Journal of Statistical Software

September 2014, Volume 60, Issue 5. *http://www.jstatsoft.org/*

Convex Optimization in R

Roger Koenker
University of Illinois

Ivan Mizera
University of Alberta

Abstract

Convex optimization now plays an essential role in many facets of statistics. We briefly survey some recent developments and describe some implementations of these methods in R. Applications of linear and quadratic programming are introduced including quantile regression, the Huber M-estimator and various penalized regression methods. Applications to additively separable convex problems subject to linear equality and inequality constraints such as nonparametric density estimation and maximum likelihood estimation of general nonparametric mixture models are described, as are several cone programming problems. We focus throughout primarily on implementations in the R environment that rely on solution methods linked to R, like **MOSEK** by the package **Rmosek**. Code is provided in R to illustrate several of these problems. Other applications are available in the R package **REBayes**, dealing with empirical Bayes estimation of nonparametric mixture models.

Keywords: convexity, optimization, linear programming, quadratic programming, second order cone programming, semidefinite programming, lasso, quantile regression, penalty methods, shape-constrained methods.

1. Introduction

Optimality, in statistics as in the rest of life, is probably over-rated; better to be “not bad” most of the time, than perfect once in a while. Tukey taught that, and Huber turned it into a higher form of optimality, although for a partial recantation, see Huber (2009). Given the apparently inescapable need for optimality, what can we do to keep the practical effort of optimization to a minimum? The answer of course, the magic elixir of optimization, is convexity. Without convexity we risk wandering around in the wilderness always looking for a higher mountain, or a deeper valley. With convexity we can proceed with confidence toward a solution.

While convexity plays an essential role in many aspects of statistical theory – it is crucial in the theory of estimation and inference for exponential family models, in experimental design, in the underpinnings of the Neyman-Pearson lemma, and in much of modern decision theory – our main objective will be to describe some recent developments in computational statistics that rely on recent *algorithmic* progress in convex optimization, and to illustrate their implementation in R (R Core Team 2014). In the remaining sections we will sketch some basic unifying theory for a variety of convex optimization problems arising in statistics and discuss some aspects of their implementation in R. In the final section we describe some future developments that we would view as desirable.

2. Convex optimization

Convex optimization seeks to minimize a convex function over a convex (constraint) set. When the constraint set consists of an entire Euclidean space such problems can be easily solved by classical Newton-type methods, and we have nothing to say about these unconstrained problems. Instead we will focus on problems with more restrictive constraint sets, usually polyhedral sets arising as the intersection of linear equality and inequality constraints. Nonlinear inequality constraints are also sometimes of interest, but we should caution that convexity of the constraint set prohibits nonlinear equality constraints.

The extremal conditions of Fermat have undergone a radical transformation since the 17th century. Classical Lagrangian conditions designed to remove equality constraints were extended by the Karush-Kuhn-Tucker (KKT) conditions to handle inequality constraints. Considerable impetus for these developments was provided by the emergence of linear programming methods in the 1940's, but it was not until barrier methods were introduced in the early 1980's that convex optimization really took off as a distinct field. Having seen how linear inequality constraints could be incorporated into the Newton framework via log barrier penalties for linear programming, there was an inevitable torrent of work designed to adapt similar methods to other convex optimization settings. Rockafellar (1993, p. 185) expressed this as

> "In fact, the great watershed in optimization isn't between linearity and nonlinearity, but between convexity and nonconvexity."

We will thus begin with a brief discussion of linear programming problems and methods, briefly describing some parallel developments for quadratic programming problems, and then turn to more general problems beyond the linear-quadratic class.

For the reader interested in a more detailed treatment we might suggest Boyd and Vandenberghe (2004), although there are now numerous treatises that offer an overview of these ideas, all of which owe an enormous debt to the seminal work of Rockafellar (1974, 1996).

2.1. Linear programming

Linear programming (LP) the optimization of a linear objective function subject to linear equality and inequality (polyhedral) constraints, has become an indispensable tool of applied mathematics. The seminal work of Kantorovich (1939) on such problems usually marks the birth of convex optimization as a distinct subject of mathematical inquiry, although the official reception of this paper in the West was delayed until its translation appeared in *Management Science*. Meanwhile Dantzig (1951) and much related work had introduced similar methods and the first revolution of large scale linear programming was well underway. These methods found immediate applications in decision theory and the emerging theory of games, as surveyed by Karlin (1959), in addition to their more remunerative applications in oil refineries and airline scheduling.

The second revolution in linear programming was sparked by Karmarkar (1984) who provided the first compelling case for the advantages of interior point methods over earlier (exterior) simplex methods. As noted by Gill, Murray, Saunders, Tomlin, and Wright (1986), Karmarkar's approach had been partially anticipated in earlier work on log-barrier methods by Frisch (1956) and Fiacco and McCormick (1968) and by the ellipsoidal method of Khachiyan (1979). But the stimulus of a more rigorous justification for the methods produced a vast outpouring of new research; freeing linear programming methods from their habitual paths along the exterior edges of the constraint set also opened the way for a panoply of new methods for non-polyhedral problems. Not only the classical task of (convex) quadratic programming, looked suddenly very similar to linear programming, but new classes of problems as catalogued by Nesterov and Nemirovskii (1987), fell under the interior point rubric in rapid succession.

Example 2.1. (Median regression) Least absolute error or median regression falls into this class of linear objective functions constrained by linear equality and inequality constraints. It has become a standard component of the statistical toolkit. Minimizing

$$\sum_{i=1}^{n} |y_i - x_i^\top b| \equiv \|y - Xb\|_1, \tag{1}$$

we obtain $\hat{\beta} \in \mathbb{R}^p$ as an estimate of the coefficients of a linear approximation of the conditional median function of y given x, just as the corresponding squared error objective yields an estimate of the coefficients of the linear approximation of the conditional expectation function. The primal form of the linear program (1) is a bit unwieldy once one introduces appropriate slack variables,

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

but its dual form,

$$\max_{a\in\mathbb{R}^n} \{y^\top a | X^\top a = \tfrac{1}{2} X^\top e,\ a \in [0,1]^n\}, \tag{3}$$

is quite convenient for both simplex and interior point implementations. Here, e denotes an n-vector of ones, and a denotes a dual vector that represents the "active set" indicators of observations that determine the solution. Solutions take the form $\hat{\beta} = b(h) = X(h)^{-1}y(h)$ where h denotes a p-element subset of the first n integers, $X(h)$ denotes a p by p submatrix of X with row indices h, and $y(h)$ the corresponding p-vector from y. Thus, $\hat{\beta}$ interpolates p observations. And at a dual solution, $\hat{\alpha}$, points above the fitted hyperplane, $H(x) = x^\top\hat{\beta}$ have $\hat{\alpha}_i = 1$, while points below $H(\cdot)$ have $\hat{\alpha}_i = 0$. Active observations that are on $H(\cdot)$ have $\hat{\alpha}_i \in (0,1)$.

When the dual problem is feasible, as it certainly is in the present instance, set $a = \frac{1}{2}e$, strong duality implies that convergence tolerances can be based upon the so-called duality gap. Thus, if we have a primal feasible point b^* with associated objective function value, $\varphi_P(b^*) = \|y - Xb^*\|_1$, and dual feasible point a^* with associated dual objective function value, $\varphi_D(a^*) = y^\top a^*$, we are assured that b^* and a^* are ϵ-suboptimal in the sense that the proposed solutions are each within $\epsilon = \varphi_P(b^*) - \varphi_D(a^*)$ of the optimal solution of the problem. This crucial feature of linear programming problems carries over to virtually all of the problems we consider below.

