Methods of Estimation II

MIT 18.655

Dr. Kempthorne

Spring 2016

Outline

1. Methods of Estimation II
    - Maximum Likelihood in Multiparameter Exponential Families
    - Algorithmic Issues

Maximum Likelihood in Exponential Families

Issues:

- Existence of MLEs
- Uniqueness of MLEs

Significant Feature of Exponential Family of Distributions

- Concavity of the log likelihood
$$l_x(\eta) = log[p(x \mid \eta)],$$
for all $x \in \mathcal{X}$, where η is the *natural* parameter in the canonical representation.

Existence and Uniqueness Theorem

Proposition 2.3.1 Suppose $X \sim P \in \{P_\theta, \theta \in \Theta\}$ with

- $\Theta \subset R^p$, an open set.
- The corresponding densitites of P_θ, $p(x \mid \theta)$, are such that for any $x \in \mathcal{X}$ the likelihood function
  $l_x(\theta) = \log[p(x \mid \theta)]$ is strictly concave in θ
- $l_x(\theta) \to -\infty$ as $\theta \to \partial\Theta$, where
  $\partial\Theta = \bar{\Theta} - \Theta$, the boundary of Θ, defined using $\bar{\Theta}$, the closure of Θ in $[-\infty, \infty]$.

Then:

- The MLE $\hat{\theta}(x)$ exists.
- The MLE $\hat{\theta}(x)$ is unique.

Proof:

- Apply properties of convexity of sets/functions.

Convexity

Definitions (Section B.9)

- A subset $S \subset R^k$ is **convex** if for every $x, y \in S$,
$$\alpha x + (1-\alpha)y \in S, \text{ for all } \alpha : 0 \leq \alpha < 1.$$
  - for $k = 1$, convex sets are intervals (finite or infinite).
  - for $k > 1$, spheres, rectangles (finite or infinite) are convex.
- $\mathbf{x}_0 \in S^0$, the interior of the convex set S if and only if
$$\{x : \mathbf{d}^T\mathbf{x} > \mathbf{d}^T\mathbf{x}_0\} \cap S^0 \neq \emptyset$$
and
$$\{x : \mathbf{d}^T\mathbf{x} < \mathbf{d}^T\mathbf{x}_0\} \cap S^0 \neq \emptyset$$
for every $\mathbf{d} \neq \mathbf{0}$.
- A function $g : S \to R$ is **convex** if
$$g(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) \leq \alpha g(\mathbf{x}) + (1-\alpha)g(\mathbf{y})$$
for all $\mathbf{x}, \mathbf{y} \in S$, and all $\alpha : 0 \leq \alpha \leq 1$.
- A function $g : S \to R$ is **strictly convex** if
$$g(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) < \alpha g(\mathbf{x}) + (1-\alpha)g(\mathbf{y})$$
for all $\mathbf{x} \neq \mathbf{y} \in S$, and all $\alpha : 0 < \alpha < 1$.

Convexity

Properties (Section B.9)

- A convex function is continuous on S^0
- For $k = 1$, if g'' exists:
  - $g''(x) \geq 0$, $x \in S \iff g(\cdot)$ is convex.
  - $g''(x) > 0$, $x \in S \iff g(\cdot)$ is strictly convex.
- For $g(\cdot) : S \to R$ convex and fixed $\mathbf{x}, \mathbf{y} \in S$,

  $h(\alpha) = g(\alpha\mathbf{x} + (1-\alpha)\mathbf{y})$ is convex in α, for $0 \leq \alpha \leq 1$.
- When $k > 1$, if $\dfrac{\partial g^2(x)}{\partial x_i \partial x_j}$ exists, convexity is equivalent to

  $$\sum_{i,j} u_i u_j \frac{\partial g^2(x)}{\partial x_i \partial x_j} \geq 0,$$

  for all $\mathbf{u} = (u_1, \ldots, u_k)^T \in R^k$, and $x \in S$.
- A function $h : S \to R$ is **(strictly) concave** if

  $g = -h$ is (strictly) convex.

Convexity

Jensen's Inequality If

- $S \subset R^k$ is convex and closed
- g is convex on S.
- U a random vector with sample space $\mathcal{U} = S$,
$$P[U \in S] = 1 \text{ and } E[U] \text{ finite}$$

Then

- $E[U] \in S$
- $E[g(U)]$ exists
- $E[g(U)] \geq g(E[U])$
- $E[g(U)] = g(E[U])$ if and only if
$$P(g(U) = a + b^T U) = 1.$$
for some fixed $a \in R$ and $\mathbf{b}(k \times 1) \in R^k$.
- If g is strictly convex, then
$$E[g(U)] = g(E[U]) \text{ if and only if } P(U = \mathbf{c}) = 1,$$
for some $\mathbf{c} \in R^k$.

Existence and Uniqueness of MLE

Proof of Proposition 2.3.1

- Because $l_x(\theta) : \Theta \to R$ is strictly concave, it follows that it is continuous on Θ.
- Because $l_x(\theta) \to -\infty$ as $\theta \to \partial\Theta$, the mle $\hat{\theta}(x)$ exists.
  This follows from
  Lemma 2.3.1:
  - Suppose the function $l : \Theta \to R$ where $\Theta \subset R^p$ is open and l is continuous.
  - If $\lim\{l(\theta) : \theta \to \partial\Theta\} = -\infty$, then
    there exists $\hat{\theta} \in \Theta$ such that: $l(\hat{\theta}) = max\{l(\theta) : \theta \in \Theta\}$
- Suppose $\hat{\theta}_1$ and $\hat{\theta}_2$ are distinct MLEs: $l_x(\hat{\theta}_1) = l_x(\hat{\theta}_2)$ and $\hat{\theta}_1 \neq \hat{\theta}_2$. By the strict concavity of l_x,
  $$l_x(\tfrac{1}{2}\hat{\theta}_1 + \tfrac{1}{2}\hat{\theta}_2) > \tfrac{1}{2}l_x(\hat{\theta}_1) + \tfrac{1}{2}l_x(\hat{\theta}_2). > l_x(\hat{\theta}_1)$$
  but this contradicts $\hat{\theta}_1$ being an MLE.

