

# MAT-269: Sesión 8

## Estimación ML bajo distribuciones de contornos elípticos III

Felipe Osorio

[fosorios.mat.utfsm.cl](mailto:fosorios.mat.utfsm.cl)

Departamento de Matemática, UTFSM



## Definición 1

Sea  $\boldsymbol{\mu} \in \mathbb{R}^p$ ,  $\boldsymbol{\Sigma}$  matriz  $p \times p$  definida positiva y  $H$  función de distribución de una variable aleatoria positiva,  $W$ . Entonces, se dice que el vector aleatorio  $\boldsymbol{x}$  sigue una **distribución de mezcla de escala normal** si su función de densidad asume la forma:

$$f(\boldsymbol{x}) = |2\pi\boldsymbol{\Sigma}|^{-1/2} \int_0^\infty \omega^{p/2} \exp(-\omega u/2) dH(\omega),$$

donde  $u = (\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})$  y anotamos  $\boldsymbol{x} \sim \text{SMN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}; H)$ .

### Observación:

Un vector aleatorio  $\boldsymbol{x} \sim \text{SMN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}; H)$  admite la representación:

$$\boldsymbol{x} \stackrel{d}{=} \boldsymbol{\mu} + W^{-1/2} \boldsymbol{z}, \tag{1}$$

donde  $\boldsymbol{z} \sim N_p(\mathbf{0}, \boldsymbol{\Sigma})$  y  $W \sim H(\delta)$  son **independientes**.



### Ejemplo 1: Distribución Slash

Un vector aleatorio  $\boldsymbol{x}$  tiene distribución **Slash** si su función de densidad es de la forma:

$$f(\boldsymbol{x}) = \nu |2\pi \boldsymbol{\Sigma}|^{-1/2} \int_0^1 \omega^{p/2+\nu-1} \exp(-\omega \boldsymbol{x}' \boldsymbol{x} / 2) d\omega.$$

Tenemos que  $h(\omega) = \nu \omega^{\nu-1}$ , para  $\omega \in (0, 1)$  y  $\nu > 0$ . Es decir,  $W \sim \text{Beta}(\nu, 1)$ .

### Ejemplo 2: Distribución Exponencial-Potencia

Se dice que un vector aleatorio  $\boldsymbol{x}$  tiene distribución **Exponencial-Potencia** (Gómez, Gómez-Villegas y Marín, 1988)<sup>1</sup>, si su función de densidad es dada por:

$$f(\boldsymbol{x}) = \frac{p \Gamma(\frac{p}{2}) \pi^{-p/2}}{\Gamma(1 + \frac{p}{2\lambda}) 2^{1 + \frac{p}{2\lambda}}} |\boldsymbol{\Sigma}|^{-1/2} \exp(-u^\lambda / 2), \quad \lambda > 0.$$

en cuyo caso anotamos  $\boldsymbol{x} \sim \text{PE}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \lambda)$ . Debemos destacar que la distribución de la **variable mezcladora**  $W$  tiene una **representación en series** y es de poco interés práctico.

---

<sup>1</sup>Esta familia pertenece a la clase SMN cuando  $\lambda \in (0, 1]$ .



### Observación:

La representación estocástica en (1), puede ser escrita de forma equivalente, como:

$$\mathbf{x}|W \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}/\omega), \quad W \sim H(\delta). \quad (2)$$

Esta representación permite, por ejemplo

$$\begin{aligned} E(\mathbf{x}) &= E(E(\mathbf{x}|W)) = \boldsymbol{\mu} \\ \text{Cov}(\mathbf{x}) &= E(\text{Cov}(\mathbf{x}|W)) + \text{Cov}(E(\mathbf{x}|W)) = E(W^{-1})\boldsymbol{\Sigma}. \end{aligned}$$

Además, la formulación condicional en (2) es muy útil para:

- ▶ Generación de dígitos pseudo-aleatorios.
- ▶ Estimación ML usando el algoritmo EM.



