Biostatistics 602 - Statistical Inference
Lecture 24
E-M Algorithm & Practice Examples

Hyun Min Kang

April 16th, 2013

Last Lecture

- What is an interval estimator?
- What is the coverage probability, confidence coefficient, and confidence interval?
- How can a $1-\alpha$ confidence interval typically be constructed?
- To obtain a lower-bounded (upper-tail) CI, whose acceptance region of a test should be inverted?
  - (a) $H_0 : \theta = \theta_0$ vs $H_0 : \theta > \theta_0$
  - (b) $H_0 : \theta = \theta_0$ vs $H_0 : \theta < \theta_0$

Interval Estimation

$\hat{\theta}(\mathbf{X})$ is usually represented as a point estimator

Interval Estimator

Let $[L(\mathbf{X}), U(\mathbf{X})]$, where $L(\mathbf{X})$ and $U(\mathbf{X})$ are functions of sample $\mathbf{X}$ and $L(\mathbf{X}) \leq U(\mathbf{X})$. Based on the observed sample $\mathbf{x}$, we can make an inference that

$$\theta \in [L(\mathbf{X}), U(\mathbf{X})]$$

Then we call $[L(\mathbf{X}), U(\mathbf{X})]$ an interval estimator of θ.

Three types of intervals

- Two-sided interval $[L(\mathbf{X}), U(\mathbf{X})]$
- One-sided (with lower-bound) interval $[L(\mathbf{X}), \infty)$
- One-sided (with upper-bound) interval $(-\infty, U(\mathbf{X})]$

Definitions

Definition : Coverage Probability

Given an interval estimator $[L(\mathbf{X}), U(\mathbf{X})]$ of θ, its *coverage probability* is defined as

$$\Pr(\theta \in [L(\mathbf{X}), U(\mathbf{X})])$$

In other words, the probability of a random variable in interval $[L(\mathbf{X}), U(\mathbf{X})]$ covers the parameter θ.

Definition: Confidence Coefficient

Confidence coefficient is defined as

$$\inf_{\theta \in \Omega} \Pr(\theta \in [L(\mathbf{X}), U(\mathbf{X})])$$

Definitions

Definition : Confidence Interval

Given an interval estimator $[L(\mathbf{X}), U(\mathbf{X})]$ of θ, if its confidence coefficient is $1-\alpha$, we call it a $(1-\alpha)$ *confidence interval*

Definition: Expected Length

Given an interval estimator $[L(\mathbf{X}), U(\mathbf{X})]$ of θ, its *expected length* is defined as

$$E[U(\mathbf{X}) - L(\mathbf{X})]$$

where $\mathbf{X}$ are random samples from $f_{\mathbf{X}}(\mathbf{x}|\theta)$. In other words, it is the average length of the interval estimator.

Confidence set and confidence interval

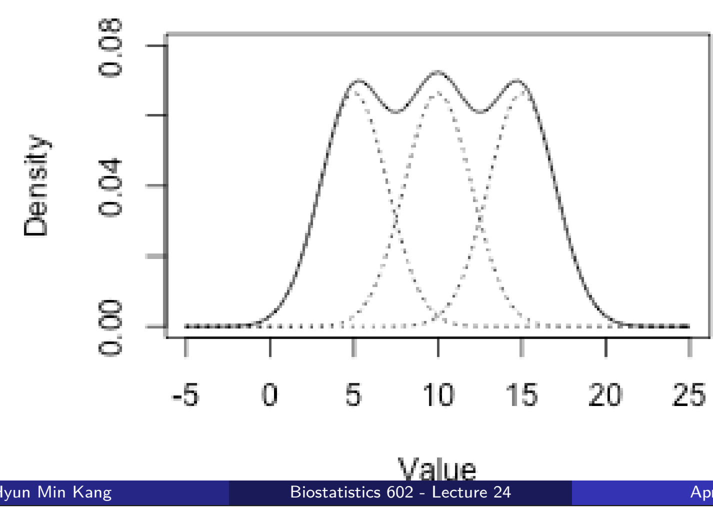
There is no guarantee that the confidence set obtained from Theorem 9.2.2 is an interval, but quite often

