Universidad Nacional Autónoma de México

Facultad de Ciencias

Estimación de datos faltantes con el Algoritmo EM

T E S I S

QUE PARA OBTENER EL TÍTULO DE:
ACTUARIO

PRESENTA:
MARÍA FERNANDA LERDO DE TEJADA PAVÓN

DIRECTOR DE TESIS:
M. EN C. JOSÉ SALVADOR ZAMORA MUÑOZ

2014

Dedicado a
mi familia y mis amigos

Índice general

Introducción

El algoritmo Esperanza-Maximización, también conocido como Algoritmo EM, permite realizar estimaciones de datos faltantes (o missing values) dentro de un conjunto de datos, vía máxima verosimilitud de una manera iterativa y más sencilla que si se utilizaran otros métodos para el mismo fin. Puesto que el proceso iterativo consta de dos pasos (Esperanza y Maximización), Dempster, Laird y Rubin (1977) nombraron al Algoritmo como EM.

El algoritmo EM puede resolver el problema de los datos faltantes partiendo de determinados valores iniciales y actualizando el valor de la estimación en cada iteración hasta que el algoritmo converge al valor óptimo.

Existe una amplia gama de problemas que el algoritmo puede resolver además del ya mencionado, problema de los datos faltantes. Algunos casos donde la falta de datos es evidente como datos agrupados, truncados o censurados; y otros donde la falta de datos no es tan evidente como modelos de mezclas finitas, estimación de hiperparámetros (modelos jerárquicos), análisis de factores, etc.

Computacionalmente el algoritmo es sencillo y rápido en comparación con otros métodos, lo cual ha permitido que se utilice en diversas ramas de aplicación de la estadística para la resolución de problemas. En algunas situaciones más complejas el algoritmo llega a ser lento para converger a la solución buscada, lo que ha llevado a que se hayan creado versiones modificadas y extendidas del mismo.

En este trabajo se presentarán los casos más comunes en donde el algoritmo se ha utilizado y se demostrará su eficiencia en comparación con otros métodos estadísticos utilizados para el mismo fin.

Capítulo 1

Datos Faltantes

La falta de datos es un problema frecuente, por ejemplo, cuando los datos no son confiables debido a fallas en los instrumentos de medición, se pierde o daña la información muestral, la información no es reportada o es reportada de una manera incorrecta, entre otros. En estos casos es necesario evaluar todas aquellas omisiones que puedan existir y qué solución se les dará previo a comenzar a trabajar con la información.

Si se tienen demasiados valores faltantes o si los datos omitidos en una variable dependen de otra o más variables se puede optar por descartarlos por completo de la información, pero esto podría conllevar a resultados sesgados o incluso inválidos al momento de realizar el análisis.

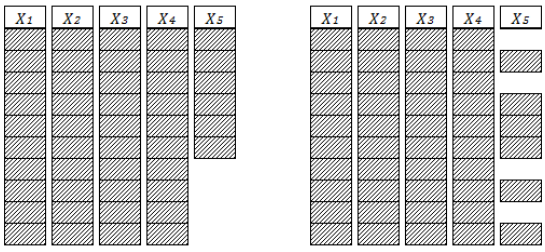
Existen diversos métodos para rellenar estos huecos en los datos con la información apropiada. Los métodos más sencillos asignan un valor fijo como la media o la mediana, otros rellenan con un valor existente de manera aleatoria o bien promedian los datos en una vecindad definida del valor faltante. Estos métodos no proporcionan una solución óptima al problema de la falta de datos y tienden a sesgar la información, por lo que los resultados posteriores al análisis podrían no ser del todo confiables.

Para obtener mejores resultados en el análisis de la información se debe hacer un estudio previo de la escala de las variables y su distribución así como del patrón que siguen los datos faltantes. Es conveniente utilizar distintos métodos para rellenar la información y realizar una evaluación de los mismos para seleccionar el método que mejor se ajusta a los datos y que minimice los posibles errores.

1.1. Patrones de Datos Faltantes

Para seleccionar un método adecuado para imputación de datos faltantes es importante encontrar el patrón que sigue la ausencia de datos. En la práctica los conjuntos de datos suelen tener un arreglo rectangular, donde los renglones corresponden a las unidades observadas y las columnas corresponden a las variables o características. Siguiendo esta línea, existen diversas formas de clasificar el patrón de ausencia de datos. Los patrones de ausencia más comunes se definen como sigue:

1. **Univariado**
   El patrón univariado es el caso más simple de presencia de valores perdidos y se identifica cuando se tienen observaciones ausentes únicamente en una variable dentro de un conjunto de datos. La ausencia de registros puede ignorarse si éstos presentan un comportamiento aleatorio, es decir, pueden considerarse como una submuestra aleatoria de la población y el análisis puede realizarse con los valores observados. Sin embargo, si la ausencia de registros depende del valor de la misma variable, un análisis sólo con los datos observados que no tome en cuenta este hecho, conduciría a un sesgo en los resultados.



(a) No aleatorio (b) Aleatorio

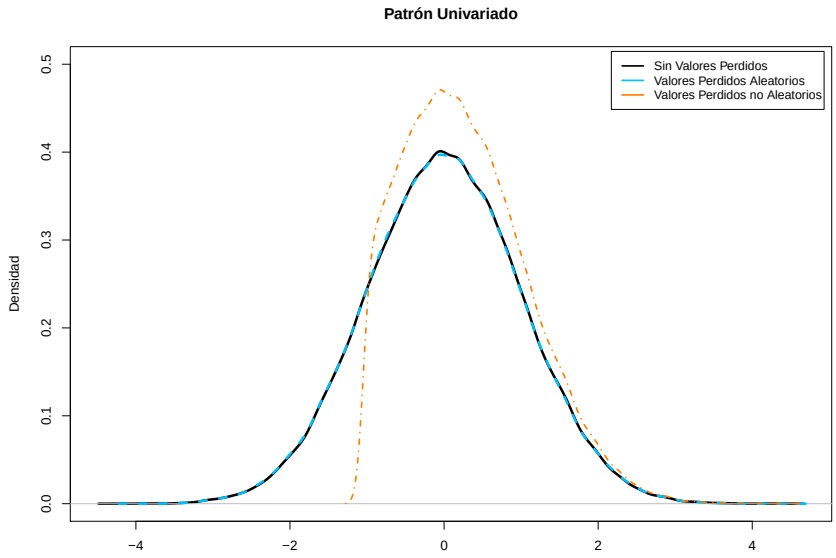
Figura 1.1: Patrón Univarado

Por ejemplo, supongamos que se tiene un conjunto de $100,000$ observaciones que provienen de una distribución Normal con media 0 y varianza 1. Si aleatoriamente se extrae el $20\,\%$ de la población y se

recalculan los parámetros, se obtiene que estos son muy similares a los de la población original.
Si por otra parte, se genera un nuevo conjunto de datos donde se excluye el 14 % de la población si el valor de las observaciones es menor a -1.5, los parámetros resultantes son significativamente diferentes, es decir, existe un sesgo en la información. A continuación se presenta un comparativo de ambos casos:

Parámetro	Sin Valores Perdidos	Valores Perdidos Aleatorios	Valores Perdidos no Aleatorios
Media	0.0037558	0.0016446	0.2694385
Varianza	0.9952137	0.9930114	0.6428224

Gráficamente las densidades original y con valores perdidos aleatorios son muy similares, mientras que la densidad del conjunto de datos con valores ausentes no aleatorios tiene un sesgo importante:



2. **Monótono**
De acuerdo con Schaffer y Graham (2002) cuando todas las variables o grupos de variables de un conjunto de datos, por decir $Y_1 \dots Y_p$,

se ordenan de tal modo que si Y_j es faltante para una observación, entonces las variables $Y_{j+1}, \ldots, Y_p$ tampoco son observadas, se dice que los valores ausentes siguen un patrón monótono.

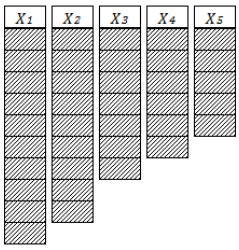


Figura 1.2: Patrón Monótono

Generalmente este patrón es común en ámbitos como la psicología, la medicina, encuestas de estudios poblacionales o en riesgo de crédito, donde se llevan a cabo estudios longitudinales o de seguimiento de una misma población a lo largo del tiempo y a partir del j-ésimo periodo comienza la pérdida de información. Los motivos de esta pérdida pueden estar asociados al fallecimiento del individuo, al cierre de un crédito o a que el encuestado no se encontraba en el momento de levantamiento de la encuesta. Este problema tradicionalmente se ha resuelto con métodos de imputación simple, que en algunos casos logran resultados aceptables,sin embargo, con el reciente descubrimiento de nuevos métodos de imputación como la imputación múltiple o el algoritmo EM se pueden lograr resultados más confiables en gran parte de las situaciones.

3. **Aleatorio**
   Cuando la ausencia de datos se presenta en una o más variables sin seguir un orden específico, entonces se tiene un patrón aleatorio.

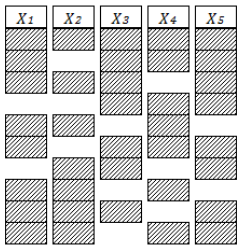


Figura 1.3: Patrón Aleatorio

4. **Parámetros no identificados**
   En este caso se tiene que la ausencia de datos se presenta en dos o más variables para tramos de observaciones excluyentes. Por ejemplo si la ausencia de datos se presenta en dos variables X_1 y X_2, las observaciones $1, \cdots, i$ estarán ausentes para la variable X_2 mientras que las restantes $i+1, \cdots, m$ estarán ausentes para la variable X_1.

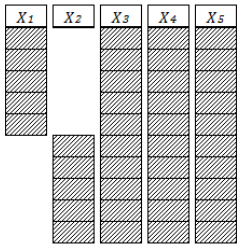


Figura 1.4: Parámetros no identificados

5. **Variables Latentes**
   Se dice que una variable es latente cuando no es observada de manera directa y tiene una relación o influencia sobre otras variables de interés. Existen numerosos métodos estadísticos que permiten encontrar la relación entre la variable latente y las demás variables, como son el análisis de factores, los modelos de mezclas, entre otros. En términos probabilísticos, las variables directamente observadas son condicionalmente independientes dada la variable latente, es decir, las variables medidas conforman un conjunto de aspectos correspondientes a una misma característica. Las variables latentes pueden ser continuas o ca-

tegóricas y donde las categorías pueden ser o no ordenadas. Un ejemplo popularmente conocido de variables latentes está ligado a la rama de la psicología, en donde se busca medir la inteligencia de un individuo a través de un conjunto de características descriptivas, sin embargo, la inteligencia de un individuo no es directamente observable, sino que a través de un conjunto de cuestionarios y modelos estadísticos se convierte en una variable medible e interpretable. Otro ejemplo ampliamente utilizado en la medicina y la fotografía es el caso de mezclas finitas, donde se asume que los sujetos provienen de subpoblaciones con distribuciones distintas dentro de una misma variable y se busca clasificarlos en una categoría latente.

1.2. Distribución

Para describir las frecuencias y los patrones de los datos faltantes, así como su relación con los datos observados y con ellos mismos, se da al comportamiento de los valores ausentes un enfoque probabilístico. De esta manera, aunque no es posible conocer de manera exacta todas las causas que originan la pérdida de datos, se puede obtener información valiosa que ayude a mitigar este problema. Supongamos que se tiene un conjunto de datos X, de los cuales X_{obs} y X_{miss} corresponden a los valores observados y faltantes respectivamente, de tal manera que $X_{obs} \bigcup X_{miss}$=X.

Rubin (1976) desarrolló una clasificación para los datos faltantes X_{miss} acuerdo a su estructura de dependencia:

- Faltante Completamente Aleatorio (MCAR)
- Faltante Aleatorio (MAR)
- Faltante No Aleatorio (MNAR)

1.2.1. Faltante Completamente Aleatorio (MCAR)

En este caso, la probabilidad de que una observación sea faltante no depende de otros datos faltantes o de los datos observados, es decir:

$$P\left(X_{miss} \mid X_{obs}, X_{miss}\right) = P\left(X_{miss}\right)$$

En el mejor de los casos se tienen observaciones perdidas del tipo MCAR. Mientras su representatividad dentro del conjunto de datos sea razonablemente baja se puede ignorar su ausencia, ya que representarían una muestra aleatoria de la población, siendo sus características heterogéneas y similares a las de la población total.

Clasificar los datos faltantes como MCAR es complicado en la práctica, ya que en general los datos faltantes presentan una relación de dependencia con datos observados y no observados. En diversas investigaciones se han desarrollado mecanismos para identificar este patrón de datos, entre los cuales se puede destacar la representación de los datos faltantes en cada variable a través de variables artificiales o dummy (Cohen y Cohen (1975)), en donde se asigna el valor de 0 a la variable artificial cuando el dato es observado y 1 cuando el dato es faltante. Las variables artificiales se utilizan como variables predictivas en un modelo de regresión y se puede evaluar si tienen una relación de dependencia con la variable dependiente a través del coeficiente regresor. Si el coeficiente resulta significativo la ausencia de ciertos datos presenta un comportamiento condicional a la variable explicada y diferente a los datos que sí fueron observados. En cambio si el coeficiente regresor no resulta significativo, se puede asumir que los datos faltantes presentan un comportamiento aleatorio e independiente a la variable dependiente.

1.2.2. Faltante Aleatorio (MAR)

Si los datos presentan el patrón MAR, la probabilidad de que un valor sea faltante depende sólo de los datos observados:

$$P\left(X_{miss} \mid X_{obs}, X_{miss}\right) = P\left(X_{miss} \mid X_{obs}\right)$$

Cuando se tienen valores ausentes del tipo MAR es posible encontrar una estructura concreta de su distribución, ya que la probabilidad de que un dato sea faltante se obtiene condicionando sobre los valores observados. En diversos estudios se ha concluido que los datos faltantes del tipo MAR pueden ser ignorados, ya que los resultados obtenidos a través de métodos basados en la versimilitud no se ven alterados por la ausencia de los MAR. Incluso Collins, Schafer y Kam (2001) demostraron con datos reales que aunque se asuma erróneamente que los datos faltantes siguen una distribución MAR, este supuesto tiene un impacto poco significativo en los estimadores y errores estándar. Asumir que los datos perdidos son MAR implica no con-

siderar en las estimaciones alguna causa o correlación debida a su ausencia.

Para identificar si los valores ausentes siguen una distribución MAR, Little (1988) propone un estadístico de prueba con distribución χ^2 con f grados de libertad, donde la hipótesis nula H_0 establece que los datos faltantes siguen una distribución MAR. La regla de decisión, como en cualquier prueba de hipótesis, es rechazar H_0 conforme a un nivel de significancia α preestablecido.

1.2.3. Faltante No Aleatorio (MNAR)

Finalmente, en el caso en que los datos son clasificados como MNAR, la probabilidad de que el dato sea missing o bien $P\left(X_{miss} \mid X_{obs}, X_{miss}\right)$ no puede ser cuantificada puesto que el motivo por el que se tiene el dato faltante depende de los datos faltantes y en algunos casos también de los observados. Por lo tanto los datos del tipo MNAR no pueden ser ignorados y siempre que se quieran realizar estimaciones con este tipo de datos es necesario incluir un modelo para la probabilidad mencionada anteriormente, que involucre las causas y relaciones de los datos faltantes con la información.

En la práctica, aún cuando los datos faltantes pueden ser ignorados, la meta es reemplazarlos por los valores apropiados con la finalidad de contrarrestar la pérdida de información, sobre todo cuando se cuenta con una muestra reducida o la información reportada es escasa.

Cuando los datos son MCAR o MAR, se dice que los motivos que ocasionaron la falta de datos son ignorables, y así es posible simplificar los métodos de estimación. Métodos como el algoritmo EM y la Imputación Múltiple trabajan bajo este supuesto. En general no es fácil obtener evidencia empírica que demuestre que los datos faltantes son MCAR o MAR sin embargo, se puede justificar la elección del método más conveniente para trabajar con ellos.

Capítulo 2

Tratamiento de los Datos Faltantes

Para trabajar de manera óptima y adecuada con un conjunto de datos que presenta valores ausentes es importante analizar la representatividad de los mismos dentro de la población total. Dependiendo de la proporción de los datos faltantes y del tamaño de la población de estudio será conveniente omitir los valores ausentes, o bien estimarlos.
Generalmente los patrones de datos faltantes no son completamente desconocidos, si no que presentan factores identificables y no identificables. Al analizar un conjunto de datos siempre existe incertidumbre acerca del comportamiento de los valores ausentes y si este es completamente identificable. El objetivo es identificar el patrón para el mayor número de valores ausentes y aún cuando una proporción de los datos faltantes siga un patrón difícil de tratar, siempre se debe dar prioridad a su tratamiento mediante métodos basados en principios estadísticos y evitar recurrir, en la medida de lo posible, a los métodos de eliminación. En este trabajo, como en estudios estadísticos previos, se mostrará que el resultado de aplicar métodos estadísticos para el tratamiento de valores ausentes es significativamente mejor que el que se obtiene eliminando información.

2.1. Métodos de eliminación

Los métodos más comunes para eliminar los datos faltantes son Listwise Deletion (LD) y Pairwise Deletion (PD), que por practicidad, son los más utilizados para el tratamiento de bases de datos.

El método LD consiste en remover observaciones completas que presenten valores ausentes en una o más variables. Por su simplicidad es el método más utilizado, sin embargo se puede llegar a perder una cantidad importante de información.
Para utilizar de manera apropiada este método se deben evaluar los valores ausentes considerando dos criterios: la aleatoriedad y la cantidad. Si los valores ausentes son completamente aleatorios y representan una proporción baja de la población total , las estimaciones realizadas con los datos no presentarán sesgos o desviaciones importantes. Aún cuando los valores ausentes no sean completamente aleatorios pero representen un porcentaje muy bajo de la población, las estimaciones resultarán confiables. Sin embargo, si el porcentaje de datos faltantes es elevado este método no es recomendado, pues los análisis estadísticos obtenidos con la información resultante presentarán sesgos y no serán representativos de la población original.

