SHARP QUADRATIC MAJORIZATION IN ONE DIMENSION

JAN DE LEEUW

Abstract. Quadratic majorizations for real-valued functions of a real variable are analyzed, and the concept of sharp majorization is introduced.

1. Introduction

Majorization algorithms, including the EM algorithm, are used for more and more computational tasks in statistcs [De Leeuw, 1994; Heiser, 1995; Lange et al., 2000]. The basic idea is simple. A function g majorizes a function f in 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 in 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 in $x^{(k+1)}$, 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 in optimization and statistics are *separable*, i.e. they are of the form

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

Date: April 26, 2006.

2000 *Mathematics Subject Classification.* 49M20,

Key words and phrases. Optimization, Methods of Successive Approximation, Methods of Relaxation Type .

and majorization of the univariate function f automatically gives a majorization of F.

Many of our results can be generalized without much trouble to real-valued functions on $\mathbb{R}^n$, and to constrained minimization over subsets of $\mathbb{R}^n$. Our univariate context is general enough to explain most of the basic ideas.

2. Majorization

We formalize the definition of majorization at a point.

Definition 2.1. Suppose f and g are real-valued functions on $\mathbb{R}^n$. 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 minimum is unique. 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 functions on $\mathbb{R}^n$ and g is a real-valued function on $\mathbb{R}^n \otimes \mathbb{R}^n$. 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 $x \neq y$.

2.1. **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 under appropriate additional compactness and continuity conditions this guarantees convergence of the algorithm. 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 $x^{(k+1)} = h(x^{(k)})$.

2.2. **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 differerentiability fails. If f has a left and right derivative in y, for instance, and g is differentiable, then

$$f'_R(y) \leq g'(y) \leq f'_L(y).$$

If g is differentiable, and f is convex, we must have

$$g'(y) \in \partial f(y),$$

with $\partial f(y)$ the subdifferential of f at y. Even more general statements are possible by using the four Dini derivates at y.

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 to convex quadratic majorizers, because concave ones have no minima and are useless 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.

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

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

which converges linearly, say to x_∞, with rate $1 - \frac{1}{B} f''(x_\infty)$. The smaller we choose B, the faster our convergence.

Example 3.1. If a quadratic g majorizes a twice-differentiable convex function f at y, then g is convex. This follows from $g''(y) \geq f''(y) \geq 0$.

Example 3.2. If a concave quadratic g majorizes a twice-differentiable function f at y, then f is concave at y, in the sense that $f''(x) \leq 0$. This follows from $0 \geq g''(y) \geq f''(y)$.

Example 3.3. Quadratic majorizers can be concave. Take $f(x) = -x^2$ and $g(x) = -x^2 + \frac{1}{2}(x - y)^2$.

Example 3.4. Quadratic majorizers may not exist anywhere. Suppose, for example, that f is a cubic. If g is quadratic, then $d = g - f$ is a cubic, at thus d is negative for at least one value of x.

Example 3.5. 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 at $y = 0$. If $y \neq 0$ we can use, following Heiser [1986], the arithmetic mean-geometric mean inequality in the form

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

and find

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

If g majorizes $|x|$ at 0, then we must have $ax^2 + bx \geq |x|$ for all $x \neq 0$, and thus $ax + b\,\textbf{sign}(x) \geq 1$ for all $x \neq 0$. But for $x < \min(0, \frac{1+b}{a})$ we have $ax + b\,\textbf{sign}(x) < 1$.

Example 3.6. For a nice regular example we use the celebrated functions

$$\begin{aligned}\phi(x) &= \frac{1}{\sqrt{2\pi}}\exp(-\frac{z^2}{2}),\\ \Phi(x) &= \int_{-\infty}^{x}\phi(z)dz.\end{aligned}$$

Then

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

It follows, by setting various derivatives to zero and checking for maxima and minima, 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

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

This is illustrated for both Φ and ϕ in the points $y = 0$ and $y = -3$ in Figures 1 and 2.

