Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Sharp quadratic majorization in one dimension

Jan de Leeuw[c,*], Kenneth Lange[a,b,c]

[a] Department of Biomathematics, University of California, Los Angeles, CA 90095, United States
[b] Department of Human Genetics, University of California, Los Angeles, CA 90095, United States
[c] Department of Statistics, University of California, Los Angeles, CA 90095, United States

ARTICLE INFO

Article history:
Received 7 November 2007
Received in revised form 4 January 2009
Accepted 6 January 2009
Available online 22 January 2009

ABSTRACT

Majorization methods solve minimization problems by replacing a complicated problem by a sequence of simpler problems. Solving the sequence of simple optimization problems guarantees convergence to a solution of the complicated original problem. Convergence is guaranteed by requiring that the approximating functions majorize the original function at the current solution. The leading examples of majorization are the EM algorithm and the SMACOF algorithm used in Multidimensional Scaling. The simplest possible majorizing subproblems are quadratic, because minimizing a quadratic is easy to do. In this paper quadratic majorizations for real-valued functions of a real variable are analyzed, and the concept of sharp majorization is introduced and studied. Applications to logit, probit, and robust loss functions are discussed.

1. Introduction

Majorization algorithms, including the EM algorithm, are used for more and more computational tasks in statistics (De Leeuw, 1994; Heiser, 1995; Hunter and Lange, 2004; Lange et al., 2000). The basic idea is simple. A function g majorizes a function f at a point y if $g \geq f$ and $g(y) = f(y)$. If we are minimizing a complicated objective function f iteratively, then we construct a majorizing function at the current best solution $x^{(k)}$. We then find a new solution $x^{(k+1)}$ by minimizing the majorization function. Then we construct a new majorizing function at $x^{(k+1)}$, and so on.

Majorization algorithms are worth considering if the majorizing functions can be chosen to be much easier to minimize than the original objective function, for instance linear or quadratic. In this paper we will look in more detail at majorization with quadratic functions. We restrict ourselves to functions of a single real variable. This is not as restrictive as it seems, because many functions $F(x_1, \ldots, x_n)$ in optimization and statistics are *separable* in the sense that

$$F(x_1, \ldots, x_n) = \sum_{i=1}^{n} f_i(x_i),$$

and majorization of the univariate functions f_i automatically gives a majorization of F.

Many of our results generalize without much trouble to real-valued functions on $\mathbb{R}^n$ and to constrained minimization over subsets of $\mathbb{R}^n$. The univariate context suffices to explain most of the basic ideas.

2. Majorization

2.1. Definitions

We formalize the definition of majorization at a point.

[*] Corresponding author. Tel.: +1 310 825 9550; fax: +1 310 206 5658.
E-mail address: deleeuw@stat.ucla.edu (J. de Leeuw).

0167-9473/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2009.01.002

Definition 2.1. Suppose f and g are real-valued functions on R^n. We say that g *majorizes* f *at* y if

- $g(x) \geq f(x)$ for all x,
- $g(y) = f(y)$.

If the first condition can be replaced by

- $g(x) > f(x)$ for all $x \neq y$,

we say that majorization is *strict*.

Thus g majorizes f at y if $d = g - f$ has a minimum, equal to zero, at y. And majorization is strict if this minimizer is unique. If g majorizes f at y, then f *minorizes* g at y. Alternatively we also say that f *supports* g at y.

It is also useful to have a global definition, which says that f can be majorized at all y.

Definition 2.2. Suppose f is a real-valued function on R^n and g is a real-valued function on $\mathrm{R}^n \otimes \mathrm{R}^n$. We say that g *majorizes* f if

- $g(x, y) \geq f(x)$ for all x and all y,
- $g(x, x) = f(x)$ for all x.

Majorization is *strict* if the first condition is

- $g(x, y) > f(x)$ for all $x \neq y$.

2.2. Majorization algorithms

The basic idea of majorization algorithms is simple. Suppose our current best approximation to the minimum of f is $x^{(k)}$, and we have a g that majorizes f in $x^{(k)}$. If $x^{(k)}$ already minimizes g we stop, otherwise we update $x^{(k)}$ to $x^{(k+1)}$ by minimizing g. If we do not stop, we have the *sandwich inequality*

$$f(x^{(k+1)}) \leq g(x^{(k+1)}) < g(x^{(k)}) = f(x^{(k)}),$$

and in the case of strict majorization

$$f(x^{(k+1)}) < g(x^{(k+1)}) < g(x^{(k)}) = f(x^{(k)}).$$

Repeating these steps produces a decreasing sequence of function values, and appropriate additional compactness and continuity conditions guarantee convergence of both sequences $x^{(k)}$ and $f(x^{(k)})$ (De Leeuw, 1994). In fact, it is not necessary to actually minimize the majorization function; it is sufficient to have a continuous update function h such that $g[h(y)] < g(y)$ for all y. In that case the sandwich inequality still applies with $x^{(k+1)} = h(x^{(k)})$.

2.3. Majorizing differentiable functions

We first show that majorization functions must have certain properties at the point where they touch the target.

Theorem 2.1. *Suppose f and g are differentiable at y. If g majorizes f at y, then*

- $g(y) = f(y)$,
- $g'(y) = f'(y)$.

If f and g are twice differentiable at y, then in addition

- $g''(y) \geq f''(y)$.

Proof. If g majorizes f at y then $d = g - f$ has a minimum at y. Now use the familiar necessary conditions for the minimum of a differentiable function, which say the derivative at the minimum is zero and the second derivative is non-negative. ■

Theorem 2.1 can be generalized in many directions if differentiability fails. If f has a left and right derivatives in y, for instance, and g is differentiable, then

$$f_R'(y) \leq g'(y) \leq f_L'(y).$$

If f is convex, then $f_L'(y) \leq f_R'(y)$, and $f'(y)$ must exist in order for a differentiable g to majorize f at y. In this case $g'(y) = f'(y)$. For nonconvex f more general differential inclusions are possible using the four Dini derivatives of f at y [see, for example, McShane (1944) Chapter V].

3. Quadratic majorizers

As we said, it is desirable that the subproblems in which we minimize the majorization function are simple. One way to guarantee this is to try to find a *convex quadratic majorizer*. We limit ourselves mostly to convex quadratic majorizers because concave ones have no minima and are of limited use for algorithmic purposes.

The first result, which has been widely applied, applies to functions with a continuous and uniformly bounded second derivative (Böhning and Lindsay, 1988).

