MAT-468: Sesión 15, Versiones estocásticas del algoritmo EM

Felipe Osorio

fosorios.mat.utfsm.cl

Departamento de Matemática, UTFSM

Monte Carlo EM (Wei y Tanner, 1990)[1]

Para facilitar el paso-E del algoritmo, es posible usar el método Monte Carlo para aproximar la función Q.

Algoritmo 1: Monte Carlo EM (MCEM).

Entrada: Conjunto de datos observados $\boldsymbol{Y}_{\textsf{obs}}$ y estimación inicial $\boldsymbol{\theta}^{(0)}$.
Salida : Estimación ML de $\boldsymbol{\theta}$.

begin

Simulación: Generar $\boldsymbol{z}_1, \dots, \boldsymbol{z}_M \overset{\textsf{IID}}{\sim} p(\boldsymbol{y}_{\textsf{mis}}|\boldsymbol{y}_{\textsf{obs}}; \boldsymbol{\theta}^{(k)})$.

Aproximación: Sea

$$\widehat{Q}_M(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)}) = \frac{1}{M}\sum_{j=1}^{M} \log p(\boldsymbol{\theta}; \boldsymbol{z}_j, \boldsymbol{y}_{\textsf{obs}}).$$

Paso-M: actualizar $\boldsymbol{\theta}^{(k+1)}$, como:

$$\boldsymbol{\theta}^{(k+1)} = \arg\max_{\theta} \widehat{Q}_M(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)}).$$

Iterar entre pasos-E y M hasta alcanzar convergencia.

end

[1] Journal of the American Statistical Association **85**, 699-704.

Note que

$$\frac{\partial \widehat{Q}_M(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)})}{\partial \boldsymbol{\theta}} = \frac{1}{M}\sum_{j=1}^{M} \frac{\partial}{\partial \boldsymbol{\theta}} \log p(\boldsymbol{\theta}; \boldsymbol{z}_j, \boldsymbol{y}_{\mathsf{obs}})$$

$$\frac{\partial^2 \widehat{Q}_M(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^\top} = \frac{1}{M}\sum_{j=1}^{M} \frac{\partial^2}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^\top} \log p(\boldsymbol{\theta}; \boldsymbol{z}_j, \boldsymbol{y}_{\mathsf{obs}}).$$

De este modo, el paso-M del algoritmo MCEM, adopta la forma

$$\boldsymbol{\theta}^{(r+1)} = \boldsymbol{\theta}^{(r)} + \Big\{ - \frac{\partial^2 \widehat{Q}_M(\boldsymbol{\theta}^{(r)};\boldsymbol{\theta}^{(k)})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^\top} \Big\}^{-1} \frac{\partial \widehat{Q}_M(\boldsymbol{\theta}^{(r)};\boldsymbol{\theta}^{(k)})}{\partial \boldsymbol{\theta}}, \qquad (1)$$

a la convergencia de (1) hacemos $\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^*$.

Observación:

Podemos usar los mismos datos simulados $\boldsymbol{z}_1, \dots, \boldsymbol{z}_n$ para aproximar

$$\partial Q(\boldsymbol{\theta};\boldsymbol{\theta})/\partial \boldsymbol{\theta}, \qquad \partial^2 Q(\boldsymbol{\theta};\boldsymbol{\theta})/\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^\top.$$

Algoritmo MH-RM (Cai, 2010)[2]

Sea

$$\boldsymbol{s}(\boldsymbol{\theta};\boldsymbol{y}_{\mathsf{com}}) = \nabla_\theta\, \ell_{\mathsf{c}}(\boldsymbol{\theta};\boldsymbol{y}_{\mathsf{com}}),$$

el gradiente de la función de log-verosimilitud de datos completos. Considere la identidad de Fisher,

$$\nabla_\theta\, \ell_{\mathsf{o}}(\boldsymbol{\theta};\boldsymbol{y}_{\mathsf{obs}}) = \int \boldsymbol{s}(\boldsymbol{\theta};\boldsymbol{y}_{\mathsf{com}})p(\boldsymbol{y}_{\mathsf{mis}}|\boldsymbol{y}_{\mathsf{obs}};\boldsymbol{\theta})\,\mathsf{d}\boldsymbol{y}_{\mathsf{mis}}.$$

