In *Maximum Entropy and Bayesian Methods*,
G. R. Heidbreder, ed., pp. 255-263,
Kluwer Academic, Dordrecht, 1996.

ESTIMATORS FOR THE CAUCHY DISTRIBUTION

K. M. Hanson and D. R. Wolf
Los Alamos National Laboratory, MS P940
Los Alamos, New Mexico 87545 USA
email: kmh@lanl.gov and wolf@lanl.gov

Abstract. We discuss the properties of various estimators of the central position of the Cauchy distribution. The performance of these estimators is evaluated for a set of simulated experiments. Estimators based on the maximum and mean the posterior density function are empirically found to be well behaved when more than two measurements are available. On the contrary, because of the infinite variance of the Cauchy distribution, the average of the measured positions is an extremely poor estimator of the location of the source. However, the median of the measured positions is well behaved. The rms errors for the various estimators are compared to the Fisher-Cramér-Rao lower bound. We find that the square root of the variance of the posterior density function is predictive of the rms error in the mean posterior estimator.

1. Introduction

We explore the properties of various estimators of the central position of the Cauchy distribution, which is notorious for the divergent nature of its first and higher moments. The results of using different kinds of estimators are evaluated by simulating a series of experiments using a Monte Carlo procedure. Investigation of the Cauchy distribution is profitable because its peculiar properties illustrate some interesting aspects of parameter estimation based on Bayesian analysis. It provides us with an example of how to properly deal with data outliers. Some aspects of this paper have been presented in [1].

2. The Cauchy Distribution

2.1. The problem

Suppose that a radioactive source, located at the position (x_0, y_0), emits gamma rays. A position-sensitive linear detector, colinear with the x axis and extending to infinity in both directions, measures the position x_i that the ith gamma ray hits the detector. The data consist of the values x_i, $i = 1, ..., N$, which we designate by the vector $\mathbf{x}$. The problem is to estimate the location of the source x_0, assuming that y_0 is known. This problem is Gull's lighthouse example [2] cast in another setting.

Assume that the gamma rays are confined to the x-y plane and are emitted uniformly in the angle θ at which they leave the source. From the relation $\tan(\theta) = -y_0/(x_i - x_0)$, which holds for $-\pi < \theta < 0$, the probability density function in x_i is obtained by using the Jacobian determinant to transform the density function dependence from θ to x

$$p(x_i|x_0, y_0) = \frac{y_0}{\pi\,[y_0^2 + (x_0 - x_i)^2]} \quad , \tag{1}$$

which is called the likelihood of measurement x_i. The formula in Eq. (1) is proper normalized, i.e. its integral with respect to x_i is unity. Viewed as a function of x_0, the likelihood for this problem is recognized to be a Cauchy distribution in x_0, which is notorious for having an undefined mean and an infinite variance. The width of this distribution may be characterized by its FWHM, which is $2y_0$.

In a Bayesian analysis the posterior probability density function for the x_0 position of the source summarizes the state of knowledge concerning x_0 by providing the probability of every possible value of x_0. The posterior of x_0, given the data $\mathbf{x}$ and the position parameter y_0, is given by Bayes's law

$$p(x_0|\mathbf{x}, y_0) \propto p(\mathbf{x}|x_0, y_0)\, p(x_0|y_0) \propto p(\mathbf{x}|x_0, y_0)\, p(x_0) \quad , \tag{2}$$

where we have assumed that the prior on x_0 is independent of y_0. Proportionality constants are always determined by normalization - the requirement that the probability that some event occurs is unity. If we suppose we have no prior information about the x_0 location of the source, then for the prior $p(x_0)$ we should use a constant over whatever sized region is required. Such a prior is noncommittal about the location of the source. Each measured x_i clearly follows the likelihood, Eq. (1), and, as the emission of one gamma ray can not effect the emission of another, the x_i are statistically independent. Thus the full posterior probability is

$$p(x_0|\mathbf{x}, y_0) \propto \; p(\mathbf{x}|x_0, y_0) = \prod_{i=1}^{N} p(x_i|x_0, y_0) \propto \prod_{i=1}^{N} \left[\frac{y_0}{y_0^2 + (x_0 - x_i)^2} \right] \quad . \tag{3}$$

Again, the normalization is determined by the requirement that the integral of $p(x_0|\mathbf{x}, y)$ over x_0 is unity. From here on, we will often drop explicit mention of y_0 and write the posterior as $p(x_0|\mathbf{x})$.