Figura 2.1: Ejemplo: Listwise Deletion

X_1	X_2	X_3	X_4	X_5
1	1	1	1	1
~~1~~	~~0~~	~~1~~	~~1~~	~~1~~
1	1	1	1	1
~~0~~	~~0~~	~~1~~	~~1~~	~~0~~
1	1	1	1	1
1	1	1	1	1
~~0~~	~~1~~	~~1~~	~~1~~	~~0~~
~~0~~	~~1~~	~~0~~	~~0~~	~~1~~
1	1	1	1	1
1	1	1	1	1

El método PD conserva toda la información disponible en cada variable, es decir, no toma en cuenta los valores ausentes dentro de cada variable para realizar estimaciones. Por ejemplo, si en una población se tienen dos variables X_i y X_j, donde ambas presentan valores ausentes y se desean calcular estadísticos como la media y la varianza se utilizaran los n_i y n_j valores disponibles para cada una de ellas. En cambio si se desea calcular la covarianza entre ambas variables, solo podrán considerarse los casos donde las variables son observadas por pares. Esto implicará trabajar con diferentes tamaños muestrales e incluso combinarlos en el cálculo de un mismo estadistico, lo que puede resultar en correlaciones fuera del intervalo $[-1, 1]$ o matrices de

correlaciones no positivas definidas, siendo éstas condiciones necesarias para diversas técnicas multivariadas.
A pesar de que la pérdida de información es menor en comparación con el método LD, cada variable tendrá un número distinto de observaciones y los resultados de estudios estadísticos podrían perder validez puesto que no son comparables entre sí. Por su simplicidad, este método está implementado de manera predeterminada en diversos programas y paqueterias de análisis estadístico.

Figura 2.2: Ejemplo: Pairwise Deletion

X_1	X_2	X_3	X_4	X_5
1	1	1	1	1
1	~~0~~	1	1	1
1	1	1	1	1
~~0~~	~~0~~	1	1	~~0~~
1	1	1	1	1
1	1	1	1	1
~~0~~	1	1	1	~~0~~
~~0~~	1	~~0~~	~~0~~	1
1	1	1	1	1
1	1	1	1	1

2.2. Métodos de Imputación

Los métodos de imputación pretenden solucionar el problema de los datos faltantes sustituyendo los mismos por valores estimados a partir de la información suministrada por la muestra. Con ello se consigue una matriz de datos completa, sobre la cual, se pueden realizar inferencias estadísticas resolviendo algunos de los problemas derivados de los métodos de eliminación. La forma de estimar o predecir los valores perdidos es lo que diferenciará unos métodos de otros, comenzando desde la estimación más simple hasta la que incorpora la mayor información posible. Dentro de los métodos de imputación que tienen una aplicabilidad general se encuentran los sigueintes:

1. *Sustitución por media no condicional*
   Dada una variable X_i dentro de un conjunto de datos que presenta

valores perdidos, se reemplaza cada uno de ellos por la media de los valores observados $\bar{X}_{i_{obs}}$. El método trabaja bajo el supuesto de que los datos faltantes siguen una distribución MCAR. Aunque esta estrategia es sencilla y puede resultar intuitivamente satisfactoria, tiende a subestimar la varianza real de la muestra al sustituir los datos faltantes por valores centrales ya que la suma de cuadrados de las desviaciones de las observaciones respecto de la media permanece inalterada pero se incrementa el tamaño de muestra, lo cual origina que la varianza de la variable disminuya. Además, en caso de que las variables imputadas bajo este método se utilicen en análisis posteriores como modelos de regresión, se ha demostrado que se alteran los valores de los parámetros estimados, así como su significancia estadística.

2. *Sustitución por media condicional*
   El objetivo de este método es imputar medias condicionadas a valores observados. Para esto se agrupan los valores observados y no observados en n categorías a partir de covariables relacionadas con la variable de interés y se imputan los valores faltantes por la media de los valores observados dentro de la misma categoría.
   Al igual que la imputación de medias, en este caso se asume que los datos faltantes siguen un patrón MCAR y existirán tantas medias como categorías se construyan, lo cual contribuye a atenuar los sesgos. En la medida que la falta de información por categoría sea baja, los sesgos disminuyen pero no desaparecen. El uso de este procedimiento no es recomendado, sobre todo si se dispone de una mejor alternativa para sustituir la información omitida.

3. *Variable artificial*
   Este método fue propuesto por Cohen y Cohen (1983) y consiste en crear una variable indicadora I que toma el valor 1 si los valores son ausentes y 0 si son observados para una determinada variable X. Posteriormente se calcula la media de I y se imputa a los valores ausentes de la variable original, creando una nueva variable imputada X^*. Si se trabaja con un modelo de regresión simple con una variable dependiente Y, tanto la variable I como X^* se introducen como variables independientes junto con las demás variables explicativas. El resultado es que la variable artificial elimina la varianza generada por los datos faltantes en la variable dependiente, mientras que los coeficientes estimados para las variables independientes tendrán un valor ajustado por

la presencia de los valores ausentes por el hecho de haber introducido la variable indicadora al modelo. La desventaja del modelo es que los coeficientes estimados resultarán iguales a los que produce el método LD y por lo tanto presentarán sesgos.

4. *Regresión Simple*
   Considérese un conjunto de datos con k variables, dentro de las cuales la variable X_i presenta n valores faltantes y $n-m$ valores observados y las restantes $k-1$ variables no presentan valores faltantes. El método consiste en estimar una regresión de la variable X_i sobre las $k-1$ variables restantes a partir de los $n-m$ casos completos y se imputa cada valor perdido con la predicción dada por la ecuación de regresión estimada. Por ejemplo, para imputar la j-éstima observación perdida de la variable X_i se aplica un modelo de la forma:

$$\hat{x}_{ij} = \hat{\beta}_{0_{obs}} + \sum_{k \neq i} \hat{\beta}_{k_{obs}} x_{kj}$$

   Donde $\hat{\beta}_{0_{obs}}$ y $\hat{\beta}_{k_{obs}}$ representan a los coeficientes estimados de la regresión basados en los $n-m$ valores observados.
   A diferencia de la sustitución por la media, este método incorpora información de la variable con valores perdidos contenida en las variables restantes.

5. *Regresión Estocástica*
   Al imputar mediante regresión simple se está reemplazando el valor perdido por una media condicionada, por lo que, como se menciona en el caso de imputación mediante la media, se tiende a subestimar la varianza. Una alternativa para atenuar este efecto consiste en añadir al valor predicho por la regresión una perturbación aleatoria, con lo que se obtiene una realización de la distribución predictiva de los valores perdidos condicionada a los valores observados.

$$\hat{x}_{ij} = \hat{\beta}_{0_{obs}} + \sum_{k \neq i} \hat{\beta}_{k_{obs}} x_{kj} + \varepsilon_{kj}$$

   Donde $\varepsilon_{kj} \sim N(0, \sigma^2_{err})$, siendo σ^2_{err} la varianza residual de la regresión.

6. *Hot-Deck*
   Este método consiste en llenar los registros vacíos (receptores) con

información de campos con información completa (donantes), y los datos faltantes se reemplazan a partir de una selección aleatoria de valores observados, lo cual no introduce sesgos en la varianza del estimador. Para ello se identifican características comunes entre donantes y receptores generando grupos homogéneos y se reemplazan los valores perdidos por valores registrados dentro de cada grupo. De esta manera se consigue que los datos faltantes sigan la misma distribución intragrupo de los valores observados. Entre las variantes de este método se destacan el algoritmo secuencial y el vecino más cercano.
Debido a que la imputación bajo esta técnica no genera sesgos y preserva la distribución de probabilidad de las variables imputadas, se considera más eficiente que los métodos de eliminación, imputación de medias y regresión.

7. *Máxima Verosimilitud*
   Rubin (1976) propone realizar inferencias en conjuntos de datos incompletos a través de la verosimilitud, incorporando un modelo para los datos faltantes. Este método incorpora el comportamiento de los datos faltantes a través de una variable indicadora de valor perdido, que toma el valor de 1 si el dato es observado y 0 si es faltante. Posteriormente se obtiene la distribución conjunta entre variable indicadora y los datos observados y se estiman los parámetros a través de la verosimilitud conjunta.
   Si la distribución de los datos faltantes es MAR, entonces la verosimilitud de los datos observados será proporcional a la conjunta y por lo tanto la estimación puede realizarse a partir de los datos observados. Por el contrario, si la distribución es NMAR, entonces debe considerarse la verosimilitud conjunta.

8. *Algoritmo EM*
   El algoritmo EM es un método iterativo que consiste en estimar parámetros desconocidos de una población con datos faltantes a partir de los valores observados. El proceso se conforma de dos pasos: paso E (Esperanza) y paso M (Maximización). En el paso E se estiman el valor esperado de los datos perdidos dada la información observada, mientras que en el paso M se maximiza la verosimilitud completando con la información calculada en el paso E. El algoritmo se repite iterativamente hasta alcanzar la convergencia.
   El algoritmo EM se ha popularizado recientemente y es ampliamente recomendado por sus propiedades y su amplia gama de aplicaciones,

además de que su aplicación es sencilla y no requiere de pesados procesos computacionales.

9. *Algoritmo EM Estocástico*
   Una gran variedad de problemas de datos faltantes pueden ser resueltos vía el algoritmo EM, por ejemplo, la estimación de parámetros desconocidos o la inferencia en variables latentes. Sin embargo, cuando se presentan casos donde los parámetros o las variables no son sistemáticamente observables y se vuelven intratables, el algoritmo EM no puede aplicarse de manera directa, ya que se requieren métodos de integración numérica en el paso E. En consecuencia, surge una variante del mismo denominada algoritmo EM estocástico.
   Similar al algoritmo EM, el algoritmo EM consiste en dos pasos: el paso S (Estocástico) y el paso M (Maximización). El objetivo del paso S es imputar los datos faltantes a partir de una muestra de la densidad de los datos faltantes dados los valores observados y el parámetro desconocido obtenido en la iteración anterior. El paso M consistirá en maximizar la verosimilitud directamente del conjunto de datos completado en el paso S. Este proceso se repetirá iterativamente hasta que la cadena de Markov generada converge a una distribución estacionaria.
   A diferencia del algoritmo EM determinístico, el resultado final de este algoritmo consistirá en una muestra de la distribución estacionaria cuya media será cercana al estimador por máxima verosimilitud del parámetro desconocido y la varianza reflejará el hecho de que se tenga información faltante.

10. *Imputación Múltiple*
    La imputación múltiple, propuesta por primera vez en Rubin (1978), reemplaza los valores ausentes por un conjunto de valores simulados, incorporando la incertidumbre generada por la ausencia de datos. El algoritmo genera múltiples conjuntos de datos completados, siendo estos repeticiones de la distribución posterior de los valores ausentes y donde cada repetición proviene de muestras independientes de los parámetros y los valores faltantes.
    Posteriormente se procede a realizar el análisis estadísitico para cada conjunto de datos y se combinan los resultados de las estimaciones. Cabe destacar que esta técnica trabaja bajo el supuesto de que los valores perdidos son MAR, además de que se debe elegir el modelo adecuado para realizar la imputación.

La desventaja de este método es que se requiere de mayor tiempo y procesos computacionales para generar los conjuntos de datos y analizar cada uno por separado, además de que por ser un proceso estocástico, en cada repetición se obtienen resultados distintos. Esto no sucede con el algoritmo EM, puesto que siempre se obtiene el mismo estimador por máxima verosimilitud.

Capítulo 3

Algoritmo EM

El algoritmo EM consiste principalmente en asociar a un problema de datos incompletos un problema de datos completados, en donde la estimación por Máxima Verosimilitud sea computacionalmente operable. Para esto se establece una relación entre las verosimilitudes de estos dos problemas con la finalidad de que la estimación vía Máxima Verosimilitud sea la más simple posible en cada iteración.

En diversos problemas estadísticos, la verosimilitud del conjunto de datos completo tiene una distribución conocida y fácil de operar, lo cual hace que el algoritmo EM sea eficiente.

El algoritmo EM consiste en dos pasos, como su mismo nombre lo indica: el paso-E y el paso-M.

El paso-E consite en estimar un conjunto de datos completo utilizando el subconjunto de datos observados y los parámetros correspondientes, de tal manera que los datos faltantes son reemplazados por su esperanza condicional dados los datos conocidos. En este paso se dan a los parámetros valores iniciales, los cuáles pueden ser asignados sin tener inferencia a priori sobre ellos.

El paso-M consiste en maximizar la función de verosimilitud con los datos rellenados en el paso-E y así obtener un nuevo conjunto de parámetros que serán utilizados para actualizar la estimación de la esperanza condicional de los datos desconocidos en la siguiente iteración.

El paso-E y el paso-M se repiten iterativamente hasta que la verosimilitud converge.

El algoritmo EM, así como muchos otros métodos, es empleado bajo el supuesto de que los datos faltantes son MAR, de esta manera los datos faltantes se modelan únicamente en función de los datos observados. Este supuesto es de suma importancia, ya que si los datos no fueran MAR, se tendrían que incluir en la modelación las causas que originaron la pérdida de información que tienen un origen en la misma información faltante, lo cual en la práctica resulta complicado. Estas causas comúnmente son difíciles de identificar y están fuera del alcance del analista estadístico.

3.1. Teoría

3.1.1. Conceptos Preliminares

Definición 3.1.1 (Estadística Suficiente)**.** *Sea $X_1, ..., X_n$ una muestra aleatoria de una población cuya distribución es $f(x, \theta)$. Se dice que el estadístico o estimador $T = T(X_1, ..., X_n)$ es suficiente para el parámetro θ si la distribución condicional de $X_1, ..., X_n$ dado $T = t$ no depende de θ.*

Teorema 3.1.1 (Teorema de Factorización)**.** *Sea $X_1, ..., X_n$ una muestra aleatoria de una población cuya distribución es $f(x, \theta)$. Se dice que la estadistica $T = T((X_1, ..., X_n))$ es una estadística suficiente para θ si y sólo sí la función de verosimilitud puede factorizarse de la forma:*

$$L(x_1, ..., x_n, \theta) = h(t, \theta)\, g(x_1, ..., x_n) \tag{3.1.1.1}$$

para cualquier valor $t = T((X_1, ..., X_n))$ y donde $g(x_1, ..., x_n)$ no contiene al parámetro θ.

Definición 3.1.2 (Función de Verosimilitud)**.** *Sea $X_1, ..., X_n$ una muestra de n variables aleatorias independientes e idénticamente distribuidas. Se define la función de verosimilitud como la función de densidad conjunta de $X_1, ..., X_n$.*

$$L(X, \theta) = L(x_1, ..., x_n, \theta) = f(x_1, ..., x_n, \theta) = \prod_{i=1}^{n} f(x_i, \theta) \tag{3.1.1.2}$$

3.1.2. La Familia Exponencial

La familia exponencial contiene una amplia gama de distribuciones, que por sus propiedades, han permitido obtener de una manera bastante práctica importantes resultados estadísticos a lo largo de la historia. Esta familia contiene algunas de las más conocidas distribuciones, tanto discretas como continuas, como son la Poisson, Binomial, Multinomial, Normal, Exponencial, Gamma, Normal Multivariada, etc.

Definición 3.1.3. *Una famila de densidades de un vector aleatorio Y, por decir $f(y \mid \theta)$, tal que $(\theta_1, \ldots, \theta_n) = \theta$ pertenece al espacio de parámetros Θ perternece a la familia exponencial si:*

$$f(Y \mid \theta) = h(y)g(\theta)e^{\sum_{i=1}^{n} w_i(\theta)T_i(y)} \tag{3.1.2.1}$$

Donde $g(\theta) \geq 0$ y $w_i(\theta)$ son funciones que no dependen de Y, y $h(y) \geq 0$ y $T_i(y)$ no dependen de θ. Esta forma corresponde a una familia exponencial con n-parámetros, donde n es el mínimo entero tal que (3.1.2.1) *se cumple.*

En particular, la distribución de Y es una familia exponencial de un parámetro si:

$$f(Y \mid \theta) = h(y)g(\theta)e^{w(\theta)T(y)} \tag{3.1.2.2}$$

Definición 3.1.4. *Sea Θ el espacio de parámetros de ϕ. La parametrización natural de la familia exponencial está definida como:*

$$f(Y \mid \theta) = h(y)g(\phi)e^{\sum_{i=1}^{n} \phi_i T_i(y)} \tag{3.1.2.3}$$

Donde $\phi \in \Theta$ se denomina como el parámetro natural que especifica todos los parámetros que definen a la densidad en su forma original.

Definición 3.1.5. *Una familia exponencial de k-parámetros es curvada si el conjunto de parámetros $\theta = (\theta_1, \ldots, \theta_k)$ se puede determinar por un subconjunto de θ de tamaño d donde $d < k$.*

Definición 3.1.6. *Una familia exponencial de k-parámetros es regular si se cumplen las siguientes condiciones:*

1. $\Theta = \{\theta \in \Theta : \frac{1}{g(\theta)} = \int_{-\infty}^{\infty} h(y)e^{\sum_{i=1}^{n} w_i(\theta)T_i(y)} < \infty\}$
2. Θ *es un conjunto abierto no vacío con k-dimensiones .*

3. *No existe dependencia lineal para el conjunto de* $T_i(Y)$ *con* $i = 1, \ldots, k$ *y* $w_i(\theta)$ *con* $i = 1, \ldots, k$.

Teorema 3.1.2 (Verosimilitud)**.** *Sean* $Y_1, \ldots, Y_n$ *variables aleatorias independientes e identicamente distribuidas de la familia exponencial. Entonces la distribución conjunta de* $Y_1, \ldots, Y_n$ *pertenece a la familia exponencial.*

Demostración.

$$
\begin{aligned}
f(y_1, \ldots, y_n) &= \prod_{i=1}^{n} f(Y_i \mid \theta) \\
&= \prod_{i=1}^{n} h(y_i) g(\theta) e^{\sum_{j=1}^{k} w_j(\theta) T_j(y_i)} \\
&= g(\theta)^n \prod_{i=1}^{n} h(y_i) e^{\sum_{i=1}^{n} \sum_{j=1}^{k} w_j(\theta) T_j(y_i)} \\
&= g(\theta)^n \prod_{i=1}^{n} h(y_i) e^{\sum_{j=1}^{k} w_j(\theta) \sum_{i=1}^{n} T_j(y_i)}
\end{aligned}
$$

Puede notarse que esta expresión es de la forma de (3.1.2.1), donde:

$$
\begin{aligned}
h^*(y_1 \ldots y_n) &= \prod_{i=1}^{n} h(y_i) \\
g^*(\theta) &= g(\theta)^n \\
w_j^*(\theta) &= w_j(\theta) \\
T_j^*(y_i) &= \sum_{i=1}^{n} T_j(y_i)
\end{aligned}
$$

□

3.1.3. Propiedades de las estadísticas $T_1(y), \ldots, T_k(y)$

Definición 3.1.7. *Supongamos que* $Y = (Y_1, \ldots, Y_n)$ *tiene una distribución de la forma* (3.1.2.2)*. Entonces la estadística* $T(Y)$ *representa a la estadística suficiente natural.*

El concepto de estadística suficiente es fundamental en la teoría estadística y sus aplicaciones, ya que formaliza la idea de ganancia informativa. Una estadística suficiente contiene toda la información acerca de los parámetros desconocidos de la distribución en cuestión y que sería proporcionada por una muestra completa. Cabe destacar que la forma de la estadística suficiente depende de la elección de la distribución con la que se desee modelar Y, sin embargo este concepto es ampliamente utilizado.

