STAT 513/413: Lecture 21
What is it good for

(An overview of most common statistical optimization tasks, with special attention to maximum likelihood)

What is most often optimized in statistics?

For instance, regression problems - this is a very common application

Something like

$$\sum_i \rho(y_i - g_i(\vartheta)) = \sum_i \rho(y_i - g(x_i, \vartheta)) \looparrowright \min_{\vartheta}!$$

- where $\rho(u)$ can be something like $\rho(u) = u^2$ (the easiest one)

- and $g(x_i, \vartheta)$ something like $x_{i1}\vartheta_1 + \cdots + x_{i1}\vartheta_p$ (the easiest one too)

The form of ρ may come from the distribution of y_i

Regression problems can be an instance of the following methodology, based on looking for the parameters that maximize *likelihood*

Maximum likelihood

Suppose that Y_i are independent and have distribution $f_i(y_i, \vartheta)$

Then the joint density of $y_1, y_2, \ldots, y_n$ is $f_1(y_1, \vartheta) f_2(y_2, \vartheta) \ldots f_n(y_n, \vartheta)$

Maximum likelihood: a good estimate of ϑ is the one that maximizes *likelihood* - the joint density of the Y_i's, viewed for the *observed* y_i's as a function of ϑ

$$f_1(y_1, \vartheta) f_2(y_2, \vartheta) \ldots f_n(y_n, \vartheta) \looparrowright \max_{\vartheta}!$$

This optimization problem equivalent to

$$-\log f_1(y_1, \vartheta) - \log f_2(y_2, \vartheta) - \cdots - \log f_n(y_n, \vartheta) \looparrowright \min_{\vartheta}!$$

An example: location-scale model

The data, $y_1, y_2, \ldots, y_n$ come as realizations of $Y_1, Y_2, \ldots, Y_n$ which are independent and have all the same distribution with density

$$f(y, \vartheta) = f(y; \mu, \sigma) = \frac{1}{\sigma}\varphi\left(\frac{y-\mu}{\sigma}\right) \qquad \text{with } \mu \text{ arbitrary, but } \sigma > 0$$

That is, Y_i is related to some "standard" distribution

its density is $\quad \varphi(y) = f(y; 0, 1)$

To obtain Y_i, we take U_i from this "standard" distribution, multiply it by σ and then shift it by μ

Or, we can view it also in the opposite direction: when μ is subtracted from Y_i and then the result is divided by σ, we obtain a quantity with a "standard" distribution

Instance: normal model

A well-known instance of this model is when the “standard” distribution is $N(0,1)$, standard normal distribution

$$f(y;\mu,\vartheta) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\dfrac{(y-\mu)^2}{2\sigma^2}}$$

The negative loglikehood, to be minimized, is then

$$\sum_{i=1}^{n} -\log f(y_i;\mu,\sigma) = n \log \sqrt{2\pi} + n \log \sigma + \sum_{i=1}^{n} \frac{(y_i-\mu)^2}{2\sigma^2}$$

and we obviously care only about the last two terms, as the first one is irrelevant regarding the parameters, it is constant with respect to to them

Differentiation yields here simple equations that typically be solved in closed form

Instance: Cauchy model

Another instance takes the Cauchy distribution for the "standard" one:

$$f(y;\mu,\vartheta) = \frac{1}{\pi\sigma\left(1+\left(\dfrac{y-\mu}{\sigma}\right)^2\right)}$$

The relevant negative loglikelihood to be minimized is

$$\sum_{i=1}^{n} -\log f(y_i;\mu,\sigma) = n\log\sigma + \log\left(1+\left(\frac{y-\mu}{\sigma}\right)^2\right)$$

Taking derivatives in μ and σ, and setting them to zero we obtain

Likelihood equations

$$\sum_{i=1}^{n} \frac{2\left(\frac{y_i-\mu}{\sigma}\right)}{1+\left(\frac{y_i-\mu}{\sigma}\right)^2}\left(-\frac{1}{\sigma}\right)=0 \quad \text{that is} \quad \sum_{i=1}^{n} \frac{y_i-\mu}{\sigma^2+(y_i-\mu)^2}=0$$

and

$$\sum_{i=1}^{n} \frac{1}{\sigma}+\frac{2\left(\frac{y_i-\mu}{\sigma}\right)\left(-\frac{y_i-\mu}{\sigma^2}\right)}{1+\left(\frac{y_i-\mu}{\sigma}\right)^2}=0 \quad \text{that is} \quad \sum_{i=1}^{n} \frac{(y_i-\mu)^2}{\sigma^2+(y_i-\mu)^2}=\frac{n}{2}$$

