J. of Basic Sciences, Cheju Nat'l Univ
13(1), 137~147, 2000

기초과학연구 제주대학교
13(1), 137~147, 2000

Algorithm choice for the Cauchy location problem

Jinhyo Kim

Abstract

Newton-Raphson methods are often used to solve the maximum likelihood estimation problems. The example in this paper illustrates a situation in which the Newton methods cannot be used to solve a maximum likelihood problem due to its chaotic behavior. Bisection, Secant-bracket, Illinois and Dichotomous search algorithms are used for speed comparison as well as guaranteed convergence.

Key Words: Cauchy Distribution; Unimodality; Dichotomous Search; Maximum Likelihood Estimation

1 Introduction

The Cauchy location problem is a thorny one numerically with potential usefulness in applications. The likelihood equation can - and for small n does - have several roots. (cf. Thisted, 1988) Barnett(1966) noted that the likelihood for the location parameter θ of Cauchy distribution is often multimodal based on simulation study. Haas, Bain and Antle(1970) suggested a method for finding the joint ML estimates. Wingo(1983) used Brent's method for the Cauchy problem. Hinkley(1978) employed Newton-Raphson iterative technique for solving the Cauchy likelihood equation without assuring its convergence. It is well-known that Newton-Raphson iterative methods has a fast convergence if it does. However, for guaranteed convergence, derivative-free algorithm(s) for the current Cauchy ML problem is suggested.

Let $X_1, \ldots, X_n$ be iid Cauchy random variables with density proportional to $1/(1+(x-\theta)^2)$ and let $R_n = R_n(X_1, \ldots, X_n)$ be the set of roots of the "Cauchy location likelihood equation":

$$R_n = \{\theta \mid \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log \frac{1}{\pi[1+(X_i-\theta)^2]} = 0\} \tag{1}$$

Let $r_n = \text{card}(R_n)$ be the number of roots. With probability 1 the likelihood equation has only simple roots, which are alternatively local maxima and minima of the likelihood function. Hence if r_n is odd, then there are $(r_n+1)/2$ local maxima and $(r_n-1)/2$ local minima. These $(r_n-1)/2$ false maxima are an embarrassment for the ML method of estimation; their number is really quite small. (cf. Reeds, 1985) The number of local maxima of the Cauchy ML function for location estimation which are not global maxima is asymptotically Poisson distributed with mean $1/\pi$ so that $\Pr\{(r_n-1)/2 = k\} = e^{-\frac{1}{\pi}}/(\pi^k k!)$. (cf. Reeds, 1985) Perlman has shown $\text{card}([-K, K] \cap R_n) \longrightarrow 1$ a.s. for each finite $K > 0$. This implies that the number of roots in R_n is, in a.s. sense, is one. A Monte Carlo study was recommendeded by Copas(1970) for further work in which the probability with which such sample configurations arise could be identified.

It is felt that the Cauchy ML case in this paper is harder than most of other statistical distributions. As suggested by Copas(1970), this paper is to present derivative-free algorithms, with guaranteed convergence based on performance comparison with other selected algorithms.

2 Dichotomous Search

In this section, we review the derivative-free Dichotomous Search method.

(Definition) (Bazarra, Sherali and Shetty, 1993)
A function f is unimodal iff for each x^1, x^2 with $f(x^1) \neq f(x^2)$ and for $0 < \lambda < 1$,

$$f(\lambda x^1 + (1-\lambda)x^2) \leq \max\{f(x^1), f(x^2)\}. \tag{2}$$

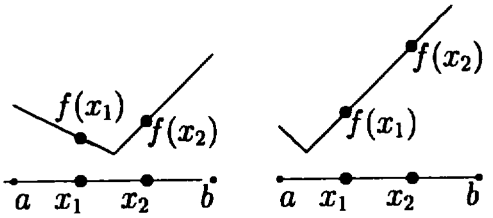

Figure 1: test points

It is well-known that the unimodality is not a sufficient condition for the correct convergence of the Newton-Raphson method. It is easy to find out a counterexample. However, as will be shown below, the unimodality guarantees the correct convergence of the Dichotomous Search method.