La esperanza y la varianza de $T(Y)$ están dadas por:

1 $E[T(y) \mid \theta] = s'(\theta)$

2 $Var[T(y) \mid \theta] = s''(\theta)$

Escribamos a $f(Y \mid \theta)$ de la forma:

$$h(y)e^{w(\theta)T(y)-s(\theta)} \tag{3.1.3.1}$$

Donde:
$s(\theta) = -\log g(\theta)$
Utilizando la función generadora de momentos:

$$\begin{aligned} m_{T(y)}(t) &= E\left[e^{tT(y)}\right] = \int e^{tT(y)}h(y)e^{w(\theta)T(y)-s(\theta)}\,dy \\ &= e^{-s(\theta)}\int h(y)e^{T(y)(t+w(\theta))}\,dy \\ &= \frac{e^{-s(\theta)}}{e^{-s(\theta+t)}}\int h(y)e^{T(y)(t+w(\theta))-s(\theta+t)}\,dy \\ &= \frac{e^{-s(\theta)}}{e^{-s(\theta+t)}} \end{aligned} \tag{3.1.3.2}$$

Utilizando este hecho podemos obtener los momentos de la siguiente manera:

$$\frac{\partial}{\partial t}e^{-s(\theta)+s(\theta+t)}$$

$$= s'(\theta+t)e^{-s(\theta)+s(\theta+t)}|_{t=0} = s'(\theta)$$

$$\frac{\partial^2}{\partial t}e^{-s(\theta)+s(\theta+t)}$$

$$= s'(\theta+t)^2e^{-s(\theta)+s(\theta+t)} + e^{-s(\theta)+s(\theta+t)}s''(\theta+t)|_{t=0} = s'(\theta)^2 + s''(\theta)$$

Por lo tanto:
$E[T(y) \mid \theta] = s'(\theta)$ y $Var[T(y) \mid \theta] = s''(\theta)$

Equivalentemente:

$$\begin{aligned} E[T(y) \mid \theta] &= -\frac{\partial}{\partial\theta}(\log g(\theta)) \\ Var[T(y) \mid \theta] &= -\frac{\partial^2}{\partial\theta}(\log g(\theta)) \end{aligned} \tag{3.1.3.3}$$

3.2. Operación del Algoritmo EM

De acuerdo con D.L.R. (1977) el procedimiento del algoritmo puede ser simplificado cuando los datos tienen una distribución de la forma de la familia exponencial. Supongamos se tiene un conjunto de datos Y con una distribución de la familia exponencial y que $\phi_i^{(p)}$ es el estimador del parámetro natural ϕ después de p iteraciones del algoritmo. Además $Y = X \cup Z$, donde X representa a los datos observados y Z a los no observados; entonces se pueden definir los pasos E y M como sigue:

Paso-E: Estimar la estadística suficiente $T(Y)$ de los datos completos calculando

$$T^{(p+1)} = E\left(T\left(Y\right) \mid \phi^{(p)}\right) \tag{3.2.0.4}$$

Paso-M: Calcular $\phi^{(p+1)}$ como la solución del sistema

$$E\left(T\left(Y\right) \mid \phi\right) = T^{(p)} \tag{3.2.0.5}$$

Este sistema corresponde a la estimación por máxima verosimilitud de ϕ del conjunto de datos completo Y, y donde $T(y)$ es reemplazado por $T^{(p)}$ en cada iteración del algoritmo. En algunos casos este sistema no tiene solución para ϕ, sin embargo, si se trata de una familia exponencial la solución puede ser encontrada explícitamente, por lo que la complejidad del proceso se concentra en efectuar los cálculos correspondientes al paso-E.

La verosimilitud de Y, considedrando que es de la forma descrita en (3.1.2.3) e ignorando los términos que no dependen de ϕ es:

$$\ell(\phi) = \log g(\phi) + \sum_{i=1}^{n} \phi_i T_i(Y)$$

Como $T_i(Y)$ no es completamente observable se tiene que reemplazar por su esperanza condicional dados los datos observados X y el parámetro $\phi^{(p)}$:

$$Q(\phi \mid \phi^{(p)}) = E[\ell(\phi) \mid X, \phi^{(p)}] = \log g(\phi) + \sum_{i=1}^{n} \phi_i T_i^{(p)}(X), \text{ donde}$$

$$T_i^{(p)}(X) = E[T_i(Y) \mid X, \phi^{(p)}]$$

El siguiente paso, es encontrar el valor $\phi^{(p+1)}$ que maximiza Q, el máximo se obtiene derivando Q respecto a ϕ:

$$0 = \frac{\partial}{\partial \phi_i} Q(\phi \mid \phi^{(p)}) = \frac{\partial}{\partial \phi_i} \log g(\phi) + T_i^{(p)}(X)$$

Entonces por (3.1.3.3):

$$E[T_i(Y) \mid \phi] = T_i^{(p)}(X), \text{ con } i = 1, \ldots, n$$

Con la finalidad de explicar por qué a traves del proceso iterativo de los pasos E y M se obtiene el valor de un parámetro, por decir θ^* ,que maximiza el logaritmo de la verosimilitud para distribuciones no necesariamente dentro de la familia exponencial, se define lo siguiente:

Utilizando una notación diferente al caso de la familia exponencial, supongamos que se tiene un conjunto de datos Z, cuya distribución no necesariamente pertenece a la familia exponencial y que depende de un parámetro θ. Denotemos como Y al conjunto de datos observados y como X al conjunto de datos faltantes, de tal manera que $X \bigcup Y = Z$. En un principio la función de densidad los datos completos Z, $f(Z \mid \theta)$, puede escribirse de la forma:

$$\begin{aligned} f(Z \mid \theta) &= f(x, y \mid \theta) \\ &= \frac{f(x, y, \theta)}{f(\theta)} \\ &= \frac{f(x \mid y, \theta) f(y, \theta)}{f(\theta)} \\ &= f(x \mid y, \theta) f(y \mid \theta) \end{aligned} \tag{3.2.0.6}$$

Si se aplica logaritmo natural en ambos lados de (3.2.0.6) se tiene que:

$$L(\theta) = \log(\theta \mid y) + \log f(x \mid y, \theta) \tag{3.2.0.7}$$

Así se puede escribir la verosimilitud de los datos observados como:

$$\log(\theta \mid y) = L(\theta) - \log f(x \mid y, \theta) \tag{3.2.0.8}$$

Si se calcula la esperanza condicional respecto a los datos desconocidos x dados los datos observados y y el parámetro $\theta^{(p)}$ en ambos lados de (3.2.0.8), entonces:

$$E\left(\log(\theta \mid y) \mid y, \theta^{(p)}\right) = E\left(L(\theta) \mid y, \theta^{(p)}\right) - E\left(\log f(x \mid y, \theta) \mid y, \theta^{(p)}\right)$$

Denotemos a

$$E\left(L(\theta) \mid y, \theta^{(p)}\right) = Q\left(\theta \mid \theta^{(p)}\right) \text{ y}$$

$$E\left(\log f(x \mid y, \theta) \mid y, \theta^{(p)}\right) = H\left(\theta \mid \theta^{(p)}\right)$$

Entonces:

$$E\left(\log(\theta \mid y) \mid y, \theta^{(p)}\right) = Q\left(\theta \mid \theta^{(p)}\right) - H\left(\theta \mid \theta^{(p)}\right) \tag{3.2.0.9}$$

Se sabe que si $g(X)$ es una función de una variable aleatoria X entonces la esperanza condicional de $g(X)$ dado el valor de una variable Y está dada por:

$$E(g(X) \mid Y = y) = \int_{-\infty}^{\infty} g(x) f_{X|Y}(x \mid y)\, dx \tag{3.2.0.10}$$

Por (3.2.0.10) tenemos que

$$Q\left(\theta \mid \theta^{(p)}\right) = \int L(\theta) f(x, y \mid y, \theta^{(p)})\, dx \text{ y} \tag{3.2.0.11}$$

$$H(\theta \mid \theta^{(p)}) = \int \log f(x \mid y, \theta) f(x \mid y, \theta^{(p)})\, dx \tag{3.2.0.12}$$

Definición 3.2.1 (Función Convexa)**.** *Se dice que una función $g(x)$ doblemente diferenciable es convexa si la segunda derivada de $g(x)$ es positiva, es decir, $g^{(2)}(x) \geq 0$.*

Teorema 3.2.1 (Desigualdad de Jensen)**.** *Si $g(x)$ es una función convexa, entonces*

$$E[g(X)] \geq f(g[X]) \tag{3.2.0.13}$$

si las esperanzas anteriores existen y son finitas. En caso de que $g(x)$ sea cóncava se invierte la desigualdad.

La finalidad del algoritmo EM es maximizar la verosimilitud de los datos observados, escribirla como la resta de dos verosimilitudes permite efectuar el paso-M en muchas ocasiones, de manera explícita.

La función $Q\left(\theta \mid \theta^{(p)}\right)$ representa a la esperanza condicional del logaritmo de la verosimilitud del conjunto de datos completo, por lo que en la práctica obtener esta verosimilitud y realizar la maximización no resulta complicado.

La función $H(\theta \mid \theta^{(p)})$ representa a la esperanza condicional del logaritmo de la verosimilitud del conjunto de datos incompleto dados los valores observados y el parámetro $\theta^{(p)}$ en la p-ésima iteración. Caso contrario al anterior, esta verosimilitud y su proceso de maximización tienen un nivel de complejidad elevado el cual impide que se puedan realizar los cálculos de manera explícita, sin embargo en el siguiente teorema se demuestra que al tener una contribución negativa en el incremento de la verosimilitud, el único factor que contribuye al incremento de la misma es Q.

Teorema 3.2.2. *La verosimilitud de los datos incompletos H cumple con la propiedad:*

$$H(\theta \mid \theta^{(p)}) \leq H(\theta^{(p)} \mid \theta^{(p)}) \tag{3.2.0.14}$$

Demostración. La desigualdad anterior se puede escribir de la siguiente manera:

$$H(\theta \mid \theta^{(p)}) - H(\theta^{(p)} \mid \theta^{(p)}) \leq 0 \tag{3.2.0.15}$$

Desarrollando (3.2.0.15) se tiene que:

$$E\left(\log f\left(x \mid y,\theta\right) \mid y,\theta^{(p)}\right) - E\left(\log f\left(x \mid y,\theta^{(p)}\right) \mid y,\theta^{(p)}\right)$$
$$E\left(\log f\left(x \mid y,\theta\right) - \log f\left(x \mid y,\theta^{(p)}\right) \mid y,\theta^{(p)}\right)$$
$$E\left(\log\left(\frac{f\left(x \mid y,\theta\right)}{f\left(x \mid y,\theta^{(p)}\right)}\right) \mid y,\theta^{(p)}\right)$$

Dado que $g(x) = \log x$ es una función cóncava, es decir:
$g''(x) = -\frac{1}{x^2} \leq 0,\ \forall x \in \Re$
y por la Desigualdad de Jensen descrita anteriormente se tiene que:

$$\leq \log\left(E\left(\frac{f\left(x \mid y,\theta\right)}{f\left(x \mid y,\theta^{(p)}\right)} \mid y,\theta^{(p)}\right)\right)$$
$$= \log\left(\int f(x \mid y,\theta^{(p)})\frac{f(x \mid y,\theta)}{f(x \mid y,\theta^{(p)})}\,dx\right)$$
$$= \log\left(\int f(x \mid y,\theta)\,dx\right)$$
$$= \log 1$$
$$= 0$$

Por lo tanto:
$H(\theta \mid \theta^{(p)}) \leq H(\theta^{(p)} \mid \theta^{(p)})$

□

Llamemos $\ell_y(\theta)$ a la verosimilitud de los datos observados:

$$\ell_y(\theta) = E\left(\log\left(\theta \mid y\right) \mid y,\theta^{(p)}\right)$$

Teorema 3.2.3. *Si $\theta^{(p)}$ converge a un valor $\theta^{(*)}$, entonces $\theta^{(*)}$ maximiza $\ell_y(\theta)$, es decir:*

$$Q\left(\theta^{(*)} \mid \theta^{(*)}\right) = \max_{\theta} Q\left(\theta \mid \theta^{(*)}\right) \tag{3.2.0.16}$$

Demostración. La ecuación (3.2.0.16) implica que cuando $\theta^{(p)} \to \theta^{(*)}$

$$\frac{\partial}{\partial\theta} Q\left(\theta^{(*)} \mid \theta^{(*)}\right) = 0$$

Maximizar la función Q implica que:

$$\begin{aligned}
0 &= \frac{\partial}{\partial\theta} Q\left(\theta \mid \theta^{(p)}\right) \\
&= \frac{\partial}{\partial\theta} H\left(\theta \mid \theta^{(p)}\right) + \frac{\partial}{\partial\theta}\ell_y(\theta) \\
&= \frac{\partial}{\partial\theta}\int \log f(x \mid y,\theta) f(x \mid y,\theta^{(p)})\, dx + \frac{\partial}{\partial\theta}\ell_y(\theta) \\
&= \int \frac{\partial}{\partial\theta}\log f(x \mid y,\theta) f(x \mid y,\theta^{(p)})\, dx + \frac{\partial}{\partial\theta}\ell_y(\theta) \\
&= \int \frac{\frac{\partial}{\partial\theta} f(x \mid y,\theta)}{f(x \mid y,\theta)} f(x \mid y,\theta^{(p)})\, dx + \frac{\partial}{\partial\theta}\ell_y(\theta)
\end{aligned}$$

Entonces bajo el supuesto de que $\theta^{(p)} \to \theta^{(*)}$:

$$\begin{aligned}
&= \int \frac{\frac{\partial}{\partial\theta} f(x \mid y,\theta^{(*)})}{f(x \mid y,\theta^{(*)})} f(x \mid y,\theta^{(*)})\, dx + \frac{\partial}{\partial\theta}\ell_y(\theta^{(*)}) \\
&= \int \frac{\partial}{\partial\theta} f(x \mid y,\theta^{(*)})\, dx + \frac{\partial}{\partial\theta}\ell_y(\theta^{(*)}) \\
&= \frac{\partial}{\partial\theta}\int f(x \mid y,\theta^{(*)})\, dx + \frac{\partial}{\partial\theta}\ell_y(\theta^{(*)}) \\
&= \frac{\partial}{\partial\theta}(1) + \frac{\partial}{\partial\theta}\ell_y(\theta^{(*)}) \\
&= \frac{\partial}{\partial\theta}\ell_y(\theta^{(*)})
\end{aligned}$$

Por lo tanto $\theta^{(*)}$ es el valor que maximiza la verosimilitud de los datos observados tal que:

$$\frac{\partial}{\partial\theta}\ell_y(\theta^{(*)}) = 0$$

□

Este resultado implica que si las iteraciones de $\theta^{(p)}$ convergen, convergen a un punto estacionario de $\ell_y(\theta)$, que en el caso de distribuciones unimodales corresponde al estimador por máxima verosimilitud de θ. Sin embargo, si la distribución es multimodal el algoritmo EM puede no converger al máximo

global, debido a la presencia de máximos locales y puntos silla. Dependiendo de la elección de los valores iniciales para el parámetro θ, el algoritmo convergerá a distintos puntos, por lo que se deben elegir de manera cuidadosa, además es recomendable que se repita la ejecución del algoritmo para distintos valores iniciales.

Entonces paso-E consistirá en calcular la esperanza condicional de la verosimilitud de los datos completos, sustituyendo el parámetro no observable θ por el estimador en la p-ésima iteración $\theta^{(p)}$. Mientras que para efectuar el paso-M únicamente se requerirá maximizar la función Q.

En resumen:
Paso-E: Calcular

$$Q\left(\theta \mid \theta^{(p)}\right) = E\left(L\left(\theta\right) \mid y, \theta^{(p)}\right)$$

Paso-M: Encontrar $\theta^{(p+1)}$ maximizando la función Q de tal manera que:

$$Q\left(\theta^{(p+1)} \mid \theta^{(p)}\right) \geq Q\left(\theta \mid \theta^{(p)}\right)$$

Uno de los puntos centrales, bajo los cuales trabaja el algoritmo EM es que en cada iteración del algoritmo hay un incremento en la verosimilitud, es decir, conforme aumentan las iteraciones del algoritmo se obtiene la mejor estimación del parámetro θ. Esto se demuestra en el siguiente teorema:

Teorema 3.2.4. *La sucesión $\ell_y(\theta^{(p)})$ es no decreciente, es decir:*

$$\ell_y(\theta^{(p+1)}) \geq \ell_y(\theta^{(p)}), \forall \theta \in \Theta \ y \ p \in N$$

Demostración. Sea $\theta^{(p)}$ un valor dado de tal manera que se puede encontrar $\theta^{(p+1)}$ que cumple:

$$Q\left(\theta^{(p+1)} \mid \theta^{(p)}\right) = \max_{\theta} Q\left(\theta \mid \theta^{(p)}\right) \tag{3.2.0.17}$$

Entonces:

$$\ell_y(\theta^{(p+1)}) = Q\left(\theta^{(p+1)} \mid \theta^{(p)}\right) - H\left(\theta^{(p+1)} \mid \theta^{(p)}\right)$$

Por (3.2.0.14)

$$\geq Q\left(\theta^{(p+1)} \mid \theta^{(p)}\right) - H\left(\theta^{(p)} \mid \theta^{(p)}\right)$$

Como $\theta^{(p+1)}$ es el valor que maximiza (3.2.0.17)

$$\begin{aligned} &\geq Q\left(\theta^{(p)} \mid \theta^{(p)}\right) - H\left(\theta^{(p)} \mid \theta^{(p)}\right) \\ &= \ell_y(\theta^{(p)}) \end{aligned}$$

Por lo tanto la verosimilitud de los datos observados es no decreciente en cada iteración del algoritmo EM. □

Por el resultado anterior y en caso de que la sucesión $\ell_y(\theta^{(p)})$ fuera acotada, entonces la sucesión converge a un valor ℓ^*. Tal y como se plantea en Wu (1983), el objetivo es saber si ℓ^* es el máximo global para $\ell_y(\theta)$ o si es un máximo local o un punto estacionario. La clave de la convergencia radica en que la sucesión de verosimilitudes esté acotada superiormente, y para que esto ocurra se deben cumplir los siguientes supuestos:

1. El espacio de parámetros Θ es un subconjunto de R^n.

2. El conjunto que contiene a los puntos iniciales $\Theta_{\theta_0}=\{\theta \in \Theta : \ell(\theta) \geq \ell(\theta_0)\}$ es compacto para toda $\ell(\theta_0) > -\infty$. Donde θ_0 representa el valor inicial para el parámetro θ en la primera iteración del algoritmo EM.