### Ejemplo 3: Distribución $t$ multivariada

Para  $\mathbf{x} \sim t_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$ , con  $\nu > 0$ , podemos escribir

$$\mathbf{x}|W \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}/\omega), \quad W \sim \text{Gamma}(\nu/2, \nu/2),$$

es decir,

$$h(\omega; \nu) = \frac{(\nu/2)^{\nu/2} \omega^{\nu/2-1}}{\Gamma(\nu/2)} \exp(-\nu\omega/2).$$

### Ejemplo 4: Distribución normal contaminada

Considere  $\mathbf{x} \sim \text{CN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \epsilon, \gamma)$  (Little, 1988) donde  $0 \leq \epsilon \leq 1$  denota el **porcentaje de contaminación** y  $0 < \gamma < 1$  corresponde a un **factor de inflación de escala**. En este caso,

$$h(\omega; \boldsymbol{\delta}) = \begin{cases} \epsilon, & \omega = \gamma \\ 1 - \epsilon & \omega = 1 \end{cases},$$

con  $\boldsymbol{\delta} = (\epsilon, \gamma)^\top$ .



# 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”**.



## Formulación de datos aumentados:

Sea  $\mathbf{Y}_{\text{obs}}$  vector de **datos observados** con función de densidad  $f(\mathbf{y}_{\text{obs}}; \boldsymbol{\theta})$ .

El objetivo es aumentar los datos observados  $\mathbf{Y}_{\text{obs}}$  con variables latentes  $\mathbf{Y}_{\text{mis}}$  (**datos perdidos**). Esto es, se considera el vector de **datos completos**

$$\mathbf{Y}_{\text{com}} = (\mathbf{Y}_{\text{obs}}^{\top}, \mathbf{Y}_{\text{mis}}^{\top})^{\top},$$

tal que la densidad  $f(\mathbf{y}_{\text{com}}; \boldsymbol{\theta})$  sea **simple**.



## Algoritmo EM (Dempster, Laird y Rubin, 1977)<sup>2</sup>

El algoritmo EM es útil cuando la función de log-verosimilitud

$$\begin{aligned}\ell_o(\boldsymbol{\theta}; \mathbf{Y}_{\text{obs}}) &= \log f(\mathbf{y}_{\text{obs}}; \boldsymbol{\theta}) \\ &= \log \int f(\mathbf{y}_{\text{com}}; \boldsymbol{\theta}) d\mathbf{y}_{\text{mis}},\end{aligned}$$

es difícil de maximizar directamente.

El algoritmo EM es un **procedimiento iterativo** que permite realizar la estimación ML basandose en la **log-verosimilitud de datos completos**:

$$\ell_c(\boldsymbol{\theta}; \mathbf{Y}_{\text{com}}) = \log f(\mathbf{y}_{\text{com}}; \boldsymbol{\theta}).$$

---

<sup>2</sup>Journal of the Royal Statistical Society, Series B **39**, 1-38.





## Algoritmo EM (Esperanza-Maximización)

El algoritmo EM permite obtener los MLE en **problemas con datos incompletos** por medio de las etapas:

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

$$\begin{aligned} Q(\theta|\theta^{(k)}) &= E\{\ell_c(\theta; \mathbf{Y}_{\text{com}}) | \mathbf{Y}_{\text{obs}}, \theta^{(k)}\} \\ &= \int \ell_c(\theta; \mathbf{Y}_{\text{com}}) f(\mathbf{y}_{\text{mis}} | \mathbf{y}_{\text{obs}}; \theta^{(k)}) d\mathbf{y}_{\text{mis}}. \end{aligned} \quad (3)$$

*Paso M:* determinar  $\theta^{(k+1)}$  como

$$\theta^{(k+1)} = \arg \max_{\theta} Q(\theta | \theta^{(k)}). \quad (4)$$



Dempster, Laird y Rubin (1977) definieron el **Algoritmo EM generalizado (GEM)**, mediante la siguiente modificación del paso M:

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

$$Q(\theta^{(k+1)} | \hat{\theta}^{(k)}) > Q(\theta^{(k)} | \theta^{(k)}).$$

*Sugerencia*: considerar **sólo un** paso Newton en la optimización de  $Q(\theta | \theta^{(k)})$ .



## Teorema (Dempster, Laird y Rubin, 1977)

Todo algoritmo EM o GEM **incrementa la log-verosimilitud de datos observados**  $\ell_o(\boldsymbol{\theta}; \mathbf{Y}_{\text{obs}})$  en cada iteración, esto es,

$$\ell_o(\boldsymbol{\theta}^{(k+1)}; \mathbf{Y}_{\text{obs}}) \geq \ell_o(\boldsymbol{\theta}^{(k)}; \mathbf{Y}_{\text{obs}}).$$

## 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  $\ell_o(\boldsymbol{\theta}; \mathbf{Y}_{\text{obs}})$ .



## *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.<sup>3</sup>
- ▶ 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).