We have to solve this system somehow: for instance, by the Netwon method (but then we have to calculate three more derivatives)

Maximum likelihood in a very simplified setting

Just assume that Y has distribution $f = f(y, \vartheta)$

If we are taking derivative, it is in ϑ:

$$\dot{f} = \dot{f}(y, \vartheta) = \frac{\partial f(y, \vartheta)}{\partial \vartheta} \qquad \ddot{f} = \ddot{f}(y, \vartheta) = \frac{\partial^2 f(y, \vartheta)}{\partial \vartheta^2}$$

If we are taking expected value, it is in y: for instance,

$$\begin{aligned} E(-\log f) &= E(-\log f(Y, \vartheta)) \\ &= \int (-\log f(y, \vartheta)) f(y, \vartheta)\, dy = -\int f(y, \vartheta) \log f(y, \vartheta))\, dy \end{aligned}$$

Another example: $$E(Y) = \int y f(y, \vartheta)\, dy$$

The crucial tricks

We need some mathematical regularity conditions - but let us suppose they are all in place... Now some tricks: we start with

$$\int f(y, \vartheta)\, dy = 1 \qquad \text{($f(y,\vartheta)$ is a probability density, right?)}$$

and then we take a derivative, in ϑ, of the both sides

$$\int \dot{f}(y, \vartheta)\, dy = \dot{1} = \frac{\partial 1}{\partial \vartheta} = 0$$

and then we can do it again

$$\int \ddot{f}(y, \vartheta)\, dy = \dot{0} = \frac{\partial 0}{\partial \vartheta} = 0$$

Log-likelihood

The important quantity, as we have seen above, is also

$$\ell(y,\vartheta) = \ell = -\log f = -\log f(y,\vartheta)$$

Again, the derivatives are in ϑ

$$\dot{\ell} = (-\log f)^{\cdot} = -\frac{\dot{f}}{f}$$

And expected values in y (or Y, as needed)

$$E(\dot{\ell}) = E\left(-\frac{\dot{f}}{f}\right) = -\int \frac{\dot{f}}{f} f\,dy = \int \dot{f}(y,\vartheta)\,dy = 0$$

so that then we have an important quantity (*Fisher information*)

$$I(\vartheta) = \operatorname{Var}\dot{\ell} = E(\dot{\ell}^2) - (E(\dot{\ell}))^2 = E(\dot{\ell}^2) = E\left(\left(-\frac{\dot{f}}{f}\right)^2\right) = E\left(\frac{\dot{f}^2}{f^2}\right)$$

A longer, but important calculation

$$E(\ddot{\ell}) = -E\left(\frac{\ddot{f}f - \dot{f}\dot{f}}{f^2}\right) = -E\left(\frac{\ddot{f}}{f} - \frac{\dot{f}^2}{f^2}\right) = E\left(\frac{\dot{f}^2}{f^2}\right) - E\left(\frac{\ddot{f}}{f}\right)$$

$$= I(\vartheta) - \int \frac{\ddot{f}}{f} f\,dy = I(\vartheta) - \int \ddot{f}\,dy = I(\vartheta)$$

And now the Newton method

We want to find $\quad \ell \looparrowright \min_{\vartheta}! \quad$ which amounts to solving $\quad \dot{\ell} = 0$

Newton method: $\quad \ddot{\ell}(\vartheta_{k+1} - \vartheta_k) = -\dot{\ell}$

We are fine with $\dot{\ell}$, but $\ddot{\ell}$ is way too many derivatives to calculate: we approximate $\ddot{\ell}$ by its expected value $\mathsf{E}(\ddot{\ell})$ - (Fisher) *scoring*

Yeah, but what is ϑ? Newton method, more precisely

$$\ddot{\ell}(y, \vartheta_k)(\vartheta_{k+1} - \vartheta_k) = -\dot{\ell}(y, \vartheta_k)$$

after the approximation

$$I(\vartheta_k)(\vartheta_{k+1} - \vartheta_k) = -\dot{\ell}(y, \vartheta_k)$$

OK, but instead of $\ddot{\ell}(y, \vartheta_k)$ we have to calculate $I(\vartheta_k)$ - is that any better?