3. ℓ es continua en Θ y diferenciable en el interior de Θ. Esto es que cuando se evalúan en $\theta^{(p)}$ las derivadas en el proceso de maximización, se asume que $\theta^{(p)}$ es un punto interior de Θ.

Algunos resultados importantes relacionados con el caso de la familia exponencial también son tratados en Wu(1983), donde se muestra que dadas las diversas propiedades que tiene esta familia, la convergencia queda garantizada. Algunos de ellos se muestran a continuación:

1. Si la distribución de los datos completos $f(Z \mid \theta)$ pertenece a la familia exponencial curvada y $\ell_y(\theta)$ está acotada, entonces $\ell_y(\theta^{(p)})$ converge a un valor estacionario ℓ^*.

2. Si la distribución de los datos completos $f(Z \mid \theta)$ pertenece a la familia exponencial regular y $\ell_y(\theta)$ está acotada, entonces $\theta^{(p)}$ converge a un valor estacionario θ^*.

Tasa de Convergencia
Un resultado interesante del algoritmo EM, está relacionado con la velocidad de convergencia del mismo. D.L.R.(1977) demuestran que entre mayor sea la proporción de datos faltantes en un conjunto de datos, la convergencia del algoritmo será menor, además la tasa de convergencia es lineal.

Definición 3.2.2. *La tasa de convergencia de un proceso iterativo se define como:*

$$r = \lim_{p\to\infty} \frac{\|\theta^{(p+1)} - \theta^*\|}{\|\theta^{(p)} - \theta^*\|} \tag{3.2.0.18}$$

Donde $\| \bullet \|$ *es una norma vectorial.*

Por otra parte, durante el proceso de maximización ,el algoritmo EM define de manera implícita un mapeo continuo y monótono de la forma:

$$\theta^{(p+1)} = M(\theta^{(p)}), \text{ con } p = 0, 1, 2, \ldots$$

Si consideramos la expansión de Taylor de $M(\theta^{(p)})$ en el punto θ^* tal que $\theta^{(p)} \to \theta^*$, se tiene lo siguiente:

$$M(\theta^{(p)}) = M(\theta^*) + (\theta^{(p)} - \theta^*)\frac{\partial M(\theta^{(p)})}{\partial \theta}|_{\theta^{(p)}=\theta^*}$$

Entonces:

$$\theta^{(p+1)} - \theta^* = (\theta^{(p)} - \theta^*)DM \tag{3.2.0.19}$$

Esto quiere decir que estando cerca de θ^*, el algoritmo EM es esencialmente una iteración lineal con una tasa definida por la matriz DM. La tasa r queda definida por el máximo eigenvalor de la matriz DM, cuando los eigenvalores de DM son menores a 1. Este resultado también muestra que, para que la convergencia del algoritmo sea rápida, DM debe tener valores cercanos a cero.

D.L.R.(1977) muestran que la matriz DM está definida de la siguiente manera:

$$DM = \frac{\partial^2}{\partial\theta} H(\theta^* \mid \theta^*) \left[\frac{\partial^2}{\partial\theta} Q(\theta^* \mid \theta^*)\right]^{-1} \tag{3.2.0.20}$$

Cabe destacar que ambos términos corresponden a medidas de información que los datos proporcionan sobre el parámetro θ. La información originada por la pérdida de información está contenida en el primer término de DM, y cuando esta información sea pequeña el algoritmo convergerá rápidamente. Algunos componentes del parámetro θ podrán reflejar diferentes proporciones de pérdida de información, en consecuencia, una parte de ellos conllevará a una convergencia rápida, mientras que otros podrán requerir de un mayor número de iteraciones.

Teorema 3.2.5. *Supongamos que $\theta^{(p)}$ con $p = 0, 1, 2, \ldots$ es una sucesión generada por el algoritmo EM, de tal manera que se cumple lo siguiente:*

1. *$\theta^{(p)}$ converge a algún θ^* en Θ.*

2. *$\frac{\partial}{\partial\theta}Q(\theta^{(p+1)} \mid \theta^{(p)}) = 0$*

3. *$\frac{\partial^2}{\partial\theta}Q(\theta^{(p+1)} \mid \theta^{(p)})$ es negativa definida, con eigenvalores distintos de cero.*

Entonces:

1. *$\frac{\partial^2}{\partial\theta}Q(\theta^* \mid \theta^*)$ es negativa definida*

2. *DM=$\frac{\partial^2}{\partial\theta}H(\theta^* \mid \theta^*)\left[\frac{\partial^2}{\partial\theta}Q(\theta^* \mid \theta^*)\right]^{-1}$*

Demostración. Como $\frac{\partial^2}{\partial\theta}Q(\theta^* \mid \theta^*)$ es el límite de $\frac{\partial^2}{\partial\theta}Q(\theta^{(p+1)} \mid \theta^{(p)})$ entonces es definida negativa ya que esta última es definida negativa por el supuesto 2. El hecho de que sea negativa definida implica que θ^* es un máximo global de la función Q.
Ahora bien, para obtener DM, hagamos la expansión de Taylor de la derivada de la función Q alrededor de θ^*:

$$\frac{\partial}{\partial\theta}Q(\theta^{(p+1)} \mid \theta^{(p)}) = \frac{\partial}{\partial\theta}Q(\theta^* \mid \theta^{(p)}) + (\theta^{(p+1)} - \theta^*)\frac{\partial^2}{\partial\theta^{(p+1)}}Q(\theta^* \mid \theta^{(p)}) + (\theta^{(p)} - \theta^*)\frac{\partial^2}{\partial\theta^{(p)}}Q(\theta^* \mid \theta^{(p)}) + \ldots$$

Tomando el límite cuando $\theta^{(p)} \to \theta^*$:

$$\frac{\partial}{\partial\theta}Q(\theta^{(*)} \mid \theta^{(*)}) = \frac{\partial}{\partial\theta}Q(\theta^{*} \mid \theta^{(*)}) + (\theta^{(p+1)} - \theta^{*})\frac{\partial^2}{\partial\theta^{(p+1)}}Q(\theta^{*} \mid \theta^{(*)}) + (\theta^{(p)} - \theta^{*})\frac{\partial^2}{\partial\theta^{(p)}}Q(\theta^{*} \mid \theta^{(*)}) + \dots$$

Sustituyendo (3.2.0.19):

$$\begin{aligned} 0 &= (\theta^{(p)} - \theta^{*})DM\frac{\partial^2}{\partial\theta^{(p+1)}}Q(\theta^{*} \mid \theta^{(*)}) + (\theta^{(p)} - \theta^{*})\frac{\partial^2}{\partial\theta^{(p)}}Q(\theta^{*} \mid \theta^{(*)}) + \dots \\ &= DM\frac{\partial^2}{\partial\theta^{(p+1)}}Q(\theta^{*} \mid \theta^{(*)}) + \frac{\partial^2}{\partial\theta^{(p)}}Q(\theta^{*} \mid \theta^{(*)}) \\ &\Longrightarrow \\ DM &= -\frac{\partial^2}{\partial\theta^{(p)}}Q(\theta^{*} \mid \theta^{(*)})\left[\frac{\partial^2}{\partial\theta^{(p+1)}}Q(\theta^{*} \mid \theta^{(*)})\right]^{-1} \end{aligned}$$

Como:

$$Q(\theta \mid \theta^{(p)}) = \ell_y(\theta) + H(\theta \mid \theta^{(p)})$$
$$\Longrightarrow$$
$$\frac{\partial}{\partial\theta^{(p)}}Q(\theta \mid \theta^{(p)}) = \frac{\partial}{\partial\theta^{(p)}}H(\theta \mid \theta^{(p)})$$

Donde :

$$\partial\theta^{(p)}H(\theta \mid \theta^{(p)}) = \partial^2\theta H(\theta \mid \theta^{(p)})$$

Por lo tanto:

$$\text{DM}=\frac{\partial^2}{\partial\theta}H(\theta^{*} \mid \theta^{*})\left[\frac{\partial^2}{\partial\theta}Q(\theta^{*} \mid \theta^{*})\right]^{-1}$$ □

3.2.1. Ejemplos

Distribución Multinomial

Este ejemplo es mostrado por D.L.R.(1977) para ilustrar de una manera sencilla el funcionamiento del algoritmo EM. De acuerdo a los datos presentados por Rao (1965, pp. 368-369) se tienen 197 animales que siguen la distribución Multinomial con cuatro categorias y_i:

$$Y = (y_1, y_2, y_3, y_4) = (125, 18, 20, 34)$$

Las probabilidades de pertenecer a cada una de las cuatro categorias están dadas por:

$$\Pi = \left(\tfrac{1}{2} + \tfrac{1}{4}\pi, \tfrac{1}{4}(1-\pi), \tfrac{1}{4}(1-\pi), \tfrac{1}{4}\pi\right), \text{ donde } 0 \leq \pi \leq 1$$

Entonces se tiene que la función de distribución de los 197 animales dados los valores de Π es:

$$g(y \mid \pi) = \frac{(y_1 + y_2 + y_3 + y_4)!}{y_1! y_2! y_3! y_4!} \left(\frac{1}{2} + \frac{1}{4}\pi\right)^{y_1} \left(\frac{1}{4}(1-\pi)\right)^{y_2} \left(\frac{1}{4}(1-\pi)\right)^{y_3} \left(\frac{1}{4}\pi\right)^{y_4} \tag{3.2.1.1}$$

Para aplicar el Algoritmo EM, se supone que la población se divide en 5 categorias x_i, en donde las primeras 2 categorías se obtienen partiendo la categoría y_1 en 2 nuevas categorias:

$$X = (x_1, x_2, x_3, x_4, x_5)$$

Con $y_1 = x_1 + x_2$, $y_2 = x_3$, $y_3 = x_4$ y $y_4 = x_5$.

Así, las probabilidades de pertenecer cada categoria son:

$$\Pi_x = \left(\tfrac{1}{2}, \tfrac{1}{4}\pi, \tfrac{1}{4}(1-\pi), \tfrac{1}{4}(1-\pi), \tfrac{1}{4}\pi\right)$$

Por lo tanto la función de distribución de X es:

$$f(x \mid \pi) = \frac{(x_1+x_2+x_3+x_4+x_5)!}{x_1!x_2!x_3!x_4!x_5!}\left(\frac{1}{2}\right)^{x_1}\left(\frac{1}{4}\pi\right)^{x_2}\left(\frac{1}{4}-\frac{\pi}{4}\right)^{x_3}\left(\frac{1}{4}-\frac{\pi}{4}\right)^{x_4}\left(\frac{1}{4}\pi\right)^{x_5} \tag{3.2.1.2}$$

El objetivo del Algoritmo EM es obtener π, el cual será calculado en cada iteración del algoritmo hasta que se alcance la convergencia.

Para ejecutar el paso-E se requiere de la estimación de algunas estadísticas suficientes de X dados los datos observados Y. En este caso los únicos valores desconocidos y por estimar son x_1 y x_2, que cumplen que $x_1+x_2 = 125$.

Está demostrado que cuando se tienen $X_1, \ldots, X_n$ variables aleatorias independientes con distribución Poisson de parámetros $\lambda_1, \ldots, \lambda_n$ no necesariamente iguales, la distribución condicional de $X = (X_1, \ldots, X_n)$ dado el total $X_1 + \ldots + X_n = n$ es Multinomial con parámetros n y $\pi = (\pi_1, \ldots, \pi_n)$, donde cada π_i es de la forma $\lambda_i/(\lambda_1, \ldots, \lambda_n)$.

Como x_1 y x_2 provienen de una distribución Poisson, se procede a calcular la esperanza condicional de x_1 dados los datos conocidos como sigue:

En primer lugar se busca obtener

$$P[x_1 = k \mid x_1+x_2 = n] = \frac{P[x_1 = k, x_1 + x_2 = n]}{P[x_1 + x_2 = n]} = \frac{P[x_1 = k]P[x_2 = n-k]}{P[x_1 + x_2 = n]} \tag{3.2.1.3}$$

En donde

$$\begin{aligned} P[x_1 = k] &= \frac{e^{-\pi_1}\pi_1^k}{k!} \\ P[x_2 = n-k] &= \frac{e^{-\pi_2}\pi_2^{n-k}}{(n-k)!} \end{aligned} \tag{3.2.1.4}$$

Además:

$$\begin{aligned} P[x_1 + x_2 = n] &= \sum_{k=0}^{n} \frac{e^{-\pi_1}\pi_1^k}{k!}\frac{e^{-\pi_2}\pi_2^{n-k}}{(n-k)!} \\ &= \frac{e^{-(\pi_1+\pi_2)}}{n!}\sum_{k=0}^{n}\frac{n!}{k!(n-k)!}\pi_1^k\pi_2^{n-k} \\ &= \frac{e^{-(\pi_1+\pi_2)}}{n!}\sum_{k=0}^{n}\binom{n}{k}\pi_1^k\pi_2^{n-k} \\ &= \frac{e^{-(\pi_1+\pi_2)}}{n!}(\pi_1+\pi_2)^n \end{aligned} \tag{3.2.1.5}$$

Completando (3.2.1.3) con (3.2.1.4) y (3.2.1.5):

$$\begin{aligned} P[x_1 = k \mid x_1 + x_2 = n] &= \frac{\dfrac{e^{-\pi_1}\pi_1^k}{k!}\dfrac{e^{-\pi_2}\pi_2^{n-k}}{(n-k)!}}{\dfrac{e^{-(\pi_1+\pi_2)}}{n!}(\pi_1+\pi_2)^n} \\ &= \binom{n}{k}\frac{\pi_1^k\pi_2^{n-k}}{(\pi_1+\pi_2)^n} \end{aligned} \tag{3.2.1.6}$$

La ecuación (3.2.1.6) pertenece a una distribución Binomial con parámetros n y $p = \frac{\pi_1}{\pi_1+\pi_2}$.

Una vez calculada la distribución condicional de x_1 y x_2 y sustituyendo los valores de n y π_i, la esperanza condicional en el paso-E se obtiene automáticamente:

$$\begin{aligned} E[x_1 \mid x_1 + x_2 = n] &= 125\frac{\frac{1}{2}}{\frac{1}{2}+\frac{1}{4\pi}} \\ E[x_2 \mid x_1 + x_2 = n] &= 125\frac{\frac{1}{4\pi}}{\frac{1}{2}+\frac{1}{4\pi}} \end{aligned} \tag{3.2.1.7}$$

En la k-ésima iteración del algoritmo el paso-E se ejecuta como:

$$\begin{aligned} x_1^{(k)} &= 125\frac{\frac{1}{2}}{\frac{1}{2}+\frac{1}{4}\pi^{(k)}} \\ x_2^{(k)} &= 125\frac{\frac{1}{4}\pi^{(k)}}{\frac{1}{2}+\frac{1}{4}\pi^{(k)}} \end{aligned} \tag{3.2.1.8}$$

El paso-M consiste en estimar π via máxima verosimilitud, utilizando los valores estimados en el paso-E y los valores ya observados, lo cual implica

que los valores estimados tomarán el papel de observados para efectuar la maximización.

El logaritmo de la función de verosimilitud está dado por:

$$\log L(\pi) = \log(x_1 + x_2 + 72)! - \log(x_1! x_2!\ 18!\ 20!\ 34!) + x_1 \log \frac{1}{2} + x_2 \log \frac{\pi}{4} + 18 \log \frac{1-\pi}{4} + 20 \log \frac{1-\pi}{4} + 34 \log \frac{\pi}{4} \tag{3.2.1.9}$$

Derivando (3.2.1.9) e igualando a cero:

$$\begin{aligned} &\frac{\partial \log L(\pi)}{\partial \pi} = \frac{\frac{x_2}{4}}{\frac{\pi}{4}} + \frac{\frac{-18}{4}}{\frac{1-\pi}{4}} + \frac{\frac{-20}{4}}{\frac{1-\pi}{4}} + \frac{\frac{34}{4}}{\frac{\pi}{4}} = 0 \\ &\implies \\ &\frac{x_2}{\pi} + \frac{34}{\pi} = \frac{18}{1-\pi} + \frac{20}{1-\pi} \\ &\frac{1}{\pi}(x_2 + 34) = \frac{1}{1-\pi}(18 + 20) \\ &x_2 + 34 = 18\pi + 34\pi + 20\pi + x_2\pi \\ &\pi = \frac{x_2 + 34}{x_2 + 18 + 20 + 34} \end{aligned} \tag{3.2.1.10}$$

Por lo tanto en el la k-ésima iteración el paso-M es:

$$\pi^{(k+1)} = \frac{x_2^{(k)} + 34}{x_2^{(k)} + 18 + 20 + 34} \tag{3.2.1.11}$$

A continuación se muestran los resultados de este procedimiento, elaborados en el Software R:

Cuadro 3.1: Resultados del Algoritmo EM para el ejemplo multinomial

Iteración	$\pi^{(k)}$	Error
0	0.5	0.108247400
1	0.6082474	0.016073630
2	0.6243211	0.002167829
3	0.6264889	0.000288443
4	0.6267773	3.83098E-05
5	0.6268156	5.08691E-06
6	0.6268207	6.75437E-07
7	0.6268214	8.96837E-08
8	0.6268215	1.19081E-08
9	0.6268215	1.58114E-09
10	0.6268215	2.09942E-10
11	0.6268215	2.78758E-11
12	0.6268215	3.70137E-12
13	0.6268215	4.91496E-13
14	0.6268215	6.51701E-14
15	0.6268215	8.65974E-15
16	0.6268215	1.22125E-15
17	0.6268215	1.11022E-16
18	0.6268215	–

Puede observarse que la verosimilitud nunca decrece en cada paso del algoritmo. Finalmente se logran obtener los estimadores para x_1 y x_2 que son 95.17 y 29.82 respectivamente.

Distribución Normal Univariada

Cuando se tienen datos faltantes, no es fácil maximizar la verosimilitud de los datos. Por ejemplo supongamos que se tiene un conjunto de datos con n variables y se desea estimar el total de una variable k, pero la variable presenta datos faltantes. Con la ausencia de algunos valores no es posible conocer el valor de la suma de la variable y en consecuencia tampoco la media. El algoritmo EM parte este problema en dos pasos. En el primero se supone que los valores de la media y la varianza de la población son conocidos, y asi se tiene un valor esperado estimado para el total de la variable. En el segundo paso el algoritmo utiliza el valor esperado estimado para obtener estimadores por máxima verosimilitud de la media y la varianza. El proceso

se repite hasta que las estimaciones convergen. En cada paso se asume que los datos faltantes son conocidos, en este caso el total de la variable, y así se estiman los parámetros restantes que son desconocidos.