1. To obtain $(1-\alpha)$ two-sided CI $[L(\mathbf{X}), U(\mathbf{X})]$, we invert the acceptance region of a level α test for $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$
2. To obtain a lower-bounded CI $[L(\mathbf{X}), \infty)$, then we invert the acceptance region of a test for $H_0 : \theta = \theta_0$ vs. $H_1 : \theta > \theta_0$, where $\Omega = \{\theta : \theta \geq \theta_0\}$.
3. To obtain a upper-bounded CI $(-\infty, U(\mathbf{X})]$, then we invert the acceptance region of a test for $H_0 : \theta = \theta_0$ vs. $H_1 : \theta < \theta_0$, where $\Omega = \{\theta : \theta \leq \theta_0\}$.

Typical strategies for finding MLEs

1. Write the joint (log-)likelihood function, $L(\theta|\mathbf{x}) = f_{\mathbf{X}}(\mathbf{x}|\theta)$.
2. Find candidates that makes first order derivative to be zero
3. Check second-order derivative to check local maximum.
   (a) For one-dimensional parameter, negative second order derivative implies local maximum.
4. Check boundary points to see whether boundary gives global maximum.

Example: A mixture distribution



A general mixture distribution

$$f(x|\pi, \phi, \eta) = \sum_{i=1}^{k} \pi_i f(x; \phi_i, \eta)$$

- x observed data
- π mixture proportion of each component
- f the probability density function
- ϕ parameters specific to each component
- η parameters shared among components
- k number of mixture components

MLE Problem for mixture of normals

Problem

$$\begin{aligned} f(x|\theta = (\pi, \mu, \sigma^2)) &= \sum_{i=1}^{k} p_i f_i(x|\mu_i, \sigma_i^2) \\ f_i(x|\mu_i, \sigma_i^2) &= \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right] \\ \sum_{i=1}^{n} \pi_i &= 1 \end{aligned}$$

Find MLEs for $\theta = (\pi, \mu, \sigma^2)$.

Solution when $k = 1$

$$f(x|\theta) = \sum_{i=1}^{k} p_i f_i(x|\mu_i, \sigma_i^2)$$

- $\pi = \pi_1 = 1$
- $\mu = \mu_1 = \overline{x}$
- $\sigma^2 = \sigma_1^2 = \sum_{i=1}^{n} (x_i - \overline{x})^2 / n$

Incomplete data problem when $k > 1$

$$f(\mathbf{x}|\theta) = \prod_{i=1}^{n} \left[\sum_{j=1}^{k} p_i f_i(x_i|\mu_j, \sigma_j^2) \right]$$

The MLE solution is not analytically tractable, because it involves multiple sums of exponential functions.

Converting to a complete data problem

Let $z_i \in \{1, \cdots, k\}$ denote the source distribution where each x_i was sampled from.

$$\begin{aligned} f(\mathbf{x}|\mathbf{z},\theta) &= \prod_{i=1}^{n} \left[\sum_{j=1}^{k} I(z_i = j) f_i(x_i|\mu_j, \sigma_j^2) \right] = \prod_{i=1}^{n} f_i(x_i|\mu_{z_i}, \sigma_{z_i}^2) \\ \hat{\pi}_i &= \frac{\sum_{i=1}^{n} I(z_i = i)}{n} \\ \hat{\mu}_i &= \frac{\sum_{i=1}^{n} I(z_i = i) x_i}{\sum_{i=1}^{n} I(z_i = i)} \\ \hat{\sigma}_i^2 &= \frac{\sum_{i=1}^{n} I(z_i = i)(x_i - \hat{\mu}_i)^2}{\sum_{i=1}^{n} I(z_i = i)} \end{aligned}$$

The MLE solution is analytically tractable, if $\mathbf{z}$ is known.