Example 2.2. (Quantile regression) The extension of this median regression dual formulation to quantiles other than the median is remarkably simple: replacing $\frac{1}{2}$ by $1 - \tau$ in (3) yields an estimate of the coefficients of the τth conditional quantile function of y given x. Standard parametric programming methods enable one to find the entire solution path $\{(\hat{\alpha}(\tau), \hat{\beta}(\tau)) : \tau \in (0,1)\}$ very efficiently. As noted by Gutenbrunner and Jureckova (1992), the dual process $\hat{\alpha}$ constitutes an important link to the classical theory of rank tests as developed by Hájek.

Example 2.3. (Piecewise linear regression) An alternative form of the linear program for the ℓ_1 regression primal minimizes the sum of u_i that dominate, for each i, both $y_i - x_i^\top b$ and its negative. This can be generalized for any piecewise-linear objective function with k pieces,

$$\min_{b \in \mathbb{R}^n} \{e^\top u \mid c_j(y - X^\top b) + d_j \preceq u \text{ for } j = 1, \ldots, k\}.$$

Of course, the objective function still needs to be convex, which places the order requirement on the slopes. The examples include not only the quantile loss function, but also $\max\{0, |\cdot| - \epsilon\}$, corresponding to so-called ϵ-insensitive, or support vector regression of Vapnik (2000); it can be also viewed as a piecewise-linear approximation of the Huber function (see the example below).

If the regression likelihood is estimated via a nonparametric estimate that constrains the shape of the estimated density to be log-concave, then the resulting loss function is convex and piecewise linear. While the combined problem of estimating regression and residual density is not convex – a demonstration of an unfortunate fact that convexity is frequently lost in more structured problems – the convex task of piecewise linear regression can be bundled with a convex task of estimating a log-concave density in an alternating backfitting iteration scheme.

Simplex algorithms for computing solutions to (3) can be formulated as moving from one $b(h)$ to another, at each step removing the element of h that allows movement in the direction of steepest descent of the objective function, among the $2p$ available directions, and then introducing a new element of h by solving a one dimensional weighted quantile problem. See the discussion in Section 6.2 of Koenker (2005) for further details.

Simplex iterations move along the exterior edges of the constraint set toward a solution vertex. This iteration is usually quite efficient: indeed *why* simplex was so efficient became an important research problem in the 1970s. Interior point algorithms take another route: starting from the center of the constraint set they take a series of Newton-like steps based on a deformed version of the constraint set. In the most common implementatons, inequality constraints are replaced by log barrier (Lagrangian) penalty functions that introduce a smooth contribution to the objective function thus ensuring a proper Hessian. Relaxing the log-barrier penalty parameter defines a homotopy, the central path, that connects the initial central point to the vertex solution on the exterior of the constraint set. As described in more detail in Portnoy and Koenker (1997), for bounded variable LP problems like (3) interior point methods begin to dominate earlier simplex methods when sample sizes exceed 100,000 or so.

The R package **quantreg** (Koenker 2013). contains implementations based on the Barrodale and Roberts (1974) version of the simplex algorithm, as well as interior point implementations based on the Mehrotra (1992) predictor-corrector approach. Several variants of the latter approach are provided. Linear inequality constraints on β can be incorporated into (1) quite easily. When the parametric dimension, p, of β is large the Newton steps of the interior point iteration may become onerous, however in many cases this can be remedied by exploiting sparsity of the design matrix, X. The difficult aspect of the interior point computation involves solving linear equations for the Newton steps. When the column dimension of X is large, and X is sparse, these solutions are typically done by Cholesky decomposition of matrices of the form $X^\top WX$ with diagonal W. These matrices are themselves (usually) sparse, and Cholesky factorization is quite efficient with current algorithms as described in more detail in Koenker and Ng (2005). When problem sizes and sparsity fail to conform to these requirements one must resort to gradient descent methods about which we will have a little more to say in the final section.

2.2. Quadratic programming

Quadratic programming involves the minimization of a positive semi-definite quadratic objective function subject to polyhedral constraints. There are many applications of quadratic programming (QP) in statistics, typically involving Gaussian likelihoods constrained by some form of linear inequalities. The Markowitz mean-variance portfolio problem is perhaps the most prominent such example. Shape constrained regression examples have gained recent attention, and the introduction of sparse regularization methods like lasso, has greatly stimulated interest in computational methods for such problems. Before considering these examples we begin with a more classical one.

Example 2.4. (Huber's M-estimator) Inspired by piecewise-linear regression formulations, one can also formulate Huber's M-estimator as a *quadratic* program. The primal formulation minimizes $\sum_i \rho_\sigma(y_i - x_i^\top b)$, where ρ_σ is the Huber function (with tuning parameter σ)

$$\rho_\sigma(u) = \begin{cases} u^2/(2\sigma) & \text{when } |u| \le \sigma, \\ |u| - \frac{1}{2}\sigma & \phantom{\text{when }} |u| > \sigma \end{cases}$$

(a more common form in the literature is equal to $\sigma\rho_\sigma$, but both obviously give the same estimates). After the inclusion of slack variables the primal is again somewhat tricky, but the dual,

$$\max_{a \in \mathbb{R}^n} \{ \tfrac{1}{2}\sigma a^\top a + y^\top a \mid X^\top a = 0, a \in [-1,1]^n \},$$

exhibits striking similarity with the version of the dual of ℓ^1 regression obtained from (3) by $a \mapsto 2a - 1$ change of variable – in which the dual variables can be interpreted as (generalized) signs of residuals.

The role of duality in quadratic programming is closely aligned with the LP case. If we write the canonical primal QP as,

$$\min\{c^\top x + \tfrac{1}{2}x^\top Qx \mid Ax - b \in T,\ x \in S\}$$

where Q is a positive semi-definite matrix and again S and T denote convex cones. The dual becomes

$$\max\{b^\top z - \tfrac{1}{2}y^\top Qy \mid c + Q^\top y - A^\top z \in S^*,\ z \in T^*\},$$

where S^* and T^* denote the dual cones, defined as $K^* = \{y \mid x^\top y \geq 0,\ x \in K\}$.

Example 2.5. (Penalized quantile regression) We can illustrate this by considering a quantile regression problem subject to quadratic constraints (Gaussian penalty) on the coefficients,

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

Problems of this type arise in both parametric and nonparametric settings. Reversing the roles of the objective and constraint we could encompass various versions of the lasso. Proceeding as in the unconstrained case we can split the residual vector in its positive and negative parts and write $x = (b, u, v)$ in the primal formulation. All is as before except that we have the quadratic term with semi-definite Q, in fact Q is only non-zero in the upper diagonal p by p block S, which we will assume is invertible, so we can further simplify the problem to obtain the dual,

$$\max\{(y + (1-\tau)XS^{-1}X^\top e)^\top a - a^\top XS^{-1}X^\top a \mid a \in [0,1]^n\}$$

The objective function is an instance of a separable quadratic problem; in fact, any quadratic problem can be reformulated as such via Cholesky factorization expressing $x^\top Qx = x^\top F^\top Fx$ as a squared quadratic norm of Fx. The problem can be then rewritten as

$$\min\{c^\top x + \tfrac{1}{2} y^\top y \mid Ax - b \in T,\ Fx = y,\ x \in S\}$$

More constraints have been added, but the separability in y is usually advantageous. Typically, we have reduced the number of non-zeros required to represent the problem by the diagonalization and this reduces the computational effort of the solution.