Supongamos que se tienen $Y_1, \ldots, Y_n$ observaciones i.i.d. con distribución $N(\mu, \sigma^2)$, en donde $Y_1, \ldots, Y_k = Y_{obs}$ son observadas y $Y_{k+1}, \ldots, Y_n = Y_{miss}$ no fueron observadas, asumiendo que el patrón de datos faltantes es MAR. Sabemos que el valor esperado de cada Y_{miss} dados Y_{obs} y los parámetros $\theta = (\mu, \sigma^2)$ es μ y que $\sum_{i=1}^{n} Y_i$ y $\sum_{i=1}^{n} Y_i^2$ son estadísticas suficientes.

El paso-E consiste en calcular dos esperanzas condicionales:

$$E[\sum_{i=1}^{n} Y_i \mid \theta, Y_{obs}] \tag{3.2.1.12}$$

$$E[\sum_{i=1}^{n} Y_i^2 \mid \theta, Y_{obs}] \tag{3.2.1.13}$$

Dado que la esperanza es un operador lineal (3.2.1.12) se puede partir en la suma de dos esperanzas, la de los datos observados y la de los datos faltantes:

$$E[\sum_{i=1}^{k} Y_i \mid \theta, Y_{obs}] + E[\sum_{i=k+1}^{n} Y_i \mid \theta, Y_{obs}] \tag{3.2.1.14}$$

Desarrollando (3.2.1.14) se tiene que:

$$\begin{gathered} \sum_{i=1}^{k} E[Y_i \mid \theta, Y_{obs}] + \sum_{i=k+1}^{n} E[Y_i \mid \theta, Y_{obs}] \\ k\frac{\sum_{i=1}^{k} Y_i}{k} + (n-k)\mu^{(t)} \\ \sum_{i=1}^{k} Y_i + (n-k)\mu^{(t)} \end{gathered} \tag{3.2.1.15}$$

De la misma manera para (3.2.1.13):

$$
\begin{aligned}
E[\sum_{i=1}^{k} Y_i^2 \mid \theta, Y_{obs}] + E[\sum_{i=k+1}^{n} Y_i^2 \mid \theta, Y_{obs}] \\
\sum_{i=1}^{k} E[Y_i^2 \mid \theta, Y_{obs}] + \sum_{i=k+1}^{n} E[Y_i^2 \mid \theta, Y_{obs}] \\
k\left(\sigma^2 + \left(\frac{\sum_{i=1}^{k} Y_i}{k}\right)^2\right) + (n-k)\left[\left(\sigma^{(t)}\right)^2 + \left(\mu^{(t)}\right)^2\right] \\
k\left(\sigma^2 + \mu^2\right) + (n-k)\left[\left(\sigma^{(t)}\right)^2 + \left(\mu^{(t)}\right)^2\right] \\
k\left(\sum_{i=1}^{k} \frac{(Y_i - \mu)^2}{k} + \mu^2\right) + (n-k)\left[\left(\sigma^{(t)}\right)^2 + \left(\mu^{(t)}\right)^2\right] \\
k\left(\frac{1}{k}\sum_{i=1}^{k}\left(Y_i^2 - 2Y_i\mu + \mu^2\right) + \mu^2\right) + (n-k)\left[\left(\sigma^{(t)}\right)^2 + \left(\mu^{(t)}\right)^2\right] \\
\sum_{i=1}^{n} Y_i^2 - 2k\mu^2 + k\mu^2 + k\mu^2 + (n-k)\left[\left(\sigma^{(t)}\right)^2 + \left(\mu^{(t)}\right)^2\right] \\
\sum_{i=1}^{n} Y_i^2 + (n-k)\left[\left(\sigma^{(t)}\right)^2 + \left(\mu^{(t)}\right)^2\right]
\end{aligned}
\tag{3.2.1.16}
$$

El paso-M consiste en obtener los estimadores por por máxima verosimilitud de la media y la varianza, los cuales son:

$$\hat{\mu} = \sum_{i=1}^{n} \frac{Y_i}{n} \quad \hat{\sigma}^2 = \sum_{i=1}^{n} \left(\frac{Y_i - \mu}{n}\right)^2$$

Sí sustituimos (3.2.1.14) en (3.2.1.17), se obtiene el estimador por máxima verosimilitud de μ en la iteración $(t+1)$ del algoritmo EM.

$$\hat{\mu}^{(t+1)} = \frac{\sum_{i=1}^{k} Y_i + (n-k)\mu^{(t)}}{n} \tag{3.2.1.17}$$

Y como

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}\left(Y_i^2 - 2Y_i\mu + \mu^2\right) \\ &= \frac{\sum_{i=1}^{n} Y_i^2}{n} - 2\mu^2 + \mu^2 \\ &= \frac{\sum_{i=1}^{n} Y_i^2}{n} - \left(\frac{\sum_{i=1}^{n} Y_i}{n}\right)^2\end{aligned} \tag{3.2.1.18}$$

Entonces sustituyendo (3.2.1.16) en (3.2.1.18), se obtiene el valor del parámetro σ^2 en la iteración $(t+1)$:

$$\left(\hat{\sigma}^{(t+1)}\right)^2 = \frac{\sum_{i=1}^{k} Y_i^2 + (n-k)\left[\left(\sigma^{(t)}\right)^2 + \left(\mu^{(t)}\right)^2\right]}{n} - \left(\hat{\mu}^{(t+1)}\right)^2 \tag{3.2.1.19}$$

Para obtener los parámetros a los que converge el algoritmo fijamos $\hat{\mu}^{(t+1)} = \hat{\mu}^{(t)} = \hat{\mu}$ y $\left(\hat{\sigma}^{(t+1)}\right)^2 = \left(\hat{\sigma}^{(t)}\right)^2 = \hat{\sigma}^2$.
Entonces se tiene que la media converge a:

$$\begin{aligned}\hat{\mu} &= \frac{\sum_{i=1}^{k} Y_i}{n} + (n-k)\frac{\hat{\mu}}{n} \\ &= \frac{\sum_{i=1}^{k} Y_i}{n} + \hat{\mu} - k\frac{\hat{\mu}}{n} \Rightarrow \\ \hat{\mu} &= \frac{\sum_{i=1}^{k} Y_i}{k}\end{aligned}$$

Mientras que la varianza converge a:

$$\begin{aligned}\hat{\sigma}^2 &= \frac{\sum_{i=1}^{k} Y_i^2 + (n-k)\left[\hat{\sigma}^2 + \hat{\mu}^2\right]}{n} - \hat{\mu}^2 \\ &= \frac{\sum_{i=1}^{k} Y_i^2}{n} + \hat{\sigma}^2 - \frac{k}{n}\hat{\mu}^2 - \frac{k}{n}\hat{\sigma}^2 \Rightarrow \\ \frac{k}{n}\hat{\sigma}^2 &= \frac{\sum_{i=1}^{k} Y_i^2}{n} - \frac{k}{n}\hat{\mu}^2 \\ \hat{\sigma}^2 &= \frac{\sum_{i=1}^{k} Y_i^2}{k} - \hat{\mu}^2\end{aligned}$$

Estos parámetros corresponden a los estimadores calculados sólamente con la población observada. Es importante destacar que esto sucede gracias a que se está trabajando bajo el supuesto de que los datos faltantes son MAR.

En este caso la aplicación del algoritmo EM no es necesaria, puesto que se pueden obtener explicitamente los parámetros μ y σ^2, sin embargo sirve para ejemplificar claramente su funcionamiento.

Ejemplo Numérico
Supongamos que se tiene una muestra aleatoria de 100 observaciones que provienen de una distribución normal con media 5 y varianza 4. A este conjunto se le extrae aleatoriamente el 15 % de las observaciones. Se procede entonces a aplicar el algoritmo EM en el Software R para estimar la media y la varianza de el conjunto de datos con observaciones faltantes. Los resultados se muestran a continuación:

Cuadro 3.2: Resultados del Algoritmo EM para el ejemplo normal univariado

Iteración	$\mu^{(k)}$	$\sigma^{(k)}$	Error $\mu^{(k)}$	Error $\sigma^{(k)}$
0	4.986371	4	2.30E-07	2.30E-07
1	4.986371	4.118588	2.99E-08	2.99E-08
2	4.986371	4.134005	3.89E-09	3.89E-09
3	4.986371	4.136009	5.05E-10	5.05E-10
4	4.986371	4.136270	6.57E-11	6.57E-11
5	4.986371	4.136303	8.54E-12	8.54E-12
6	4.986371	4.136308	1.11E-12	1.11E-12
7	4.986371	4.136308	1.44E-13	1.44E-13
8	4.986371	4.136308	1.87E-14	1.87E-14
9	4.986371	4.136309	2.66E-15	2.66E-15
10	4.986371	4.136309	0	0

Puede observarse en el cuadro (3.2) que las estimaciones obtenidas por el algoritmo EM son muy similares a las reales, habiendo generado una media de 4.98 y una varianza de 4.13. Estos parámetros corresponden a la media y varianza de la población observada. En el caso de la media se tomó como punto inicial el estimador con los datos observados, mientras que para la varianza se tomó un valor arbirtrario. Si se toman como valores iniciales para la media y la varianza los estimadores de la población observada, el algoritmo converge a estos mismos.

Utilizar el algoritmo EM no produce valores para observaciones individuales faltantes. Con este método se obtienen estimadores de las medias y

varianzas de las variables de interés los cuales se utilizan como parámetros en modelos de estudio de la población.

Distribución Normal Bivariada

En este ejemplo se presenta la aplicación del Algoritmo EM en una distribución normal bivariada con un patrón de datos faltantes general en ambas variables:

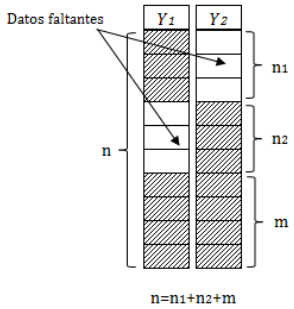


Figura 3.1: Patrón de datos faltantes para Y_1 y Y_2

Sean Y_1 y Y_2 dos variables aleatorias con densidad conjunta de la forma:

$$f(y_1, y_2) = \frac{1}{2\pi\sigma_{y_1}\sigma_{y_2}\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left(\frac{(y_1-\mu_1)^2}{\sigma_1^2}+\frac{(y_2-\mu_2)^2}{\sigma_2^2}-2\rho\frac{(y_1-\mu_1)(y_2-\mu_2)}{\sigma_1\sigma_2}\right)} \tag{3.2.1.20}$$

Donde $\rho = \frac{\sigma_{12}}{\sigma_1\sigma_2}$ es el coeficiente de correlación.

Llamémosle G_1 al grupo de observaciones con datos faltantes en Y_1, G_2 al grupo con observaciones faltantes en Y_2 y G_3 al grupo sin observaciones faltantes. El objetivo es estimar la media y la matriz de covarianza de Y_1 y Y_2 dados los valores observados.

En contraste con el ejemplo multinomial, no pueden rellenarse los valores faltantes individuales en el paso-E debido a que la función de verosimilitud no es lineal en los datos, sin embargo es lineal sobre las siguientes estadísticas suficientes:

$$s_1 = \sum_{j=1}^{n} y_{j1}, \quad s_2 = \sum_{j=1}^{n} y_{j2}, \quad s_{11} = \sum_{j=1}^{n} y_{j1}^2, \quad s_{22} = \sum_{j=1}^{n} y_{j2}^2, \quad s_{12} = \sum_{j=1}^{n} y_{j1} y_{j2}$$

(3.2.1.21)

La propiedad de linealidad es importante, ya que gracias a este hecho se puede dividir el problema de estimación en dos partes: la observada y la no observada. Las estadísticas suficientes en (3.2.1.21) se pueden desagregar de la siguiente manera:

El paso-E consiste en resolver las esperanzas condicionales de las estadísticas suficientes en (3.2.1.21) para cada uno de los grupos definidos anteriormente: G_1, G_2 y G_3. En el caso de G_3 la esperanza condicional corresponde a los valores observados. Para el grupo G_1 se requiere el cálculo de:

$$E[Y_{j1} \mid Y_{j2}, \mu, \Sigma], \quad E[Y_{j1}^2 \mid Y_{j2}, \mu, \Sigma] \text{ y } \quad E[Y_{j1} Y_{j2} \mid Y_{j2}, \mu, \Sigma] \qquad (3.2.1.22)$$

Mientras que para G_2 se requieren:

$$E[Y_{j2} \mid Y_{j1}, \mu, \Sigma], \quad E[Y_{j2}^2 \mid Y_{j1}, \mu, \Sigma] \text{ y } \quad E[Y_{j2} Y_{j1} \mid Y_{j1}, \mu, \Sigma] \qquad (3.2.1.23)$$

Para calcular las esperanzas en (3.2.1.22) y (3.2.1.23) se requiere de la densidad condicional de Y_1 dado Y_2 y la de Y_2 dado Y_1. Entonces procedemos a calcular:

$$
\begin{aligned}
f_{Y_2|Y_1}(y_2 \mid y_1) &= \frac{f(y_1, y_2)}{f_{Y_1}(y_1)} \\
&= C_1 f(y_1, y_2) \\
&= C_2 e^{-\frac{1}{2(1-\rho^2)}\left(\frac{(y_2-\mu_2)^2}{\sigma_2^2} - 2\rho\frac{(y_1-\mu_1)(y_2-\mu_2)}{\sigma_1\sigma_2}\right)} \\
&= C_3 e^{-\frac{1}{2(1-\rho^2)}\left(\frac{(y_2-\mu_2)^2}{\sigma_2^2} - 2\rho\frac{y_2(y_1-\mu_1)}{\sigma_1\sigma_2}\right)} \\
&= C_3 e^{-\frac{1}{2(1-\rho^2)}\left(\frac{y_2^2-2y_2\mu_2+\mu_2^2}{\sigma_2^2} - \frac{2\rho}{\sigma_1\sigma_2}(y_2y_1-y_2\mu_1)\right)} \\
&= C_3 e^{-\frac{1}{2(1-\rho^2)}\left(\frac{1}{\sigma_2^2}\left(y_2^2-2y_2\mu_2+\mu_2^2-2\rho\frac{\sigma_2}{\sigma_1}y_2y_1+2\rho\frac{\sigma_2}{\sigma_1}y_2\mu_1\right)\right)} \\
&= C_4 e^{-\frac{1}{2(1-\rho^2)}\left(\frac{1}{\sigma_2^2}\left(y_2^2-2y_2(\mu_2+\rho\frac{\sigma_2}{\sigma_1}y_1-\rho\frac{\sigma_2}{\sigma_1}\mu_1)\right)\right)} \\
&= C_4 e^{-\frac{1}{2(1-\rho^2)}\left(\frac{1}{\sigma_2^2}\left(y_2^2-2y_2(\mu_2+\rho\frac{\sigma_2}{\sigma_1}(y_1-\mu_1))\right)\right)}
\end{aligned}
$$

Completando el cuadrado se tiene que:

$$
= C_5 e^{-\frac{1}{2\sigma_2^2(1-\rho^2)}\left(y_2-\left(\mu_2+\rho\frac{\sigma_2}{\sigma_1}(y_1-\mu_1)\right)\right)^2} \tag{3.2.1.24}
$$

Donde $C_1, \ldots, C_5$ representan constantes que no dependen de y_2.

Por lo tanto por (3.2.1.24):
$f_{Y_2|Y_1}(y_2 \mid y_1) \sim N\left(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(y_1 - \mu_1), \sigma_2^2(1-\rho^2)\right)$

Análogamente:
$f_{Y_1|Y_2}(y_1 \mid y_2) \sim N\left(\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(y_2 - \mu_2), \sigma_1^2(1-\rho^2)\right)$

De esta manera el paso-E se resuelve sencillamente, obteniendo las siguientes esperanzas para cada parámetro:

$$
\begin{aligned}
E[Y_{j1} \mid Y_{j2}, \mu, \Sigma] &= \mu_1 + \rho\frac{\sigma_1}{\sigma_2}(y_2 - \mu_2) \\
E[Y_{j1}^2 \mid Y_{j2}, \mu, \Sigma] &= \left(\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(y_2 - \mu_2)\right)^2 + \sigma_1^2(1-\rho^2) \\
E[Y_{j1}Y_{j2} \mid Y_{j2}, \mu, \Sigma] &= Y_{j2}\left(\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(y_2 - \mu_2)\right) \\
E[Y_{j2} \mid Y_{j1}, \mu, \Sigma] &= \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(y_1 - \mu_1) \\
E[Y_{j2}^2 \mid Y_{j1}, \mu, \Sigma] &= \left(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(y_1 - \mu_1)\right)^2 + \sigma_2^2(1-\rho^2) \\
E[Y_{j2}Y_{j1} \mid Y_{j1}, \mu, \Sigma] &= Y_{j1}\left(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(y_1 - \mu_1)\right)
\end{aligned}
$$

El paso-M consistirá en maximizar la función de verosimilitud dados los datos observados. Para esto se procede a maximizar Q, que corresponde a la función de verosimilitud para los datos completos:

$$
\begin{aligned}
L(\theta) &= \prod_{j=1}^{n} f(y_1, y_2) \\
&= \frac{1}{2\pi\sigma_{y_1}\sigma_{y_2}\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\sum_{j=1}^{n}\left(\frac{(y_{j1}-\mu_1)^2}{\sigma_1^2} + \frac{(y_{j2}-\mu_2)^2}{\sigma_2^2} - 2\rho\frac{(y_{j1}-\mu_1)(y_{j2}-\mu_2)}{\sigma_1\sigma_2}\right)}
\end{aligned}
\tag{3.2.1.25}
$$

Esta verosimilitud puede ser escrita en forma matricial como sigue:

$$
L(\theta) = |2\pi\Sigma|^{-\frac{n}{2}}\, e^{-\frac{1}{2}\sum_{j=1}^{n}(Y_j - M)^T\Sigma^{-1}(Y_j - M)} \tag{3.2.1.26}
$$

Donde:

$$
\begin{aligned}
Y &= \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \\
M &= \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \\
\Sigma &= \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}
\end{aligned}
\tag{3.2.1.27}
$$

Entonces el logaritmo de la verosimilitud queda de la forma:

$$\log L(\theta) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log|\Sigma| - \frac{1}{2}\sum_{j=1}^{n}(Y_j - M)^T\Sigma^{-1}(Y_j - M)$$