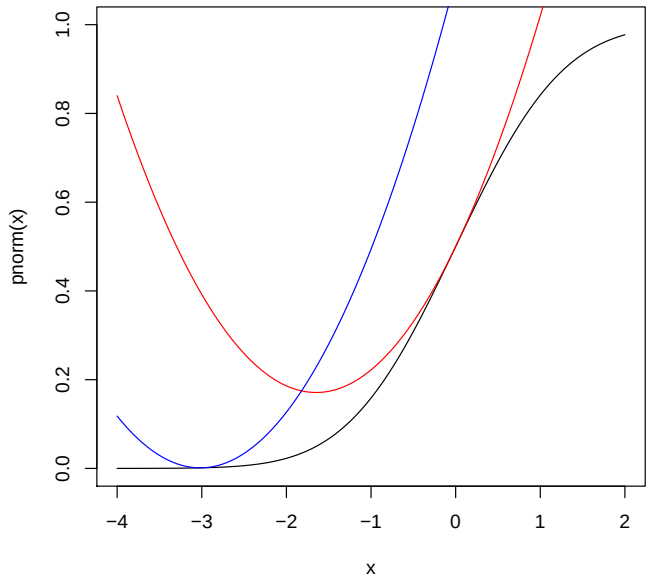

FIGURE 1. Quadratic majorization of cumulative normal

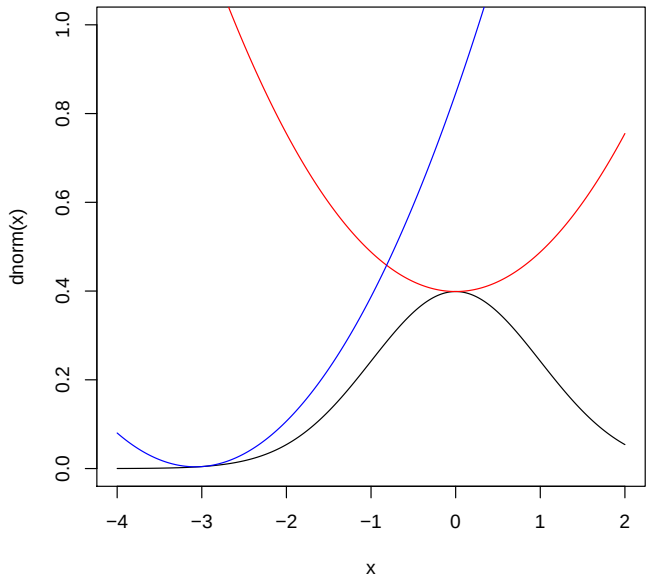

FIGURE 2. Quadratic majorization of normal density

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

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

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

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

Define the function

$$\delta(x,y) = \frac{f(x) - f(y) - f'(y)(x-y)}{\frac{1}{2}(x-y)^2} \tag{3}$$

for all $x \neq y$. The inequalities (2) have 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 (2). 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. If f is convex we have $\delta(x,y) \geq 0$ and for a concave f we have $\delta(x,y) \leq 0$. For strictly concave and convex 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.

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

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

If f is three times differentiable this can be sharpened to

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

Moreover, in the twice differentiable case, by the mean value theorem there is a ξ in the interval with endpoints x and y such that $\delta(x,y)=f''(\xi)$. Tom Ferguson (personal communication, March, 2004) has shown, more precisely, that

$$\delta(x,y)=\mathbf{E}\{f''(\underline{w}y+(1-\underline{w})x)\},$$

where $\underline{w}$ is a random variable with a **beta**$(2,1)$ distribution. Thus δ can be interpreted as a smoothed version of f''.