Example 2.6. (Support vector machine with hinge loss) In the the field of statistical learning QP became a standard technology long ago: binary classification by support vector machines combines quadratic penalization with the so-called hinge loss function. The dual formulations is

$$\min\{\sum_i a_i - \tfrac{1}{2}\sum_{i,j} a_i a_j y_i y_j K(x_i, x_j) \mid y^\top a = 0, a \succeq 0\},$$

where y_i are ±1 indicators and $K(x_i, x_j)$ are the values of the positive definite function K (Mercer kernel) at the corresponding classifiers x_i. See Vapnik (2000). Many instances of this scheme can be found in the R package **kernlab** (Karatzoglou, Smola, Hornik, and Zeileis 2004).

2.3. Second-order cone programming

Modern QP implementations reformulate problems in terms of generalized inequalities employing second-order cone constraints. The QP domain is thus extended to a more general class of quadratic programs with quadratic constraints; the latter were not permitted in traditional QP formulations that allowed quadratic components only in the objective function. These so-called second order cone programs (SOCP) include constraints of the form

$$\|Ax + b\| \leq c^\top x + d,$$

which require that the affine function $(y,t) \equiv (Ax+b, c^\top x + d)$ lie in the second order cone defined as $K = \{(y,t) \mid \|y\| \le t\}$. A canonical formulation of SOCPs can be expressed as

$$\min_{x \in \mathbb{R}^n} \{c_0^\top x \mid A_0 x = b_0, \|A_i x + b_i\| \le c_i^\top x + d_i, i = 1, 2, \cdots, m\}.$$

Using cone constraints, quadratic problems can be formulated in “epigraph” form: the objective is linear, and all quadratic terms appear in the constraints. The form of the quadratic cone K as defined above suffices for all these needs; however, sometimes it may be more straightforward to use so-called rotated quadratic cones, defined as $K = \{(y,t,s) \mid \|y\|^2 \le 2ts\}$. Compare the implementations of the lasso and group lasso in the Appendix.

Example 2.7. (Random linear constraints) To illustrate such SOCP methods, suppose that we wish to minimize a linear objective function subject to random linear constraints in of the form:

$$\mathbb{P}(a_i^\top x \le b_i) \ge \eta \quad i = 1, \cdots, m,$$

where the vectors a_i are multivariate Gaussian with means α_i and covariance matrix Ω_i. Thus, our constraints may be rewritten as,

$$\alpha_i^\top x + \Phi^{-1}(\eta)\|\Omega^{1/2} x\| \le b_i,$$

so provided that $\eta \ge 1/2$ we have a SOCP.

2.4. Generalized inequalities and cone programming

Cone programming further expands the class of convex optimization problems and leads to many important applications. Following Boyd and Vandenberghe (2004), a cone, $K \in \mathbb{R}^n$ is *proper* if it is closed, convex, has non-empty interior, and is pointed, i.e., $x \in K$ and $-x \in K$ implies $x = 0$. Proper cones can be used to define partial orderings on $\mathbb{R}^n$ via generalized inequalities: $x \preceq_K y$ iff $y - x \in K$ and similarly, $x \prec_K y$ iff $y - x \in \text{int}(K)$.

An important example is the familiar partial order of positive semidefinite (psd) matrices. Let S^n denote the set of n-dimensional symmetric matrices, and $S_+^n \subset S^n$ the set of psd matrices. The set S_+^n is a convex cone since for any A and B in S_+^n and positive real numbers, θ_1 and θ_2,

$$x^\top(\theta_1 A + \theta_2 B)x = \theta_1 x^\top A x + \theta_2 x^\top B x \ge 0$$

for all $x \in \mathbb{R}^n$. S_+^n is a proper cone with interior consisting of the positive definite matrices, so we can write for $K = S_+^n$, $A \succeq_K B$ to mean $A - B$ is psd.

Generalized inequalities can be used to formulate conic optimization problems like,

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

As a special case, if $K = S_+^n$ we obtain semi-definite programs, SDP, of the form,

$$\min\{c^\top x \mid A_0 x = b, \sum_{i=1}^n x_i A_i \preceq_K B\}.$$

where B and $A_i : i = 1, \cdots, n$ are all symmetric matrices. Another subclass of conic problems are the SOCP’s, since we can write,

$$\min\{c_0^\top x \mid A_0 x = b, \ -(A_i x + b_i, c_i^\top x + d_i) \succeq_{K_i} 0, i = 1, \cdots, m\}.$$

for $K_i = \{(y,t) \in \mathbb{R}^{k_i+1} \mid \|y\| \leq t\}$.

Matrix norm optimization can also be cast into conic form. Suppose $A(x) = A_0 + \sum_{i=1}^{n} x_i A_i$ for $A_i \in \mathbb{R}^{p \times n}$ and we wish to minimize the spectral norm $\|A(x)\|$, the maximum singular value of $A(x)$. We can rewrite this as

$$\min\{a \in \mathbb{R} \mid A(x)^\top A(x) \preceq aI\}.$$

Such problems are closely related to recent work on matrix completion and robust principal component analysis. Suppose in the spirit of the recent Netflix competition, Feuerverger, He, and Khatri (2012), we could like to minimize,

$$f(A) = \sum_{(i,j) \in \mathcal{C}} |Y_{ij} - A_{ij}|$$

where $\mathcal{C}$ denotes the set of complete entries provided for the matrix, $A \in \mathbb{R}^{m \times n}$. Some form of regularization is necessary: motivated by lasso methods for regression, several authors have recently considered penalization of $f(A)$ by the nuclear norm $\|A\|_*$, i.e., the sum of of the singular values of A, yielding,

$$\min_{A \in \mathbb{R}^{m \times n}} \{f(A) \mid \|A\|_* \leq t/2\}$$

This formulation can be viewed as an attempt to find a minimal rank decomposition, $A = UV^\top$ to approximate the matrix A. This problem can be reformulated as a semi-definite program as follows: let $A = UV^\top$ with $U \in \mathbb{R}^{n \times r}$, $V \in \mathbb{R}^{m \times r}$ and

$$Z = \begin{pmatrix} U \\ V \end{pmatrix} (U^\top \; V^\top) = \begin{pmatrix} UU^\top & UV^\top \\ VU^\top & VV^\top \end{pmatrix} \equiv \begin{pmatrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \end{pmatrix}.$$

Then $f(A) = g(Z) \equiv f(Z_{12})$ and since $\|A\|_* = \min_{A=UV^\top} \frac{1}{2}(\mathrm{Tr}(UU^\top) + \mathrm{Tr}(VV^\top))$, we have the SDP,

$$\min\{g(Z) \mid Z \succeq 0, \mathrm{Tr}(Z) = t\}.$$

Linear inequality constraints are typically transformed by interior point methods into log barrier constraints; generalized inequality constraints require an analogous treatment. For this purpose it is convenient to define "generalized logarithms" for proper cones, $K \subset \mathbb{R}^n$. The function, $\lambda : \mathbb{R}^n \to \mathbb{R}$, is a generalized logarithm for K if (i) λ is closed, concave, twice continuously differentiable on $\mathrm{dom}(\lambda) = \mathrm{int}(K)$, and $\nabla^2 \lambda(x) \prec 0$ for $x \in \mathrm{int}(K)$, and (ii) there exists a constant $\theta > 0$ such that for all $x \succ_K 0$ and $y > 0$, $\lambda(yx) = \lambda(x) + \theta \log(y)$. When $K = \mathbb{R}^n_+$ we can use $\lambda(x) = \sum \log(x_i)$, while for the second order cone $K = \{(x_0, x) \in \mathbb{R}^{n+1} \mid \|x\| \leq x_0\}$ we can employ $\lambda(x_0, x) = \log(x_0^2 - \|x\|^2)$, and for the positive semi-definite cone $K = S^n_+$, we have $\lambda(A) = \log(|A|)$. For further details, consult Boyd and Vandenberghe (2004) and Nesterov and Nemirovskii (1987).

