Taylor & Francis
Taylor & Francis Group

Communications in Statistics - Simulation and Computation

ISSN: 0361-0918 (Print) 1532-4141 (Online) Journal homepage: www.tandfonline.com/journals/lssp20

A Simple New Algorithm for Quadratic Programming with Applications in Statistics

Mary C. Meyer

To cite this article: Mary C. Meyer (2013) A Simple New Algorithm for Quadratic Programming with Applications in Statistics, Communications in Statistics - Simulation and Computation, 42:5, 1126-1139, DOI: 10.1080/03610918.2012.659820

To link to this article: https://doi.org/10.1080/03610918.2012.659820

Communications in Statistics—Simulation and Computation®, 42: 1126–1139, 2013

ISSN: 0361-0918 print / 1532-4141 online
DOI: 10.1080/03610918.2012.659820

Taylor & Francis
Taylor & Francis Group

A Simple New Algorithm for Quadratic Programming with Applications in Statistics

MARY C. MEYER

Department of Statistics, Colorado State University, Fort Collins, Colorado, USA

Problems involving estimation and inference under linear inequality constraints arise often in statistical modeling. In this article, we propose an algorithm to solve the quadratic programming problem of minimizing $\psi(\boldsymbol{\theta}) = \boldsymbol{\theta}'Q\boldsymbol{\theta} - 2\boldsymbol{c}'\boldsymbol{\theta}$ *for positive definite* Q*, where* $\boldsymbol{\theta}$ *is constrained to be in a closed polyhedral convex cone* $\mathcal{C} = \{\boldsymbol{\theta} : \boldsymbol{A}\boldsymbol{\theta} \geq \boldsymbol{d}\}$*, and the* $m \times n$ *matrix* $\boldsymbol{A}$ *is not necessarily full row rank. The three-step algorithm is intuitive and easy to code. Code is provided in the* R *programming language.*

Keywords Cone projection; Constrained parameter estimation; Dual algorithm; Inequality constraints; Polar cone; Primal-dual methods; Restricted least squares; Shape-restricted regression

Mathematics Subject Classification Primary 65K05; Secondary 62G05

1. Motivation and Background

Let $\boldsymbol{Q}$ be a positive definite $n \times n$ matrix, let $\boldsymbol{c}$ be a vector in $\mathbb{R}^n$, let $\boldsymbol{A}_0$ be an $m \times n$ irreducible matrix, and let $\boldsymbol{d}$ be a vector in the column space of $\boldsymbol{A}$. The term "irreducible" means "nonredundant" and will be defined formally in Sec. 2. The quadratic programming problem

$$\text{find } \hat{\boldsymbol{\theta}} \text{ to minimize } \boldsymbol{\theta}'Q\boldsymbol{\theta} - 2\boldsymbol{c}'\boldsymbol{\theta} \text{ subject to } \boldsymbol{A}_0\boldsymbol{\theta} \geq \boldsymbol{d} \tag{1}$$

may be readily transformed into a "cone projection" problem by considering the Cholesky decomposition $\boldsymbol{U}'\boldsymbol{U} = \boldsymbol{Q}$ and finding $\boldsymbol{\theta}_0$ such that $\boldsymbol{A}_0\boldsymbol{\theta}_0 = \boldsymbol{d}$. Then define $\boldsymbol{\phi} = \boldsymbol{U}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$, $\boldsymbol{z} = (\boldsymbol{U}^{-1})'(\boldsymbol{c} - \boldsymbol{Q}\boldsymbol{\theta}_0)$, and $\boldsymbol{A} = \boldsymbol{A}_0\boldsymbol{U}^{-1}$ to get

$$\text{find } \hat{\boldsymbol{\phi}} \text{ to minimize } \| \boldsymbol{z} - \boldsymbol{\phi} \|^2 \text{ subject to } \boldsymbol{A}\boldsymbol{\phi} \geq \boldsymbol{0}, \tag{2}$$

hence $\hat{\boldsymbol{\theta}} = \boldsymbol{U}^{-1}\hat{\boldsymbol{\phi}} + \boldsymbol{\theta}_0$. The set

$$\mathcal{C} = \{\boldsymbol{\phi} \in \mathbb{R}^n : \boldsymbol{A}\boldsymbol{\phi} \geq \boldsymbol{0}\} \tag{3}$$

is easily observed to be a polyhedral convex cone in $\mathbb{R}^n$, as each row of $\boldsymbol{A}$ defines a half-space, and $\mathcal{C}$ is the intersection of these half-spaces. Because the convex objective function is to be minimized over a convex set, a unique solution $\hat{\boldsymbol{\phi}}$ exists.

Received December 3, 2009; Accepted January 10, 2012
Address correspondence to Mary C. Meyer, Department of Statistics, Colorado State University, 102 Statistics Building, Fort Collins, CO 80523-1877, USA; E-mail: meyer@stat.colostate.edu

An early algorithm for cone projection was proposed by Dykstra (1983). Using that $\mathcal{C}$ is the intersection of half-spaces, it sequentially projects a residual vector onto the half-spaces, updating the residual vector at each step, and then repeats the process until convergence. The interior point algorithm for minimizing a quadratic function over a convex set is a gradient-based algorithm, first proposed by Karmarkar (1984) for linear programming. For more details, see Fang and Puthenpura (1993, chaps. 9 and 10). Fraser and Massam (1989) developed the mixed primal-dual bases algorithm for cone projections and applied it to concave nonparametric regression.

The interior point algorithm and the sequential projection algorithm are considered to converge in "infinitely many" steps, because the true solution is approached asymptotically and reached within a user-defined tolerance. In contrast, other algorithms that exploit the edges of the cone are guaranteed to produce the solution in a finite number of steps. These include the "nearest point" algorithm of Wilhelmsen (1976), the dual method of Goldfarb and Idnani (1983), the mixed primal-dual bases (MPDB) algorithm of Fraser and Massam (1989), and the nonnegative least squares (NNLS) algorithm of Lawson and Hanson (1974). The Goldfarb and Idnani algorithm is used in the `R` code `quadprog`; see http://CRAN.R-project.org/package=quadprog. The "hinge" algorithm proposed here can be summarized in three simple steps, and from a statistician's point of view, it is more intuitive than other methods. The derivations and proofs use results from linear models rather than optimization theory.