MLEs for Canonical Exponential Family

Theorem 2.3.1 Suppose $\mathcal{P}$ is the canonical exponential family generated by (T, h), and that

- The natural parameter space $\mathcal{E}$ is open
- The family is of rank k.

(a). If $t_0 \in R^k$ satisfies:

$$P[c^T T(X) > c^T t_0] > 0 \text{ for all } c \neq 0, \quad (*)$$

then the MLE $\hat{\eta}$ exists, is unique,
and is a solution to the equation

$$\dot{A}(\eta) = E(T(X) \mid \eta) = t_0. \quad (**)$$

(b). If $t_0 \in R^k$ does not satisfy $(*)$, then the MLE does not exist and $(**)$ has no solution.

Recall canonical exponential family generated by (T, h):

- *Natural Sufficient Statistic:* $\mathbf{T}(\mathbf{X}) = (T_1(X), \dots, T_k(X))^T$
- *Natural Parameter:* $\boldsymbol{\eta} = (\eta_1, \dots, \eta_k)^T$
- Density function
  $$p(x \mid \boldsymbol{\eta}) = h(x) exp\{\mathbf{T}^T(x)\boldsymbol{\eta}) - A(\boldsymbol{\eta})\}$$
  where $A(\cdot)$ is defined to normalize the density:
  $$A(\boldsymbol{\eta}) = log \int \cdots \int h(x) exp\{\mathbf{T}^T(x)\boldsymbol{\eta}\} dx$$
  or
  $$A(\boldsymbol{\eta}) = log[\ \underset{x\in\mathcal{X}}{\qquad}\ h(x) exp\{\mathbf{T}^T(x)\boldsymbol{\eta}\}]$$
- *Natural Parameter space:* $\boldsymbol{\mathcal{E}} = \{\boldsymbol{\eta} \in R^k : -\infty < A(\boldsymbol{\eta}) < \infty\}$.

Proof.

- We can suppose that $h(x) = p(x \mid \eta_0)$ for some reference $\eta_0 \in \mathcal{E}$.
    - The canonical family generated by $(T(x), h(x))$ with natural parameter η and normalization term $A(\eta)$, is identical to the family generated by $(T(x), h_0(x))$ with $h_0(x) = p(x \mid \eta_0)$ and natural parameter η^* and normalization term $A^*(\eta^*)$.
    - $\eta^* = \eta - \eta_0$
    - $A^*(\eta^*) = A(\eta^* + \eta_0) - A(\eta_0)$
      (Problem 1.6.27)
- We can also assume that $t_0 = T(x) = 0$. (N.B. x is fixed)
    - The class $\mathcal{P}$ is the same exponential family generated by
      $$T^*(X) = T(X) - t_0.$$
- The likelihood function for x is
  $$l_x(\eta) = log[p(x \mid \eta)] = -A(\eta) + \log[h(x)]$$
  since $T(x) = 0$.

Proof (continued)

Claim: If $\{\eta_m\}$ has no subsequence converging to a point in $\mathcal{E}$, then for any convergent subsequence $\{\eta_{m_k}\}$:

$$\lim_{k\to\infty} l_x(\eta_{m_k}) = -\infty.$$

- Any sub-sequence that has a limit is on the boundary of $\mathcal{E}$, outside $\mathcal{E}$.
- The existence of the MLE $\hat{\eta}(x)$ is guaranteed by Lemma 2.3.1.

Proof of Claim: Let $\{\eta_m\}$ be a sequence with no subsequence converging to a point in $\mathcal{E}$ and let $\{\eta_{m_k}\}$ be convergent.
Express the η_m in terms of scalars λ_m and unit k-vectors $u_m \in R^k$:

$$\eta_m = \lambda_m u_m,$$

where $u_m = \eta_m/|\eta_m|$ and $\lambda_m = |\eta_m|$

Two cases to consider:

Case 1: $\lambda_{m_k} \to \infty$, and $u_{m_k} \to u$ $\quad (|\eta_{m_k}| \to \infty)$

Case 2: $\lambda_{m_k} \to \lambda$, and $u_{m_k} \to u$ $\quad (\eta_{m_k} \to \lambda\mu \notin \mathcal{E})$

Proof (continued)

Case 1: $\lambda_{m_k} \to \infty$, and $u_{m_k} \to u$. Writing E_0 for $E[\cdot \mid \eta_0]$, and P_0 for P_{η_0}, then for some $\delta > 0$:

$$\begin{aligned} \lim_{k\to\infty} \int e^{\eta_{m_k}^T T(x)} h(x)dx &= \lim_{k\to\infty} E_0[e^{\lambda_{m_k} u_{m_k}^T T(x)}] \\ &\geq \lim_{k\to\infty} E_0[e^{\lambda_{m_k} u_{m_k}^T T(x)} \times \mathbf{1}(\{u_{m_k}^T T(X) > \delta\})] \\ &\geq \lim_{k\to\infty} e^{\lambda_{m_k}\delta} E_0[\mathbf{1}(\{u_{m_k}^T T(X) > \delta\})] \\ &= \lim_{k\to\infty} e^{\lambda_{m_k}\delta} P_0[\{u_{m_k}^T T(X) > \delta\}] \\ &= \lim_{k\to\infty} e^{\lambda_{m_k}\delta} P_0[\{u^T T(X) > \delta\}] \\ &= +\infty \end{aligned}$$

