1

Roger Koenker

Computational Methods for Quantile Regression

[1]Version: August 3, 2016. This research was partially supported by NSF grant SES-11-53548.

0.1 Introduction

The earliest computation of a median regression estimator, is usually attributed to the Croatian Jesuit, Rudjer Boscovich. In 1760 Boscovich visited London, and as recounted by Stigler (1984) and Farebrother (1990), posed the problem of computing it to Thomas Simpson. In Boscovich's version of the problem the mean residual was constrained to be zero, a requirement that conveniently reduces the problem to finding a (scalar) weighted median. Thus, the bivariate median regression problem of minimizing sum of absolute residuals,

$$\hat{\beta} = \text{argmin}_{(b_0,b_1)\in\mathbb{R}^2}\{\sum_{i=1}^{n} |y_i - b_0 - b_1 x_i|\},$$

is reduced to solving,

$$\hat{\beta}_1 = \text{argmin}_{b_1\in\mathbb{R}}\{\sum_{i=1}^{n} w_i |z_i - b_1|\},$$

where $z_i = (y_i - \bar{y})/(x_i - \bar{x})$ and $w_i = |x_i - \bar{x}|$, for $i = 1, \cdots, n$, and $\hat{\beta}_0 = \bar{y} - \bar{x}\hat{\beta}_1$. Simpson apparently recognized that the solution to the constrained problem could be found as,

$$\hat{\beta}_1 = z_{(j^*)},$$

where $z_{(1)}, \cdots, z_{(n)}$ denote the order statistics of the z_i's, $w_{(1)}, \cdots, w_{(n)}$ denote the correspondingly ordered weights and

$$j^* = \min\{j | \sum_{i=1}^{j} w_{(i)} > \tfrac{1}{2}\sum_{i=1}^{n} w_{(i)}\}.$$

This formulation was maintained by Laplace and became known as the *Methode de Situation.*

Gauss (1809, §186), noting that the Boscovich/Laplace proposal could be generalized by removing the zero mean residual constraint and including more than a single covariate, makes several remarkably astute observations about the resulting procedure:

> It can be easily shown, that a system of values of unknown quantities, derived from this principle alone, must necessarily exactly satisfy as many equations out of the number proposed, as there are unknown quantities, so that the remaining equations come under consideration only so far as they help to *determine the choice*: if, therefore, the equation $V = M$, for example, is of the number of those which are not satisfied, the system of values found according to this principle would in no respect be changed even if any other value N had been observed instead of M, provided that, denoting the computed value by n, the differences $M - n$, $N - n$, were affected by the same signs.

Not only does Gauss recognize in this brief passage that minimizing the sum of absolute residuals yields solutions determined by an exact fit of p observations when there are p parameters to be estimated, but also that these solutions are insensitive to perturbations that do not alter the signs of the residuals. Whether Gauss had further algorithmic ideas is unclear, but he seems well on the way to a full understanding of the linear programming structure of the problem. In his later memoir on least squares fitting, Gauss(1823, §7) seems more on the defensive,

> Laplace has also considered the problem in a similar manner, but he adopted the absolute value of the error as his measure of loss. Now if I am not mistaken this convention is no less arbitrary than mine. Should an error of double size be considered as tolerable as a single error twice repeated or worse? Is it better to assign only twice as much influence to a double error or more? The answers are not self evident, and the problem cannot be resolved by mathematical proofs, but only by an arbitrary decision. Moreover, it cannot be denied that Laplace's convention violates continuity and hence resists analytic treatment, while the results that my convention leads to are distinguished by their wonderful simplicity and generality.

Perhaps by 1823 he had forgotten the wonderful simplicity and generality of the Laplace method that he had grasped so easily earlier?

In a series of papers in the 1880's Edgeworth also proposed removing the constraint on the mean residual, and defined a "plural median" generalizing the original Boscovich formulation to multiple covariates. Edgeworth (1888) suggested an ingenious geometric strategy for the case of bivariate regression that anticipated later development of the simplex algorithm. Noting that points in sample space $(x_i, y_i) \mapsto \{(\alpha, \beta) : \alpha = y_i - x_i\beta\}$ map to lines in parameter space, and thus lines through pairs of points in sample space map to points in parameter space, Edgeworth proposed starting at one of these intersections in parameter space, choosing a direction of steepest descent, and proceeding to the next intersection. Continuing in this fashion eventually leads to a solution characterized by a pair of points that determine the optimal solution. This approach can be generalized to additional covariates, indeed Edgeworth notes this himself, but rather apologetically observes that it may require "the attention of a mathematician ... with some power of hypergeometrical conception."