Theorem 3.1. *If f is twice differentiable and there is an $B > 0$ such that $f''(x) \leq B$ for all x, then for each y the convex quadratic function*

$$g(x) = f(y) + f'(y)(x-y) + \frac{1}{2}B(x-y)^2$$

majorizes f at y.

Proof. Use Taylor's theorem in the form

$$f(x) = f(y) + f'(y)(x-y) + \frac{1}{2}f''(\xi)(x-y)^2,$$

with ξ on the line connecting x and y. Because $f''(\xi) \leq B$, this implies $f(x) \leq g(x)$, where g is defined above. ■

This result is very useful, but it has some limitations. In the first place we would like a similar result for functions that are not everywhere twice differentiable, or even those that are not everywhere differentiable. Second, the bound does take into account that we only need to bound the second derivative on the interval between x and y, and not on the whole line. This may result in a bound which is not sharp. In particular we shall see below that substantial improvements can result from a non-uniform bound $B(y)$ that depends on the support point y.

Why do we want the bounds on the second derivative to be sharp? The majorization algorithm corresponding to this result is

$$x^{(k+1)} = x^{(k)} - \frac{1}{B}f'(x^{(k)}),$$

which converges linearly, say to x_∞, by Ostrowski's Theorem (De Leeuw, 1994). More precisely

$$\lim_{k\to\infty} \frac{|x^{(k+1)} - x_\infty|}{|x^{(k)} - x_\infty|} = 1 - \frac{1}{B}f''(y).$$

Thus the smaller we choose B, the faster our convergence. We mention some simple properties of quadratic majorizers.

Property 1. *If a quadratic g majorizes a twice-differentiable convex function f at y, then g is convex. This follows from $g''(y) \geq f''(y) \geq 0$.*

Property 2. *Quadratic majorizers are not necessarily convex. In fact, they can even be concave. Take $f(x) = -x^2$ and $g(x) = -x^2 + \frac{1}{2}(x-y)^2$.*

Property 3. *If a concave quadratic g majorizes a twice-differentiable function f at y, then f is concave at y. This follows from $0 \geq g''(y) \geq f''(y)$.*

Property 4. *For some functions quadratic majorizers may not exist. Suppose, for example, that f is a cubic. If g is quadratic and majorizes f, then we must have $d = g - f \geq 0$. But $d = g - f$ is a cubic, and thus $d < 0$ for at least one value of x.*

Property 5. *Quadratic majorizers may exist almost everywhere, but not everywhere. Suppose, for example, that $f(x) = |x|$. Then f has a quadratic majorizer at each y except $y = 0$. If $y \neq 0$ we can use, following Heiser (1986), the arithmetic mean-geometric mean inequality in the form*

$$\sqrt{x^2y^2} \leq \frac{1}{2}(x^2 + y^2),$$

and find

$$|x| \leq \frac{1}{2|y|}x^2 + \frac{1}{2}|y|.$$

That a quadratic majorizer does not exist at $y = 0$ follows from the discussion at the end of Section 2: f is convex and $f'_L(0) = -1 < f'_R(0) = +1$.

Example 1. For a nice regular example we use the celebrated functions

$$\phi(x) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-z^2/2},$$

$$\Phi(x) = \int_{-\infty}^{x} \phi(z)\,\mathrm{d}z.$$

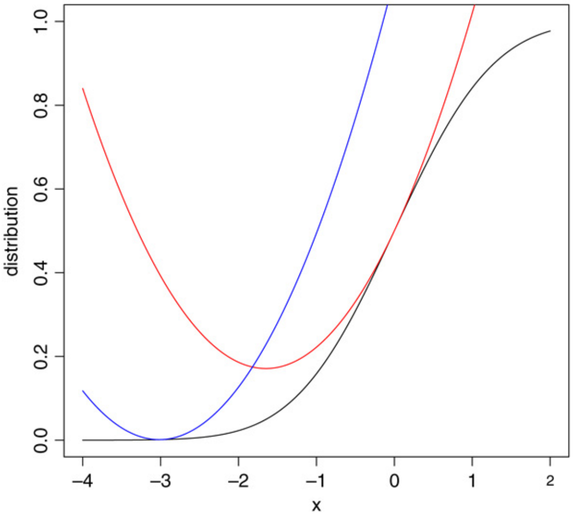

Fig. 1. Quadratic majorization of cumulative normal.

Then

$$\begin{aligned}
&\Phi'(x) = \phi(x),\\
&\Phi''(x) = \phi'(x) = -x\phi(x),\\
&\Phi'''(x) = \phi''(x) = -(1 - x^2)\phi(x),\\
&\Phi''''(x) = \phi'''(x) = -x(x^2 - 3)\phi(x).
\end{aligned}$$

To obtain quadratic majorizers we must bound the second derivatives. We can bound $\Phi''(x)$ by setting its derivative equal to zero. We have $\Phi'''(x) = 0$ for $x = \pm 1$. Moreover $\Phi''''(1) < 0$ and thus $\Phi''(x) \leq \phi(1)$. In the same way $\phi'''(x) = 0$ for $x = 0$ and $x = \pm\sqrt{3}$. At $x = 0$ the function $\phi''(x)$ has a minimum, at $x = \pm\sqrt{3}$ it has two maxima. Thus $\phi''(x) \leq 2\phi(\sqrt{3})$. More precisely, it follows that

$$\begin{aligned}
&0 \leq \Phi'(x) = \phi(x) \leq \phi(0),\\
&-\phi(1) \leq \Phi''(x) = \phi'(x) \leq \phi(1),\\
&-\phi(0) \leq \Phi'''(x) = \phi''(x) \leq 2\phi(\sqrt{3}).
\end{aligned}$$

Thus we have the quadratic majorizers

$$\Phi(x) \leq \Phi(y) + \phi(y)(x - y) + \frac{1}{2}\phi(1)(x - y)^2,$$

and

$$\phi(x) \leq \phi(y) - y\phi(y)(x - y) + \phi(\sqrt{3})(x - y)^2.$$

The majorizers are illustrated for both Φ and ϕ at the points $y = 0$ and $y = -3$ in Figs. 1 and 2. The inequalities in this section may be useful in majorizing multivariate functions involving ϕ and Φ. They are mainly intended, however, to illustrate construction of quadratic majorizers in the smooth case.

4. Sharp quadratic majorization

We now drop the assumption that the objective function is twice differentiable, even locally, and we try to improve our bound estimates at the same time.