The first inequality follows because under condition **(a)** of the theorem, we are given that $t_0 \in R^k$ satisfies:

$$P[c^T T(X) > c^T t_0] > 0 \text{ for all } c \neq 0, \quad (*)$$

So, with $t_0 = 0$, and $c = u$ $(\neq 0)$, it must be that for some $\delta > 0$,

$$P_0(u^T T(X) > \delta) > 0.$$

$$A(\eta_{m_k}) = \log[\int e^{\eta_{m_k}^T T(x)} h(x)dx] \to \infty \implies l_x(\eta_{m_k}) \to -\infty$$

Proof (continued)

Case 2: $\lambda_{m_k} \to \lambda$, and $u_{m_k} \to u$, with $\eta^* = \lambda\mu \notin \mathcal{E}$.

$$\begin{aligned} \lim_{k\to\infty} \int e^{\eta_{m_k}^T T(x)} h(x)dx &= \lim_{k\to\infty} E_0[e^{\lambda_{m_k} u_{m_k}^T T(x)}] \\ &= E_0[e^{\lambda u^T T(X)}] = \log A(\eta^*), \end{aligned}$$

But $A(\eta^*) = +\infty$ since $\eta^* \notin \mathcal{E} = \{\eta : A(\eta) < \infty\}$. So

$$\begin{aligned} A(\eta_{m_k}) = \log[\textstyle\int e^{\eta_{m_k}^T T(x)} h(x)dx] &\to \infty \\ \Longrightarrow \quad l_x(\eta_{m_k}) &\to -\infty \end{aligned}$$

We can conclude:

- Under both Cases 1 and 2, $\lim_k l_x(\eta_{m_k}) \to -\infty$ so it must be that $l_x(\eta_n) \to -\infty$. By Lemma 2.3.1 it must be that $\hat{\eta}(x)$ exists.
- By Theorem 1.6.4, the mle $\hat{\eta}(x)$ is unique and satisfies:
$$\dot{A}(\eta) = E(T(X) \mid \eta) = t_0. \quad (**)$$

Proof (continued)

Nonexistence:

(b). Suppose no $t_0 \in R^k$ satisfies:

$$P[c^T T(X) > c^T t_0] > 0 \text{ for all } c \neq 0. \quad (*)$$

Then, with $t_0 = 0$, there exists a $c \neq 0$ such that

$$P[c^T T(X) > 0] = 0$$

equivalently

$$P_0[c^T T(X) \leq 0] = 1.$$

It follows that:

$$E_\eta[c^T T(X)] \leq 0 \text{ for all } \eta.$$

If $\hat{\eta}$ exists, then it solves $E_\eta(T(X)) = t_0 = 0$ which means there is an η such that

$E_\eta(c^T T(X)) = 0$. But for this η, it would have to be that
$P_\eta(c^T T(X) = 0) = 1$.

and this contradicts the assumption that the family is of rank k.

Corollary 2.3.1 Under the conditions of Theorem 2.3.1, if

C_T is the convex support of the distribution of $T(X)$.

then $\hat{\eta}(x)$ exists and is unique if and only if

$t_0 = T(x) \in C_T^0$, the interior of C_T.

Proof: A point t_0 is in the interior of C_T if and only if there exist points in C_T^0 on either side of it; that is, for all $d \neq 0$:

$$\{t : d^T t > d^T t_0\} \cap C_T^0 \neq \emptyset$$

and

$$\{t : d^T t < d^T t_0\} \cap C_T^0 \neq \emptyset$$

and that the two sets are open.

It follows that condition (a) of Theorem 2.3.1 is satisfied:

$$P[c^T T(X) > c^T t_0] > 0 \text{ for all } c \neq 0.$$

Example 2.3.1 The Gaussian Model.

- $X_1, \ldots, X_n$ iid $N(\mu, \sigma^2)$, with $\mu \in R$, and $\sigma^2 > 0$
- $T(X) = (\sum_1^n X_i, \sum_1^n X_i^2)$ is the natural sufficieint statistic.
- $C_T = R \times R^+$.
- The density of $T(X)$ can be derived for $n = 1, 2, \ldots$
- For $n \geq 2$, $C_T = C_T^0$ and the mle of the natural parameter η exists (and thus of $\theta = (\mu, \sigma^2)$.
- For $n = 1$, $T(X)$ is a parabola in x_1 and $T(x)$ is a point. So $C_T^0 = \emptyset$ and the MLE does not exist.
  ($\hat{\mu} = X_1$ and the likelihood becomes unbounded as $\hat{\sigma} \to 0^+$.)

Theorem 2.3.2 Suppose the conditions of Theorem 2.3.1 hold and T $(k \times 1)$ has a continuous case density on R^k. Then the MLE $\hat{\eta}$ exists with probability 1 and necessarily satisfies (2.3.3)

$$\dot{A}(\eta) = E(T(X) \mid \eta) = t_0. \quad (**)$$

Proof. The boundary of a convex set necessarily has volume 0. If T has continuous density $P_T(t)$, then

$$P(T \in \partial C_T) = \int_{\partial C_T} p_T(t)dt = 0.$$

By Corollary 2.3.1, $T(X)$ is in the interior of C_T with probability 1 and in that case, the MLE exists and is unique.

Notes:

- Generalized method-of-moments principle. For exponential families, the MLE solves
  $$E_\eta[T(X)] = t_0, \text{ for } \eta \text{ given } T(x) = t_0,$$
  which matches moments because:
  $$E_\eta[T(X)] = \dot{A}(\eta).$$
- MLEs are generally best; the better method-of-moments estimators are often those that are equivalent to MLEs.

Example 2.3.2 Two-Parameter Gamma Family.

$X_1, \ldots, X_n$ are iid $Gamma(p, \lambda)$ random variables:

$$p(x \mid p, \lambda) = \frac{\lambda^p x^{p-1} e^{-\lambda x}}{\Gamma(p)}$$

where $x > 0,, \ p > 0,, \ \lambda > 0$.