In the above derivation the posterior probability is the same as the likelihood because the prior is assumed to be a constant. The likelihood expresses the probability of obtaining the specific set of measurements, given a particular x_0. We emphasize that Bayes's law is necessary to gain information about x_0 from the likelihood [2].

If it were known that the source position was limited to a specific region, an appropriate prior would consist of a function that is a nonzero constant inside the region and zero outside. This prior would have the effect of eliminating the tails of the posterior probability in (2) outside the legitimate region. This prior would alleviate any problem that might exist with the normalization of the prior.

2.2. Monte Carlo simulation

To numerically test how well various estimators of x_0 perform, we need to generate measurements that simulate a series of experiments. The cumulative probably (also called the distribution function), the probability of a measurement with $x_i < u$, is given by

$$P(x_i < u) = \int_{-\infty}^{u} p(x|x_0, y_0)\, dx = \frac{1}{\pi} \tan^{-1} \left(\frac{u - x_0}{y_0} \right) + \frac{1}{2} \quad . \tag{4}$$

To generate measurements from the Cauchy distribution, one uses a pseudorandom number generator that provides a number r_i in the interval (0,1) and then maps the result into the x_i value using the inverse of (4), $x_i = x_0 + y_0 \tan[\pi(r_i - \frac{1}{2})]$.

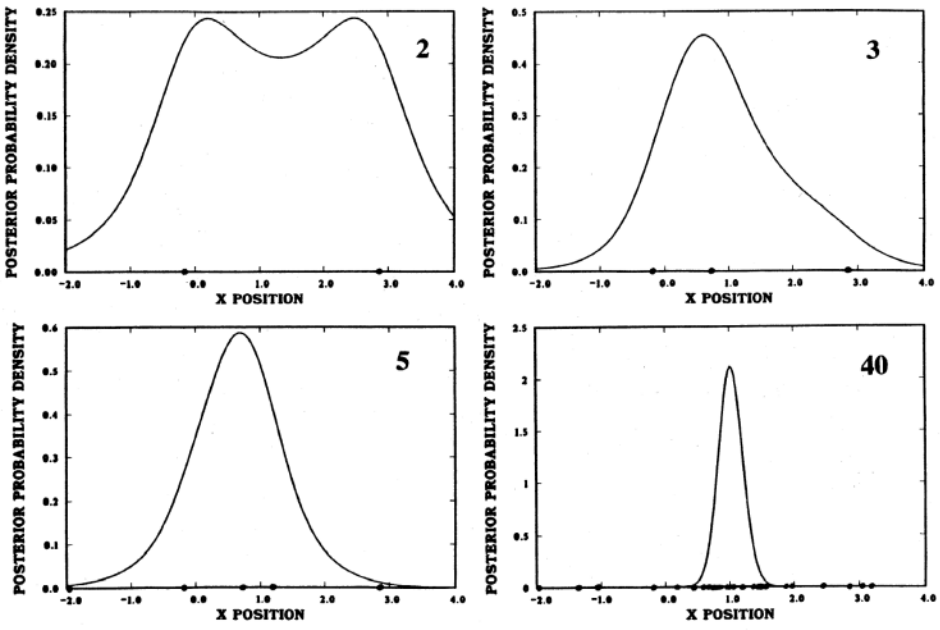

Figure 1: The posterior probability density function for the x_0 position of the radioactive source assuming that the correct y_0 is known. Each of these plots is shown for one simulated experiment; the measured x_i are displayed on the horizontal axis. The number of measurements is noted in the upper-right corner of each plot.

Note that when $u - x_0 \gg y_0$, $P(x_i \geq u) = 1 - P(x_i < u) \approx y_0/\pi(u - x_0)$. The probability of getting an x_i value that is greater than 1000 times the FWHM of the distribution ($2y_0$) is roughly 1/1000. Thus the Cauchy distribution offers a superb example of a data distribution with outliers.