El método de aproximación estocástica propuesto por Cai (2010) está basado en el algoritmo de Robbins-Monró, usando $\{\gamma_k\}_{k\geq 1}$, tal que

$$\sum_{k=1}^{\infty}\gamma_k = \infty, \qquad \sum_{k=1}^{\infty}\gamma_k^2 < \infty.$$

Sea

$$\mathsf{H}_{\mathsf{o}}(\boldsymbol{\theta};\boldsymbol{y}_{\mathsf{com}}) = -\frac{\partial^2 \ell_{\mathsf{c}}(\boldsymbol{\theta};\boldsymbol{y}_{\mathsf{com}})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top},$$

la matriz de información de datos completos.

[2] Psychometrika **75**, 33-57.

Algoritmo 2: Algoritmo MH-RM.

Entrada: Estimación inicial $\boldsymbol{\theta}^{(0)}$ y $\boldsymbol{\Gamma}_0$ matriz definida positiva.

begin

Simulación: Generar $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_M \overset{\text{IID}}{\sim} p(\boldsymbol{y}_{\text{mis}} | \boldsymbol{y}_{\text{obs}}; \boldsymbol{\theta}^{(k)})$, y formar el conjunto de datos completos

$$\boldsymbol{y}_{1,\text{com}}^{(k+1)}, \ldots, \boldsymbol{y}_{M,\text{com}}^{(k+1)}, \qquad \boldsymbol{y}_{i,\text{com}}^{(k+1)} = (\boldsymbol{z}_i, \boldsymbol{y}_{\text{obs}})$$

Aproximación: de $\nabla_\theta\, \ell_{\text{o}}(\boldsymbol{\theta}^{(k)}; \boldsymbol{y}_{\text{obs}})$ y de la matriz de información de datos completos

$$\boldsymbol{s}_{k+1} = \frac{1}{M} \sum_{j=1}^{M} \boldsymbol{s}(\boldsymbol{\theta}^{(k)}; \boldsymbol{y}_{j,\text{com}}^{(k+1)})$$

$$\boldsymbol{\Gamma}_{k+1} = \boldsymbol{\Gamma}_k + \gamma_k \Big\{ \frac{1}{M} \sum_{j=1}^{M} \mathsf{H}_{\text{o}}(\boldsymbol{\theta}^{(k)}; \boldsymbol{y}_{\text{com}}^{(k+1)}) - \boldsymbol{\Gamma}_k \Big\}.$$

Actualización: Usamos Robbins-Monró para actualizar $\boldsymbol{\theta}^{(k+1)}$,

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} + \gamma_k \, \boldsymbol{\Gamma}_{k+1}^{-1} \boldsymbol{s}_{k+1}$$

Iterar entre las etapas anteriores hasta alcanzar convergencia.

end

Algoritmo SAEM (Delyon, Lavielle y Moulines, 1999)[3]

Algoritmo 3: Algoritmo SAEM.

Entrada: Estimación inicial $\boldsymbol{\theta}^{(0)}$.

begin

Simulación: Generar $\boldsymbol{z}_1, \dots, \boldsymbol{z}_M \stackrel{\text{IID}}{\sim} p(\boldsymbol{y}_{\text{mis}} | \boldsymbol{y}_{\text{obs}}; \boldsymbol{\theta}^{(k-1)})$, y formar el conjunto de datos completos

$$\boldsymbol{y}_{1,\text{com}}^{(k+1)}, \dots, \boldsymbol{y}_{M,\text{com}}^{(k+1)}, \qquad \boldsymbol{y}_{i,\text{com}}^{(k+1)} = (\boldsymbol{z}_i, \boldsymbol{y}_{\text{obs}})$$

Aproximación: Actualizar $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)})$,

$$Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)}) = Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k-1)}) + \gamma_k \Big\{ \frac{1}{M} \sum_{j=1}^{M} \log p(\boldsymbol{y}_{j,\text{com}}^{(k)}; \boldsymbol{\theta}^{(k-1)}) - Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k-1)}) \Big\}$$

Actualización: Resolver el problema,

$$\boldsymbol{\theta}^{(k+1)} = \arg\max_{\theta} Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)}).$$

Iterar entre las etapas anteriores hasta alcanzar convergencia.

end

[3] The Annals of Statistics **27**, 94-128.