---

<sup>3</sup> puede ser **extremadamente** lento.



Sea  $\mathbf{x}_1, \dots, \mathbf{x}_n$  vectores aleatorios IID desde  $\text{SMN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \text{H})$ . Se llevará a cabo la **estimación ML** usando el **algoritmo EM**.

De este modo, tenemos el siguiente **modelo jerárquico**:

$$\mathbf{x}_i | W_i \sim \text{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}/\omega_i), \quad W_i \sim \text{H}(\boldsymbol{\delta}).$$

En este caso el vector de **datos completos** es  $\mathbf{x}_{\text{com}} = (\mathbf{x}^\top, \boldsymbol{\omega}^\top)^\top$ , donde

$$\mathbf{x} = (\mathbf{x}_1^\top, \dots, \mathbf{x}_n^\top)^\top, \quad \boldsymbol{\omega} = (\omega_1, \dots, \omega_n)^\top.$$

En este contexto,  $\mathbf{x}$  corresponde a los **datos observados**, mientras que  $\boldsymbol{\omega}$  serán asumidos como **datos perdidos**.



## Estimación ML usando mezclas de escala normal

Primeramente asumiremos que  $\delta$  es conocido. La función de log-verosimilitud de datos completos adopta la forma:

$$\begin{aligned}\ell_c(\boldsymbol{\theta}; \mathbf{x}_{\text{com}}) &= \log f(\mathbf{x}_{\text{com}}; \boldsymbol{\theta}) = \sum_{i=1}^n \log f(\mathbf{x}_i, \omega_i; \boldsymbol{\theta}) \\ &= \sum_{i=1}^n \log f(\mathbf{x}_i | \omega_i; \boldsymbol{\theta}) + \sum_{i=1}^n \log h(\omega_i; \boldsymbol{\delta}) \\ &= -\frac{n}{2} \log |2\pi \boldsymbol{\Sigma}| - \frac{1}{2} \sum_{i=1}^n \omega_i (\mathbf{x}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \\ &\quad - \frac{p}{2} \sum_{i=1}^n \log \omega_i + \log h^{(n)}(\boldsymbol{\omega}; \boldsymbol{\delta}),\end{aligned}$$

donde  $h^{(n)}(\boldsymbol{\omega}; \boldsymbol{\delta})$  denota la densidad conjunta para las variables de mezcla  $\boldsymbol{\omega} = (\omega_1, \dots, \omega_n)^\top$ .



Considere una estimación para  $\theta = \theta^{(k)}$ , entonces

$$Q(\theta; \theta^{(k)}) = E\{\ell_c(\theta; \mathbf{x}_{\text{com}}) | \mathbf{x}; \theta^{(k)}\} = Q_1(\theta; \theta^{(k)}) + Q_2(\delta; \theta^{(k)}),$$

donde

$$Q_1(\theta; \theta^{(k)}) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log |\Sigma| - \frac{1}{2} \sum_{i=1}^n \omega_i^{(k)} (\mathbf{x}_i - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{x}_i - \boldsymbol{\mu}),$$

$$Q_2(\theta; \theta^{(k)}) = E\{\log h^{(n)}(\boldsymbol{\omega}; \boldsymbol{\delta}) | \mathbf{x}; \theta^{(k)}\},$$

con  $\omega_i^{(k)} = E(\omega_i | \mathbf{x}_i; \theta^{(k)})$  para  $i = 1, \dots, n$ . En general la forma para la esperanza condicional requerida en el paso-E del algoritmo EM es dada por:

$$E(\omega_i | \mathbf{x}_i; \theta) = \frac{\int_0^\infty \omega_i^{p/2+1} \exp(-\omega_i u_i/2) dH(\boldsymbol{\delta})}{\int_0^\infty \omega_i^{p/2} \exp(-\omega_i u_i/2) dH(\boldsymbol{\delta})},$$

con  $u_i = (\mathbf{x}_i - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{x}_i - \boldsymbol{\mu})$ ,  $i = 1, \dots, n$ .