An example of an application in statistics is the least-squares regression problem with linear inequality constraints on the coefficients. Let $\boldsymbol{X}$ be an $n \times k$ fixed design matrix and $\boldsymbol{\beta}$ be a k-dimensional parameter vector, and consider the model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{4}$$

where $E(\boldsymbol{\epsilon}) = \mathbf{0}$ and $\mathrm{cov}(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{I}$. Suppose that it is reasonable to impose constraints on the parameter vector in the form $\boldsymbol{A}_0\boldsymbol{\beta} \geq \boldsymbol{d}$, where $\boldsymbol{A}_0$ is an $m \times k$ irreducible matrix. The problem of minimizing $\| \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} \|^2$ subject to the constraints is a quadratic programming problem with $\boldsymbol{Q} = \boldsymbol{X}'\boldsymbol{X}$ and $\boldsymbol{c} = \boldsymbol{X}'\boldsymbol{y}$. Methods for minimizing $\| \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} \|^2$ subject to various linear inequality and equality constraints were proposed by Judge and Takayama (1966); they used a simplex algorithm to obtain the optimal $\boldsymbol{\beta}$. Liew (1976) considered constraints $\boldsymbol{A}\boldsymbol{\beta} \geq \boldsymbol{d}$, and used the Dantzig–Cottle "principal pivoting" algorithm and provided an approximate variance-covariance matrix for $\hat{\boldsymbol{\beta}}$. Hawkins (1994) considered the problem of fitting monotonic polynomials to data using a primal-dual algorithm to adjust the coefficients iteratively. An exposition of estimation and testing under linear inequality constraints can be found in chapter 21 of Gourieroux and Monfort (1995), chapter 4 of Silvapulle and Sen (2005), and chapter 2 of Robertson et al. (1988).

Another traditional use of constrained estimation and inference is for nonparametric regression with shape restrictions. Suppose that interest is in fitting a scatterplot generated from

$$y_i = f(x_i) + \epsilon_i, \quad i = 1, \ldots, n, \tag{5}$$

where a parametric form is unknown for f, but it may be assumed that f is nondecreasing. Defining $\theta_i = f(x_i)$, the problem is to find $\boldsymbol{\theta} \in \mathbb{R}^n$ to minimize $\| \boldsymbol{y} - \boldsymbol{\theta} \|^2$ given $\boldsymbol{A}\boldsymbol{\theta} \geq \mathbf{0}$, where the nonzero elements of the $(n-1) \times n$ matrix $\boldsymbol{A}$ are $A_{i,i} = -1$ and $A_{i,i+1} = 1$, $i = 1, \ldots, n-1$. This model was first considered by Brunk (1955) who proposed the pooled adjacent violators algorithm (PAVA) to obtain the solution. For estimation, assuming f is convex, the nonzero elements of the $(n-2) \times n$ constraint matrix are $A_{i,i} = t_{i+2} - t_{i+1}$,

$A_{i,i+1} = t_i - t_{i+2}$, and $A_{i,i+2} = t_{i+1} - t_i$. Fraser and Massam (1989) proposed the MPDB algorithm to solve the concave regression problem.

This article introduces a simple, three-step algorithm for cone projection that is fast and intuitive. The code, shown in the `R` programming language in Appendix A, is only a couple dozen lines. The speed is important for applications in iteratively reweighted cone projection algorithms, such as for generalized constrained regression, and in estimating the mixing distribution used for the tests of $H_0 : \boldsymbol{A\theta} = \boldsymbol{b}$ versus $H_0 : \boldsymbol{A\theta} > \boldsymbol{b}$ [see Gouriéroux et al. (1982), Raubertas et al. (1986), or Silvapulle and Sen (2005, chap. 4), or Robertson et al. (1988)]. The `R` package `ic.infer` supports inequality-constrained estimation and testing, and is described in Grömping (2010). The rest of the article is organized as follows. Some necessary theory about cones and projections onto cones is outlined in the next section, which is used to explain the steps of the algorithm and also used in the convergence results (Appendix B). This is meant to be accessible to people with some background in linear spaces, who are not necessarily familiar with convex optimization theory. For more details and proofs of the claims, see Silvapulle and Sen (2005, chap. 3), and van Eeden (2006) also provides theory for constrained estimation. The proposed cone projection ("hinge") algorithm is described in Sec. 3, and some useful applications in statistics are given in Sec. 4.

2. The Constraint Cone and Its Polar Cone

The convex polyhedral cone (3) is considered for an $m \times n$ irreducible constraint matrix $\boldsymbol{A}$, where "irreducible" means that no row of $\boldsymbol{A}$ is a positive linear combination of other rows, and there is not a positive linear combination of rows of $\boldsymbol{A}$ that equals the zero vector. Note that if $\boldsymbol{A}$ is full row rank, then it is irreducible. If a row is a positive linear combination of other rows, it can be removed without affecting the problem, and if the origin can be written as a positive linear combination of rows, then there is an implicit equality constraint in the matrix $\boldsymbol{A}$, which can be dealt with separately. Silvapulle and Sen (2005) use the term "tight" to describe this nonredundancy of $\boldsymbol{A}$.

Let m_1 be the dimension of the space spanned by the rows of A. If m_1 is less than n, then the cone contains a linear space V of dimension $n - m_1$; this is the null space of A. Let $\Omega = \mathcal{C} \cap V^{\perp}$, where $V^{\perp}$ is the linear space orthogonal to V. Then Ω is a polyhedral convex cone that does not contain a linear space of dimension one or greater, and sits in an m_1-dimensional subspace of $\mathbb{R}^n$. The "edges" or "generators" of Ω are vectors $\boldsymbol{\delta}^1, \ldots, \boldsymbol{\delta}^{m_2}$ in Ω such that

$$\Omega = \left\{ \boldsymbol{\phi} \in \mathbb{R}^n : \boldsymbol{\phi} = \sum_{i=1}^{m_2} b_j \boldsymbol{\delta}^j, \quad \text{where} \quad b_j \geq 0, \quad j = 1, \ldots, m_2 \right\},$$

and hence

$$\mathcal{C} = \left\{ \boldsymbol{\phi} \in \mathbb{R}^n : \boldsymbol{\phi} = \boldsymbol{v} + \sum_{i=1}^{m_2} b_j \boldsymbol{\delta}^j, \quad \text{where} \quad b_j \geq 0, \quad j = 1, \ldots, m_2 \quad \text{and} \quad \boldsymbol{v} \in V \right\}. \quad (6)$$

If $m_1 = m$, then $m_2 = m$ and the edges of Ω are the columns of the matrix $\Delta = A'(AA')^{-1}$. For the case of more constraints than dimensions ($m > m_1$), the edges of the cone are obtained as follows (see Meyer, 1999, for proof). Define $\boldsymbol{\gamma}^1, \ldots, \boldsymbol{\gamma}^m$ to be the rows of $-\boldsymbol{A}$. Suppose $J \subseteq \{1, \ldots, m\}$ and let $S = \mathcal{L}(\{\boldsymbol{\gamma}^j, j \in J\})$, where $\mathcal{L}$ denotes "the space spanned by." If $\dim(S) = m_1 - 1$, then $S^{\perp} \cap V^{\perp}$ is a line through the origin containing the vectors $\boldsymbol{\delta}$ and $-\boldsymbol{\delta}$, say, where $\boldsymbol{\delta} \perp \boldsymbol{\gamma}^j$ for all $j \in J$. If $\langle \boldsymbol{\delta}, \boldsymbol{\gamma}^i \rangle \leq 0$ for all $i \notin J$, then $\boldsymbol{\delta}$ is an edge of

$\boldsymbol{\Omega}$. Conversely, all edges are of this form. In the case of more constraints than dimensions, the number m_2 of edges may be considerably larger than m.

