methods for linear programming, but probably also deserves further exploration.

6.2. **Confidence Bands.** The confidence bands used in our reported computational experiments are of the standard Scheffé type. Under iid error assumptions the covariance matrix of the initial solution is given by

$$V = \omega^2 (X'X)^{-1}$$

where $\omega^2 = \tau(1-\tau)/f^2(F^{-1}(\tau))$; the reciprocal of the error density at the τth quantile is estimated using the Hall-Sheather (1986) bandwidth for Siddiqui's(1960) estimator. Quantiles of the residuals from the initial fit are computed using the Floyd-Rivest(1975) algorithm. We then pass through the entire sample computing the intervals,

$$B_i = (x_i'\hat{\beta} - \zeta ||\hat{V}^{1/2} x_i||, x_i'\hat{\beta} + \zeta ||\hat{V}^{1/2} x_i||).$$

The parameter ζ is currently set naively, at 2, but could, more generally, be set as $\zeta = (\Phi^{-1}(1-\alpha) + \sqrt{2p-1})/\sqrt{2} = \mathcal{O}(\sqrt{p})$ to achieve $(1-\alpha)$ coverage for the band, and thus assures that the number of cycles is geometric. Since, under the moment condition of Lemma A.1, if $p \to \infty$, the quantity $||\hat{V}^{1/2} x_i||$ also behaves like the square-root of a χ^2 random variable, the width of the confidence band is of $\mathcal{O}_p(p/\sqrt{m})$.

Unfortunately, using the Scheffé bands requires $\mathcal{O}(np^2)$ operations, a computation of the same order as that required by least-squares estimation of the model. It seems reasonable, therefore, to consider alternatives. One possibility, suggested by the Studentized range, is to base intervals on the inequality,

$$|x_i'\hat{\beta}| \le \max_j \left\{ \left| \hat{\beta}_j \right| / s_j \right\} \times \sum_{j=1}^{p} |x_{ij}| \, s_j \, , \tag{6.1}$$

where s_j is $\hat{\omega}$ times the jth diagonal element of the $(X'X)^{-1}$ matrix, and $\hat{\omega}$ is computed as for the Scheffé intervals. This approach provides conservative (though not "exact") confidence bands with width $c_q \sum_{j=1}^{p} |x_j| \, s_j$. Note that this requires only $\mathcal{O}(np)$ operations, thus providing an improved rate. Choice of the constant, c_q, is somewhat problematic, but some experimentation with simulated data showed that c_q could be taken conservatively to be approximately one, and that the algorithm was remarkably independent of the precise value of c_q. For these bands the width is again $\mathcal{O}_p(p/\sqrt{m})$, as for the Scheffé bands. Although these $\mathcal{O}(np)$ confidence bands worked well in simulation experiments, and thus merit further study, the computational experience reported here is based entirely on the more traditional Scheffé bands.

After creating the globbed sample, we again solve the quantile regression problem, this time with the M observations of the globbed sample. Finally we check the signs of the globbed observations. If they all agree with the signs predicted by the confidence band we may declare victory and return the optimal solution. If there are only a few incorrect signs we have found it expedient to adjust the globs, reintroduce these observations into the new globbed sample and resolve. If there are too many incorrect signs, we return to the initial phase, increasing the initial sample size somewhat, and repeat the process. One or two repetitions of the inner (fixup) loop are not unusual; more than two cycles of the outer loop is highly unusual given current settings of the confidence band parameters.

6.3. **Choosing** m. The choice of the initial subsample size, m, and its implications for the complexity of an interior point algorithm for quantile regression *with preprocessing* is resolved by the next lemma.

Theorem 6.1. *Under the conditions of Lemma A.1, for any nonrecursive quantile regression algorithm with complexity,* $\mathcal{O}_p(n^\alpha p^\beta \log n)$*, for problems with dimension* (n, p)*, there exists a confidence band construction based on an initial subsample of size* m *with expected width,* $\mathcal{O}_p(p/\sqrt{m})$*, and consequently, the optimal initial subsample size is* $m^* = \mathcal{O}((np)^{2/3})$*. With this choice of* m^**,* M *is also* $\mathcal{O}((np)^{2/3})$*. Then, with* $\alpha = 1 + a$*, and* $\beta = 3$*, from Theorem 5.1, the overall complexity of the algorithm with preprocessing is, for any* n^a *underlying interior point algorithm,*

$$\mathcal{O}_p((np)^{2(1+a)/3} p^3 \log n) + \mathcal{O}_p(np).$$

For $a < 1/2$*,* n *sufficiently large, and* p *fixed, this complexity is dominated by the complexity of the confidence band computation, and is strictly smaller than the* $\mathcal{O}(np^2)$ *complexity of least-squares.*