Para obtener $\widehat{M}$ derivamos con respecto a M e igualamos a 0:

$$\frac{\partial}{\partial M}\log L(\theta) = -\frac{1}{2}2\Sigma^{-1}\sum_{j=1}^{n}(Y_j - M) = 0$$

$$\Sigma^{-1}\sum_{j=1}^{n}(Y_j - M) = 0$$

$$\Sigma^{-1}\sum_{j=1}^{n}Y_j = n\Sigma^{-1}M$$

$$\Sigma\Sigma^{-1}\sum_{j=1}^{n}Y_j = n\Sigma\Sigma^{-1}M$$

Por lo tanto:

$$\widehat{M} = \bar{Y_j} = \begin{pmatrix} \sum_{j=1}^{n}\frac{Y_{j1}}{n} \\ \\ \sum_{j=1}^{n}\frac{Y_{j2}}{n} \end{pmatrix}$$

Análogamente para obtener $\widehat{\Sigma}$:

$$\frac{\partial}{\partial\Sigma}\log L(\theta) = -\frac{n}{2}\Sigma^{-1} + \frac{1}{2}\Sigma^{-1}\sum_{j=1}^{n}(Y_j - M)(Y_j - M)^T\Sigma^{-1} = 0$$

$$n\Sigma^{-1} = \Sigma^{-1}\sum_{j=1}^{n}(Y_j - M)(Y_j - M)^T\Sigma^{-1}$$

$$n\Sigma\Sigma^{-1} = \Sigma\Sigma^{-1}\sum_{j=1}^{n}(Y_j - M)(Y_j - M)^T\Sigma^{-1}$$

$$nI\Sigma = \sum_{j=1}^{n}(Y_j - M)(Y_j - M)^T\Sigma^{-1}\Sigma$$

Por lo tanto:

$$\widehat{\Sigma} = \frac{\sum_{j=1}^{n}(Y_j - M)(Y_j - M)^T}{n} = \begin{pmatrix} \frac{1}{n}\sum_{j=1}^{n}(Y_{j1} - \mu_1)^2 & \frac{1}{n}\sum_{j=1}^{n}(Y_{j1} - \mu_1)(Y_{j2} - \mu_2) \\ \frac{1}{n}\sum_{j=1}^{n}(Y_{j1} - \mu_1)(Y_{j2} - \mu_2) & \frac{1}{n}\sum_{j=1}^{n}(Y_{j2} - \mu_2)^2 \end{pmatrix}$$

Finalmente el paso-E en la p-ésima iteración se efectúa como sigue:

Para Y_1:

$$E[\sum_j Y_{j1} \mid Y_{j2}, \mu^{(p)}, \Sigma^{(p)}] = \sum_{G_2 \cup G_3} Y_{j1} + \sum_{G_1} \left(\mu_1^{(p)} + \rho \frac{\sigma_1^{(p)}}{\sigma_2^{(p)}} (y_2 - \mu_2^{(p)}) \right)$$

$$E[\sum_j Y_{j1}^2 \mid Y_{j2}, \mu^{(p)}, \Sigma^{(p)}] = \sum_{G_2 \cup G_3} Y_{j1}^2 + \sum_{G_1} \left[\left(\mu_1^{(p)} + \rho^{(p)} \frac{\sigma_1^{(p)}}{\sigma_2^{(p)}} (y_2 - \mu_2^{(p)}) \right)^2 + \sigma_1^{2(p)} (1 - \rho^{2(p)}) \right]$$

$$E[\sum_j Y_{j1} Y_{j2} \mid Y_{j2}, \mu^{(p)}, \Sigma^{(p)}] = \sum_{G_1} \left[Y_{j2} \left(\mu_1^{(p)} + \rho^{(p)} \frac{\sigma_1^{(p)}}{\sigma_2^{(p)}} (y_2 - \mu_2^{(p)}) \right) \right]$$

Y los denotamos como $E_{11}^{(p)}$, $E_{12}^{(p)}$ y $E_{13}^{(p)}$ respectivamente.

Para Y_2:

$$E[\sum_j Y_{j2} \mid Y_{j1}, \mu^{(p)}, \Sigma^{(p)}] = \sum_{G_1 \cup G_3} Y_{j2} + \sum_{G_2} \left(\mu_2^{(p)} + \rho^{(p)} \frac{\sigma_2^{(p)}}{\sigma_1^{(p)}} (y_1 - \mu_1^{(p)}) \right)$$

$$E[\sum_j Y_{j2}^2 \mid Y_{j1}, \mu^{(p)}, \Sigma^{(p)}] = \sum_{G_1 \cup G_3} Y_{j2}^2 + \sum_{G_2} \left[\left(\mu_2 + \rho^{(p)} \frac{\sigma_2^{(p)}}{\sigma_1^{(p)}} (y_1 - \mu_1^{(p)}) \right)^2 + \sigma_2^{2(p)} (1 - \rho^{2(p)}) \right]$$

$$E[\sum_j Y_{j2} Y_{j1} \mid Y_{j1}, \mu^{(p)}, \Sigma^{(p)}] = \sum_{G_2} \left[Y_{j1} \left(\mu_2^{(p)} + \rho^{(p)} \frac{\sigma_2^{(p)}}{\sigma_1^{(p)}} (y_1 - \mu_1^{(p)}) \right) \right]$$

Y los denotamos como $E_{21}^{(p)}$, $E_{22}^{(p)}$ y $E_{23}^{(p)}$ respectivamente.

El paso-M consistirá en calcular:

Para Y_1:

$$\widehat{\mu_1}^{(p+1)} = \frac{1}{n} E_{11}^{(p)}$$

$$\widehat{\sigma_1}^{(p+1)} = \frac{1}{n} \left(E_{12}^{(p)} - 2\widehat{\mu_1}^{(p+1)} E_{11}^{(p)} + n \left(\widehat{\mu_1}^{(p+1)} \right)^2 \right) = \frac{1}{n} E_{12}^{(p)} - \left(\widehat{\mu_1}^{(p+1)} \right)^2$$

Para Y_2:

$$\widehat{\mu_2}^{(p+1)} = \frac{1}{n} E_{21}^{(p)}$$
$$\widehat{\sigma_2}^{(p+1)} = \frac{1}{n}\left(E_{22}^{(p)} - 2\widehat{\mu_2}^{(p+1)} E_{21}^{(p)} + n\left(\widehat{\mu_2}^{(p+1)}\right)^2\right) = \frac{1}{n} E_{22}^{(p)} - \left(\widehat{\mu_2}^{(p+1)}\right)^2$$

Y la covarianza:

$$\begin{aligned}\widehat{\sigma_{12}}^{(p+1)} &= \frac{1}{n}\left(\sum_{G_3} Y_{j1}Y_{j2} + E_{13}^{(p)} + E_{23}^{(p)} - E_{11}^{(p)}\widehat{\mu_2}^{(p+1)} + E_{21}^{(p)}\widehat{\mu_1}^{(p+1)} + n\widehat{\mu_1}^{(p+1)}\widehat{\mu_2}^{(p+1)}\right) \\ &= \frac{1}{n}\left(\sum_{G_3} Y_{j1}Y_{j2} + E_{13}^{(p)} + E_{23}^{(p)}\right) - \widehat{\mu_1}^{(p+1)}\widehat{\mu_2}^{(p+1)}\end{aligned}$$

A continuación se presentan dos ejemplos numéricos con diferentes estructuras de datos faltantes.

Caso 1: Valores perdidos en una variable
Este conjunto de datos es mostrado por Rubin(1987) y consiste en 10 observaciones provenientes de una normal bivariada en donde la variable 1 presenta dos valores faltantes, y la variable 2 no presenta datos faltantes.

Variable 1	10	14	16	15	20	4	18	22	–	–
Variable 2	8	11	16	18	6	4	20	25	9	13

En este caso, dado que la variable 2 es completamente observada, sus estimadores corresponden a los estimadores poblacionales. En consecuencia no se requiere el cómputo de dos densidades condicionales como se explicó anteriormente, sino que se requiere la densidad condicional de la variable 1 dados los valores de la variable 2.

Se programó en el Software R el algoritmo EM para efectuar la estimación. En los resultados que se muestran en el cuadro (3.3) se puede observar que los estimadores para la variable 1 se mantienen constantes durante cada iteración debido a que son completamente observados. Los valores iniciales del algoritmo se tomaron como los estimadores de cada variable excluyendo los valores ausentes, si bien, pueden elegirse de manera arbitraria, es recomendable elegirlos utilizando la información a priori.

Cuadro 3.3: Resultados del Algoritmo EM para caso 1

Iteración	$\mu_1^{(k)}$	$\mu_2^{(k)}$	$\sigma_1^{(k)}$	$\sigma_2^{(k)}$	$\rho^{(k)}$	$\sigma_1 2^{(k)}$
0	14.87500	13	32.98214	44.66667	0.6700216	25.71698
1	14.64470	13	27.19995	40.20000	0.6311760	20.87121
2	14.62127	13	26.83346	40.20000	0.6355207	20.87282
3	14.61656	13	26.76949	40.20000	0.6365672	20.88225
4	14.61553	13	26.75707	40.20000	0.6367837	20.88451
5	14.61530	13	26.75464	40.20000	0.6368280	20.88501
6	14.61525	13	26.75417	40.20000	0.6368370	20.88512
7	14.61524	13	26.75408	40.20000	0.6368389	20.88515
8	14.61524	13	26.75406	40.20000	0.6368392	20.88515
9	14.61523	13	26.75406	40.20000	0.6368393	20.88516
10	14.61523	13	26.75406	40.20000	0.6368393	20.88516
11	14.61523	13	26.75406	40.20000	0.6368393	20.88516
12	14.61523	13	26.75406	40.20000	0.6368393	20.88516
13	14.61523	13	26.75406	40.20000	0.6368393	20.88516
14	14.61523	13	26.75406	40.20000	0.6368393	20.88516
15	14.61523	13	26.75406	40.20000	0.6368393	20.88516
16	14.61523	13	26.75406	40.20000	0.6368393	20.88516
17	14.61523	13	26.75406	40.20000	0.6368393	20.88516
18	14.61523	13	26.75406	40.20000	0.6368393	20.88516
19	14.61523	13	26.75406	40.20000	0.6368393	20.88516
20	14.61523	13	26.75406	40.20000	0.6368393	20.88516

Caso 2: Valores perdidos en ambas variables

Este ejemplo es presentado por Murray(1977) como una contribución a la discusión sobre el trabajo de D.L.R.(1987). El autor muestra que con estos datos, la verosimilitud tiene un punto silla y dos máximos y que dependiendo de la elección de los puntos iniciales, el algoritmo EM converge a uno u otro punto. Los datos están conformados por 12 observaciones provenientes de una normal bivariada con medias iguales a cero y en donde ambas variables presentan valores perdidos. Cabe destacar que este patrón es igual al que se describe al principio de esta sección, por lo que el procedimiento para estimar los parámetros se realiza conforme a lo explicado anteriormete.

Variable 1	1	1	-1	-1	2	2	-2	-2	–	–	–	–
Variable 2	1	-1	1	-1	–	–	–	–	2	2	-2	-2

Una vez más se programó en el Software R el algoritmo EM para ejemplificar el comportamiento de los parámetros estimados. Los resultados se muestran a continuación para 3 distintos puntos iniciales:

Cuadro 3.4: Resultados del Algoritmo EM con punto inicial θ_0=(0, 0, 0.5, 0.5, 0.5)

Iteración	$\mu_1^{(k)}$	$\mu_2^{(k)}$	$\sigma_1^{(k)}$	$\sigma_2^{(k)}$	$\rho^{(k)}$
0	0	0	0.50000	0.50000	0.500000
1	0	0	2.12500	2.12500	0.627451
2	0	0	2.62106	2.62106	0.638369
3	0	0	2.72767	2.72767	0.624093
4	0	0	2.74108	2.74108	0.607151
5	0	0	2.73505	2.73505	0.591970
6	0	0	2.72611	2.72611	0.579063
7	0	0	2.71775	2.71775	0.568178
8	0	0	2.71057	2.71057	0.558976
9	0	0	2.70448	2.70448	0.551159
10	0	0	2.69934	2.69934	0.544487
11	0	0	2.69498	2.69498	0.538766
12	0	0	2.69126	2.69126	0.533843
13	0	0	2.68808	2.68808	0.529590
14	0	0	2.68534	2.68534	0.525907
15	0	0	2.68298	2.68298	0.522709
16	0	0	2.68094	2.68094	0.519926
17	0	0	2.67917	2.67917	0.517499
18	0	0	2.67763	2.67763	0.515380
19	0	0	2.67629	2.67629	0.513526
20	0	0	2.67512	2.67512	0.511903
21	0	0	2.67410	2.67410	0.510480
22	0	0	2.67321	2.67321	0.509232
23	0	0	2.67242	2.67242	0.508135
24	0	0	2.67173	2.67173	0.507171
25	0	0	2.67113	2.67113	0.506323

Cuadro 3.5: Resultados del Algoritmo EM con punto inicial θ_0=(0, 0, 1, 2, -0.1)

Iteración	$\mu_1^{(k)}$	$\mu_2^{(k)}$	$\sigma_1^{(k)}$	$\sigma_2^{(k)}$	$\rho^{(k)}$
0	0	0	1.00000	2.00000	-0.100000
1	0	0	2.00333	2.35333	-0.130265
2	0	0	2.34237	2.46438	-0.145051
3	0	0	2.45769	2.50036	-0.156086
4	0	0	2.49787	2.51286	-0.166142
5	0	0	2.51289	2.51819	-0.176124
6	0	0	2.51959	2.52147	-0.186336
7	0	0	2.52363	2.52430	-0.196871
8	0	0	2.52694	2.52718	-0.207747
9	0	0	2.53016	2.53025	-0.218951
10	0	0	2.53354	2.53357	-0.230454
11	0	0	2.53714	2.53715	-0.242219
12	0	0	2.54099	2.54099	-0.254199
13	0	0	2.54509	2.54509	-0.266342
14	0	0	2.54943	2.54943	-0.278590
15	0	0	2.55400	2.55400	-0.290879
16	0	0	2.55878	2.55878	-0.303143
17	0	0	2.56374	2.56374	-0.315313
18	0	0	2.56885	2.56885	-0.327320
19	0	0	2.57406	2.57406	-0.339096
20	0	0	2.57934	2.57934	-0.350576
21	0	0	2.58465	2.58465	-0.361701
22	0	0	2.58994	2.58994	-0.372417
23	0	0	2.59517	2.59517	-0.382677
24	0	0	2.60030	2.60030	-0.392444
25	0	0	2.60529	2.60529	-0.401689

Cuadro 3.6: Resultados del Algoritmo EM con punto inicial θ_0=(0, 0, 2 ,1, 0)

Iteración	$\mu_1^{(k)}$	$\mu_2^{(k)}$	$\sigma_1^{(k)}$	$\sigma_2^{(k)}$	$\rho^{(k)}$
0	0	0	2.00000	1.00000	0
1	0	0	2.33333	2.00000	0
2	0	0	2.44444	2.33333	0
3	0	0	2.48148	2.44444	0
4	0	0	2.49383	2.48148	0
5	0	0	2.49794	2.49383	0
6	0	0	2.49931	2.49794	0
7	0	0	2.49977	2.49931	0
8	0	0	2.49992	2.49977	0
9	0	0	2.49997	2.49992	0
10	0	0	2.49999	2.49997	0
11	0	0	2.50000	2.49999	0
12	0	0	2.50000	2.50000	0

Murray (1977) demuestra que el algoritmo EM converge al punto silla tomando como punto inicial $\rho = 0$ y para cualquier otro valor inicial para los parámetros restantes. El punto silla se genera para los valores $\rho = 0$ y $\sigma_1 = \sigma_2 = \frac{2}{5}$, mientras que los máximos se generan en los puntos $\rho = \pm\frac{1}{2}$ y $\sigma_1 = \sigma_2 = \frac{8}{3}$. Este ejemplo muestra que el algoritmo puede converger a diferentes puntos estacionarios, tal y como se mencionaba anteriormente. La solución a este problema es repetir el algoritmo para diferentes valores un suficiente número de veces y tomar los parámetros que resulten en la mayor proporción de casos. Para cualquier valor de $\rho \neq 0$ el algoritmo diverge del punto silla.

Distribución Normal Multivariada

El ejemplo anterior se puede generalizar a una distribución normal multivariada. El procedimiento a efectuar es análogo al del caso Bivariado.