Because V is orthogonal to Ω, $\hat{\boldsymbol{\phi}}$ is the sum of the projections of $\boldsymbol{y}$ onto V and Ω so that $\hat{\boldsymbol{\phi}}$ can be written as $\boldsymbol{v} + \Delta\boldsymbol{b}$, where $\boldsymbol{v} \in V$, the columns of $\boldsymbol{\Delta}$ are $\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_m$, and $\boldsymbol{b} \geq 0$. Note that this representation is not necessarily unique if there are more constraints than dimensions, but $\hat{\boldsymbol{\phi}}$ is unique. Necessary and sufficient conditions for $\hat{\boldsymbol{\phi}} \in \mathcal{C}$ to minimize $\| \boldsymbol{y} - \boldsymbol{\phi} \|^2$ are $\langle \boldsymbol{y} - \hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\phi}} \rangle = 0$ and $\langle \boldsymbol{y} - \hat{\boldsymbol{\phi}}, \boldsymbol{\phi} \rangle \leq 0$, for all $\boldsymbol{\phi} \in \mathcal{C}$; because the $\boldsymbol{\delta}$ vectors are generators of the cone, these conditions may be written as

$$\langle \boldsymbol{y} - \hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\phi}} \rangle = 0; \quad \langle \boldsymbol{y} - \hat{\boldsymbol{\phi}}, \boldsymbol{v} \rangle \leq 0 \ \text{ for all } \ \boldsymbol{v} \in V; \quad \text{and} \quad \langle \boldsymbol{y} - \hat{\boldsymbol{\phi}}, \boldsymbol{\delta}^j \rangle \leq 0 \ \text{ for } \ j = 1, \ldots, m_2. \tag{7}$$

Proposition 2.1. *Let $\hat{\boldsymbol{\phi}}$ be the unique minimizer of $\| \boldsymbol{y} - \boldsymbol{\phi} \|^2$ over $\boldsymbol{\phi} \in \mathcal{C}$, and write $\hat{\boldsymbol{\phi}} = \boldsymbol{v} + \Delta\boldsymbol{b}$ for $\boldsymbol{b} \geq 0$ and $\boldsymbol{v} \in V$. Let $J \subseteq \{1, \ldots, m\}$ index the nonzero elements of $\boldsymbol{b}$; that is, $j \in J$ if $b_j > 0$. Then $\hat{\boldsymbol{\phi}}$ is the projection of $\boldsymbol{y}$ onto the linear space spanned by $\boldsymbol{\delta}^j$, $j \in J$, and the basis vectors for* V.

Hence, the projection of a vector onto a polyhedral convex cone is the ordinary least-squares projection onto the linear space spanned by a basis for V and a subset of cone edges indexed by a subset J of $\{1, \ldots, m\}$. The projection lands on a *face* of the cone; the faces can be defined as

$$\mathcal{F}_J = \left\{ \boldsymbol{\phi} \in \mathbb{R}^n : \boldsymbol{\phi} = \boldsymbol{v} + \sum_{j \in J} b_j \boldsymbol{\delta}_j, \quad b_j > 0, \quad \boldsymbol{v} \in V \right\}.$$

Note that $\mathcal{F}_\emptyset = V$ (where $\emptyset$ denotes the empty set), and the interior of the constraint cone is the face where $J = \{1, \ldots, m\}$.

The set of points whose projection onto $\mathcal{C}$ lands in V (or whose projection onto Ω is the origin) is also called the *polar cone*. The edges of the polar cone are $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_m$, that is, the rows of $-\boldsymbol{A}$, and the polar cone may be seen to be the collection of vectors in $V^\perp$ that make obtuse angles with all vectors in $\mathcal{C}$. The next proposition implies that the polar cone is analogous to the orthogonal space in the linear model.

Proposition 2.2. *Let $\hat{\boldsymbol{\rho}}$ be the projection of $\boldsymbol{y}$ onto the polar cone, and let $\hat{\boldsymbol{\phi}}$ be the projection of $\boldsymbol{y}$ onto the constraint cone. Then $\hat{\boldsymbol{\phi}} + \hat{\boldsymbol{\rho}} = \boldsymbol{y}$.*

Therefore, the projection onto the constraint cone may alternatively be found via projection onto the polar cone, and that the polar cone is analogous to the orthogonal space in the linear model.

3. The Hinge Algorithm

The hinge algorithm for cone projection uses either the constraint cone edges $\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_{m_2}$ or the polar cone edges $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_m$, so the cone can be specified either as (3) or as (6). In the following, the $\boldsymbol{\delta}$ vectors are used, but the algorithm is identical if the $\boldsymbol{\gamma}$ vectors are substituted, except that the residual of the projection onto the constraint cone is obtained. This is useful for problems with more constraints than dimensions, as the number of constraint cone edges can be large and finding them may be computationally intensive.

