MAT-468: Sesión 14, Algoritmo EM

Felipe Osorio

fosorios.mat.utfsm.cl

Departamento de Matemática, UTFSM

Algoritmo EM (Esperanza-Maximización)

Consideraciones:

- Algoritmo para el cálculo iterativo de estimadores ML en modelos con datos incompletos.
- Requiere de una formulación de datos aumentados.
- Reemplaza una optimización "compleja" (estimación ML) por una serie de maximizaciones "simples".

Algoritmo EM (Dempster, Laird y Rubin, 1977)[1]

Sea $\boldsymbol{Y}_{\mathsf{com}} = (\boldsymbol{Y}_{\mathsf{obs}}, \boldsymbol{Y}_{\mathsf{mis}})$ vector de datos completos con función de densidad

$$p(\boldsymbol{y}_{\mathsf{com}}; \boldsymbol{\theta}),$$

donde $\boldsymbol{Y}_{\mathsf{obs}}$ y $\boldsymbol{Y}_{\mathsf{mis}}$ denotan los datos observados y perdidos, respectivamente.

En modelos con datos incompletos, la log-verosimilitud de datos observados

$$\ell_{\mathsf{o}}(\boldsymbol{\theta}; \boldsymbol{Y}_{\mathsf{obs}}) = \log p(\boldsymbol{y}_{\mathsf{obs}}; \boldsymbol{\theta}) = \log \int p(\boldsymbol{y}_{\mathsf{com}}; \boldsymbol{\theta}) \, \mathsf{d}\, \boldsymbol{y}_{\mathsf{mis}},$$

puede ser difícil de maximizar directamente.

El algoritmo EM aumenta los datos $\boldsymbol{Y}_{\mathsf{obs}}$ con variables latentes permitiendo que la log-verosimilitud de datos completos

$$\ell_{\mathsf{c}}(\boldsymbol{\theta}; \boldsymbol{Y}_{\mathsf{com}}) = \log p(\boldsymbol{y}_{\mathsf{com}}; \boldsymbol{\theta}),$$

sea bastante simple para muchas aplicaciones en Estadística.

[1] Journal of the Royal Statistical Society, Series B **39**, 1-38.

Algoritmo EM

Algoritmo 1: Algoritmo EM (Esperanza-Maximización).

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

begin

Paso-E: para $\boldsymbol{\theta}^{(k)}$ estimación de $\boldsymbol{\theta}$ en la k-ésima iteración, calcular:

$$\begin{aligned} Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)}) &= \mathsf{E}\{\ell_{\mathsf{c}}(\boldsymbol{\theta};\boldsymbol{Y}_{\mathsf{com}})|\boldsymbol{Y}_{\mathsf{obs}},\boldsymbol{\theta}^{(k)}\} \\ &= \int \ell_c(\boldsymbol{\theta};\boldsymbol{Y}_{\mathsf{com}}) f(\boldsymbol{y}_{\mathsf{mis}}|\boldsymbol{y}_{\mathsf{obs}};\boldsymbol{\theta}^{(k)})\,\mathsf{d}\boldsymbol{y}_{\mathsf{mis}}. \end{aligned} \tag{1}$$

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

$$\boldsymbol{\theta}^{(k+1)} = \arg\max_{\theta} Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)}). \tag{2}$$

Iterar entre pasos-E y M hasta alcanzar convergencia.

end

Algoritmo EM generalizado

Algoritmo 2: Algoritmo GEM.

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

begin

Paso-E: para $\boldsymbol{\theta}^{(k)}$ estimación de $\boldsymbol{\theta}$ en la k-ésima iteración, calcular:

$$\begin{aligned} Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)}) &= \mathsf{E}\{\ell_{\mathsf{c}}(\boldsymbol{\theta};\boldsymbol{Y}_{\text{com}})|\boldsymbol{Y}_{\text{obs}},\boldsymbol{\theta}^{(k)}\} \\ &= \int \ell_c(\boldsymbol{\theta};\boldsymbol{Y}_{\text{com}}) f(\boldsymbol{y}_{\text{mis}}|\boldsymbol{y}_{\text{obs}};\boldsymbol{\theta}^{(k)})\, \mathrm{d}\boldsymbol{y}_{\text{mis}}. \end{aligned}$$

Paso-M*: seleccionar $\boldsymbol{\theta}^{(k+1)}$ satisfaciendo,

$$Q(\boldsymbol{\theta}^{(k+1)};\boldsymbol{\theta}^{(k)}) > Q(\boldsymbol{\theta}^{(k)};\boldsymbol{\theta}^{(k)}).$$

Iterar entre pasos-E y M hasta alcanzar convergencia.

end

Sugerencia:

Considerar una única iteración Newton en la optimización de $Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)})$ requerida en el Paso-M*.