En este caso suponemos que se tiene un conjunto de datos de tamaño n con k-variables $Y = (Y_1, \ldots, Y_k)$ y que tienen una distribución normal multivariada con media $\mu = (\mu_1, \ldots, \mu_k)$ y matriz de covarianza $\Sigma = (\sigma_{ik})$. La distribución de los datos completos Y pertenece a la familia exponencial con las estadísticas suficientes:

$\sum_{i=1}^{n} y_{ij}$ con $j = 1, \ldots, k$ y $\sum_{i=1}^{n} y_{ij} y_{ik}$ con $j, k = 1, \ldots, k$

El problema se divide en dos casos, y_{ij} es observado o y_{ij} no es observado de tal manera que el paso E consiste en calcular:

$E\left(\sum_{i=1}^{n} y_{ij} \mid Y_{obs}, \theta^{(p)}\right) = \sum_{i=1}^{n} y_{ij}^{(p)}$ y
$E\left(\sum_{i=1}^{n} y_{ij}y_{ik} \mid Y_{obs}, \theta^{(p)}\right) = \sum_{i=1}^{n} \left(y_{ij}^{(p)} y_{ik}^{(p)} + c_{jki}^{(p)}\right)$

Donde

$$y_{ij}^{(p)} = \begin{cases} y_{ij} & \text{si } y_{ij} \text{ es observada} \\ E\left(y_{ij} \mid Y_{obs_i}, \theta^{(p)}\right) & \text{si } y_{ij} \text{ es faltante} \end{cases}$$

y

$$c_{jki}^{(p)} = \begin{cases} 0 & \text{si } y_{ij} \text{ o } y_{ik} \text{ son observadas} \\ Cov\left(y_{ij}, y_{ik} \mid Y_{obs_i}, \theta^{(p)}\right) & \text{si } y_{ij} \text{ y } y_{ik} \text{ son faltantes} \end{cases}$$

El paso M consistirá en calcular los estimadores por máxima verosimilitud, donde la verosimilitud de una normal multivariada es de la forma:

$$L(\theta) = |2\pi\Sigma|^{-\frac{n}{2}} \, e^{-\frac{1}{2}\sum_{j=1}^{n}(Y_j - M)^T \Sigma^{-1}(Y_j - M)} \tag{3.2.1.28}$$

Donde:

$$\begin{aligned} Y &= \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \\ M &= \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix} \\ \Sigma &= \begin{pmatrix} \sigma_1^2 & \dots & \rho\sigma_1\sigma_k \\ \vdots & \ddots & \vdots \\ \rho\sigma_k\sigma_1 & \dots & \sigma_k^2 \end{pmatrix} \end{aligned} \tag{3.2.1.29}$$

Calculando el logaritmo de la verosimilitud se tiene lo siguiente:

$$\log L(\theta) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log|\Sigma| - \frac{1}{2}\sum_{j=1}^{n}(Y_j - M)^T\Sigma^{-1}(Y_j - M)$$

Derivando respecto a M e igualando a 0:

$$\frac{\partial}{\partial M}\log L(\theta) = -\frac{1}{2}2\Sigma^{-1}\sum_{j=1}^{n}(Y_j - M) = 0$$

$$\Sigma^{-1}\sum_{j=1}^{n}(Y_j - M) = 0$$

$$\Sigma^{-1}\sum_{j=1}^{n}Y_j = n\Sigma^{-1}M$$

$$\Sigma\Sigma^{-1}\sum_{j=1}^{n}Y_j = n\Sigma\Sigma^{-1}M$$

Por lo tanto:

$$\widehat{M} = \bar{Y_j} = \begin{pmatrix} \sum_{j=1}^{n}\frac{Y_{j1}}{n} \\ \\ \vdots \\ \sum_{j=1}^{n}\frac{Y_{jk}}{n} \end{pmatrix}$$

De forma similiar, derivando respecto a Σ:

$$\frac{\partial}{\partial\Sigma}\log L(\theta) = -\frac{n}{2}\Sigma^{-1} + \frac{1}{2}\Sigma^{-1}\sum_{j=1}^{n}(Y_j - M)(Y_j - M)^T\Sigma^{-1} = 0$$

$$n\Sigma^{-1} = \Sigma^{-1}\sum_{j=1}^{n}(Y_j - M)(Y_j - M)^T\Sigma^{-1}$$

$$n\Sigma\Sigma^{-1} = \Sigma\Sigma^{-1}\sum_{j=1}^{n}(Y_j - M)(Y_j - M)^T\Sigma^{-1}$$

$$nI\Sigma = \sum_{j=1}^{n}(Y_j - M)(Y_j - M)^T\Sigma^{-1}\Sigma$$

Por lo tanto:

$$\widehat{\Sigma} = \frac{\sum_{j=1}^{n}(Y_j - M)(Y_j - M)^T}{n}$$

$$= \begin{pmatrix} \frac{1}{n}\sum_{j=1}^{n}(Y_{j1} - \mu_1)^2 & \cdots & \frac{1}{n}\sum_{j=1}^{n}(Y_{j1} - \mu_1)(Y_{jk} - \mu_k) \\ \vdots & \ddots & \vdots \\ \frac{1}{n}\sum_{j=1}^{n}(Y_{j1} - \mu_1)(Y_{jk} - \mu_k) & \cdots & \frac{1}{n}\sum_{j=1}^{n}(Y_{jk} - \mu_k)^2 \end{pmatrix}$$

De esta manera, se tiene que el paso M consistirá en obtener:

$$\mu_j^{(p+1)} = \frac{1}{n}\sum_{i=1}^{n} y_{ij}^{(p)}$$
$$\sigma_{jk}^{(p+1)} = \frac{1}{n}E\left(\sum_{i=1}^{n} y_{ij}y_{ik} \mid Y_{obs}\right) - \mu_j^{(p+1)}\mu_k^{(p+1)}$$

Little y Rubin (1987) proponen cuatro alternativas para los parámetros iniciales en el caso multivariado:

1. Estimar los parámetros haciendo un Listwise Deletion: En este caso los valores iniciales serán consistentes en el caso de que los datos faltantes sean MCAR y se tengan al menos $k+1$ observaciones completamente observadas.

2. Estimar los parámetros haciendo un Pairwise Deletion: Este caso no es recomendado, pues puede haber un conflicto en la matriz de covarianzas, ya que como cada variable tendrá un número de valores observados diferente, la matriz puede no ser positiva definida.

3. Conformar las media y matriz de covarianzas muestrales completando los datos con algún método de imputación: Los estimadores de la matriz de covarianzas pueden resultar inconsistentes, sin embargo la matriz resulta positiva semidefinida, por lo que para comenzar con el algoritmo, funciona regularmente.

4. Estimar los parámetros con los valores observados de cada variable y comenzar con las correlaciones iguales a cero: Ocurre lo mismo que en el caso anterior.

En general, se pueden obtener computacionalmente diferentes parámetros iniciales para comenzar con el algoritmo EM, y los métodos para obtenerlos no necesariamente tienen que ser los propuestos. El objetivo es encontrar aquellos parámetros iniciales que mejor se ajusten a las circunstancias a prueba y error.

Cabe destacar que el algoritmo EM en el caso de una distribución Normal funciona bien, por el hecho de que las esperanzas condicionales del paso E, pueden ser interpretadas como los mejores predictores lineales de los valores no observados dado el parámetro obtenido en cada iteración. Además se

obtiene el factor de ajuste c_{jk} para la matriz de covarianzas, que incluye en las estimaciones el hecho de la ausencia de información.

Capítulo 4

Aplicación

En esta sección se mostrarán dos casos en los que puede aplicarse el Algoritmo EM para la estimación de valores perdidos. En el primer caso los valores perdidos son explícitos, mientras que en el segundo caso se tendrán datos faltantes latentes.

Ambas aplicaciones se eligieron porque representan una amplia gama de situaciones en las que puede utilizarse el Algoritmo EM. Tanto en la literatura como en problemas de la vida cotidiana es de gran interes encontrar soluciones para ambas clases de datos faltantes.

4.1. Caso 1: Datos faltantes explícitos

Se simularon 400 observaciones pertenecientes a cuatro variables con distribución normal multivariada, con vector de medias y matriz de covarianzas muestrales:

$$M = \begin{pmatrix} 29.09 \\ -0.92 \\ 55.38 \\ -0.25 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} 5.21 & 2.01 & 10.15 & 0.47 \\ 2.01 & 1.99 & 4.77 & -0.59 \\ 10.15 & 4.77 & 24.47 & -3.74 \\ 0.47 & -0.59 & -3.74 & 5.64 \end{pmatrix}$$

En este caso se mostrará la eficiencia del Algoritmo EM en comparación con algunos de los métodos de imputación mencionados en el capítulo 2.

El ejercicio consiste en eliminar información de manera aleatoria del conjunto de datos multivariado, aplicar diferentes métodos de eliminación e imputación, estimar parámetros y realizar un comparativo de los resultados.

Recordemos que el Algoritmo EM no es un método de imputación, sino un medio de estimación de parámetros cuando existen valores perdidos, sin embargo, se puede comparar su eficiencia con diferentes métodos de imputación y eliminación una vez que se obtiene el conjunto de datos completado.
Las cuatro variables que conforman el conjunto de datos presentan los siguientes porcentajes de datos faltantes:

Variable	% Valores Perdidos
V1	10
V2	15
V3	6
V4	30

Se efectuaron en el software R los siguientes métodos de eliminación e imputación para ser comparados con el Algoritmo EM:

1. Listwise Deletion
2. Pairwise Deletion
3. Imputación de la media
4. Imputación aleatoria
5. Regresión iterativa
6. Regresión iterativa estocástica
7. Imputación Múltiple

Los valores estimados de las medias para cada componente se muestran en el siguiente cuadro:

Método	μ_1	μ_2	μ_3	μ_4
Listwise	29.0726964	-9.7854814	55.4328533	-0.2880953
Pairwise	29.1218411	-9.9128168	55.3082139	-0.1969287
Imputación de la Media	29.1218411	-9.9128168	55.3082139	-0.1969287
Imputación Aleatoria	29.1853214	-9.9001322	55.2904607	-0.1627755
Regresión Iterativa	29.1390442	-9.9248722	55.3297273	-0.2068088
Regresión Iterativa Estocástica	29.1497535	-9.9109328	55.3462225	-0.2064243
Imputación Múltiple 1	29.0975606	-9.8742167	55.3862039	-0.1760776
Imputación Múltiple 2	29.0675875	-9.8781054	55.3616618	-0.1895219
Imputación Múltiple 3	29.0735267	-9.8947865	55.407493	-0.2067784
Imputación Múltiple 4	29.082719	-9.938572	55.384776	-0.19916
Algoritmo EM	29.1007251	-9.9102715	55.3787711	-0.1885119
Valores Muestrales	29.0929043	-9.9275803	55.385787	-0.2542211

Cuadro 4.1: Cuadro comparativo de las medias

Puede observarse que los resolutados obtenidos con el método Listwise se asemejan a las medias muestrales, sin embargo, en 3 de las 4 variables se sobreestiman las medias. El método Pairwise también resulta eficiente, pero hay que destacar que ambos métodos fueron efectivos dado el bajo porcentaje de valores perdidos y por la aleatoriedad de los mismos.

En el caso de las imputaciones por media y valores aleatorios se tienen desviaciones grandes respecto a los valores muestrales. Como se mencionó en el capítulo 2, estos métodos tienden a sesgar significativamente las estimaciones generando resultados que distan mucho de los valores originales.

Las diferencias son más marcadas para las variables 2 y 4, donde los porcentajes de valores faltantes son 15 y 30 respectivamente, lo que conlleva a concluir que entre mayor sea la cantidad de valores perdidos imputados, se generarán sesgos mayores. En cuanto a los métodos de regresión se puede concluir que son eficientes, dado que las estimaciones se asemejan más a los valores muestrales que con los métodos mencionados anteriormente por el hecho de incorporar información de las otras variables para la imputación. Por otra parte, mediante el método de imputación múltiple, se generaron 4 conjuntos de datos imputados para luego ser combinados en un único resultado, sin embargo, se puede analizar qué tan eficientes son las imputaciones de cada conjunto.

En este caso, las medias de los conjuntos de datos obtenidos se encuentran alrededor de las medias muestrales, por lo que al combinar las 4 imputaciones se tendrán resultados robustos y confiables. La desventaja de la imputación múltiple es que cada conjunto de datos generado depende de la aleatoriedad, por lo que se necesitan generar varios conjuntos de datos para poder alcanzar un resultado óptimo. El último método y el de mayor interés, el Algorimo EM, arroja los mejores resultados para las 4 variables en conjunto.

La mayor ventaja del algoritmo es que siempre proporciona estimadores únicos y estos son los mejores estimadores que pueden obtenerse. Los errores promedio de estimación de cada variable para cada método se muestran a continuación:

Método	Error Promedio
Listwise	0.0608
Pairwise	0.0446
Imputación de la Media	0.0446
Imputación Aleatoria	0.0767
Regresión Iterativa	0.0381
Regresión Iterativa Estocástica	0.0402
Imputación Múltiple 1	0.0341
Imputación Múltiple 2	0.0409
Imputación Múltiple 3	0.0303
Imputación Múltiple 4	0.0193
Algoritmo EM	0.0245

Cuadro 4.2: Errores promedio por método

Aunque el error promedio mínimo corresponde a la cuarta imputación múltiple, como se mencionó anteriormente, no indica que sea el mejor modelo, puesto que cada conjunto de datos es generado estocásticamente. Puede observarse que el Algoritmo EM presenta el error más bajo contemplando los errores de las 4 variables.

Para el caso de la matriz de covarianzas se tiene lo siguiente:

Listwise				Muestral			
5.3627	1.9397	10.3222	0.5231	5.2109	2.0083	10.1506	0.4661
1.9397	1.8658	4.5864	-0.6141	2.0083	1.9866	4.7735	-0.5914
10.3222	4.5864	24.6854	-3.6713	10.1506	4.7735	24.4700	-3.7372
0.5231	-0.6141	-3.6713	5.6772	0.4661	-0.5914	-3.7372	5.6379

Cuadro 4.3: Matriz de covarianzas

Pairwise				Muestral			
5.2151	1.9703	10.1173	0.6031	5.2109	2.0083	10.1506	0.4661
1.9703	1.9464	4.8555	-0.5822	2.0083	1.9866	4.7735	-0.5914
10.1173	4.8555	24.6869	-3.5787	10.1506	4.7735	24.4700	-3.7372
0.6031	-0.5822	-3.5787	5.4924	0.4661	-0.5914	-3.7372	5.6379

Cuadro 4.4: Matriz de covarianzas

Imputación de la media				Muestral			
4.6661	1.5006	8.4931	0.3929	5.2109	2.0083	10.1506	0.4661
1.5006	1.6586	3.9185	-0.3602	2.0083	1.9866	4.7735	-0.5914
8.4931	3.9185	23.2638	-2.3769	10.1506	4.7735	24.4700	-3.7372
0.3929	-0.3602	-2.3769	3.8956	0.4661	-0.5914	-3.7372	5.6379

Cuadro 4.5: Matriz de covarianzas

Imputación Aleatoria				Muestral			
5.1330	1.5549	8.4665	0.4751	5.2109	2.0083	10.1506	0.4661
1.5549	1.9595	4.1288	-0.4694	2.0083	1.9866	4.7735	-0.5914
8.4665	4.1288	24.4998	-2.5059	10.1506	4.7735	24.4700	-3.7372
0.4751	-0.4694	-2.5059	5.4733	0.4661	-0.5914	-3.7372	5.6379

Cuadro 4.6: Matriz de covarianzas

Regresión				Muestral			
5.2165	1.5002	8.5094	0.2892	5.2109	2.0083	10.1506	0.4661
1.5002	1.8004	3.9491	-0.3678	2.0083	1.9866	4.7735	-0.5914
8.5094	3.9491	24.5236	-2.5986	10.1506	4.7735	24.4700	-3.7372
0.2892	-0.3678	-2.5986	4.2381	0.4661	-0.5914	-3.7372	5.6379

Cuadro 4.7: Matriz de covarianzas

Regresión Estocástica				Muestral			
5.2525	1.4910	8.6180	0.2064	5.2109	2.0083	10.1506	0.4661
1.4910	2.0075	3.9142	-0.4325	2.0083	1.9866	4.7735	-0.5914
8.6180	3.9142	24.2724	-2.9225	10.1506	4.7735	24.4700	-3.7372
0.2064	-0.4325	-2.9225	5.5686	0.4661	-0.5914	-3.7372	5.6379

Cuadro 4.8: Matriz de covarianzas

Imputación Múltiple 1				Muestral			
5.2514	1.9694	10.2203	0.6581	5.2109	2.0083	10.1506	0.4661
1.9694	1.9427	4.7422	-0.5669	2.0083	1.9866	4.7735	-0.5914
10.2203	4.7422	24.6694	-3.1413	10.1506	4.7735	24.4700	-3.7372
0.6581	-0.5669	-3.1413	5.2413	0.4661	-0.5914	-3.7372	5.6379

Cuadro 4.9: Matriz de covarianzas

Imputación Múltiple 2				Muestral			
5.1412	1.9409	9.9423	0.3471	5.2109	2.0083	10.1506	0.4661
1.9409	1.9392	4.7242	-0.8021	2.0083	1.9866	4.7735	-0.5914
9.9423	4.7242	24.2697	-4.1931	10.1506	4.7735	24.4700	-3.7372
0.3471	-0.8021	-4.1931	5.8296	0.4661	-0.5914	-3.7372	5.6379

Cuadro 4.10: Matriz de covarianzas

Imputación Múltiple 3				Muestral			
5.1577	1.8811	9.9545	0.5013	5.2109	2.0083	10.1506	0.4661
1.8811	1.8816	4.5239	-0.6895	2.0083	1.9866	4.7735	-0.5914
9.9545	4.5239	24.2561	-3.7590	10.1506	4.7735	24.4700	-3.7372
0.5013	-0.6895	-3.7590	5.4924	0.4661	-0.5914	-3.7372	5.6379

Cuadro 4.11: Matriz de covarianzas

Imputación Múltiple 4				Muestral			
5.1872	1.8350	10.0856	0.3503	5.2109	2.0083	10.1506	0.4661
1.8350	1.9222	4.4803	-0.6622	2.0083	1.9866	4.7735	-0.5914
10.0856	4.4803	24.4774	-3.9980	10.1506	4.7735	24.4700	-3.7372
0.3503	-0.6622	-3.9980	5.5198	0.4661	-0.5914	-3.7372	5.6379

Cuadro 4.12: Matriz de covarianzas

Algoritmo EM				Muestral			
5.1924	1.9222	10.0664	0.4894	5.2109	2.0083	10.1506	0.4661
1.9222	1.9258	4.6740	-0.6925	2.0083	1.9866	4.7735	-0.5914
10.0664	4.6740	24.4518	-3.7682	10.1506	4.7735	24.4700	-3.7372
0.4894	-0.6925	-3.7682	5.5872	0.4661	-0.5914	-3.7372	5.6379

Cuadro 4.13: Matriz de covarianzas

La coherencia en la estimación de la matriz de covarianzas es crucial para determinar la eficiencia de los métodos, ya que ésta es utilizada en diversos análisis estadísticos posteriores a la imputación. Las mejores estimaciones son generadas por el Algoritmo EM y son únicas, es decir, siempre que se replique el algorimo se obtendrán los mismos resultados, lo cual representa una gran ventaja sobre el método de imputación múltiple. En los casos de ambos tipos de regresión, puede notarse que las covarianzas (en valor absoluto) entre las 4 variables son siempre subestimadas en una proporción mayor que las covarianzas arrojadas tanto por el Algoritmo EM como por cada una de las 4 imputaciones múltiples. Inclusive, los métodos de eliminación presentan valores no tan alejados de los datos reales, a diferencia de los métodos de imputación por media y aleatorio, que presentan valores

subestimados para las varianzas y covarianzas como se había mencionado en el capítulo 2.

4.2. Caso 2: Datos faltantes latentes

El objetivo de esta sección es mostrar la funcionalidad del Algorimo EM en la clasificación de datos. En ámbitos como la fotografía o la medicina es común que se requiera clasificar en grupos a los elementos de una población a partir de sus características. En este caso, podemos definir como datos faltantes o latentes a las probabilidades de pertenencia de los sujetos a un determinado grupo y estimar simultáneamente los parámetros que describen la distribución de cada grupo.

El conjunto de datos a clasificar consiste en los créditos de la cartera de Consumo No Revolvente de facturación mensual otorgados por instituciones bancarias del sistema financiero mexicano. La finalidad de este ejercicio es clasificar a partir de características de comportamiento crediticio cada uno de los créditos en dos categorías: *crédito bueno* y *crédito malo*. Definiremos un *crédito malo* como aquel el en donde el acreditado incumple con sus obligaciones de pago durante 90 días o más, y aun *crédito bueno* donde el acreditado no presenta esta característica. Con este modelo se espera que, dependiendo de la distribución de la varible de comportamiento, un crédito pertenezca al grupo de créditos buenos o al de créditos malos.

