1. Análisis exploratorio

Cuadro 1: Estadísticos Iniciales

variable	media	varianza
Cred_perd	0.67	1.36
Cred_PF	1.56	9.04
Electronico	16.23	1565.86
Ing_tot	5175.00	4990698327.63
Monto_prom	1296.89	7208400.90
Saldo	640.16	236844874.62

Cuadro 2: Summary

	Cred_perd	Cred_PF	Electronico	Ing_tot	Monto_prom	Saldo
Min. :	0	0	0	-251	0	0
1st Qu.:	0	0	1	100	500	0
Median :	0	1	7	407	800	0
Mean :	0.6707	1.56	16.23	5175	1297	640.2
3rd Qu.:	1	2	18	1631	1350	0
Max. :	59	168	3439	11024580	143881	2674250

2. Desarrollo del Modelo de mezclas

Se tiene un conjunto de datos compuesto de p variables de conteo y d variables continuas, con n observaciones.

Las variables continuas se distribuyen Normal-Multivariada, $X^c \sim N(\mu_j, \Sigma_j)$, mientras que cada una de las variables de conteo tiene una distribución Poisson, $X_l^d \sim Po(\lambda_{lj})$.

Por lo que nuestro modelo de mezclas con K componentes se define de la siguiente manera:

$$\sum_{j=1}^{K}(\pi_j \cdot N(\mu_j, \Sigma_j) \cdot \prod_{l=1}^{p} Po(\lambda_{lj})) \tag{1}$$

Ahora incluimos la variable latente z_j al modelo. Debido a que x_i sólo puede pertenecer a un componente, se definine a z_{ij} como la esperanza condicional,

$$E[z_{ij}|x] = \frac{\pi_j^{(t)} \cdot N(\mu_j, \Sigma_j) \cdot \prod_{l=1}^{p} Po(\lambda_{lj})}{\sum_{j=1}^{K} \pi_j^{(t)} \cdot N(\mu_j, \Sigma_j) \cdot \prod_{l=1}^{p} Po(\lambda_{lj})} \tag{2}$$

Para obtener la función de distribución conjunta, tenemos que $P(x, z; \theta) = P(z, \theta) \cdot P(x|z, \theta)$ con $\theta = (\pi, \mu, \Sigma, \lambda)$, entonces

$$P(z,\theta) = \prod_{j=1}^{K} \pi_j^{z_j} \tag{3}$$

$$P(x|z_j,\theta) = (N(\mu_j,\Sigma_j) \cdot \prod_{l=1}^{p} Po(\lambda_{lj}))^{z_j} \tag{4}$$

$$P(x,z;\theta) = \prod_{j=1}^{K}\prod_{i=1}^{n} \left(\pi_j \cdot N(\mu_j,\Sigma_j) \cdot \prod_{l=1}^{p} Po(\lambda_{lj})\right)^{z_{ij}} \tag{5}$$

De manera que al desarrollarlo obtenemos lo siguiente,

$$L(\theta) = \prod_{j=1}^{K} \pi_j^{\bar{z}_j} \cdot |\Sigma_j^{-1}|^{-\bar{z}_j} \exp\left\{-\frac{1}{2} trace\left[\Sigma_j^{-1}\left(\bar{z}_j(\bar{x}_j-\mu_j)(\bar{x}_j-\mu_j)' + \sum_{i=1}^{n}(x_i-\bar{x}_j)(x_i-\bar{x}_j)'\right)\right]\right\} \tag{6}$$

$$\cdot \left[\prod_{l=1}^{p} \lambda_{lj}^{\bar{z}_j\bar{x}_{lj}} exp\left\{-\bar{z}_j\lambda_{lj}\right\}\right]$$

(El desarrollo completo para llegar a la ecuación 6 se encuentra en el pdf anexo que está a mano)

La distribución posterior está compuesta de la función de verosimilitud y las funciones previas de los parámetros. Así que primero se definen las distribuciones previas de los parámetros y sus hiperparámetros correspondientes:

- $\Sigma_j \sim W^{-1}(\Lambda_j, v_j)$ con una función de densidad,

$$P(\Sigma_j|\Lambda_j,v_j) \propto |\Sigma_j^{-1}|^{-(v_j+d+1)/2} \exp\left\{-\frac{1}{2} trace\left[\Sigma_j^{-1}\Lambda_j\right]\right\}$$

