Cases where estimation behaves strangely or badly

In optimization, we're trying to maximize or minimize some function. For example, we may be looking for the maximum likelihood estimate (MLE) of one or more parameters. But is this always a good thing to do?

1. Infinity of the likelihood

Suppose for some likelihood function $f(\theta, X)$, where $X = (X_1, ..., X_n)$ is a vector of observations, the supremum of f is $+\infty$ at or approaching some value(s) of the parameter θ. Can one still define an MLE?

1.1. **Unique point where the likelihood goes to $+\infty$; example.** Suppose the supremum of the likelihood is $+\infty$ and there a parameter point θ_0 such that for every $\delta > 0$, there is an $M < +\infty$ such that for every θ such that $f(\theta, X) > M$ we have $|\theta - \theta_0| < \delta$. Thus, the likelihood becomes arbitrarily large only in the neighborhood of θ_0 or as $\theta \to \theta_0$. In such a case, we can define θ_0 as an MLE of θ in a generalized sense.

Such cases do arise for reasonable likelihoods. For any probability density function f on the real line, we can form a so-called *location-scale* family, where for every real m and every $\sigma > 0$, $(1/\sigma)f((x-m)/\sigma)$ is also a probability density. We can extend the family to the boundary case $\sigma = 0$ by defining the distribution then to be the point mass δ_m at m even though this, of course, does not have a density.

Let $f(x) = C/(1+x^2)^2$ where C is the normalizing constant making f a density. (This is a t density with 3 degrees of freedom; the outer exponent is $(\nu+1)/2$ where ν is the degrees of freedom; C can be expressed in terms of gamma functions and is given in many statistics textbooks.) Suppose for the location-scale family based on this density we have n observations $X_1, ..., X_n$ giving a likelihood function

$$f((m,\sigma), X) = \left(\frac{C}{\sigma}\right)^n \prod_{j=1}^{n} \frac{\sigma^4}{(\sigma^2 + (X_j - m)^2)^2} = C^n \frac{\sigma^{3n}}{\prod_{j=1}^{n}(\sigma^2 + (X_j - m)^2)^2}.$$

Now, suppose X_j are actually sampled from a discrete distribution with $X_j = 0$ for k values of $j \le n$. (This could happen as t distributions are sometimes used in defining nonparametric estimators.) Then when $m = 0$, the likelihood function is asymptotic as $\sigma \downarrow 0$ to

$$C^n \sigma^{3n-4k} \prod_{X_j \neq 0} \frac{1}{X_j^4} \tag{1}$$

which will go to $+\infty$ if and only if $k > 3n/4$. Then, $m = \sigma = 0$ will be the unique parameter pair for which the likelihood becomes arbitrarily large in its neighborhood, as there can be only one m such that $3/4$ or more of the X_j equal m. Specific discrete distributions that can easily give such data would be the geometric distribution with $p > 3/4$ or a Poisson distribution with $e^{-\lambda} > 3/4$.

1.2. **Nonparametric location and scale estimators.** Given observations $X_1, ..., X_n$ real-valued, the usual nonparametric estimator of location is the sample median $m(X)$ and a usual nonparametric estimator of scale is the MAD (median absolute deviation), namely the sample median of $|X_j - m(X)|$ (sometimes multiplied by a constant to make it match the standard deviation in the normal case). If more than half of the X_j are equal to some m_0 then m_0 will be the sample median and the MAD will be 0.