4.1. Differentiable case

Let us first deal with the case in which f is differentiable in y. Consider all $a > 0$ for which

$$f(x) \leq f(y) + f'(y)(x - y) + \frac{1}{2}a(x - y)^2$$

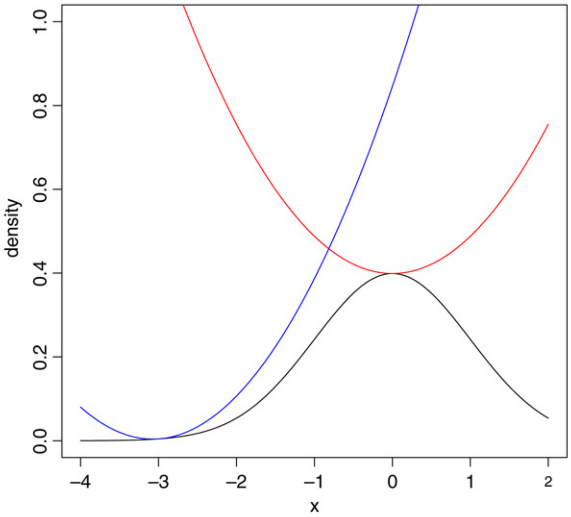

Fig. 2. Quadratic majorization of normal density.

for a fixed y and for all x. Equivalently, we must have, for all x,

$$a \geq \frac{f(x) - f(y) - f'(y)(x-y)}{\frac{1}{2}(x-y)^2}. \tag{1}$$

Define the function

$$\delta(x, y) = \frac{f(x) - f(y) - f'(y)(x-y)}{\frac{1}{2}(x-y)^2}$$

for all $x \neq y$. The system of inequalities (1) has a solution if and only if

$$A(y) = \sup_x \delta(x, y) < \infty.$$

If this is the case, then any $a \geq A(y)$ will satisfy (1). Because we want a to be as small as possible, we will usually prefer to choose $a = A(y)$. This is what we mean by the *sharp quadratic majorization*. If the second derivative is uniformly bounded by B, we have $A(y) \leq B$, and thus our bound improves on the uniform bound considered before.

The function δ has some interesting properties. For differentiable convex f we have $f(x) \geq f(y) + f'(y)(x - y)$ and thus $\delta(x, y) \geq 0$. In the same way for concave f we have $\delta(x, y) \leq 0$. For strictly convex and concave f these inequalities are strict. If $\delta(x, y) \leq 0$ for all x and y, then f must be concave. Consequently $A(y) \leq 0$ only if f is concave, and without loss of generality we can exclude this case from consideration.

The function $\delta(x, y)$ is closely related to the second derivative at or near y. If f is twice differentiable at y, then, by the definition of the second derivative,

$$\lim_{x \to y} \delta(x, y) = f''(y). \tag{2}$$

If f is three times differentiable, we can use the Taylor Expansion to sharpen this to

$$\lim_{x \to y} \frac{\delta(x, y) - f''(y)}{x - y} = \frac{1}{6} f'''(y).$$

Moreover, in the twice differentiable case, the Mean Value Theorem implies there is a ξ in the interval extending from x to y with $\delta(x, y) = f''(\xi)$. We can also derive an integral representation of $\delta(x, y)$ and its first derivative with respect to x (Tom Ferguson, Personal Communication, 03/12/04).

Lemma 4.1. *$\delta(x, y)$ can written as the expectation*

$$\delta(x, y) = \mathrm{E}\{f''[Vy + (1 - V)x]\},$$

where the random variable V follows a $\beta(2, 1)$ distribution. Likewise

$$\delta'(x, y) = \frac{1}{3}\mathrm{E}\{f'''[Wy + (1 - W)x]\},$$

where the random variable W follows a $\beta(2, 2)$ distribution. Thus $\delta(x, y)$ and $\delta'(x, y)$ can be interpreted as smoothed versions of f'' and f'''.

Proof. The first representation follows from the second-order Taylor's expansion

$$f(x) = f(y) + f'(y)(x-y) + (x-y)^2 \int_0^1 f''[vy + (1-v)x]v \, dv$$

with integral remainder (Lange, 2004). This can be rewritten as

$$\delta(x,y) = 2\int_0^1 f''[vy + (1-v)x]v \, dv. \tag{3}$$

Since the density of $\beta(2, 1)$ at v is $2v$ this gives the first result in the lemma. Differentiation under the integral sign of (3) yields the second representation. ■

In view of Lemma 4.1, $\delta(x, y)$ is jointly continuous in x and y when $f''(x)$ is continuous. Furthermore, if $f''(x)$ tends to ∞ as x tends to $-\infty$ or $+\infty$, then $\delta(x, y)$ is unbounded in x for each fixed y. Thus, quadratic majorizations do not exist for any y if the second derivative grows unboundedly. It also follows from Lemma 4.1 that the best quadratic majorization does not exist if the third derivative f''' is always positive (or always negative). This happens, for instance, if the first derivative f' is strictly convex or strictly concave. Thus as mentioned earlier, cubics do not have quadratic majorizations.

Property 6. *Majorization may be possible at all points y without the function $A(y)$ being bounded. Suppose the graph of $f''(x)$ is 0 except for an isosceles triangle centered at each integer $n \geq 2$. If we let the base of the triangle be $2n^{-3}$ and the height of the triangle be n, then the area under the triangle is n^{-2}. The formulas*

$$f'(x) = \int_0^x f''(y)\, dy, \qquad f(x) = \int_0^x f'(y)\, dy$$

define a nonnegative convex function $f(x)$ satisfying

$$f'(x) \leq \sum_{n=2}^{\infty} \frac{1}{n^2} < \infty.$$

To prove the $A(y)$ is finite for every y, recall the limit (2) and observe that

$$\delta(x,y) = \frac{f'(w)(x-y) - f'(y)(x-y)}{\frac{1}{2}(x-y)^2} = \frac{f'(w) - f'(y)}{\frac{1}{2}(x-y)}$$

for some w between x and y. It follows that $\delta(x, y)$ tends to 0 as $|x|$ tends to ∞. Because $A(n) \geq f''(n) = n$, it is clear that $A(y)$ is unbounded.

