Lecture 2: Maximum Likelihood and Friends

Chris Conlon

February 6, 2021

NYU Stern

Computing Maximum Likelihood Estimators

Newton's Method for Root Finding

Consider the Taylor series for $f(x)$ approximated around $f(x_0)$:

$$f(x) \approx f(x_0) + f'(x_0) \cdot (x - x_0) + f''(x_0) \cdot (x - x_0)^2 + o_p(3)$$

Suppose we wanted to find a root of the equation where $f(x^*) = 0$ and solve for x:

$$0 = f(x_0) + f'(x_0) \cdot (x - x_0)$$
$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$$

This gives us an iterative scheme to find x^*:

1. Start with some x_k. Calculate $f(x_k), f'(x_k)$
2. Update using $x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}$
3. Stop when $|x_{k+1} - x_k| < \epsilon_{tol}$.

Newton-Raphson for Minimization

We can re-write optimization as root finding;

- We want to know $\hat{\theta} = \arg\max_{\theta} \ell(\theta)$.
- Construct the FOCs $\frac{\partial \ell}{\partial \theta} = 0 \rightarrow$ and find the zeros.
- How? using Newton's method! Set $f(\theta) = \frac{\partial \ell}{\partial \theta}$

$$\theta_{k+1} = \theta_k - \left[\frac{\partial^2 \ell}{\partial \theta^2}(\theta_k) \right]^{-1} \cdot \frac{\partial \ell}{\partial \theta}(\theta_k)$$

The SOC is that $\frac{\partial^2 \ell}{\partial \theta^2} > 0$. Ideally at all θ_k.
This is all for a single variable but the multivariate version is basically the same.

Newton's Method: Multivariate

Start with the objective $Q(\theta) = -\ell(\theta)$:

- Approximate $Q(\theta)$ around some initial guess θ_0 with a quadratic function
- Minimize the quadratic function (because that is easy) call that θ_1
- Update the approximation and repeat.

$$\theta_{k+1} = \theta_k - \left[\frac{\partial^2 Q}{\partial\theta\partial\theta'}\right]^{-1} \frac{\partial Q}{\partial\theta}(\theta_k)$$

- The equivalent SOC is that the Hessian Matrix is positive semi-definite (ideally at all θ).
- In that case the problem is globally convex and has a unique maximum that is easy to find.

Newton's Method

We can generalize to Quasi-Newton methods:

$$\theta_{k+1} = \theta_k - \lambda_k \underbrace{\left[\frac{\partial^2 Q}{\partial\theta\partial\theta'}\right]^{-1}}_{A_k} \frac{\partial Q}{\partial\theta}(\theta_k)$$

Two Choices:

- Step length λ_k
- Step direction $d_k = A_k \frac{\partial Q}{\partial\theta}(\theta_k)$
- Often rescale the direction to be unit length $\frac{d_k}{\|d_k\|}$.
- If we use A_k as the true Hessian and $\lambda_k = 1$ this is a full Newton step.

Newton's Method: Alternatives

Choices for A_k

- $A_k = I_k$ (Identity) is known as gradient descent or steepest descent
- BHHH. Specific to MLE. Exploits the Fisher Information.

$$A_k = \left[\frac{1}{N} \sum_{i=1}^{N} \frac{\partial \ln f}{\partial \theta}(\theta_k) \frac{\partial \ln f}{\partial \theta'}(\theta_k) \right]^{-1}$$
$$= -\mathbb{E}\left[\frac{\partial^2 \ln f}{\partial \theta \partial \theta'}(Z, \theta^*) \right] = \mathbb{E}\left[\frac{\partial \ln f}{\partial \theta}(Z, \theta^*) \frac{\partial \ln f}{\partial \theta'}(Z, \theta^*) \right]$$

- Alternatives SR1 and DFP rely on an initial estimate of the Hessian matrix and then approximate an update to A_k.
- Usually updating the Hessian is the costly step.
- Non invertible Hessians are bad news.

EM Algorithm and Mixtures

Estimating Finite Mixtures

- In practice estimating finite mixture models can be tricky.
- A simple example is the mixture of normals (incomplete data likelihood)

$$f(x_1,\ldots,x_n|\theta) = \prod_{i=1}^{N}\sum_{k=1}^{K}\pi_k f(x_i|\mu_k,\sigma_k)$$

- We need to find both mixture weights $\pi_k = Pr(z_k)$ and the components (μ_k,σ_k) the weights define a valid probabiltiy measure $\sum_k \pi_k = 1$.
- Easy problem is label switching. Usually it helps to order the components by say decreasing $\pi_1 > \pi_2 > \ldots$ or $\mu_1 > \mu_2 > \ldots$
- The real problem is that which component you belong to is unobserved. We can add an extra indicator variable $z_{ik} \in \{0,1\}$.
- We don't care about z_{ik} per-se so they are nuisance parameters.

Estimating Finite Mixtures

- We can write the complete data log-likelihood (as if we observed z_{ik}):

$$\ell(x_1, \ldots, x_n|\theta) = \sum_{i=1}^{N} \log\left(\sum_{k=1}^{K} I[z_i = k]\pi_k f(x_i, \mu_k, \sigma_k)\right)$$

- We can instead maximized the expected log-likelihood where we take the expectation $E_{z|\theta}$

$$\alpha_{ik}(\theta) = Pr(z_{ik} = 1|x_i, \theta) = \frac{f_k(x_i, z_k, \mu_k, \sigma_k)\pi_k}{\sum_{m=1}^{K} f_m(x_i, z_m, \mu_m, \sigma_m)\pi_m}$$

- Now we have a probability $\hat{\alpha}_{ik}$ that gives us the probability that i came from component k. We also compute $\hat{\pi}_k = \frac{1}{N}\sum_{i=1}^{N} \alpha_{ik}$

EM Algorithm

- Treat the $\hat{\alpha}_k(\theta^{(q)})$ as data and maximize to find μ_k, σ_k for each k

$$\hat{\theta}^{(q+1)} = \arg\max_{\theta} \sum_{i=1}^{N} \log \left(\sum_{k=1}^{K} \hat{\alpha}_k(\theta^{(q)}) f(x_i | z_{ik}, \theta) \right)$$

- We iterate between updating $\hat{\alpha}_k(\theta^{(q)})$ (E-step) and $\hat{\theta}^{(q+1)}$ (M-step)
- For the mixture of normals we can compute the M-step very easily:

$$\begin{aligned} \mu_k^{(q+1)} &= \frac{1}{N} \sum_{i=1}^{N} \hat{\alpha}_k(\theta^{(q)}) x_i \\ \sigma_k^{(q+1)} &= \frac{1}{N} \sum_{i=1}^{N} \hat{\alpha}_k(\theta^{(q)}) (x_i - \overline{x})^2 \end{aligned}$$

EM Algorithm

- EM algorithm has the advantage that it avoids complicated integrals in computing the expected log-likelihood over the missing data.
- For a large set of families it is proven to converge to the MLE
- That convergence is monotonic and linear. (Newton's method is quadratic)
- This means it can be slow, but sometimes $\nabla_\theta f(\cdot)$ is really complicated.

Thanks!