CHAPTER 5: MAXIMUM LIKELIHOOD ESTIMATION
Introduction to Efficient Estimation
• Goal

MLE is asymptotically efficient estimator under some regularity conditions.
• Basic setting

Suppose $X_1, \ldots, X_n$ are i.i.d from $P_{\theta_0}$ in the model $\mathcal{P}$.

(A0). $\theta \neq \theta^*$ implies $P_{\theta} \neq P_{\theta^*}$ (identifiability).

(A1). $P_{\theta}$ has a density function $p_{\theta}$ with respect to a dominating $\sigma$-finite measure $\mu$.

(A2). The set $\{x : p_{\theta}(x) > 0\}$ does not depend on $\theta$. 
• MLE definition

\[ L_n(\theta) = \prod_{i=1}^{n} p_{\theta}(X_i), \quad l_n(\theta) = \sum_{i=1}^{n} \log p_{\theta}(X_i). \]

\( L_n(\theta) \) and \( l_n(\theta) \) are called the \textit{likelihood function} and the \textit{log-likelihood function} of \( \theta \), respectively.

An estimator \( \hat{\theta}_n \) of \( \theta_0 \) is the maximum likelihood estimator (MLE) of \( \theta_0 \) if it maximizes the likelihood function \( L_n(\theta) \).
Ad Hoc Arguments
\[
\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, I(\theta_0)^{-1})
\]

- Consistency: \( \hat{\theta}_n \to \theta_0 \) (no asymptotic bias)

- Efficiency: asymptotic variance attains efficiency bound \( I(\theta_0)^{-1} \).
• Consistency

Definition 5.1 Let $P$ be a probability measure and let $Q$ be another measure on $(\Omega, \mathcal{A})$ with densities $p$ and $q$ with respect to a $\sigma$-finite measure $\mu$ ($\mu = P + Q$ always works). $P(\Omega) = 1$ and $Q(\Omega) \leq 1$. Then the Kullback-Leibler information $K(P, Q)$ is

$$K(P, Q) = E_P[\log \frac{p(X)}{q(X)}].$$
Proposition 5.1 \( K(P,Q) \) is well-defined, and \( K(P,Q) \geq 0 \). \( K(P,Q) = 0 \) if and only if \( P = Q \).

Proof

By the Jensen’s inequality,

\[
K(P,Q) = E_P[-\log \frac{q(X)}{p(X)}] \geq -\log E_P\left[\frac{q(X)}{p(X)}\right] = -\log Q(\Omega) \geq 0.
\]

The equality holds if and only if \( p(x) = Mq(x) \) almost surely with respect \( P \) and \( Q(\Omega) = 1 \)

\( \Rightarrow P = Q. \)
• Why is MLE consistent?

\( \hat{\theta}_n \) maximizes \( l_n(\theta) \),

\[
\frac{1}{n} \sum_{i=1}^{n} p_{\hat{\theta}_n}(X_i) \geq \frac{1}{n} \sum_{i=1}^{n} p_{\theta_0}(X_i).
\]

Suppose \( \hat{\theta}_n \to \theta^* \). Then we would expect to the both sides converge to

\[
E_{\theta_0}[p_{\theta^*}(X)] \geq E_{\theta_0}[p_{\theta_0}(X)],
\]

which implies \( K(P_{\theta_0}, P_{\theta^*}) \leq 0 \).

From Prop. 5.1, \( P_{\theta_0} = P_{\theta^*} \). From A0, \( \theta^* = \theta_0 \). That is, \( \hat{\theta}_n \) converges to \( \theta_0 \).
• Why is MLE efficient?

Suppose \( \hat{\theta}_n \to \theta_0 \). \( \hat{\theta}_n \) solves the following likelihood (or score) equations

\[
\dot{l}_n(\hat{\theta}_n) = \sum_{i=1}^{n} \dot{l}_{\hat{\theta}_n}(X_i) = 0.
\]

Taylor expansion at \( \theta_0 \):

\[
- \sum_{i=1}^{n} \dot{l}_{\theta_0}(X_i) = - \sum_{i=1}^{n} \ddot{l}_{\theta^*}(X_i)(\hat{\theta} - \theta_0),
\]

where \( \theta^* \) is between \( \theta_0 \) and \( \hat{\theta} \).

\[
\sqrt{n}(\hat{\theta} - \theta_0) = - \frac{1}{\sqrt{n}} \left\{ n^{-1} \sum_{i=1}^{n} \ddot{l}_{\theta^*}(X_i) \right\} \left\{ \sum_{i=1}^{n} \dot{l}_{\theta_0}(X_i) \right\}.
\]
\( \sqrt{n}(\hat{\theta}_n - \theta_0) \) is asymptotically equivalent to
\[
\frac{1}{\sqrt{n}} \sum_{i=1}^{n} I(\theta_0)^{-1} \dot{l}_{\theta_0}(X_i). 
\]

Then \( \hat{\theta}_n \) is an asymptotically linear estimator of \( \theta_0 \) with the influence function \( I(\theta_0)^{-1} \dot{l}_{\theta_0} = \tilde{l}(\cdot, P_{\theta_0}|\theta, \mathcal{P}) \).
Consistency Results
Theorem 5.1 Consistency with dominating function

Suppose that
(a) \( \Theta \) is compact.
(b) \( \log p_\theta(x) \) is continuous in \( \theta \) for all \( x \).
(c) There exists a function \( F(x) \) such that 
\[ E_{\theta_0}[F(X)] < \infty \quad \text{and} \quad |\log p_\theta(x)| \leq F(x) \quad \text{for all} \quad x \quad \text{and} \quad \theta. \]
Then \( \hat{\theta}_n \to_{a.s.} \theta_0. \)
Proof

For any sample $\omega \in \Omega$, $\hat{\theta}_n$ is compact. By choosing a subsequence, $\hat{\theta}_n \rightarrow \theta^*$.

