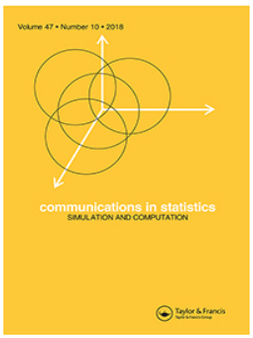

Taylor & Francis
Taylor & Francis Group

Communications in Statistics - Simulation and Computation

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

Estimating the location of the cauchy distribution by numerical global optimization

Dallas R. Wingo

To cite this article: Dallas R. Wingo (1983) Estimating the location of the cauchy distribution by numerical global optimization, Communications in Statistics - Simulation and Computation, 12:2, 201-212, DOI: 10.1080/03610918308812311

To link to this article: https://doi.org/10.1080/03610918308812311

COMMUN. STATIST.-SIMULA. COMPUTA., 12(2), 201-212 (1983)

ESTIMATING THE LOCATION OF THE CAUCHY DISTRIBUTION BY NUMERICAL GLOBAL OPTIMIZATION

Dallas R. Wingo

MIS, Algorithm Development, AT&T Long Lines
Bedminster, New Jersey

Key Words and Phrases: Cauchy distribution; parameter estimation; miximum likelihood; global optimization; numerical methods.

ABSTRACT

The log-likelihood function (LLF) of the single (location) parameter Cauchy distribution can exhibit up to n relative maxima, where n is the sample size. To compute the maximum likelihood estimate of the location parameter, previously published methods have advocated scanning the LLF over a sufficiently large portion of the real line to locate the absolute maximum. This note shows that, given an easily derived upper bound on the second derivative of the negative LLF, Brent's univariate numerical global optimization method can be used to locate the absolute maximum among several relative maxima of the LLF without performing an exhaustive search over the real line.

1. PROBLEM

The single (location) parameter Cauchy distribution has, for a complete sample of n independent observations x_1, $x_2,\ldots,x_n$, probability density

$$pdf\{x;\theta\} = 1/[\pi(1+(x-\theta)^2)], \qquad (1)$$

 0361-0918/83/1202-0201$3.50/0

and log-likelihood function (LLF)

$$L(\theta) = -n \log(\pi) - \Sigma \log[1+(x_i-\theta)^2], \qquad (2)$$

where $x,\theta \in (-\infty,+\infty)$. This model is a member of the family of symmetric, stable Paretian distributions and, as such, is receiving increasing attention as an alternative to the Normal distribution in modelling experimental error. Because it is potentially useful in applications, the Cauchy model has generated much interest in the problem of estimating its location θ.

We note, parenthetically, that the maximum likelihood (ML) estimation of the location and scale parameters of the two-parameter Cauchy distribution (Ferguson, 1978) is a well-conditioned optimization problem, as Copas (1975) has shown that the joint LLF of the two parameters is unimodal. Hence, the numerical maximization of the LLF for the two-parameter Cauchy distribution is numerically straightforward. However, estimation of θ in the single (location) parameter Cauchy distribution (1) is far from routine. The method of moments cannot be applied, and if the method of ML is attempted, one finds that (2) is frequently multimodal. Indeed, Barnett (1966a), Ferguson (1978) and Edwards (1972) have observed that if the observations x_i $(i = 1,2,\ldots,n)$ are sufficiently far apart, (2) will exhibit a local maximum near each $\theta = x_i$. This observation led Barnett (1966a) to conclude that the only "sure" way to obtain the maximum likelihood estimate (MLE) $\hat{\theta}$ is to perform a complete scan of (2) on a sufficiently large portion of the real line. It would thus appear that (2) exhibits a structure that has defied attempts to develop more efficient solution methods. Fortunately, this dilemma can be resolved by drawing freely upon recent algorithmic and computational advances in nonlinear, nonconvex optimization techniques. Specifically, we show that, given an easily derived upper bound on the second derivative of the negative LLF, a method of unidimensional global optimization due to Brent (1973) can be used to determine $\hat{\theta}$. Brent's method is an efficient and

finitely convergent algorithm that is guaranteed to find $\hat{\theta}$ without having to scan the LLF over the real line. Moreover, although a positive upper bound on the second derivative of the negative LLF is assumed known, the method does not require the explicit evaluation of any derivatives whatsoever.