4.2. **Computing the Sharp Quadratic Majorization.** Let's study the case in which the supremum of $\delta(x,y)$ over x is attained at, say, $\hat{x}$. Thus $A(y)=\delta(\hat{x},y)$. Differentiating δ gives us

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

and thus we must have

$$\frac{f(\hat{x})-f(y)}{\hat{x}-y}=\frac{1}{2}(f'(\hat{x})+f'(y))). \tag{4}$$

At $\hat{x}$ we find

$$\delta(\hat{x})=\frac{f'(\hat{x})-f'(y)}{\hat{x}-y}. \tag{5}$$

If f is convex, then $\delta(\hat{x})\geq 0$. For the second derivative at $\hat{x}$ we find

$$\delta''(\hat{x})=\frac{(\hat{x}-y)^2f''(\hat{x})-(f'(\hat{x})-f'(y))(\hat{x}-y)}{\frac{1}{4}(\hat{x}-y)^4}.$$

At a maximum we must have $\delta''(\hat{x})\geq 0$, and thus

$$f''(\hat{x})\geq\frac{f'(\hat{x})-f'(y)}{\hat{x}-y}=\delta(\hat{x}). \tag{6}$$

4.3. **Non-differentiable case.** If f is not differentiable at y 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 differentible case we define

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

as well as

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

and

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

This gives us the sharpest quadratic majorization in the non-differentiable case.

5. Examples

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

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

This shows that $f(x)$ is strictly convex. Clearly $f''(x) \leq \frac{1}{4}$, so a uniform bound is readily available.

Moreover we have the symmetry relations

$$\begin{aligned} f(-x) &= x + f(x), \\ f'(-x) &= -(1 + f'(x)) = -\Psi(x), \\ f''(-x) &= f''(x). \end{aligned}$$

These can be used to show that $\hat{x} = -y$ gives satisfies (4), and gives a maximizer of δ, with optimum value given by (5)

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

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

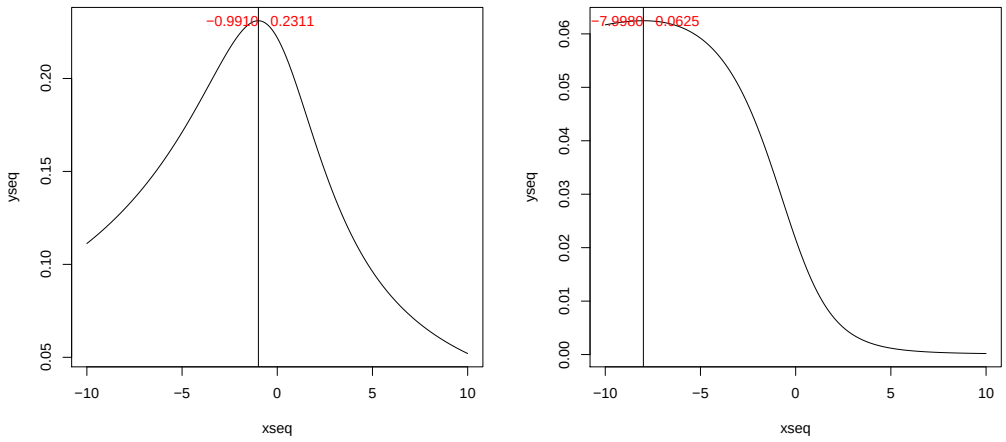

FIGURE 3. δ for logistic at $y = 1$ (left) and $y = 8$ (right).

We show the function δ for $y = 1$ and $y = 8$ in Figure 3. Observe that the uniform bound 0.25 is not improved much for y close to 0, but for large values of y the improvement is huge. This is because $A(y) \approx (2y)^{-1}$ for large y, and $A(y) \to 0$ for $y \to \pm\infty$.

