The Newton-Raphson algorithm: Computing the MLE of the Cauchy distribution

The Newton-Raphson algorithm

The Newton-Raphson algorithm is a general purpose method for solving equations of the form $g(x^*) = 0$ where $g(x)$ is a (non-linear) differentiable function with derivative $g'(x)$. This algorithm solves for x^* by computing successive approximations $x^{(1)}, x^{(2)}, \cdots$ via the formula

$$x^{(k)} = x^{(k-1)} + \frac{g(x^{(k-1)})}{g'(x^{(k-1)})}.$$

The success of the Newton-Raphson algorithm depends on the choice of an initial estimate $x^{(0)}$.

A natural application of the Newton-Raphson algorithm is the computation of maximum likelihood estimates (MLEs). Suppose that θ is a real-valued parameter lying in a open parameter space Θ; if $\ln \mathcal{L}(\theta)$ is the log-likelihood function and $S(\theta)$ is its derivative with respect to θ then the MLE $\widehat{\theta}$ satisfies the likelihood equation

$$S(\widehat{\theta}) = 0.$$

We can compute the MLE iteratively by

$$\widehat{\theta}^{(k)} = \widehat{\theta}^{(k-1)} + \frac{S(\widehat{\theta}^{(k-1)})}{I(\widehat{\theta}^{(k-1)})}$$

where $I(\theta) = -S'(\theta)$; note that $I(\widehat{\theta})$ is simply the observed Fisher information.

The choice of initial estimate $\widehat{\theta}^{(0)}$ can be very important in the success of the Newton-Raphson algorithm and if this estimate is chosen appropriately, the one-step estimator

$$\widehat{\theta}^{(1)} = \widehat{\theta}^{(0)} + \frac{S(\widehat{\theta}^{(0)})}{I(\widehat{\theta}^{(0)})}$$

can be as good as the MLE $\widehat{\theta}$ itself. In the next section, we will try to illustrate this using a simple Cauchy model with an unknown location parameter.

MLE for the Cauchy location model

Suppose that $X_1, \cdots, X_n$ are independent Cauchy random variables with common density

$$f(x; \theta) = \frac{1}{\pi \{1 + (x - \theta)^2\}}$$

where θ is an unknown location parameter describing the centre of the distribution. The likelihood equation for MLE is given by

$$S(\widehat{\theta}) = 2 \sum_{i=1}^{n} \frac{x_i - \widehat{\theta}}{1 + (x_i - \widehat{\theta})^2} = 0$$

and the Newton-Raphson iteration for computing $\widehat{\theta}$ is given by

$$\widehat{\theta}^{(k)} = \widehat{\theta}^{(k-1)} + \frac{S(\widehat{\theta}^{(k-1)})}{I(\widehat{\theta}^{(k-1)})}$$

where

$$I(\theta) = -S'(\theta) = 2\sum_{i=1}^{n} \frac{1-(x_i-\theta)^2}{\{1+(x_i-\theta)^2\}^2}.$$

A little bit of algebra reveals that the solution of the likelihood equation is equivalent to the solution of a polynomial of degree $2n-1$, which suggests that the likelihood equation may have multiple solutions — as many as $2n-1$. Indeed, the likelihood equation $S(\widehat{\theta}) = 0$ has an odd number of solutions, specifically $2\ell+1$ solutions for some $\ell \geq 0$ (where ℓ depends on the data), $\ell+1$ of which are local maxima of the log-likelihood function. In fact, Reeds (1985)[1] shows that when n is large then ℓ has approximately a Poisson distribution with mean $1/\pi$; therefore, when n is large, the probability of a single local (and hence global) maximum is approximately $\exp(-1/\pi) \approx 0.727$ and so the probability of multiple local maxima is therefore approximately 0.273. Therefore, it is important to choose a good initial estimate for the algorithm in order to avoid converging to a solution of the likelihood equation that does not maximize the likelihood function.

Generally speaking, initial estimates $\widehat{\theta}^{(0)}$ should satisfy two conditions:

(a) They should be easy to compute, and
(b) they should be themselves good estimates of the parameter[2].

In the case of the Cauchy distribution, the parameter θ represents the centre of a symmetric distribution and so we can use, for example, the sample median or any trimmed mean computed by trimming a significant number of the extreme observations; for example, if $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ are the ordered data (the order statistics) then we can define the trimmed mean

$$\widetilde{\theta} = \frac{1}{n-2g}\sum_{i=g+1}^{n-g} x_{(i)}$$

where $g \geq 1$ represents the number of extreme observations trimmed from each end of the data. The sample mean is not a good initial estimate due to the fact that the mean (expected value) of the Cauchy distribution does not exist, which means that the sample mean will not converge to θ as n increases. We will examine later what happens to the Newton-Raphson algorithm when the sample mean is used as an initial estimate.