- Natural Sufficient Statistic: $T = (\sum_1^n \log X_i, \sum_1^n X_i)$
- Natural Parameters: $\eta = (p, -\lambda)$
- $A(\eta_1, \eta_2) = n(\log [\Gamma(\eta_1) - \eta_1 log(-\eta_2)]$
- The likelihood equations:

$$\frac{\Gamma'}{\Gamma}(\hat{p}) - \log \hat{\lambda} = \overline{\log (X)}$$

$$\frac{\hat{p}}{\hat{\lambda}} = \overline{X}$$

where $\overline{\log(X)} = \sum_1^n \log X_i / n$.

To apply the theorems we need to demonstrate that the distribution of T has a continuous density.

Example 2.3.3 Multinomial Trials. Recall:

$$\begin{aligned}
p(x \mid \theta) &= \tfrac{n}{x_1!\cdots x_q!}\theta_1^{x_1}\theta_2^{x_2}\cdots\theta_q^{x_q}, \quad x_i \geq 0,\ \textstyle\sum_1^q x_i = n \\
&= \tfrac{n}{x_1!\cdots x_q!} \times exp\{log(\theta_1)x_1 + \cdots + log(\theta_{q-1})x_{q-1} \\
&\qquad\qquad + log(1 - \textstyle\sum_1^{q-1}\theta_j)[n - \textstyle\sum_1^{q-1} x_j]\} \\
&= h(x) exp\{\textstyle\sum_{j=1}^{q-1}\eta_j(\theta)T_j(x) - B(\theta)\} \\
&= h(x) exp\{\textstyle\sum_{j=1}^{q-1}\eta_j T_j(x) - A(\eta)\}
\end{aligned}$$

where:

- $h(x) = \frac{n}{x_1!\cdots x_q!}$
- $\eta(\theta) = (\eta_1(\theta), \eta_2(\theta), \ldots, \eta_{q-1}(\theta))$
  $$\eta_j(\theta) = log(\theta_j/(1 - \textstyle\sum_1^{q-1}\theta_j)),\ j = 1, \ldots, q-1$$
- $T(x) = (X_1, X_2, \ldots, X_{q-1}) = (T_1(x), T_2(x), \ldots, T_{q-1}(x))$.
- $B(\theta) = -nlog(1 - \sum_{j=1}^{q-1}\theta_j)$ and $A(\boldsymbol{\eta}) = +nlog(1 + \sum_{j=1}^{q-1} e^{\eta_j})$

$$\dot{A}(\boldsymbol{\eta})_j = n\frac{e^{\eta_j}}{1+\sum_{j=1}^{q-1} e^{\eta_j}} = n\frac{\theta_j/(1-\sum_1^{q-1}\theta_k)}{1+\sum_1^{q-1}\theta_k/(1-\sum_1^{q-1}\theta_k)} = n\theta_j$$

$$\ddot{A}(\boldsymbol{\eta})_{i,j} = -n\theta_i\theta_j,\ (i \neq j) \text{ and } \ddot{A}(\boldsymbol{\eta})_{i,i} = n\theta_i(1-\theta_i),$$

Multinomial Example (continued)

Note: MLE for θ exists only if $X_i > 0$ for all $i = 1, \ldots, q$

Argument:

- The condition of Theorem 2.3.1 (2.3.2) for existence of MLE is
$$P[c^T T(X) > c^T t_0] > 0, \text{ for all } c \neq 0.$$
- For any given c, decompose:
$$c^T t_0 = \sum_{c_i > 0} c_i [t_0]_i + \sum_{c_j < 0} c_j [t_0]_j$$
- To have positive probability that $c^T T(X)$ is larger than $c^T t_0$, we need to have:
$$T(x)_i < n \text{ for } i : c_i > 0$$
and
$$T(x)_i > 0 \text{ for } j : c_j < 0$$
- Varying c leads to the condition that $0 < X_i < n$ for all i.

Corollary 2.3.2 Consider the exponential family:

$$p(x \mid \theta) = h(x) exp\{\sum_{j=1}^{k} c_j(\theta) T_j(x) - B(\theta)\},\ x \in \mathcal{X},\ \theta \in \Theta.$$

- Let C^0 be the interior of the range of $(c_1(\theta), \ldots, c_k(\theta))^T$
- Let x be the observed data.

If the equations

$$E_\theta T_j(X) = T_j(x),\ i = 1, \ldots, k$$

have a solution

$$\hat{\theta}(x) \in C^0,$$

then $\hat{\theta}(x)$ is the unique MLE of θ.

Outline

1. Methods of Estimation II
   - Maximum Likelihood in Multiparameter Exponential Families
   - Algorithmic Issues

Algorithmic Issues

Bisection Method: Root Solution to Equation

Consider the problem of solving: $f(x) = 0$ for x.

- Function $f(\cdot)$: continuous for $x \in (a, b)$
- $f(a^+) < 0$ and $f(b^-) > 0$
- Intermediate value theorem of calculus:
$$\exists x^* \in (a, b) : f(x^*) = 0.$$
- If $f(\cdot)$ is strictly increasing then x^* is unique.

Bisection Algorithm

1. Find $x_0 < x_1 : f(x_0) < 0 < f(x_1)$.
2. Evaluate $f(x_*)$ for $x_* = (x_0 + x_1)/2$.
3. If $f(x_*) < 0$, replace x_0 with x_* or
   if $f(x_*) > 0$, replace x_1 with x_*
4. Go back to step 2 until $|x_1 - x_0| < \epsilon$ for some fixed $\epsilon > 0$
5. Return x_* as the approximate solution $(|x_* - x^*| < \epsilon)$

Theorem 2.4.1

- $p(x \mid \eta)$ is the density/pmf function of a one-parameter canonical exponential family generated by $(T(X), h(x))$
- The conditions of Theorem 2.3.1 are satisfied:
  - Natural parameter space $\mathcal{E}$ is open
  - Family is of rank k
- $T(x) = t_0 \in C_T^0$, the interior of convex support for $p(t \mid \eta)$, the density/pmf of $T(X)$.