5.2. **The absolute value function.** 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 $\overline{a}(b)$. If $y < 0$ then $b = -1$ and thus

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

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

$$\overline{a} = \sup_{x\neq y} \frac{|x| - x}{(x-y)^2} = \frac{1}{2}\frac{1}{|y|}.$$

If $y = 0$ then we must look at

$$\overline{a}(b) = \sup_{x\neq 0} \frac{|x| - bx}{x^2} = \sup_{x\neq 0} \frac{\mathbf{sign}(x) - b}{x},$$

which is clearly $+\infty$. For $y \neq 0$ we see that

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

Thus for $y \neq 0$ the best quadratic majorization is given by the AM/GM inequality, while for $y = 0$ no quadratic majorization exists.

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 with our technique, which is tedious but straightforward, and 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 functions, one for each $c > 0$. The Huber functions are differentiable, with derivative

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

We can find all quadratic majorizers by making a table of $\frac{f(x)-f(y)-f'(y)(x-y)}{(x-y)^2}$.

	$x \leq -c$	$\lvert x\rvert < c$	$x \geq +c$
$y \leq -c$	0	$\frac{1}{2}\frac{(x+c)^2}{(x-y)^2}$	$\frac{2cx}{(x-y)^2}$
$\lvert y\rvert < c$	$\frac{1}{2}(1-\frac{(x+c)^2}{(x-y)^2})$	$\frac{1}{2}$	$\frac{1}{2}(1-\frac{(x-c)^2}{(x-y)^2})$
$y \geq +c$	$-\frac{2cx}{(x-y)^2}$	$\frac{1}{2}\frac{(x-c)^2}{(x-y)^2}$	0

The sup over x in each cell is

	$x \leq -c$	$\lvert x\rvert < c$	$x \geq +c$
$y \leq -c$	0	$\frac{2c^2}{(c-y)^2}$	$\frac{1}{2}\frac{c}{\lvert y\rvert}$
$\lvert y\rvert < c$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$
$y \geq +c$	$\frac{1}{2}\frac{c}{\lvert y\rvert}$	$\frac{2c^2}{(c+y)^2}$	0

Finally, taking the sup of the rows shows that

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

References

D. Böhning and B.G. Lindsay. Monotonicity of Quadratic-approximation Algorithms. *Annals of the Institute of Statistical Mathematics*, 40(4): 641–663, 1988.

J. De Leeuw. Block Relaxation Methods in Statistics. In H.H. Bock, W. Lenski, and M.M. Richter, editors, *Information Systems and Data Analysis*, Berlin, 1994. Springer Verlag.

P.J.F. Groenen, P. Giaquinto, and H.L Kiers. Weighted Majorization Algorithms for Weighted Least Squares Decomposition Models. Technical Report EI 2003-09, Econometric Institute, Erasmus University, Rotterdam, Netherlands, 2003.

W.J. Heiser. Convergent Computing by Iterative Majorization: Theory and Applications in Multidimensional Data Analysis. In W.J. Krzanowski, editor, *Recent Advanmtages in Descriptive Multivariate Analysis*, pages 157–189. Oxford: Clarendon Press, 1995.

W.J. Heiser. Correspondence Analysis with Least Absolute Residuals. *Computational Statistica and Data Analysis*, 5:357–356, 1987.

W.J. Heiser. A Majorization Algorithm for the Reciprocal Location Problem. Technical Report RR-86-12, Department of Data Theory, University of Leiden, 1986.

T.S. Jaakkola and M. I. Jordan. Bayesian Parameter Estimation via Variational Methods. *Statistics and Computing*, 10:25–37, 2000.

K. Lange, D.R. Hunter, and I. Yang. Optimization Transfer Using Surrogate Objective Functions. *Journal of Computational and Graphical Statistics*, 9:1–20, 2000.

P. Verboon and W.J. Heiser. Resistant Lower Rank Approximation of Matrices by Iterative Majorization. *Computational Statistics and Data Analysis*, 18:457–467, 1994.

Department of Statistics, University of California, Los Angeles, CA 90095-1554

E-mail address, Jan de Leeuw: deleeuw@stat.ucla.edu

URL, Jan de Leeuw: http://gifi.stat.ucla.edu