Propiedades del algoritmo EM

La clave del algoritmo EM está basada en el hecho que mediante maximizar $Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)})$ permite incrementar la verosimilitud en cada etapa.

La identidad básica del algoritmo EM es

$$p(\boldsymbol{y}_{\mathsf{mis}}|\boldsymbol{y}_{\mathsf{obs}};\boldsymbol{\theta}) = \frac{p(\boldsymbol{y}_{\mathsf{com}};\boldsymbol{\theta})}{p(\boldsymbol{y}_{\mathsf{obs}};\boldsymbol{\theta})} \quad \Big(\frac{p(\boldsymbol{y}_{\mathsf{obs}},\boldsymbol{y}_{\mathsf{mis}};\boldsymbol{\theta})}{p(\boldsymbol{y}_{\mathsf{obs}};\boldsymbol{\theta})}\Big),$$

de ahí que

$$\log p(\boldsymbol{y}_{\mathsf{obs}};\boldsymbol{\theta}) = \log p(\boldsymbol{y}_{\mathsf{com}};\boldsymbol{\theta}) - \log p(\boldsymbol{y}_{\mathsf{mis}}|\boldsymbol{y}_{\mathsf{obs}};\boldsymbol{\theta}).$$

De este modo, para cualquier $\boldsymbol{\theta}_0$

$$\begin{aligned}\log p(\boldsymbol{y}_{\mathsf{obs}};\boldsymbol{\theta}) &= \mathsf{E}\{\log p(\boldsymbol{y}_{\mathsf{com}};\boldsymbol{\theta})|\boldsymbol{y}_{\mathsf{obs}},\boldsymbol{\theta}_0\} - \mathsf{E}\{\log p(\boldsymbol{y}_{\mathsf{mis}}|\boldsymbol{y}_{\mathsf{obs}};\boldsymbol{\theta})|\boldsymbol{y}_{\mathsf{obs}},\boldsymbol{\theta}_0\} \\ &= Q(\boldsymbol{\theta};\boldsymbol{\theta}_0) - H(\boldsymbol{\theta};\boldsymbol{\theta}_0).\end{aligned}$$

(estas esperanzas son tomadas con relación a la distribución condicional $p(\boldsymbol{y}_{\mathsf{mis}}|\boldsymbol{y}_{\mathsf{obs}},\boldsymbol{\theta}_0)$). De este modo, se tiene que

$$\begin{aligned}\ell_{\mathsf{o}}(\boldsymbol{\theta}^{(k+1)}) - \ell_{\mathsf{o}}(\boldsymbol{\theta}^{(k)}) &= Q(\boldsymbol{\theta}^{(k+1)};\boldsymbol{\theta}^{(k)}) - Q(\boldsymbol{\theta}^{(k)};\boldsymbol{\theta}^{(k)}) \\ &\quad - \{H(\boldsymbol{\theta}^{(k+1)};\boldsymbol{\theta}^{(k)}) - H(\boldsymbol{\theta}^{(k)};\boldsymbol{\theta}^{(k)})\}.\end{aligned}$$

EX UMBRA IN SOLEM

Propiedades del algoritmo EM

La primera diferencia es no negativa pues

$$Q(\boldsymbol{\theta}^{(k+1)}; \boldsymbol{\theta}^{(k)}) \geq Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)}), \qquad \forall\, \boldsymbol{\theta},$$