Edgeworth's formulation of what has become known as the "dual plot," e.g. Rousseeuw and Hubert (1999) incorporates the essential features of the Barrodale and Roberts (1974) algorithm for median regression. Starting from an initial basic solution consisting of an exact fit to p observations, we consider the local consequences of dropping each of the p observations and moving in either a positive or negative direction. Choosing the steepest of the possible directions of descent these choices present, we then decide how far to go by solving a one dimensional weighted median problem of the same type as that originally formulated by Boscovich. This identifies a new observation to replace the one removed by our determination of the descent direction, and the procedure continues until we can no longer find a direction of descent. For problems of modest size, up to a few thousand observations and a few dozen parameters this form of the algorithm is extremely efficient. However, in very large problems we now have available a new arsenal of techniques that can be adapted to various forms of larger problems. We will briefly survey some of these techniques in the sequel.

0.2 Exterior Point Methods

Linear programming and the associated simplex solution method emerged out of the fog of World War II, as did many other important statistical ideas. Dantzig's (2002) memoir recounts that his simplex method ideas arose in 1947 as an attempt to solve a class of military planning problems using methods similar to those he had employed in earlier work with Wald and Neyman on the Neyman-Pearson Lemma. Kantorovich's (1939) contributions were not appreciated in the west until they appeared in translated form in 1960. Following these developments it was quickly recognized that the median regression problem fit nicely

into the linear programming framework; Charnes et al. (1955) appears to be the first explicit use of simplex to solve the median regression problem.

The algorithm of Barrodale and Roberts (1974) was the first to exploit the bounded variables dual form of the median regression problem. The primal median regression problem can be formulated as,

$$\min\{1_n^\top u + 1_n^\top v \mid y - Xb = u - v,\ (u, v) \geq 0\}$$

and seems a bit unwieldy since the minimization is over a $2n + p$ dimensional vector. In contrast the dual problem has the simpler form,

$$\max_a\{y^\top a \mid X^\top a = \tfrac{1}{2}X^\top 1_n,\ a \in [0, 1]\}$$

In effect, Barrodale and Roberts implemented a general form of the Edgeworth dual plot strategy. Given a basic solution, which we can write as

$$b(h) = X(h)^{-1}y(h)$$

where h indexes p element subsets of the integers $\mathbb{N} = \{1, 2, \cdots, n\}$, $X(h)$ denotes the submatrix of X consisting of the rows h, and $y(h)$ is the corresponding subvector of the response y, we need to find the direction of steepest descent. In the original Edgeworth bivariate setting this amounts to looking in one of four possible choices: starting from an intersection in the dual plot we consider dropping one of the two basic observations in h, and moving away from the intersection along the line representing the chosen observation. Thus, we need only look at four possible direction, and among those with negative slope choose the steepest. The Barrodale and Roberts innovation was rather than stopping at the next adjacent vertex to continue in this direction as long as such motion reduced the objective function. This is just the weighted median problem that we have already described. When we are estimating $p > 2$ parameters the situation is essentially the same except that we have $2p$ directions to examine in order to select the descent direction. See Bloomfield and Steiger (1983) for a more detailed investigation of simplex-based algorithms for median regression.

Modification of this approach to compute quantile regression models other than the median is straightforward. In the primal we only replace the 1_n's by appropriate asymmetric weights, in the dual we simply change the $\frac{1}{2}$ to $(1-\tau)$ to obtain the τth regression quantile estimate. Some further details are provided in Koenker and d'Orey (1987). It may seem alarming that there are to be a continuum of problems of this form: Do we really need to solve such problems for every $\tau \in [0, 1]$? Fortunately, the answer to this question is "no"; there are only a finite set of distinct solutions, and they are easily found by classical parametric linear programming methods. Given a vertex solution at a particular, $b(h)$, small changes in τ have no impact on the solution, eventually τ changes enough that the hyperplane representing the objective function is no longer "tangent" to the constraint set at a unique vertex, but now coincides with the constraint set along an entire edge of that set. It is easy to compute these "critical" τ's at which the solution jumps and thereby produce the entire solution path for $\tau \in [0, 1]$. Portnoy (1989) has shown that the expected number of distinct solutions along this path is $O(n \log n)$, of course in the one sample setting there are always precisely n distinct solutions provided that observations are themselves distinct.

