ML Procedure, Quadratic Approximation, Exponential Family Models

Surya Tokdar

ML calculations for normal model

- Consider the model $X_1, \cdots, X_n \overset{\text{IID}}{\sim} N(\mu, \sigma^2)$, $\mu \in (-\infty, \infty)$
- σ is fixed, not a parameter
- The log-likelihood function equals

$$\ell_x(\mu) = \text{const} - \frac{n(\bar{x} - \mu)^2}{2\sigma^2}, \quad \mu \in (-\infty, \infty)$$

- Because this is a quadratic function in μ, we have

$$\hat{\mu}_{\text{MLE}}(x) = \bar{x}$$
$$B_c(x) \overset{\text{def}}{=} \left\{\mu : \ell_x(\mu) \geq \ell_x(\hat{\mu}_{\text{MLE}}(x)) - \frac{c^2}{2}\right\}$$
$$= \left(\bar{x} - c\frac{\sigma}{\sqrt{n}}, \bar{x} + c\frac{\sigma}{\sqrt{n}}\right)$$

- Note: $B_c(x) = A_k(x)$ with $k = e^{-c^2/2}$.

Quadratic log-likelihood

- The quadratic form of the log-likelihood is a signature of normal "location" models

$$Y_i \overset{\text{IND}}{\sim} N(\beta^T z_i, \sigma^2)$$

with fixed σ
- Even when σ is an unknown parameter, the profile log-likelihood in β remains quadratic up to a monotone transform.

Normal model with unknown variance

- Now consider $X_1, \cdots, X_n \overset{\text{IID}}{\sim} N(\mu, \sigma^2)$
- Both $\mu \in (-\infty, \infty)$ and $\sigma^2 > 0$ are model parameters
- The profile log-likelihood in μ equals

$$\ell_x^*(\mu) = \text{const} - \tfrac{n}{2}\log\{\tfrac{n-1}{n}s_x^2 + (\bar{x} - \mu)^2\}$$

- Consequently,

$$\hat{\mu}_{\text{MLE}}(x) = \bar{x}$$
$$B_c(x) = \bar{x} \mp c_n\frac{s_x}{\sqrt{n}}$$

where $s_x^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2$ and $c_n = \sqrt{(n-1)(e^{c^2/n} - 1)}$

Non-normal models: MLE

- Models involving non-normal pdfs or pmfs usually have log-likelihood function that are fairly more complex
- Binomial model $X \sim Bin(n, p)$, $p \in [0, 1]$

$$\ell_x(p) = \text{const} + x\log p + (n - x)\log(1 - p)$$

- But it could still be possible to calculate MLE fairly easily

Normal equations

- Assume $\ell_x(\theta)$ is continuously differentiable and unimodal
- Then $\hat{\theta}_{\text{MLE}}(x)$ solves (uniquely) the first order condition

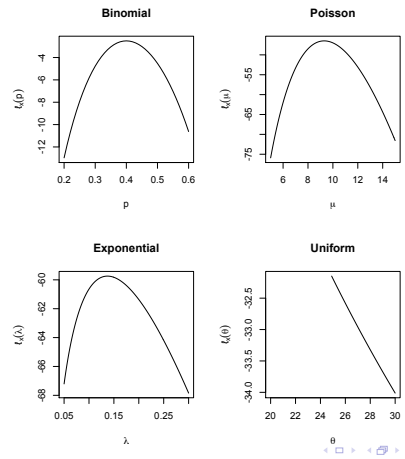
$$\dot{\ell}_x(\theta) = 0$$

where for a
  - scalar θ, $\dot{\ell}_x(\theta) = \frac{\partial}{\partial\theta}\ell_x(\theta)$
  - vector $\theta = (\theta_1, \cdots, \theta_d)^T$, $\dot{\ell}_x(\theta) = (\frac{\partial}{\partial\theta_1}\ell_x(\theta), \cdots, \frac{\partial}{\partial\theta_d}\ell_x(\theta))^T$.
- For θ with $\dim(\theta) = d$, this is really a system of d equations
- Important question: when do these assumptions hold?

ML intervals

- Even if we can compute the MLE, calculating ML intervals $B_c(x) = \{\theta : \ell_x(\theta) \geq \ell_x(\hat{\theta}_{\text{MLE}}(x)) - c^2/2\}$ can be tricky
- Remarkably, for many useful models, the log-likelihood is approximately quadratic near the MLE
- So we can get a closed form approximation to $B_c(x)$
- Important question: When does this happen?

