Quadratic and Cubic Majori

Jan de Leeuw

Majorization algorithms generalize the EM algorithm. In this paper we discuss and compare various quadratic and cubic majorization methods, and apply them to logistic regression.

Majorization

Majorization algorithms [De Leeuw, 1994; Heiser, 1995; Lange et al., 2000] are used with increasing frequency in statistical computation. They generalize the EM algorithm to a much broader class of problems and they can usually be tailored to handle very high-dimensional problems.

The general idea is simple. If we want to minimize f over $X \subseteq \mathbb{R}^n$, we construct a majorization function g on X x X such that

$$
\begin{aligned}
f(x) &\le g(x,y) \qquad && \forall x, y \in X \\
f(x) &= g(x) \qquad && \forall x \in X
\end{aligned}
$$

The majorization algorithm corresponding with this majorization function g updates x at iteration k by

$$x^{(k+1)} \in \underset{x \in X}{\mathbf{argmin}}\, g(x, x^{(k)})$$

unless we already have

$$x^{(k)} \in \underset{x \in X}{\mathbf{argmin}}\, g(x, x^{(k)})$$

in which case we stop. Convergence follows, under some additional simple conditions, from the sandwich inequality, which says that if we do not stop at iteration k, then

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

Quadratic majorization

Majorization functions can be constructed by using classic inequalities such as Cauchy-Schwartz or Jensen, by using convexity, or by using Taylor's theorem. Taylor's theorem usually lead to *quadratic majorization algorithms* (Böhning and Lindsay, 1988), which we explain next. By Taylor's theorem

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

for some ξ on the line between x and y. Now suppose there is a matrix B such that $\partial^2 f(\xi) \lesssim B$, in the sense that

ation

$B-\partial^2 f(\xi)$ is positive semi-definite for all ξ. Then clearly

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

is a majorization function for *f*. We also write $g_y(x)$ for *g(x,y)* to emphasize that g_y is a function of x that majorizes *f* at *y*. Observe that g_y has both the same function value and the same derivative as *f* at *y*. By defining the current *target*

$$z=y-B^{-1}\partial f(y)$$

and by completing the square, we see that

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

Thus step *k* of the majorization algorithm solves the least squares problem

$$\min_{x\in X}(x-z^{(k)})'B(x-z^{(k)})$$

If we choose the matrix B to be scalar, for instance by using an upper bound for the largest eigenvalue of $\partial^2 f(\xi)$ In that case computing the target simplifies, and all majorization subproblems are unweighted least squares problems.

In the case in which *X* is all of $\mathbb{R}^n$ the quadratic majorization algorithm simply becomes

$$x^{(k+1)}=x^{(k)}-B^{-1}\partial f(x^{(k)})$$

By Ostrowski's theorem (Ortega and Rheinboldt, 1970, page 300-301) if this algorithm converges to x^{∞}, then it will in general have a linear convergence rate $1-\lambda(x^{\infty})$, where $\lambda(x^{\infty})$ is the largest eigenvalue $B^{-1}\partial^2 f(x^{\infty})$ A smaller B will give a more rapid convergence rate, but in general we cannot expect to see anything faster than linear convergence. If our bound B is really bad, then we may see very slow linear convergence.

Example

We use maximum likelihood logistic regression as an example because it leads to relatively simple computations and clearly illustrates the principles of algorithm construction. In logistic regression we have an n x p matrix Z with regressors, and we solve for the p regression coefficients β. The negative log-likelihood is

$$f(\beta)=-\sum_{i=1}^{n}n_i z_i'\beta-\sum_{i=1}^{n}N_i\log(1-\pi_i(\beta))$$

where

$$\pi_i(\beta)=\frac{1}{1+\exp(-z_i'\beta)}$$

Since

$$\partial\pi_i(\beta)=\pi_i(\beta)(1-\pi_i(\beta))x_i$$

we have

$$\partial f(\beta)=\sum_{i=1}^{n}N_i\pi_i(\beta)-n_i)z_i$$

and

$$\partial^2 f(\beta)=\sum_{i=1}^{n}N_i\pi_i(\beta)(1-\pi_i(\beta))z_i\otimes z_i$$