If $\frac{1}{n} \sum_{i=1}^{n} l_{\hat{\theta}_n}(X_i) \rightarrow E_{\theta_0}[l_{\theta^*}(X)]$, then since

$$\frac{1}{n} \sum_{i=1}^{n} l_{\hat{\theta}_n}(X_i) \geq \frac{1}{n} \sum_{i=1}^{n} l_{\theta_0}(X_i),$$

$\Rightarrow E_{\theta_0}[l_{\theta^*}(X)] \geq E_{\theta_0}[l_{\theta_0}(X)].$

$\Rightarrow \theta^* = \theta_0$. Done!

It remains to show $P_n[l_{\hat{\theta}_n}(X)] \equiv \frac{1}{n} \sum_{i=1}^{n} l_{\hat{\theta}_n}(X_i) \rightarrow E_{\theta_0}[l_{\theta^*}(X)].$

It suffices to show

$$|P_n[l_{\hat{\theta}}(X)] - E_{\theta_0}[l_{\hat{\theta}}(X)]| \rightarrow 0.$$
We can even prove the following uniform convergence result

\[
\sup_{\theta \in \Theta} |P_n[l_\theta(X)] - E_{\theta_0}[l_\theta(X)]| \to 0.
\]

Define

\[
\psi(x, \theta, \rho) = \sup_{|\theta' - \theta| < \rho} (l_{\theta'}(x) - E_{\theta_0}[l_{\theta'}(X)]).
\]

Since \( l_\theta \) is continuous, \( \psi(x, \theta, \rho) \) is measurable and by the DCT, \( E_{\theta_0}[\psi(X, \theta, \rho)] \) decreases to \( E_{\theta_0}[l_\theta(x) - E_{\theta_0}[l_\theta(X)]] = 0 \).

\[\Rightarrow\] for \( \epsilon > 0 \), for any \( \theta \in \Theta \), there exists a \( \rho_\theta \) such that

\[
E_{\theta_0}[\psi(X, \theta, \rho_\theta)] < \epsilon.
\]
The union of \( \{ \theta' : |\theta' - \theta| < \rho_\theta \} \) covers \( \Theta \). By the compactness of \( \Theta \), there exists a finite number of \( \theta_1, ..., \theta_m \) such that
\[
\Theta \subset \bigcup_{i=1}^{m} \{ \theta' : |\theta' - \theta_i| < \rho_{\theta_i} \}.
\]

\[\Rightarrow\]
\[
\sup_{\theta \in \Theta} \{ P_n[l_\theta(X)] - E_{\theta_0}[l_\theta(X)] \} \leq \sup_{1 \leq i \leq m} P_n[\psi(X, \theta_i, \rho_{\theta_i})].
\]

\[
\limsup \sup_{n} \sup_{\theta \in \Theta} \{ P_n[l_\theta(X)] - E_{\theta_0}[l_\theta(X)] \} \leq \sup_{1 \leq i \leq m} P_{\theta}[\psi(X, \theta_i, \rho_{\theta_i})] \leq \epsilon.
\]

\[\Rightarrow\]
\[
\limsup \sup_{n} \sup_{\theta \in \Theta} \{ P_n[l_\theta(X)] - E_{\theta_0}[l_\theta(X)] \} \leq 0. \text{ Similarly,} \]
\[
\limsup \sup_{n} \sup_{\theta \in \Theta} \{ -P_n[l_\theta(X)] + E_{\theta_0}[l_\theta(X)] \} \geq 0.
\]

\[\Rightarrow\]
\[
\limsup \sup_{n} \sup_{\theta \in \Theta} |P_n[l_\theta(X)] - E_{\theta_0}[l_\theta(X)]| \to 0.
\]
Theorem 5.2 Wald’s Consistency \( \Theta \) is compact. Suppose \( \theta \mapsto l_\theta(x) = \log p_\theta(x) \) is upper-semicontinuous for all \( x \), in the sense \( \limsup_{\theta' \to \theta} l_{\theta'}(x) \leq l_\theta(x) \). Suppose for every sufficient small ball \( U \subset \Theta \), 
\[ E_{\theta_0} \left[ \sup_{\theta' \in U} l_{\theta'}(X) \right] < \infty. \] Then \( \hat{\theta}_n \to_p \theta_0 \).
Proof

\[ E_{\theta_0}[l_{\theta_0}(X)] > E_{\theta_0}[l_{\theta'}(X)] \] for any \( \theta' \neq \theta_0 \)

\[ \Rightarrow \] there exists a ball \( U_{\theta'} \) containing \( \theta' \) such that

\[ E_{\theta_0}[l_{\theta_0}(X)] > E_{\theta_0}[\sup_{\theta^* \in U_{\theta'}} l_{\theta^*}(X)]. \]

Otherwise, there exists a sequence \( \theta_m^* \to \theta' \) but

\[ E_{\theta_0}[l_{\theta_0}(X)] \leq E_{\theta_0}[l_{\theta_m^*}(X)]. \] Since \( l_{\theta_m^*}(x) \leq \sup_{U'} l_{\theta'}(X) \) where \( U' \) is the ball satisfying the condition,

\[ \limsup_m E_{\theta_0}[l_{\theta_m^*}(X)] \leq E_{\theta_0}[\limsup_m l_{\theta_m^*}(X)] \leq E_{\theta_0}[l_{\theta'}(X)]. \]

\[ \Rightarrow E_{\theta_0}[l_{\theta_0}(X)] \leq E_{\theta_0}[l_{\theta'}(X)] \] contradiction!
For any $\epsilon$, the balls $\bigcup_{\theta'} U_{\theta'}$ covers the compact set $\Theta \cap \{ |\theta' - \theta_0| > \epsilon \}$ \Rightarrow there exists a finite covering balls, $U_1, \ldots, U_m$.