Algoritmo SAEM (Gu y Kong, 1998)[4]

Algoritmo 4: Algoritmo SAEM.

Entrada: Estimación inicial $\boldsymbol{\theta}^{(0)}$ y $\boldsymbol{\Gamma}_0$ matriz definida positiva.

begin

Simulación: Generar $\boldsymbol{z}_1, \dots, \boldsymbol{z}_M \overset{\text{IID}}{\sim} p(\boldsymbol{y}_{\text{mis}} | \boldsymbol{y}_{\text{obs}}; \boldsymbol{\theta}^{(k)})$, y formar el conjunto de datos completos

$$\boldsymbol{y}_{1,\text{com}}^{(k+1)}, \dots, \boldsymbol{y}_{M,\text{com}}^{(k+1)}, \qquad \boldsymbol{y}_{i,\text{com}}^{(k+1)} = (\boldsymbol{z}_i, \boldsymbol{y}_{\text{obs}})$$

Aproximación: del score y la matriz de información de datos completos

$$\boldsymbol{g}_{k+1} = \boldsymbol{g}_k + \gamma_k \Big\{ \frac{1}{M} \sum_{j=1}^{M} \boldsymbol{s}(\boldsymbol{\theta}^{(k)}; \boldsymbol{y}_{j,\text{com}}^{(k+1)}) - \boldsymbol{g}_k \Big\}$$

$$\boldsymbol{\Gamma}_{k+1} = \boldsymbol{\Gamma}_k + \gamma_k \Big\{ \frac{1}{M} \sum_{j=1}^{M} \big[\mathsf{H}_{\mathsf{c}}(\boldsymbol{\theta}^{(k)}; \boldsymbol{y}_{\text{com}}^{(k+1)}) - \boldsymbol{s}(\boldsymbol{\theta}^{(k)}; \boldsymbol{y}_{j,\text{com}}^{(k+1)}) \boldsymbol{s}^\top (\boldsymbol{\theta}^{(k)}; \boldsymbol{y}_{j,\text{com}}^{(k+1)}) \big] - \boldsymbol{\Gamma}_k \Big\}.$$

Actualización: Usar Robbins-Monró para actualizar $\boldsymbol{\theta}^{(k+1)}$,

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} + \gamma_k \, \boldsymbol{\Gamma}_{k+1}^{-1} \boldsymbol{g}_{k+1}$$

Iterar entre las etapas anteriores hasta alcanzar convergencia.

end

[4] Proceedings of the National Academy of Sciences of USA **95**, 7270-7274.

Error estándar en el contexto de EM

Disponemos de varias alternativas para estimar el error estándar de $\widehat{\boldsymbol{\theta}}$. En particular, podemos usar:

- Matriz de información de Fisher

$$\boldsymbol{F}_{\mathsf{o}}(\boldsymbol{\theta}) = \mathsf{E}\{\boldsymbol{s}_{\mathsf{o}}(\boldsymbol{\theta})\boldsymbol{s}_{\mathsf{o}}^{\top}(\boldsymbol{\theta})\},$$

  cuando $\ell_{\mathsf{o}}(\boldsymbol{\theta})$ es dos veces diferenciable podemos usar

$$\boldsymbol{F}_{\mathsf{o}}(\boldsymbol{\theta}) = \mathsf{E}\Big\{-\frac{\partial^2 \ell_{\mathsf{o}}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\top}}\Big\}.$$

- Matriz de información observada

$$\mathsf{H}_{\mathsf{o}}(\boldsymbol{\theta}) = -\frac{\partial^2 \ell_{\mathsf{o}}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\top}}.$$