para cualquier $\boldsymbol{\theta}$

$$\begin{aligned} H(\boldsymbol{\theta}^{(k+1)}; \boldsymbol{\theta}^{(k)}) - H(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(k)}) &= \mathsf{E}\left\{ \log \frac{p(\boldsymbol{y}_{\mathsf{mis}}|\boldsymbol{y}_{\mathsf{obs}}; \boldsymbol{\theta})}{p(\boldsymbol{y}_{\mathsf{mis}}|\boldsymbol{y}_{\mathsf{obs}}; \boldsymbol{\theta}^{(k)})} \Big| \boldsymbol{y}_{\mathsf{obs}}, \boldsymbol{\theta}^{(k)} \right\} \\ &\leq \log \mathsf{E}\left\{ \frac{p(\boldsymbol{y}_{\mathsf{mis}}|\boldsymbol{y}_{\mathsf{obs}}; \boldsymbol{\theta})}{p(\boldsymbol{y}_{\mathsf{mis}}|\boldsymbol{y}_{\mathsf{obs}}; \boldsymbol{\theta}^{(k)})} \Big| \boldsymbol{y}_{\mathsf{obs}}, \boldsymbol{\theta}^{(k)} \right\} \\ &= \log \int p(\boldsymbol{y}_{\mathsf{mis}}|\boldsymbol{y}_{\mathsf{obs}}; \boldsymbol{\theta})\, \mathsf{d}\boldsymbol{y}_{\mathsf{mis}} = 0. \end{aligned}$$

De este modo la secuencia $\{\boldsymbol{\theta}^{(k)}\}$ satisface

$$\ell_{\mathsf{o}}(\boldsymbol{\theta}^{(k+1)}; \boldsymbol{y}_{\mathsf{obs}}) \geq \ell_{\mathsf{o}}(\boldsymbol{\theta}^{(k)}; \boldsymbol{y}_{\mathsf{obs}}).$$

Observación:

Ésta propiedad aún es válida si se selecciona un $\boldsymbol{\theta}^{(k+1)}$ tal que:

$$Q(\boldsymbol{\theta}^{(k+1)}; \boldsymbol{\theta}^{(k)}) \geq Q(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(k)})$$

EX UMBRA IN SOLEM

Teorema (Dempster, Laird y Rubin, 1977)

Todo algoritmo EM o GEM incrementa la log-verosimilitud de datos observados $\ell_{\mathsf{o}}(\boldsymbol{\theta}; \boldsymbol{y}_{\mathsf{obs}})$ en cada iteración, esto es,

$$\ell_{\mathsf{o}}(\boldsymbol{\theta}^{(k+1)}; \boldsymbol{y}_{\mathsf{obs}}) \geq \ell_{\mathsf{o}}(\boldsymbol{\theta}^{(k)}; \boldsymbol{y}_{\mathsf{obs}}).$$

Convergencia (Wu, 1983; Little y Rubin, 1987)

Bajo condiciones suaves, la secuencia $\{\boldsymbol{\theta}^{(k)}\}_{k\geq 0}$ generada por el algoritmo EM (GEM). Converge a un punto estacionario de $\ell_{\mathsf{o}}(\boldsymbol{\theta}; \boldsymbol{Y}_{\mathsf{obs}})$.

Propiedades del algoritmo EM:

- Frecuentemente el algoritmo EM es simple, de bajo costo computacional y numéricamente estable.
- Dempster, Laird y Rubin (1977) mostraron que el algoritmo EM converge con velocidad lineal, que depende de la proporción de información perdida.[2]
- Para modelos con datos aumentados con densidad en la familia exponencial, el algoritmo EM reduce a actualizar las estadísticas suficientes.
- Errores estándar pueden ser obtenidos por cálculo directo, diferenciación numérica o usando el Principio de Información Perdida (Louis, 1982).

[2] puede ser **extremamente** lento.

Un vector aleatorio $\boldsymbol{y}$ tiene una distribución t multivariada con posición $\boldsymbol{\mu}$ y escala $\boldsymbol{\Sigma}$, esto es $\boldsymbol{y} \sim t_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$, si su función de densidad asume la forma

$$f(\boldsymbol{y}) = \frac{\Gamma(\frac{\nu+p}{2})}{\Gamma(\frac{\nu}{2})\pi^{p/2}}\, \nu^{-p/2}|\boldsymbol{\Sigma}|^{-1/2}\Big(1 + \frac{u}{\nu}\Big)^{-\frac{1}{2}(\nu+p)},$$

con $u = (\boldsymbol{y} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{\mu})$. Esta distribución puede ser escrito como

$$\boldsymbol{Y}|W \sim \mathsf{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}/\omega), \qquad W \sim \mathsf{Gamma}\Big(\frac{\nu}{2}, \frac{\nu}{2}\Big).$$

Observación:

Recuerde que Gamma(a, b) tiene densidad

$$f(\omega) = \frac{b^a}{\Gamma(a)}\, \omega^{a-1} \exp(-b\omega), \qquad \omega \geq 0, a, b > 0.$$