E-M Algorithm

E-M (Expectation-Maximization) algorithm is

- A procedure for typically solving for the MLE.
- Guaranteed to converge the MLE (!)
- Particularly suited to the "missing data" problems where analytic solution of MLE is not tractable

The algorithm was derived and used in various special cases by a number of authors, but it was not identified as a general algorithm until the seminal paper by Dempster, Laird, and Rubin in Journal of Royal Statistical Society Series B (1977).

Overview of E-M Algorithm

Basic Structure

- $\mathbf{y}$ is observed (or incomplete) data
- $\mathbf{z}$ is missing (or augmented) data
- $\mathbf{x} = (\mathbf{y}, \mathbf{z})$ is complete data

Complete and incomplete data likelihood

- Complete data likelihood : $f(\mathbf{x}|\theta) = f(\mathbf{y}, \mathbf{z}|\theta)$
- Incomplete data likelihood : $g(\mathbf{y}|\theta) = \int f(\mathbf{y}, \mathbf{z}|\theta) \, d\mathbf{z}$

We are interested in MLE for $L(\theta|\mathbf{y}) = g(\mathbf{y}|\theta)$.

Maximizing incomplete data likelihood

$$\begin{aligned} L(\theta|\mathbf{y},\mathbf{z}) &= f(\mathbf{y},\mathbf{z}|\theta) \\ L(\theta|\mathbf{y}) &= g(\mathbf{y}|\theta) \\ k(\mathbf{z}|\theta,\mathbf{y}) &= \frac{f(\mathbf{y},\mathbf{z}|\theta)}{g(\mathbf{y}|\theta)} \\ \log L(\theta|\mathbf{y}) &= \log L(\theta|\mathbf{y},\mathbf{z}) - \log k(\mathbf{z}|\theta,\mathbf{y}) \end{aligned}$$

Because $\mathbf{z}$ is missing data, we replace the right side with its expectation under $k(\mathbf{z}|\theta', \mathbf{y})$, creating the new identity

$$\log L(\theta|\mathbf{y}) = \mathrm{E}\left[\log L(\theta|\mathbf{y},\mathbf{Z})|\theta',\mathbf{y}\right] - \mathrm{E}\left[\log k(\mathbf{Z}|\theta,\mathbf{y})|\theta',\mathbf{y}\right]$$

Iteratively maximizing the first term in the right-hand side results in E-M algorithm.

Overview of E-M Algorithm (cont'd)

Objective

- Maximize $L(\theta|\mathbf{y})$ or $l(\theta|\mathbf{y})$.
- Let $f(\mathbf{y},\mathbf{z}|\theta)$ denotes the pdf of complete data. In E-M algorithm, rather than working with $l(\theta|\mathbf{y})$ directly, we work with the surrogate function

$$Q(\theta|\theta^{(r)}) = \mathrm{E}\left[\log f(\mathbf{y},\mathbf{Z}|\theta)|\mathbf{y},\theta^{(r)}\right]$$

where $\theta^{(r)}$ is the estimation of θ in r-th iteration.
- $Q(\theta|\theta^{(r)})$ is the *expected log-likelihood of complete data*, conditioning on the observed data and $\theta^{(r)}$.

Key Steps of E-M algorithm

Expectation Step

- Compute $Q(\theta|\theta^{(r)})$.
- This typically involves in estimating the conditional distribution $\mathbf{Z}|\mathbf{Y}$, assuming $\theta = \theta^{(r)}$.
- After computing $Q(\theta|\theta^{(r)})$, move to the M-step

Maximization Step

- Maximize $Q(\theta|\theta^{(r)})$ with respect to θ.
- The $\arg\max_\theta Q(\theta|\theta^{(r)})$ will be the $(r+1)$-th θ to be fed into the E-step.
- Repeat E-step until convergence