Given an initial region of interest $\mathcal{I} = [a, b]$, as described in Figure 1, we evaluate the values at the two test points x_1 and x_2 with $x_1 < x_2$. If $f(x_1) \leq f(x_2)$, then the new interval of 'uncertainty' becomes $[a, x_2]$ since the optimum point cannot exist in $(x_2, b]$. Otherwise if $f(x_1) > f(x_2)$, then the new interval of uncertainty is $[x_1, b]$. Notice that depending on the value comparison of f at x_1 and x_2, the length of the new interval of uncertainty is either equal to $b - x_1$ or $x_2 - a$, which is less than $b - a$. In selecting x_1 and x_2, one usually takes them symmetrically around the midpoint $(b+a)/2$ of a and b with certain distance $\epsilon > 0$. Depending on the values of f at x_1 and x_2, as mentioned above, a new interval of uncertainty is determined. This procedure is repeated with placing two new observations x_1 and x_2 for the next iteration until it terminates. In fact, this procedure works for any $a < x_1 < x_2 < b$. However, a particular choice of $\epsilon = x_2 - (b - a)/2$ yields an optimal algorithm, so called the 'Golden Section Search', with the 'golden number' $\alpha \equiv (x_2 - a)/(b - a) = (\sqrt{5} - 1)/2 \cong 0.618$.

Algorithm for the Dichotomous Search

Initialization step

Choose a constant α (=0.618) and an allowable final length of uncertainty, $l > 0$. Let $[a_1, b_1] = [a, b]$ be the initial interval of uncertainty, and note that the initial interval $\mathcal{I} = [a_1, b_1]$ includes the optimum point, and let $k = 1$ and go to the main step.

Main Step

1. If $b_k - a_k < l$, then stop; the minimum point lies in $[a_k, b_k]$. Otherwise consider x_1 and x_2 defined in Formulae (3) and (4) ; go to step 2.

$$x_1 = a_k + \alpha(b_k - a_k) = (1 - \alpha)a_k + \alpha b_k \qquad (3)$$

$$x_2 = b_k - \alpha(b_k - a_k) = \alpha a_k + (1 - \alpha)b_k \qquad (4)$$

2. If $f(x_1) < f(x_2)$, let $a_{k+1} = a_k$ and $b_{k+1} = x_2$. Otherwise let $a_{k+1} = x_1$ and $b_{k+1} = b_k$. Replace k by $k + 1$, go to step 1.

In Formulae (3) and (4), it should be noted that one of the values at the two test points in current iteration can be reused in the next iteration if the optimal constant $\alpha = 0.618\ldots$ is given, with which the the Golden Section Search is achieved as illustrated in Figures 2. It implies that the Golden Section Search requires only one additional test point in each iteration step. However, due to its floating-point representation in a digital computer (the digital computer cannot recognize irrational numbers), in practice one cannot use the above algorithm as precisely as above. Given $\alpha = 0.618\ldots$, the test point in next iteration step will be only mathematically coincided, but not computationally. Therefore, it is recommended to store 1)the locations x_1 and x_2 and 2)the values $f(x_1)$ and $f(x_2)$, and to reuse one of them without evaluating over again in the next iteration.

The following two theorems, which can be found in Bazarra, Sherali and Shetty (1993), validate the convergence of the Dichotomous Search method.

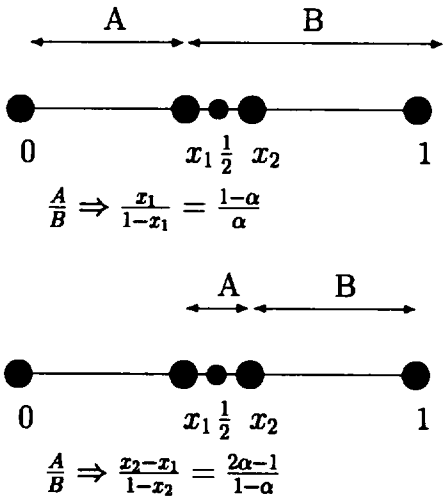
A
B
0
$x_1 \frac{1}{2} x_2$
1
$\frac{A}{B} \Rightarrow \frac{x_1}{1-x_1} = \frac{1-\alpha}{\alpha}$
A
B
0
$x_1 \frac{1}{2} x_2$
1
$\frac{A}{B} \Rightarrow \frac{x_2-x_1}{1-x_2} = \frac{2\alpha-1}{1-\alpha}$