2. METHOD OF SOLUTION

The MLE $\hat{\theta}$ can be determined by solving the unidimensional optimization problem

$$\text{minimize } S(\theta) \equiv -L(\theta), \quad \theta \in [a,b], \quad -\infty < a,b < \infty \tag{3}$$

for the global minimum on $[a,b]$. Since $S(\theta)$ can possess up to n local minima on the real line (Barnett, 1966a), the global minimization of this function can be difficult and time-consuming unless suitable a priori information regarding the properties of the objective function, such as an upper bound M on the second derivative of $S(\theta)$, is available and used.

A bound on the second derivative of $S(\theta)$ implies a limit on the number of local minima that can occur in $[a,b]$. It also implies the existence of a minimum separation between different local minima. The lower the bound, the fewer the number of potential local minima and the larger the interval about each point which cannot contain a minimum.

Fortunately, an upper bound on the second derivative of $S(\theta)$ can be derived rather easily. Observe that

$$-(x_i-\theta)^2 + 1 \leq [(x_i-\theta)^2 + 1]^2, \quad (i=1,\ldots,n).$$

Upon dividing throughout by the right-hand side of this inequality, summing over i, and multiplying by 2, we have that $\partial^2 S(\theta)/\partial\theta^2 \leq 2n$. Thus, $M = 2n$.

In view of the availability of M, problem (3) can be solved by using a finitely convergent method of univariate global optimization due to Brent (1973). Brent's method will find $\hat{\theta} \in [a,b]$ and $\hat{S} = S(\hat{\theta})$ satisfying $|\hat{S} - S_{min}| \leq t$ for a given positive tolerance t, where $S_{min} = \min S(\theta)$ for all $\theta \in [a,b]$. The function $S(\theta)$ is assumed to be twice continuously differen-

tiable on an open interval of the real line containing [a,b], and the second derivative of $S(\theta)$ is assumed to satisfy $\partial^2 S(\theta)/\partial\theta^2 \leq M$ for all $\theta \in [a,b]$, where $M > 0$.

Brent's method starts with a current estimate $\hat{S}$ of the global minimum value of $S(\theta)$ on $[a,b]$ and a corresponding $\hat{\theta} \in [a,b]$ satisfying $\hat{S} = S(\hat{\theta})$. The algorithm seeks to successively improve $\hat{S}$ by evaluating $S(\theta)$ at judiciously chosen points a_3 in $[a,b]$. This is done by setting $a_2 \leftarrow a$ at the start of the algorithm and, on successive steps, attempting to find a point $a_3 \in (a_2,b]$ that, hopefully, will result in a decrease in $S(\theta)$. Any such point a_3, even if it does not result in an immediate decrease in $S(\theta)$, is acceptable if the convex parabola $P(\theta)$, with leading term $M\theta^2/2$ and which coincides with $S(\theta)$ at the points $(a_2, S(a_2))$ and $(a_3, S(a_3))$, satisfies $\hat{S} - P(\theta) \leq t$ for all $\theta \in [a_2,a_3]$. However, if this inequality is not satisfied for the current a_3, then a_3 is successively reduced by setting $a_3 \leftarrow (a_3+a_2)/2$ and the above inequality is re-tested until an acceptable a_3 is found. Once an acceptable a_3 is determined, a_2 is updated by setting $a_2 \leftarrow a_3$. If this a_3 also satisfies $S(a_3) < \hat{S}$, then $\hat{\theta}$ and $\hat{S}$ are also updated by setting $\hat{\theta} \leftarrow a_3$ and $\hat{S} \leftarrow S(a_3)$. The search for a new $a_3 \in (a_2,b]$ is begun from the current a_2. This cycle is repeated until $a_2 \geq b$.

The choice of a_3 should, of course, be made with a view toward keeping the number of evaluations of $S(\theta)$ to a bare minimum. In particular, a_3 should be chosen as large as possible, but not so large that it has to be reduced, since each reduction of a_3 and subsequent test of the inequality $\hat{S} - P(\theta) \leq t$ for all $\theta \in [a_2,a_3]$ will necessitate additional evaluations of $S(\theta)$. Brent (1973; p. 87) has shown that there is no risk that a_3 will have to be reduced if we always choose $a_3 = a_3^*$, where

$$a_3^* = \min\{b,\ a_2 + [(S(a_2) - \hat{S} + t)/(M/2)]^{1/2}\}.$$

Brent (1973; p. 88) has also noted that when $a_2 > a$ and $S(\theta)$ is decreasing rapidly at a_2, it is possible to speed up the

algorithm by choosing $a_3 > a_3^*$. However, the choice $a_3 = a_3^*$ is always safe and guarantees that a_3 will never have to be reduced, even when $a_2 = a$.