Estimación ML en la distribución t multivariada

Un vector aleatorio $\boldsymbol{y}$ tiene una distribución t multivariada con posición $\boldsymbol{\mu}$ y escala $\boldsymbol{\Sigma}$, esto es $\boldsymbol{y} \sim t_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$, si su función de densidad asume la forma

$$f(\boldsymbol{y}) = \frac{\Gamma(\frac{\nu+p}{2})}{\Gamma(\frac{\nu}{2})\pi^{p/2}}\, \nu^{-p/2}|\boldsymbol{\Sigma}|^{-1/2}\Big(1 + \frac{u}{\nu}\Big)^{-\frac{1}{2}(\nu+p)},$$

con $u = (\boldsymbol{y} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{\mu})$. Esta distribución puede ser escrito como

$$\boldsymbol{Y}|W \sim \mathsf{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}/\omega), \qquad W \sim \mathsf{Gamma}\Big(\frac{\nu}{2}, \frac{\nu}{2}\Big).$$

Observación:

Recuerde que Gamma(a, b) tiene densidad

$$f(\omega) = \frac{b^a}{\Gamma(a)}\, \omega^{a-1} \exp(-b\omega), \qquad \omega \geq 0, a, b > 0.$$

Note que, la función de log-verosimilitud de datos observados para una muestra aleatoria de tamaño n, desde $t_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$ adopta la forma:

$$
\begin{aligned}
\ell_{\mathsf{o}}(\boldsymbol{\theta}; \boldsymbol{y}_{\mathsf{obs}}) &= \sum_{i=1}^{n} \log f(\boldsymbol{y}_i; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu) \\
&= n \log \Gamma\Big(\frac{\nu + p}{2}\Big) - n \log \Gamma\Big(\frac{\nu}{2}\Big) - \frac{np}{2} \log \nu + \frac{np}{2} \log \pi \\
&- \frac{n}{2} \log |\boldsymbol{\Sigma}| - \frac{\nu + p}{2} \sum_{i=1}^{n} \log \Big(1 + \frac{u_i}{\nu}\Big).
\end{aligned}
$$

A continuación revisaremos la estimación ML en la distribución t usando el algoritmo EM.

Estimación ML en la distribución t multivariada

Es conveniente revisar la estimación ML para la distribución t como un problema de datos incompletos. Sea,

$$\boldsymbol{y}_{\mathsf{com}} = (\boldsymbol{y}^\top, \boldsymbol{\omega}^\top)^\top, \qquad \boldsymbol{y} = (\boldsymbol{y}_1^\top, \dots, \boldsymbol{y}_n^\top)^\top,$$

con $\boldsymbol{\omega} = (\omega_1, \dots, \omega_n)^\top$ representa los datos perdidos. De este modo,

$$\ell_{\mathsf{c}}(\boldsymbol{\theta}; \boldsymbol{y}_{\mathsf{com}}) = \log p(\boldsymbol{y}|\boldsymbol{\omega}; \boldsymbol{\theta}) + \log p(\boldsymbol{\omega}; \boldsymbol{\theta}),$$

donde

$$\log p(\boldsymbol{y}|\boldsymbol{\omega}; \boldsymbol{\theta}) = -\frac{np}{2} \log 2\pi - \frac{n}{2} \log |\boldsymbol{\Sigma}| - \frac{1}{2} \sum_{i=1}^n \omega_i (\boldsymbol{y}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu})$$

$$\log p(\boldsymbol{\omega}; \boldsymbol{\theta}) = -n \log \Gamma\Big(\frac{\nu}{2}\Big) + \frac{n\nu}{2} \log \Big(\frac{\nu}{2}\Big) + \frac{\nu}{2} \sum_{i=1}^n (\log \omega_i - \omega_i) - \sum_{i=1}^n \log \omega_i$$

Por otro lado, se puede mostrar de forma sencilla que la distribución condicional de $\omega_i|\boldsymbol{y}_i$ es

$$\omega_i|\boldsymbol{y}_i \sim \mathsf{Gamma}\Big(\frac{\nu+p}{2}, \frac{\nu+u_i}{2}\Big), \qquad i=1,\dots,n,$$

y sigue que

$$\mathsf{E}(\omega_i|\boldsymbol{y}_i) = \frac{\nu+p}{\nu+u_i}.$$

De este modo,

$$\omega_i^{(k)} = \mathsf{E}(\omega_i|\boldsymbol{y}_i, \boldsymbol{\theta}^{(k)}) = \frac{\nu+p}{\nu+u_i^{(k)}}, \qquad i=1,\dots,n,$$

donde $u_i^{(k)} = (\boldsymbol{y}_i - \boldsymbol{\mu}^{(k)})^\top \{\boldsymbol{\Sigma}^{(k)}\}^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}^{(k)})$.