Well, often it is: note that after all, $I(\vartheta_k) = \mathsf{E}(\dot{\ell}^2)$, so at least we do not have to evaluate the second derivative

Nonlinear least-squares in a simplified setting

Consider the problem $$\sum_i (y_i - g(x_i, \vartheta))^2 \looparrowright \min_\vartheta !$$

equivalent to $$\sum_i \frac{1}{2}(y_i - g(x_i, \vartheta))^2 \looparrowright \min_\vartheta !$$

(this 1/2 is added for purely aesthetic reasons, as will be seen later)

However, ϑ is not a vector $\vartheta = (\vartheta_1, \ldots, \vartheta_p)^\mathsf{T}$ now - we would have to use partial derivatives, gradients, Hessians and all that -

but ϑ is only a single number, so we can again write $g_i = g(x_i, \vartheta)$, with $\dot{g}_i$ and $\ddot{g}_i$ being the corresponding partial derivatives in ϑ

The problem to is solve is $$\ell = \sum_i \ell_i = \sum_i \frac{1}{2}(y_i - g_i)^2 \looparrowright \min_\vartheta !$$

that is, $$\sum_i \ell_i(y_i, \vartheta) = \sum_i \frac{1}{2}(y_i - g_i(y_i, \vartheta))^2 \looparrowright \min_\vartheta !$$

Well... are there any distributions so that the $\ell_i(y_i, \vartheta)$ above come as $-\log f_i(y_i, \vartheta)$, where $f_i(y_i, \vartheta)$ is a density of Y_i? Turns out that yes (...), but we have to worry about that later, if at all; so far, we have just to assume that $E(Y_i) = g_i$

Let us use the rules

We are to solve

$$\dot{\ell} = \sum_i \dot{\ell}_i = \sum_i (y_i - g_i)(-\dot{g}_i) = 0$$

We take $\ddot{\ell}$

$$\ddot{\ell} = \sum_i (-\dot{g}_i(-\dot{g}_i) - (y_i - g_i)(-\ddot{g}_i)) = \sum_i \left(\dot{g}_i{}^2 + (y_i - g_i)(\ddot{g}_i)\right)$$

and replace it by $\displaystyle\sum_i \dot{g}_i{}^2$

which is in fact - note: the expected value is still in y_i (or Y_i)

$$\mathsf{E}(\ddot{\ell}) = \sum_i \left(\mathsf{E}(\dot{g}_i{}^2) + \mathsf{E}((Y_i - g_i)\ddot{g})\right) = \sum_i \left(\dot{g}_i{}^2 + \mathsf{E}(Y_i - g_i)\ddot{g}\right) = \sum_i \dot{g}_i{}^2$$

The expected value is in y_i, and $g_i = g(x_i, \vartheta)$ does not contain any y_i - it is a constant with respect to y_i, and hence $\mathsf{E}(\text{const}) = \text{const}$

So the *Gauss-Newton* method amounts to iterating along the rule

$$\left(\sum_i \dot{g}_i^2\right)(\vartheta_{k+1} - \vartheta_k) = -\sum_i \dot{g}_i$$

General expression

The problem is

$$\sum_i (y_i - g(x_i, \vartheta))^2 =\!\!\hookrightarrow \min_\vartheta !$$

where $\vartheta = (\vartheta^1, \vartheta^2, \ldots, \vartheta^p)^\mathsf{T}$ (note: superscripts)

We calculate the sum of gradients (the sum of $p \times 1$ vectors)

$$b(\vartheta_k) = \sum_i \nabla g(x_i, \vartheta_k) \qquad \text{(note: subscripts)}$$

and the sum of $p \times p$ matrices (Fisher information *matrix*)

$$A(\vartheta_k) = \sum_i \big(\nabla g(x_i, \vartheta_k)\big)\big(\nabla g(x_i, \vartheta_k)\big)^\mathsf{T}$$

The Gauss-Newton iteration ϑ_{k+1} solves the system

$$A(\vartheta_k)(\vartheta_{k+1} - \vartheta_k) = -b(\vartheta_k)$$

(and also Fisher information matrix at the solution is then ready - which is a path to evaluate the variance of the solution...)