- Versión empírica de la matriz de información

$$\widehat{\boldsymbol{F}}_{\mathsf{o}}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{s}_i(\boldsymbol{\theta})\boldsymbol{s}_i^{\top}(\boldsymbol{\theta}).$$

Principio de información perdida (Orchand y Woodbury, 1972)

Sabemos que

$$\ell_{\mathsf{o}}(\boldsymbol{\theta};\boldsymbol{y}_{\mathsf{obs}}) = \ell_{\mathsf{c}}(\boldsymbol{\theta};\boldsymbol{y}_{\mathsf{com}}) - \log k(\boldsymbol{y}_{\mathsf{mis}}|\boldsymbol{y}_{\mathsf{obs}};\boldsymbol{\theta}),$$

tomando segundas derivadas del negativo de la expresión anterior, tenemos

$$\mathsf{H}_{\mathsf{o}}(\boldsymbol{\theta};\boldsymbol{y}_{\mathsf{obs}}) = \mathsf{H}_{\mathsf{c}}(\boldsymbol{\theta};\boldsymbol{y}_{\mathsf{com}}) + \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\log k(\boldsymbol{y}_{\mathsf{mis}}|\boldsymbol{y}_{\mathsf{obs}};\boldsymbol{\theta}),$$

donde

$$\mathsf{H}_{\mathsf{o}}(\boldsymbol{\theta};\boldsymbol{y}_{\mathsf{obs}}) = -\frac{\partial^2\ell_{\mathsf{o}}(\boldsymbol{\theta};\boldsymbol{y}_{\mathsf{obs}})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}, \qquad \mathsf{H}_{\mathsf{c}}(\boldsymbol{\theta};\boldsymbol{y}_{\mathsf{obs}}) = -\frac{\partial^2\ell_{\mathsf{c}}(\boldsymbol{\theta};\boldsymbol{y}_{\mathsf{com}})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top},$$

son las matrices de información (observadas) para el modelo de datos observados y completos, respectivamente.

Principio de información perdida (Orchand y Woodbury, 1972)

Tomando esperanzas con relación a la distribución condicional de $\boldsymbol{y}_{\mathsf{mis}}|\boldsymbol{y}_{\mathsf{obs}}$, obtenemos

$$\mathsf{H}_{\mathsf{o}}(\boldsymbol{\theta}; \boldsymbol{y}_{\mathsf{obs}}) = \boldsymbol{F}_{\mathsf{c}}(\boldsymbol{\theta}) - \boldsymbol{F}_{\mathsf{m}}(\boldsymbol{\theta}), \qquad (2)$$

con

$$\boldsymbol{F}_{\mathsf{c}}(\boldsymbol{\theta}) = \mathsf{E}\{\mathsf{H}_{\mathsf{o}}(\boldsymbol{\theta}; \boldsymbol{y}_{\mathsf{obs}})|\boldsymbol{y}_{\mathsf{obs}}\}$$

y

$$\boldsymbol{F}_{\mathsf{m}}(\boldsymbol{\theta}) = \mathsf{E}\Big\{ -\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top} \log k(\boldsymbol{y}_{\mathsf{mis}}|\boldsymbol{y}_{\mathsf{obs}}; \boldsymbol{\theta})\Big|\boldsymbol{y}_{\mathsf{obs}}\Big\}.$$

Finalmente, integrando sobre la distribución de $\boldsymbol{y}_{\mathsf{obs}}$,

$$\boldsymbol{F}_{\mathsf{o}}(\boldsymbol{\theta}) = \boldsymbol{F}_{\mathsf{c}}(\boldsymbol{\theta}) - \mathsf{E}\{\boldsymbol{F}_{\mathsf{m}}(\boldsymbol{\theta})\}.$$

Note también que, siempre que sea posible intercambiar las operaciones de integración y diferenciación, tenemos

$$\boldsymbol{F}_{\mathsf{c}}(\boldsymbol{\theta}) = -\frac{\partial^2 Q(\boldsymbol{\theta}; \widehat{\boldsymbol{\theta}})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}.$$