Estimación ML en la distribución t multivariada

Por tanto, la parte relevante de la función $Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)})$ está dada por (estamos asumiendo ν conocido):

$$Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)}) = -\frac{n}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{i=1}^{n}\omega_i^{(k)}(\boldsymbol{y}_i-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i-\boldsymbol{\mu}).$$

Diferenciando con relación a $\boldsymbol{\mu}$ y $\boldsymbol{\Sigma}$ y resolviendo la condición de primer orden, tenemos

$$\boldsymbol{\mu}^{(k+1)} = \frac{1}{\sum_{i=1}^{n}\omega_i^{(k)}}\sum_{i=1}^{n}\omega_i^{(k)}\boldsymbol{y}_i,$$

$$\boldsymbol{\Sigma}^{(k+1)} = \frac{1}{n}\sum_{i=1}^{n}\omega_i^{(k)}(\boldsymbol{y}_i-\boldsymbol{\mu}^{(k+1)})(\boldsymbol{y}_i-\boldsymbol{\mu}^{(k+1)})^\top.$$

EX UMBRA IN SOLEM

Estimación ML en la distribución t multivariada

Algoritmo 3: Algoritmo EM en la t multivariada.

Entrada: Conjunto de datos observados $\boldsymbol{y}_1, \dots, \boldsymbol{y}_n$ y estimación inicial $\boldsymbol{\theta}^{(0)}$.
Salida : Estimación ML de $\boldsymbol{\mu}$ y $\boldsymbol{\Sigma}$.

begin

Paso-E: para $\boldsymbol{\theta}^{(k)}$, calcular:

$$\omega_i^{(k)} = \frac{\nu + p}{\nu + u_i^{(k)}}, \qquad i = 1, \dots, n.$$

Paso-M: actualizar $\boldsymbol{\mu}^{(k+1)}$ y $\boldsymbol{\Sigma}^{(k+1)}$ como:

$$\boldsymbol{\mu}^{(k+1)} = \frac{1}{\sum_{i=1}^n \omega_i(\boldsymbol{\theta}^{(k)})} \sum_{i=1}^n \omega_i(\boldsymbol{\theta}^{(k)}) \boldsymbol{x}_i, \tag{3}$$

$$\boldsymbol{\Sigma}^{(k+1)} = \frac{1}{n} \sum_{i=1}^n \omega_i(\boldsymbol{\theta}^{(k)})(\boldsymbol{x}_i - \boldsymbol{\mu}^{(k+1)})(\boldsymbol{x}_i - \boldsymbol{\mu}^{(k+1)})^\top. \tag{4}$$

Iterar entre pasos-E y M hasta alcanzar convergencia.

end

A la convergencia del algoritmo hacemos $\boldsymbol{\mu} = \widehat{\boldsymbol{\mu}}$ y $\boldsymbol{\Sigma} = \widehat{\boldsymbol{\Sigma}}$.

Una curiosa propiedad de la distribución t multivariada

Desde (4), debemos tener que a la convergencia del algoritmo:

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n} \widehat{\omega}_i(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}})(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}})^\top,$$

así premultiplicando por $\widehat{\boldsymbol{\Sigma}}^{-1}$ y aplicando traza, tenemos:

$$\begin{aligned} \operatorname{tr} \boldsymbol{I}_p &= \frac{1}{n}\sum_{i=1}^{n} \widehat{\omega}_i \operatorname{tr} \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}})(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}})^\top \\ p &= \frac{1}{n}\sum_{i=1}^{n} \widehat{\omega}_i \widehat{u}_i, \qquad (5) \end{aligned}$$

usando la función de pesos asociada a la distribución t, tenemos que:

$$\nu + p = \widehat{\omega}_i(\nu + \widehat{u}_i) = \widehat{\omega}_i\nu + \widehat{\omega}_i\widehat{u}_i,$$

promediando y usando (5), lleva a:

$$\nu + p = \nu\,\frac{1}{n}\sum_{i=1}^{n}\widehat{\omega}_i + p, \qquad \Rightarrow \qquad \frac{1}{n}\sum_{i=1}^{n}\widehat{\omega}_i = 1.$$

Estimación ML en la distribución t multivariada

La consideración anterior llevó a Kent, Tyler y Vardi (1994)[3] a proponer la siguiente variante del Algoritmo 3:

Algoritmo 4: Algoritmo EM en la t multivariada.