Many but not all



When quadratic approximation holds

- If $\ell_x(\theta)$ is approximately quadratic around $\hat{\theta}_{\text{MLE}}(x)$ then we can use a two-term Taylor approximation:

$$\begin{aligned}\ell_x(\theta) \approx\ & \ell_x(\hat{\theta}_{\text{MLE}}(x)) \\ & + (\theta - \hat{\theta}_{\text{MLE}}(x))^T \dot{\ell}_x(\hat{\theta}_{\text{MLE}}(x)) \\ & + \frac{1}{2}(\theta - \hat{\theta}_{\text{MLE}}(x))^T \ddot{\ell}_x(\hat{\theta}_{\text{MLE}}(x))(\theta - \hat{\theta}_{\text{MLE}}(x))\end{aligned}$$

- Note: $\dot{\ell}_x(\hat{\theta}_{\text{MLE}}(x)) = 0$ and $I_x = -\ddot{\ell}_x(\hat{\theta}_{\text{MLE}}(x))$ positive definitie
- So: $\ell_x(\theta) \approx \ell_x(\hat{\theta}_{\text{MLE}}(x)) - \frac{1}{2}(\theta - \hat{\theta}_{\text{MLE}}(x))^T I_x(\theta - \hat{\theta}_{\text{MLE}}(x))$.

Approximate ML intervals

- $B_c(x) \approx \{\theta : (\theta - \hat{\theta}_{\text{MLE}}(x))^T I_x(\theta - \hat{\theta}_{\text{MLE}}(x)) \leq c\}$
- For scalar θ, $B_c(x) \approx \hat{\theta}_{\text{MLE}}(x) \mp c/\sqrt{I_x}$
- For vector θ the above set is an ellipsoid centered at $\hat{\theta}_{\text{MLE}}(x)$

Profile likelihood derivations

- Consider a vector valued θ
- Suppose we are interested in the scalar quantity $\eta = a^T\theta$ where a is a known vector
- The profile log-likelihood in η equals:

$$\ell_x^*(\eta) \approx \text{const} - \frac{(\eta - \hat{\eta}_{\text{MLE}}(x))^2}{a' I_x^{-1} a}$$

  where $\hat{\eta}_{\text{MLE}}(x) = a'\hat{\theta}_{\text{MLE}}(x)$
- So $B_c(x) \approx \hat{\eta}_{\text{MLE}}(x) \mp c\sqrt{a' I_x^{-1} a}$

When do assumptions hold?

- The three assumptions (log-likelihood is unimodal, continuously differentiable and approximately quadratic near the maxima) hold for many statistical models and verification can be done case by case
- A general class of models for which these hold is given by the exponential families

Exponential family of pdfs/pmfs

- A family of pdfs (or pmfs) $\{f(x|\theta) : \theta \in \Theta\}$ on S is called an exponential family if one can write

$$f(x|\theta) = h(x) \exp\left\{\eta(\theta)^T T(x) - B(\theta)\right\},$$

  for every $\theta \in \Theta$, $x \in S$, for some functions $h(x)$, $T(x)$, $\eta(\theta)$ and $B(\theta)$.
- Must have
  - $B(\theta) = \log \int_S h(x) e^{\eta(\theta)^T T(x)} dx$
  - $h(x) > 0 \implies f(x|\theta) > 0$ for all θ,
  - $h(x) = 0 \implies f(x|\theta) = 0$ for all θ

From EF pieces to a joint EF

- Suppose $X = (X_1, \cdots, X_n)$ and

$$X_i \overset{\text{IID}}{\sim} g(x_i|\theta) = h(x_i) e^{\eta(\theta)^T T(x_i) - B(\theta)}$$

- Then the pdfs for X are

$$f(x|\theta) = \left\{\prod_{i=1}^n h(x_i)\right\} e^{\eta(\theta)^T \sum_{i=1}^n T(x_i) - nB(\theta)}$$
$$= h_n(x) e^{\eta(\theta)^T T_n(x) - B_n(\theta)}$$

One more

- Suppose $X = (X_1, \cdots, X_n)$ and

$$X_i \overset{\text{IND}}{\sim} g(x_i|z_i, \theta) = h(x_i, z_i) e^{\eta(\theta)^T T(x_i, z_i) - B_{z_i}(\theta)}$$

- Then the pdfs for X are

$$f(x|\theta) = \left\{\prod_{i=1}^n h(x_i, z_i)\right\} e^{\eta(\theta)^T \sum_{i=1}^n T(x_i, z_i) - \sum_{i=1}^n B_{z_i}(\theta)}$$
$$= h_{n,z}(x) e^{\eta(\theta)^T T_{n,z}(x) - B_{n,z}(\theta)}$$

Assumptions hold...

- If $\{f(x|\theta) : \theta \in \Theta\}$ is an exponential family, i.e., $f(x|\theta) = h(x) e^{\eta(\theta)^T T(x) - B(\theta)}$ where $\eta(\theta)$ is a one-to-one function of θ with a continuous second derivative
- Only a sufficient condition (non-exponential families can exhibit same behaviors)