Louis (1982) mostró que la matriz de información perdida, puede ser escrita como:

$$
\begin{aligned}
\boldsymbol{F}_{\mathsf{m}}(\boldsymbol{\theta}) &= \operatorname{Cov}(\boldsymbol{s}_{\mathsf{c}}(\boldsymbol{\theta};\boldsymbol{y}_{\mathsf{com}})|\boldsymbol{y}_{\mathsf{obs}}) \\
&= \mathsf{E}\{\boldsymbol{s}_{\mathsf{c}}(\boldsymbol{\theta};\boldsymbol{y}_{\mathsf{com}})\boldsymbol{s}_{\mathsf{c}}^{\top}(\boldsymbol{\theta};\boldsymbol{y}_{\mathsf{com}})|\boldsymbol{y}_{\mathsf{obs}}\} - \mathsf{E}\{\boldsymbol{s}_{\mathsf{c}}(\boldsymbol{\theta};\boldsymbol{y}_{\mathsf{com}})|\boldsymbol{y}_{\mathsf{obs}}\}\,\mathsf{E}^{\top}\{\boldsymbol{s}_{\mathsf{c}}(\boldsymbol{\theta};\boldsymbol{y}_{\mathsf{com}})|\boldsymbol{y}_{\mathsf{obs}}\} \\
&= \mathsf{E}\{\boldsymbol{s}_{\mathsf{c}}(\boldsymbol{\theta};\boldsymbol{y}_{\mathsf{com}})\boldsymbol{s}_{\mathsf{c}}^{\top}(\boldsymbol{\theta};\boldsymbol{y}_{\mathsf{com}})|\boldsymbol{y}_{\mathsf{obs}}\} - \boldsymbol{s}_{\mathsf{o}}(\boldsymbol{\theta};\boldsymbol{y}_{\mathsf{obs}})\boldsymbol{s}_{\mathsf{o}}^{\top}(\boldsymbol{\theta};\boldsymbol{y}_{\mathsf{obs}})
\end{aligned}
$$

Substituyendo en Ecuación (2), tenemos

$$
\begin{aligned}
\mathsf{H}_{\mathsf{o}}(\boldsymbol{\theta};\boldsymbol{y}_{\mathsf{obs}}) &= \boldsymbol{F}_{\mathsf{c}}(\boldsymbol{\theta}) - \boldsymbol{F}_{\mathsf{m}}(\boldsymbol{\theta}) \\
&= \boldsymbol{F}_{\mathsf{c}}(\boldsymbol{\theta}) - \mathsf{E}\{\boldsymbol{s}_{\mathsf{c}}(\boldsymbol{\theta};\boldsymbol{y}_{\mathsf{com}})\boldsymbol{s}_{\mathsf{c}}^{\top}(\boldsymbol{\theta};\boldsymbol{y}_{\mathsf{com}})|\boldsymbol{y}_{\mathsf{obs}}\} - \boldsymbol{s}_{\mathsf{o}}(\boldsymbol{\theta};\boldsymbol{y}_{\mathsf{obs}})\boldsymbol{s}_{\mathsf{o}}^{\top}(\boldsymbol{\theta};\boldsymbol{y}_{\mathsf{obs}}).
\end{aligned}
$$

A la convergencia del algoritmo EM podemos considerar,

$$
\mathsf{H}_{\mathsf{o}}(\boldsymbol{\theta};\boldsymbol{y}_{\mathsf{obs}}) = \boldsymbol{F}_{\mathsf{c}}(\boldsymbol{\theta}) - \mathsf{E}\{\boldsymbol{s}_{\mathsf{c}}(\boldsymbol{\theta};\boldsymbol{y}_{\mathsf{com}})\boldsymbol{s}_{\mathsf{c}}^{\top}(\boldsymbol{\theta};\boldsymbol{y}_{\mathsf{com}})|\boldsymbol{y}_{\mathsf{obs}}\}.
$$