Large Sample ML-based Methods I

Feng-Chang Lin

Department of Biostatistics
University of North Carolina at Chapel Hill

flin@bios.unc.edu

(C&B §10)

Notations

- $X_1, \ldots, X_n$ be iid random variables from a family indexed by θ.
- Log-likelihood: $\ell(\theta) = \sum_{i=1}^n \ell_i(\theta|x_i)$, where
$$\ell_i(\theta|x_i) = \log f(x_i|\theta).$$
- Score function: $U(\theta) = \sum_{i=1}^n U_i(\theta|x_i)$, where
$$U_i(\theta|x_i) = (\partial/\partial\theta)\ell_i(\theta|x_i).$$
- Observed information: $J(\theta) = \sum_{i=1}^n J_i(\theta|x_i)$, where
$$J_i(\theta|x_i) = -(\partial^2/\partial\theta^2)\ell_i(\theta|x_i).$$

Notations (cont'd)

- Expected information: $I_n(\theta) = nI_1(\theta)$, where

$$I_1(\theta) = EJ_i(\theta|x_i) = E\{-(\partial^2/\partial\theta^2)\ell_i(\theta|x_i)\}.$$

- $\ell_1, \ldots, \ell_n$ are iid.
- $U_1, \ldots, U_n$ are iid mean 0 and variance $I_1(\theta)$.

$$\begin{aligned} E(U_i) &= E\left\{\frac{\partial}{\partial\theta}\ell_i(\theta|x_i)\right\} = E\left\{\frac{\partial}{\partial\theta}\log f(X_i|\theta)\right\} \\ &= E\left\{\frac{\frac{\partial}{\partial\theta}f(x_i|\theta)}{f(x_i|\theta)}\right\} = \int_{\mathcal{X}}\frac{\partial}{\partial\theta}f(x|\theta)dx = \frac{\partial}{\partial\theta}(1) = 0 \end{aligned}$$

Notations (cont'd)

- You may find the proof of $Var(U_i) = I_1(\theta)$ in Exercise 7.39 in C&B. Here are some outlines:

$$
\begin{aligned}
I_1(\theta) &= -E\left\{\frac{\partial^2}{\partial\theta^2}\log f(x_i|\theta)\right\} = -E\left[\frac{\partial}{\partial\theta}\left\{\frac{\partial}{\partial\theta}\log f(x_i|\theta)\right\}\right] \\
&= -E\left[\frac{\partial}{\partial\theta}\left\{\frac{\frac{\partial}{\partial\theta}f(x_i|\theta)}{f(x_i|\theta)}\right\}\right] = E\left\{\frac{\frac{\partial}{\partial\theta}f(x_i|\theta)}{f(x_i|\theta)}\right\}^2 \\
&= E\left\{\frac{\partial}{\partial\theta}\log f(x_i|\theta)\right\}^2 = Var(U_i)
\end{aligned}
$$

Notations (cont'd)

- $J_1, \ldots, J_n$ are iid mean $I_1(\theta)$.
- $I_1(\theta)$ is the expected (Fisher) information in one observation.
- We call $I_1(\theta)$ **information number**.
- $I_n(\theta) = nI_1(\theta)$ is the expected information in n observation.

Bernoulli Example

- Let $X_1, \ldots, X_n$ be iid Bernoulli(θ), $\theta \in (0, 1)$.
- The log-likelihood is $\ell(\theta) = \sum_{i=1}^n \ell_i(\theta|x_i)$, where

$$\ell_i(\theta|x_i) = x_i \log \frac{\theta}{1-\theta} + \log(1-\theta).$$

- The score function is $U(\theta) = \sum_{i=1}^n U_i(\theta|x_i)$, where

$$U_i(\theta|x_i) = \frac{x_i}{\theta(1-\theta)} - \frac{1}{1-\theta} = \frac{x_i - \theta}{\theta(1-\theta)}.$$

- The observed information is $J(\theta) = \sum_{i=1}^n J_i(\theta|x_i)$

$$J_i(\theta|x_i) = \frac{1}{\theta^2(1-\theta)^2}(x_i - 2x_i\theta + \theta^2).$$

Bernoulli Example (cont'd)

- The expected information is $nI_1(\theta)$, where

$$I_1(\theta) = EJ_i(\theta|x_i) = \frac{1}{\theta(1-\theta)}.$$

- Check: $E\{U_i(\theta|x_i)\} = 0$.
- Check: $Var\{U_i(\theta|x_i)\} = I_1(\theta)$.