Proof: Formally, we treat only the case of p fixed, but we have tried to indicate the role of p in the determination of the constants, where possible. Thus, for example, for $p \to \infty$, we have suggested above that the width of both the Scheffé bands and the Studentized range bands are $\mathcal{O}_p(p/\sqrt{m})$. For p fixed this condition is trivially satisfied. By independence we may conclude that the number of observations inside such a confidence band will be,

$$M = \mathcal{O}_p(np/\sqrt{m}),$$

and minimizing, for any constant c,

$$m^\alpha p^\beta \log m + (cnp/\sqrt{m})^\alpha p^\beta \log(cnp/\sqrt{m}) \tag{6.2}$$

yields,

$$m^* = \mathcal{O}((np)^{2/3}).$$

Substituting this m^* back into (6.2), Theorem 5.1 implies that we have complexity,

$$\mathcal{O}((np)^{2(1+a)/3} p^3 \log n),$$

for each cycle of the preprocessing. The number of cycles required is bounded in probability since it is a realization of a geometrically distributed random variable with a finite expectation. The complexity computation for the algorithm as a whole is completed by observing that the required residual checking is $\mathcal{O}(np)$ for each cycle, and employing the Studentized range confidence bands also requires $\mathcal{O}(np)$ operations per cycle. Thus the contribution of the confidence band construction and residual checking is precisely $\mathcal{O}_p(np)$, and for any $a < 1/2$ the complexity of the ℓ_1 algorithm is therefore dominated by this term for any fixed p and n sufficiently large. ∎

Remarks. (1.) Clearly these results above apply not only to median regression, but to quantile regression in general. (2.) If the explicit rates in p of the Theorem hold for $p \to \infty$, and if the Mizuno-Todd-Ye conjecture that n^a can be improved to $\log n$ holds, then the complexity of the algorithm becomes,

$$\mathcal{O}(n^{2/3}p^3 \log^2 n) + \mathcal{O}_p(np).$$

The contribution of the first term in this expression would then assure an improvement over least squares for n sufficiently large, provided $p = o(n^{1/5})$, a rate approaching the domain of nonparametric regression applications. (3.) It is tempting to consider the recursive application of the preprocessing approach described above, and this can be effective in reducing the complexity of the solution of the initial subsample m problem, but it does not appear possible to make it effective in dealing with the globbed sample. This accounts for the qualifier "nonrecursive" in the statement of the theorem.

7. Computational Experience

In this section we provide some further evidence on the performance of our implementation of the algorithm on both simulated and real data. In Figure 7.1 we compare the performance of `l1fit` with the new `prqfn`, which combines the primal-dual algorithm with preprocessing. With the range of sample sizes 20,000 - 120,000, the clear superiority of `prqfn` is very striking. At $n = 20,000$ `prqfn` is faster than `l1fit` by a factor of about 10, and it is faster by a factor of 100 at $n = 120,000$. The quadratic growth in the `l1fit` timings is also quite apparent in this figure.

In Figure 7.2 we illustrate another small experiment to compare `rqfn` and `prqfn` with `lm` for n up to 180,000. Patience, or more accurately the lack thereof, however, doesn't permit us to include further comparisons with `l1fit`. Figure 7.2 displays the improvement provided by preprocessing, and shows that `prqfn` is actually slightly faster than `lm` at for $p = 4$ and quite close to least squares speed for $p = 8$ for this range of sample sizes. It may be noted that internal Fortran timings of of `prqfn` have shown that most of the time is spent in the primal-dual routine `rqfn` for $n < 200,000$. The results of Sections 5-6 suggest that the greatest value of preprocessing appears when n is large enough that the time needed to create the globs and check residuals is comparable to that spent in `rqfn`.

Finally, we report some experience with a moderately large econometric application. This is a fairly typical wage equation as employed in the labor economics literature. See Buchinsky (1994,1995) for a much more extensive discussion of related results. The data are from the five percent sample of the 1990 U.S. Census, and consists of annual salary and related characteristics on 113,547 men from the state of Illinois who