4.2. Computing the sharp quadratic majorization

Let us study the case in which the supremum of $\delta(x, y)$ over $x \neq y$ is attained at, say, $z \neq y$. In our earlier notation $A(y) = \delta(z, y)$. Differentiating $\delta(x, y)$ with respect to x gives

$$\delta'(x,y) = \frac{\frac{1}{2}(x-y)^2[f'(x) + f'(y)] - (x-y)[f(x) - f(y)]}{\frac{1}{4}(x-y)^4},$$

and

$$\frac{f(z) - f(y)}{z - y} = \frac{1}{2}[f'(z) + f'(y)] \tag{4}$$

is a necessary and sufficient condition for $\delta'(z, y)$ to vanish. At the optimal z we have

$$A(y) = \delta(z,y) = \frac{f'(z) - f'(y)}{z - y}. \tag{5}$$

It is interesting that the fundamental theorem of calculus allows us to recast Eqs. (4) and (5) as

$$\frac{1}{2}[f'(z) + f'(y)] = \int_0^1 f'[z + t(y-z)]\, dt$$

$$A(y) = \int_0^1 f''[z + t(y-z)]\, dt.$$

When f is convex, $A(y) \geq 0$. For the second derivative at z, we have

$$\delta''(z,y) = \frac{(z-y)^2 f''(z) - [f'(z) - f'(y)](z-y)}{\frac{1}{2}(z-y)^4}.$$

At a maximum we must have $\delta''(z, y) \leq 0$, which is equivalent to

$$f''(z) \leq \frac{f'(z) - f'(y)}{z - y} = A(y). \tag{6}$$

We can achieve more clarity by viewing these questions from a different angle. If the quadratic g majorizes f at y, then it satisfies

$$g(x) = f(y) + f'(y)(x - y) + \frac{1}{2}a(x - y)^2$$

for some a. If z is a second support point, then g not only intersects f at z, but it also majorizes f at z. The condition $g'(z) = f'(z)$ yields

$$a = \frac{f'(z) - f'(y)}{z - y}.$$

If we match this value with the requirement $\delta(z, y) = a$, then we recover the second equality in (5). Conversely, if a point z satisfies the second equality in (5), then it is a second support point. In this case, one can easily check condition (4) guaranteeing that z is a stationary point of $\delta(x, y)$.

4.3. Optimality with two support points

Building on earlier work by Groenen et al. (2003), Van Ruitenburg (2005) proves that a quadratic function g majorizing a differentiable function f at two points must be a sharp majorizer. The idea of looking for quadratic majorizers with two support points has been used earlier by Heiser (1986) and others. Van Ruitenburg, however, is the first to present the result in full generality. Our approach is more analytical and computational, and designed to be applied eventually to multivariate quadratic majorizations. For completeness, we now summarize in our language Van Ruitenburg's (2005) lovely proof of the two-point property.

Lemma 4.2. *Suppose two quadratic functions $g_1 \neq g_2$ both majorize the differentiable function f at y. Then either g_1 strictly majorizes g_2 at y or g_1 strictly majorizes g_2 at y.*

Proof. We have

$$g_1(x) = f(y) + f'(y)(x - y) + \frac{1}{2}a_1(x - y)^2, \tag{7}$$

$$g_2(x) = f(y) + f'(y)(x - y) + \frac{1}{2}a_2(x - y)^2, \tag{8}$$

with $a_1 \neq a_2$. Subtracting (7) and (8) proves the lemma. ■

Lemma 4.3. *Suppose the quadratic function g_1 majorizes a differentiable function f at y and $z_1 \neq y$ and that the quadratic function g_2 majorizes f at y and $z_2 \neq y$. Then $g_1 = g_2$.*

Proof. Suppose $g_1 \neq g_2$. Since both g_1 and g_2 majorize f at y, Lemma 4.2 applies. If g_2 strictly majorizes g_1 at y, then $g_1(z_2) < g_2(z_2) = f(z_2)$, and g_1 does not majorize f. If g_1 strictly majorizes g_2 at y, then similarly $g_2(z_1) < g_1(z_1) = f(z_1)$, and g_2 does not majorize f. Unless $g_1 = g_2$, we reach a contradiction. ■

We now come to Van Ruitenburg's main result.

Theorem 4.4. *Suppose a quadratic function g_1 majorizes a differentiable function f at y and at $z \neq y$, and suppose $g_2 \neq g_1$ is a quadratic majorizing f at y. Then g_2 strictly majorizes g_1 at y.*

Proof. Suppose g_1 strictly majorizes g_2. Then $g_2(z) < g_1(z) = f(z)$ and thus g_2 does not majorize f. The result now follows from Lemma 4.2. ■

Property 7. *It is not true, by the way, that a quadratic majorizer can have at most two support points. There can even be an infinite number of them. Consider the function $h(x) = c \sin^2(x)$ for some $c > 0$. Clearly $h(x) \geq 0$ and $h(x) = 0$ for all integer multiples of π. Now define $f(x) = x^2 - h(x)$ and $g(x) = x^2$. Then g is a quadratic majorizer of f at all integer multiples of π. This is plotted in Fig. 3 for $c = 10$.*

Property 8. *There is no guarantee that a second support point $z \neq y$ exists. Consider the continuously differentiable convex function*

$$f(x) = \begin{cases} x^2 & x \leq 1 \\ 2x - 1 & x > 1, \end{cases}$$

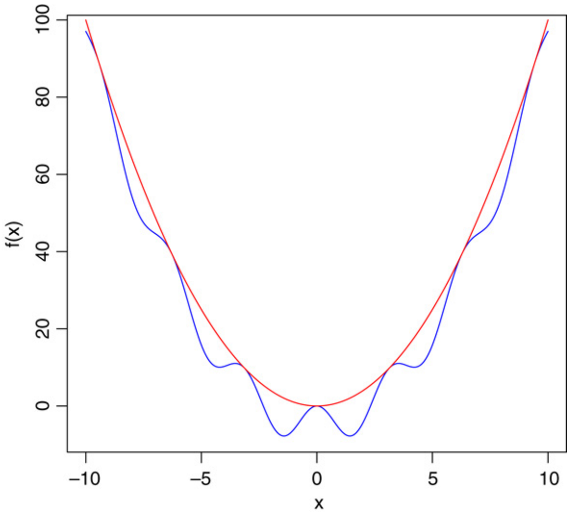

Fig. 3. Many support points.

and fix $y > 1$. For $x > 1$

$$\delta(x,y) = \frac{2x - 1 - 2y + 1 - 2(x-y)}{\frac{1}{2}(x-y)^2} = 0.$$

For $x \leq 1$

$$\delta(x,y) = \frac{x^2 - 2y + 1 - 2(x-y)}{\frac{1}{2}(x-y)^2} = \frac{(x-1)^2}{\frac{1}{2}(x-y)^2}.$$

It follows that $\lim_{x\to-\infty}\delta(x,y) = 2$. On the other hand, one can easily demonstrate that $\delta(x,y) < 2$ whenever $x \leq 1$. Hence, $A(y) = 2$, but $\delta(x,y) < 2$ for all $x \neq y$.

