

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**.

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

- 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

- 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:

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

- 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!