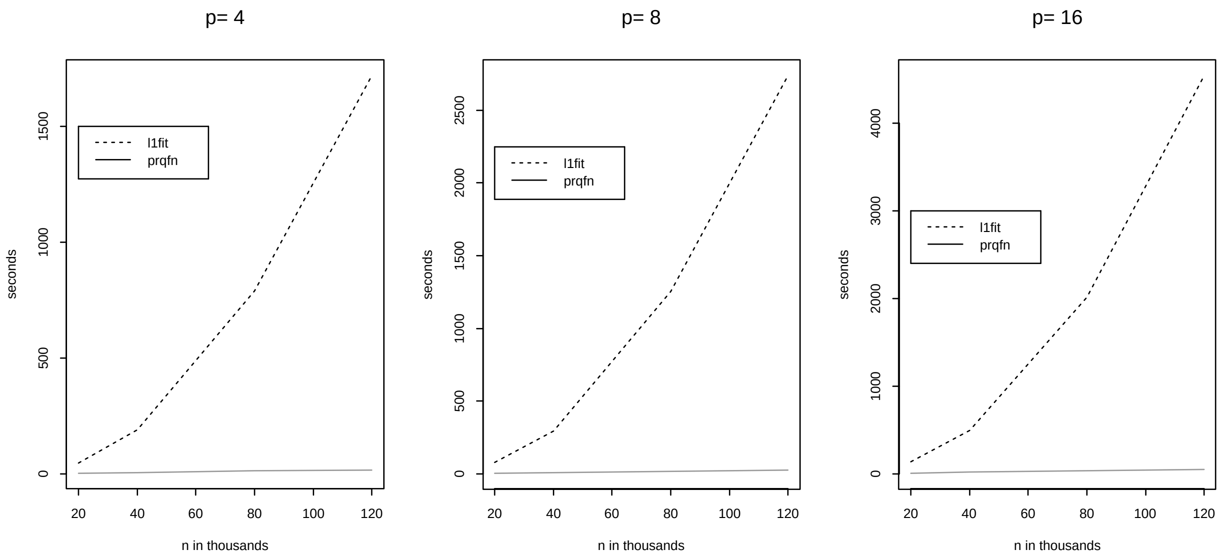

Figure 7.1. Timing comparison of two ℓ_1-algorithms for median regression: Times are in seconds for the mean of five replications for iid Gaussian data. The parametric dimension of the models is $p + 1$ with p indicated above each plot, p columns are generated randomly and an intercept parameter is appended to the resulting design. Timings were made at 4 design points in n: 20000, 40000, 80000, 120000. The dotted line represents the results for the simplex-based Barrodale and Roberts algorithm `l1fit`, which increases roughly quadratically in n. The solid line represents `prqfn`, the timings of the Frisch-Newton interior point algorithm, with preprocessing,

responded that they worked 40 or more weeks in the previous year and who worked on average 35 or more hours per week.

We seek to investigate the determinants of the logarithm of individuals' reported wage or salary income in 1989 based on their attained educational level, a quadratic labor market experience effect, and other characteristics. Results are reported for five distinct quantiles. Least squares results for the same model appear in the final column of the table. The standard errors reported in parentheses were computed by the sparsity method described in Koenker (1994) using the Hall-Sheather bandwidth. There are a number of interesting findings. The experience profile of salaries is quite consistent across quantiles, with salary increasing with experience at a decreasing rate. There is a very moderate tendency toward more deceleration in salary growth with experience at the lower quantiles. The white-nonwhite salary gap is highest at the first quartile, with whites receiving a 17 percent premium over non-whites with similar characteristics, but this appears to decline both in the lower tail and for higher quantiles. Marriage appears to entail an enormous premium at the lower quantiles, nearly a 30 percent premium at the fifth percentile for example, but this premium declines somewhat as salary rises. The least squares results are quite consistent with the median regession results, but we should emphasize that the pattern of estimated quantile regression coefficients in the table as a whole is quite inconsistent with the classical iid-error linear model, or indeed, any of the conventional models accomodating some form of parametric heteroscedasticity.

In Table 7.2 we report the time (in seconds) required to produce the estimates in the previous table, using three alternative quantile regression algorithms. The time required for the least squares estimates reported in the last column of Table 7.1 was 7.8 seconds, roughly comparable to the `prqfn` times. Again, the interior-point approach with preprocessing as incorporated in `prqfn`, is considerably quicker than