\[
P(|\hat{\theta}_n - \theta_0| > \epsilon) \leq P(\sup_{|\theta' - \theta_0| > \epsilon} P_n[l_{\theta'}(X)] \geq P_n[l_{\theta_0}(X)])
\]
\[
\leq P(\max_{1 \leq i \leq m} P_n[\sup_{\theta' \in U_i} l_{\theta'}(X)] \geq P_n[l_{\theta_0}(X)])
\]
\[
\leq \sum_{i=1}^{m} P(P_n[\sup_{\theta' \in U_i} l_{\theta'}(X)] \geq P_n[l_{\theta_0}(X)]).
\]

Since

\[
P_n[\sup_{\theta' \in U_i} l_{\theta'}(X)] \rightarrow_{a.s.} E_{\theta_0}[\sup_{\theta' \in U_i} l_{\theta'}(X)] < E_{\theta_0}[l_{\theta_0}(X)],
\]

the right-hand side converges to zero. \Rightarrow $\hat{\theta}_n \rightarrow_p \theta_0$. 
Asymptotic Efficiency Result
Theorem 5.3 Suppose that the model \( \mathcal{P} = \{ P_\theta : \theta \in \Theta \} \) is Hellinger differentiable at an inner point \( \theta_0 \) of \( \Theta \subset R^k \). Furthermore, suppose that there exists a measurable function \( F \) with \( E_{\theta_0}[F^2] < \infty \) such that for every \( \theta_1 \) and \( \theta_2 \) in a neighborhood of \( \theta_0 \),

\[
| \log p_{\theta_1}(x) - \log p_{\theta_2}(x) | \leq F(x)|\theta_1 - \theta_2|.
\]

If the Fisher information matrix \( I(\theta_0) \) is nonsingular and \( \hat{\theta}_n \) is consistent, then

\[
\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I(\theta_0)^{-1} i_{\theta_0}(X_i) + o_{p\theta_0}(1).
\]

In particular, \( \sqrt{n}(\hat{\theta}_n - \theta_0) \rightarrow_d N(0, I(\theta_0)^{-1}) \).
Proof

For any $h_n \to h$, by the Hellinger differentiability,

$$W_n = 2 \left( \sqrt{\frac{p_{\theta_0} + h_n / \sqrt{n}}{p_{\theta_0}}} - 1 \right) \to h^T \dot{l}_{\theta_0}, \text{ in } L_2(P_{\theta_0}).$$

⇒

$$\sqrt{n}(\log p_{\theta_0 + h_n / \sqrt{n}} - \log p_{\theta_0}) = 2\sqrt{n} \log(1 + W_n/2) \to_p h^T \dot{l}_{\theta_0}.$$

⇒

$$E_{\theta_0} \left[ \sqrt{n}(P_n - P) \left[ \sqrt{n}(\log p_{\theta_0 + h_n / \sqrt{n}} - \log p_{\theta_0}) - h^T \dot{l}_{\theta_0} \right] \right] \to 0$$

$$Var_{\theta_0} \left[ \sqrt{n}(P_n - P) \left[ \sqrt{n}(\log p_{\theta_0 + h_n / \sqrt{n}} - \log p_{\theta_0}) - h^T \dot{l}_{\theta_0} \right] \right] \to 0.$$

⇒

$$\sqrt{n}(P_n - P) \left[ \sqrt{n}(\log p_{\theta_0 + h_n / \sqrt{n}} - \log p_{\theta_0}) - h^T \dot{l}_{\theta_0} \right] \to_p 0.$$
From Step I in proving Theorem 4.1,

\[
\log \prod_{i=1}^{n} \frac{\log p_{\theta_0 + h_n/\sqrt{n}}}{\log p_{\theta_0}} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} h^T l_{\theta_0}(X_i) - \frac{1}{2} h^T I(\theta_0) h + o_{\theta_0}(1).
\]

\[
nE_{\theta_0} \left[ \log p_{\theta_0 + h_n/\sqrt{n}} - \log p_{\theta_0} \right] \to -h^T I(\theta_0) h / 2.
\]

\[
\Rightarrow \quad nP_n \left[ \log p_{\theta_0 + h_n/\sqrt{n}} - \log p_{\theta_0} \right] = -\frac{1}{2} h^T n I(\theta_0) h_n + h_n \sqrt{n} (P_n - P) [l_{\theta_0}] + o_{\theta_0}(1).
\]

Choose \( h_n = \sqrt{n} (\hat{\theta}_n - \theta_0) \) and \( h_n = I(\theta_0)^{-1} \sqrt{n} (P_n - P) [l_{\theta_0}] \). \( \Rightarrow \)

\[
nP_n \left[ \log p_{\hat{\theta}_n} - \log p_{\theta_0} \right] = \frac{1}{2} \left\{ \sqrt{n} (P_n - P) [l_{\theta_0}] \right\}^T I(\theta_0)^{-1} \left\{ \sqrt{n} (P_n - P) [l_{\theta_0}] \right\} + o_{\theta_0}(1).
\]
Compare the above two equations:

\[-\frac{1}{2} \left\{ \sqrt{n}(\hat{\theta}_n - \theta_0) + I(\theta_0)^{-1} \sqrt{n}(P_n - \mathcal{P})[\dot{\theta}_0] \right\}^T I(\theta_0) \]