4.4. Even functions

Assuming that $f(x)$ is even, i.e. $f(x) = f(-x)$ for all x, simplifies the construction of quadratic majorizers. If an even quadratic g satisfies $g(y) = f(y)$ and $g'(y) = f'(y)$, then it also satisfies $g(-y) = f(-y)$ and $g'(-y) = f'(-y)$. If in addition g majorizes f at either y or $-y$, then it majorizes f at both y and $-y$, and Theorem 4.4 implies that it is the best possible quadratic majorization at both points. This means we only need an extra condition to guarantee that g majorizes f. The next theorem, essentially proved in the references (Groenen et al., 2003; Jaakkola and Jordan, 2000; Hunter and Li, 2005) by other techniques, highlights an important sufficient condition.

Theorem 4.5. *Suppose $f(x)$ is an even, differentiable function on $\mathbb{R}$ such that the ratio $f'(x)/x$ is decreasing on $(0,\infty)$. Then the even quadratic*

$$g(x) = \frac{f'(y)}{2y}(x^2 - y^2) + f(y)$$

is the best quadratic majorizer of $f(x)$ at the point y.

Proof. It is obvious that $g(x)$ is even and satisfies the tangency conditions $g(y) = f(y)$ and $g'(y) = f'(y)$. For the case $0 \leq x \leq y$, we have

$$\begin{aligned} f(y) - f(x) &= \int_x^y f'(z)\,\mathrm{d}z \\ &= \int_x^y \frac{f'(z)}{z} z\,\mathrm{d}z \\ &\geq \frac{f'(y)}{y}\int_x^y z\,\mathrm{d}z \\ &= \frac{f'(y)}{y}\frac{1}{2}(y^2 - x^2) \\ &= f(y) - g(x), \end{aligned}$$

where the inequality comes from the assumption that $f'(x)/x$ is decreasing. It follows that $g(x) \geq f(x)$. The case $0 \leq y \leq x$ is proved in similar fashion, and all other cases reduce to these two cases given that $f(x)$ and $g(x)$ are even. ■

There is an condition equivalent to the sufficient condition of Theorem 4.5 that is sometimes easier to check.

Theorem 4.6. *The ratio $f'(x)/x$ is decreasing on $(0, \infty)$ if and only $f(\sqrt{x})$ is concave. The set of functions satisfying this condition is closed under the formation of* (a) *positive multiples,* (b) *convex combinations,* (c) *limits, and* (d) *composition with a concave increasing function $g(x)$.*

Proof. Suppose $f(\sqrt{x})$ is concave in x and $x > y$. Then the two inequalities

$$f(\sqrt{x}) \leq f(\sqrt{y}) + \frac{f'(\sqrt{y})}{2\sqrt{y}}(x-y)$$
$$f(\sqrt{y}) \leq f(\sqrt{x}) + \frac{f'(\sqrt{x})}{2\sqrt{x}}(y-x)$$

are valid. Adding these, subtracting the common sum $f(\sqrt{x}) + f(\sqrt{y})$ from both sides, and rearranging give

$$\frac{f'(\sqrt{x})}{2\sqrt{x}}(x-y) \leq \frac{f'(\sqrt{y})}{2\sqrt{y}}(x-y).$$

Dividing by $(x-y)/2$ yields the desired result

$$\frac{f'(\sqrt{x})}{\sqrt{x}} \leq \frac{f'(\sqrt{y})}{\sqrt{y}}.$$

Conversely, suppose the ratio is decreasing and $x > y$. Then the mean value expansion

$$f(\sqrt{x}) = f(\sqrt{y}) + \frac{f'(\sqrt{z})}{2\sqrt{z}}(x-y)$$

for $z \in (y, x)$ leads to the concavity inequality.

$$f(\sqrt{x}) \leq f(\sqrt{y}) + \frac{f'(\sqrt{y})}{2\sqrt{y}}(x-y).$$

The asserted closure properties are all easy to check. ■

As examples of property (d) of Theorem 4.6, note that the functions $g(x) = \ln x$ and $g(x) = \sqrt{x}$ are concave and increasing. Hence, if $f(\sqrt{x})$ is concave, then $\ln f(\sqrt{x})$ and $f(\sqrt{x})^{1/2}$ are concave as well.

The above discussion suggests that we look at more general transformations of the argument of f. If we define $\tilde{f}(x) = f(\alpha + \beta x)$ for an arbitrary function $f(x)$, then a brief calculation shows that

$$\tilde{A}(y) = \beta^2 A(\alpha + \beta y)$$
$$\tilde{z}(y) = \frac{z(\alpha + \beta y) - \alpha}{\beta}$$

using the identity $\tilde{\delta}(x, y) = \beta^2 \delta(\alpha + \beta x, \alpha + \beta y)$. An even function $f(x)$ satisfies $\tilde{f}(x) = f(x)$ for $\alpha = 0$ and $\beta = -1$.

4.5. Non-differentiable functions

If f is not differentiable at y, then we must find a and b such that

$$f(x) \leq f(y) + b(x-y) + \frac{1}{2}a(x-y)^2$$

for all x. This is an infinite system of linear inequalities in a and b, which means that the solution set is a closed convex subset of the plane.

Analogous to the differentiable case we define

$$\delta(x, y, b) = \frac{f(x) - f(y) - b(x-y)}{\frac{1}{2}(x-y)^2},$$

as well as

$$A(y, b) = \sup_x \delta(x, y, b).$$

If $A(y,b) < +\infty$, we have the sharpest quadratic majorization for given y and b. The sharpest quadratic majorization at y is given by

$$A(y) = \inf_b A(y,b).$$

5. Examples

As we explained in the introduction, majorizing univariate functions is usually not useful in itself. The results become relevant for statistics if they are used in the context of separable multivariate problems. In this section we first illustrate how to compute sharp quadratic majorizers for some common univariate function occurring in maximum likelihood problems, and then we apply these majorizers to the likelihood problems themselves.

5.1. Logistic

Our first example is the negative logarithm of the logistic cdf

$$\Psi(x) = \frac{1}{1+e^{-x}}.$$

Thus

$$f(x) = \log(1+e^{-x}).$$

Clearly

$$f'(x) = -\frac{e^{-x}}{1+e^{-x}} = \Psi(x) - 1,$$

and

$$f''(x) = \frac{e^{-x}}{(1+e^{-x})^2} = \Psi(x)[1-\Psi(x)].$$

Thus $f''(x) > 0$ and $f(x)$ is strictly convex. Since $f''(x) \leq 1/4$, a uniform bound is readily available.