- $\mu_j|\Sigma_j \sim N(\varepsilon_j, \frac{\Sigma_j}{n_j})$ con función de densidad,

$$P(\mu_j|\varepsilon_j, \frac{\Sigma_j}{n_j}) \propto |\Sigma_j^{-1}|^{-\frac{1}{2}} \exp\left\{-\frac{n_j}{2} trace\left[(\mu_j - \varepsilon_j)'\Sigma_j^{-1}(\mu_j - \varepsilon_j)\right]\right\}$$

- $\lambda_{lj} \sim G(a_{lj}, S_{lj})$ con función de densidad,

$$P(\lambda_{lj}|a_{lj}, S_{lj}) \propto \lambda_{lj}^{a_{lj}} \exp\{-S_{lj}\lambda_{lj}\}$$

- $\pi \sim Dir(\alpha_1 = 1/k, ..., \alpha_k = 1/k)$ con función de densidad,

$$P(\pi|\alpha_1, ..., \alpha_k) \propto \prod_{j=1}^{K} \pi_j^{\alpha_j}$$

Tomando en cuenta lo anterior, se tiene la siguiente función de distribución posterior,

$$P(\theta|x, z) = L(\theta) \cdot P(\pi|\alpha_1, ..., \alpha_k) \cdot \prod_{j=1}^{K}\left[P(\Sigma_j|\Lambda_j, v_j) \cdot P(\mu_j|\varepsilon_j, \frac{\Sigma_j}{n_j}) \cdot \prod_{l=1}^{p} P(\lambda_{lj}|a_{lj}, S_{lj})\right] \tag{7}$$

Con el fin de ser más claros en el desarrollo de la distribución posterior, separaremos la parte de las variables continuas y la parte de las variables de conteo.

1. Variables continuas

$$
\begin{aligned}
& L(\theta) \cdot P(\Sigma_j|\Lambda_j, v_j) \cdot P(\mu_j|\varepsilon_j, \frac{\Sigma_j}{n_j}) \qquad (8)\\
& \propto |\Sigma_j^{-1}|^{-\bar{z}_j} \exp\left\{-\frac{1}{2} trace\left[\Sigma_j^{-1}\left(\bar{z}_j(\bar{x}_j-\mu_j)(\bar{x}_j-\mu_j)' + \sum_{i=1}^{n}(x_i-\bar{x}_j)(x_i-\bar{x}_j)'\right)\right]\right\}\\
& \cdot |\Sigma_j^{-1}|^{-\frac{1}{2}} \exp\left\{-\frac{n_j}{2} trace\left[(\mu_j-\varepsilon_j)'\Sigma_j^{-1}(\mu_j-\varepsilon_j)\right]\right\}\\
& \cdot |\Sigma_j^{-1}|^{-(v_j+d+1)/2} \exp\left\{-\frac{1}{2} trace\left[\Sigma_j^{-1}\Lambda_j\right]\right\}\\
& = |\Sigma_j^{-1}|^{-(\tilde{v}_j+d+1)/2} \exp\left\{-\frac{1}{2} trace\left[\Sigma_j^{-1}\tilde{\Lambda}_j\right]\right\}\\
& \cdot |\Sigma_j^{-1}|^{-\frac{1}{2}} \exp\left\{-\frac{\tilde{n}_j}{2} trace\left[(\mu_j-\tilde{\varepsilon}_j)'\Sigma_j^{-1}(\mu_j-\tilde{\varepsilon}_j)\right]\right\}\\
& = N(\mu_j|\tilde{\varepsilon}_j, \frac{\Sigma_j}{\tilde{n}_j}) \cdot W^{-1}(\Sigma_j|\tilde{v}_j, \tilde{\Lambda}_j)
\end{aligned}
$$

2. Variables de conteo

$$
\begin{aligned}
& L(\theta) \cdot \prod_{l=1}^{p} P(\lambda_{lj}|a_{lj}, S_{lj}) \qquad (9)\\
& \propto \prod_{l=1}^{p} \lambda_{lj}^{\bar{z}_j\bar{x}_{lj}} exp\left\{-\bar{z}_j\lambda_{lj}\right\} \cdot \lambda_{lj}^{a_{lj}} \exp\left\{-S_{lj}\lambda_{lj}\right\}\\
& = \prod_{l=1}^{p} \lambda_{lj}^{\bar{z}_j\bar{x}_{lj}+a_{lj}} \cdot exp\left\{-\lambda_{lj}(\bar{z}_j+S_{lj})\right\}\\
& = \prod_{l=1}^{p} Ga(\tilde{a}_{lj}, \tilde{S}_{lj})
\end{aligned}
$$