\[ \times \left\{ \sqrt{n}(\hat{\theta}_n - \theta_0) + I(\theta_0)^{-1} \sqrt{n}(P_n - \mathcal{P})[\dot{\theta}_0] \right\} \]

\[ + o_p(1) \geq 0. \]

\[ \Rightarrow \]

\[ \sqrt{n}(\hat{\theta}_n - \theta_0) = -I(\theta_0)^{-1} \sqrt{n}(P_n - \mathcal{P})[\dot{\theta}_0] + o_p(1). \]
**Theorem 5.4** For each \( \theta \) in an open subset of Euclidean space. Let \( \theta \mapsto \dot{l}_\theta(x) = \log p_\theta(x) \) be twice continuously differentiable for every \( x \). Suppose \( E_{\theta_0}[\dot{l}_{\theta_0} \dot{l}'_{\theta_0}] < \infty \) and \( E[\ddot{l}_{\theta_0}] \) exists and is nonsingular. Assume that the second partial derivative of \( \dot{l}_\theta(x) \) is dominated by a fixed integrable function \( F(x) \) for every \( \theta \) in a neighborhood of \( \theta_0 \). Suppose \( \hat{\theta}_n \rightarrow_p \theta_0 \). Then

\[
\sqrt{n}(\hat{\theta}_n - \theta_0) = -(E_{\theta_0}[\ddot{l}_{\theta_0}])^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \dot{l}_{\theta_0}(X_i) + o_{p_{\theta_0}}(1).
\]
Proof

\( \hat{\theta}_n \) solves \( 0 = \sum_{i=1}^{n} i_{\hat{\theta}}(X_i) \).

\[ 0 = \sum_{i=1}^{n} i_{\theta_0}(X_i) + \sum_{i=1}^{n} i_{\theta_0}(\hat{\theta}_n - \theta_0) + \frac{1}{2}(\hat{\theta}_n - \theta_0)^T \left\{ \sum_{i=1}^{n} l^{(3)}_{\hat{\theta}_n} \right\} (\hat{\theta}_n - \theta_0). \]

\[ \Rightarrow \left| \left\{ \frac{1}{n} \sum_{i=1}^{n} i_{\theta_0}(X_i) \right\} (\hat{\theta}_n - \theta_0) - \frac{1}{n} \sum_{i=1}^{n} i_{\theta_0}(X_i) \right| \leq \frac{1}{n} \sum_{i=1}^{n} |F(X_i)|. \]

\[ \Rightarrow (\hat{\theta}_n - \theta_0) = o_p(1). \]

\[ \Rightarrow \sqrt{n}(\hat{\theta}_n - \theta_0) \left\{ \frac{1}{n} \sum_{i=1}^{n} i_{\theta_0}(X_i) + o_p(1) \right\} = -\frac{1}{\sqrt{n}} \sum_{i=1}^{n} l_{\theta_0}(X_i). \]
Computation of MLE
• Solve likelihood equation

\[ \sum_{i=1}^{n} \dot{l}_\theta(X_i) = 0. \]

– Newton-Raphson iteration: at kth iteration,

\[ \theta^{(k+1)} = \theta^{(k)} - \left\{ \frac{1}{n} \sum_{i=1}^{n} i_{\theta^{(k)}}(X_i) \right\}^{-1} \left\{ \frac{1}{n} \sum_{i=1}^{n} i_{\theta^{(k)}}(X_i) \right\}. \]

– Note \( -\frac{1}{n} \sum_{i=1}^{n} i_{\theta^{(k)}}(X_i) \approx I(\theta^{(k)}). \Rightarrow Fisher scoring algorithm:

\[ \theta^{(k+1)} = \theta^{(k)} + I(\theta^{(k)})^{-1} \left\{ \frac{1}{n} \sum_{i=1}^{n} i_{\theta^{(k)}}(X_i) \right\}. \]
• Optimize the likelihood function

optimum search algorithm: grid search, quasi-Newton method (gradient decent algorithm), MCMC, simulation annealing
EM Algorithm for Missing Data
When part of data is missing or some mis-measured data is observed, a commonly used algorithm is called the \textit{expectation-maximization} (EM) algorithm.

- **Framework of EM algorithm**
  - $Y = (Y_{mis}, Y_{obs})$.
  - $R$ is a vector of 0/1 indicating which subjects are missing/not missing. Then $Y_{obs} = RY$.
  - The density function for the observed data $(Y_{obs}, R)$
    \[
    \int_{Y_{mis}} f(Y; \theta) P(R|Y) dY_{mis}.
    \]
• **Missing mechanism**

Missing at random assumption (MAR):

\[ P(R|Y) = P(R|Y_{obs}) \quad \text{and} \quad P(R|Y) \text{ does not depend on } \theta; \]

i.e., the missing probability only depends on the observed data and it is informative about \( \theta \).