Similar parametric programming techniques may be employed to study the solution path for penalized smoothing problems, or lasso-type penalized estimators. They have also proven very useful in inferential applications such as the inversion of the rank tests proposed by Gutenbrunner and Jurečková (1992) which can be carried out by parametric programming to produce confidence intervals for quantile regression coefficients. While it has become common to encounter paeans to the computation of the "entire regularization path," it

should also be recognized that such computations quickly become burdensome in large data applications. It is a great virtue of exterior point methods like simplex that it is easy to trace out trajectory of solutions for parametric families of problems, but the number of distinct solutions can easily become overwhelming and in such cases we need to find better ways to approximate the path. Unfortunately, the great advances made in the development of interior point methods for linear programming and discussed in the next section do not easily lend themselves to this task.

0.3 Interior Point Methods

In contrast to the "exterior point" algorithms exemplified by the Edgeworth procedure and its simplex progeny that move from vertex to vertex on the exterior of the constraint set, interior point methods move from the center of the constraint set toward a vertex solution. Although prior work in the Soviet literature offered theoretical support for the idea that polynomial algorithms for linear programming could be structured in this way, Karmarker (1984) constituted a pivotal moment in the development of optimization tools for linear programs and convex problems more generally. It was quickly recognized that Karmarker's ideas were closely connected to earlier work on barrier methods for nonlinear programming as developed by Fiacco and McCormick (1968) and even earlier for linear programs by Frisch (1956).

The logarithmic barrier method of Frisch for the canonical linear program,

$$\min\{c^\top x \mid Ax = b,\ x \geq 0\}$$

simply replaces the inequality constraints with a penalty term that forces x to stay in the positive orthant,

$$\min\{c^\top x - \mu \sum_{j=1}^{p} \log x_j \mid Ax = b\}.$$

By gradually relaxing the penalty parameter, μ, we can approach a vertex solution as $\mu \to 0$. The modified problem has the obvious advantage that it has a smooth objective that for any fixed μ generates Newton steps. Following the exposition in Portnoy and Koenker (1997), and denoting diagonal matrices by upper-case letters corresponding to lower-case vectors, e.g. $X = \text{diag}(x)$, and letting e denote a p-vector of ones, we can write the quadratic (Newton) problem for a direction of descent, p, starting from x as,

$$\min\{c^\top p - \mu p^\top X^{-1} e + \tfrac{1}{2}\mu p^\top X^{-2} p \mid Ap = b\}.$$

Denoting a vector of Lagrange multipliers for the equality constraint by y, this problem yields first order conditions,

$$\{c - \mu X^{-1} e + \mu X^{-2} p = A^\top y,\ Ap = 0\},$$

which, multiplying through by AX^2, can be reformulated as,

$$AX^2 A^\top y = AX^2 c - \mu A X^1 e.$$

Solving for y and substituting back into the first order conditions yields a Newton direction, δ. The inherent difficulty of each step of this primal log barrier method thus lies in solving

the p by p linear system in this equation. As long as p is modest, or the matrix $AX^2A^\top$ is sparse, this can be done quite efficiently.

Some improvement in performance can be achieved by exploiting both the primal and dual formulations of the problem. The dual of our canonical problem may be expressed as,

$$\max_y \{b^\top y \mid A^\top y + z = c,\ z \geq 0\}.$$

Optimality in the primal implies that, $c - \mu X^{-1}e = A^\top y$, so we can set $z = \mu X^{-1}e$ to satisfy the dual constraint and obtain the system,

$$\begin{aligned} Ax &= b \qquad x \geq 0 \\ A^\top y + z &= c \qquad z \geq 0 \\ Xz &= \mu e. \end{aligned}$$

The parametric trajectory $(x(\mu), y(\mu), z(\mu))$ describes the "central path" from the center of the constraint set to a solution on the boundary of the constraint set satisfying the classical complementary slackness condition, $Xz = 0$ when $\mu = 0$. As described in more detail in Portnoy and Koenker (1997), this primal-dual formulation yields a slighted perturbed version of the primal Newton step described above, but again results in a p by p linear system that requires the same computational effort at each iteration.

To complete the description of the primal-dual method we would need to specify how far to go in the direction, p, how to adjust μ as we proceed along the central path and how to stop. Each of these aspects are addressed in Section 4 of Portnoy and Koenker (1997) where the bounded variables approach of Lustig et al. (1994) and Mehrotra (1992) is adapted to the quantile regression dual problem. This approach has been implemented in Fortran in several variants in the R package `quantreg`, Koenker (2015).