Figure 2: Using the fixed ratio α at k<u>th</u> and *(k+1)*<u>th</u> iteration

(Theorem 1) *Let $f : R^1 \longrightarrow R^1$ be unimodal over $[a, b]$ with $x_1 \leq x_2 \in [a, b]$. If $f(x_1) \leq f(x_2)$, then $f(z) \geq f(x_1)$ for all $z \in (x_2, b]$; and if $f(x_1) > f(x_2)$, then $f(z) \geq f(x_2)$ for all $z \in [a, x_1)$.*

From Theorem 1, we know that if $f(x_1) < f(x_2)$, then there must not exist an optimum point in $[x_2, b]$ since $f(z) \geq f(x_1)$ for all $z \in [x_2, b)$. Now we eliminate the region $[x_2, b)$ to get the new interval of uncertainty $[a, x_2]$ for the next iteration step. In this way, our region of interest will be reduced in each step until we reach the optimum point within an allowable final length of uncertainty. A similar argument follows for the case of $f(x_1) \geq f(x_2)$.

(Theorem 2) *Consider the problem of minimizing a unimodal function $f(x)$ defined on an open set $S \subset \mathcal{R}^p$. If x is a local optimal solution using the Dichotomous Search method, then x is also the global solution.*

By Theorem 2, we are convinced that if we find *one* local optimum point of a unimodal function $f(x)$, then it is necessarily *the* global optimum point. In the above two theorems, whenever a function to be minimized is unimodal, the Dichotomous Search always finds the optimum point. This is definitely a great advantage over the Newton-Raphson method, since Theorem 2 does not hold for the Newton-Raphson method.

3 Simulation Study

Set the parameter $\theta = 2$, for example, and generate a random sample from the Cauchy distribution using the known formula:

$$\tan[\pi(U - \frac{1}{2})] + \theta \sim \text{Cauchy}(\theta) \tag{5}$$

where U has the *Uniform*(0,1) distribution. The pre-setted value $\theta = 2$ is irrelevant since it is location-equivariant. In Figure 3, the object function $l(\theta)$ of the log-likelihood is unimodal on the range of $-10 \leq \theta \leq 10$ and has minimum value around $\theta = 2$, as is expected from the presetting of $\theta = 2$. The graphs of $\ell(\theta)$ are observed with unimodality by inspection. But unfortunately no analytic method was found to verify its unimodality. Even though it is not proved or verified that f is always unimodal, we assume it is unimodal in the current simulation study. The plots are visually examined more than 100 times; all of them exhibits unimodality.

Newton-Raphson method

With an initial value of $\theta_{INIT} = 3$, it converges to $\hat{\theta} = 2.1501426$ after 6 iterations. However, this method will diverge if we roughly set $|\theta_{INIT}| \geq 3$; the Newton-Raphson for a unimodal function may diverge. Given an initial value of 1, the iteration ends to $\hat{\theta}_{\text{MLE}} = 0.1501426$ after 6 iteration steps.

Bisection

With initial two values of $x_0 = -10$ and $x_1 = 10$, it converges to $\hat{\theta} = 2.1501426$ after 29 iterations. We notice that the condition $f(x_0)f(x_1) < 0$ guarantees the convergence by the Mean Value Theorem. Suppose we are given the condition $f(x_0)f(x_1) < 0$ for initial values of x_0 and x_1, then this

	bisection		secant-Bracket		Illinois		golden section	
initial interval	[1]	[2]	[1]	[2]	[1]	[2]	[1]	[2]
$[-1,1]$	26	27	8	6	6	5	34	35
$[-10,10]$	29	29	15	14	10	12	39	37
$[-50,50]$	31	32	19	19	13	12	43	42
$[-100,100]$	32	33	21	20	13	13	44	42

Table 1: (Comparison of speed) [1]:$n = 15$ [2]:$n = 30$

method has a guaranteed convergence unlike the Newton-Raphson method.