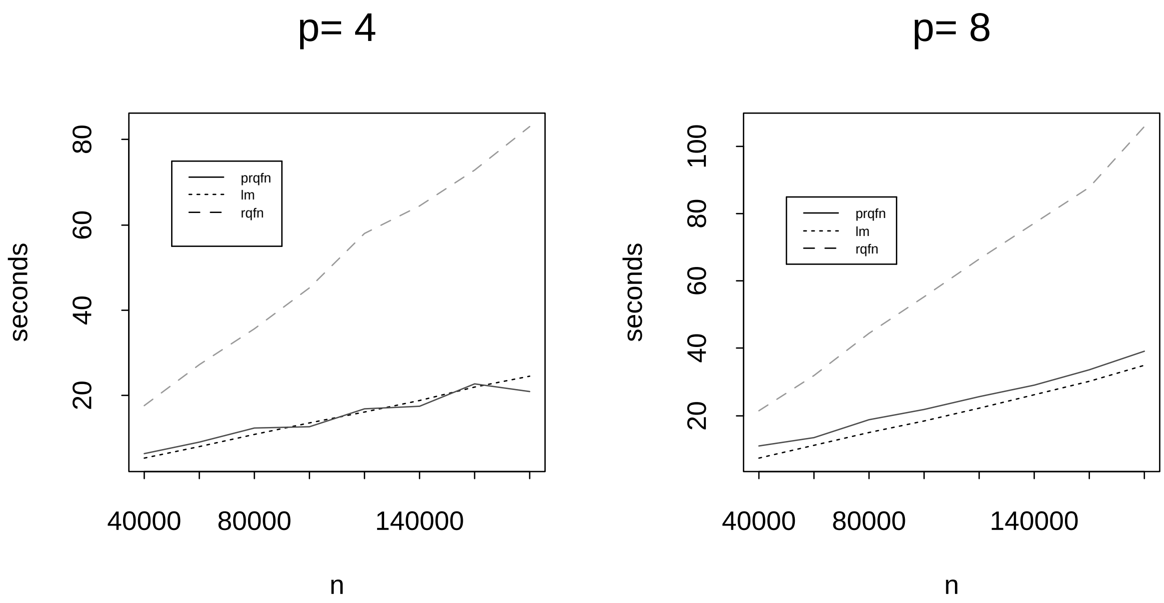

FIGURE 7.2. Timing comparison of two ℓ_1-algorithms for median regression: Times are in seconds for the mean of ten replications for iid Gaussian data. The parametric dimension of the models is $p+1$ with p indicated above each plot, p columns are generated randomly and an intercept parameter is appended to the resulting design. Timings were made at 8 design points in n: 40,000, 60,000, 80,000, 100,000, 120,000, 140,000, 160,000, 180,000. The `rqfn` dashed line represents a primal-dual interior point algorithm, `prqfn` is `rqfn` with preprocessing, and the dotted line represents least squares timings based on `lm(y ~ x)` as a benchmark.

covariate	τ =0.05	τ =0.25	τ =0.5	τ =0.75	τ =0.95	ols
intercept	7.60598 (0.028468)	7.95888 (0.012609)	8.27162 (0.009886)	8.52930 (0.010909)	8.54327 (0.025368)	8.21327 (0.010672)
exp	0.04596 (0.001502)	0.04839 (0.000665)	0.04676 (0.000522)	0.04461 (0.000576)	0.05062 (0.001339)	0.04582 (0.000563)
exp2	−0.00080 (0.000031)	−0.00075 (0.000014)	−0.00069 (0.000011)	−0.00062 (0.000012)	−0.00056 (0.000028)	−0.00067 (0.000012)
education	0.07034 (0.001770)	0.08423 (0.000784)	0.08780 (0.000615)	0.09269 (0.000678)	0.11953 (0.001577)	0.09007 (0.000664)
white	0.14202 (0.014001)	0.17084 (0.006201)	0.15655 (0.004862)	0.13930 (0.005365)	0.10262 (0.012476)	0.14694 (0.005249)
married	0.28577 (0.011013)	0.24069 (0.004878)	0.20120 (0.003824)	0.18083 (0.004220)	0.20773 (0.009814)	0.21624 (0.004129)

TABLE 7.1. Quantile Regression Results for a U.S. Wage Equation

method	τ =0.05	τ =0.25	τ =0.5	τ =0.75	τ =0.95
prqfn	9.92	9.78	19.91	7.68	8.64
rqfn	41.07	42.34	28.33	40.87	59.69
rq	565.97	2545.42	3907.42	3704.50	3410.49

TABLE 7.2. Timing Comparisons for 3 Methods in Wage Equation Example: Results are given in seconds for three different quantile regression algorithms described in the text.

the interior point algorithm applied to the full data set in `rqfn`. The simplex approach to computing quantile regression estimates is represented here by the modification of the Barrodale and Roberts(1974) algorithm described in Koenker and d'Orey (1987), and denoted by `rq` in the table. There is obviously a very substantial gain in moving away from the simplex approach to computation in large problems of this type.