The symmetry relations

$$f(-x) = x + f(x),$$
$$f'(-x) = -[1+f'(x)] = -\Psi(x),$$
$$f''(-x) = f''(x)$$

demonstrate that $z = -y$ satisfies Eq. (4) and hence maximizes $\delta(x,y)$. The optimum value is determined by (5) as

$$A(y) = \delta(z,y) = \frac{2\Psi(y)-1}{2y}.$$

The same result was derived, using quite different methods, by Jaakkola and Jordan (2000) and Groenen et al. (2003).

We plot the function $\delta(x,y)$ for $y = 1$ and $y = 8$ in Fig. 4. Observe that the uniform bound $1/4$ is not improved much for y close to 0, but for large values of y the improvement is huge. This is because $A(y) \approx (2|y|)^{-1}$ for large $|y|$. Thus for large values of y we will see close to superlinear convergence if we use $A(y)$.

Alternatively, we can majorize $f(x) = \log(1+e^{-x})$ by writing

$$\log(1+e^{-x}) = -\frac{1}{2}x + \log(e^{x/2} + e^{-x/2})$$

and majorizing the even function $h(x) = \log(e^{x/2} + e^{-x/2})$. Straightforward but tedious differentiation shows that

$$\begin{aligned}\left[\frac{h'(x)}{x}\right]' &= \frac{1 - e^{2x} + 2xe^x}{2x^2(1+e^x)^2}\\ &= \frac{1}{2x^2(1+e^x)^2}\sum_{k=2}^{\infty}\left[2x\frac{x^k}{k!} - \frac{(2x)^{k+1}}{(k+1)!}\right]\\ &= \frac{2}{2x^2(1+e^x)^2}\sum_{k=2}^{\infty}\frac{x^{k+1}}{k!}\left[1 - \frac{2^k}{k+1}\right]\\ &\leq 0.\end{aligned}$$

Hence, $h'(x)/x$ is decreasing on $(0,\infty)$, and Theorem 4.5 applies.

5.2. The absolute value function

Because $|x|$ is even, Theorem 4.5 yields the majorization

$$g(x) = \frac{1}{2|y|}(x^2 - y^2) + |y| = \frac{1}{2|y|}x^2 + \frac{1}{2}|y|,$$

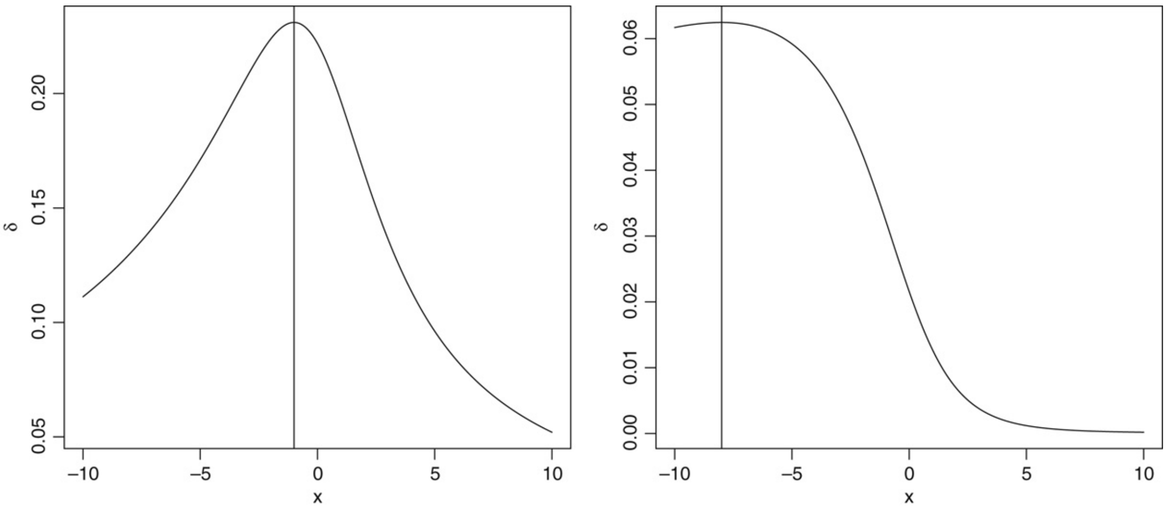

Fig. 4. δ for logistic at $y = 1$ (left) and $y = 8$ (right).

which is just the result given by the arithmetic/geometric mean inequality in Property 5. When $y = 0$, recall that no quadratic majorization exists.

If we approach majorization of $|x|$ directly, we need to find $a > 0$ and b such that

$$a(x-y)^2 + b(x-y) + |y| \geq |x|$$

for all x. Let us compute $A(y, b)$. If $y < 0$ then $b = -1$, and thus

$$A(y, -1) = \sup_{x \neq y} \frac{|x| + x}{\frac{1}{2}(x-y)^2} = \frac{1}{|y|}.$$

If $y > 0$ then $b = +1$, and again

$$A(y, +1) = \sup_{x \neq y} \frac{|x| - x}{\frac{1}{2}(x-y)^2} = \frac{1}{|y|}.$$

In both cases, the best quadratic majorizer can be expressed as

$$\begin{aligned} g(x) &= \frac{1}{2}\frac{1}{|y|}(x-y)^2 + \operatorname{sign}(y)(x-y) + |y| \\ &= \frac{1}{2|y|}x^2 + \frac{1}{2}|y|. \end{aligned}$$

5.3. The Huber function

Majorization for the Huber function, specifically quadratic majorization, has been studied earlier by Heiser (1987) and Verboon and Heiser (1994). In those papers quadratic majorization functions appear more or less out of the blue, and it is then verified that they are indeed majorization functions. This is not completely satisfactory. Here we attack the problem by applying Theorem 4.5. This leads to the sharpest quadratic majorization.

The Huber function is defined by

$$f(x) = \begin{cases} \frac{1}{2}x^2 & \text{if } |x| < c, \\ c|x| - \frac{1}{2}c^2 & \text{otherwise.} \end{cases}$$

Thus we really deal with a family of even functions, one for each $c > 0$. The Huber functions are differentiable with derivative

$$f'(x) = \begin{cases} x & \text{if } |x| < c, \\ c & \text{if } x \geq c, \\ -c & \text{if } x \leq -c. \end{cases}$$

Since it is obvious that $f'(x)/x$ is decreasing $(0, \infty)$, Theorem 4.5 immediately gives the sharpest majorizer

$$g(x) = \begin{cases} \frac{1}{2}\frac{c}{|y|}(x-y)^2 - cx - \frac{1}{2}c^2 & \text{if } y \leq -c, \\ \frac{1}{2}x^2 & \text{if } |y| < c, \\ \frac{1}{2}\frac{c}{|y|}(x-y)^2 + cx - \frac{1}{2}c^2 & \text{if } y \geq +c. \end{cases}$$