- ▶ **t-Student:**  $\mathbf{x} \sim t_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$ , en cuyo caso

$$E(\omega_i | \mathbf{x}_i; \boldsymbol{\theta}) = \frac{\nu + p}{\nu + u_i}.$$

- ▶ **Slash:**  $\mathbf{x} \sim \text{Slash}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$ . De este modo,

$$E(\omega_i | \mathbf{x}_i; \boldsymbol{\theta}) = \left( \frac{p + 2\nu}{u_i} \right) \frac{P_1(p/2 + \nu + 1, u_i/2)}{P_1(p/2 + \nu, u_i/2)},$$

donde

$$P_z(a, b) = \frac{b^a}{\Gamma(a)} \int_0^z t^{a-1} e^{-bt} dt,$$

es la función gama incompleta (regularizada).

- ▶ **Exponencial Potencia:**  $\mathbf{x} \sim \text{PE}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \lambda)$ , donde

$$E(\omega_i | \mathbf{x}_i; \boldsymbol{\theta}) = \lambda u_i^{\lambda-1}, \quad u_i \neq 0, \lambda \neq \frac{1}{2}.$$





## Estimación ML usando mezclas de escala normal

Finalmente, el algoritmo EM obtener los MLEs en una muestra aleatoria  $\mathbf{x}_1, \dots, \mathbf{x}_n$  desde  $\text{SMN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, H)$  adopta la forma:

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

$$\omega_i^{(k)} = \mathbb{E}(\omega_i | \mathbf{x}_i; \boldsymbol{\theta}^{(k)}), \quad i = 1, \dots, n.$$

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

$$\boldsymbol{\mu}^{(k+1)} = \frac{1}{\sum_{i=1}^n \omega_i(\boldsymbol{\theta}^{(k)})} \sum_{i=1}^n \omega_i(\boldsymbol{\theta}^{(k)}) \mathbf{x}_i, \quad (5)$$

$$\boldsymbol{\Sigma}^{(k+1)} = \frac{1}{n} \sum_{i=1}^n \omega_i(\boldsymbol{\theta}^{(k)}) (\mathbf{x}_i - \boldsymbol{\mu}^{(k+1)}) (\mathbf{x}_i - \boldsymbol{\mu}^{(k+1)})^\top. \quad (6)$$

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



## Una curiosa propiedad de la distribución t multivariada

Desde (6), debemos tener que a la convergencia del algoritmo:

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

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

$$\begin{aligned} \text{tr } \mathbf{I}_p &= \frac{1}{n} \sum_{i=1}^n \widehat{\omega}_i \text{tr } \Sigma^{-1} (\mathbf{x}_i - \widehat{\boldsymbol{\mu}})(\mathbf{x}_i - \widehat{\boldsymbol{\mu}})^\top \\ p &= \frac{1}{n} \sum_{i=1}^n \widehat{\omega}_i \widehat{u}_i, \end{aligned} \tag{7}$$

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 (7), lleva a:

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



## Estimación ML usando mezclas de escala normal

La consideración anterior llevó a Kent, Tyler y Vardi (1994)<sup>4</sup> a proponer la siguiente variante del algoritmo EM, válido para la distribución  $t$ :

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

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

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

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

Posteriormente, Liu, Rubin y Wu (1998)<sup>5</sup> identificaron esta variante en la clase de algoritmos EM (PX-EM) de parámetros-expandidos.

---

<sup>4</sup>Communications in Statistics - Simulation and Computation **23**, 441-453.

<sup>5</sup>Biometrika **85**, 755-770.



Para ejemplificar la **estimación de los parámetros de la variable de mezcla**, considere:

► **t-Student:** En este caso,

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

con  $\psi(z) = d \log \Gamma(z) / dz$  la función digama, y actualizamos  $\nu^{(k+1)}$  usando un método de optimización uni-dimensional.



- **Slash:** Tenemos que,

$$Q_2(\nu; \boldsymbol{\theta}^{(k)}) = n \log \nu + \nu \sum_{i=1}^n \mathbb{E}(\log \omega_i | \mathbf{x}_i; \boldsymbol{\theta}^{(k)}),$$

y actualizamos  $\nu^{(k+1)}$  como

$$\nu^{(k+1)} = - \left\{ \frac{1}{n} \sum_{i=1}^n \mathbb{E}(\log \omega_i | \mathbf{x}_i; \boldsymbol{\theta}^{(k)}) \right\}^{-1},$$

con

$$\mathbb{E}(\log \omega_i | \mathbf{x}_i; \boldsymbol{\theta}^{(k)}) = \psi(\nu + p/2) - \log(u_i^2/2) + \frac{\partial P_1(\nu + p/2, u_i^2/2)/\partial \nu}{P_1(\nu + p/2, u_i^2/2)}.$$



## 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.