The unique MLE $\hat{\eta}$ (by Theorem 2.3.1) may be approximated by the bisection method applied to

$$f(\eta) = E[T(X) \mid \eta] - t_0.$$

Proof

- $f(\eta)$ is strictly increasing because $f'(\eta) = Var[T(X) \mid \eta] > 0$.
- $f(\eta)$ is continuous .
- The existence of the MLE $\hat{\eta}$ implies that with $\mathcal{E} = (a, b)$, it must be that

$$f(a^+) < 0 < f(b^-).$$

Other Algorithms

- Coordinate Ascent
  - Line search: coordinate by coordinate
- Newton-Raphson Algorithm
  - Iterative solution of quadratic approximations of $f(\eta)$.
- Expectation-Maximization (EM) Algorithm
  - Problems where likelihood function easily maximized if observed variables extended to include additional variables (missing data/latent variables).
  - Iterative solution alternates:
    E-Step: estimating unobserved variables given a preliminary estimate $\hat{\eta}_j$
    M-Step: maximizing the full-data likelihood to obtain an updated estimate $\hat{\eta}_{j+1}$

EM Algorithm

Preliminaries

- Complete Data: $X \sim P_\theta$, with density $p(x \mid \theta), \theta \in \Theta \subset R^d$.
- Log likelihood: $l_{p,x}(\theta)$ easy to maximize.
  Suppose the distribution is a member of the canonical exponential family with
  - Natural parameter $\eta(\theta)$
  - Natural sufficient statistic: $T(X) = (T_1(X), \dots, T_k(X))$
  - $E[T(X) \mid \eta] = \dot{A}(\eta)$
  - Given $T(x) = t_0$, the mle for η is the solution to:
    $$\dot{A}(\eta) = E(T(X) \mid \eta) = t_0. \quad (**)$$
- Incomplete Data / Observed Data:
  $$S = S(X) \sim Q_\theta \text{ with density } q(s \mid \theta).$$
- Log likelihood: $l_{q,s}(\theta)$ is hard to maximize.

EM Algorithm

Example 2.4.5 Mixture of Gaussians. Let $S_1, \ldots, S_n$ be iid P with density

$$p(s \mid \theta) = \lambda \phi_{\sigma_1}(s - \mu_1) + (1 - \lambda)\phi_{\sigma_2}(s - \mu_2)$$

where

- $\lambda : 0 \leq \lambda \leq 1$.
- $\phi_\sigma(\cdot)$ is the density of a Gaussian distribution with mean zero and variance σ^2, i.e., $\phi_\sigma(s) = \frac{1}{\sigma}\phi(s/\sigma))$ where $\phi(\cdot)$ is the density of a standard Gaussian distribution (mean 0 and variance 1).
- $\theta = (\lambda, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$

The $\{S_i\}$ are a sample from a Gaussian-mixture distribution which is $N(\mu_1, \sigma_1^2)$ with probability λ and is $N(\mu_2, \sigma_2^2)$ with probability $(1 - \lambda)$.

EM Algorithm: Gaussian Mixture

Consider adding to $\{S_i\}$ the variables $(\Delta_1, \ldots, \Delta_n)$ indicating whether or not case i came from the first Gaussian distribution $(\Delta_i = 1)$ or the second $(\Delta_i = 0)$. The complete data are thus

$$\{X_i = (\Delta_i, S_i), i = 1, \ldots, n\}$$

and

- Δ_i are iid *Bernoulli*(λ), i.e., $P(\Delta_i = 1) = \lambda = 1 - P(\Delta_i = 0)$.
- Given Δ_i, the density of S_i is
  $$p(s \mid \Delta_i, \theta) = \phi_{\sigma_*}(s - \mu_*)$$
  where
  $$\begin{aligned} \mu_* &= \Delta_i \mu_1 + (1 - \Delta_i)\mu_2, \quad \text{and} \\ \sigma_*^2 &= \Delta_i \sigma_1^2 + (1 - \Delta_i)\sigma_2^2. \end{aligned}$$

Consider inference about $\theta = (\lambda, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$ observing

$$S(\mathbf{X}) = (S_1, \ldots, S_n)$$

rather than

$$\mathbf{X} = (X_1, \ldots, X_n) = ((\Delta_1, S_1), \ldots, (\Delta_n, S_n))$$

EM Algorithm: Theoretical Basis

For complete data X and incomplete data $S(X)$, the complete-data density $p(x \mid \theta)$ satisfies

$$p(x \mid \theta) = q(s \mid \theta) r(x \mid s, \theta)$$

where

- $q(s \mid \theta)$ is the density of $S(X) = s$ given θ, and
- $r(x \mid s, \theta)$ is the density of the conditional distribution of X given $S(x) = s$, and θ.

Claim 1: The likelihood ratio of θ to θ_0 based on $S(X)$ is the conditional expectation of the likelihood ratio based on X given $S(X) = s$ and θ_0.

$$\frac{q(s \mid \theta)}{q(s \mid \theta_0)} = E\left[\frac{p(x \mid \theta)}{p(x \mid \theta_0)} | S(X) = s, \theta_0\right]$$

EM Algorithm: Theoretical Basis

Proof of Claim 1:

$$
\begin{aligned}
E\left[\frac{p(x \mid \theta)}{p(x \mid \theta_0)} | S(X) = s, \theta_0\right] &= E\left[\frac{q(s \mid \theta) r(x \mid s, \theta)}{q(s \mid \theta_0) r(x \mid s, \theta_0)} | S(X) = s, \theta_0\right] \\
&= \frac{q(s \mid \theta)}{q(s \mid \theta_0)} \cdot E\left[\frac{r(x \mid s, \theta)}{r(x \mid s, \theta_0)} | S(X) = s, \theta_0\right] \\
&= \frac{q(s \mid \theta)}{q(s \mid \theta_0)} \cdot \sum_{\{x : S(x) = s\}} \left[\frac{r(x \mid s, \theta)}{r(x \mid s, \theta_0)}\right] r(x \mid s, \theta_0) \\
&= \frac{q(s \mid \theta)}{q(s \mid \theta_0)} \cdot \sum_{\{x : S(x) = s\}} [r(x \mid s, \theta)] \\
&= \frac{q(s \mid \theta)}{q(s \mid \theta_0)}.
\end{aligned}
$$