Large Sample Properties of MLE

- When $\theta = \theta_0$ and $n \to \infty$,

$$\sqrt{n}\left\{\frac{1}{n}U(\theta_0) - 0\right\} = \frac{1}{\sqrt{n}}U(\theta_0) \to_d N\{0, I_1(\theta_0)\}.$$

- $n^{-1}J(\theta_0) \to_p I_1(\theta_0)$.
- Let $K(\theta_0) = \sum_{i=1}^n K_i(\theta_0|x_i)$, where $K_i(\theta|x_i) = (\partial^3/\partial\theta^3)\ell_i(\theta|x_i)$.
- $n^{-1}\sum_{i=1}^n K_i(\theta_0|x_i) \to_p E\{K_i(\theta_0)|x_i\}$.

Large Sample Properties of MLE (cont'd)

- Let $\hat{\theta}$ be MLE of θ based on n observations (also denoted by $\hat{\theta}$).
- **Theorem (Consistency)**:

$$\hat{\theta} \to_p \theta_0 \text{ as } n \to \infty.$$

- **Theorem (Asymptotic Normality)**:

$$\sqrt{n}(\hat{\theta} - \theta_0) \to_d N\{0, I_1(\theta_0)^{-1}\} \text{ as } n \to \infty.$$

- This implies: $\tau(\hat{\theta}) \to_p \tau(\theta_0)$, and

$$\sqrt{n}\left\{\tau(\hat{\theta}) - \tau(\theta_0)\right\} \to_d N\left[0, \frac{\{\tau'(\theta_0)\}^2}{I_1(\theta_0)}\right]$$

- What method did we use? It requires $\tau(\cdot)$ is a continuous function and $\tau'(\theta_0) \neq 0$.

Asymptotic Efficiency

- T_n is an "asymptotically efficient" estimator of $\tau(\theta)$ if

$$\sqrt{n}\{T_n - \tau(\theta)\} \to_d N(0, v(\theta)),$$

  and

$$v(\theta) = \frac{\{\tau'(\theta_0)\}^2}{I_1(\theta_0)}.$$

- That means, asymptotic variance = CRLB
- MLE $\tau(\hat{\theta})$ is asymptotically efficient.

Asymptotic Relative Efficiency

- Definitions: If

$$\sqrt{n}(T_{1n} - \theta) \rightarrow_d N(0, \sigma_1^2), \text{ and}$$
$$\sqrt{n}(T_{2n} - \theta) \rightarrow_d N(0, \sigma_2^2), \text{ as } n \rightarrow \infty.$$

- The asymptotic relative efficiency of T_{1n} with respect to T_{2n} is

$$\text{ARE}(T_{1n}, T_{2n}) = \frac{\sigma_2^2}{\sigma_1^2}.$$

Asymptotic Relative Efficiency (cont'd)

- **Example**: $X_1, \ldots, X_n$ be iid logistic(θ) with $EX_i = \theta$ and $VarX_i = \pi^2/3$.
- We have

$$\sqrt{n}(\bar{X} - \theta) \to_d N(0, \pi^2/3), \text{ and}$$
$$\sqrt{n}(\hat{\theta} - \theta) \to_d N(0, 3), \text{ as } n \to \infty.$$

  by CLT and asymptotic normality of the MLE, respectively.
- Note:

$$I_1(\theta) = \frac{1}{3} = -E\left\{\frac{\partial^2}{\partial\theta^2} \log f(x|\theta)\right\}.$$

- ARE$(\bar{X}, \hat{\theta}) = \frac{3}{\pi^2/3} = 9/\pi^2 \approx 0.91$.