2.5. Separable convex programs

Interior point strategies for convex optimization crucially rely on sparsity of the relevant Hessian matrix. Given sparsity, Cholesky decomposition offers an efficient means of computing Newton type steps, without sparsity problems quickly become intractable as problem size grows. Another large class of sparse convex problems that is particularly well suited to statistical applications is the class of separable nonlinear programs of the form,

$$\min\{\sum \varphi(x_i) \parallel Ax = b,\ x \succeq_K 0\}.$$

Example 2.8. (Mixture models) An important instance of such problems is the classical Kiefer and Wolfowitz (1956) nonparametric maximum likelihood estimator for mixture models. Additive separability of the log likelihood ensures that the objective function contributes sparsely to the Hessian. Suppose we observe $X_1, \cdots, X_n$ drawn iid-ly from the mixture density,

$$g(x) = \int \varphi(x, \theta) dF(\theta),$$

where φ is a known parametric density and F is an unknown mixing distribution. Koenker and Mizera (2014) propose a discrete formulation of the Kiefer-Wolfowitz MLE as,

$$\min\{-\sum \log(g_i) \mid Af = g,\ f \in \mathcal{S}\}.$$

where A is an n by m matrix with typical element $(\varphi(x_i, \theta_j))$, with the θ_j defined on a relatively fine grid, and $\mathcal{S}$ denotes the unit simplex in $\mathbb{R}^m$. Convexity of $-\log(\cdot)$ implies that we have a convex object subject to linear equality and inequality constraints thereby ensuring a unique solution. This is quite remarkable in view of the notorious multimodality observed in most finite dimensional mixture settings. The dual to the original variational problem can be written as,

$$\max\{\sum_{i=1}^{n} \log(\nu_i) \mid \sum_{i=1}^{n} \nu_i \varphi(x_i, \theta) \leq n, \text{ for all } \theta \in \mathbb{R}\}.$$

and a discretized version of this dual has proven to be very efficient in a wide variety of mixture problems. Since Laird (1978) the EM algorithm has been used to compute the Kiefer-Wolfowitz estimator, however it is our experience that modern interior point methods are vastly superior, both in terms of accuracy and computational effort.

Example 2.9. (Penalized density estimation) Maximum penalized likelihood methods for density estimation also fall nicely into the separable convex programming framework. Good (1971) proposed a nonparametric maximum likelihood estimator of a univariate density subject to the smoothness constraint,

$$P(f) = \int ((\sqrt{f(x)})')^2 dx.$$

This penalty has the effect of shrinking $\hat{f}$ toward densities with minimal Fisher information for location. How natural this might be is an open question, and subsequent authors have focused on other roughness penalties imposed on the log density. For example, Silverman (1982) proposed the penalty,

$$P(f) = \int ((\log f(x))''')^2 dx,$$

which shrinks toward the normal density. Koenker and Mizera (2006) proposed total variation penalties of the form,

$$P(f) = TV((\log f)') = \int |(\log f)''| dx.$$

If we reparameterize the problem in terms of $h(x) = -\log f(x)$, we can write the variational form of the problem as

$$\min\{\sum h(x_i) \mid \int |h''(x)|dx \leq M, \int e^{-h(x)}dx = 1\}.$$

Again the problem can be discretized on a grid to enforce the integrability constraint and we are left with a strictly convex objective subject to linear constraints.

Regularization of likelihood methods for density estimation by such seminorm penalties has the feature that it require some form of λ-selection to control the degree of smoothness. This often imposes an outer layer of optimization over the problems described above, a layer whose non-convexity that (fortunately) places it outside the scope of the present survey. An alternative regularization strategy that has attracted substantial recent attention replaces the seminorm penalties by shape constraints. Imposing log-concavity, for example, is sufficient to regularize density estimation by maximum likelihood. Recent work by Cule, Samworth, and Stewart (2010), Dümbgen and Rufibach (2009), Koenker and Mizera (2010), and Seregin and Wellner (2010) have explored such methods and there is continuing work seeking to link this approach to more complicated semi-parametric models.

3. Convex optimization in R

Many *unconstrained* convex optimization problems in statistics can be very efficiently solved by some form of iteratively reweighted least squares. This class includes base R's `glm()` family and its many extensions. Linear equality constraints can generally be incorporated by some form of projected gradient method. However, inequality constraints, whether linear or nonlinear, pose new difficulties and require more sophisticated methods. We will briefly survey existing R functionality for such problems with apologies in advance for contributions we have overlooked.

In addition to the standalone R packages described below there are an increasing number of R packages that provide interfaces to independent convex optimization solvers. We will focus primarily on the **Rmosek** (Friberg 2014) interface to **MOSEK** (**MOSEK** ApS, Denmark 2011), available from the Comprehensive R Archive Network (CRAN) at `http://CRAN.R-project.org/`, but similar interfaces are available for **CPLEX** (IBM 2009), **CSDP** (Borchers 1999), and **Gurobi** (**Gurobi** Optimization, Inc. 2014). We emphasize that we are restricting attention to software intended for convex optimization; a broader overview of the optimization options in R is available from the task views on optimization (Theussl 2014) and machine learning (Hothorn 2014).

3.1. Linear programming

Most of the existing LP functionality in R is based on **lpSolve** Berkelaar *et al.* (2014) which in turn is based on the open source project **lp_solve** Berkelaar, Eikland, and Notebaert (2012), employing a revised simplex algorithm and supporting problem formulation in most standard formats. Alternative interior point methods are available via the **quantreg** package, like lpSolve, problems can be formulated to exploit sparsity of the constraint matrix. But problem formulations are currently confined to bounded variable forms that arise in the parametric and nonparametric quantile regression setting.

Example 3.1. (Penalized quantile regression) Standard quantile regression models can be estimated with the `rq()` function of the **quantreg** package. The syntax is similar to the `lm()` function for mean regression in base R, and associated inference apparatus is also similar: `summary()`, `anova()`, `predict()`, etc. The computational method for this estimation is controlled by the `method` argument. By default, on small problems a variant of the Barrodale and Roberts (1974) bounded variables simplex algorithm is used; on larger problems – by default for $n > 1000$ – the interior point algorithm `method = "fn"` is used. Linear inequality constraints of the general form $Rb = r$ can be introduced by specifying the `R` and `r` arguments to `rq()`. Model selection with the lasso penalty can be invoked with the `method = "lasso"`.

Nonparametric quantile regression methods are invoked with the `rqss()` function of **quantreg**, which can be used to fit additive smoothing spline models much like the **mgcv** package of Wood (2006) for mean models, except that the Sobolev penalties in **mgcv** are replaced by total variation penalties for `rqss()`. See Koenker (2011) for further details and examples.

Example 3.2. Another leading example of such problems is the Dantzig Selector of Candes and Tao (2007):

$$\min\{\|\beta\|_1 \mid \|X^\top(y - X\beta)\|_\infty < K\},$$

where $K = \lambda\sigma$ with σ being the scale of the noise. This problem is closely related to the standard lasso penalized mean regression problem as can be seen by noting that the vector appearing in the ℓ_∞ norm is just the usual gradient of the unpenalized least squares problem. It is easy to see that this can be reformulated as a linear inequality constrained ℓ_1 regression problem, and as along as X is reasonably sparse can be solved efficiently by the interior point methods available in the **quantreg** package. See the appendix for an implementation of this. This formulation is used in the R package **hdlm** (Arnold 2013) for high dimensional linear model estimation.