E-M algorithm for mixture of normals

E-step

$$\begin{aligned}
Q(\theta|\theta^{(r)}) &= \mathrm{E}\left[\log f(\mathbf{y},\mathbf{Z}|\theta)|\mathbf{y},\theta^{(r)}\right] \\
&= \sum_{\mathbf{z}} k(\mathbf{z}|\theta^{(r)},\mathbf{y}) \log f(\mathbf{y},\mathbf{z}|\theta) \\
&= \sum_{i=1}^{n}\sum_{z_i=1}^{k} k(z_i|\theta^{(r)},y_i) \log f(y_i,z_i|\theta) \\
&= \sum_{i=1}^{n}\sum_{z_i=1}^{k} \frac{f(y_i,z_i|\theta^{(r)})}{g(y_i|\theta^{(r)})} \log f(y_i,z_i|\theta) \\
y_i,z_i|\theta &\sim \mathcal{N}(\mu_{z_i},\sigma^2_{z_i}) \\
g(y_i|\theta) &= \sum_{j=1}^{k} \pi_i f(y_i,z_i=j|\theta)
\end{aligned}$$

E-M algorithm for mixture of normals (cont'd)

M-step

$$\begin{aligned}
Q(\theta|\theta^{(r)}) &= \sum_{i=1}^{n}\sum_{z_i=1}^{k} \frac{f(y_i,z_i|\theta^{(r)})}{g(y_i|\theta^{(r)})} \log f(y_i,z_i|\theta) \\
\pi_j^{(r+1)} &= \frac{1}{n}\sum_{i=1}^{n} k(z_i=j|y_i,\theta^{(r)}) = \frac{1}{n}\frac{f(y_i,z_i=j|\theta^{(r)})}{g(y_i|\theta^{(r)})} \\
\mu_j^{(r+1)} &= \frac{\sum_{i=1}^n x_i k(z_i=j|y_i,\theta^{(r)})}{k(z_i=j|y_i,\theta^{(r)})} = \frac{\sum_{i=1}^n x_i k(z_i=j|y_i,\theta^{(r)})}{n\pi_j^{(r+1)}} \\
\sigma_j^{2,(r+1)} &= \frac{\sum_{i=1}^n (x_i-\mu_j^{(r+1)})^2 k(z_i=j|y_i,\theta^{(r)})}{k(z_i=j|y_i,\theta^{(r)})} \\
&= \frac{\sum_{i=1}^n (x_i-\mu_j^{(r+1)})^2 k(z_i=j|y_i,\theta^{(r)})}{n\pi_j^{(r+1)}}
\end{aligned}$$

Does E-M iteration converge to MLE?

Theorem 7.2.20 - Monotonic EM sequence

The sequence $\{\hat{\theta}^{(r)}\}$ defined by the E-M procedure satisfies

$$L\left(\hat{\theta}^{(r+1)}|\mathbf{y}\right) \geq L\left(\hat{\theta}^{(r)}|\mathbf{y}\right)$$

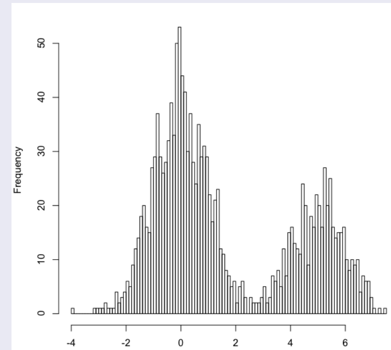
with equality holding if and only if successive iterations yield the same value of the maximized expected complete-data log likelihood, that is

$$E\left[\log L\left(\hat{\theta}^{(r+1)}|\mathbf{y},\mathbf{Z}\right)|\hat{\theta}^{(r)},\mathbf{y}\right] = E\left[\log L\left(\hat{\theta}^{(r)}|\mathbf{y},\mathbf{Z}\right)|\hat{\theta}^{(r)},\mathbf{y}\right]$$

Theorem 7.5.2 further guarantees that $L(\hat{\theta}^{(r)}|\mathbf{y})$ converges monotonically to $L(\hat{\theta}|\mathbf{y})$ for some stationary point $\hat{\theta}$.