Under MAR,

\[
\int_{Y_{mis}} f(Y; \theta) \, dY_{mis} \, P(R|Y).
\]

We maximize

\[
\int_{Y_{mis}} f(Y; \theta) \, dY_{mis} \quad \text{or} \quad \log \int_{Y_{mis}} f(Y; \theta) \, dY_{mis}
\]
• Details of EM algorithm

We start from any initial value of $\theta^{(1)}$ and use the following iterations. The $k$th iteration consists both E-step and M-step:

**E-step.** We evaluate the conditional expectation

$$E \left[ \log f(Y; \theta) | Y_{obs}, \theta^{(k)} \right].$$

$$E \left[ \log f(Y; \theta) | Y_{obs}, \theta^{(k)} \right] = \frac{\int_{Y_{mis}} [\log f(Y; \theta)] f(Y; \theta^{(k)}) dY_{mis}}{\int_{Y_{mis}} f(Y; \theta^{(k)}) dY_{mis}}.$$
$M$-step. We obtain $\theta^{(k+1)}$ by maximizing

$$E \left[ \log f(Y; \theta) | Y_{obs}, \theta^{(k)} \right].$$

We then iterate till the convergence of $\theta$; i.e., the difference between $\theta^{(k+1)}$ and $\theta^{(k)}$ is less than a given criteria.
• **Rationale why EM works**

**Theorem 5.5** At each iteration of the EM algorithm,

\[ \log f(Y_{obs}; \theta^{(k+1)}) \geq \log f(Y_{obs}, \theta^{(k)}) \]

and the equality holds if and only if \( \theta^{(k+1)} = \theta^{(k)} \).
Proof

\[ E \left[ \log f(Y; \theta^{(k+1)}) | Y_{obs}, \theta^{(k)} \right] \geq E \left[ \log f(Y; \theta^{(k)}) | Y_{obs}, \theta^{(k)} \right]. \]

\[ \Rightarrow \]

\[ E \left[ \log f(Y_{mis} | Y_{obs}; \theta^{(k+1)}) | Y_{obs}, \theta^{(k)} \right] + \log f(Y_{obs}; \theta^{(k+1)}) \]

\[ \geq E \left[ \log f(Y_{mis} | Y_{obs}, \theta^{(k)}) | Y_{obs}, \theta^{(k)} \right] + \log f(Y_{obs}; \theta^{(k)}). \]

\[ E \left[ \log f(Y_{mis} | Y_{obs}; \theta^{(k+1)}) | Y_{obs}, \theta^{(k)} \right] \]

\[ \leq E \left[ \log f(Y_{mis} | Y_{obs}, \theta^{(k)}) | Y_{obs}, \theta^{(k)} \right], \]

\[ \Rightarrow \log f(Y_{obs}; \theta^{(k+1)}) \geq \log f(Y_{obs}, \theta^{(k)}). \] The equality holds iff

\[ \log f(Y_{mis} | Y_{obs}, \theta^{(k+1)}) = \log f(Y_{mis} | Y_{obs}, \theta^{(k)}), \]

\[ \Rightarrow \log f(Y; \theta^{(k+1)}) = \log f(Y; \theta^{(k)}). \]
Incorporating Newton-Raphson in EM

E-step. We evaluate the conditional expectation

\[ E \left[ \frac{\partial}{\partial \theta} \log f(Y; \theta) | Y_{obs}, \theta^{(k)} \right] \]

and

\[ E \left[ \frac{\partial^2}{\partial \theta^2} \log f(Y; \theta) | Y_{obs}, \theta^{(k)} \right] \]
\( M\)-step. We obtain \( \theta^{(k+1)} \) by solving

\[
0 = E \left[ \frac{\partial}{\partial \theta} \log f(Y; \theta) | Y_{obs}, \theta^{(k)} \right]
\]

using one-step Newton-Raphson iteration:

\[
\theta^{(k+1)} = \theta^{(k)} - \left\{ E \left[ \frac{\partial^2}{\partial \theta^2} \log f(Y; \theta) | Y_{obs}, \theta^{(k)} \right] \right\}^{-1} E \left[ \frac{\partial}{\partial \theta} \log f(Y; \theta) | Y_{obs}, \theta^{(k)} \right] \bigg|_{\theta=\theta^{(k)}}.
\]
Example

Suppose a random vector $Y$ has a multinomial distribution with $n = 197$ and

$$p = \left( \frac{1}{2} + \frac{\theta}{4}, \frac{1 - \theta}{4}, \frac{1 - \theta}{4}, \frac{\theta}{4} \right).$$

Then the probability for $Y = (y_1, y_2, y_3, y_4)$ is given by

$$\frac{n!}{y_1!y_2!y_3!y_4!} \left( \frac{1}{2} + \frac{\theta}{4} \right)^{y_1} \left( \frac{1 - \theta}{4} \right)^{y_2} \left( \frac{1 - \theta}{4} \right)^{y_3} \left( \frac{\theta}{4} \right)^{y_4}.$$

Suppose we observe $Y = (125, 18, 20, 34)$. If we start with $\theta^{(1)} = 0.5$, after the convergence in the Newton-Raphson iteration, we obtain $\theta^{(k)} = 0.6268215$. 
– EM algorithm: the full data is $X$ has a multivariate normal distribution with $n$ and the $p = (1/2, \theta/4, (1 - \theta)/4, (1 - \theta)/4, \theta/4)$.

$Y = (X_1 + X_2, X_3, X_4, X_5)$. 
The score equation for the complete data $X$ is simple
\[ 0 = \frac{X_2 + X_5}{\theta} - \frac{X_3 + X_4}{1 - \theta}. \]

M-step of the EM algorithm needs to solve the equation
\[ 0 = E \left[ \frac{X_2 + X_5}{\theta} - \frac{X_3 + X_4}{1 - \theta} \mid Y, \theta^{(k)} \right]; \]
while the E-step evaluates the above expectation.

\[ E[X \mid Y, \theta^{(k)}] = (Y_1 \frac{1/2}{1/2 + \theta^{(k)}/4}, Y_1 \frac{\theta^{(k)}/4}{1/2 + \theta^{(k)}/4}, Y_2, Y_3, Y_4). \]

\[ \theta^{(k+1)} = \frac{E[X_2 + X_5 \mid Y, \theta^{(k)}]}{E[X_2 + X_5 + X_3 + X_4 \mid Y, \theta^{(k)}]} = \frac{Y_1 \frac{\theta^{(k)}/4}{1/2 + \theta^{(k)}/4} + Y_4}{Y_1 \frac{\theta^{(k)}/4}{1/2 + \theta^{(k)}/4} + Y_2 + Y_3 + Y_4}. \]