8. CONCLUSIONS

There is a compelling general case for the superiority of interior point methods over traditional simplex methods for large linear programming problems, and for large quantile regression applications in particular. We have shown that preprocessing can effectively reduce the sample size dimension of quantile regression problems from n to $\mathcal{O}_p(n^{2/3})$, thereby enabling computational speed comparable to that of least squares for some large quantile regression problems. There are many possible refinements of the basic approach investigated here, but the message for that Gaussian rabbit who has been frolicking in the flowers, confident of victory, is clear. Laplace's old tortoise, despite the house he wears on his back to protect him from inclement statistical weather, has a few new tricks and the race is far from over.

APPENDIX A.

Lemma A.1. *In the linear model* $Y_i = x_i'\beta + u_i \quad i = 1, \dots, n$, *assume:*

(i) $\{(x_i, Y_i), i = 1, \dots, n\}$ *are iid with a bounded continuous density in* $\Re^{p+1}$.

(ii) $E|x_{ij}|^p < \infty$ *and* $E|Y_i|^a < \infty$, *for some* $a > 0$.

Then the duality gap of the median regression estimator at the second best vertex exceeds $n^{-(p+5)}$ *with probability tending to 1 as* $n \to \infty$, *and the initial duality gap* Δ_0 *satisfies* $\log \Delta_0 = \mathcal{O}_p(\log n)$.

Proof: Let $\hat{\beta}$ denote the optimal median regression solution based on the data $\{(x_i, Y_i),\ i = 1, ..., n\}$ and $\hat{d}$ denote the corresponding dual solution. Consider the duality gap at another trial solution pair, $(\tilde{\beta}, \tilde{d})$, which we can write,

$$\Delta^* = \sum_{i=1}^{n} |Y_i - x_i'\tilde{\beta}| \,-\, Y'\tilde{d} = \sum_{i=1}^{n} |Y_i - x_i'\tilde{\beta}| - \sum_{i=1}^{n} |Y_i - x_i'\hat{\beta}| \,+\, Y'\hat{d} - Y'\tilde{d}. \tag{A.1}$$

since the duality gap at $(\hat{\beta}, \hat{d})$, is zero. Now, as in Koenker and Bassett (1978), let $h = \{i_1, , \cdots, i_p\}$, denote a subset of $N = \{1, 2, ..., n\}$ consisting of p distinct indices. Define $X(h)$ be the $p \times p$ matrix with rows $\{x_i' \,:\, i \in h\}$, $Y(h)$ to be the vector with coordinates $\{Y_i \,:\, i \in h\}$, and let $\beta(h) = X^{-1}(h)Y(h)$. Now suppose $\hat{h}$ is the subset defining the optimal solution, i.e. $\hat{\beta} = \beta(\hat{h})$, and that $\tilde{h}$ represents the vertex of the constraint set (distinct from $\hat{h}$) for which $\tilde{\beta} = \beta(\tilde{h})$ is nearest to $\hat{\beta}$ in terms of the primal objective function.

Note that $\tilde{h}$ is also a subset of size p and differs from $\hat{h}$ in exactly one element. Let i_1 be the (unique) index in $\hat{h} \cap \tilde{h}^c$, and let $i_2 \in \tilde{h} \cap \hat{h}^c$, where h^c denotes the complement of h Let $\hat{r}_i$ and $\tilde{r}_i$ denote the respective ith residuals for $\hat{\beta}$ and $\tilde{\beta}$, and note the following: $\hat{r}_i = 0$ for $i \in \hat{h}$, $\tilde{r}_i = 0$ for $i \in \tilde{h}$, $\mathrm{sgn}(\hat{r}_i) = \mathrm{sgn}(\tilde{r}_i)$ for $i \notin \hat{h} \cup \tilde{h}$, and $\mathrm{sgn}(\hat{r}_{i_2}) = -\mathrm{sgn}(\tilde{r}_{i_1})$. Therefore, since the dual contribution to Δ^* is positive, we can write,

$$\begin{aligned}
\Delta^* \geq \Delta_1^* &\equiv \sum_{i \notin h} |Y_i - x_i'\tilde{\beta}| \,-\, \sum_{i \notin \tilde{h}} |Y_i - x_i'\hat{\beta}| \\
&= \left(Y_{i_1} - x_{i_1}'\tilde{\beta}\right) \mathrm{sgn}(\tilde{r}_{i_1}) - \left(Y_{i_2} - x_{i_2}'\hat{\beta}\right) \mathrm{sgn}(\hat{r}_{i_2}) \\
&\quad - \sum_{i \notin \hat{h} \cup \tilde{h}} x_i'\tilde{\beta}\, \mathrm{sgn}(\tilde{r}_i) \,+\, \sum_{i \notin \hat{h} \cup \tilde{h}} x_i'\hat{\beta}\, \mathrm{sgn}(\hat{r}_i) \\
&= (Y_{i_1} + Y_{i_2})\, \mathrm{sgn}(\tilde{r}_{i_1}) \,-\, \sum_{i \notin \tilde{h}} \mathrm{sgn}(\tilde{r}_i)\, x_i' X^{-1}(\tilde{h}) Y(\tilde{h}) \\
&\quad + \sum_{i \notin \hat{h}} \mathrm{sgn}(\hat{r}_i)\, x_i' X^{-1}(\hat{h}) Y(\hat{h}) \\
&\equiv \sum_{j=1}^{p+1} b_j\, Y_{i_j}
\end{aligned} \tag{A.2}$$