Entrada: Conjunto de datos observados $\boldsymbol{y}_1, \dots, \boldsymbol{y}_n$ y estimación inicial $\boldsymbol{\theta}^{(0)}$.
Salida : Estimación ML de $\boldsymbol{\mu}$ y $\boldsymbol{\Sigma}$.

begin

Paso-E: para $\boldsymbol{\theta}^{(k)}$, calcular:

$$\omega_i^{(k)} = \frac{\nu + p}{\nu + u_i^{(k)}}, \qquad i = 1, \dots, n.$$

Paso-M: actualizar $\boldsymbol{\mu}^{(k+1)}$ y $\boldsymbol{\Sigma}^{(k+1)}$ como:

$$\boldsymbol{\mu}^{(k+1)} = \frac{1}{\sum_{i=1}^n \omega_i(\boldsymbol{\theta}^{(k)})} \sum_{i=1}^n \omega_i(\boldsymbol{\theta}^{(k)}) \boldsymbol{x}_i,$$

$$\boldsymbol{\Sigma}^{(k+1)} = \frac{1}{\sum_{i=1}^n \omega_i(\boldsymbol{\theta}^{(k)})} \sum_{i=1}^n \omega_i(\boldsymbol{\theta}^{(k)})(\boldsymbol{x}_i - \boldsymbol{\mu}^{(k+1)})(\boldsymbol{x}_i - \boldsymbol{\mu}^{(k+1)})^\top.$$

Iterar entre pasos-E y M hasta alcanzar convergencia.

end

Posteriormente, Liu, Rubin y Wu (1998)[4] identificaron esta variante en la clase de algoritmos EM (PX-EM) de parámetros-expandidos.

[3] Communications in Statistics - Simulation and Computation **23**, 441-453.
[4] Biometrika **85**, 755-770.

Cuando los grados de libertad ν son desconocidos, debemos añadir la siguiente sub-etapa al Paso-M del Algoritmo 3 ó 4.

$$\nu^{(k+1)} = \arg\max_{\nu} \, Q_2(\nu; \boldsymbol{\theta}^{(k)}),$$

donde

$$\begin{aligned} Q_2(\nu; \boldsymbol{\theta}^{(k)}) = \frac{n\nu}{2} \log\Big(\frac{\nu}{2}\Big) - n \log\Gamma\Big(\frac{\nu}{2}\Big) + \frac{n\nu}{2}\Big\{\frac{1}{n}\sum_{i=1}^{n}(\log\omega_i^{(k)} - \omega_i^{(k)}) \\ + \psi\Big(\frac{\nu^{(k)}+p}{2}\Big) - \log\Big(\frac{\nu^{(k)}+p}{2}\Big)\Big\}, \end{aligned}$$

con $\psi(z) = \mathrm{d}\log\Gamma(z)/\,\mathrm{d}z$ la función digama.

Observación:

Podemos actualizar $\nu^{(k+1)}$ usando un método de optimización uni-dimensional.

Referencias bibliográficas

Dempster, A.P., Laird, N.M., Rubin, D.B. (1977).
Maximum likelihood from incomplete data via the EM algorithm (with discussion)
Journal of the Royal Statistical Society, Series B **39**, 1-38.

Kent, J.T., Tyler, D.E., Vardi, Y. (1994).
A curious likelihood identity for the multivariate t-distribution.
Communication in Statistics: Simulation and Computation **23**, 441-453.

Lange, K., Sinsheimer, J.S. (1993).
Normal/independent distributions and their applications in robust regression.
Journal of Computational and Graphical Statistics **2**, 175-198.

Little, R.J.A. (1988).
Robust estimation of the mean and covariance matrix from data with missing values.
Applied Statistics **37**, 23-38.

Liu, C., Rubin, D.B., Wu, Y.N. (1998).
Parameter expansion to accelerate EM: The PX-EM algorithm.
Biometrika **85**, 775-770.