Stat 8053 Lecture Notes

An Example that Shows that Statistical Obnoxiousness is not Related to Dimension

Charles J. Geyer

November 17, 2014

We are all familiar with the "usual" mathematics of maximum likelihood. Under the "usual regularity conditions" the MLE is consistent and asymptotically normal, and the inverse of the Fisher information matrix estimates the asymptotic variance of the MLE. The phrase "usual regularity conditions" sweeps a lot under the rug, since every theory book seems to have slightly different regularity conditions and each list of regularity conditions is too complicated and technical to provide any insight into what is going on.

But it gets worse. One can radically weaken the "usual" regularity conditions and get the same asymptotics for some estimators other than the MLE. The main source for this is Lucian Le Cam (Le Cam and Yang, 2000). His theory is notoriously hard to read and far beyond the scope of this course. The purpose of this note is just to present one concrete example.

The integral

$$\int_0^\infty e^{-x^\alpha} \, dx$$

is evaluated using the change of variable

$$y = x^\alpha \qquad x = y^{1/\alpha} \qquad dx = \frac{1}{\alpha} y^{1/\alpha - 1} \, dy$$

which gives

$$\int_0^\infty e^{-x^\alpha} \, dx = \frac{1}{\alpha} \int_0^\infty y^{1/\alpha - 1} e^{-y} \, dy = \frac{1}{\alpha} \Gamma\left(\frac{1}{\alpha}\right) = \Gamma\left(1 + \frac{1}{\alpha}\right)$$

Hence for any real θ, the function f_θ defined by

$$f_\theta(x) = \frac{e^{-|x - \theta|^\alpha}}{2\Gamma\left(1 + \frac{1}{\alpha}\right)}, \qquad -\infty < x < \infty \tag{1}$$

is a probability density and the family of densities $\{ \, f_\theta : \theta \in \mathbb{R} \, \}$ is a statistical model (a location family).

Having derived this density, we see that it can be derived from the gamma distribution by change of variable. Suppose X is a random variable with

density (1), then, defining $Y = |X - \theta|^\alpha$, we see that the distribution of Y does not depend on θ and has density

$$g(y) = \frac{e^{-y}}{\Gamma\left(1 + \frac{1}{\alpha}\right)} \cdot \frac{1}{\alpha} y^{1/\alpha - 1}$$

that is, $Y \sim \text{Gamma}(1/\alpha)$, and $X = \theta + SY^{1/\alpha}$, where S is a random sign (equal to -1 or $+1$ with equal probabilities).

```
>   set.seed(42)
>   n <- 1e2
>   alpha <- 3 / 4
>   theta <- 0
>   y <- rgamma(n, shape = 1 / alpha)
>   x <- theta + y^(1 / alpha) * sample(c(-1, 1), n, replace = TRUE)
```

If we treat α as a known constant, so θ is the only unknown parameter, the log likelihood is

$$l_n(\theta) = -\sum_{i=1}^{n} |x_i - \theta|^\alpha \tag{2}$$

If we choose $\alpha < 1$, then the log likelihood has a cusp at every x_i. And the "usual" notions of maximum likelihood go out the window. But Le Cam and Yang (2000) assert that for $1/2 < \alpha$ the log likelihood is asymptotically quadratic at a root n rate. The MLE is useless, but other likelihood-based estimators keep going.

Evaluate the log likelihood at a prespecified finite set of points ("prespecified" meaning specified before any data are observed). Fit a quadratic function to this log likelihood evaluated on a grid by any reasonable method, say robust regression, take the maximum of the quadratic approximation as the parameter estimate, and take the inverse of minus the second derivative of the quadratic approximation evaluated at the maximum, that is, we replace the actual log likelihood with a good quadratic approximation and then proceed to do likelihood inference as if the quadratic approximation were a log likelihood satisfying the "usual regularity conditions."

```
> xgrid <- pretty(x, n = 20)
> l <- function(theta) sum(- abs(x - theta)^alpha)
> vl <- Vectorize(l)
> lgrid <- vl(xgrid)
> # xgrid
> # lgrid
```

```
> library(MASS)
> rout <- rlm(lgrid ~ xgrid + I(xgrid^2))
> beta <- rout$coefficients
> theta.hat <- as.numeric(- beta[2] / (2 * beta[3]))
> theta.hat

[1] -0.2854965

> fish.hat <- as.numeric(- 2 * beta[3])
> theta.hat + c(- 1, 1) * qnorm(0.975) * sqrt(1 / fish.hat)

[1] -0.9902583  0.4192653
```

Now we try to picture what is going on. The picture (Figure 1) is given by the following code

```
> plot(xgrid, lgrid, ylab = "log likelihood", xlab = "theta")
> curve(predict(rout, newdata = data.frame(xgrid = x)),
+     add = TRUE, lty = "dashed")
> # and now for actual log likelihood
> foo <- par("usr")
> xgrid2 <- seq(foo[1], foo[2], length = 10001)
> xgrid2 <- sort(c(x, xgrid2))
> lgrid2 <- vl(xgrid2)
> lines(xgrid2, lgrid2)
```