Proposition 1 tells us that the minimizer of (2) subject to $\boldsymbol{A\phi} \geq \mathbf{0}$ with irreducible $m \times n$ $\boldsymbol{A}$ can be solved by finding $J \subseteq \{1, \ldots, m\}$ where $\hat{\boldsymbol{\phi}}$ lands on $\mathcal{F}_J$. The hinge algorithm arrives at the appropriate set J through a series of guesses J_k. At a typical iteration, the current estimate $\boldsymbol{\phi}^k$ is the least-squares regression of $\boldsymbol{z}$ on the space spanned by $\boldsymbol{\delta}^j$, for $j \in J_k$. (The $\boldsymbol{\delta}^j$, $j \in J$, were originally called "hinges" since in the convex regression problem, for which the algorithm was initially devised, the $(t_j, \hat{\phi}_j)$, $j \in J$, are the points at which the line segments in the piecewise linear fit change slope, and the bends are allowed to go only one way.) The initial guess J_0 can be any subset of $\{1, \ldots, m_2\}$ for which the corresponding $\boldsymbol{\delta}^j$, $j \in J$, form a linearly independent set. The hinge algorithm can be summarized in three steps. At the kth iteration:

1. Project $\boldsymbol{z}$ onto the linear space spanned by $\{\boldsymbol{\delta}^j, j \in J_k\}$, to get $\boldsymbol{\phi}^k = \sum_{j \in J_k} b_j^{(k)} \boldsymbol{\delta}_j$.
2. Check to see if $\boldsymbol{\phi}^k$ satisfies the constraints, that is, if all $b_j^{(k)}$ are nonnegative:
   - If yes, go to step 3.
   - If no, choose j for which $b_j^{(k)}$ is smallest, and remove it from the set; go to step 1.
3. Compute $\langle \boldsymbol{y} - \boldsymbol{\phi}^k, \boldsymbol{\delta}^j \rangle$ for each $j \notin J_k$. If these are all nonpositive, then stop. If not, choose j for which this inner product is largest, add it to the set, and go to step 1.

Intuitively, at each stage, the new edge is added where it is "most needed," and other edges are removed if the new fit does not satisfy the constraints. Although the $\boldsymbol{\delta}^j$, $j = 1, \ldots, m_2$, might not form a linearly independent set, at each step, the $\boldsymbol{\delta}^j$, $j \in J_k$ are linearly independent. Suppose that at step 2, we have that J_k defines a linearly independent set of $\boldsymbol{\delta}^j$ vectors. For all vectors $\boldsymbol{\delta}^j$ such that the indices $J_k \cup \{j\}$ do not define a linearly independent set, we have $\langle \boldsymbol{y} - \boldsymbol{\phi}^k, \boldsymbol{\delta}^j \rangle = 0$, so these j are not added. Since the stopping criteria are defined by (7), it is clear that if the algorithm ends, it gives the correct solution. The only thing that requires proof is that the algorithm does end, that is, it does not produce the same set of edges twice, which would result in an infinite loop. The proofs are deferred to Appendix B.

4. Applications in Statistics

4.1. *Constrained Least-Squares Regression*

Consider the model (4) where it is known that $\boldsymbol{A\beta} \geq \mathbf{0}$ for irreducible $\boldsymbol{A}$. Let $\hat{\boldsymbol{\beta}}$ be the constrained least-squares estimator for $\boldsymbol{\beta}$, and let $\tilde{\boldsymbol{\beta}}$ be the unconstrained estimator $(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$. The conditions (7) lead to $(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'\boldsymbol{X}\hat{\boldsymbol{\beta}} = 0$ and $(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'\boldsymbol{X}\boldsymbol{\beta} \leq 0$ for all $\boldsymbol{\beta}$ such that $\boldsymbol{A\beta} \geq \mathbf{0}$. Now

$$\| \boldsymbol{X}\tilde{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta} \|^2 = \| \boldsymbol{X}\tilde{\boldsymbol{\beta}} - \boldsymbol{X}\hat{\boldsymbol{\beta}} \|^2 + \| \boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta} \|^2 + 2(\boldsymbol{X}\tilde{\boldsymbol{\beta}} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}).$$

The last term is

$$(\boldsymbol{X}\tilde{\boldsymbol{\beta}} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}) - (\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}})'(\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}),$$

where the second term on the right is zero by principles of ordinary least-squares regression, and the first is positive. Hence, if the true $\boldsymbol{\beta}$ satisfies the constraints, the constrained estimate of $E(\boldsymbol{y})$ is closer to the truth than the unconstrained estimate, with equality only if the two estimates coincide.

A well-known example is the simple linear regression model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, where the slope β_1 must be nonnegative. The constrained least-squares estimate is easy to obtain: it is the ordinary least-squares estimate unless that does not provide a positive slope,

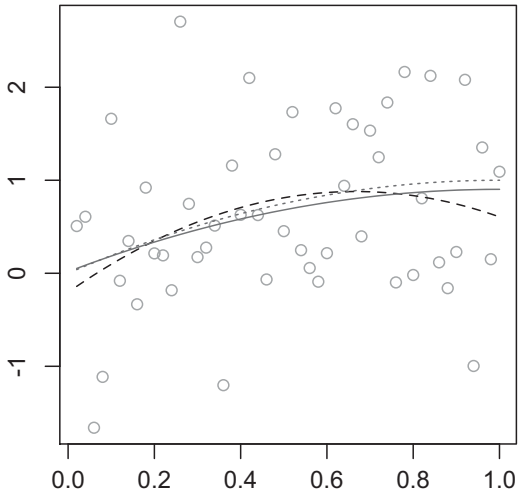

Figure 1. Quadratic fits to a scatterplot simulated using $f(x) = 1 - (1 - x)^2$ (shown as the dotted curve) and $n = 50$ observations. The dashed curve is the unconstrained fit, and the solid curve is constrained to be nondecreasing.

in which case $\hat{\beta}_1 = 0$ and $\hat{\beta}_0 = \bar{y}$. Now suppose that the researcher wishes to allow some curvature in the regression function, so the quadratic model $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$ is considered, retaining the assumption that the expected value of y is nondecreasing in x. The constraint is $\beta_1 + 2\beta_2 x \geq 0$ over the range of the data, so if $x \in [0, 1]$, these can be written as $\boldsymbol{A}\boldsymbol{\beta} \geq \mathbf{0}$ with

$$\boldsymbol{A} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 1 & 2 \end{pmatrix},$$

and $\hat{\boldsymbol{\beta}}$ is obtained by using the hinge algorithm. To illustrate, the scatterplot of $n = 50$ points shown in Fig. 1 was generated using equally spaced x_i, the regression function $f(x_i) = 1 - (1 - x_i)^2$, and independent standard normal errors. Least-squares quadratic fits are shown, where the dashed curve represents the unconstrained fit and the solid curve is constrained to be monotone. The dotted curve is the true regression function.

Another important example is the "warped plane" model $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + \epsilon_i$, where $E(y)$ is constrained to be increasing in both variables. The constraints are $\beta_1 + \beta_3 x_{2i} \geq 0$, for all x_{2i}, and $\beta_2 + \beta_3 x_{1i} \geq 0$, for all x_{1i}. If both predictors are confined to the unit interval, this provides four constraints on three parameters. Constrained and unconstrained fits are shown in Fig. 2 for data simulated from the surface $f(x_1, x_2) = x_1 x_2$, unit error variance, and $n = 100$, with the predictor values forming a grid in the unit intervals.