3.2. Quadratic programming

Several R packages include some QP functionality: the **quadprog** package of Turlach and Weingessel (2013), and the `ipop` function in the **kernlab** package of Karatzoglou *et al.* (2004) both provide interior point methods for a somewhat restrictive class of QPs. The function `solve_QP_SOCP` of the package **DWD** of Huang, Haaland, Lu, Liu, and Marron (2013) implements the full QP formulation via the SOCP algorithm of Toh, Todd, and Tutuncu (1999), used in **SDPT3**, the open-source software project for semidefinite programming (however, the interface to its full capabilities seems to be still missing in R). More general formulations are provided by the commercial developers. Several such applications using **Rmosek** will be described below.

Example 3.3. (The lasso) Perhaps the most familiar statistical QP application in recent times has been the lasso estimator of Tibshirani (1996) and Donoho, Chen, and Saunders (1998),

$$\hat{\beta} = \operatorname{argmin}\{\|y - Xb\|_2^2 + \lambda\|b\|_1\}.$$

In effect this is an attempt to approximate the solution of the combinatorial problem with non-convex penalty, $\|b\|_0 = \text{Card}\{b_i \neq 0\}$, by solutions for the convex ℓ_1 penalty. There are a variety of methods that have been proposed for computing $\hat{\beta}$, including the LARS approach of Efron, Hastie, Johnstone, and Tibshirani (2004). To illustrate the conic programming approach to such QPs, we rewrite the problem as

$$\min\{\lambda\|b\|_1 + t \mid v = y - Xb,\ \|(v,(t-1)/2)\|_2 \le (t+1)/2\}.$$

This equivalence follows immediately by noting that the constraint, $\|v\|_2^2 \le t$, is equivalent to the second-order cone constraint, $\|(v,(t-1)/2)\|_2 \le (t+1)/2$. An implementation of this formulation, as an **Rmosek** conic program, is given in the Appendix; the test problem was again adapted from Belloni, Chernozhukov, and Wang (2011).

Example 3.4. (Square-root lasso) Belloni *et al.* (2011) have recently proposed a modified version of the lasso in which the fidelity term of the objective function is replaced by its square root,

$$\hat{\beta} = \operatorname{argmin}\{\|y - Xb\|_2 + \lambda\|b\|_1\}.$$

This replacement, albeit numerically equivalent to lasso (with different λ, but in one-to-one correspondence), has several advantages, most notably it simplifies the choice of λ. In this form the conic problem is even simpler,

$$\min\{\lambda\|b\|_1 + t \mid v = y - Xb,\ \|v\|_2 \le t\},$$

and again we have illustrated an R implementation in the Appendix; the test problem is taken from Belloni *et al.* (2011).

Example 3.5. (Group lasso) For more structured problems, Bakin (1999), Yuan and Lin (2006), and others proposed the so-called group lasso, to achieve sparse selection not of individual covariates but rather their groups. Group lasso replaces the lasso's usual ℓ^1 penalty by a combined ℓ^1/ℓ^2 one, $\sum_j \|b_j\|_2$, where b_j are subvectors of b, the elements of a specific partitioning of the parameter vector b. The applications are manifest, e.g., in fitting ANOVA models, when the group consist of all parameters related to the given factor or interaction. Group lasso can be implemented as a special instance of SOCP, minimizing the square of quadratic norm of the residuals while either bounding the penalty or adding it multiplied by a Lagrange multiplier λ to the objective function. Example code for the latter is given in the Appendix; the test problem is once again adapted from Belloni *et al.* (2011).

Example 3.6. (Convex regression) Suppose we would like to estimate the nonparametric (mean) regression model,

$$y_i = g(x_i) + u_i, \qquad i = 1, \cdots, n,$$

subject to the constraint that the function $g : \mathbb{R} \to \mathbb{R}$ is convex. This is a QP of the form,

$$\min\{\tfrac{1}{2}\|g\|^2 - y^\top g \mid Dg \ge 0\},$$

where D denotes an n by $n-2$ matrix that computes second differences of the function values $g = g(x_i)$. There are a variety of reformulations of this problem that lend themselves to interior point methods. For **MOSEK**, it is convenient to introduce new variables, u and v and write the quadratic term as a rotated quadratic cone constraint,

$$\min\{v - y^\top g \mid Dg \ge 0, u = 1, 2uv \ge \sum g_i^2, v \in \mathbb{R}_+, g \in \mathbb{R}_+^n\}.$$

In this form the problem is easily implemented in **Rmosek**. We include illustrative code in the Appendix, tested on a problem taken from Groeneboom, Jongbloed, and Wellner (2008). The latter have proposed an alternative "support reduction" active-set algorithm; they compare performance of this algorithm with their coding of a log-barrier interior point method. For their test problems with $n = 10,000$ and equispaced x_i on $[-1, 1]$, they report timings of 0.24, 0.36, and 0.95 seconds for three variants with iid Gaussian u_i with standard deviations 1, 0.1, and 0.01 respectively. Their version of the interior point algorithm required, in contrast, about 4.5 seconds for each of these variants. Our implementation in **Rmosek** required 0.30, 0.31, and 0.26 seconds respectively, on a 2009 Mac Pro (2.93GHz). Corresponding timings for the support reduction algorithm on this machine were 0.12, 0.44 and 1.54.

In our experience **MOSEK** provides a very flexible, reliable platform for convex optimization. Relative performance of the commercial packages is a matter of some controversy, but for LP and QP problems it appears that **Gurobi** has a slight edge, see e.g., Mittleman (2012). We have no direct experience with either **CPLEX** or **Gurobi**, but at this point it is important to stress that they are both limited to LP and QP problems and related integer programming problems so they are not available for the separable convex optimization problems that we have described in the previous section. Finally, the realm of semidefinite programming was so far open to R users only via **Rcsdp** interface to **CSDP**; however, the new version 7 of **MOSEK** and **Rmosek**, still in beta stage, seems to open new possibilities.

3.3. Separable convex optimization

The options in R for separable convex optimization are much more limited than those for the prior categories. Again, **MOSEK** provides a quite general framework for such problems. The recent CVX Research, Inc. (2012) initiative advocating "disciplined convex programming" includes an interface to both **MOSEK** and **Gurobi**. Since additively separable convex problems play a natural role in many statistical applications. We will briefly describe two recent examples with which we have some personal experience.

Example 3.7. There is a growing literature on density estimation subject to shape constraints. An important example involves estimating log-concave densities: given a random sample, $x_1, \cdots, x_n$ from a distribution F, believed to have a log-concave density, we would like to solve,

$$\min\{-\sum \log g(x_i) \mid \int g = 1, \log g \text{ convex}\}.$$

Reparameterizing, so $h(x) = -\log g(x)$, the problem becomes,

$$\min\{\sum h(x_i) \mid \int e^{-h(x)} dx, h \text{ convex}\}.$$

Thus we have a linear objective, subject to a strictly convex nonlinear constraint, and linear inequalities defining a convex cone constraint. Various strategies for solving such problems have been proposed, see e.g., Cule *et al.* (2010), Dümbgen and Rufibach (2009) and Koenker and Mizera (2010). In the last of these we introduce auxiliary variables, undata, on relatively fine grid, to approximate the integral as a Riemann sum. This approach can be extended to bivariate densities, where the solutions take the form of piecewise linear $\hat{h}$ defined on triangulations of the observed points. Log-concavity is evidently a strong restriction, and similar methods can be adapted to replace the shape constraint with some form of norm constraint that penalizes roughness of the fitted density. Total variation of $\nabla \log g$ provides one potentially attractive possibility, as described by Koenker and Mizera (2006). One can also relax the log-concavity condition, which implicitly imposes exponential tails. Instead we might want to require concavity of $1/\sqrt{g}$, as discussed in Koenker and Mizera (2010) and Seregin and Wellner (2010). Methods of this type are available in the function `medde()` in the R package **REBayes** (Koenker 2014).