Asymptotic Distribution of LRT

- The likelihood ratio statistic can be shown as

$$-2\log\lambda(\boldsymbol{x}) = 2\{\ell(\hat{\theta}) - \ell(\theta_0)\}.$$

- Taylor expansion of $\ell(\theta_0)$ around $\hat{\theta}$ leads to

$$\ell(\theta_0) = \ell(\hat{\theta}) + (\theta_0 - \hat{\theta})U(\hat{\theta}) - \frac{1}{2}(\theta_0 - \hat{\theta})^2 J(\hat{\theta}) + \frac{1}{6}(\theta_0 - \hat{\theta})^3 K(\theta^*).$$

- This implies

$$\begin{aligned} -2\log\lambda(\boldsymbol{x}) &= 2\{\ell(\hat{\theta}) - \ell(\theta_0)\} \\ &= \left\{\sqrt{n}(\hat{\theta} - \theta_0)\sqrt{\frac{J(\hat{\theta})}{n}}\right\}^2 + \frac{1}{3\sqrt{n}}\left\{\sqrt{n}(\hat{\theta} - \theta_0)\right\}^3 \frac{K(\theta^*)}{n} \end{aligned}$$

Asymptotic Distribution of LRT (cont'd)

- What is the asymptotic distribution of $\sqrt{n}(\hat{\theta} - \theta_0)$?
- What does $\sqrt{J(\hat{\theta})/n}$ converge in probability to?
- What does the first term converge in distribution to?
- One can see that $\frac{1}{3\sqrt{n}}$ converges to 0 and $K(\theta^*)/n$ converges almost surely to $EK_1(\theta_0)$.
- What does the second term converge in probability to?
- Combining the convergence of both terms, we may prove

$$-2\log\lambda(\boldsymbol{x}) \rightarrow_d \chi_1^2 \quad \text{as} \quad n \rightarrow \infty.$$

- One may have **Signed Likelihood Ratio Statistic**

$$\text{sign}(\hat{\theta} - \theta_0)\sqrt{-2\log\lambda(\boldsymbol{x})} \rightarrow_d N(0,1) \quad \text{as} \quad n \rightarrow \infty.$$

Hypothesis Tests in Large Samples

- When testing $H_0 : \theta = \theta_0$ and $H_1 : \theta \neq \theta_0$, we have
  (a) Likelihood ratio test: under H_0,
  $$2\{\ell(\hat{\theta}) - \ell(\theta_0)\} = -2\log\lambda(\boldsymbol{x}) \to_d \chi_1^2, \quad \text{as} \quad n \to \infty.$$
  (b) Score test: under H_0,
  $$\frac{U(\theta_0)}{\sqrt{nI_1(\theta_0)}} = \frac{U(\theta_0)}{\sqrt{I_n(\theta_0)}} \to_d N(0,1).$$
  (c) Wald test: under H_0, we have two options
  $$\sqrt{nI_1(\hat{\theta})}(\hat{\theta} - \theta_0) \to_d N(0,1), \quad \text{as} \quad n \to \infty,$$
  and
  $$\sqrt{J(\hat{\theta})}(\hat{\theta} - \theta_0) \to_d N(0,1), \quad \text{as} \quad n \to \infty,$$

Bernoulli Example

- Let $X_1, \ldots, X_n$ be iid Bernoulli(θ), $\theta \in (0, 1)$.
- The log-likelihood is $\ell(\theta) = \sum_{i=1}^{n} \ell_i(\theta|x_i)$, where

$$\ell_i(\theta|x_i) = x_i \log \frac{\theta}{1-\theta} + \log(1-\theta).$$

- The score function is $U(\theta) = \sum_{i=1}^{n} U_i(\theta|x_i)$, where

$$U_i(\theta|x_i) = \frac{x_i}{\theta(1-\theta)} - \frac{1}{1-\theta} = \frac{x_i - \theta}{\theta(1-\theta)}.$$

- The observed information is $J(\theta) = \sum_{i=1}^{n} J_i(\theta|x_i)$

$$J_i(\theta|x_i) = \frac{1}{\theta^2(1-\theta)^2}(x_i - 2x_i\theta + \theta^2).$$

Bernoulli Example (cont'd)

- Information number: $I_1(\theta) = E\{J_1(\theta|x_1)\} = \theta^{-1}(1-\theta)^{-1}$.
- To test $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$:
- Under the null hypothesis, we have

$$\sqrt{n}(\hat{\theta} - \theta_0) \to_d N(0, I_1(\theta_0)^{-1}), \text{ as } n \to \infty.$$