4.2. *Monotone Regression*

There is an elegant closed-form solution for the monotone regression problem provided by Brunk (1955):

$$\hat{\theta}_i = \min_{v \geq i} \max_{u \leq i} \frac{1}{v - u + 1} \sum_{j=u}^{v} y_j.$$

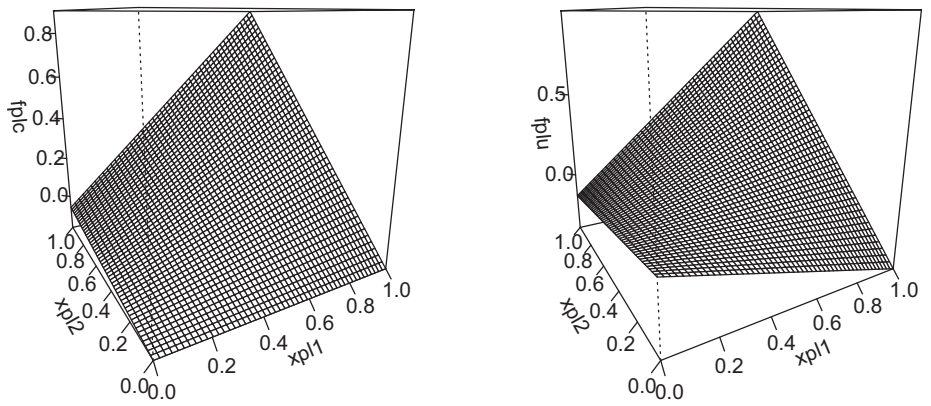

Figure 2. (a) Warped plane fit to dataset $y_i = x_1 x_2 + \epsilon_i$, where the expected value of y is constrained to be increasing in both predictors. (b) Unconstrained fit.

PAVA is a well-known method for finding this solution; see Silvapulle and Sen (2005, p. 47) for more details. The PAVA algorithm begins by setting $\theta = y$, $l = 1$, and $u = 2$. Then the following steps are performed in a loop: if $\theta_l \le \theta_u$, then $l = u$, $u = u + 1$. Else, the θ_i, $i = l, \ldots, u$, are set to the average of the y_i, for $i = l, \ldots, u$, and l and u are updated as follows. If $l > 1$, then $l = l - 1$, but if $l = 1$ then $u = u + 1$. The loop ends when $u > n$.

The monotone regression problem is an interesting application of the hinge algorithm when the edges of the constraint cone are used. During the implementation, no hinge indices are removed from the sets J_k at any iteration, so the number of iterations is the number of jumps in the monotone regression estimator. The $\boldsymbol{\delta}$ vectors can be written as $\delta_i^j = j - n$, for $i = 1, \ldots, j$, and $\delta_i^j = j$, for $i = j + 1, \ldots, n$, for $j = 1, \ldots, m$, and V is spanned by the one vector. Suppose at some iteration, we have $J_k = \{j_1, \ldots, j_p\}$, where the elements of J_k are ordered so that $j_1 < \cdots < j_p$. Then it is easily observed that for $1 \le l \le p$,

$$\bar{y} + b_{j_1} + \cdots + b_{j_l} = \frac{1}{j_{l+1} - j_k} \sum_{i=j_l}^{j_{l+1}-1} y_i.$$

It can be shown that the first hinge index, say l, is chosen so that for any $j_1 < l$ and $j_2 > l$,

$$\frac{1}{l - j_1} \sum_{i=j_1}^{l-1} y_i \le \frac{1}{j_2 - l} \sum_{i=l}^{j_2-1} y_i.$$

This means that the coefficient on $\boldsymbol{\delta}^l$ remains positive after any subsequent additions of edges, and a similar argument can be applied to coefficients of those edges, so all coefficients remain positive throughout the implementation of the algorithm, and their indices are never removed from the sets J_k. From any J, the projection of $\boldsymbol{y}$ onto the face $\mathcal{F}_J$ is easily found by averaging the y values between $j \in J$.

4.3. *Shape-Restricted Regression Splines*

The monotone regression estimator is a step function, and the convex regression estimator is piecewise linear—neither is satisfactory if f known to be smooth. For model (5), a

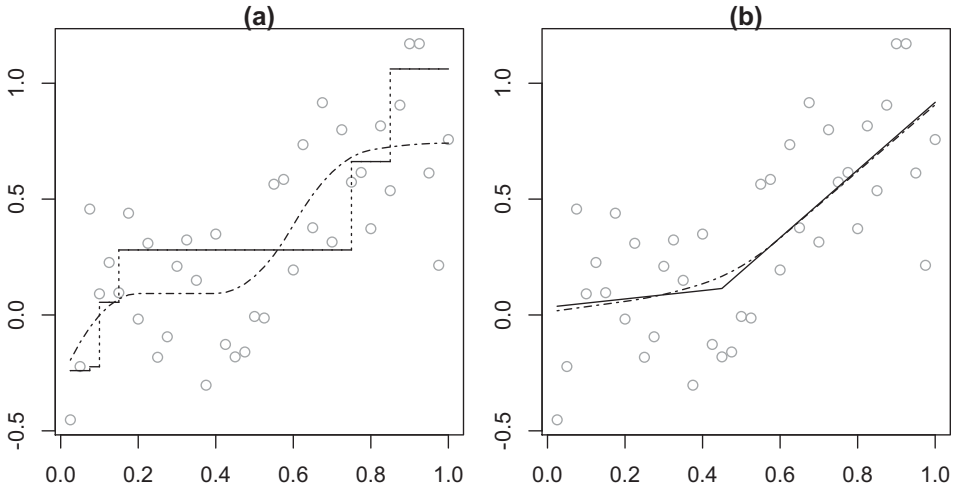

Figure 3. (a) Unsmoothed least-squares monotone fits a dataset (step function) and quadratic monotone regression spline fit (dot–dash curve). (b) Least-squares convex fit (solid) and cubic convex spline (dot–dash). Both spline fits use three interior knots.

method for estimating f with smoothed monotone regression using I-splines was given by Ramsay (1988) using regression splines, and Meyer (2008) extended the method to convex restrictions using C-splines. An alternative method for either is outlined here using the more familiar B-splines (see de Boor, 2001, for details). Given a set of k distinct knots over the range of the x values, a set of m piecewise degree-p spline basis functions are to be defined that span the space of such piecewise polynomials. The spline basis vectors contain the values of these functions at the observed x_i; let the columns of the matrix $\boldsymbol{X}$ contain the basis vectors. The unconstrained fit is obtained by minimizing $\| \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} \|^2$ over $\beta \in \mathbb{R}^m$.