Comparison of performance of the modified Barrodale and Roberts algorithm, Koenker and d'Orey (1987), for quantile regression with the interior point implementation indicates that the exterior point (simplex) approach has a clear advantage for relatively small problems with sample size, n, less than a few thousand and parametric dimension, p, also modest, say less than 20. However, for larger problems IP is substantially quicker and also more accurate than BR. Accuracy of the BR solutions for large problems could be improved by periodically reinverting the current basic solution, since extensive pivoting can produce substantial accumulated error. The IP algorithm typically requires at most only a few dozen iterations and accuracy is easily monitored by the duality gap in the primal dual formulation.

A natural extension of the basic quantile regression problem that maintains its linear programming structure involves the imposition of additional linear inequality constraints on the model parameters. Such constraints arise in a variety of contexts including portfolio optimization and the introduction of shape constraints in nonparametric regression. Koenker and Ng (2005) describe a modified version of the interior point method that is implemented in the `quantreg` package. The only potential difficulty with adding such constraints is lack of an initial feasible solution, in contrast to the original dual problem where the center of the unit cube is always feasible.

When the parametric dimension of the model is large then the original implementations of both the BR and IP methods can be quite slow, so it is worthwhile to consider other options. The first question in such circumstances should always be: How sparse is the design matrix X? In most nonparametric applications like those encompassed by the total variation penalized additive models described in Koenker (2011) and implemented in the `quantreg` function `rqss`, the design matrix is extremely sparse, typically with only 1 or 2 percent nonzero entries. In such cases sparse linear algebra comes to the rescue, and in particular

sparse Cholesky factorization as described in Koenker and Ng (2003) makes the interior point approach entirely feasible even for problems with several thousand parameters.

0.4 Preprocessing

In many linear programming applications we can profitably remove dominated constraints and thereby reduce the effective dimensionality and consequently the effort required to solve the problem. In large quantile regression problems it is worthwhile to consider strategies that might be able to reduce both the column and row dimension of the design matrix. Especially in dense design settings with large column dimension, p, it is natural to consider lasso methods to reduce the column dimension in a preliminary phase. This tactic is described in some detail in Chernozhukov et al. (2016), so I won't dwell on it here, instead I will briefly describe a strategy for reducing the row dimension.

In Portnoy and Koenker (1997) we considered a relatively simple strategy for reducing the row dimension of large, dense problems. Variants of this technique are likely to prove helpful in many applications. An important feature of the linear quantile regression problem – already apparent to Gauss, as we have seen – is that the subgradient condition for optimality of a solution depends only on the signs of the residuals. More explicitly, the directional derivative of the objective,

$$R(b) = \sum_{i=1}^{n} \rho_\tau(y_i - x_i^\top b),$$

in the direction, δ, is

$$\partial R(b, \delta) = -\sum_{i=1}^{n} x_i^\top \delta(\tau - \text{sgn}^*(y_i - x_i^\top b,\ -x_i\delta),$$

where $\text{sgn}^*(u, v) = \text{sgn}(u)I(u \neq 0) + \text{sgn}(v)I(u = 0)$. Optimality at b requires that $\partial R(b, \delta) \geq 0$ for all δ on the unit sphere in $\mathbb{R}^p$. Thus, if we had a way to predict that a group of observations, say $J_L \subset \mathbb{N}$ would be below the optimal hyperplane $\hat{h}(x) = x^\top \hat{\beta}$, and another group $J_H \subset \mathbb{N}$ would be above, we would also know precisely how these observations would contribute to the subgradient. This implies that we could aggregate the observations in J_L and J_H and treat them as two globbed observations with the new objective function,

$$\tilde{R}(b) = \sum_{\mathcal{J}} \rho_\tau(y_i - x_i^\top b),$$

where the index set $\mathcal{J} = \{\mathbb{N} \backslash J_L \backslash J_H, L, H\}$ and $y_j = \sum_{i \in J_j} y_i$ and $x_j = \sum_{i \in J_j} x_i$ for $j = L, H$. If the number of elements of J_L and J_H is large relative to n we have significantly reduced the row dimension of the problem.

Of course, no one is likely to do our predictions for us, but we can easily do them ourselves using a subset of m of the n observations. As shown in Portnoy and Koenker (1997) standard confidence band procedures yield bands with expected width $\mathcal{O}(p/\sqrt{m})$ and it is optimal to choose $\mathcal{O}((np)^{2/3})$ to balance coverage and the complexity of the band construction. Given a band it is easy to determine how many of the original points lie within the band; this number M is of order $\mathcal{O}(np/\sqrt{m})$. Reestimating using the globbed sample of $M + 2 = \mathcal{O}((np)^{2/3})$ observations we have a trial solution. It must now be verified that the globbed observations do indeed lie above or below the fitted hyperplane as predicted.