3. Variable latente

$$
\begin{aligned}
& L(\theta) \cdot P(\pi|\alpha_1, ..., \alpha_k) \qquad (10) \\
& \propto \prod_{j=1}^{K} \pi_j^{\bar{z}_j} \cdot \pi_j^{\alpha_j} \\
& = Dir(\tilde{\alpha}_1 = \bar{z}_1 + \alpha_1, ..., \tilde{\alpha}_1 = \bar{z}_1 + \alpha_1)
\end{aligned}
$$

3. Algoritmo

Para desarrollar el algoritmo del modelo basado en mezclas se toma como referencia el Algortimo 6 propuesto por Liang en su artículo *'On simulation methods for two component normal mixture model "*

1. Obtener los valores iniciales $\{\theta_j^{(0)} = (\pi_j^{(0)}, \mu_j^{(0)}, \Sigma_j^{(0)}, \lambda_j^{(0)})\}_{j=1}^{K}$ de los parámetros con base en las distribuciones previas definidas en la sección anterior.

   - $\Sigma_j \sim W^{-1}(\Lambda_j, v_j)$ donde Λ_j es la matriz de covarianzas de los datos observados y v_j es el número de variables continuas mas uno, $(d+1)$.
   - $\mu_j|\Sigma_j \sim N(\varepsilon_j, \frac{\Sigma_j}{n_j})$ donde ε_j es un vector de las medias observas de las variables continuas, y n_j para fines prácticos se tomo como el

número total de observaciones n entre el número de componentes K.

- $\lambda_{lj} \sim G(a_{lj}, S_{lj})$ donde S_{lj} es la varianza observada de la l-ésima variable discreta y a_{lj} el parámetro de forma.
- $\pi \sim Dir(\alpha_1 = 1/k, ..., \alpha_k = 1/k)$

2. Repetir para $t = 1, 2, ..., T$, siendo T el número de iteraciones.

   a) Generar $z_{ij}^{(t)} \in \{0, 1\}$ para $i = 1, ..., n$, donde n es el número de observaciones, y

$$z_{ij}^{(t)} \sim Ber\left(\frac{\pi_j^{(t)} \cdot N(\mu_j, \Sigma_j) \cdot \prod_{l=1}^{p} Po(\lambda_{lj})}{\sum_{j=1}^{K} \pi_j^{(t)} \cdot N(\mu_j, \Sigma_j) \cdot \prod_{l=1}^{p} Po(\lambda_{lj})}\right)$$

   b) Generar las distribuciones posteriores de los parámetros para cada componente j, con $j = 1, ..., K$.

   - $\Sigma_j^{(t+1)} \sim W^{-1}(\tilde{\Lambda}_j, \tilde{v}_j)$, se definen $\tilde{\Lambda}_j$ y $\tilde{v}_j$ como,

$$\tilde{\Lambda}_j = \Lambda_j + \sum_{i=1}^{n} z_{ij}(x_i - \bar{x}_j)(x_i - \bar{x}_j)' + \frac{n_j \bar{z}_j}{n_j + \bar{z}_j}(\bar{x}_j - \varepsilon_j)(\bar{x}_j - \varepsilon_j)'$$

$$\tilde{v}_j = v_j + \bar{z}_j$$

     donde $\bar{z}_j = \sum_{i=1}^{n} z_{ij}$ y $\bar{x}_j = \sum_{i=1}^{n} \frac{z_{ij} x_i}{\bar{z}_j}$

- $\mu_j^{(t+1)} \sim N(\tilde{\varepsilon}_j, \frac{\Sigma_j}{\tilde{n}_j})$ donde,

$$\tilde{\varepsilon}_j = \frac{\bar{z}_j \bar{x}_j + n_j \varepsilon_j}{\bar{z}_j + n_j}$$

$$\tilde{n}_j = \bar{z}_j + n_j$$

- $\lambda_{lj}^{(t+1)} \sim G(\tilde{a}_j, \tilde{S}_j)$, se definen $\tilde{a}_{lj}$ y $\tilde{S}_{lj}$ como,

$$\tilde{a}_{lj} = \bar{z}_j \bar{x}_{lj} + a_{lj}$$

$$\tilde{S}_{lj} = \bar{z}_j + S_{lj}$$

  donde $\bar{x}_{lj} = \sum_{i=1}^{n} \frac{z_{ij} x_{li}}{\bar{z}_j}$

- $\pi_j^{(t+1)} \sim Dir(\bar{z}_1 + \alpha_1, ..., \bar{z}_k + \alpha_k)$