where for $j = 2, \cdots, p$,

$$
\begin{aligned}
b_1 &= \operatorname{sgn}(\tilde{r}_{i_1}) - \left(\sum_{i \notin \tilde{h}} \operatorname{sgn}(\tilde{r}_i)\, x_i' X^{-1}(\tilde{h}) \right)_1 \\
b_2 &= \operatorname{sgn}(\tilde{r}_{i_1}) - \left(\sum_{i \notin \hat{h}} \operatorname{sgn}(\hat{r}_i)\, x_i' X^{-1}(\hat{h}) \right)_1 \qquad \text{(A.3)} \\
b_{j+1} &= \left(\sum_{i \notin \hat{h}} \operatorname{sgn}(\hat{r}_i)\, x_i' X^{-1}(\hat{h}) \right)_j - \left(\sum_{i \notin \tilde{h}} \operatorname{sgn}(\tilde{r}_i)\, x_i' X^{-1}(\tilde{h}) \right)_j .
\end{aligned}
$$

Now let $\mathcal{H}$ be the set of all pairs (h, h^*) where h and h^* are p-element subsets of indices that differ in exactly one element. Note that there are $\binom{n}{p+1}\binom{p+1}{2}$ such pairs. For $(h, h^*) \in \mathcal{H}$, let $\Delta_1(h, h^*) = \sum_{j=1}^{p+1} b_j Y_{i_j}$, where $\{b_j\}$ are defined for arbitrary $(h, h^*) \in \mathcal{H}$ as in (A.3), and where $\{i_j\}$ ranges over $h \cup h^*$. Then, clearly,

$$\Delta_1^* \geq \inf\{\Delta_1(h, h^*) : (h, h^*) \in \mathcal{H}\} . \qquad \text{(A.4)}$$

Now fix an arbitrary pair $(h, h^*) \in \mathcal{H}$,

$$
\begin{aligned}
\mathbf{P}\left\{ |\Delta_1(h, h^*)| \leq \frac{1}{n^{p+5}} \right\} &\leq \mathbf{P}\left\{ |b_j| \leq \frac{1}{n^3},\ j = 2, \cdots, p+1 \right\} \\
&\quad + \mathbf{P}\left\{ for\ some\ j : |b_j| > \frac{1}{n^3} \wedge |Y_{i_j} + \sum_{k \neq j} \frac{b_k}{b_j} Y_{i_k}| \leq \frac{1}{n^{p+2}} \right\} \qquad \text{(A.5)} \\
&\equiv P_A + P_B . \qquad \text{(A.6)}
\end{aligned}
$$

Now note that P_A is the probability that the vector whose coordinates are the "sum over $i \notin \hat{h}$" terms in $(b_2, \cdots, b_{p+1})$ (see (A.3)) lies in a specified cube with sides of length $2/n^3$ (centered at the other terms in b_j). That is, if $\mathcal{C}$ denotes the set of all cubes in $\Re^p$ with sides of length $2/n^3$, then

$$P_A \leq \sup\left\{ \mathbf{P}\left\{ \left(\sum_{i \notin h} \operatorname{sgn}(r_i)\, x_i' \right) X^{-1}(h) \in C \right\} : C \in \mathcal{C} \right\} \qquad \text{(A.7)}$$

Now multiplying the condition in P_A through by $X(h)$ and noting that the density of the sum in P_A is bounded by the density of a single x_i (A.7) becomes

$$P_A \leq E_h a_0\, Vol(C) \det(X(h)) \leq a_1 \left(\frac{2}{n^3} \right)^p \leq \frac{a_2}{n^{p+2}} \qquad \text{(A.8)}$$

where a_0, a_1 and a_2 are constants, $\det(X(h))$ is bounded by the moment condition on $\{x_i\}$, and $3p \geq p + 2$ for $p \geq 1$.