We start form $\theta^{(1)} = 0.5$. 
<table>
<thead>
<tr>
<th>$k$</th>
<th>$\theta^{(k+1)}$</th>
<th>$\theta^{(k+1)} - \theta^{(k)}$</th>
<th>$\frac{\theta^{(k+1)} - \hat{\theta}_n}{\theta^{(k)} - \hat{\theta}_n}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>.5000000000</td>
<td>.126821498</td>
<td>.1465</td>
</tr>
<tr>
<td>1</td>
<td>.608247423</td>
<td>.018574075</td>
<td>.1346</td>
</tr>
<tr>
<td>2</td>
<td>.624321051</td>
<td>.002500447</td>
<td>.1330</td>
</tr>
<tr>
<td>3</td>
<td>.626488879</td>
<td>.000332619</td>
<td>.1328</td>
</tr>
<tr>
<td>4</td>
<td>.626777323</td>
<td>.000044176</td>
<td>.1328</td>
</tr>
<tr>
<td>5</td>
<td>.626815632</td>
<td>.000005866</td>
<td>.1328</td>
</tr>
<tr>
<td>6</td>
<td>.626820719</td>
<td>.000000779</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>.626821395</td>
<td>.000000104</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>.626821484</td>
<td>.000000014</td>
<td></td>
</tr>
</tbody>
</table>
• Conclusions

– the EM converges and the result agrees with what is obtained from the Newton-Raphson iteration;

– the EM convergence is linear as
  \[(\theta^{(k+1)} - \hat{\theta}_n)/(\theta^{(k)} - \hat{\theta}_n)\]
  becomes a constant when convergence;

– the convergence in the Newton-Raphson iteration is quadratic in the sense
  \[(\theta^{(k+1)} - \hat{\theta}_n)/(\theta^{(k)} - \hat{\theta}_n)^2\]
  becomes a constant when convergence;

– the EM is much less complex than the Newton-Raphson iteration and this is the advantage of using the EM algorithm.
• **More example**

  – the example of exponential mixture model: Suppose $Y \sim P_\theta$ where $P_\theta$ has density

  $$p_\theta(y) = \left\{ p\lambda e^{-\lambda y} + (1 - p)\mu e^{-\mu y} \right\} I(y > 0)$$

  and $\theta = (p, \lambda, \mu) \in (0, 1) \times (0, \infty) \times (0, \infty)$. Consider estimation of $\theta$ based on $Y_1, \ldots, Y_n$ i.i.d $p_\theta(y)$. Solving the likelihood equation using the Newton-Raphson is much computation involved.
EM algorithm: the complete data \( X = (Y, \Delta) \sim p_\theta(x) \) where

\[
p_\theta(x) = p_\theta(y, \delta) = (pye^{-\lambda y})^\delta ((1 - p)\mu e^{-\mu y})^{1-\delta}.
\]

This is natural from the following mechanism: \( \Delta \) is a bernoulli variable with \( P(\Delta = 1) = p \) and we generate \( Y \) from Exp(\( \lambda \)) if \( \Delta = 1 \) and from Exp(\( \mu \)) if \( \Delta = 0 \). Thus, \( \Delta \) is missing. The score equation for \( \theta \) based on \( X \) is equal to

\[
0 = \dot{l}_p(X_1, ..., X_n) = \sum_{i=1}^{n} \left\{ \frac{\Delta_i}{p} - \frac{1 - \Delta_i}{1 - p} \right\},
\]

\[
0 = \dot{l}_\lambda(X_1, ..., X_n) = \sum_{i=1}^{n} \Delta_i \left( \frac{1}{\lambda} - Y_i \right),
\]

\[
0 = \dot{l}_\mu(X_1, ..., X_n) = \sum_{i=1}^{n} (1 - \Delta_i) \left( \frac{1}{\mu} - Y_i \right).
\]
M-step solves the equations

\[
0 = \sum_{i=1}^{n} E \left[ \left\{ \frac{\Delta_i}{p} - \frac{1 - \Delta_i}{1 - p} \right\} | Y_1, \ldots, Y_n, p^{(k)}, \lambda^{(k)}, \mu^{(k)} \right] \\
= \sum_{i=1}^{n} E \left[ \left\{ \frac{\Delta_i}{p} - \frac{1 - \Delta_i}{1 - p} \right\} | Y_i, p^{(k)}, \lambda^{(k)}, \mu^{(k)} \right],
\]

\[
0 = \sum_{i=1}^{n} E \left[ \Delta_i \left( \frac{1}{\lambda} - Y_i \right) | Y_1, \ldots, Y_n, p^{(k)}, \lambda^{(k)}, \mu^{(k)} \right] \\
= \sum_{i=1}^{n} E \left[ \Delta_i \left( \frac{1}{\lambda} - Y_i \right) | Y_i, p^{(k)}, \lambda^{(k)}, \mu^{(k)} \right],
\]

\[
0 = \sum_{i=1}^{n} E \left[ 1 - \Delta_i \left( \frac{1}{\mu} - Y_i \right) | Y_1, \ldots, Y_n, p^{(k)}, \lambda^{(k)}, \mu^{(k)} \right] \\
= \sum_{i=1}^{n} E \left[ 1 - \Delta_i \left( \frac{1}{\mu} - Y_i \right) | Y_i, p^{(k)}, \lambda^{(k)}, \mu^{(k)} \right].
\]
This immediately gives