The posterior probability given by Eq. (3) is plotted in Fig. 1 for specific measurements generated using the Monte Carlo technique described above. The plot for two measurements is bimodal, making it ambiguous to use the maximum posterior probability (see Sect. 3.2) to estimate x_0. As the number of measurements increases, the width of the posterior density function decreases, indicating less uncertainty in the knowledge of x_0. The broad tail of the Cauchy likelihood is suppressed as the number of measurements increases because the posterior probability involves a product of likelihoods of the individual measurements.

3. Estimation of Location

3.1. Mean and median of the measurements

The average of the x_i measurements (or samples) is often used to estimate the central position of their distribution:

$$\hat{x}_{0(\text{samp mean})} = \frac{1}{N}\sum_{i=1}^{N} x_i \ . \tag{5}$$

The variance of the average of N samples taken randomly and independently from an arbitrary density function is easily shown to be N^{-1} times the variance of the original density function, provided such variance exists. Curiously the density function for the average of N samples from the Cauchy distribution is identical to that for one sample. Because the variance of the Cauchy density function is infinite, so will be the variance of the average of any finite number of samples. However, for a Gaussian density function, this estimator would be the sufficient statistic for the central position and would be optimal in many ways.

An alternative estimator of the center of a sampled distribution is the sample median $\hat{x}_{0(\text{samp med})}$, which is supposed to be robust against outliers [3, p. 232]. For odd N, the median is defined as the $\frac{1}{2}(N+1)$th sample in the list of magnitude-ordered measurements; for even N, it is defined as the average of the $(N/2)$th and the $(N/2+1)$th samples from such a list.

3.2. Bayesian estimators

The Bayesian viewpoint is that the posterior probability density function for x_0 summarizes our state of knowledge of x_0 in probabilistic terms. Various types of estimators can be formed from the posterior. The choice of estimator can be based on how the cost of making an error in the estimated quantity depends on the size of the error [1]. The most commonly used estimator in Bayesian analysis is the x_0 value at the maximum of the posterior probability, which we designate by $\hat{x}_{0(\text{MAP})}$, because it is usually called the maximum a posteriori estimator. The MAP estimator minimizes a cost function that is zero for no error and a positive constant for any finite error.

The estimate $\hat{x}_0$ that minimizes the expected mean-square error, that is $\int(\hat{x}_0 - x_0)^2\, p(x_0|\mathbf{x})\, dx_0$, is the mean of the posterior density function:

$$\hat{x}_{0(\text{post mean})} = \int x_0\, p(x_0|\mathbf{x})\, dx_0 \ . \tag{6}$$

Defining an integral that is proportional to the kth moment of the posterior given in Eq. (3)

$$I_k(\mathbf{x}) = \frac{y_0}{\pi}\int_{-\infty}^{+\infty} x_0^k \prod_{i=1}^{N} \frac{1}{[y_0^2 + (x_i - x_0)^2]}\, dx_0 \ , \tag{7}$$

the mean (or first moment[1]) of the posterior is

$$\hat{x}_{0(\text{post mean})} = \frac{I_1(\mathbf{x})}{I_0(\mathbf{x})} \ . \tag{8}$$

[1]The kth moment of the posterior is $I_k(\mathbf{x})/I_0(\mathbf{x})$.

The integrand in Eq. (7) has simple[2] poles at $x_i^{\pm} \equiv x_i \pm \imath y_0$. By interpreting the integral as one along the real axis in the complex plane and closing the contour at ∞ in the upper half plane (which contributes nothing provided the integrand falls off faster than x^{-1}), the desired result is found using the Cauchy residue theorem

$$I_k(\mathbf{x}) = \sum_{i=1}^{N} (x_i^+)^k \prod_{j \neq i} \frac{1}{(x_i^+ - x_j^+)(x_i^+ - x_j^-)} \ , \quad 0 \leq k < 2N-1 \tag{9}$$

$$= \sum_{i=1}^{N} (x_i^+)^k \prod_{j \neq i} \left[\frac{1}{(x_i - x_j)^2 + 4y_0^2} \right] \left[1 - \frac{2\imath y_0}{x_i - x_j} \right] \ , \quad 0 \leq k < 2N-1, \tag{10}$$

where the second expression is obtained by simply rearranging the product.