5.4. General logit and probit problems

Suppose we observe independent counts from n binomials with parameters N_i and $\pi_i(x)$, where

$$\pi_i(x) = \frac{1}{1 + \exp(-h_i(x))}$$

for given functions $h_i(x)$ of p unknown parameters x. This covers both linear and nonlinear logistic regression problems. The deviance, i.e. twice the negative log-likelihood, is

$$\begin{aligned} \mathcal{D}(x) &= -2\sum_{i=1}^{N} n_i \log \pi_i + (N_i - n_i)\log(1 - \pi_i(x)) \\ &= 2\sum_{i=1}^{n} N_i\{f(h_i(x)) - (p_i - 1)h_i(x)\}, \end{aligned}$$

where as before, $f(x) = \log(1 + \exp(-x))$. Using $f'(-h_i(x)) = -\Psi(h_i(x)) = -\pi_i(x)$ we see that a quadratic majorization of f

$$f(h_i(x)) \leq f(h_i(y)) + f'(h_i(y))(h_i(x) - h_i(y)) + \frac{1}{2}A_i(y)(h_i(x) - h_i(y))^2$$

leads to the quadratic majorization of the deviance by a weighted least squares function of the form

$$\sigma(x) = \sum_{i=1}^{n} N_i A_i(y)(h_i(x) - z_i(y))^2,$$

where

$$z_i(y) = h_i(y) + \frac{\pi_i(y) - p_i}{A_i(y)}.$$

This means we can solve the logistic problem by solving a sequence of weighted least squares problems. If the h_i are linear, then these are just linear regression problems. If the h_i are bilinear, the subproblems are weighted singular value decompositions, and if the h_i are distances the subproblems are least squares multidimensional scaling problems. If we have algorithms to solve the weighted least squares problems, then we automatically have an algorithm to solve the corresponding logistic maximum likelihood problem.

In fact, it is shown by De Leeuw (2006) that essentially the same results apply if we replace the logit Ψ by the probit function Φ. The difference is that for the probit we have $A(y) \equiv 1$ and thus uniform quadratic majorization is sharp.

As one of the reviewers correctly points out, sharp univariate quadratic majorization of f does not imply sharp multivariate quadratic majorization of $\mathcal{D}$. Although, it is of course true that sharp univariate majorization gives better results that unsharp univariate majorization. The problem of sharp multivariate majorization is basically unexplored, although we do have some tentative results.

5.5. Application to discriminant analysis

Discriminant analysis is another attractive application. In discriminant analysis with two categories, each case i is characterized by a feature vector z_i and a category membership indicator y_i taking the values -1 or 1. In the machine learning approach to discriminant analysis (Vapnik, 1995), the hinge loss function $[1 - y_i(\alpha + z_i^t\beta)]_+$ plays a prominent role. Here $(u)_+$ is shorthand for the convex function $\max\{u, 0\}$. Just as in ordinary regression, we can penalize the overall separable loss

$$g(\theta) = \sum_{i=1}^{m}[1 - y_i(\alpha + z_i^t\beta)]_+,$$

where $\theta = (\alpha, \beta)$, by imposing a lasso or ridge penalty $+\lambda\theta'\theta$.

Most strategies for estimating θ pass to the dual of the original minimization problem. A simpler strategy, proposed by Groenen et al. (2007) is to majorize each contribution to the loss by a quadratic and minimize the surrogate loss plus penalty. In Groenen et al. (2008) this approach is extended to quadratic and Huber hinges, still maintaining the idea of using quadratic majorizers. A little calculus shows that the absolute value hinge $(u)_+$ is majorized at $u_n \neq 0$ by the quadratic

$$q(u \mid u_n) = \frac{1}{4|u_n|}\left(u + |u_n|\right)^2 . \tag{9}$$

In fact, by the same reasoning as for the absolute value function, this is the best quadratic majorizer. To avoid the singularity at 0, we recommend replacing $q(u \mid u_n)$ by

$$r(u \mid u_n) = \frac{1}{4|u_n| + \varepsilon}\left(u + |u_n|\right)^2 .$$

In double precision, a good choice of ε is 10^{-5}. Of course, the dummy variable u is identified in case i with $1 - y_i(\alpha + z_i^t\beta)$. If we impose a ridge penalty, then the majorization (9) leads to a majorization algorithm exploiting weighted least squares.

6. Iterative computation of $A(y)$

In general, one must find $A(y)$ numerically. We do not suggest here that the combination of finding $A(y)$ by an iterative algorithm, and then using this $A(y)$ in the iterative quadratic majorization algorithm, is necessarily an efficient way to construct overall algorithms. It requires an infinite number of iterations within each of an infinite number of iterations. The results in this section can be used, however, for computing $A(y)$ for specific functions to find out how it behaves as a function of y. This has been helpful to us in finding $A(y)$ for the logit and probit, where the computations suggested the final analytic result.

For a convex function f, two similar iterative algorithms are available. They both depend on minorizing f by the linear function $f(x^{(k)}) + f'(x^{(k)})(x - x^{(k)})$ at the current point $x^{(k)}$ in the search for the maximum z of $\delta(x, y)$. This minorization propels the further minorization

$$\begin{aligned}\delta(x, y) &\geq \frac{f(x^{(k)}) + f'(x^{(k)})(x - x^{(k)}) - f(y) - f'(y)(x - y)}{\frac{1}{2}(x - y)^2}\\ &= \frac{[f'(x^{(k)}) - f'(y)](x - y) + f(x^{(k)}) + f'(x^{(k)})(y - x^{(k)}) - f(y)}{\frac{1}{2}(x - y)^2}.\end{aligned}$$

Maximizing the displayed minorizer drives $\delta(x, y)$ uphill. Fortunately, the minorizer is a function of the form

$$h(w) = \frac{cw + d}{w^2} = \frac{c}{w} + \frac{d}{w^2}$$

with $w = x - y$. The stationary point $w = -2d/c$ furnishes the maximum of $h(w)$ provided

$$h''\left(-\frac{2d}{d}\right) = \left.\frac{2c}{w^3}\right|_{w=-2d/c} + \left.\frac{6d}{w^4}\right|_{w=-2d/c} = \frac{c^4}{8d^3}$$

is negative. If $f(x)$ is strictly convex, then

$$d = 2\left[f(x^{(k)}) + f'(x^{(k)})(y - x^{(k)}) - f(y)\right],$$

is negative, and the test for a maximum succeeds. The update can be phrased as

$$x^{(k+1)} = y - 2\,\frac{f(x^{(k)}) + f'(x^{(k)})(y - x^{(k)}) - f(y)}{f'(x^{(k)}) - f'(y)}.$$

A brief calculation based on Eqs. (4) and (5) shows that the iteration map $x^{(k+1)} = g(x^{(k)})$ has derivative

$$g'(z) = \frac{f''(z)(z - y)}{f'(z) - f'(y)} = \frac{f''(z)}{A(y)}$$

at the optimal point z.