\[ p^{(k+1)} = \frac{1}{n} \sum_{i=1}^{n} E[\Delta_i | Y_i, p^{(k)}, \lambda^{(k)}, \mu^{(k)}], \]

\[ \lambda^{(k+1)} = \frac{\sum_{i=1}^{n} E[\Delta_i | Y_i, p^{(k)}, \lambda^{(k)}, \mu^{(k)}]}{\sum_{i=1}^{n} Y_i E[\Delta_i | Y_i, p^{(k)}, \lambda^{(k)}, \mu^{(k)}]}, \]

\[ \mu^{(k+1)} = \frac{\sum_{i=1}^{n} E[(1 - \Delta_i) | Y_i, p^{(k)}, \lambda^{(k)}, \mu^{(k)}]}{\sum_{i=1}^{n} Y_i E[(1 - \Delta_i) | Y_i, p^{(k)}, \lambda^{(k)}, \mu^{(k)}]}. \]

The conditional expectation

\[ E[\Delta | Y, \theta] = \frac{p\lambda e^{-\lambda Y}}{p\lambda e^{-\lambda Y} + (1 - p)\mu e^{-\mu Y}}. \]

As seen above, the EM algorithm facilitates the computation.
Information Calculation in EM
• **Notation**

- $\dot{l}_c$ as the score function for $\theta$ in the full data;
- $\dot{l}_{mis|obs}$ as the score for $\theta$ in the conditional distribution of $Y_{mis}$ given $Y_{obs}$;
- $\dot{l}_{obs}$ as the score for $\theta$ in the distribution of $Y_{obs}$.

\[
\dot{l}_c = \dot{l}_{mis|obs} + \dot{l}_{obs}.
\]

\[
Var(\dot{l}_c) = Var(E[\dot{l}_c|Y_{obs}]) + E[Var(\dot{l}_c|Y_{obs})].
\]
• **Information in the EM algorithm**

We obtain the following Louis formula

\[ I_c(\theta) = I_{obs}(\theta) + E[I_{mis|obs}(\theta, Y_{obs})]. \]

Thus, the complete information is the summation of the observed information and the missing information.

One can even show when the EM converges, the convergence linear rate, denote as \((\theta^{(k+1)} - \hat{\theta}_n)/(\theta^{(k)} - \hat{\theta}_n)\) approximates the \(1 - I_{obs}(\hat{\theta}_n)/I_c(\hat{\theta}_n)\).
Nonparametric Maximum Likelihood Estimation
• First example

Let $X_1, \ldots, X_n$ be i.i.d random variables with common distribution $F$, where $F$ is any unknown distribution function. The likelihood function for $F$ is given by

$$L_n(F) = \prod_{i=1}^{n} f(X_i),$$

where $f(X_i)$ is the density function of $F$ with respect to some dominating measure.

However, the maximum of $L_n(F)$ does not exists.

We instead maximize an alternative function

$$\tilde{L}_n(F) = \prod_{i=1}^{n} F\{X_i\},$$
where $F\{X_i\}$ denotes the value $F(X_i) - F(X_i^-)$. 

• Second example

Suppose $X_1, ..., X_n$ are i.i.d $F$ and $Y_1, ..., Y_n$ are i.i.d $G$. We observe i.i.d pairs $(Z_1, \Delta_1), ..., (Z_n, \Delta_n)$, where $Z_i = \min(X_i, Y_i)$ and $\Delta_i = I(X_i \leq Y_i)$. We can think $X_i$ as survival time and $Y_i$ as censoring time. Then it is easy to calculate the joint distributions for $(Z_i, \Delta_i)$, $i = 1, ..., n$, is equal to

$$L_n(F, G) = \prod_{i=1}^{n} \{f(Z_i)(1 - G(Z_i))\}^{\Delta_i} \{(1 - F(Z_i))g(Z_i)\}^{1-\Delta_i}$$

$L_n(F, G)$ does not have the maximum so we consider an alternative function

$$\prod_{i=1}^{n} \{F\{Z_i\}(1 - G(Z_i))\}^{\Delta_i} \{(1 - F(Z_i))G\{Z_i\}\}^{1-\Delta_i}.$$
• Third example

Suppose $T$ is survival time and $Z$ is covariate. Assume $T|Z$ has a conditional hazard function

$$
\lambda(t|Z) = \lambda(t)e^{\theta^T Z}.
$$

Then the likelihood function from $n$ i.i.d $(T_i, Z_i), i = 1, ..., n$ is given by

$$
L_n(\theta, \Lambda) = \prod_{i=1}^{n} \left\{ \lambda(T_i) \exp\{-\Lambda(T_i)e^{\theta^T Z_i}\} f(Z_i) \right\}.
$$

Note $f(Z_i)$ is not informative about $\theta$ and $\lambda$ so we can discard it from the likelihood function. Again, we replace
\[ \lambda \{ T_i \} \] by \( \Lambda \{ T_i \} \) and obtain a modified function

\[ \tilde{L}_n(\theta, \Lambda) = \prod_{i=1}^{n} \left\{ \Lambda \{ T_i \} \exp\{-\Lambda(T_i)e^{\theta^TZ_i}\} \right\}. \]

Let \( p_i = \Lambda \{ T_i \} \) we maximize

\[ \prod_{i=1}^{n} \left\{ p_i \exp\{-\left( \sum_{Y_j \leq Y_i} p_j \right)e^{\theta^TZ_i}\} \right\} \]

or its logarithm as

\[ \sum_{i=1}^{n} \left\{ \theta^T Z_i - \exp\{\theta^T Z_i\} \sum_{Y_j \leq Y_i} p_j + \log p_j \right\}. \]
• **Fourth example**