A quadratic ($p = 2$) spline function is increasing if and only if it is increasing at the knots; hence, if A_{ij} is the slope of the jth spline basis function at the ith knot, the monotone spline estimator minimizes the sum of squared residuals subject to $\boldsymbol{A}\boldsymbol{\beta} \geq \boldsymbol{0}$. Similarly, a cubic spline function is convex if and only if it is convex at the knots, and the cubic B-spline basis functions may be used with A_{ij} equal to the second derivative of the jth spline basis function at the ith knot. Examples of constrained fits to a scatterplot are shown in Fig. 3. In Fig. 3(a), the unsmoothed monotone regression is shown as the piecewise constant function, and the monotone quadratic spline with four interior knots is shown as the dot–dash curve. In Fig. 3(b), the unsmoothed convex regression function is shown as the solid piecewise linear function, and the convex cubic regression spline is the dot–dash curve. For either monotone or convex constraints, we can demonstrate that when the constraints hold, the constrained version of the estimator provides smaller squared error loss, by an argument similar to that for the parametric case (see Meyer, 2008).

4.4. *Weighted and Iteratively Reweighted Least Squares*

Constrained estimation for the regression models (4) or (5) may be accomplished when the errors have an arbitrary positive definite covariance matrix, through weighted regression. Specifically, for $\boldsymbol{y} = \boldsymbol{\theta} + \sigma\boldsymbol{\epsilon}$ with $\boldsymbol{A}\boldsymbol{\theta} \geq \boldsymbol{0}$, suppose $\mathrm{cov}(\boldsymbol{\epsilon}) = \boldsymbol{\Sigma}$. The model equation is multiplied through by $\boldsymbol{L}^{-1}$, where $\boldsymbol{L}\boldsymbol{L}' = \boldsymbol{\Sigma}$, to get $z = \boldsymbol{\phi} + \sigma\boldsymbol{\xi}$ where $z = \boldsymbol{L}^{-1}\boldsymbol{y}$, $\boldsymbol{\phi} =$

$\boldsymbol{L}^{-1}\boldsymbol{\theta}$, and cov($\boldsymbol{\xi}$) is the identity matrix. The new constraint matrix is $\boldsymbol{AL}$. The projection is accomplished using the transformed model, and the reverse transformation provides the solution.

More general constrained maximum-likelihood estimation problems may be solved through iteratively reweighted cone projections. For example, we consider a constrained generalized regression model, where the response is a vector $\boldsymbol{y}$ of independent observations from a distribution written in the form of an exponential family:

$$f(y_i) = \exp\{[y_i\theta_i - b(\theta_i)]/\tau^2 - c(y_i, \tau)\},$$

where the specifications of b and c determine the subfamily of models. See Silvapulle (1994) for examples including one-sided testing problems. Common examples are $b(\theta) = \log(1 + e^{\theta})$ for the Bernoulli and $b(\theta) = \exp(\theta)$ for the Poisson model. The log-likelihood function

$$\ell(\boldsymbol{\theta}, \tau) = \sum_{i=1}^{n} \left[\frac{y_i\theta_i - b(\theta_i)}{\tau^2} \right]$$

is to be maximized over appropriate constraints on $\boldsymbol{\theta}$.

The vector $\boldsymbol{\mu} = E(\boldsymbol{y})$, can be observed to be $\mu_i = b'(\theta_i)$, and the variance of y_i is $b''(\theta_i)\tau$. The mean vector is related to the predictor variables through a link function $g(\mu_i) = \eta_i$. If $\boldsymbol{\eta} = \boldsymbol{X\beta}$, constraints of the form $\boldsymbol{A\beta} \geq \boldsymbol{d}$ may be considered. Further, if the link function is one-to-one, then monotonicity constraints imposed on the η values also constrain the mean function to be monotone. Constrained regression splines may be used to model $\boldsymbol{\eta}$, under assumptions of monotonicity and smoothness. In any of these applications, the log-likelihood function is maximized over $\boldsymbol{\eta} \in \mathcal{C}$ where $\mathcal{C}$ is of the form (6). The algorithm involves iteratively reweighted cone projections and follows the same ideas for the generalized linear model as found in McCullagh and Nelder (1989). Starting with $\boldsymbol{\eta}^0 \in \mathcal{C}$, the estimate $\boldsymbol{\eta}^{k+1}$ is obtained from $\boldsymbol{\eta}^k$ by constructing $\boldsymbol{z}$:

$$z_i = \eta_i^k + \left(y_i - \mu_i^k\right)\left(\frac{d\eta}{d\mu}\right)_{ki},$$

where $\mu_i^k = g^{-1}(\eta_i^k)$ and the derivative of the link function is evaluated at μ_i^k. The weighted projection of $\boldsymbol{z}$ onto $\mathcal{C}$ is obtained with weight vector $\boldsymbol{w}$, where $1/w_i^k = (d\eta/d\mu)_k^2 V_k$, and V_k is the variance function evaluated at μ_i^k. This scheme can be shown to converge to the value of $\boldsymbol{\eta}$ that maximizes ℓ over $\boldsymbol{\eta} \in \mathcal{C}$. The proof is similar to that of theorem 1 in Meyer and Woodroofe (2004).

To illustrate, we use a data-set given by Ruppert et al. (2003), concerning incidence of bronchopulmonary dysplasia in $n = 223$ low birth weight infants. Suppose experts believe that the probability of the condition is a smooth, decreasing function of birth weight. The data are shown in Fig. 4 as tick marks at one for infants suffering from the condition and at zero otherwise. The ordinary logistic regression is shown as the dashed curve, but perhaps there is no reason to believe that the assumption of linear log odds holds. The unconstrained spline estimator with five equally spaced knots is shown as the dotted curve; this is unsatisfactory because it violates the assumption of decreasing probability. The constrained spline estimator using the same knots is shown as the solid curve; this shows a steeper descent than the fit assuming linear log odds, and a leveling off above zero. If the only valid assumptions are that the probability of the condition is a smooth, decreasing function of birth weight, the more flexible nonparametric fit may be preferred.