[1]Reeds, J. (1985) Asymptotic number of roots of Cauchy location likelihood equations. *Annals of Statistics.* **13**, 775–784.

[2]In practice, this condition may be somewhat complicated to satisfy — statistical models are typically only crude approximations to reality and so some care must be taken to insure that an initial estimate is, in fact, estimating roughly the same characteristic of the model as represented by the parameter.

The function below computes the MLE using the Newton-Raphson algorithm. The option `start` allows the user to specify an initial estimate of θ; if this is missing then the initial estimate is defined to be the sample median.

```
cauchy.mle <-  function(x,start,eps=1.e-8,max.iter=50){
    if (missing(start)) start <- median(x)
      theta <- start
      n <- length(x)
      score <- sum(2*(x-theta)/(1+(x-theta)^2))
      iter <- 1
      conv <- T
      while (abs(score)>eps && iter<=max.iter){
         info <- sum((2-2*(x-theta)^2)/(1+(x-theta)^2)^2)
         theta <- theta + score/info
         iter <- iter + 1
         score <- sum(2*(x-theta)/(1+(x-theta)^2))
         }
 if (abs(score)>eps) {
            print("No Convergence")
            conv <- F
            }
         loglik <- -sum(log(1+(x-theta)^2))
         info <- sum((2-2*(x-theta)^2)/(1+(x-theta)^2)^2)
         r <- list(theta=theta,loglik=loglik,info=info,convergence=conv)
         r
 }
```

The function can now be used as follows:

```
> x <- rcauchy(100) + 5 # 100 observations with theta = 5
> r <- cauchy.mle(x,start=median(x))
>  r
$theta
[1] 4.983086
$loglik
[1] -163.3458
$info
[1] 35.52706
$convergence
[1] TRUE
```

The component `$convergence` denotes whether the algorithm has converged (`TRUE`) or not (`FALSE`). Using the observed Fisher information contained in `r$info`, we can obtain an approximate 95% confidence interval for θ by

$$\texttt{r\$theta} \pm 1.96 \times \left\{\frac{1}{\texttt{r\$info}}\right\}^{1/2} = 4.983086 \pm 1.96 \times \left\{\frac{1}{35.52706}\right\}^{1/2} = 4.983086 \pm 0.3288338.$$

One-step estimates

In the previous section, we noted that the convergence of the Newton-Raphson algorithm to the MLE can be facilitated by an appropriate choice of the initial estimate. In fact, if $\widehat{\theta}^{(0)} = \widehat{\theta}_n^{(0)}$ satisfies certain properties then the one-step Newton-Raphson estimator

$$\widehat{\theta}_n^{(1)} = \widehat{\theta}_n^{(0)} + \frac{S(\widehat{\theta}_n^{(0)})}{I(\widehat{\theta}_n^{(0)})}$$

will have essentially the same statistical properties as the MLE itself.

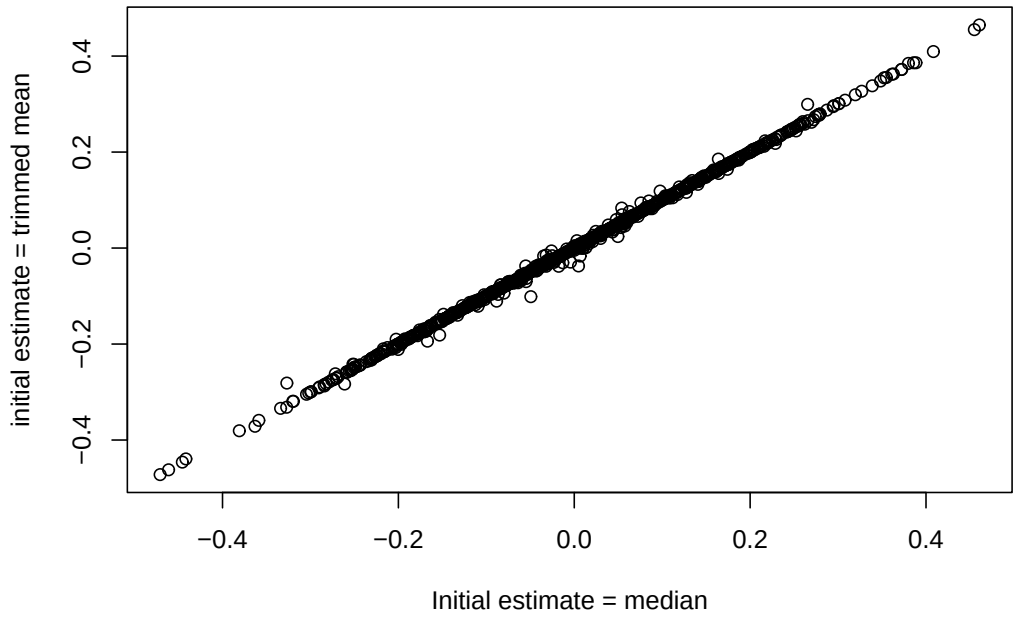

Figure 1: Scatterplot of one-step Newton-Raphson estimates for $n = 100$ using sample medians and 20% trimmed means as initial estimates.