On the other hand, the Dinkelbach (1967) maneuver for increasing $h(w)$ considers the function $e(w) = cw + d - h(w^{(k)})w^2$ with value $e(w^{(k)}) = 0$. If we choose

$$w^{(k+1)} = \frac{c}{2h(w^{(k)})}$$

to maximize $e(w)$, then it is obvious that $h(w^{(k+1)}) \geq h(w^{(k)})$. This gives the iteration map

$$x_{n+1} = y + \frac{\frac{1}{2}[f'(x^{(k)}) - f'(y)](x^{(k)} - y)^2}{f(x^{(k)}) - f(y) - f'(y)(x^{(k)} - y)} = y + \frac{f'(x^{(k)}) - f'(y)}{\delta(x^{(k)}, y)}$$

with derivative at z equal to $f''(z)/A(y)$ by virtue of Eqs. (4) and (5). Hence, the two algorithms have the same local rate of convergence. We recommend starting both algorithms near y. In the case of the Dinkelbach algorithm, this entails

$$h(w) \approx \delta(x,y) \approx f''(y) > 0$$

for $f(x)$ strictly convex. Positivity of $h(w^{(0)})$ is required for proper functioning of the algorithm.

In view of the convexity of $f(x)$, it is clear that $f''(z)/A(y) \geq 0$. The inequality $f''(z) \leq A(y)$ follows from the condition $A(y) = A(z)$ determined by Theorem 4.4 and inequality (6). Ordinarily, strict inequality $f''(z) < A(y)$ prevails, and the two iteration maps just defined are locally contractive. Globally, the standard convergence theory for iterative majorization (MM algorithms) suggests that $\lim_{n\to\infty} |x^{(k+1)} - x^{(k)}| = 0$ and that the limit of every convergent subsequence must be a stationary point of $\delta(x,y)$ (Lange, 2004).

7. Discussion

In separable problems in which quadratic majorizers exist we have shown that it is often possible to use univariate sharp quadratic majorizers to increase the convergence speed of iterative majorization (or MM) algorithms. This is true even for multivariate problems with a potentially very large number of parameters, such as the logit and probit problems in Section 5.4.

There is, however, still plenty of room for improvement. We do not have, at the moment, a satisfactory theory of multivariate sharp quadratic majorization, and such a theory would obviously help to boost convergence rates even more. Such a theory should be based on the fact that a multivariate quadratic majorizer must satisfy

$$f(x) - f(y) - f'(y)(x-y) - \frac{1}{2}(x-y)'A(x-y) \leq 0$$

for all x. This is an infinite system of linear inqualities in A, and thus the solution set is either empty or convex. Sharp quadratic majorization will concentrate on the extreme points of this convex set, although it is unrealistic to expect that we can find an $A(y)$ which is sharp in all directions. Clearly both the theoretical and the implementation aspects of sharp multivariate majorization are a useful topic for further research.

Acknowledgments

This research was supported in part by NIH grants GM53275 and MH59490 to KL. Comments by two anonymous reviewers have been very helpful in preparing this final version.

References

Böhning, D., Lindsay, B.G., 1988. Monotonicity of quadratic approximation algorithms. Ann. Inst. Statist. Math. 40, 641–663.

De Leeuw, J., 1994. Block relaxation algorithms in statistics. In: Bock, H.H., Lenski, W., Richter, M.M. (Eds.), Information Systems and Data Analysis. Springer-Verlag, Berlin, pp. 308–325.

De Leeuw, J., 2006. Principal component analysis of binary data by iterated singular value decomposition. Comput. Statist. Data Anal. 50, 21–39.

Dinkelbach, W., 1967. On nonlinear fractional programming. Manage. Sci. 13, 492–498.

Groenen, P.J.F., Giaquinto, P., Kiers, H.A.L., 2003. Weighted majorization algorithms for weighted least squares decomposition models. Technical Report EI 2003–09, Econometric Institute, Erasmus University, Rotterdam, Netherlands.

Groenen, P.J.F., Nalbantov, G., Bioch, J.C., 2007. Nonlinear support vector machines through iterative majorization and I-splines. In: Lenz, H.J., Decker, R. (Eds.), Studies in Classification, Data Analysis, and Knowledge Organization. Springer-Verlag, Heidelberg, Berlin, pp. 149–161.

Groenen, P.J.F., Nalbantov, G., Bioch, J.C., 2008. SVM-Maj: A majorization approach to linear support vector machines with different hinge errors. Adv. Data Anal. Classification 2, 17–43.

Heiser, W.J., 1986. A majorization algorithm for the reciprocal location problem. Technical Report RR-86-12. Department of Data Theory, University of Leiden, Leiden, Netherlands.

Heiser, W.J., 1987. Correspondence analysis with least absolute residuals. Comput. Statist. Data Anal. 5, 337–356.

Heiser, W.J., 1995. Convergent computation by iterative majorization: Theory and applications in multidimensional data analysis. In: Krzanowski, W.J. (Ed.), Recent Advances in Descriptive Multivariate Analysis. Clarendon Press, Oxford, pp. 157–189.

Hunter, D.R., Lange, K., 2004. A tutorial on MM algorithms. Amer. Statist. 58, 30–37.

Hunter, D.R., Li, R., 2005. Variable selection using MM algorithms. Ann. Statist. 33, 1617–1642.

Jaakkola, T.S.W., Jordan, M.I.W., 2000. Bayesian parameter estimation via variational methods. Statist. Comput. 10, 25–37.

Lange, K., 2004. Optimization. Springer-Verlag, New York.

Lange, K., Hunter, D.R., Yang, I., 2000. Optimization transfer using surrogate objective functions (with discussion). J. Comput. Graph. Statist. 9, 1–59.

McShane, E.J., 1944. Integration. Princeton University Press, Princeton.

Van Ruitenburg, J., 2005. Algorithms for parameter estimation in the Rasch model. Measurement and Research Department Reports 2005–04, Arnhem, Netherlands.

Vapnik, V., 1995. The Nature of Statistical Learning Theory. Springer-Verlag, New York.

Verboon, P., Heiser, W.J., 1994. Resistant lower rank approximation of matrices by iterative majorization. Comput. Statist. Data Anal. 18, 457–467.