# MAT-269: Estimación ML bajo distribuciones de mezcla de escala normal

**Felipe Osorio**

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:

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

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

*Observación:*

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

$$\boldsymbol{x} \stackrel{\mathsf{d}}{=} \boldsymbol{\mu} + W^{-1/2}\boldsymbol{z}, \qquad (1)$$

donde $\boldsymbol{z} \sim \mathsf{N}_p(\boldsymbol{0}, \boldsymbol{\Sigma})$ y $W \sim \mathsf{H}(\boldsymbol{\delta})$ son independientes.

*Ejemplo 1: Distribución Slash*

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

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

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)[1], si su función de densidad es dada por:

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

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.

[1] 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:

$$\boldsymbol{x}|W \sim \mathsf{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}/\omega), \qquad W \sim \mathsf{H}(\boldsymbol{\delta}). \tag{2}$$

Esta representación permite, por ejemplo

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

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

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

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 $\boldsymbol{x} \sim \mathsf{CN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \epsilon, \gamma)$ (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 $\boldsymbol{\delta} = (\epsilon, \gamma)^\top$.

EX UMBRA IN SOLEM

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

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

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

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

# Algoritmo EM (Dempster, Laird y Rubin, 1977)[2]

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

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

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:

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

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

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

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

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) &= \mathsf{E}\{\ell_{\mathsf{c}}(\boldsymbol{\theta};\boldsymbol{Y}_{\mathsf{com}})|\boldsymbol{Y}_{\mathsf{obs}},\boldsymbol{\theta}^{(k)}\} \\ &= \int \ell_c(\boldsymbol{\theta};\boldsymbol{Y}_{\mathsf{com}}) f(\boldsymbol{y}_{\mathsf{mis}}|\boldsymbol{y}_{\mathsf{obs}};\boldsymbol{\theta}^{(k)})\,\mathsf{d}\boldsymbol{y}_{\mathsf{mis}}. \end{aligned} \tag{3}$$

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

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

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

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

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

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

**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.[3]
- 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).

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

# Estimación ML usando mezclas de escala normal

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

De este modo, tenemos el siguiente modelo jerárquico:

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

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

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

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

## Estimación ML usando mezclas de escala normal

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

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

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 $\boldsymbol{\theta} = \boldsymbol{\theta}^{(k)}$, entonces

$$Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)}) = \mathsf{E}\{\ell_{\mathsf{c}}(\boldsymbol{\theta}; \boldsymbol{x}_{\mathsf{com}})|\boldsymbol{x}; \boldsymbol{\theta}^{(k)}\} = Q_1(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)}) + Q_2(\boldsymbol{\delta}; \boldsymbol{\theta}^{(k)}),$$

donde

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

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

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

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

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

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

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

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

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

donde

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

es la función gama incompleta (regularizada).

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

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

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

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

$$\omega_i^{(k)} = \mathsf{E}(\omega_i|\boldsymbol{x}_i; \boldsymbol{\theta}^{(k)}), \qquad i = 1, \ldots, n.$$

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

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

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

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 (6), 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:

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

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 (7), lleva a:

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

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

*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:

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

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

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

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

# Estimación ML usando mezclas de escala normal

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

- $t$-Student: En este caso,

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

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

- Slash: Tenemos que,

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

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

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

con

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

## 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.


EX UMBRA IN SOLEM