We consider $X_1, \ldots, X_n$ are i.i.d $F$ and $Y_1, \ldots, Y_n$ are i.i.d $G$. We only observe $(Y_i, \Delta_i)$ where $\Delta_i = I(X_i \leq Y_i)$ for $i = 1, \ldots, n$. This data is one type of interval censored data (or current status data). The likelihood for the observations is

$$
\prod_{i=1}^{n} \left\{ F(Y_i)^{\Delta_i} (1 - F(Y_i))^{1-\Delta_i} g(Y_i) \right\}.
$$

To derive the NPMLE for $F$ and $G$, we instead maximize

$$
\prod_{i=1}^{n} \left\{ P_i^{\Delta_i} (1 - P_i)^{1-\Delta_i} q_i \right\},
$$

subject to the constraint that $\sum q_i = 1$ and $0 \leq P_i \leq 1$ increases with $Y_i$. 
Clearly, $\hat{q}_i = 1/n$ (suppose $Y_i$ are all different). This constrained maximization turns out to be solved by the following steps:

(i) Plot the points $(i, \sum_{Y_j \leq Y_i} \Delta j)$, $i = 1, \ldots, n$. This is called the cumulative sum diagram.

(ii) Form the $H^*(t)$, the greatest the convex minorant of the cumulative sum diagram.

(iii) Let $\hat{P}_i$ be the left derivative of $H^*$ at $i$. Then $(\hat{P}_1, \ldots, \hat{P}_n)$ maximizes the object function.
• **Summary of NPMLE**

  - The NPMLE is a generalization of the maximum likelihood estimation in the parametric model; the semiparametric or nonparametric models.

  - We replace the functional parameter by an empirical function with jumps only at observed data and maximize a modified likelihood function.

  - Both computation of the NPMLE and the asymptotic property of the NPMLE can be difficult and vary for different specific problems.
CHAPTER 5 MAXIMUM LIKELIHOOD ESTIMATION

Alternative Efficient Estimation
• **One-step efficient estimation**
  
  – start from a strongly consistent estimator for parameter $\theta$, denoted by $\tilde{\theta}_n$, assuming that $|\tilde{\theta}_n - \theta_0| = O_p(n^{-1/2})$.
  
  – One-step procedure is a one-step Newton-Raphson iteration in solving the likelihood score equation;
    
    $$\hat{\theta}_n = \tilde{\theta}_n - \left\{ \tilde{l}_n(\tilde{\theta}_n) \right\}^{-1} \tilde{l}_n(\tilde{\theta}_n),$$
    
    where $\tilde{l}_n(\theta)$ is the sore function and $\tilde{l}_n(\theta)$ is the derivative of $\tilde{l}_n(\theta)$. 
• Result about the one-step estimation

**Theorem 5.6** Let $l_\theta(X)$ be the log-likelihood function of $\theta$. Assume that there exists a neighborhood of $\theta_0$ such that in this neighborhood, $|l_\theta^{(3)}(X)| \leq F(X)$ with $E[F(X)] < \infty$. Then

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \rightarrow_d N(0, I(\theta_0)^{-1}),$$

where $I(\theta_0)$ is the Fisher information.
**Proof** Since $\tilde{\theta}_n \rightarrow_{a.s.} \theta_0$, we perform the Taylor expansion on the right-hand side of the one-step equation and obtain

$$
\hat{\theta}_n = \tilde{\theta}_n - \left\{ \frac{1}{n} \dot{\theta}_n(\tilde{\theta}_n) \right\} \left\{ \frac{1}{n} \ddot{\theta}_n(\theta_0) + \ddot{\theta}_n(\theta^*)(\tilde{\theta}_n - \theta_0) \right\}
$$

where $\theta^*$ is between $\tilde{\theta}_n$ and $\theta_0$. \Rightarrow

$$
\hat{\theta}_n - \theta_0 = \left[ I - \left\{ \frac{1}{n} \dot{\theta}_n(\tilde{\theta}_n) \right\}^{-1} \dot{\theta}_n(\theta^*) \right] (\tilde{\theta}_n - \theta_0) - \left\{ \frac{1}{n} \dot{\theta}_n(\tilde{\theta}_n) \right\} \dot{\theta}_n(\theta_0).
$$

On the other hand, by the condition that $|l_\theta^{(3)}(X)| \leq F(X)$ with $E[F(X)] < \infty$,

$$
\frac{1}{n} \dot{\theta}_n(\theta^*) \rightarrow_{a.s.} E[\dot{\theta}_0(X)], \quad \frac{1}{n} \dot{\theta}_n(\tilde{\theta}_n) \rightarrow_{a.s.} E[\dot{\theta}_0(X)].
$$

\Rightarrow

$$
\hat{\theta}_n - \theta_0 = o_p(|\tilde{\theta}_n - \theta_0|) - \left\{ E[\dot{\theta}_0(X)] + o_p(1) \right\}^{-1} \frac{1}{n} \dot{\theta}_n(\theta_0).
$$
• Slightly different one-step estimation

\[ \hat{\theta}_n = \tilde{\theta}_n + I(\tilde{\theta}_n)^{-1} i(\tilde{\theta}_n). \]

• Other efficient estimation

the Bayesian estimation method (posterior mode, minimax estimator etc.)
• **Conclusions**

  – The maximum likelihood approach provides a natural and simple way of deriving an efficient estimator.

  – Other estimation approaches are possible for efficient estimation such as one-step estimation, Bayesian estimation etc.

  – Generalization from parametric models to semiparametric or nonparametric models. How?
READING MATERIALS: Ferguson, Sections 16-20, Lehmann and Casella, Sections 6.2-6.7