Similarly, P_B is bounded by

$$P_B \leq \sup_t \left\{ \mathbf{P}\left\{ Y_{i_j} \in [t, t + n^{-(p+2)}] \right\} \right\} \leq \frac{a_3}{n^{p+2}} \qquad \text{(A.9)}$$

(where a_3 is a bound on the density of Y_i). Thus, from (A.5), (A.8), and (A.9),

$$\mathbf{P}\left\{ |\Delta_1^*| \leq n^{-(p+5)} \right\} \leq \binom{n}{p+1} \binom{p+1}{2} \frac{a_2 + a_3}{n^{p+2}} \to 0 . \qquad \text{(A.10)}$$

That is, $\Delta_1^* > n^{-(p+5)}$ with probability tending to one.

Finally, taking the initial β and d to be 0, the initial duality gap is bounded by $\Delta_0 = \sum |Y_i|$. But a simple Chebychev-type inequality based on the moment condition $(E|Y_i|^a < +\infty)$ yields

$$\mathbf{P}\left\{\sum_{i=1}^{n} |Y_i| \geq n^{1+2/a}\right\} \leq n\mathbf{P}\left\{|Y_1| \geq n^{2/a}\right\} \leq \frac{n\, E|Y_1|^a}{n^{2a/a}} \to 0\,.$$

■

REFERENCES

ANDERSON, E. *et al.* (1995). *LAPACK Users' Guide*, SIAM.

BARRODALE, I. AND F.D.K. ROBERTS (1974). Solution of an overdetermined system of equations in the ℓ_1 norm, *Communications ACM,* **17**, 319-320.

BARTELS, R. AND A. CONN (1980). Linearly constrained discrete ℓ_1 problems, *Transactions of the ACM on Mathematical Software*, **6**, 594-608.

BLOOMFIELD, P. AND W.L. STEIGER (1983). *Least Absolute Deviations: Theory, Applications, and Algorithms*, Birkhauser: Boston.

BUCHINSKY, M. (1994). Changes in US wage structure 1963-87: An application of quantile regression, *Econometrica*, **62**, 405-458.

BUCHINSKY, M. (1995). Quantile regression, the Box-Cox transformation model and U.S. wage structure 1963-1987, *Journal of Econometrics*, **65**, 109-154.

CHAMBERS, J.M. (1992). Linear Models, in *Statistical Models in S*, J.M. Chambers and T.J. Hastie, (eds.) Wadsworth: Pacific Grove.

CHAMBERLAIN, G. (1994). Quantile regression, censoring and the structure of wages, *Advances in Econometrics*, C. Sims (ed.), Elsevier: New York.

CHARNES, A., W.W. COOPER AND R.O. FERGUSON, (1955). Optimal estimation of executive compensation by linear programming, *Management Science*, **1**, 138-151.

CHAUDHURI, P. (1992). Generalized regression quantiles *Proceedings of the Second Conference on Data Analysis Based on the L_1 Norm and Related Methods*, North-Holland: New York.

DIKIN, I.I. (1967). Iterative solution of problems of linear and quadratic programming, *Soviet Mat. Dokl.* **8**, 674-675.

CHEN, S., AND D.L. DONOHO (1995). Atomic decomposition by basis pursuit, preprint.

EDGEWORTH, F.Y. (1887). On observations relating to several quantities, *Hermathena*, **6**, 279-285.

EDGEWORTH, F.Y. (1888), On a new method of reducing observations relating to several quantities, *Philosophical Magazine*, **25**, 184-191.

FAN, J. AND I. GIJBELS (1996). *Local Polynomial Modelling and Its Applications*, Chapman and Hall: London.

FIACCO, A.V. AND G.P. MCCORMICK (1968). *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*, Wiley: New York.

FLOYD, R.W. AND R.L. RIVEST (1975) Expected Time Bounds for Selection, *Communications of the ACM*, **18**, 165-173.

FRISCH, R. (1956). La Résolution des problèmes de programme linéaire par la méthode du potential logarithmique, *Cahiers du Séminaire d'Econometrie*, **4**, 7-20.

GAUSS, C.F. (1821). *Theoria combinationis observationum erroribus minimis obnoxiae: Pars prior.* translated by G.W. Stewart(1995) as *Theory of the Combination of Observations Least Subject to Error*, SIAM: Philadelphia.

GILL, P., W. MURRAY, M. SAUNDERS, T. TOMLIN, AND M. WRIGHT (1986). On projected Newton barrier methods for linear programming and an equivalence to Karmarkar's projective method, *Mathematical Programming*, **36**, 183-209.