The following is based on Dudley, Sidenko and Wang (2009, Theorem 12). Som parts were known earlier and mentioned in references given by Dudley et al. If we take the t density with $\nu > 1$ degrees of freedom (where ν need not be an integer) and its location-scale family, then for arbitrary $X = (X_1, ..., X_n)$, the maximum likelihood estimates of m and σ exist and are unique, where that of σ is in the extended sense if its MLE is 0. For $\nu = 1$ we get the Cauchy distribution, for which the MLEs are not unique, specifically in case n is even, half the X_j have one value and the other half have another. If a fraction $\nu/(\nu + 1)$ or more of the X_j have the same value m_0, then m_0 is the MLE of m and that of σ is 0. For an arbitrary probability distribution Q on the real line, t_ν location and scale functionals $m_\nu(Q)$ and $\sigma_\nu(Q)$ exist. If $X_1, ..., X_n$ are i.i.d. (Q), then the MLEs of m and σ will converge with probability 1 as $n \to \infty$ to $m_\nu(Q)$ and $\sigma_\nu(Q)$. On the class of Q having no atom of size $\nu/(\nu + 1)$ or larger, the functionals μ_ν and σ_ν are highly smooth (analytic) with respect to suitable norms. If there are such large atoms, however, the functionals are only continuous, not differentiable.

As ν decreases down toward 1, the cases where we estimate $(m, \sigma) = (m_0, 0)$ converge toward those for using the median and MAD. More generally, like the median and MAD, the t estimates of location and scale m and σ are not sensitive to outliers and can serve as nonparametric estimators. As ν becomes large, the t distribution approaches normal and the t estimators of location and scale approach the parametric estimators, the sample mean and standard deviation.

1.3. **Likelihood going to infinity, not only near one point.** Suppose the likelihood function comes from a mixture of two normal distributions, having a density

$$f(x) = \lambda\sigma^{-1}\phi((x - \mu)/\sigma) + (1 - \lambda)\tau^{-1}\phi((x - \nu)/\tau) \tag{2}$$

where ϕ is the standard normal density, μ and ν can be any real numbers, $\sigma > 0$, $\tau > 0$, and $0 < \lambda < 1$. Suppose we observe $X_1, ..., X_n$ all distinct, and we'd like to find maximum likelihood estimates of the five parameters. To avoid a symmetry giving equal values of the likelihood let's assume that $\lambda < 1/2$. Consider the likelihood function

$$f(\theta, X) = \prod_{j=1}^{n} f(X_j)$$

where $\theta = (\mu, \sigma, \nu, \tau, \lambda)$. Suppose we take μ equal to some X_j. Let ν and $\tau > 0$ have any fixed values. As $\sigma \downarrow 0$, in the jth factor, we will have

$$\phi((X_j - \mu)/\sigma) = \phi(0/\sigma) = \phi(0) = 1/\sqrt{2\pi}$$

while the term $(1 - \lambda)\tau^{-1}\phi(X_j - \nu)/\tau)$ has a constant, positive value and the factor σ^{-1} goes to $+\infty$, so the whole jth factor goes to $+\infty$. In the ith factor for $i \neq j$, $\exp(-(X_i - \mu)^2/\sigma^2)$ will go to 0 very fast as $\sigma \to 0$ because $X_i \neq X_j = \mu$, and divided by σ it will still go to 0, so the ith factor will converge to the term with coefficient

$1-\lambda$ which is strictly positive. Thus the product of all n factors will go to $+\infty$. Since this can occur for n values of μ and arbitrary values of $\lambda \in (0, 1/2)$, $\tau > 0$ and ν, the "maximum likelihood estimate" is highly non-unique in this case, and maximum likelihood estimation is not useful.

How, then, should parameters in the normal mixture model be estimated? One might consider Bayesian methods. The Wikipedia article on the EM (Expectation-maximization) algorithm gives an estimation procedure in which Bayesian expectation steps alternate with maximum likelihood steps. The example treats a mixture of two multivariate normal distributions.

In "Solari's example" regarding regression with errors in both variables (on which there is a handout), something similar occurs. There are two critical points of the likelihood function, but they are both saddle points. For infinitely many values of some of the parameters, when one scale parameter σ decreases to 0, the likelihood goes to $+\infty$, so again maximum likelihood estimation is not useful.

In that case, there are a lot of parameters. For n observations (X_i, Y_i), $i = 1, ..., n$, in the plane, the X_i are modeled as $X_i = a_i + \xi_i$ where a_i are unknown true values and ξ_i are i.i.d. $N(0, \sigma^2)$ errors, and $Y_i = ba_i + \eta_i$ where η_i are i.i.d. $N(0, \tau^2)$ and independent of the ξ_j. The regression line is assumed to pass through $(0, 0)$, so that the true intercept is 0 and is not estimated. There are $n+3$ parameters b, σ, τ, $a_1, ..., a_n$, of which b is the main parameter of interest.