where ⊗ is used for the outer product. If we collect the $N_i\pi_i(\beta)(1-\pi_i(\beta))$ in a diagonal matrix $V(\beta)$, then the Hessian is simply $X'V(\beta)X$. Observe that the Hessian is positive semi-definite, and thus the negative log-likelihood is convex (strictly convex if Z is of full column-rank). The gradient is $Z'u(\beta)$, where $u(\beta)$has elements $N_i\pi_i(\beta)-n_i$

The most straightforward algorithm to minimize f is Newton's method, which is simply

$$\beta^{(k+1)}=\beta^{(k)}-(Z'V(\beta^{(k)})Z)^{-1}Z'u(\beta^{(k)})$$

This requires solving a linear system in each iteration, but it will give superlinear, in fact quadratic, convergence to the unique minimum in most cases.

Quadratic majorization can be based on $\pi(\beta)1-\pi_i(\beta))\leq 0.25$ from which $V(\beta))\leq 0.25N$, where N is the diagonal matrix with the N_i. Thus

$$\beta^{(k+1)}=\beta^{(k)}-4(Z'NZ)^{-1}Z'u(\beta^{(k)})$$

which only requires a single matrix inversion. It does still require a matrix multiplication in each iteration.

If K is a bound for the largest eigenvalue of $Z'NZ$, i.e. if $b'Z'NZb\leq Kb'b$ for all b, then

$$\beta^{(k+1)}=\beta^{(k)}-\frac{4}{K}Z'u(\beta^{(k)})$$

will have slower convergence, but will require fewer multiplications per iteration.

Computation

Consider the data from Maxwell (1961, page 64) in Table 1. They indicate the number of boys in a clinic classified as inveterate liars by the resident psychiatrist. We do a simple logistic regression on age, which means Z has a column of ones and a column with the numbers one to five. The maximum likelihood solution for the intercept is $\beta_{(0)}\approx$ -1.1971, while that for the slope is $\beta_{(1)}\approx$0.2737.

Newton's method does not converge if we start too far from the solution, for instance at (1, 1). This is because the Hessian is numerically singular. The algorithm could easily be fixed by building in some safeguards. If we start closer Newton typically converges in less than 10 iterations.

The majorization method using 0.25 $Z'NZ$ as the upper bound for the Hessian typically also uses less than ten iterations, and it converges from any starting point. If we start very far away, for example in (10, 10), then majorization takes about thirty iterations. Asymptotically majorization converges with a rate of about 0.0105, which is very fast and implies an additional two decimals precision in each iteration.

In this example majorization looks particulary good, because the $\pi_i(\beta)$ are all fairly close to 0.5, which implies that the bound we use is quite sharp. If we use the much poorer bound 0.25 KI, using the largest eigenvalue of $Z'NZ$ for K, then the convergence rate becomes 0.9917, we need about 275 iterations for an additional decimal of precision, and the algorithm typically takes around two thousand iterations.

age	n	N-n	N
5-7	6	15	21
8-9	18	31	49
10-11	19	31	50
12-13	27	32	59
14-15	25	19	44

Tabel 1 Boy's Ratings on a Lie Scale

Alternatively, we can use the non-uniform bound $Z'W(\beta)Z$, with $W(\beta)$ the diagonal matrix with elements

$$W_i(\beta)=N_i\frac{2\pi_i(\beta)-1}{2z_i'\beta}$$

These non-uniform bounds for the logit were discovered, independently, by Jaakkola and Jordan (2000) and Groenen et al. (2003). The resulting quadratic majorization algorithm

$$\beta^{(k+1)}=\beta^{(k)}-(Z'W(\beta^{(k)})Z)^{-1}Z'u(\beta^{(k)})$$

converges in less than ten iterations, even if we start at (10, 10), and has a rate of 0.0070.