Secant-bracket method

An alternative to Newton-Raphson method is to approximate the derivative by a finite difference such that $f'(x) \approx (f(x_i) - f(x_{i-1}))/(x_i - x_{i-1})$. It uses the secant line through two successive points. This method also has no guaranteed convergence. Given two values x_i and x_{i-1}, the secant of $f(x)$ is the line intersecting the curve $f(x)$ at $(x_i, f(x_i))$ and $(x_{i-1}, f(x_{i-1}))$. However it exhibits relatively fast convergence if it converges. Note that the derivatives are estimated from the previous iterations rather than being supplied analytically. For initial value of $x_0 = -10$ and $x_1 = 10$, $\hat{\theta}_{\text{MLE}}$ converges to $\hat{\theta} = 0.1501426$ after 14 iteration steps. For same initial two point same as in Bisection method, the Secant-Bracket exhibits fast convergence.

Illinois method

For initial value of $x_0 = -10$ and $x_1 = 10$, $\hat{\theta}_{\text{MLE}}$ converges to $\hat{\theta} = 0.1501426$. This Illinois method shows convergence after 12 iteration steps, which is faster than Bisection and Secant-Bracket.

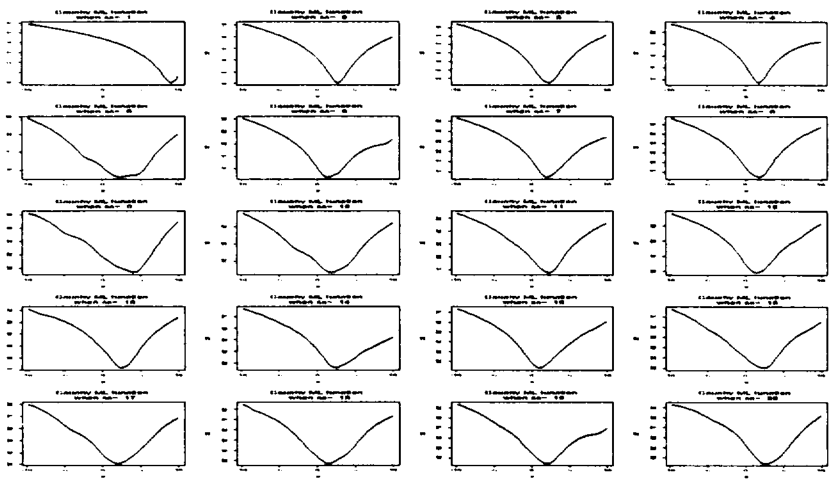

Figure 3: Cauchy ML surface with respect to sample size

4 Concluding remarks and future study

As we can see, neither the Newton method nor the Secant method is recommended. The Secant method does not have a guaranteed convergence of the ML function with multiple local maxima. It is well-known that the Newton method is not reliable since it sometimes diverges to the choice of the initial value and sometimes converges too slowly; for the current Cauchy location problem, the Newton method does not converge to the choice of the initial value. But the Dichotomous Search method can be used with guaranteed convergence whenever the graph shows unimodality and furthermore we can even predict how long it takes. This particular simulation study shows that the Illinois, secant-bracket, bisection, golden-section methods are in descending order in speed comparison. The Illinois algorithm performs speediest convergence than other selected methods. Evidently the Newton-Raphson algorithm has worst convergence. The number n of simulated data was not relevant in the situation. However, for guaranteed convergence, the Dichotomous Search was recommended even with relatively slower speed.

Based on the above 4 algorithms, we summarize the result in Table 2. It is evident that the Newton-Raphson method cannot be a proper choice for the Cauchy location problem. The Illinois method has faster convergence than the Bisection method and Secant-Bracket methods. The golden section search with slower convergence than the bisection and the secant-bracket method has a guaranteed convegence. The number of data is not relevant for the convergence speed.

An extension of the one-dimensional Dichotomous Search to the multi-dimensional parameter case, the 'Iterated Grid Search', can be applied to the Cauchy location-scale problem described in Appendix. However the analytic argument why the Cauchy ML function with location and scale parameters shows mostly unimodality phenomenon is not satisfactorily solved yet. Similar results can be derived in many other cases; 1)replace the Cauchy distribution of the data by some other iid distributions and 2)replace ML method of estimation by another M-estimate method. In many cases simpler arguments than in the Cauchy distribution are sufficient to count the number of false maxima. Typically LLN alone are sufficient. Stochastic optimization tools including simulated annealing is suggested for future study.