It may be that trying to estimate all the a_i individually is a bad idea. It seems that regression in such generality is not well-advised. If $\sigma = 0$ or σ/τ is very small, one has basically observed design points $X_i = a_i$ and can do classical y-on-x regression. Similarly if τ/σ is very small one can interchange the variables, i.e. do x-on-y regression. If both X_i and Y_i are observed with error, and are measured in the same units, one can minimize the sum of squared distances from the points to the line (best fit by squared distance [bfsd] regression, as I teach in 18.443).

2. Identifiability

A simple form of unidentifiability is as follows. Suppose we have a family of probability distributions P_θ depending on a parameter θ. We can say that θ is unidentifiable if $P_\theta = P_\phi$ for some $\theta \neq \phi$. If we have even a large sample size n of many observations, and find that the value θ fits the data well in whatever sense, then ϕ will fit it equally well. For example, if we have in (2 some $\theta = (\lambda, \mu, \sigma, \nu, \tau)$ with $\mu \neq \nu$ or $\sigma \neq \tau$ or $\lambda \neq 1/2$, then for $\phi = (1-\lambda, \nu, \tau, \mu, \sigma) \neq \phi$ we get $P_\theta = P_\phi$.

Bickel and Doksum (2001, p. 6) define identifiability only in this sense, that $P_\theta \neq P_\phi$ for all $\theta \neq \phi$ in the parameter space. But there are broader definitions, which may depend on the sample size n. Suppose we say that a real parameter θ is identifiable for a given n if, given $X_1, ..., X_n$, we can give a two-sided $1-\alpha$ confidence interval for θ. We would like if possible that the width of this confidence interval should shrink to 0 as $n \to \infty$ if the data are really i.i.d. for a distribution P_θ in the given parametric family.

2.1. Non-unique choices of θ.

A parameter θ may be said to be unidentified or only partially identified if one wishes say to minimize a certain function $Q(\theta, P)$ with respect to θ where P is the distribution of observations X, but the set Θ_0 of minimizing θ in

the parameter space Θ contains more than one point. For example, for distributions P on the real line, let $Q(\theta, P) = \int_{-\infty}^{\infty} |x - \theta| - |x| dP(x)$. This is minimized at θ if and only if θ is a median of P, and for some P, whose distribution functions equal $1/2$ on a non-degenerate interval, the median is not unique.

Based on a vector X of observations, one wishes to define a subset $\widehat{\Theta}_0$ which is a confidence set in the sense that for some $\beta > 0$, for each $\theta \in \Theta_0$, $\theta \in \widehat{\Theta}_0$ with probability at least β. This is the situation studied in the paper by Romano and Shaikh (2008) which has already been rather much cited.

Earlier in the course we saw how to get a nonparametric confidence interval, say a 95% interval, for "the" median using order statistics, namely $[X_{(j)}, X_{(k)}]$ for the smallest k such that $E(k, n, 1/2) \leq 0.025$ and the largest j such that $B(j-1, n, 1/2)$. If there is a non-degenerate interval Θ_0 of medians, it will be included in $[X_{(j)}, X_{(k)}]$ with probability at least 0.95. Romano and Shaikh are treating multidimensional parameter spaces and say their methods are "computationally intensive yet feasible."

REFERENCES

Bickel, P. J., and Doksum, K. A. (2001). *Mathematical Statistics: Basic Ideas and Selected Topics*, Vol. I, 2d ed. Prentice Hall.

Dudley, R. M., Sidenko, S., and Wang, Z. (2009), "Differentiability of t-functionals of location and scatter," *Ann. Statistics* **37**, pp. 939-960.

Romano, J. P., and Shaikh, A. M. (2008), "Inference for identifiable parameters in partially identified econometric models," *J. Statistical Planning and Inference* **138**, pp. 2786-2807.