This example illustrates the main points we want to make. Newton's method without safeguards can go astray, while majorization methods will converge but can be very slow if the bound for the Hessian is imprecise. Logistic regression is, in a sense, atypical because we have a convex negative log-likelihood, and thus a unique optimum. Newton's method is very well behaved, and difficult to improve upon. Also, our example has only two parameters. More complicated functions will no doubt behave quite differently.

Cubic majorization

Next we show that majorization can be used to obtain super-linear convergence under some conditions. Alternatively, we think of this particular use of majorization as a way to safeguard Newton's method. The technique and most of the analysis we use is due to Nesterov and Polyak (2003), although our implementation is quite different.

Let us go back to the general problem of minimizing f on $\mathbb{R}^n$ We can write, in obvious notation,

$$f(x) = f(y) + (x-y)'\partial f(y) + \frac{1}{2}(x-y)'\partial^2 f(y)(x-y) + \\ + \frac{1}{6}\partial^3 f(\xi)(x-y)\otimes(x-y)\otimes(x-y)$$

Now if we can find $K > 0$ such that

$$\partial^3 f(\xi)(x-y)\otimes(x-y)\otimes(x-y) \leq K\|x-y\|^3$$

then

$$g(x,y) = f(y) + (x-y)'\partial f(y) + \frac{1}{2}(x-y)'\partial^2 f(y)(x-y) + \frac{1}{6}\|x-y\|^3$$

is a majorization function of f. Observe that for each y the function g_y is twice continuously differentiable, and has the same first and second derivatives at y as f. This immediately implies super-linear convergence, and in fact quadratic convergence in regular cases (Ortega and Rheinboldt, 1970, page 304).

It may appear, at first sight, that minimizing the majorization function is a complicated non-convex problem, But we shall show, following Nesterov and Polyak (2003), that it can actually be solved by finding the solution of a single well-behaved univariate equation.

Example

Before we discuss how to implement the majorization algorithm associated with this majorization function, we first apply the basic idea to the logistic negative log-likelihood. We find

$$\partial^3 f(\beta) = -\sum_{i=1}^{n} N_i\pi_i(\beta)(1-\pi_i(\beta))(1-2\pi(\beta))z_i \otimes z_i \otimes z_i$$

and since

$$\pi_i(\beta)(1-\pi_i(\beta))(1-2\pi(\beta)) \leq \frac{1}{18}\sqrt{3}$$

we can choose

$$K_T = \frac{1}{18}\sqrt{3}\sum_{i=1}^{n} N_i\|z_i\|^3$$

A more precise bound is

$$K_S = \frac{1}{18}\sqrt{3}\sum_{i=1}^{n}\sqrt{3}\|C\|_\infty$$

where $\|C\|_\infty$ is the largest eigenvalue of

$$C = \sum_{i=1}^{n} N_i\|z_i\|z_iz_i'$$

Observe that K_T uses the trace of C, and thus $K_T \geq K_S$. A smaller bound will mean faster convergence, in particular in regions where quadratic approximation, and consequently Newton's method, is poor.

Implementation

Let us now look in more detail at minimizing the cubic majorization function. The problem is to minimize a function of the form

$$f(x) = x'b + \frac{1}{2}x'Ax + \frac{1}{6}K\|x\|^3$$

over $x \in \mathbb{R}^n$ .We use decomposition to reformulate this problem. Define h on $\mathbb{R}\otimes\mathbb{R}^n$ by $h(\tau,y) = f(\tau y)$. Let S be the half-sphere $\{y \mid \|y\| = 1 \wedge y'b < 0\}$. Then

$$\min_x f(x) = \min_{y\in S}\min_t h(t,y) = \min_{y\in S}\min_t\{t\,y\text{í}b + \frac{1}{2}t^2Ay + \frac{1}{6}Kt^3\}$$

A necessary condition for optimality of (τ, y) is obviously that τ minimizes h for fixed y. Thus τ must be a root of the quadratic equation

$$\frac{1}{2}K\tau^2 + y'Ay\tau + y'b = 0 \tag{1}$$

Because K is positive and $y'b$ is negative, this equation has two real roots, with the positive root corresponding to the minimum. Thus at an optimum $\tau > 0$.