A working example (from BIOSTAT615/815 Fall 2012)

Example Data (n=1,500)



Running example of implemented software

```
user@host~/> ./mixEM ./mix.dat
Maximum log-likelihood = 3043.46, at pi = (0.667842,0.332158)
between N(-0.0299457,1.00791) and N(5.0128,0.913825)
```


Practice Problem 1

Problem

Let $X_1, \cdots, X_n$ be a random sample from a population with pdf

$$f(x|\theta) = \frac{1}{2\theta} \qquad -\theta < x < \theta,\ \theta > 0$$

Find, if one exists, a best unbiased estimator of θ.

Strategy to solve the problem

- Can we use the Cramer-Rao bound? No, because the interchangeability condition does not hold
- Then, can we use complete sufficient statistics?
  1. Find a complete sufficient statistic T.
  2. For a trivial unbiased estimator of θ, and compute $\phi(T) = \mathrm{E}[W|T]$ or
  3. Make a function $\phi(T)$ such that $\mathrm{E}[\phi(T)] = \theta$.

Solution

First, we need to find a complete sufficient statistic.

$$f_X(x|\theta) = \frac{1}{2\theta}I(|x| < \theta)$$

$$f_{\mathbf{X}}(\mathbf{x}|\theta) = \frac{1}{(2\theta)^n}I(\max_i |x_i| < \theta)$$

Let $T(\mathbf{X}) = \max_i |X_i|$, then $f_T(t|\theta) = \frac{nt^{n-1}}{\theta^n}I(0 < t < \theta)$

$$E[g(T)] = \int_0^\theta \frac{nt^{n-1}g(t)}{\theta^n}dt = 0$$

$$\int_0^\theta t^{n-1}g(t)\,dt = 0$$

$$\theta^{n-1}g(\theta) = 0$$

$$g(\theta) = 0$$

Therefore the family of T is complete.

Solution

We need to make a $\phi(T)$ such that $E[\phi(T)] = \theta$.
First, let's see what the expectation of T is

$$E[g(T)] = \int_0^\theta t\frac{nt^{n-1}}{\theta^n}dt = \int_0^\theta \frac{nt^n}{\theta^n}dt = \frac{n}{n+1}\theta$$

$\phi(T) = \frac{n+1}{n}T$ is an unbiased estimator and a function of a complete sufficient statistic.
Therefore, $\phi(T)$ is the best unbiased estimator by Theorem 7.3.23.

Practice Problem 2

Problem

Let $X_1, \cdots, X_{n+1}$ be the iid $\text{Bernoulli}(p)$, and define the function $h(p)$ by

$$h(p) = \Pr\left(\sum_{i=1}^n X_i > X_{n+1} \middle| p\right)$$

the probability that the first n observations exceed the $(n+1)st$.

1. Show that
$$W(X_1, \cdots, X_{n+1}) = I\left(\sum_{i=1}^n X_i > X_{n+1}\right)$$
is an unbiased estimator of $h(p)$.
2. Find the best unbiased estimator of $h(p)$.

Solution for (a)

$$\begin{aligned} E[W] &= \sum_{\mathbf{x}} W(\mathbf{X})\Pr(\mathbf{X}) \\ &= \sum_{\mathbf{x}} I\left(\sum_{i=1}^n X_i > X_{n+1}\right)\Pr(\mathbf{x}) \\ &= \sum_{\sum_{i=1}^n X_i > X_{n+1}} \Pr(\mathbf{x}) \\ &= \Pr\left(\sum_{i=1}^n X_i > X_{n+1}\right) = h(p) \end{aligned}$$

Therefore T is an unbiased estimator of $h(p)$.