EM Algorithm: Theoretical Basis

Claim 2: Suppose $\theta = \theta_0$ is not the MLE $\hat{\theta}(S)$ for $S(X) = s$. As a function of θ, the likelihood ratio based on S at θ versus θ_0

$$\frac{q(s \mid \theta)}{q(s \mid \theta_0)}$$

will increase (above 1) for θ^* maximizing:

$$J(\theta \mid \theta_0) = E\left[\log\left(\frac{p(x|\theta)}{p(x|\theta_0)}\right) \mid S(X) = s, \theta_0\right] \quad (***)$$

Proof: Substitute $p(x \mid \theta) = q(s \mid \theta) r(x \mid S(X) = s, \theta)$ in $(***)$ to give

$$J(\theta \mid \theta_0) = \log \frac{q(s \mid \theta)}{q(s \mid \theta_0)} + E\left[\log \frac{r(X \mid s, \theta)}{r(X \mid s, \theta_0)} \mid S(X) = s, \theta_0\right]$$

By Jensen's inequality, since *log*() is a concave function:

$$\begin{aligned} E\left[\log \frac{r(X \mid s, \theta)}{r(X \mid s, \theta_0)} \mid S(X) = s, \theta_0\right] &\leq \log\left(E\left[\frac{r(X \mid s, \theta)}{r(X \mid s, \theta_0)} \mid S(X) = s, \theta_0\right]\right) \\ &\leq \log(1) = 0 \end{aligned}$$

It follows that: $\log \dfrac{q(s \mid \theta^*)}{q(s \mid \theta_0)} \geq J(\theta^* \mid \theta_0) > 0$, since $J(\theta_0 \mid \theta_0) = 0$.

EM Algorithm: Theoretical Basis

Claim 3: Under suitable regularity conditions,

- $\frac{\partial}{\partial\theta}\log q(s \mid \theta)$, the gradient of the log likelihood for the incomplete data S, and
- $\frac{\partial}{\partial\theta}J(\theta \mid \theta_0)$, the gradient of the conditional expectation of the complete-data log likelihood ratio given θ_0

are identical when evaluated at $\theta = \theta_0$.

Proof: From Claim 1:

$$
\begin{aligned}
\frac{q(s \mid \theta)}{q(s \mid \theta_0)} &= E\left[\frac{p(x|\theta)}{p(x|\theta_0)}|S(X) = s, \theta_0\right] \\
\Longrightarrow \frac{\partial}{\partial\theta}[\frac{q(s \mid \theta)}{q(s \mid \theta_0)}] &= \frac{\partial}{\partial\theta}\left(E\left[\frac{p(x|\theta)}{p(x|\theta_0)}|S(X) = s, \theta_0\right]\right) \\
\Longrightarrow \frac{\partial}{\partial\theta}[\log q(s \mid \theta)]|_{\theta=\theta_0} &= E\left[\frac{\partial}{\partial\theta}\left(\frac{p(x|\theta)}{p(x|\theta_0)}\right)|S(X) = s, \theta_0\right] \\
&= E\left[\frac{\partial}{\partial\theta}[\log\left(p(x \mid \theta)\right)]|S(X) = s, \theta_0\right]|_{\theta=\theta_0} \\
&= \frac{\partial}{\partial\theta}\left(E\left[\log\left(p(x \mid \theta)\right)\right]|S(X) = s, \theta_0]\right)|_{\theta=\theta_0} \\
&= \frac{\partial}{\partial\theta}J(\theta \mid \theta_0)|_{\theta=\theta_0}
\end{aligned}
$$

EM Algorithm: Practical Implementation

Theorem 2.4.3. Suppose $\{P_\theta, \theta \in \Theta\}$ is a canonical exponential family generated by (T, h) satisfying (conditions of Theorem 2.3.1):

- The natural parameter space $\mathcal{E}$ is open
- The family is of rank k.
- For complete data X, if $T(X) = t_0 \in R^k$, and
$$P[c^T T(X) > c^T t_0] > 0, \text{ for all } c = 0.$$
and the MLE $\hat{\eta}$ exists, is unique and the solution to the equation:
$$\dot{A}(\eta) = E[T(X) \mid \eta] = t_0.$$

Let $S(X)$ be any statistic (incomplete-data version of X), then the EM Algorithm given $S(X) = s$ consists of:

1. Initialize $\eta = \eta_0$
2. Solve $\dot{A}(\eta) = E[T(X) \mid \eta_0, S(X) = s]$ for η^*
3. Replace η_0 with η^*, and return to step 2.

EM Algorithm: Theorem 2.4.3

Theorem 2.4.3 (continued). If

- The sequence $\{\hat{\eta}_n\}$ obtained from the EM algorithm is bounded.
- The equation $\dot{A}(\eta) = E[T(X) \mid \eta S(X) = s]$ has a unique solution

Then the limit of $\hat{\eta}_n$ exists and is a local maximum of $q(s, \theta)$.