If they do, we are done; if not we can expand m and try again. The probability of failing this check, π, can be controlled, and the number of required repetitions of this cycle is a geometric random variable with expectation π^{-1}, so we can assure that only a small number of cycles is needed. Each cycle operates on a significantly reduced sample, reducing a sample one million observations for example to only 10 to 20 thousand. The entire strategy is implemented in the ``pfn'' option for the `rq` fitting function of the `quantreg` package in R.

0.5 First-order, Proximal Methods

However pleased we might be with the performance of interior point methods and pre-processing for large problems, there may come a time when the parametric dimension of new problems stretches the effort required for Cholesky factorization at each iteration to the breaking point. When this happens it is time to reconsider first-order, gradient descent methods. Fortunately, here too we find that great progress has been made in recent years and a unified, efficient approach has emerged well suited to modern parallelized computation.

0.5.1 Proximal Operators and the Moreau Envelope

Proximal algorithms for convex optimization rely on additive separability of the objective function and efficient computation of optima for separable components of the problem. This structure is well adapted to a wide variety of statistical applications including quantile regression. A brief introduction to these methods will be sketched here, for further details the reader is encouraged to consult Parikh and Boyd (2013) and the extensive references provided there.

Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be a closed, proper convex function with effective domain, $\mathrm{dom} f = \{x \in \mathbb{R}^n \mid f(x) < \infty\}$. The proximal operator $P_f : \mathbb{R}^n \to \mathbb{R}$ of f is

$$P_f = \mathrm{argmin}_x \{f(x) + \tfrac{1}{2}\|x - v\|_2^2\},$$

where $\|\cdot\|_2$ denotes the usual Euclidean norm. $P_f(v)$ can be interpreted as seeking to minimize f without allowing the solution to move too far away from v. By rescaling the function f, so,

$$P_{\lambda f}(v) = \mathrm{argmin}_x \{f(x) + \tfrac{1}{2\lambda}\|x - v\|_2^2\},$$

we can control the relative strength of this tradeoff. $P_f(v)$ can be viewed as a generalized projection: if f is simply the indicator of a convex set $\mathcal{C}$, so $f(x) = 0$ if $x \in \mathcal{C}$ and $f(x) = \infty$ otherwise, we have,

$$P_f = \mathrm{argmin}_{x \in \mathcal{C}} \|x - v\|_2^2,$$

the Euclidean projection of v onto $\mathcal{C}$.

To illustrate the behavior of the proximal operator, P_f a bit further, Figure 1 depicts the action of P_f for the function,

$$f(x) = \begin{cases} \|x\|_2 & \text{if } x_1 x_2 \geq 1 \\ \infty & \text{otherwise.} \end{cases}$$

Points v are mapped by the proximal operator toward the constrained optimum at $(1, 1)$: when v lies outside the constraint set it is sent to the nearest boundary point, when v lies inside the constraint set it is directed toward the optimum by an amount controlled by λ.