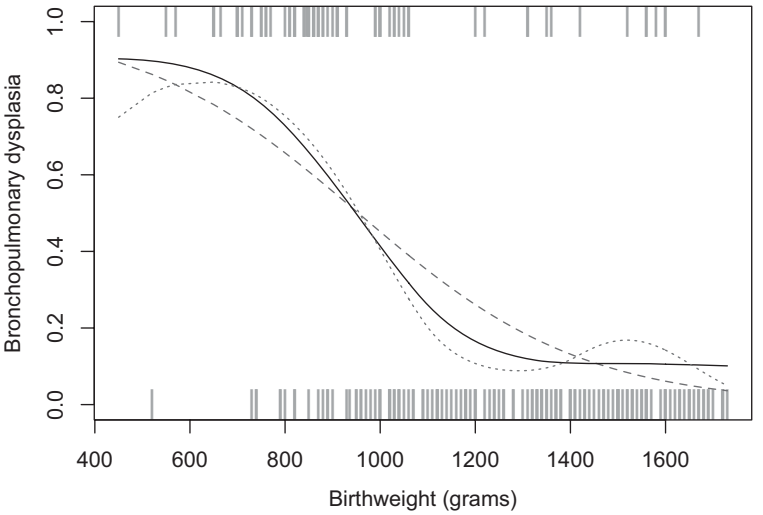

Figure 4. The estimated probability of bronchopulmonary dysplasia in low birth weight infants, as a function of birth weight. The dashed curve is the ordinary logistic regression, the dotted curve is the unconstrained regression spline estimate, and the solid curve is the monotone decreasing regression spline estimate.

5. Discussion

Constrained estimation is useful for both parametric and nonparametric function estimation. When constraints are valid, their use reduces squared error loss of the estimates and increases power of the tests for significance of the predictor. The hinge algorithm is intuitively simple and fast. It can be readily understood using ideas from linear models and does not require a background in optimization theory. The code provided in Appendix A is short and uses only basic matrix operations. More specific code for the other applications listed in this article, including the general quadratic programming problem, constrained parametric regression, constrained spline regression, and the generalized regression models, can be found at http://www.stat.colostate.edu/~meyer/code.htm.

Acknowledgment

This work was partially supported by NSF DMS 0905656.

Appendix A: Code

The hinge algorithm to solve the generic cone projection problem of minimizing $\| \boldsymbol{y} - \boldsymbol{\theta} \|^2$ subject to $\boldsymbol{A\theta} \geq \boldsymbol{0}$ is provided here in `R` code, where `amat` is the irreducible constraint matrix $\boldsymbol{A}$. Specifically, $\boldsymbol{y}$ is projected onto the polar cone to get $\hat{\boldsymbol{\rho}}$, and Proposition 2.2 is used to get $\hat{\boldsymbol{\theta}}$.

```
coneproj=function(y,amat){
    n=length(y);m=length(amat)/n
    sm=1e-8;h=1:m<0;obs=1:m;check=0
    delta=-amat;b2=delta%*%y
    if(max(b2)>sm){
```

```
            i=min(obs[b2==max(b2)])
            h[i]=TRUE
        }else{check=1;theta=1:n*0}
        while(check==0){
            xmat=matrix(delta[h,],ncol=n)
            a=solve(xmat%*%t(xmat))%*%xmat%*%y
            if(min(a)<(-sm)){
                avec=1:m*0;avec[h]=a
                i=min(obs[avec==min(avec)])
                h[i]=FALSE;check=0
            }else{
                check=1
                theta=t(xmat)%*%a
                b2=delta%*%(y-theta)
                if(max(b2)>sm){
                    i=min(obs[b2==max(b2)])
                    h[i]=TRUE;check=0}}
            }
    return(y-theta)}
```

Appendix B: Proofs

It is clear that if the hinge algorithm ends, it gives the correct solution. The algorithm ends because it does not choose the same set of edges twice. The sum of squared errors (SSE) $\| z - \hat{\boldsymbol{\phi}} \|^2$ decreases for subsequent iterations with the same number of edges, so the algorithm cannot produce an infinite loop. First we show that the simplest type of loop does not occur.

Proposition B.1. *The algorithm does not remove the edge that it just added.*

Proof. Suppose at the start of some iteration of the algorithm, the set of edges is J_k, and at the end it is $J_{k+1} = J_k \cup \{l\}$, so $\boldsymbol{\delta}^l$ is the most recently added edge. The coefficient of $\boldsymbol{\delta}^l$ is

$$b_l = \frac{\langle z, \tilde{\boldsymbol{\delta}}^l \rangle}{\| \tilde{\boldsymbol{\delta}}^l \|^2},$$

where $\tilde{\boldsymbol{\delta}}^l$ is the residual from the regression of $\boldsymbol{\delta}^l$ on the other regressors $\{\boldsymbol{\delta}^j, j \in J_k\}$. The numerator of the right-hand side is equivalent to $\langle z - \hat{\boldsymbol{\phi}}^k, \boldsymbol{\delta}^l \rangle$, because of orthogonality:

$$\langle z, \tilde{\boldsymbol{\delta}}^l \rangle = \langle z - \hat{\boldsymbol{\phi}}^k, \tilde{\boldsymbol{\delta}}^l \rangle = \langle z - \hat{\boldsymbol{\phi}}^k, \boldsymbol{\delta}^l \rangle > 0.$$

The first equality is because $\hat{\boldsymbol{\phi}}^k \perp \tilde{\boldsymbol{\delta}}^l$ and the second because $\boldsymbol{\delta}^l - \tilde{\boldsymbol{\delta}}^l \perp z - \hat{\boldsymbol{\phi}}$.

The idea for proving that the algorithm stops is to show that the SSE at a given iteration with n_h edges is less than that of the last iteration with n_h edges. Suppose that at the beginning of some iteration of the algorithm, we have the solution:

$$\hat{\boldsymbol{\phi}}^B = \sum_{j \in \boldsymbol{J}_B} b_j^B \boldsymbol{\delta}^j,$$

and that this solution satisfies the constraints. Suppose that it is not optimal and the algorithm adds the vector $\boldsymbol{\delta}^l$ to the set of regressors. The least-squares fit produced is then

$$\hat{\boldsymbol{\phi}}^M = \sum_{j \in \boldsymbol{J}_B} b_j^M \boldsymbol{\delta}^j + b_l^M \boldsymbol{\delta}^l.$$

Further, suppose that this $\hat{\boldsymbol{\phi}}^M$ does not satisfy the constraints, so b_i^M, say, is negative. The algorithm will then remove $\boldsymbol{\delta}^i$ from the set of regressors in step 2 and go to step 1 to refit the data. The next proposition shows that the new solution

$$\hat{\boldsymbol{\phi}}^N = \sum_{j \in \boldsymbol{J}_N} b_j^N \boldsymbol{\delta}^j,$$

where $J_N = J_B \cup \{l\} - \{i\}$, has $\mathrm{SSE}(\hat{\boldsymbol{\phi}}^N) < \mathrm{SSE}(\hat{\boldsymbol{\phi}}^B)$.