The maximum number of function evaluations necessary to minimize $S(\theta)$ on $[a,b]$ to an accuracy t is of order $(M/t)^{1/2}$ (Brent, 1973; p. 84). However, Brent has incorporated into his algorithm several heuristic strategies to improve the overall computational efficiency of the method, with the result that this upper bound is rarely, if ever, achieved in practice. The aim of these heuristics is to reduce, as far as possible, the number of function evaluations required to give an answer which is guaranteed to be accurate to within the prescribed tolerance t. These strategies include (i) the use of a crude random search during the early stages of the algorithm, (ii) intermittent local parabolic interpolation using the three most recent points at which $S(\theta)$ has been evaluated, and (iii) the option to input an estimate of the position of the global minimum of $S(\theta)$. Strategies (i) and (iii) aim to reduce $S(\theta)$ quickly in the initial stages of the search, while strategy (ii) helps to locate the local minima which are in the interior of $[a,b]$. Unless the global minimum of $S(\theta)$ is at one of the endpoints of $[a,b]$, one of these local minima is the global minimum of $S(\theta)$ on $[a,b]$. When $S(\theta)$ is sufficiently well-behaved in the vicinity of the global minimum, the use of strategy (ii) will also enable the global minimum of $S(\theta)$ to be determined more accurately than would be expected in theory.

Further theoretical and practical implementational details are given by Brent, who also gives a Fortran function subprogram (GLOMIN) implementation of his method. The figures reported in subsequent sections of this paper were obtained using GLOMIN.

3. NUMERICAL EXAMPLES

As illustrations of the practical application of the methods described in the previous section of this note, we consider three

sets of data. The first set, Sample A, consists of four observations {3, 7, 12, 17} given by Edwards (1972, p.163). The remaining two sets of observations, Sample B and Sample C, are given in Table I. The results of using GLOMIN to solve (3) for each of these samples is summarized in Table II. Here, N_f is the number of evaluations of $S(\theta)$ that were required to satisfy $|\hat{S}-S_{min}| \leq t$, where $S_{min} \equiv \min S(\theta)$ for $\theta \in [a, b]$. Also, $M = 2\pi^2$ is the upper bound on the second derivative of $S(\theta)$. The value $t = 10^{-8}$ was chosen so as to yield approximately three decimal digits of accuracy in computed values of $\hat{\theta}$.

Although in principle the starting value for GLOMIN can be any value on [a, b], the sample median was used as a starting value θ°. This was done for two reasons. First, the results of a simulation study carried out by Barnett (1966a) indicate that the global maximum of the LLF is frequently the local maximum closest to the sample median. Therefore, conceivably, fewer evaluations of the LLF would be required to obtain the MLE $\hat{\theta}$ when the sample median is used as a starting value.

Second, although the sample median is a consistent, unbiased and easily calculated estimator of θ, the choice between the ML and median estimator of θ is clear cut when their respective efficiencies are compared. Barnett (1966b) has shown that the ML estimator of θ is more than 90% efficient for $n = 20$, whereas the median estimator is only about 72% efficient for the same value of n. However, as Table 4 of Barnett (1966b) also shows, even for $n < 20$, the efficiency of the ML estimator is significantly greater than the corresponding efficiency of the median estimator of θ. Moreover, the ML estimator becomes fully efficient asymptotically, whereas the median estimator of θ achieves only about 80% efficiency, asymptotically. Therefore, inasmuch as the numerical methods described in this paper now make the estimation of θ a routine matter, the superior efficiency of the ML estimator vis-à-vis the efficiency of the median estimator of θ justifies the computation required to compute the MLE $\hat{\theta}$.