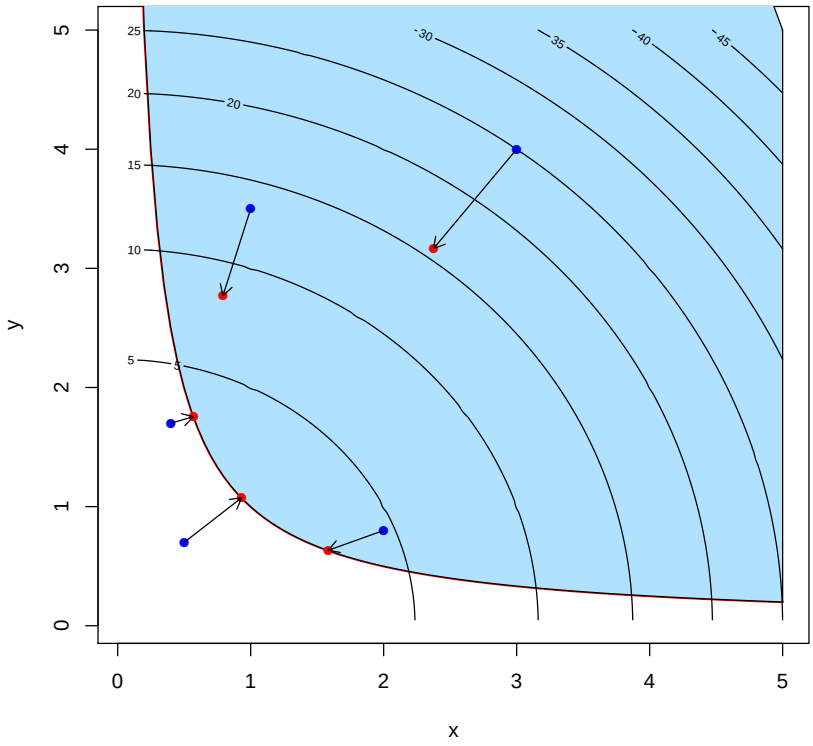

FIGURE 1
The proximal operator $P_f(v)$ projects points, v outside the shaded constraint set to the constraint boundary, while points inside the constraint set are mapped toward the boundary by an amount controlled by the choice of λ.

To pursue the connection to projection a bit further, recall that the infimal convolution of two closed proper convex functions f and g on $\mathbb{R}^n$ is,

$$(f \square g)(v) = \inf_x \{f(x) + g(v - x)\}.$$

If we take $g(x) = \frac{1}{2}\|\cdot\|_2^2$, then,

$$M_{\lambda f}(v) = \inf_x \{f(x) + \tfrac{1}{2\lambda}\|x - v\|_2^2\},$$

is called the Moreau envelope of the function λf. We may view $M_{\lambda f}(v)$ as a smoothed, or regularized version of f and as such it has several advantages. It has domain $\mathbb{R}^n$ even when f does not, it is continuously differentiable even though f may not be, and perhaps most importantly, f and $M_{\lambda f}$ have the same minimizers. Parikh and Boyd (2013) interpret M_f in the following way: letting f^* denote the convex conjugate of f, that is $f^*(y) = \sup_x\{y^\top x - f(x)\}$, we may write

$$M_f = (f^* + \tfrac{1}{2}\|\cdot\|_2^2)^*,$$

so M_f results from adding a smooth regularization to f^*, and then transforming back to obtain a smooth approximation of f. The connections between P_f and M_f are obviously very intimate: $P_f(x)$ is the unique point that achieves the infimum of M_f, that is,

$$M_f(x) = f(P_f(x)) + \tfrac{1}{2}\|x - P_f(x)\|_2^2,$$

and

$$\nabla M_{\lambda f}(x) = \tfrac{1}{\lambda}(x - P_{\lambda f}(x)).$$

The latter expression, when rewritten as,

$$P_{\lambda f}(x) = x - \lambda \nabla M_{\lambda f}(x),$$

reveals that $P_{\lambda f}(x)$ can be viewed as a gradient step of length λ for the regularized function $M_{\lambda f}$. This interpretation also suggests the fixed point iteration,

$$x^{k+1} = P_{\lambda f}(x^k),$$

which can be shown to converge under quite general conditions. As noted by Parikh and Boyd (2013) such methods are closely related to gradient flow methods for solving differential equations, and in special cases to the well-known EM and MM algorithms that have been extensively employed in the statistics literature.

0.5.2 Alternating Direction Method of Multipliers

It is common in statistical applications to encounter optimization problems with additively separable convex components. Suppose for the moment we consider only two components,

$$\min_x \{f(x) + g(x)\},$$

one, or even both, components may represent constraints since they may take on infinite values. A familiar example would be f as (negative) log likelihood and g a lasso-like parametric penalty. When P_f and P_g are easily computed, but P_{f+g} is not, the following iteration is attractive:

$$\begin{aligned} x^{k+1} &= P_{\lambda f}(z^k - u^k) \\ z^{k+1} &= P_{\lambda g}(x^k + u^k) \\ u^{k+1} &= u^k + x^{k+1} - z^{k+1}. \end{aligned}$$

This alternating direction method of multipliers (ADMM) algorithm has broad applicability and has been shown to converge under very mild conditions.