Solution for (b)

$T = \frac{1}{n+1}\sum_{i=1}^{n+1} X_i$ is complete sufficient statistic for p.

$$\begin{aligned} \phi(T) &= E[W|T] = \Pr(W = 1|T) \\ &= \Pr\left(\sum_{i=1}^n X_i > X_{n+1}|T\right) \end{aligned}$$

- If $T = 0$, then $\sum_{i=1}^n X_i = X_{n+1}$
- If $T = 1$, then
  - $\Pr(\sum_{i=1}^n X_i = 1 > X_{n+1} = 0) = n/(n+1)$
  - $\Pr(\sum_{i=1}^n X_i = 0 < X_{n+1} = 1) = 1/(n+1)$
- If $T = 2$ then
  - $\Pr(\sum_{i=1}^n X_i = 2 > X_{n+1} = 0) = \binom{n}{2}/\binom{n+1}{2} = (n-1)/(n+1)$
  - $\Pr(\sum_{i=1}^n X_i = 1 = X_{n+1} = 1) = 2/(n+1)$
- If $T > 2$, then $\sum_{i=1}^n X_i \geq 2 > 1 \geq X_{n+1}$

Solution for (b) (cont'd)

Therefore, the best unbiased estimator is

$$\begin{aligned} \phi(T) &= \Pr\left(\sum_{i=1}^n X_i > X_{n+1} | T\right) \\ &= \begin{cases} 0 & T = 0 \\ n/(n+1) & T = 1 \\ (n-1)/(n+1) & T = 2 \\ 1 & T \geq 3 \end{cases} \end{aligned}$$

Practice Problem 3

Problem

Suppose $X_1, \cdots, X_n$ are iid samples from $f(x|\theta) = \theta \exp(-\theta x)$. Suppose the prior distribution of θ is

$$\pi(\theta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} \theta^{\alpha-1} e^{-\theta/\beta}$$

where α, β are known.

(a) Derive the posterior distribution of θ.

(b) If we use the loss function $L(\theta, a) = (a - \theta)^2$, what is the Bayes rule estimator for θ?

(a) Posterior distribution of θ

$$\begin{aligned} f(\mathbf{x}, \theta) &= \pi(\theta) f(\mathbf{x}|\theta) \pi(\theta) \\ &= \frac{1}{\Gamma(\alpha)\beta^\alpha} \theta^{\alpha-1} e^{-\theta/\beta} \prod_{i=1}^n \left[\theta \exp\left(-\theta x_i\right)\right] \\ &= \frac{1}{\Gamma(\alpha)\beta^\alpha} \theta^{\alpha-1} e^{-\theta/\beta} \theta^n \exp\left(-\theta \sum_{i=1}^n x_i\right) \\ &= \frac{1}{\Gamma(\alpha)\beta^\alpha} \theta^{\alpha+n-1} \exp\left[-\theta\left(1/\beta + \sum_{i=1}^n x_i\right)\right] \\ &\propto \text{Gamma}\left(\alpha + n - 1, \frac{1}{\beta^{-1} + \sum_{i=1}^n x_i}\right) \\ \pi(\theta|\mathbf{x}) &= \text{Gamma}\left(\alpha + n - 1, \frac{1}{\beta^{-1} + \sum_{i=1}^n x_i}\right) \end{aligned}$$

(b) Bayes' rule estimator with squared error loss

Bayes' rule estimator with squared error loss is posterior mean. Note that the mean of $\text{Gamma}(\alpha, \beta)$ is $\alpha\beta$.

$$\begin{aligned} \pi(\theta|\mathbf{x}) &= \text{Gamma}\left(\alpha + n - 1, \frac{1}{\beta^{-1} + \sum_{i=1}^n x_i}\right) \\ E[\theta|\mathbf{x}] &= E[\pi(\theta|\mathbf{x})] \\ &= \frac{\alpha + n - 1}{\beta^{-1} + \sum_{i=1}^n x_i} \end{aligned}$$

Summary

Today

- E-M Algorithm
- Practice Problems for the Final Exam

Next Lectures

- Bayesian Tests
- Bayesian Intervals
- More practice problems