Note that by its definition (7), I_k is real for all allowed k. In particular for $k = 1$, the factor $(x_i^+)^k = x_i + \imath y_0$ in Eq. (10) contributes to two summations, one summation with factor x_i and the other summation with factor $\imath y_0$. The summation with factor $\imath y_0$ is identically $\imath y_0 I_0$. Because I_0 is real, $\imath y_0 I_0$ is imaginary. Thus the $\imath y_0$-factor summation must be exactly cancelled by the imaginary part of the x_i-factor summation and we may write

$$I_1(\mathbf{x}) = \Re \left\{ \sum_{i=1}^{N} x_i \prod_{j \neq i} \left[\frac{1}{(x_i - x_j)^2 + 4y_0^2} \right] \left[1 - \frac{2\imath y_0}{(x_i - x_j)} \right] \right\} \ , \quad 0 \leq k < 2N-1. \tag{11}$$

Therefore, the posterior mean estimator (8) has the form of a weighted average of the x_i, $\hat{x}_{0(\text{post mean})} = \sum w_i x_i$, where the sum of the weights is unity. Although this expression looks like a simple variation on the sample average (5), the weights behave in a very complicated manner. The net effect of the first factor in the product in (11) leads to a diminished contribution from an outlier. But it is very difficult to conceptually grasp the effect of the second factor owing to its complex nature.

Figure 2 shows the behavior of the various estimators when a new measurement is combined with five existing measurements. As the value of the new measurement moves away from the other measurements, its net effect on $\hat{x}_{0(\text{post mean})}$ goes to zero. Thus the estimator minimizes the contribution of any measurement that lies far from a cluster of other measurements, which seems to be an ideal treatment of outliers. Because the posterior is independent of the order of the measurements, the same behaviour is expected for any measurement. The posterior maximum estimator behaves similarly to the posterior mean. A new measurement affects the sample mean in a linear fashion because it is just a linear combination of all measurements. The outlier sample can drastically affect the sample mean. The sample median behaves quite differently. The change in the median reamins constant as long as the $(N+1)$th sample lies outside the central-most two or three samples, depending on whether N is even or odd, respectively. The estimators based on the posterior are the only ones for which the effect of a single disparate measurement decreases as its descrepancy from the others increases.

The variance of the posterior density function of x_0 for a particular data vector $\mathbf{x}$ is

[2]The procedure described here must be trivially modified when $x_i = x_j$ for $i \neq j$. However, we need not consider such coincident measurements because they represent a set of zero probability.

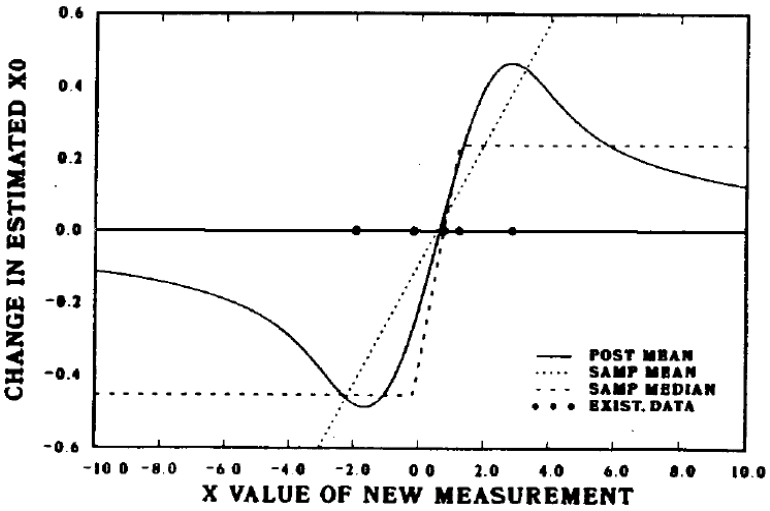

Figure 2: The change in the estimated position of the Cauchy distribution caused by adding a new measurement x_6 to five existing measurements as a function of the value of the new measurement. The results are shown for several kinds of estimators.

$$var\{p(x_0|\mathbf{x})\} = \int [x_0 - \hat{x}_{0(\text{post mean})}]^2 p(x_0|\mathbf{x})\, dx_0 = \frac{I_2(\mathbf{x})}{I_0(\mathbf{x})} - \left[\frac{I_1(\mathbf{x})}{I_0(\mathbf{x})}\right]^2 \ , N \geq 2 \ . \quad (12)$$

An interesting property of the posterior probability is that its shape depends on the data $\mathbf{x}$ and is hence different for every experiment. See Sect. 4 for its relationship to rms error for $\hat{x}_{0(\text{post mean})}$.