TABLE I

Sample Data

Sample B										
	2.0	5.0	7.0	8.0	11.0	15.0	17.0	21.0	23.0	26.0
Sample C										
	4.1	7.7	17.5	31.4	32.7	92.4	115.3	118.3	119.0	129.6
	198.6	200.7	242.5	255.0	274.7	274.7	303.8	334.1	430.0	489.1
	703.4	978.0	1656.0	1697.8	2745.6					

TABLE II

Results of Estimating $\hat{\theta}$ Using GLOMIN for Samples A-C

Sample	n	a	b	t	M	θ^{o}	$\hat{\theta}$	S_{min}	N_f	N_e
A	4	3	17	10^{-8}	8	9.5	7.062	-15.28	71	14×10^3
B	10	2	26	10^{-8}	20	13.0	7.728	-44.95	107	24×10^3
C	25	4.1	2745.6	10^{-8}	50	242.5	118.497	-261.78	1048	2731.5×10^3

The interval endpoints a and b are set, respectively, to the smallest and largest sample order statistic since, as pointed out by Barnett (1966), the global maximum of $L(\theta)$ will not occur at any value of θ less (greater) than the smallest (largest) sample observation. For comparative purposes, Table II also gives figures for N_e, where N_e represents the number of evaluations of $S(\theta)$ that are required to scan $S(\theta)$ on $[a, b]$ in steps of 10^{-3}. The large values of N_e presented in Table II make it rather clear that scanning $S(\theta)$ to locate $\hat{\theta}$ is neither an attractive nor a practical proposition, particularly when the sample range is large. Moreover, such an approach introduces the complication of having to choose a grid spacing of sufficient size to locate $\hat{\theta}$ to the desired degree of precision, and in as few evaluations of $S(\theta)$ as practicable.

The MLE $\hat{\theta}$ can also be determined by finding all of the real zeros of the equation $\partial L/\partial\theta = 0$ and then selecting the zero(s) for which $L(\theta)$ is a global maximum on $[a, b]$. In explicit form, this equation is

$$\Sigma(x_i-\theta)/[(x_i-\theta)^2 + 1] = 0. \tag{4}$$

Equation (4), when rationalized, becomes

$$\Sigma_i [(x_i-\theta) \prod_{\substack{j=1 \\ j\neq i}}^{n} \{(x_j-\theta)^2 + 1\}] = 0. \tag{5}$$

Equation (5) is a polynomial of degree $2n-1$ in θ. Unfortunately, finding all of the real zeros of (4) or (5) is easier said than done. Finding all of the zeros of a polynomial to a specified accuracy is a nontrivial problem computationally. This is especially so for large n. Moreover, the use of (5) is not recommended since the product expression in (5), when evaluated, can lead to computer floating point overflow problems. Nevertheless, using the sample median as a starting value θ^o, an attempt was made to find all of the distinct, real zeros of $\partial L/\partial\theta = 0$ by applying Newton's and Muller's methods (with deflation) to (4) for each of the Samples A-C. Muller's method

(Muller, 1956) possesses the following desirable attributes: (i) it does not require the evaluation of any derivatives of (4); (ii) it is almost as fast as Newton's method, since its asymptotic convergence rate is 1.84 as compared to 2 for Newton's method; (iii) it tends to be less sensitive to bad starting values than is Newton's method.

Equation (4) possesses seven distinct, real zeros on [3, 17] for Sample A. Those zeros at 3.709, 7.062, 11.878 and 16.523 are relative maxima of $L(\theta)$, while those at 4.004, 9.580 and 15.490 are relative minima. The global maximum is, of course, at 7.062. These findings are in complete agreement with those of Edwards (1972, pp. 163-164), who demonstrates graphically that $L(\theta)$ possesses a local, relative maximum near each of the four observations of Sample A and a local, relative minimum between each of the relative maxima. It can also be verified numerically that, on $[x_{(1)}, x_{(n)}]$, $L(\theta)$ possesses 4 maxima and 3 minima for Sample B, and 17 maxima and 18 minima for Sample C. Here, $x_{(1)}$ and $x_{(n)}$ are, respectively, the smallest and largest sample order statistics. The exact values to three decimal places, of the relative optima of $L(\theta)$ for Samples A-C are summarized in Table III.

Unfortunately, neither Newton's method nor Muller's method worked consistently well in solving (4). Both methods required several applications of the algorithm, and the use of starting values other than the sample median, to obtain convergence to a zero. Newton's method diverged frequently, and for none of the samples was it able to find all of the zeros of (4) in a single run of the algorithm. On the other hand, Muller's method, while able to find all of the real, distinct zeros in a single run for Sample A, was able to find at most 4 zeros in a single run for Sample B and at most 5 zeros in a a single run for Sample C.