Proposition B.2. *If the algorithm replaces a edge, then the SSE after is less than the SSE before.*

Proof. Let

$$\tilde{\boldsymbol{\delta}}^l = \boldsymbol{\delta}^l - \sum_{j \in \boldsymbol{J}_B} \alpha_j^l \boldsymbol{\delta}^j,$$

where the second term is the projection of $\boldsymbol{\delta}^l$ onto the space spanned by $\{\boldsymbol{\delta}^j, j \in J_B\}$. Then, $\tilde{\boldsymbol{\delta}}^l \perp \hat{\boldsymbol{\phi}}^B$, so we can write

$$\begin{aligned} \hat{\boldsymbol{\phi}}^M &= \hat{\boldsymbol{\phi}}^B + b_l^M \tilde{\boldsymbol{\delta}}^l \\ &= \sum_{j \in \boldsymbol{J}_B} \left(b_j^B - \alpha_j^l b_l^M\right) \boldsymbol{\delta}^j + b_l^M \boldsymbol{\delta}^l. \end{aligned}$$

We know that $b_i^B > 0$, since $\hat{\boldsymbol{\phi}}^B$ satisfies the constraints, and $b_l^M > 0$, by Proposition B.1. Further,

$$b_i^M = b_i^B - \alpha_i^l b_l^M < 0,$$

so $\alpha_i^l > 0$. Let

$$\boldsymbol{\phi}(x) = \sum_{j \in \boldsymbol{J}_B} \left(b_j^B - \alpha_j^l x\right) \boldsymbol{\delta}^j + x \boldsymbol{\delta}^l.$$

Note that $\boldsymbol{\phi}(0) = \hat{\boldsymbol{\phi}}^B$ and $\boldsymbol{\phi}(b_l^M) = \hat{\boldsymbol{\phi}}^M$. When $x = b_i^B/\alpha_i^l$, the coefficient of $\boldsymbol{\delta}^i$ in $\boldsymbol{\phi}(x)$ disappears. Further, $0 < b_i^B/\alpha_i^l < b_l^M$, since $b_i^B - \alpha_i^l 0 > 0$ and $b_i^B - \alpha_i^l b_l^M < 0$. Since b_l^M minimizes $\| z - \boldsymbol{\phi}(x) \|^2$, we have

$$\| z - \boldsymbol{\phi}(0) \|^2 > \| z - \boldsymbol{\phi}\left(\frac{b_i^B}{\alpha_i^l}\right) \|^2 > \| z - \boldsymbol{\phi}(b_l^M) \|^2 .$$

Further,

$$\| z - \hat{\boldsymbol{\phi}}^N \|^2 < \| z - \boldsymbol{\phi} \left(\frac{b_i^B}{\alpha_i^l} \right) \|^2$$

since $\hat{\boldsymbol{\phi}}^N$ is the least-squares fit with the same regressors. So, finally,

$$\| z - \hat{\boldsymbol{\phi}}^N \|^2 < \| z - \hat{\boldsymbol{\phi}}^B \|^2 .$$

The proof that the sum of squares decreases if *two* edges are replaced is similar in idea but more tedious, and can be found at http://www.stat.colostate.edu/~meyer/quadprog.htm.

References

Brunk, H. D. (1955). Maximum likelihood estimates of monotone parameters. *Annals of Mathematical Statistics* 26(4):607–616.

de Boor, C. (2001). *A Practical Guide to Splines.* Rev. ed. New York: Springer.

Dykstra, R. J. (1983). An algorithm for restricted least squares regression. *Journal of the American Statistical Association* 78:837–842.

Fang, S. C., Puthenpura, S. (1993). *Linear Optimization and Extensions. Theory and Algorithms*. Englewood Cliffs NJ: Prentice Hall.

Fraser, D. A. S., Massam, H. (1989). A mixed primal-dual bases algorithm for regression under inequality constraints. Application to convex regression. *Scandinavian Journal of Statistics* 16:65–74.

Goldfarb, D., Idnani, A. (1983). A numerically stable dual method for solving strictly convex quadratic programs. *Mathematical Programming* 27:1–33.

Gouriéroux, C., Holly, A., Monfort, A. (1982). Likelihood ratio test, Wald test, and Kuhn-Tucker test in linear models with inequality constraints on the regression parameters. *Econometrica* 50(1):63–80.

Gourieroux, C., Monfort, A. (1995). *Statistics and Econometric Models*. Cambridge: Cambridge University Press.

Grömping, U. (2010). Inference with linear equality and inequality constraints using R: The package ic.infer. *Journal of Statistical Software* 33(10):1–31.

Hawkins, D. M. (1994). Fitting monotonic polynomials to data. *Computational Statistics* 9:233–247.

Judge, G. G., Takayama, T. (1966). Inequality restrictions in regression analysis. *Journal of the American Statistical Association* 61:166–181.

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

Lawson, C. L., Hanson, R. J. (1974). *Solving Least Squares Problems*. Englewood Cliffs NJ: Prentice Hall.

Liew, C. K. (1976). Inequality constrained least-squares estimation. *Journal of the American Statistical Association* 71:746–751.

McCullagh, P., Nelder, J. A. (1989). *Generalized Linear Models.* 2nd ed. New York: Chapman & Hall.

Meyer, M. C. (1999). An extension of the mixed primal-dual bases algorithm to the case of more constraints than dimensions. *Journal of Statistical Planning and Inference* 81:13–31.

Meyer, M. C. (2008). Inference using shape-restricted regression splines. *Annals of Applied Statistics* 2(3):1013–1033.

Meyer, M. C., Woodroofe, M. (2004). Estimation of a unimodal density using shape restrictions. *Canadian Journal of Statistics* 32(1):85–100.

Ramsay, J. O. (1988). Monotone regression splines in action. *Statistical Science* 3(4):425–461.

Raubertas, R. F., Lee, C. I. C., Nordheim, E. V. (1986). Hypothesis tests for normal means constrained by linear inequalities. *Communications in Statistics—Theory and Methods* 15(9):2809–2833.

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

Ruppert, D., Wand, M. P., Carroll, R. J. (2003). *Semiparametric Regression (Cambridge Series in Statistical and Probabilistic Mathematics)*. Cambridge: Cambridge University Press.

Silvapulle, M. J. (1994). On tests against one-sided hypotheses in some generalized linear models. *Biometrics* 50(3):853–858.

Silvapulle, M. J., Sen, P. K. (2005). *Constrained Statistical Inference*. New York: Wiley.

van Eeden, C. (2006). *Restricted Parameter Space Estimation Problems*. New York: Springer.

Wilhelmsen, D. R. (1976). A nearest point algorithm for convex polyhedral cones and applications to positive linear approximation. *Mathematics of Computation* 30(133):48–57.