Fougner and Boyd (2015) discuss implementation details for an extension of the ADMM approach introduced in Parikh and Boyd (2014) to problems in the following "graph form,"

$$\min_{(x,y)}\{f(y) + g(x) \mid y = Ax\}.$$

Now, (x, y) is constrained to the graph $\mathcal{G} = \{(x, y) \in \mathbb{R}^{n+m} \mid y = Ax\}$. The modified ADMM algorithm becomes:

$$\begin{aligned}(x^{k+1/2}, y^{k+1/2}) &= (P_{\lambda g}(x^k - \tilde{x}^k), P_{\lambda f}(y^k - \tilde{y}^k))\\ (x^{k+1}, y^{k+1}) &= \Pi_A(x^{k+1/2} - \tilde{x}^k, y^{k+1/2} - \tilde{y}^k)\\ (\tilde{x}^{k+1}, \tilde{y}^{k+1}) &= (\tilde{x}^k + x^{k+1/2} - x^{k+1}, \tilde{y}^{k+1/2} + y^{k+1/2} - y^{k+1})\end{aligned}$$

where Π_A denotes the (Euclidean) projection operator into the set $\mathcal{G}$. This projection has a relatively simple structure and has the advantage that the linear system representing the solution need only be solved once. See Appendix A of Parikh and Boyd (2014) for full details. This contrasts sharply with interior point methods where we repeatedly need to solve linear systems involving a diagonally weight moment matrix.

ADMM algorithms like other first-order gradient type methods have the advantage that they are efficiently parallelizable, all we need to be able to do is compute the proximal operators for f and g and we have a gradient descent strategy based the regularized problem, a strategy that avoids the burden of computing Cholesky factorization at each iteration. Fougner and Boyd (2015) discuss some implementation issues and Fougner (2014) describes a *C++* library with both a Matlab and an R interface. In the next subsection some computational experience with this approach for quantile regression is described.

0.5.3 Proximal Performance

To evaluate performance of the ADMM approach for large quantile regression problems I have carried out some very limited tests based on the GPU implementation of Fougner (2014). These tests were conducted on an unloaded IBM x3400M3 machine running linux with an NVIDIA Tesla C2050 graphics card.

In Figure 0.5.3 we compare the timing and accuracy tradeoff for the interior point solver described above and the GPU implementation of the POGS solver in three large quantile regression problems. Each setting has sample size $n = 10,000$, but p varies from 100 to 300. In each case entries of X are iid standard Gaussian, except for the appended intercept. By adjusting the convergence tolerance we can control the accuracy of both methods. Accuracy is measured in decimal digits relative to the interior point solution solution with the (default) tolerance of $\epsilon = 10^{-6}$, as (minus) the logarithm (base 10) of root mean squared error. The interior point solution yields essentially single precision accuracy averaged over the p estimated coefficients. This can be relatively easily evaluated by tightening the convergence tolerance of the interior point algorithm. The CPU effort required for this benchmark solution is indicated in the figure by the horizontal (red) lines. Further testing revealed that there was little reduction in CPU effort achieved by further relaxation of the interior point tolerance, a finding that is easily explained by examining the number of iterations required. The interior point method is doing at most a few dozen iterations, and relaxing the convergence tolerance saves a few of these; however, the POGS procedure required 1421, 6065, 28867 and 120002 iterations respectively for the four ascending points appearing in the right-most panel of the figure. And the efficacy of the GPU notwithstanding, this takes some time.

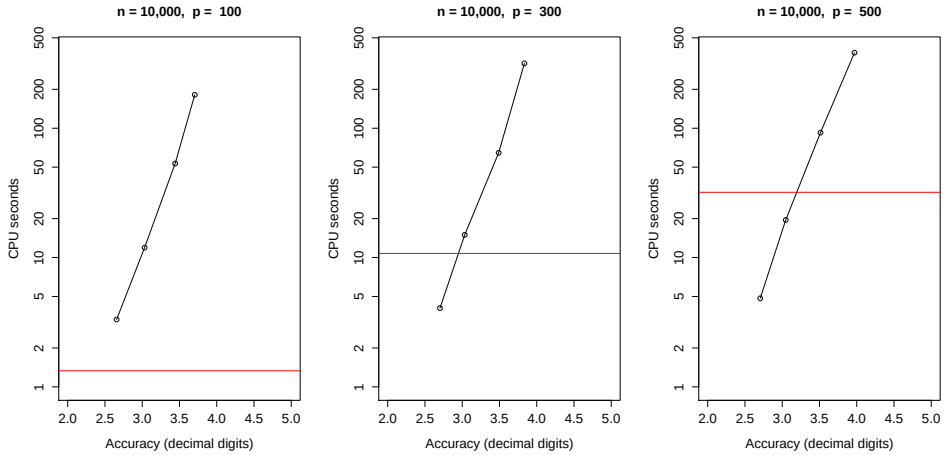

FIGURE 2
Accuracy vs. Computational Effort: CPU effort (in seconds) is plotted against accuracy in the number of correct decimal digits, averaged over the p coefficients for the POGS GPU solutions to the primal quantile regression problem. Baseline accuracy is determined by the interior point solution depicted by the horizontal (red) line, which is accurate to at six decimal digits. Although the POGS procedure is quite quick to produce a solution with two to three digit accuracy, the effort required to produce better accuracy increases rapidly. In contrast, there is little advantage observed in the interior point timings when the convergence tolerance is relaxed.