3.3. Fisher-Cramér-Rao lower bound and Fisher information

The Fisher-Cramér-Rao bound[3] places a lower bound on the variance of any unbiased estimator $\hat{x}(\mathbf{x})$, $var(\hat{x}) \geq \mathcal{I}_N^{-1}$, where $\mathcal{I}_N$ is the Fisher information

$$\mathcal{I}_N \equiv \int \frac{\partial^2 \log[p(x_0|\mathbf{x})]}{\partial x_0^2} p(x_0|\mathbf{x})\, d\mathbf{x} \ , \quad (13)$$

and where N, the number of measurements, is the dimension of $\mathbf{x}$. Because the posterior (3) factors, we have $\mathcal{I}_N = N\mathcal{I}_1$, where $\mathcal{I}_1$ is the single-sample Fisher information given by

$$\mathcal{I}_1 = \frac{4y_0}{\pi} \int_{-\infty}^{+\infty} \frac{(x - x_0)^2}{[y_0^2 + (x - x_0)^2]^3} dx = \frac{1}{2y_0^2} \ . \quad (14)$$

In the last step the integral is evaluated by applying the Cauchy residue theorem (as in Sect. 3.2). Thus the Fisher-Cramér-Rao lower bound on the variance of any unbiased estimator of x_0 is

$$var(\hat{x}) \geq [N\mathcal{I}_1]^{-1} = \frac{2y_0^2}{N} \ . \quad (15)$$

[3]Fisher stated this lower bound many years before Cramér and Rao [4, p. 66].

Table 1: Summary of the performance of several estimators of the central position of a Cauchy distribution observed in 10^5 trials for a fixed number of samples per trial N. The estimators used are the mean and the median of the samples, and the maximum and mean of the posterior probability density function. The last two columns give the Fisher-Cramér-Rao lower bound on the rms error and the rms width of the posterior probability.

	rms error in estimated position					
N	samp mean	samp median	post max	post mean	CR	rms post
1	2.85×10^{10}	2.85×10^{10}	2.85×10^{10}	2.85×10^{10}	1.414	∞
2	1.43×10^{10}	1.43×10^{10}	$--$	1.43×10^{10}	1.000	1.43×10^{10}
3	9.52×10^{9}	2.828	2.825	2.768	0.816	2.616
5	5.71×10^{9}	1.103	1.070	0.958	0.632	0.963
10	2.86×10^{9}	0.578	0.538	0.522	0.447	0.523
20	1.43×10^{9}	0.373	0.341	0.339	0.316	0.339
40	7.14×10^{8}	0.256	0.236	0.232	0.224	0.232

It is important to note that this lower bound is valid only for unbiased estimators, i.e. when averaged over all possible data, it yields the correct result $\int \hat{x}_0(\mathbf{x})\, p(\mathbf{x}|x_0)\, d\mathbf{x} = x_0$. We have established through subsidary calculations that both the sample median and posterior mean are unbiased for $N \geq 3$ and that their variances exist for $N \geq 4$.

4. Simulation Results

The performance of the above estimators for x_0 is tested by simulating 10^5 experiments, each involving a fixed number of measurements, which are independently drawn from a Cauchy distribution as indicated in Sect. 2.2. The parameters are held fixed at $x_0 = 1$ and $y_0 = 1$ throughout. The results are summarized in Table 1. In these numerical experiments, except for the sample mean, the bias is always observed to be consistent with zero to within its statistical uncertainty, i.e. on the order of the [rms error of the estimator] $/\sqrt{T}$, where T is the number of trials.

We observe that the average value of the measurements performs terribly! This poor performance was anticipated, owing to the infinite variance of the Cauchy distribution. The only reason that the rms error in $\hat{x}_{0(\text{samp mean})}$ is not infinite, as mentioned, is that only a finite number of trials are included. The largest x_i in the particular sequence of pseudorandom numbers used to generate 4×10^6 measurements for the $N = 40$ test is 9.03×10^{12}. Because of the symmetry of the likelihood (1) for one and two measurements, all the estimators are identical for $N = 1$ and 2. The posterior mean and maximum perform much better than the sample average for three or more measurements.