Suppose that $\widehat{\theta} = \widehat{\theta}_n$ is the MLE of θ based on $X_1, \cdots, X_n$. Then under regularity conditions (which hold for the Cauchy distribution),

$$\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0, 1/\mathcal{I}(\theta));$$

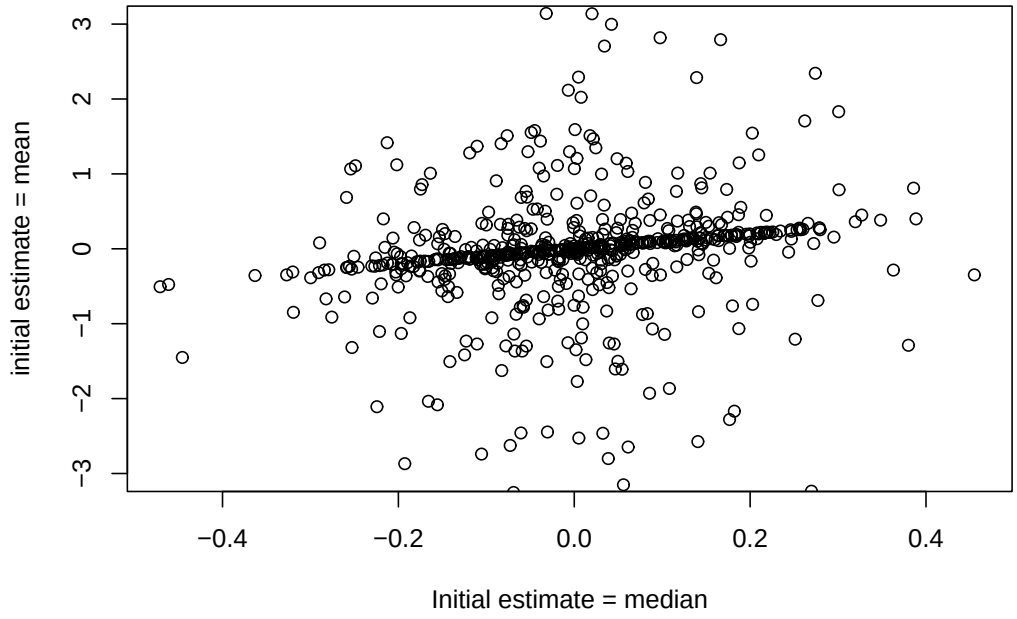

Figure 2: Scatterplot of one-step Newton-Raphson estimates for $n = 100$ using sample medians and sample means as initial estimates.

that is, the distribution of $\widehat{\theta}_n$ is approximately normal with mean θ (the true parameter value) and variance $\{n\mathcal{I}(\theta)\}^{-1}$ where

$$\mathcal{I}(\theta) = -E_\theta \left\{ \frac{\partial^2}{\partial \theta^2} \ln f(X;\theta) \right\} = \mathrm{Var}_\theta \left\{ \frac{\partial}{\partial \theta} \ln f(X;\theta) \right\}.$$

For the Cauchy distribution, $\mathcal{I}(\theta) = 1/2$ and so the variance of the MLE is approximatedly $2/n$. If $\sqrt{n}(\widehat{\theta}_n^{(0)} - \theta)$ converges in distribution (where the limiting distribution need not be normal) then the limiting distribution of $\sqrt{n}(\widehat{\theta}_n^{(1)} - \theta)$ is exactly the same as that of the MLE:

$$\sqrt{n}(\widehat{\theta}_n^{(1)} - \theta) \xrightarrow{d} \mathcal{N}(0, 1/\mathcal{I}(\theta))$$

The following R code looks at the distribution of the one-step estimators when the initial estimator is the sample median, 20% trimmed mean, and the sample mean.

```
> onestep.mean <- NULL
> onestep.median <- NULL
> onestep.trim <- NULL
> for (i in 1:1000) {
+   x <- rcauchy(100)
+   r1 <- cauchy.mle(x,start=mean(x),max.iter=1)
```

```
+   r2 <- cauchy.mle(x,start=median(x),max.iter=1)
+   r3 <- cauchy.mle(x,start=mean(x,trim=0.2),max.iter=1)
+   onestep.mean <- c(onestep.mean,r1$theta)
+   onestep.median <- c(onestep.median,r2$theta)
+   onestep.trim <- c(onestep.trim,r3$theta)
+  }
> var(onestep.median)
[1] 0.02070975
> var(onestep.trim)
[1] 0.02075797
> var(onestep.mean)
[1] 3707.28
> cor(cbind(onestep.median,onestep.trim,onestep.mean))
               onestep.median onestep.trim onestep.mean
onestep.median     1.00000000   0.99943565  -0.03235976
onestep.trim       0.99943565   1.00000000  -0.03345622
onestep.mean      -0.03235976  -0.03345622   1.00000000
```