Perhaps we should have fit the quadratic just to the middle part of the curve.

```
> xgrid <- theta.hat + c(- 1, 1) * qnorm(0.995) * sqrt(1 / fish.hat)
> lgrid <- vl(xgrid)
> xgrid <- pretty(xgrid2[min(lgrid) <= lgrid2], n = 20)
> lgrid <- vl(xgrid)
> rout <- rlm(lgrid ~ xgrid + I(xgrid^2))
> beta <- rout$coefficients
> theta.hat <- as.numeric(- beta[2] / (2 * beta[3]))
> theta.hat

[1] -0.01002563

> fish.hat <- as.numeric(- 2 * beta[3])
> theta.hat + c(- 1, 1) * qnorm(0.975) * sqrt(1 / fish.hat)
```

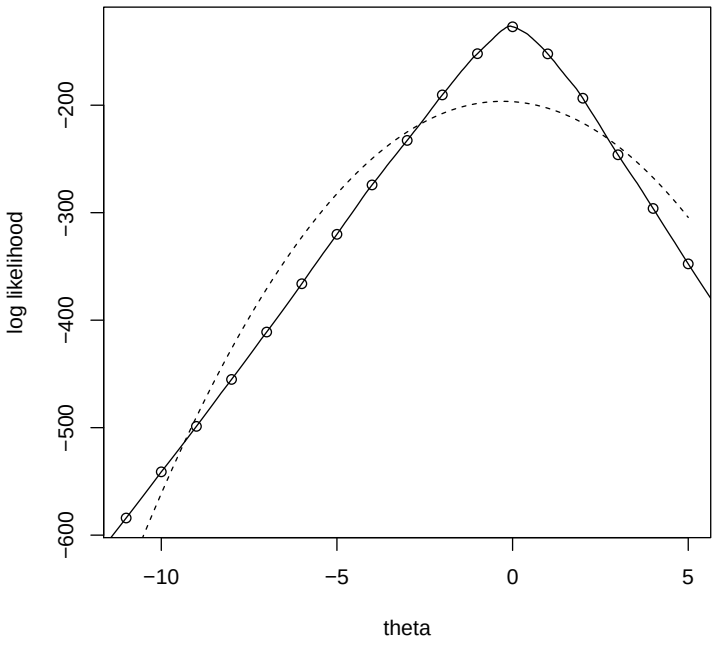

Figure 1: Log likelihood and quadratic approximation.

```
[1] -0.3076781  0.2876269
```

Of course, this is not what the theory says to do. We were suppose to choose the points where the log likelihood is evaluated before we saw the data, but this is impossible in practice. We would need more complicated theory to say that what we are doing here is o. k. Since this method is not used by applied people (only by theoreticians), there is no theory (AFAIK) that justifies what we have to do to get a practical method.

Now the picture (Figure 2) is given by the following code

```
> plot(xgrid, lgrid, ylab = "log likelihood", xlab = "theta",
+     xlim = range(xgrid))
> curve(predict(rout, newdata = data.frame(xgrid = x)),
+     add = TRUE, lty = "dashed")
> foo <- par("usr")
> xgrid2 <- seq(foo[1], foo[2], length = 10001)
> xgrid2 <- sort(c(x, xgrid2))
> xgrid2 <- xgrid2[foo[1] < xgrid2 & xgrid2 < foo[2]]
> lgrid2 <- vl(xgrid2)
> lines(xgrid2, lgrid2)
```

References

Le Cam, L. and Yang, G. L. (2000). *Asymptotics in Statistics: Some Basic Concepts*, 2nd ed. Springer-Verlag, New York.

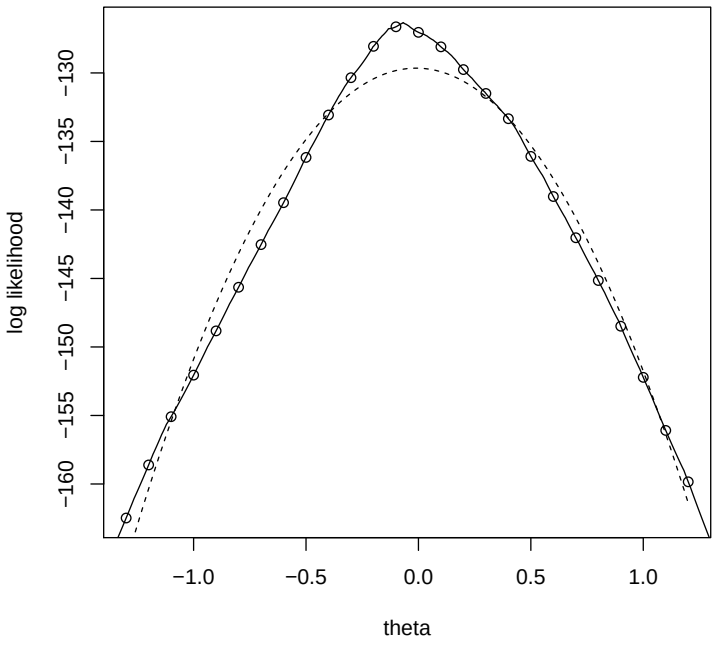

Figure 2: Log likelihood and quadratic approximation.