A second necessary condition for optimality is that y minimizes h over S for fixed τ. We forget about the constraint $y'b < 0$ for the moment, and minimize over $\|y\| = 1$. This leads to

$$\tau Ay + b = \lambda y \tag{2a}$$

$$y'y=1 \tag{2b}$$

where λ is a Lagrange multiplier. It follows that at a solution

$$y = -(\tau A - \lambda I)^{-1} b \tag{3}$$

so that λ must be chosen to satisfy

$$b'(\tau A - \lambda I)^{-2} b = 1 \tag{4}$$

Equation (4) is a so-called *secular equation*, which has been studied in various contexts. Relevant results and computational methods are discussed in Melman (1997, 1998). We know from these results that the optimum λ is strictly smaller than the smallest eigenvalue of τA, which implies that y defined by (3) satisfies $y'b < 0$ and is consequently in S.

At any solution of the stationary equations we have, from (2a), that $\tau y'Ay + y'b = \lambda$, which implies, using (1), that

$$\tau = \sqrt{-\frac{2}{K}\lambda} \tag{5}$$

and thus $\lambda < 0$. In the algorithm we use (5) to eliminate τ from (4). Letting $\theta = \sqrt{-\lambda}$ and $B = \sqrt{\frac{2}{K}}A$ we see that (4) becomes

$$b'(B + \theta I)^{-2} b = \theta^2 \tag{6}$$

Solve Equation (6) for θ, and then compute the optimal y and τ from there.

Computation

If we use analyze the Maxwell example with third order majorization, we find global convergence from any starting point. The number of iterations is around 20, which is not particularly impressive. It can probably be improved by computing a smaller K, or by using nonuniform cubic majorization along the lines of the Jaakola-Jordan method.

About the author

Jan de Leeuw is Professor of Statistics at the University of California at Los Angeles. His research interests include multivariate analysis and and computationa statistics. On June 9th professor De Leeuw will present a paper at the "50 Years of Econometrics" conference and on June 12th and June 13th he will give a workshop entitled "Majorization Algorithms for Logit, Probit, and Tobit Models" at the Erasmus University Rotterdam.

References

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

J. De Leeuw (1994) Block Relaxation Methods in Statistics. *Information Systems and Data Analysis*, H.H. Bock, W. Lenski, & M.M. Richter (eds.) Springer Verlag.

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

W.J. Heiser (1995) Convergent Computing by Iterative Majorization: Theory and Applications in Multidimensional Data Analysis. *Recent Advantages in Descriptive Multivariate Analysis*, W.J. Krzanowski (ed.), 157-189. Clarendon Press.

T.S. Jaakkola & M.I. Jordan (2000) Bayesian Parameter Estimation via Variational Methods. *Statistics and Computing* 10, 25-37.

K. Lange, D.R. Hunter & I. Yang (2000) Optimization Transfer Using Surrogate Objective Functions. *Journal of*

Computational and Graphical Statistics 9, 1-20.

A.E. Maxwell (1961) *Analysing Qualitative Data.* Chapman & Hall.

A. Melman (1997) A Unifying Convergence Analysis of Second-Order Methods for Secular Equations. *Mathematics of Computation* 66, 333-344.

A. Melman (1998) Analysis of Third-order Methods for Secular Equations. *Mathematics of Computation* 67, 271-286.

Y. Nesterov & B.T. Polyak (2003) Cubic Regularization of a Newton Scheme and its Global Performance. CORE Discussion Paper 41, Catholic University of Louvain.

J. M. Ortega & W. C. Rheinboldt (1970) *Iterative Solution of Nonlinear Equations in Several Variables.* Academic Press.