Note that there is essentially no difference between the one-step estimators using the sample median and 20% trimmed mean as initial estimates — their variances are essentially the same (and approximately equal to 2/100) and their correlation is almost 1. A scatterplot of these two one-step estimates from the 1000 samples is given in Figure 1. On the other hand, the sample mean does not work well as an initial estimate in the one-step approach as evidenced by the very large variance from the 1000 samples. A scatterplot of the one-step estimates using the sample mean and sample median is shown in Figure 2.

More on the sample mean as the initial estimate

In the previous section, we saw that using the sample mean as the initial estimate for a one-step Newton-Raphson estimate is a very bad idea! However, it is possible that the Newton-Raphson algorithm will still converge to the MLE if the sample mean is used as the initial estimate.

To investigate this, we will draw 1000 samples with $n = 100$ from a Cauchy distribution (with $\theta = 0$) and see how often the Newton-Raphson algorithm converges when the sample mean is used as the initial estimate. Using the sample median as a "gold standard" initial estimate, we will say that the algorithm using the sample mean initial estimate converges if the two estimates are with 0.001 of each other.

```
> number <- 0 # this variable will record the number of convergences
> for (i in 1:1000) {
```

```
+    x <- rcauchy(100)
+    r <- cauchy.mle(x,start=mean(x))
+    r1 <- cauchy.mle(x,start=median(x))
+    if (abs(r$theta-r1$theta)<1.e-3) number <- number + 1
+    }
> number
[1] 524
```

For this simulation, we obtain convergence in 524 out of 1000 cases. Repeating the simulation for $n = 1000$, we obtained convergence in 520 out of 1000 cases.

It turns out to be straightforward to analyze the behaviour of the Newton-Raphson algorithm when the initial estimate is the sample mean. Assume for simplicity that the true value of θ is 0; the more general case follows similarly. In this case, the distribution of the sample mean $\bar{X}$ (for any value of n) is itself Cauchy with $\theta = 0$. Define T_n to be the one-step Newton-Raphson estimate using the sample mean:

$$T_n = \bar{X} + \frac{S(\bar{X})}{I(\bar{X})}.$$

When $\theta = 0$, we have (using the Weak Law of Large Numbers),

$$\begin{aligned}
\frac{1}{n}S(t) &\xrightarrow{p} 2\int_{-\infty}^{\infty}\left\{\frac{(x-t)}{1+(x-t)^2}\right\}\left\{\frac{1}{\pi(1+x^2)}\right\}dx = -2\frac{t}{t^2+4} \\
\frac{1}{n}I(t) &\xrightarrow{p} 2\int_{-\infty}^{\infty}\left\{\frac{1-(x-t)^2}{\{1+(x-t)^2\}^2}\right\}\left\{\frac{1}{\pi(1+x^2)}\right\}dx = -2\frac{(t^2-4)}{(t^2+4)^2},
\end{aligned}$$

which suggests that

$$T_n \xrightarrow{d} Z + \frac{Z(Z^2+4)}{Z^2-4} = g(Z)$$

where Z has a Cauchy distribution with $\theta = 0$. (In general, $T_n - \theta$ would converge in distribution to $g(Z)$.) Likewise, a two-step Newton-Raphson estimator (using the sample mean as the initial estimator) would converge in distribution to $g(g(Z)) = g \circ g(Z)$ and so on for multi-step Newton-Raphson estimators. In order to obtain convergence to 0 of the iterated function $g \circ \cdots \circ g(Z)$, we essentially need $|g(Z)| < |Z|$, that is,

$$\left|Z + \frac{Z(Z^2+4)}{Z^2-4}\right| = |Z|\left|1 + \frac{Z^2+4}{Z^2-4}\right| < |Z|,$$

which gives

$$\left|1 + \frac{Z^2+4}{Z^2-4}\right| < 1.$$

Solving for Z, we find that $|g(Z)| < |Z|$ if $|Z| < \sqrt{4/3} \approx 1.1547$. This suggests that the probability of convergence of the Newton-Raphson algorithm is

$$P(|Z| < \sqrt{4/3}) = \int_{-\sqrt{4/3}}^{\sqrt{4/3}} \frac{1}{\pi(1+x^2)}\,dx = 0.546,$$

which is close to the convergence proportion observed in the simulations.