The estimators based on the sample median and the maximum of the posterior probability density function perform only slightly worse than the one based on the posterior mean. Just as they demonstrate the weakness of the sample mean estimator, these

results underscore the value of the sample median as a simple estimator that is robust against outliers. The table indicates that $rms(\hat{x}_{0(\text{samp med})}) > rms(\hat{x}_{0(\text{post max})}) > rms(\hat{x}_{0(\text{post mean})})$

The Fisher-Cramér-Rao lower bound on the rms error is seen to be a valid lower bound for the estimators summarized in the table, which only begin to approach the lower bound for $N \geq 20$.

It is natural to ask whether the posterior is predictive of the uncertainty in an estimator. The rms width of the posterior for our Cauchy problem $rms\{p(x_0|\mathbf{x})\}$ may be calculated by taking the square root of the variance in x_0, calculated using Eq. (12). The results for the simulated experiments, shown in the last column of the table, indicate that this calculation does predict the rms error in the $\hat{x}_{0(\text{post mean})}$ estimator.

We note that the shape of the posterior depends on the measured data and hence is different for each experiment. Furthermore, the shape of the likelihood depends on the relative positions of the x_i, as inferred from Fig. 1. This behavior is different for a Gaussian likelihood with a uniform prior, for which the width and shape of the posterior for a fixed number of data samples does not depend on the actual data values. We find in 10^4 trials for $N = 5$ that when the trials are selected on basis of $rms\{p(x_0|\mathbf{x})\}$, the rms error in the estimator $\hat{x}_{0(\text{post max})}$ for those trials reproduces the chosen $rms\{p(x_0|\mathbf{x})\}$. This result indicates that the posterior prabability density function derived for each experiment provides information about the certainy in inferences made on the basis of that experiment. It is clear that such information can be used to make decisions about whether more data should be taken to achieve a desired accuracy of interpretation.

5. Discussion

The prior used in the Bayesian analysis is the uniform prior. Because the uniform prior on the real number line is not normalizable, the analysis must be viewed as a limit over normalized priors [5]. In practice, the prior should reflect the state of prior knowledge.

Our analysis assumes that the measurement interval is the complete x axis. When the measurement interval is finite, and assuming that a fixed number of measurements are made, the posterior has asymptotically nonzero constant tails. The reason for this is that the probability of the measurements is then simply the product of the probabilities $p(x_i|x_0, y_0)$ of Eq. (1), with each normalized to unity over the finite measurement interval. For large x_0, the normalization constant is effectively the width of the measurement interval times the value of the Cauchy distribution tail in that interval (Eq. (2) may be used to establish the precise relationship). Thus, the normalization constant has the same large x_0 behavior as the Cauchy distribution that it normalizes, and this gives rise to the asymptotically nonzero tail.

For a source of fixed intensity, the assumption of fixing the number of measurements corresponds to varying the measurement interval until the specified number of photons is gathered. A more reasonable assumption might be to consider a fixed time interval. In this case, the number of measurements follows a Poisson distribution, so that the likelihood for N measurements discussed the last paragraph is modified by the factor $P(N|x_0, y_0) = e^{-\lambda}\lambda^N/N!$, where $\lambda = \lambda(x_0, y_0)$ is the source intensity times the total probability that a measurement occurs in the measurement interval.

Acknowledgments

We acknowledge many helpful discussions with Gregory S. Cunningham who suggested using the sample-median estimator. This work was supported by the United States Department of Energy under contract number W-7405-ENG-36.

References

1. K. M. Hanson. Introduction to Bayesian image analysis. *Proc. SPIE*, vol. 1898:716–731, 1993.
2. S. F. Gull. Bayesian inductive inference and maximum entropy. In G. J. Erickson and C. R. Smith, editors, *Maximum Entropy and Bayesian Methods in Science and Engineering (Vol. 1)*, pages 53–74. Kluwer Academic, 1989.
3. J. L. Devore. *Probability and Statistics for Engineering and the Sciences*. Brooks/Cole, Monterey, 1987.
4. H. L. Van Trees. *Detection, Estimation, and Modulation Theory - Part I*. John Wiley and Sons, New York, 1968.
5. G. E. P. Box and G. C. Tiao. *Inference in Statistical Analysis*. John Wiley and Sons, New York, 1973 (reprinted 1992).