While no general conclusions valid for all sample sizes can be drawn from the computational results presented for Samples A-C, it is not unreasonable to conclude that the use of GLOMIN to locate $\hat{\theta}$ is generally preferable to the alternative strategy of

TABLE III

Zeros of (4) and Relative Optima of $L(\theta)$

	Maxima	Minima
Sample A		
	3.709	4.004
	7.062 (G)	9.580
	11.878	15.490
	16.523	
Sample B		
	7.728 (G)	9.871
	10.759	13.083
	15.143	19.740
	20.757	
Sample C		
	7.714	11.780
	17.558	22.607
	31.982	57.368
	118.497 (G)	97.710
	129.333	126.871
	200.052	168.090
	242.576	223.096
	254.976	248.334
	274.654	264.024
	303.681	297.307
	333.977	327.111
	429.945	414.829
	489.031	476.149
	703.363	678.028
	977.973	942.673
	1656.008	1554.422
	1697.761	1680.120
(G) global maximum		2656.816

attempting to compute zeros of (4). This conclusion seems reasonable because of the generally wasted computational overhead that is incurred by having to compute all of the distinct, real zeros of (4) and then having to evaluate $L(\theta)$ for those zeros which correspond to relative maxima of $L(\theta)$. While this overhead is generally negligible for small samples, it is likely to be significant to the point of dominating the total computational effort for large samples.

Finally, it must be stressed that convergence of GLOMIN to the global maximum of $L(\theta)$ is guaranteed (Brent, 1973). Convergence reliability can be an important practical consideration

when the problem of estimating $\hat{\theta}$ is part of a larger, more complex calculation as, for example, in a statistical simulation system. In such applications, it is essential that the calculations proceed without user intervention or monitoring.

4. CONCLUSIONS

The work described in this paper was motivated by the desire to avoid certain computational limitations of earlier proposed methods for estimating the location parameter of the Cauchy distribution. These methods include computing real zeros of (4) or scanning (2) over a portion of the real line using a sufficiently fine grid (Barnett, 1966a). Since it is not usually known which of the zeros of (4) corresponds to $\hat{\theta}$, the former method entails a considerable waste of computational effort, because of the necessity to compute zeros which may correspond to relative maxima and minima of (2).

It is difficult to make a general statement regarding the number of evaluations of $S(\theta)$ required by GLOMIN. In particular, it is not currently known how N_f varies with n. Additional computational tests were performed for samples of size $n = 50$ and $n = 100$ but because of space limitations, the sample data have been omitted from Table I. The values of N_f for $t = 10^{-8}$ and these sample sizes are 132 and 180, respectively. The comparatively large value of N_f presented in Table II for Sample C seems to imply that the algorithm is not entirely immune to the effect on N_f of a large sample range. An investigation of this issue, as well as the manner in which N_f varies with n, is in progress and the results will be reported at a later date.

The figures presented in Table II show that scanning (2) over a sufficiently large portion of the real line may entail an enormous number of function evaluations. Such an approach also introduces the complication of having to choose a grid spacing of sufficient size to locate $\hat{\theta}$ to the desired degree of precision, and in as few function evaluations as practicable.

In view of the availability of a bound on the second

derivative of the LLF, the use of Brent's method to maximize the LLF offers a computationally less costly means of locating $\hat{\theta}$. Extensive computational experience with the method seems to indicate that the method can be especially useful in the small sample case. Because of convergence and other computational difficulties, this case is frequently the most troublesome for root-finding approaches to parameter estimation.

BIBLIOGRAPHY

Barnett, V. D. (1966a). Evaluation of the maximum-likelihood estimator where the likelihood equation has multiple roots. Biometrika 53, 151-165.

Barnett, V. D. (1966b). Order statistics estimators of the location of the Cauchy distribution. J. Amer. Statist. Assoc. 61, 1205-1218.

Brent, R. P. (1973). Algorithms for Minimization without Derivatives. Englewood Cliffs: Prentice-Hall Inc.

Copas, J. B. (1975). On the unimodality of the likelihood for the Cauchy distribution. Biometrika 62, 701-704.

Edwards, A. W. F. (1972). Likelihood. Cambridge: Cambridge University Press.

Ferguson, T. S. (1978). Maximum likelihood estimates of the parameters of the Cauchy distribution for samples of size 3 and 4. J. Amer. Statist. Assoc. 73, 211-213.

Muller, D. E. (1956). A method for solving algebraic equations using an automatic computer. MTAC 10, 208-215.

Received by Editorial Board member December, 1981; Revised July, 1982.

Recommended by Wm. A. Coberly, University of Tulsa, Tulsa, OK

Refereed anonymously.