- Hence, the Wald test statistic is

$$\frac{\sqrt{n}(\hat{\theta} - \theta_0)}{\sqrt{I_1(\hat{\theta})^{-1}}} = \frac{\sqrt{n}(\hat{\theta} - \theta_0)}{\sqrt{\hat{\theta}(1-\hat{\theta})}}.$$

- Reject H_0 if

$$\left| \frac{\sqrt{n}(\bar{x} - \theta_0)}{\sqrt{\bar{x}(1-\bar{x})}} \right| \geq z_{1-\alpha/2}.$$

Bernoulli Example (cont'd)

- By the large sample normality of the score function, we have

$$n^{-1/2}U(\theta_0) \rightarrow_d N(0, I_1(\theta_0)).$$

- Hence, the score test statistic is

$$\frac{U(\theta_0)}{\sqrt{nI_1(\theta_0)}} = \frac{\sum_{i=1}^n (x_i - \theta_0)/\{\theta_0(1-\theta_0)\}}{\sqrt{n\theta_0^{-1}(1-\theta_0)^{-1}}} = \frac{\sqrt{n}(\bar{x} - \theta_0)}{\sqrt{\theta_0(1-\theta_0)}}.$$

- Reject H_0 if

$$\left| \frac{\sqrt{n}(\bar{x} - \theta_0)}{\sqrt{\theta_0(1-\theta_0)}} \right| \geq z_{1-\alpha/2}.$$

Bernoulli Example (cont'd)

- Based on LRT, we reject H_0 if $-2\log\lambda(\boldsymbol{x}) \geq \chi^2_{1,1-\alpha}$.
- Note that, in this example,

$$I_n(\hat{\theta}) = nI_1(\hat{\theta}) = \frac{n}{\bar{x}(1-\bar{x})},$$

and

$$J(\hat{\theta}) = \sum_{i=1}^{n} J_i(\theta|x_i) = \frac{1}{\bar{x}^2(1-\bar{x})^2}\sum_{i=1}^{n}(x_i - 2x_i\bar{x} + \bar{x}^2) = \frac{n}{\bar{x}(1-\bar{x})}.$$

- Here $I_n(\hat{\theta}) = J(\hat{\theta})$, not true in general.

Numerical Example

- Test $H_0 : \theta = 0.5$ versus $H_1 : \theta \neq 0.5$ given $\alpha = 0.05$.
- $n = 10$, $\sum x_i = 3$, $\hat{\theta} = \bar{x} = 0.3$.
- Likelihood ratio test:

$$-2\log\lambda(\boldsymbol{x}) = 2(10)\left(0.3\log\frac{0.3}{0.5} + 0.7\log\frac{0.7}{0.5}\right) \approx 1.646 < \chi^2_{1,1-\alpha} = 3.84.$$

- Score test:

$$\left|\frac{\sqrt{10}(0.3-0.5)}{\sqrt{0.5(1-0.5)}}\right| \approx 1.265 < z_{1-\alpha/2} = 1.96$$

- Wald test:

$$\left|\frac{\sqrt{10}(0.3-0.5)}{\sqrt{0.3(1-0.3)}}\right| \approx 1.38 < z_{1-\alpha/2} = 1.96$$

Intervals

- How do we derive interval estimators?
- Inverting acceptance regions:

$$\{\theta_0 : \delta(\boldsymbol{X}, \theta_0, \alpha) = 0\},$$

where δ may be one of the three tests.

Intervals: Bernoulli Example

- Likelihood ratio:

$$\left\{\theta_0 : 20\left[0.3\log\frac{0.3}{\theta_0} + 0.7\log\frac{0.7}{1-\theta_0}\right] \leq 3.84\right\} = (0.085, 0.606).$$

- Score test:

$$\left\{\theta_0 : \left|\frac{\sqrt{10}(0.3-\theta_0)}{\sqrt{\theta_0(1-\theta_0)}}\right| \leq 1.96\right\} = (0.108, 0.603).$$

- Wald test:

$$0.3 \pm 1.96\sqrt{\frac{0.3(1-0.3)}{10}} = (0.016, 0.584).$$

- These are large sample approximate 95% confidence intervals for θ. The “exact interval" (using CDF as a pivot) is (0.067,0.652).