This performance tradeoff should not be entirely surprising since it is already apparent in other gradient descent algorithms, and has been often remarked upon in applications of the closely related EM algorithm. In some applications it can be easily disregarded since decisions based on such data analysis only require a couple of digits accuracy. Nevertheless, it is somewhat disconcerting in view of our usual obsessions with rates of convergence of statistical procedures. There is an increasing tendency in statistical research to explicitly consider computational effort as well as statistical performance in the evaluation of procedures. It would be nice to understand this better in the present context. Larger sample sizes may not offer the precision we have come to expect, if we cannot reliably estimate models with them.

Bibliography

I. Barrodale and F. Roberts. Solution of an overdetermined system of equations in the ℓ_1 norm. *Communications of the ACM*, 17:319–320, 1974.

P. Bloomfield and W. S. Steiger. *Least Absolute Deviations. Theory, Applications and Algorithms*. Birkhäuser, 1983.

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

Victor Chernozhukov, Kengo Kato, and Alex Belloni. High dimensional quantile regression. In *Handbook of Quantile Regression*, 2016. forthcoming.

G. Dantzig. Linear programming. *Operations Research*, 50, 2002.

F.Y. Edgeworth. On a new method of reducing observations relating to several quantities. *Philosophical Magazine*, 25:184–191, 1888.

R. W. Farebrother. Further details of contacts between boscovich, simpson in june 1760. *Biometrika*, 77:397–400, 1990.

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

Christopher Fougner. *POGS: Proximal Operator Graph Solver*, 2014. `https://github.com/foges/pogs`.

Christopher Fougner and Stephen Boyd. Parameter selection and pre-conditioning for a graph form solver. `http://stanford.edu/~boyd/papers/pogs.html`, 2015.

R. Frisch. La résolution des problèmes de programme linèaire par la mèthode du potential logarithmique. *Cahiers du Seminaire d'Econometrie*, 4:7–20, 1956.

Carl F. Gauss. *Theoria Motus Corporum Celestium*. Perthes et Besser, Hamburg, 1809a. Translated, 1857, as *Theory of Motion of the Heavenly Bodies Moving about the Sun in Conic Sections*, trans. C. H. Davis. Boston, Little, Brown. Reprinted, 1963; New York, Dover.

Carl F. Gauss. *Theoria Combinationis Observationum Erroribus Minimis Obnoxiae*. Dieterich, Gottingen, 1809b. Translated as *Theory of the Combination of Observations Least Subject to Errors*, trans. G. W. Stewart. Siam 1995.

C. Gutenbrunner and J. Jurečková. Regression quantile and regression rank score process in the linear model and derived statistics. *Ann. Statist.*, 20:305–330, 1992.

L. V. Kantorovich. Mathematical methods of organizing and planning production. *Management Science*, 6:pp. 366–422, 1960.

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

R. Koenker. Additive models for quantile regression: Model selection and confidence bandaids. *Brazilian J. of Probability and Statistics*, 25:239–262, 2011.

R. Koenker and V. d'Orey. Computing regression quantiles. *Applied Statistics*, 36:383–393, 1987.

R. Koenker and P. Ng. Sparsem: A sparse linear algebra package for r. *Journal of Statistical Software*, 8, 2003.

R. Koenker and P. Ng. Inequality constrained quantile regression. *Sankhyā*, 67:418–440, 2005.

Roger Koenker. *Quantreg: An R package for quantile regression*, 2015. http://www.R-project.org.

I.J. Lustig, R.E. Marsden, and D.F. Shanno. Interior point methods for linear programming: computational state of the art with discussion. *ORSA J. on Computing*, 6:1–36, 1994.

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

Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, 1:123–221, 2013.

Neal Parikh and Stephen Boyd. Block splitting for distributed optimization. *Math Prog. Comp.*, 6:77–102, 2014.

S. Portnoy. Asymptotic behavior of the number of regression quantile breakpoints. *SIAM J. Science Statistical Computing,*, 12:867–883, 1989.

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

P.J. Rousseeuw and M. Hubert. Regression depth. *J. of Am. Stat. Assoc.*, 94:388–433, 1999.

S. Stigler. Boscovich, simpson and a 1760 manuscript note on fitting a linear relation. *Biometrika*, 71:615–620, 1984.