Proof:

$$\begin{aligned} J(\eta \mid \eta_0) &= E\left[(\eta - \eta_0)^T T(X) - [A(\eta) - A(\eta_0)] \mid S(X) = s, \eta_0\right] \\ &= (\eta - \eta_0)^T E\left[T(X) \mid S(X) = s, \eta_0\right] - [A(\eta) - A(\eta_0)] \end{aligned}$$

So, $\frac{\partial}{\partial \eta}[J(\eta \mid \eta_0)] = 0$ yields the equation:

$$E\left[T(X) \mid S(X) = s, \eta_0\right] = \dot{A}(\eta)$$

EM Algorithm: Gaussian Mixture

For the Gaussian Mixture (Example 2.4.5) derive the EM Algorithm.
The complete-data likelihood of $X_i = (\Delta_i, S_i)$ for $\theta = (\lambda, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$ is:

$$\begin{aligned}
p(\Delta_i, S_i \mid \theta) &= p(\Delta_i \mid \theta) p(S_i \mid \theta, \Delta_i) \\
&= \lambda^{\Delta_i} p(S_i \mid \theta, \Delta_i)^{\Delta_i} (1-\lambda)^{(1-\Delta_i)} p(S_i \mid \theta, \Delta_i)^{(1-\Delta_i)} \\
&= exp\{\Delta_i \log\left(\tfrac{\lambda}{1-\lambda}\right) - [-log(1-\lambda)] \\
&\quad + \Delta_i \left[\tfrac{\mu_1}{\sigma_1^2} S_i + \left(-\tfrac{1}{2\sigma_1^2}\right) S_i^2 - \tfrac{1}{2}\left(\tfrac{\mu_1^2}{\sigma_1^2} + \log\left(2\pi\sigma_1^2\right)\right)\right] + \\
&\quad (1-\Delta_i)\left[\tfrac{\mu_2}{\sigma_2^2} S_i + \left(-\tfrac{1}{2\sigma_2^2}\right) S_i^2 - \tfrac{1}{2}\left(\tfrac{\mu_2^2}{\sigma_2^2} + \log\left(2\pi\sigma_2^2\right)\right)\right] \\
&\qquad \}
\end{aligned}$$

EM Algorithm: Gaussian Mixture

Complete-Data Natural Sufficient Statistic and Expectation:

$$\mathbf{T}(X_i) = \begin{bmatrix} \Delta_i \\ \Delta_i S_i \\ \Delta_i S_i^2 \\ (1-\Delta_i)S_i \\ (1-\Delta_i)S_i^2 \end{bmatrix} \text{ and } E[\mathbf{T}(X_i) \mid \theta] = \begin{bmatrix} \lambda \\ \lambda\mu_1 \\ \lambda(\sigma_1^2+\mu_1^2) \\ (1-\lambda)\mu_2 \\ (1-\lambda)(\sigma_2^2+\mu_2^2) \end{bmatrix}$$

Compute the MLE $\hat{\theta}$ by solving

$$\mathbf{T}(\mathbf{X}) = {}_{1}^{n}\mathbf{T}(X_i) = nE[T(X_i \mid \theta)] \ (*)$$

EM Algorithm:

1. Initialize estimate $\tilde{\theta}_n$, $n = 1$
2. Given preliminary estimate $\tilde{\theta}_n$ solve $(*)$ for θ^* using $E[\mathbf{T}(\mathbf{X}) \mid S(X), \theta = \tilde{\theta}_n]$ in place of $\mathbf{T}(\mathbf{X})$.
3. Replace θ_n with $\theta_{n+1} = \theta^*$ and return to step 2.

Finite Mixture Model

- $S_1, S_2, \ldots, S_n$ $i.i.d.$ with density $p(s_i \mid \theta)$, $s_i \in R^d$.
- $p(s_i \mid \theta) = \sum_{j=1}^m \lambda_j \phi_j(s_i)$ where

  $\{\phi_1(\cdot), \ldots, \phi_m(\cdot)\}$ are densities of mixture *components*

  $\{\lambda_1, \ldots, \lambda_m\}$: $\lambda_j > 0$, and $\sum_{j=1}^m \lambda_j = 1$. are *component* probabilities of the model

  $\theta = (\lambda_1, \ldots, \lambda_m, \phi_1, \ldots, \phi_m)$, (mixture model parameter)
- Assume every $\phi_j \in \mathcal{P}$, a given family of models

  E.g. 1: Gaussian Mixtures

  $$\mathcal{P} = \{N(\mu, \sigma^2), (\mu, \sigma^2) \in R \times R^+\}$$

  E.g. 2: p-parameter family given by $\phi(\cdot \mid \cdot)$

  $$\mathcal{P} = \{\phi(\cdot \mid \xi), \xi \in \mathcal{E} \subset R^p\}$$

  E.g. 3: Conditionally $i.i.d.$ coordinates of S_i

  $$\mathcal{P} = \{\phi(s_i) = \prod_{k=1}^d f(s_{i,k}), non - parametric \;\; f\}.$$

Complete Data Augmentation for Finite Mixtures

Observed Data: $S_1, S_2, \ldots, S_n$

Missing Data: $Z_1, Z_2, \ldots, Z_n$, which are $i.i.d.$

$$Multinomial(N = 1, probs = (\lambda_1, \ldots, \lambda_m)), \text{ i.e.,}$$

$$Z_i = (Z_{i,1}, Z_{i,2}, \ldots, Z_{i,m})$$

$$\begin{aligned}
&Z_{i,j} = 1 \text{ if case } i \text{ drawn from component } j \\
&\qquad\qquad \text{(otherwise 0)} \\
&Z_{i,j} \in \{0, 1\} \text{ (Bernoulli)} \\
&P(Z_{i,j} = 1) = \lambda_j, \\
&\lambda_j > 0,\ j = 1, \ldots, m, \text{ and } \textstyle\sum_{j=1}^m \lambda_j = 1.
\end{aligned}$$

Complete Data: $X_1, X_2, \ldots, X_n$

$X_i = (S_i, Z_i),\ i = 1, \ldots, n$ with density

$$\begin{aligned}
p(x_i \mid \theta) &= p(S_i, Z_i \mid \theta) \\
&= p(Z_i \mid \theta) p(S_i \mid Z_i, \theta) \\
&= \textstyle\sum_{j=1}^m I_{Z_{i,j}} p(Z_{i,j} = 1 \mid \theta) p(S_i \mid Z_{i,j} = 1, \theta) \quad] \\
&= \textstyle\sum_{j=1}^m I_{Z_{i,j}} \lambda_j \phi_j(S_i)
\end{aligned}$$

with: $\theta = (\lambda_1, \ldots, \lambda_m, \phi_1, \ldots, \phi_m)$.