Example 3.8. Suppose we have iid observations from the mixture model,

$$g(x) = \int \varphi(x, \theta) dF(\theta),$$

where φ denotes a known parametric density and $\theta \in \Theta \subset \mathbb{R}$. Kiefer and Wolfowitz (1956) proposed estimating g by maximum likelihood,

$$\min_{F \in \mathcal{F}} \{ - \sum \log g(x_i) \mid g(x_i) = \int \varphi(x_i, \theta) dF(\theta), \qquad i = 1, \cdots, n \},$$

where $\mathcal{F}$ is the convex set of distribution functions on $\mathbb{R}$. Laird (1978) was apparently the first to propose a viable algorithm to implement the Kiefer-Wolfowitz MLE. However, the EM algorithm has proven to be frustratingly slow for problems of this type. Fortunately, they are easy to reformulate as separable convex optimization problems. Groeneboom *et al.* (2008) report on an example due originally to Richard Gill in which their support reduction method reduced a five hour EM computation to about five minutes. An equivalent problem formulated in the R package **REBayes** has been solved in less than five seconds. Mixture problems of this type play an important role in many empirical Bayes and hierarchical model problems and there are many opportunities to exploit improved methods of estimating the nonparametric MLE in these models.

4. What's next?

Up to this point we have emphasized the crucial role played by interior point methods in unifying the field of convex optimization. But already there are signs that this foundation may be crumbling. Massive new problems make the second order (Newton-type) methods underlying interior point impractical, and this has led researchers back to first order methods of the type pioneered by Shor and Nesterov. These methods can be seen as generalizations of familiar projected gradient methods for solving large ℓ_2 problems, adapted to the nonsmooth settings typically encountered with polyhedral constraints. Drawing on earlier work of Nesterov, Candès and coauthors have pioneered first-order methods for a variety of problems, see e.g., Becker, Bobin, and Candès (2011a), Cai, Candès, and Shen (2010), and Becker, Candès, and Grant (2011b). It would be desirable to see further development of these methods in R.

In the domain of interior point methods, there is still some scope for a "grand unification": the combination of separable nonlinearity with various types of conic constraints, in particular of the semidefinite type is still missing, even in commercial implementations. Such methods would find applications in multivariate density estimation with shape constraints, or in quadratically penalized logistic regression.

Acknowledgments

This research was partially supported by NSF grant SES-11-53548 and the NSERC of Canada. We thank the Associate Editor and three referees for helpful comments on the manuscript.

References

Arnold TB (2013). ***hdlm**: Fitting High Dimensional Linear Models.* R package version 1.2, URL http://CRAN.R-project.org/package=hdlm.

Bakin S (1999). *Adaptive Regression and Model Selection in Data Mining Problems.* Ph.D. thesis, ANU.

Barrodale I, Roberts FDK (1974). "Solution of an Overdetermined System of Equations in the ℓ_1 Norm." *Communications of the ACM*, **17**, 319–320.

Becker S, Bobin J, Candès EJ (2011a). "NESTA: A Fast and Accurate First-Order Method for Sparse Recovery." *SIAM Journal on Imaging Sciences*, **4**, 1–39.

Becker SR, Candès EJ, Grant MC (2011b). "Templates for Convex Cone Problems with Applications to Sparse Signal Recovery." *Mathematical Programming Computation*, **3**, 165–218.

Belloni A, Chernozhukov V, Wang L (2011). "Square-Root Lasso: Pivotal Recovery of Sparse Signals via Conic Programming." *Biometrika*, **98**, 791–806.

Berkelaar M, Eikland K, Notebaert P (2012). ***lp_solve**: Linear Programming System.* Version 5.5, URL http://lpsolve.sourceforge.net/5.5/.

Berkelaar M, *et al.* (2014). ***lpSolve**: Interface to **lp_solve** to Solve Linear-Integer Program.* R package version 5.6.10, URL http://CRAN.R-project.org/package=lpSolve.

Borchers B (1999). "**CSDP**, a C Library for Semidefinite Programming." *Optimization Methods and Software*, **11**(1), 613–623.

Boyd S, Vandenberghe L (2004). *Convex Optimization.* Cambridge University Press.

Cai JF, Candès EJ, Shen Z (2010). "A Singular Value Thresholding Algorithm for Matrix Completion." *SIAM Journal on Optimization*, **20**, 1956–1982.

Candes E, Tao T (2007). "The Dantzig Selector: Statistical Estimation When p is Much Larger Than n." *The Annals of Statistics*, **35**, 2313–2351.

Cule M, Samworth R, Stewart M (2010). "Computing the Maximum Likelihood Estimator of a Multidimensional Log-Concave Density." *Journal of the Royal Statistical Society B*, **72**, 545–600.

CVX Research, Inc (2012). "**CVX**: MATLAB Software for Disciplined Convex Programming, Version 2.0." URL http://cvxr.com/cvx/.

Dantzig GB (1951). "Maximization of a Linear Function of Variables Subject to Linear Inequalities." In TC Koopmans (ed.), *Activity Analysis of Production and Allocation*. John Wiley & Sons.

Donoho D, Chen SS, Saunders M (1998). "Atomic Decomposition by Basis Pursuit." *SIAM Journal of Scientific Computing*, **20**, 33–61.

Dümbgen L, Rufibach K (2009). "Maximum Likelihood Estimation of a Log-Concave Density: Basic Properties and Uniform Consistency." *Bernoulli*, **15**, 40–68.

Efron B, Hastie T, Johnstone I, Tibshirani R (2004). "Least Angle Regression." *The Annals of Statistics*, **32**, 407–499.

Feuerverger A, He Y, Khatri S (2012). "Statistical Significance of the Netflix Challenge." *Statistical Science*, **27**, 202–231.

Fiacco AV, McCormick GP (1968). *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*. Research Analysis Corp., McLean. Reprinted in SIAM Classics in Applied Mathematics, 1990.

Friberg HA (2014). ***Rmosek**: The R-to-**MOSEK** Optimization Interface*. R package version 1.2.5.1, URL http://CRAN.R-project.org/package=Rmosek.

Frisch R (1956). "La Résolution des Problèmes de Programme Linéaire par la Méthode du Potential Logarithmique." *Cahiers du Séminaire d'Econometrie*, **4**, 7–20.

Gill PE, Murray W, Saunders MA, Tomlin JA, Wright MH (1986). "On Projected Newton Barrier Methods for Linear Programming and an Equivalence to Karmarkar's Projective Method." *Mathematical Programming*, **36**(2), 183–209.

Good IJ (1971). "A Nonparametric Roughness Penalty for Probability Densities." *Nature*, **229**, 29–30.

Groeneboom P, Jongbloed G, Wellner JA (2008). "The Support Reduction Algorithm for Computing Non-Parametric Function Estimates in Mixture Models." *Scandinavian Journal of Statistics*, **35**, 385–399.

Gurobi Optimization, Inc (2014). ***Gurobi** Optimizer Reference Manual*. URL http://www.gurobi.com/.

Gutenbrunner C, Jureckova J (1992). "Regression Quantile and Regression Rank Score Process in the Linear Model and Derived Statistics." *The Annals of Statistics*, **20**, 305–330.