GONZAGA, C.C. (1992). Path-following methods for linear programming, *SIAM Review*, **34**, 167-224.

GREEN, P.J. AND B.W. SILVERMAN (1994). *Nonparametric Regression and Generalized Linear Models*, Chapman and Hall: London.

GUTENBRUNNER, C. AND J. JUREČKOVÁ (1992). Regression quantile and regression rank score process in the linear model and derived statistics, *Annals of Statistics* **20**, 305-330.

GUTENBRUNNER, C., J. JUREČKOVÁ , R. KOENKER, S. PORTNOY (1993). Tests of linear hypotheses based on regression rank scores, *Journal of Nonparametric Statistics*, **2**, 307-33.

HÁJEK , J. AND Z. ŠIDÁK (1967). *Theory of Rank Tests*, Academia: Prague.

HALL, P. AND S. SHEATHER (1988). On the distribution of a studentized quantile, *JRSS-B*, **50**, 381-391.

KARMARKAR, N. (1984). A new polynomial time algorithm for linear programming, *Combinatorica*, **4**, 373-395.

KOENKER R., P. NG AND S. PORTNOY (1995). Quantile Smoothing Splines, *Biometrika,* **81**, 673-80.

KOENKER, R. AND G. BASSETT (1978). Regression quantiles, *Econometrica*, 46, 33-50.

KOENKER, R. AND V. D'OREY (1987,1993). Computing regression quantiles, *Applied Statistics*, **36**, 383-393, and **43**, 410-414.

LAPLACE, P.-S. (1789). Sur quelques points du système du monde, *Memoires de l'Academie des Sciences de Paris*, reprinted in **11** of *Œvres Complétes*, Gauthier-Villars: Paris.

LUSTIG, I.J., R.E. MARSDEN AND D.F. SHANNO (1992). On implementing Mehrotra's predictor-corrector interior-point method for linear programming, *SIAM Journal of Optimization*, **2**, 435-449.

LUSTIG, I.J., R.E. MARSDEN AND D.F. SHANNO (1994). Interior point methods for linear programming: computational state of the art, with discussion. *ORSA Journal on Computing*, **6**, 1-36.

MANNING, W., L. BLUMBERG AND L.H. MOULTON (1995). The demand for alcohol: the differential response to price, *Journal of Health Economics*, **14**, 123-148.

MEHROTRA, S. (1992). On the implementation of a primal-dual interior point method, *SIAM Journal of Optimization*, **2**, 575-601.

MEKETON, M.S.(1986). Least Absolute Value Regression, Bell Labs Technical Report.

MIZUNO, S., M.J. TODD, AND Y. YE (1993) On adaptive-step primal dual interior point algorithms for linear programming, *Mathematics of Operations Research*, **18**, 964-981.

OJA, H. (1983). Descriptive statistics for multivariate distributions, *Statistics and Probability Letters*, **1**, 327-332.

PORTNOY, S. (1991). Asymptotic behavior of the number of regression quantile breakpoints, *SIAM Journal of Scientific and Statistical Computing,* **12**, 867-883.

POWELL, J.L. (1986). Censored regression quantiles, *Journal of Econometrics, 32*, 143-155.

SHAMIR, R. (1993). Probabilistic analysis in linear programming, *Statistical Science*,**8**, 57-64.

SONNEVEND, G., J. STOER, AND G. ZHAO (1991). On the complexity of following the central path of linear programs by linear extrapolation II, *Mathematical Programming*, **52**, 527-553.

STIGLER, S. M. (1984). Boscovich, Simpson and a 1760 manuscript note on fitting a linear relation, *Biometrika*, **71**, 615-620.

STIGLER, S. M. (1986). *The History of Statistics: Measurement of Uncertainty before 1900*, Harvard University Press:Cambridge.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso, *Journal of Royal Statistical Society (B)*, forthcoming.

WAGNER, H.M., (1959). Linear programming techniques for regression analysis, *Journal of American Statistical Association* **54**, 206-212.

WELSH, A.H. (1996). Robust estimation of smooth regression and spread functions and their derivatives, *Statistica Sinica*, forthcoming?

WRIGHT, M.H. (1992). Interior methods for constrained optimization, *Acta Numerica*, 341-407.

VANDERBEI, R.J., M.J. MEKETON AND B.A. FREEDMAN (1986). A modification of Karmarkar's linear programming algorithm, *Algorithmica*, **1**, 395-407.

ZHANG, Y. (1992) Primal-dual interior point approach fo computing ℓ_1-solutions and ℓ_∞-solutions of overdetermined linear systems, *J. Optimization Theory and Applications*, **77**, 323-341.