How much calculus is necessary to understand maximum likelihood estimation?

Asked 10 years, 10 months ago Modified 4 years, 9 months ago Viewed 2k times

11

I am trying to plan out a study plan for learning MLE. In order to do this I am trying to figure out what is the minimum level of calculus that is necessary to understand MLE.

Is it sufficient to understand the basics of calculus (i.e. finding the minimum and maximum of functions) in order to understand MLE?

estimation mathematical-statistics maximum-likelihood

Share Cite Edit Follow Flag

edited Jul 4, 2014 at 9:22
Glen_b
279k 37 615 1024

asked Feb 7, 2013 at 21:32
histelheim
2,983 6 25 46

2 As always, *it depends*. If you're only trying to comprehend the basics, being able to find extrema of functions gets you a fair way (though in many practical cases of MLE, the L is M'd numerically, in which case you need some other skills as well as some basic calculus). – Glen_b Feb 7, 2013 at 21:54

It depends ... conceptually calculus is not needed stats.stackexchange.com/questions/112451/… – kjetil b halvorsen ♦ Dec 9 at 18:45

2 Answers

Sorted by: Highest score (default)

21

To expand on my comment - it depends. If you're only trying to comprehend the basics, being able to find extrema of functions gets you a fair way (though in many practical cases of MLE, the likelihood is maximized numerically, in which case you need some other skills as well as some basic calculus).

I'll leave aside the nice simple cases where you get explicit algebraic solutions. Even so, calculus is often very useful.

I'll assume independence throughout. Let's take the simplest possible case of 1-parameter optimization. First we'll look at a case where we can take derivatives and separate out a function of the parameter and a statistic.

Consider the $\text{Gamma}(\alpha, 1)$ density

$$f_X(x;\alpha) = \frac{1}{\Gamma(\alpha)} x^{\alpha-1}\exp(-x);\;\;\; x>0;\;\; \alpha>0$$

Then for a sample of size n, the likelihood is

$$\mathcal{L}(\alpha;\mathbf{x}) = \prod_{i=1}^n f_X(x_i;\alpha)$$

and so the log-likelihood is

$$l(\alpha;\mathbf{x}) = \sum_{i=1}^n \ln f_X(x_i;\alpha)$$

$$= \sum_{i=1}^n \ln\left(\frac{1}{\Gamma(\alpha)} x_i^{\alpha-1}\exp(-x_i)\right)$$

$$= \sum_{i=1}^n -\ln\Gamma(\alpha) + (\alpha-1)\ln x_i - x_i$$

$$= -n\ln\Gamma(\alpha) + (\alpha-1)S_x - n\bar{x}$$

where $S_x = \sum_{i=1}^n \ln x_i$. Taking derivatives,

$$\frac{d}{d\alpha} l(\alpha;\mathbf{x}) = \frac{d}{d\alpha}(-n\ln\Gamma(\alpha) + (\alpha-1)S_x - n\bar{x})$$

$$= -n\frac{\Gamma'(\alpha)}{\Gamma(\alpha)} + S_x$$

$$= -n\psi(\alpha) + S_x$$

So if we set that to zero and try to solve for $\hat{\alpha}$, we can get this:

$$\psi(\hat{\alpha}) = \ln G(\mathbf{x})$$

where $\psi(\cdot)$ is the digamma function and $G(\cdot)$ is the geometric mean. We must not forget that in general you can't just set the derivative to zero and be confident you will locate the argmax; you still have to show in some way that the solution is a maximum (in this case it is). More generally, you may get minima, or horizontal points of inflexion, and even if you have a local maximum, you may not have a global maximum (which I touch on near the end).

So our task is now to find the value of $\hat{\alpha}$ for which

$$\psi(\hat{\alpha}) = g$$

where $g = \ln G(\mathbf{x})$.

This doesn't have a solution in terms of elementary functions, it must be calculated numerically; at least we were able to get a function of the parameter on one side and a

function of the data on the other. There are various zero-finding algorithms that might be used if you don't have an explicit way of solving the equation (even if you are without derivatives, there's binary section, for example).
Often, it's not so nice as that. Consider the logistic density with unit scale:

$$f(x;\mu) = \frac{1}{4}\operatorname{sech}^2\left(\frac{x-\mu}{2}\right).$$

Neither the argmax of the likelihood nor of the log-likelihood function can be readily obtained algebraically - you have to use numerical optimization methods. In this case, the function is fairly well behaved and the Newton-Raphson method should usually suffice to locate the ML estimate of μ. If the derivative was unavailable or if Newton-Raphson doesn't converge, other numerical optimization methods may be needed, such as golden-section (this is not intended to be an overview of the best available methods, just mentioning some methods you are more likely to encounter at a basic level).

More generally, you may not even be able to do that much. Consider a Cauchy with median θ and unit scale:

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

In general the likelihood here doesn't have a unique local maximum, but several local maxima. If you find *a* local maximum, there may be another, bigger one elsewhere. (Sometimes people focus on identifying the local maximum closest to the median, or some-such.)

It is easy for beginners to assume that if they find a concave turning point that they have the argmax of the function, but besides multiple modes (already discussed), there may be maxima that are not associated with turning points at all. Taking derivatives and setting them to zero is not sufficient; consider estimating the parameter for a uniform on $(0,\theta)$ for example.

In other cases, the parameter space may be discrete.

Sometimes finding the maximum may be quite involved.

And that's just a sampling of the issues with a single parameter. When you have multiple parameters, things get more involved again.

edited Mar 9, 2019 at 22:57

answered Feb 7, 2013 at 23:09

Glen_b
279k 37 615 1024

4

Yes. Of course, we are not talking about one-dimensional functions, but functions $\mathbb{R}^p \to \mathbb{R}$ to be maximized (viz., the likelihood), so this is slightly more advanced than the one-dimensional case.

Some facility with logarithms will definitely be helpful, since maximizing the logarithm of the

likelihood is usually much easier than maximizing the likelihood itself.

Quite a lot more than simple MLE can be understood (information matrices etc.) if you can deal with second derivatives of $\mathbb{R}^p \to \mathbb{R}$ functions, i.e., the Hessian matrix.

Share Cite Edit Follow Flag

answered Feb 7, 2013 at 21:36

Stephan Kolassa
119k 18 234 453