Hothorn T (2014). "CRAN Task View: Machine Learning & Statistical Learning." Version 2014-08-30, URL http://CRAN.R-project.org/view=MachineLearning.

Huang H, Haaland P, Lu X, Liu Y, Marron JS (2013). ***DWD**: DWD Implementation Based on A IPM SOCP Solver*. R package version 0.11, URL http://CRAN.R-project.org/package=DWD.

Huber PJ (2009). "On the Non-Optimality of Optimal Procedures." In *Optimality: The Third Erich L. Lehmann Symposium, IMS Lecture Notes-Monograph Series*, volume 57, pp. 31–46. Institute of Mathematical Statistics.

IBM (2009). *User Manual for* ***CPLEX***. URL `http://www.ibm.com/software/integration/optimization/cplex-optimizer/`.

Kantorovich LV (1939). *Mathematical Methods of Organizing and Planning Production.* Leningrad State University Publishers. Translation in *Management Science*, 6, 366–422 (1960).

Karatzoglou A, Smola A, Hornik K, Zeileis A (2004). "**kernlab** – An S4 Package for Kernel Methods in R." *Journal of Statistical Software*, **11**(9), 1–20. URL `http://www.jstatsoft.org/v11/i09/`.

Karlin S (1959). *Mathematical Methods and Theory in Games, Programming, and Economics.* Addison-Wesley, Reading.

Karmarkar NA (1984). "A New Polynomial-Time Algorithm for Linear Programming." *Combinatorica*, **4**, 373–395.

Khachiyan LG (1979). "A Polynomial Algorithm in Linear Programming." *Soviet Mathematics Doklady*, **20**, 191–194.

Kiefer J, Wolfowitz J (1956). "Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Incidental Parameters." *The Annals of Mathematical Statistics*, **27**, 887–906.

Koenker R (2005). *Quantile Regression.* Cambridge University Press.

Koenker R (2011). "Additive Models for Quantile Regression: Model Selection and Confidence Bandaids." *Brazilian Journal of Probability and Statistics*, **25**, 239–262.

Koenker R (2013). ***quantreg**: Quantile Regression.* R package version 5.05, URL `http://CRAN.R-project.org/package=quantreg`.

Koenker R (2014). ***REBayes**: Empirical Bayes Estimation and Inference in R.* R package version 0.50, URL `http://CRAN.R-project.org/package=REBayes`.

Koenker R, Mizera I (2006). "Density Estimation by Total Variation Regularization." In V Nair (ed.), *Advances in Statistical Modeling and Inference, Essays in Honor of Kjell A. Doksum.* World Scientific, Singapore.

Koenker R, Mizera I (2010). "Quasi-Concave Density Estimation." *The Annals of Statistics*, **38**(5), 2998–3027.

Koenker R, Mizera I (2014). "Convex Optimization, Shape Constraints, Compound Decisions, and Empirical Bayes Rules." *Journal of the American Statistical Association*, **109**(506), 674–685.

Koenker R, Ng P (2005). "A Frisch-Newton Algorithm for Sparse Quantile Regression." *Acta Mathematicae Applicatae Sinica*, **21**, 225–236.

Laird N (1978). "Nonparametric Maximum Likelihood Estimation of a Mixing Distribution." *Journal of the American Statistical Association*, **73**, 805–811.

Mehrotra S (1992). "On the Implementation of a Primal-Dual Interior Point Method." *SIAM Journal of Optimization*, **2**, 575–601.

Mittleman H (2012). "Mixed Integer Linear Programming Benchmark." URL http://plato.asu.edu/ftp/milpc.html.

MOSEK ApS, Denmark (2011). *The **MOSEK** Optimization Tools Manual.* Version 6.0, URL http://www.mosek.com/.

Nesterov Y, Nemirovskii A (1987). *Interior Point Polynomial Algorithms in Convex Programming.* Society for Industrial and Applied Mathematics.

Portnoy S, Koenker R (1997). "The Gaussian Hare and the Laplacian Tortoise: Computability of Squared-Error Versus Absolute-Error Estimators." *Statistical Science*, **12**, 279–300.

R Core Team (2014). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Rockafellar RT (1974). *Conjugate Duality and Optimization.* SIAM, Philadelphia.

Rockafellar RT (1996). *Convex Analysis.* Princeton University Press, Princeton.

Rockafellar T (1993). "Lagrange Multipliers and Optimality." *SIAM Review*, **35**, 183–238.

Seregin A, Wellner JA (2010). "Nonparametric Estimation of Multivariate Convex-Transformed Densities." *The Annals of Statistics*, **38**, 3751–3781.

Silverman BW (1982). "On the Estimation of a Probability Density Function by the Maximum Penalized Likelihood Method." *The Annals of Statistics*, **10**, 795–810.

Theussl S (2014). "CRAN Task View: Optimization and Mathematical Programming." Version 2014-08-08, URL http://CRAN.R-project.org/view=Optimization.

Tibshirani R (1996). "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society B*, **58**, 267–288.

Toh KC, Todd MJ, Tutuncu RH (1999). "**SDPT3** – A MATLAB Software Package for Semidefinite Programming." *Optimization Methods and Software*, **11**, 545–581.

Turlach BA, Weingessel A (2013). ***quadprog**: Functions to Solve Quadratic Programming Problems.* R package version 1.5-5, URL http://CRAN.R-project.org/package=quadprog.

Vapnik VN (2000). *The Nature of Statistical Learning Theory.* 2nd edition. Springer-Verlag, New York.

Wood SN (2006). *Generalized Additive Models: An Introduction with R.* Chapman & Hall/CRC.

Yuan M, Lin Y (2006). "Model Selection and Estimation in Regression with Grouped Variables." *Journal of Royal Statistical Society B*, **68**, 49–67.

A. The lasso

```
lasso <- function(x, y, sigma = 1, alpha = 0.05, c = 1.1, rtol = 1e-6,
  verb = 5)
{
  n <- nrow(x)
  p <- ncol(x)
  lambda <- c * sigma * 2 * sqrt(n) * qnorm(1 - alpha/(2*p))
  P <- list(sense = "min")
  P$c <- c(rep(lambda, 2*p), rep(0, n), 1, 0, 0)/n
  A <- as.matrix.csr(x)
  A <- cbind(A, -A, as(n, "matrix.diag.csr"), as.matrix.csr(0, n, 3))
  A <- rbind(A,cbind(as.matrix.csr(0, 2, 2*p + n),
    as.matrix.csr(c(-.5,-.5,1,0,0,1), 2, 3)))
  P$A <- as(A," CsparseMatrix")
  P$bc <- rbind(c(y, -0.5, 0.5), c(y, -0.5, 0.5))
  P$bx <- rbind(c(rep(0, 2 * p), rep(-Inf, n), rep(0, 3)),
                c(rep(Inf, 2 * p + n + 3)))
  P$cones <- matrix(list("QUAD",
    c(n + 2 * p + 3, (2 * p + 1):(2 * p + n), n + 2 * p + 2)), 2, 1)
  rownames(P$cones) <- c("type", "sub")
  P$dparam$intpnt_nl_tol_rel_gap <- rtol
  z <- mosek(P, opts = list(verbose = verb))
  status <- z$sol$itr$solsta
  f <- z$sol$itr$xx
  coef <- f[1:p] - f[(p + 1):(2 * p)]
  resid <- f[(2 * p + 1):(2 * p + n)]
  list(coef = coef, resid = resid, status = status)
}

library("SparseM")
library("Rmosek")
library("mvtnorm")
n <- 500
p <- 50
S <- 0.5^toeplitz(1:p)
X <- rmvnorm(n, sigma = S)
y <- apply(X[,1:5], 1, sum) + rnorm(n)
f <- lasso(X, y)
```