EM Algorithm for Finite Mixtures

Log-Likelihood of Observed Data $S = (S_1, \dots, S_n)$

$$\ell_S(\theta) = \sum_{i=1}^n \log p(S_i \mid \theta) = \sum_{i=1}^n \log[\sum_{j=1}^m \lambda_j \phi_j(S_i)]$$

Conditional Expectation of Complete-Data Log-Likelihood

$$J(\theta \mid \theta^{(t)}) = E\left(\sum_{i=1}^n \log[p(X_i \mid \theta) \mid S, \theta^{(t)}]\right)$$

EM Algorithm

- Generate sequence of parameter estimates $\{\theta^{(t)}, t = 1, 2, \dots\}$
- Initialize $\theta^{(t)}$ for $t = 1$.
- Given $\theta^{(t)}$, generate $\theta^{(t+1)}$ as follows:

  E-Step: Compute $J(\theta \mid \theta^{(t)})$.

  M-Step: Set $\theta^{(t+1)} = argmax_\theta J(\theta \mid \theta^{(t)})$.
- Repeat previous step until successive changes in $\theta^{(t)}$ indicate convergence

E-Step in EM Algorithm for Finite Mixtures

Conditional Expectation of Complete-Data Log-Likelihood

$$
\begin{aligned}
J(\theta \mid \theta^{(t)}) &= E\left(\sum_{i=1}^n \log[p(X_i \mid \theta)] \mid S, \theta^{(t)}\right) \\
&= E\left(\sum_{i=1}^n \log[\sum_{j=1}^m I_{Z_{i,j}} \lambda_j \phi_j(S_i)] \mid S, \theta^{(t)}\right) \\
&= E\left(\sum_{i=1}^n \sum_{j=1}^m I_{Z_{i,j}} \log[\lambda_j \phi_j(S_i)] \mid S, \theta^{(t)}\right) \\
&= \sum_{i=1}^n \sum_{j=1}^m E\left(I_{Z_{i,j}} \log[\lambda_j \phi_j(S_i)] \mid S, \theta^{(t)}\right) \\
&= \sum_{i=1}^n \sum_{j=1}^m [E\left(I_{Z_{i,j}} \mid S, \theta^{(t)}\right)] \log[\lambda_j \phi_j(S_i)] \\
&= \sum_{i=1}^n \sum_{j=1}^m [P(Z_{i,j} = 1 \mid S, \theta^{(t)})] \log[\lambda_j \phi_j(S_i)] \\
&= \sum_{i=1}^n \sum_{j=1}^m p_{i,j}^{(t)} \log[\lambda_j \phi_j(S_i)] \\
&= [\sum_{j=1}^m \log(\lambda_j)(\sum_{i=1}^n p_{i,j}^{(t)})] \\
&\qquad + \quad [\sum_{j=1}^m (\sum_{i=1}^n p_{i,j}^{(t)} \log[\phi_j(S_i)])]
\end{aligned}
$$

where $p_{i,j}^{(t)} = P(Z_{i,j} = 1 \mid S, \theta^{(t)}) = \dfrac{\lambda_j^{(t)} \phi_j^{(t)}(S_i)}{\sum_{j^*=1}^m \lambda_{j^*} j^{(t)} \phi_{j^*} j^{(t)}(S_i)}$

M-Step in EM Algorithm for Finite Mixtures

Solve for $\theta = (\lambda_1, \dots, \lambda_m, \phi_1, \dots, \phi_m)$ maximizing

$$\begin{array}{rcl} J(\theta \mid \theta^{(t)}) & = & E\left(\sum_{i=1}^n \log[p(X_i \mid \theta)] \mid S, \theta^{(t)}\right) \\ & = & [\sum_{j=1}^m \log(\lambda_j)(\sum_{i=1}^n p_{i,j}^{(t)})] \\ & & \quad + \quad [\sum_{j=1}^m(\sum_{i=1}^n p_{i,j}^{(t)} \log[\phi_j(S_i)])] \end{array}$$

where $p_{i,j}^{(t)} = P(Z_{i,j} = 1 \mid S, \theta^{(t)}) = \dfrac{\lambda_j^{(t)}\phi_j^{(t)}(S_i)}{\sum_{j^*=1}^m \lambda_{j^*}j^{(t)}\phi_{j^*}j^{(t)}(S_i)}$

M-Step for $\lambda_1, \dots, \lambda_m$: $\quad \lambda_j^{(t+1)} = \frac{1}{n}\sum_{i=1}^n p_{i,j}^{(t)}$

(same formula for all $\phi_j^{(t)}$)

M-Step for $\phi_1, \dots, \phi_m$**:** maximize sum of case-weighted conditional-log-likelihoods of the $\phi_j(\cdot)$

$$[\sum_{j=1}^m \ (\sum_{i=1}^n p_{i,j}^{(t)} \log[\phi_j(S_i)]) \ ]$$

References

Dempster, AP, Laird, NM, and Rubin, DB (1977). "Maximum Likelihood from Incomplete Data Via the EM Algorithm." *Journal of the Royal Statistial Society. Series B (Methodological),* **39***(1), 1-38.*

Bengalia, T., Chauveau, D. Hunter, D.R., and Young, D.S. "mixtools: An R Package for Analyzing Finite Mixture Models" *Journal of Statistial Software,* October 2009, Volume 32, Issue 6, 1-29, https://www.jstatsoft.org/article/view/v032i06

18.655 Mathematical Statistics
Spring 2016