5 Appendix

Consider $X_1, \ldots, X_n$ be iid random variables from the Cauchy distribution with density proportional to $\sigma/\{\sigma^2 + (x-\theta)^2\}$ which produces log-likelihood $\ell(\theta, \sigma; x_1, \ldots, x_n) = n \log \sigma - \sum \log\{\sigma^2 + (x_i - \theta)^2\}$. Setting $\partial\ell/\partial\theta = \partial\ell/\partial\sigma = 0$ gives the set of nonlinear equations $\sum_{i=1}^n \psi_1(\theta, \sigma; x_i) = 0$ and $\sum_{i=1}^n \psi_2(\theta, \sigma; x_i) = 0$ with respect to θ and σ where

$$\psi_1(\theta, \sigma; x_i) = \frac{x_i - \theta}{\sigma^2 + (x_i - \theta)^2} \tag{6}$$

$$\psi_2(\theta, \sigma; x_i) = \frac{\sigma^2}{\sigma^2 + (x_i - \theta)^2} - \frac{1}{2} \tag{7}$$

5.1 Scale estimation with fixed location parameter

When equation (7) is satisfied, it is easy to show

$$\frac{\partial^2 \ell}{\partial \sigma^2} = -4\sum_{i=1}^{n} \frac{(x_i - \theta)^2}{\left[\sigma^2 + (x_i - \theta)^2\right]^2} < 0 \tag{8}$$

$$\frac{\partial^2 \ell}{\partial \theta^2} = \sum_{i=1}^{n} \frac{-\sigma^2 + (x_i - \theta)^2}{\left[\sigma^2 + (x_i - \theta)^2\right]^2} \tag{9}$$

$$\frac{\partial^2 \ell}{\partial \sigma \partial \theta} = \sum_{i=1}^{n} \frac{-2\sigma(x_i - \theta)}{\left[\sigma^2 + (x_i - \theta)^2\right]^2} \tag{10}$$

Therefore, for $\theta = \theta_0$ fixed, $L(\theta, \sigma) = L(\theta_0, \cdot)$ as a function of σ alone is unimodal upwardand even convex.

5.2 Location estimation with fixed scale parameter

The maximum likelihood estimation problem in the Cauchy distribution for the location, when evaluated at the MLE of the scale, is unimodal considerably more often than might have been anticipated from the single-parameter case. If the scale parameter is eliminated by a conditional ML argument, giving the maximized likelihood function, then the resulting function is always unimodal.

References

Bai, Z. and Fu, J. (1987), "On the maximum-likelihood estimator for the location parameter of a Cauchy distribution," *The Canadian journal of Statistics.* **15**(2):137-146

Balmer, D., Boulton, M. and Sack, R. (1974), "Optimal solutions in parameter estimation problem for the Cauchy distribution," *Journal of the American Statistical Association.* **69**:238-242

Bazarra, M., Sherali, H. and Shetty, C. (1993), *Nonlinear Programming: Theory and Algorithms.* 2nd edition. John Wiley & Sons.

Copas, J. (1975), "On the Unimodality of the likelihood for the Cauchy distribution," *Biometrika.* **62**:701-704

Ferguson, T. (1978), "Maximum likelihood estimates of the parameters of the Cauchy distribution for samples of size 3 and 4," *Journal of the American Statistical Association.* **73**:211-213

Haas, G. and Bain, L. (1970), "Inferences for the Cauchy distribution based on maximum likelihood estimates," *Biometrika.* **57**:403-408

Hinkley, D. (1978), "Likelihood inference about location and scale parameters," *Biometrika.* **65**:253-261

Kim, J. (1997), *Iterated Grid Search Algorithm on Unimodal Criteria.* Doctoral dissertation, Department of Statistics, Virginia Tech.

—— (1999), "An iterative algorithm for the Cramer-von Mises estimator," *InterStat.* March 1999
`http://interstat.stat.vt.edu/interstat`

Reeds, J. (1985), "Asymptotic number of roots of Cauchy location likelihood equations," *Annals of Statistics.* **13**:775-784

Thisted, R. (1988), *Elements of Statistical Computing -Numerical Computation.* Chapman and Hall.

Wingo, D. (1983), "Estimating the location of the Cauchy distribution by Numerical Global Optimization," *Communications in Statistics -Simulation and Computation.* **12**:201-212