B. The square-root lasso

```
rooto <- function(x, y, alpha = 0.05, c = 1.1, rtol = 1e-6, verb = 5)
{
  n <- nrow(x)
  p <- ncol(x)
  lambda <- c * sqrt(n) * qnorm(1 - alpha/(2 * p))
  P <- list(sense = "min")
  P$c <- c(rep(lambda, 2 * p), rep(0, n), sqrt(n))/n
  A <- as.matrix.csr(x)
  A <- cbind(A, -A, as(n, "matrix.diag.csr"), as.matrix.csr(0, n, 1))
  P$A <- as(A, "CsparseMatrix")
  P$bc <- rbind(y, y)
  P$bx <- rbind(c(rep(0, 2 * p), rep(-Inf, n), 0),
    c(rep(Inf, 2 * p + n + 1)))
  P$cones <- matrix(list("QUAD",
    c(n + 2 * p + 1, (2 * p + 1):(2 * p + n))), 2, 1)
  rownames(P$cones) <- c("type", "sub")
  P$dparam$intpnt_nl_tol_rel_gap <- rtol
  z <- mosek(P, opts = list(verbose = verb))
  status <- z$sol$itr$solsta
  f <- z$sol$itr$xx
  coef <- f[1:p] - f[(p + 1):(2 * p)]
  resid <- f[(2 * p + 1):(2 * p + n)]
  list(coef = coef, resid = resid, status = status)
}

library("SparseM")
library("Rmosek")
library("mvtnorm")
n <- 500
p <- 50
S <- 0.5^toeplitz(1:p)
X <- rmvnorm(n, sigma = S)
y <- apply(X[, 1:5], 1, sum) + rnorm(n)
f <- rooto(X, y)
```

C. The group lasso

```
grupo <- function(x, y, alpha = 0.1, c = 1.1, rtol = 1e-6, verb = 5)
{
  n <- nrow(x)
  p <- ncol(x)/2
  lambda <- c * sqrt(n) * qnorm(1 - alpha/(2 * p))
  pr <- list(sense = "min")
  pr$c <- c(rep(0, 2*p), rep(lambda, p), rep(0, n), 1, 0)/n
  A <- as.matrix.csr(x)
  A <- cbind(A, as.matrix.csr(0, n, p),
    as(n, "matrix.diag.csr"), as.matrix.csr(0, n, 2))
  pr$A <- as(A, "CsparseMatrix")
  pr$bc <- rbind(y, y)
  pr$bx <- rbind(c(rep(-Inf, 2 * p), rep(0, p), rep(-Inf, n), 0, 0.5),
    c(rep(Inf, 3 * p + n + 1), 0.5))
  pr$cones <- matrix(list(), 2, p + 1)
  pr$cones[1,1:p] <- "QUAD"
  for (k in 1:p) pr$cones[2,k] <- list(c(2 * p + k, 2 * k - 1, 2 * k))
  pr$cones[1, (p + 1)] <- "RQUAD"
  pr$cones[2, (p + 1)] <- list(c(3 * p + n + 1, 3 * p + n + 2,
    (3 * p + 1):(3 * p + n)))
  pr$dparam$intpnt_nl_tol_rel_gap <- rtol
  z <- mosek(pr, opts = list(verbose = verb))
  status <- z$sol$itr$solsta
  f <- z$sol$itr$xx
  coef <- f[1:(2 * p)]
  resid <- f[(3 * p + 1):(3 * p + n)]
  list(coef = coef, resid = resid, status = status)
}

library("SparseM")
library("Rmosek")
library("mvtnorm")
n <- 500
p <- 25
S <- 0.5^toeplitz(1:p)
X <- rmvnorm(n, sigma = S)
X <- matrix(cbind(X, X^2), n)
y <- as.vector(X[, 1:10] %*% c(1, 1, 0, 1, 1, 0, 1, 1, 0.5, 0.5) + rnorm(n))
f <- grupo(X, y)
```

D. The Dantzig selector

```
DSelector <- function(X, y, sigma = 1, lambda = 3.5, sparse = TRUE)
{
  n <- nrow(X)
  p <- ncol(X)
  K <- lambda * sigma
  A <- t(X) %*% X
  R <- rbind(A, -A)
  a <- c(as.matrix(t(X) %*% y))
  r <- c(a - K, -a - K)
  zp <- rep(0, p)
  if(sparse){
    Ip <- as(p, "matrix.diag.csr")
    R <- as.matrix.csr(R)
    f <- rq.fit.sfnc(Ip, zp, R = R, r = r)
  }
  else{
    Ip <- diag(p)
    f <- rq.fit.fnc(Ip, zp, R = R, r = r)
  }
  return(f)
}

library("quantreg")
n <- 100
p <- 50
X <- matrix(rnorm(p*n), n, p)
X <- cbind(1, X)
y <- apply(X[, 1:5], 1, sum) + rnorm(n)
f <- DSelector(X, y)
```

E. Convex regression as a conic program

```
creg <- function(x, y, rtol = 1e-6, verb = 5) {
  n <- length(x)
  o <- order(x)
  x <- x[o]
  y <- y[o]
  P <- list(sense = "min")
  P$c <- c(0, 1, -y)
  P$A <- Diff2(x)
  P$bc <- rbind(rep(0, n - 2), rep(Inf, n - 2))
  P$bx <- rbind(c(1, 0, rep(-Inf,n)), c(1, rep(Inf, n + 1)))
  P$cones <- matrix(list("RQUAD", c(1:(n + 2))), 2, 1)
  rownames(P$cones) <- c("type", "sub")
  P$dparam$intpnt_nl_tol_rel_gap <- rtol
  z <- mosek(P, opts = list(verbose = verb))
  status <- z$sol$itr$solsta
  f <- z$sol$itr$xx[-(1:2)]
  S <- 2 * z$sol$itr$xx[2]
  list(f = f, S = S, status = status)
}
Diff2 <- function(x) {
  p <- length(x)
  h <- diff(x)
  s <- 1/h
  q <-  p-2
  ia <- c(1:(p - 2), 1:(p - 2), 1:(p - 2))
  ja <- c(1:(p - 2), 2:(p - 1), 3:p)
  xa <- c(-s[1:(p - 2)], s[1:(p - 2)] + s[2:(p - 1)], -s[2:(p - 1)])
  D  <- 0.5 * (h[1:(p - 2)]+h[2:(p - 1)])
  xa <- -xa / c(D, D, D)
  A <- new("matrix.coo", ra = xa, ja = as.integer(ja),
    ia = as.integer(ia), dimension = as.integer(c(p - 2, p)))
  A <- cbind(as.matrix.csr(0, q, 2), as.matrix.csr(A))
  as(A, "CsparseMatrix")
}

library("SparseM")
library("Rmosek")
n <- 10000
sd <- 0.1
x <- seq(-1, 1, length = n)
y <- x^2 + sd * rnorm(n)
f <- creg(x, y)
```

Affiliation:

Roger Koenker
Department of Economics
University of Illinois
Urbana, Illinois, 61801, United States of America
E-mail: rkoenker@uiuc.edu
URL: http://www.econ.uiuc.edu/~roger/

Ivan Mizera
Department of Mathematical and Statistical Sciences
University of Alberta
Edmonton, Alberta T6G 2G1, Canada
E-mail: imizera@ualberta.ca
URL: http://www.stat.ualberta.ca/~mizera/

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 60, Issue 5 *Submitted:* 2013-01-24
September 2014 *Accepted:* 2014-08-05