Las variables que describen el comportamiento de los créditos se construyeron a partir de los insumos que se muestran a continuación y cuya descripción se muestra en el artículo 91 de la Circular Única de Bancos:

Monto Exigible
Monto que corresponde cubrir al acreditado en el Periodo de Facturación pactado. Tratándose de créditos con Periodos de Facturación semanal y quincenal, no se deberá incluir el acumulado de importes exigibles anteriores no pagados. Para créditos con Periodo de Facturación mensual, el Monto Exigible deberá considerar tanto el importe correspondiente al mes como los importes exigibles anteriores no pagados, si los hubiera. Las bonificaciones y descuentos podrán disminuir el Monto Exigible, únicamente cuando el acreditado cumpla con las condiciones requeridas en el contrato crediticio para la realización de los mismos.

Pago Realizado
Monto correspondiente a la suma de los pagos realizados por el acreditado en el Periodo de Facturación. No se consideran pagos a los castigos, quitas, condonaciones, bonificaciones y descuentos que se efectúen al crédito o grupo de créditos. El valor de esta variable deberá ser mayor o igual a cero.

Días de Atraso
Número de días naturales a la fecha de la calificación, durante los cuales el acreditado no haya liquidado en su totalidad el Monto Exigible en los términos pactados originalmente.

Plazo Total
Número de Periodos de Facturación (semanales, quincenales o mensuales) establecido contractualmente en el que debe liquidarse el crédito.

Plazo Remanente
Número de Periodos de Facturación semanales, quincenales o mensuales que, de acuerdo con lo establecido contractualmente, resta para liquidar el crédito a la fecha de calificación de la cartera. En el caso de créditos cuya fecha de vencimiento hubiera pasado sin que el acreditado realizara la liquidación correspondiente, el plazo remanente deberá ser igual al Plazo Total del crédito.

Importe Original del Crédito
Monto correspondiente al importe total del crédito en el momento de su otorgamiento.

Valor Original del Bien
Monto correspondiente al valor del bien financiado que tenga la institución registrado en el momento del otorgamiento del crédito. En caso de que el crédito no sea para financiar la compra o adquisición de un bien, el Valor Original del Bien será igual al Importe Original del Crédito. Asimismo, se podrá utilizar el Importe Original del Crédito para créditos que no cuenten con el Valor Original del Bien y que hayan sido otorgados con anterioridad a la entrada en vigor de las presentes disposiciones.

Saldo del Crédito
Al saldo insoluto a la fecha de la calificación, el cual representa el monto de crédito efectivamente otorgado al acreditado, ajustado por los intereses devengados, menos los pagos al seguro que en su caso se hubiera financiado, los cobros de principal e intereses, así como por las quitas, condonaciones,

bonificaciones y descuentos que en su caso se hayan otorgado. En todo caso, el monto sujeto a la calificación no deberá incluir los intereses devengados no cobrados, reconocidos en cuentas de orden dentro del balance, de créditos que estén en cartera vencida.

Ingreso
Ingreso percibido por el acreditado al momento de la originación del crédito.

A partir de los insumos mencionados anteriormente, se construyeron 10 variables que describen el comportamiento de pago de cada crédito. Su descripción se muestra a continuación:

- EXG_ING_3M: Promedio de la proporción que representa el monto exigible respecto al ingreso del acreditado durante los últimos 3 meses.

- EXG_ING_6M: Promedio de la proporción que representa el monto exigible respecto al ingreso del acreditado durante los últimos 6 meses.

- PAGO_EXG_3M: Promedio de la proporción que representa el pago realizado respecto al monto exigible durante los últimos 3 meses.

- PAGO_EXG_6M: Promedio de la proporción que representa el pago realizado respecto al monto exigible durante los últimos 6 meses.

- VECES_PAGO: Número de veces que el acreditado paga el valor del bien o el importe original del crédito, en caso de que no exista bien financiado. Se obtiene como el cociente de la suma de todos los pagos programados del crédito y el valor del bien.

- IMP_INGR: Proporción que representa el importe original del crédito respecto al ingreso del acreditado.

- SALDO_IMP: Proporción que representa el saldo del crédito al momento de la calificación respecto al importe otorgado originalmente.

- VAL_BIEN_ING: Proporción que representa el valor del bien financiado respecto al ingreso del acreditado.

- PAGO_SALDO_T0: Proporción que representa el monto exigible respecto al saldo del crédito al momento de la calificación.

- PAGO_SALDO_3M: Promedio de la proporción que representa el monto exigible respecto al saldo del crédito durante los últimos 3 meses.

Si cada variable está relacionada con el incumplimiento de un crédito, se esperaría tener dos posibles comportamientos o distribuciones dentro de la misma variable que diferencien entre un comportamiento bueno de uno malo. Siguiendo esta línea, estaríamos hablando de que se tienen mezclas de poblaciones con diferentes comportamientos y podemos utilizar en algoritmo EM para estimar los parámetros que las describen, así como las categorías a las que pertenecen. En caso de no encontrar un modelo de mezclas que se ajuste a alguna variable, se podrá concluir que la variable no funciona para diferenciar un crédito bueno de uno malo.

4.2.1. Análisis de los datos

Para trabajar con las variables construidas, se realizaron varios filtros numéricos debido a que se presentaron valores extremos y perdidos originados por un mal reporte de la estimación. En estos casos no se consideró conveniente realizar una estimación para estos valores dado que representaban una cantidad no significativa dentro de la población y se desconoce el origen de los errores. Por lo tanto, se incluyeron en el ejercicio únicamente los créditos que cumplían con las siguientes condiciones:

1. PAGO_EXG_3M ≤ 3

2. PAGO_EXG_6M ≤ 3

3. IMP_INGR ≤ 3

4. VAL_BIEN_ING ≤ 3

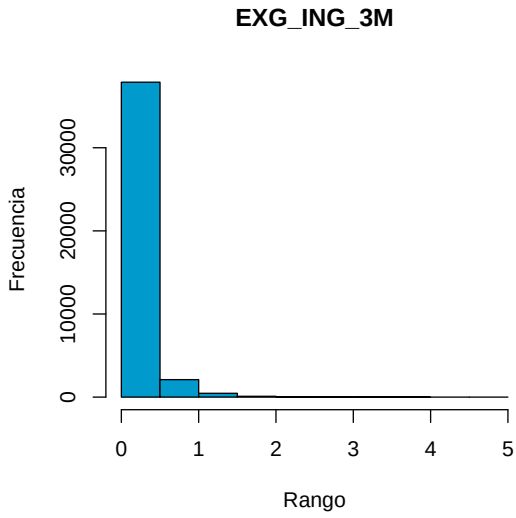
Con la finalidad de tener información suficiente para construir las variables históricas, se consideraron solo aquellos créditos con una antigüedad mayor o igual a 6 meses. De ésta manera, el ejercicio se realizó con un total de 40,580 créditos vigentes entre septiembre del 2005 y agosto del 2008.

A continuación se presenta un análisis descriptivo para cada una de las variables que conforman el ejercicio:

Exigible respecto al ingreso durante los últimos 3 meses (EXG_ING_3M)

Intuitivamente, se espera que si el monto exigible representa una proporción pequeña de su ingreso, el acreditado no presente dificultades para saldar su deuda. Al tomar la historia de 3 meses, podría localizarse un posible deterioro o endeudamiento del acreditado, lo cual desembocaría en el impago del crédito. Esta variable también permite analizar si la deuda adquirida por un individuo es coherente con su nivel de ingreso, ya que de lo contrario es altamente probable que el crédito no pueda ser pagado. En el histograma puede observarse que la mayor parte de los exigibles representan menos del 20 por ciento del salario, por lo que los montos fuera de este rango podrían generar una señal de alerta.

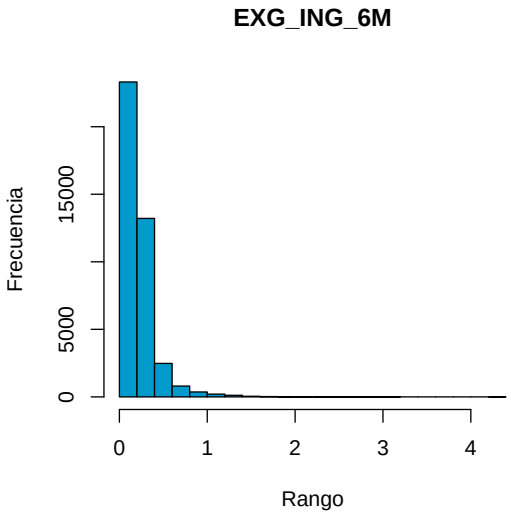
	Variable 1
Mínimo	0.00
Q_{25}	0.08
Mediana	0.15
Media	0.16
Q_{75}	0.22
Máximo	2.35



Exigible respecto al ingreso durante los últimos 6 meses (EXG_ING_6M)

Como en el caso de la variable anterior, se busca medir el endeudamiento del acreditado. Al tomar una historia más larga se puede capturar mejor el deterioro. Si bien, el salario no es una variable que cambie a corto plazo, una aproximación del monto exigible al salario implicaría que el acreditado no está pagando una cantidad suficiente tal que le permita mantener su deuda en un nivel estable. La distribución de esta variable es muy similar a la que sólo involucra 3 meses de historia.

	Variable 2
Mínimo	0.00
Q_{25}	0.08
Mediana	0.15
Media	0.17
Q_{75}	0.23
Máximo	1.73

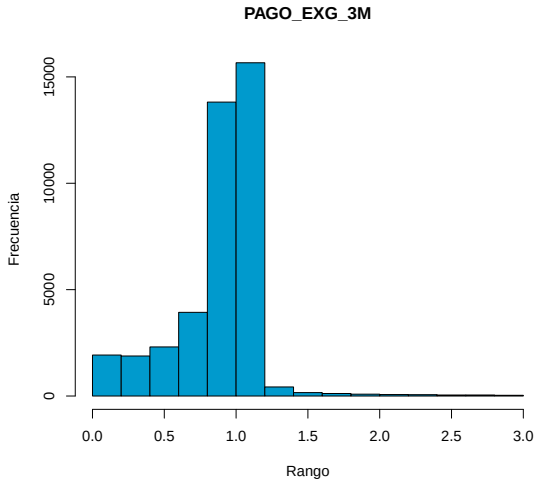


Pago respecto al monto exigible durante los últimos 3 meses (PAGO_EXG_3M)

El sentido de esta variable radica en capturar si un acreditado paga a penas lo suficiente, paga más de lo suficiente o simplemente no paga. Valores cercanos a cero de esta variable indican que el acreditado está incumpliendo con sus obligaciones de pago, mientras que valores mayores o iguales a 1 indican un excelente comportamiento de pago. Es importante considerar esta variable de manera histórica para incorporar deterioros de pago, ya que el hecho de que un individuo pague puntualmente su deuda en un mes, no garantiza que el mes siguiente cumpla con sus obligaciones de pago, puesto que está sometido a factores externos y que sólo pueden ser capturados

considerando un periodo histórico.

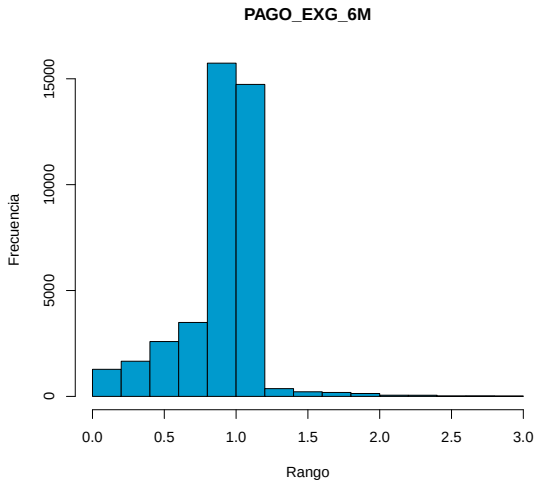
	Variable 3
Mínimo	0.00
Q_{25}	0.81
Mediana	1.00
Media	0.88
Q_{75}	1.00
Máximo	2.99



Pago respecto al monto exigible durante los últimos 6 meses (PAGO_EXG_6M)

Al considerar una historia más larga puede penalizarse más el incumplimiento de pago del individuo, identificando a a quellos que no pagan el mínimo exigible durante varios periodos, o bien, premiarse la constancia de pago. Esto permitiría diferenciar mejor un crédito malo de uno bueno.

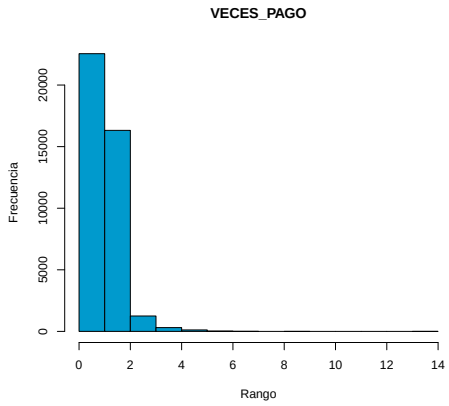
	Variable 4
Mínimo	0.000
Q_{25}	0.833
Mediana	1.000
Media	0.892
Q_{75}	1.009
Máximo	2.999



Veces que se paga el importe del crédito (VECES_PAGO)

La finalidad de esta variable es encontrar aquellos casos en donde el individuo termina pagando un importe mucho mayor al importe originalmente pactado, lo cual es un indicio de deterioro en la calidad crediticia, ya que mientras más reducido haya sido el pago del crédito durante todda su historia, la deuda se irá incrementando a lo largo del tiempo.

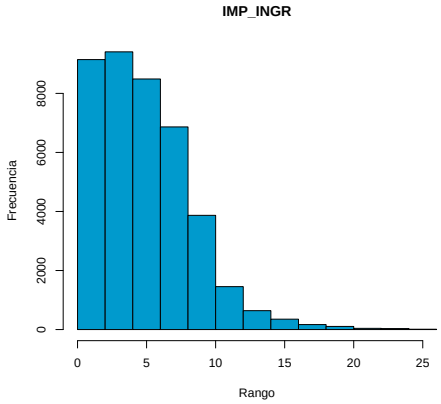
	Variable 5
Mínimo	0.043
Q_{25}	0.579
Mediana	0.913
Media	1.002
Q_{75}	1.310
Máximo	13.65



Importe del crédito respecto al ingreso (IMP_INGR)

Esta variable busca relacionar el tamaño de la deuda adquirida respecto al ingreso del individuo con su calidad crediticia. Aquellos créditos cuyo importe exceda considerablemente el salario del acreditado serán más riesgosos, ya que los pagos a los que el individuo se hace acreedor podrían superar los ingresos con los que cuenta para solventar su deuda. Se espera que los individuos que adquieren una deuda de manera conservadora cumplan de manera más puntual con sus obligaciones de pago.

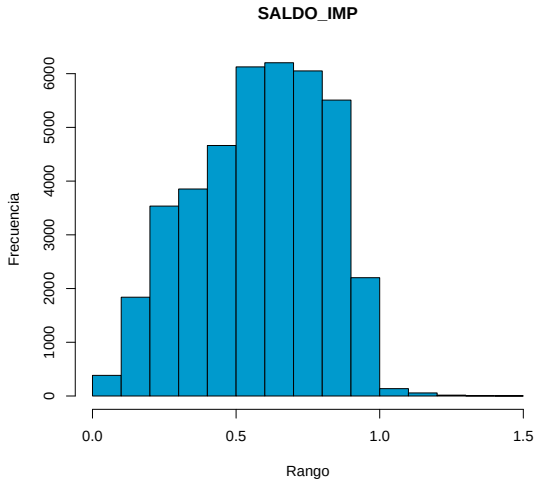
	Variable 6
Mínimo	0.013
Q_{25}	2.244
Mediana	4.407
Media	4.905
Q_{75}	6.883
Máximo	25.00



Saldo respecto al importe original del crédito (SALDO_IMP)

Al medir la proporción que representa el saldo del crédito respecto al importe original del mismo se busca capturar el tamaño de la deuda del acreditado al momento de evaluarlo. Los individuos que estén cerca de liquidar el crédito serán menos riesgosos que aquellos que han pagado una proporción menor de su deuda.

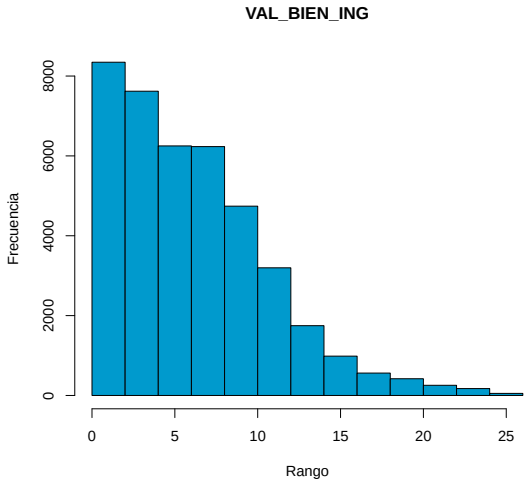
	Variable 7
Mínimo	9.9e-06
Q_{25}	0.412
Mediana	0.598
Media	0.581
Q_{75}	0.765
Máximo	1.407



Valor del bien respecto al ingreso (VAL_BIEN_ING)

Esta variable tiene el mismo objetivo que la variable 6. Se busca discriminar aquellos individuos que adquieren deuda de manera conservadora, de aquellos que adquieren créditos que sobrepasan sus posibilidades de pago.

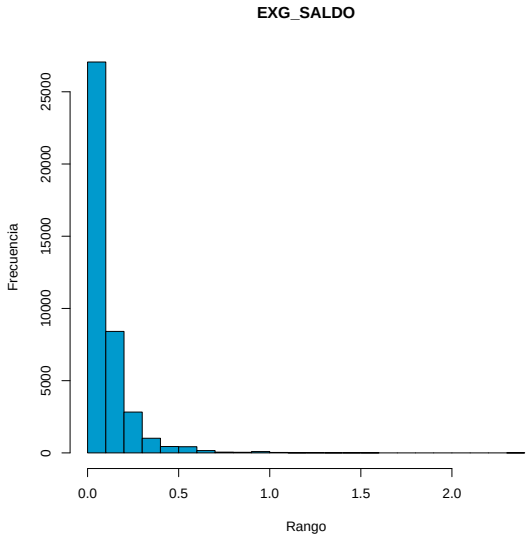
	Variable 8
Mínimo	0.13
Q_{25}	2.50
Mediana	5.38
Media	6.14
Q_{75}	8.73
Máximo	25.00



Pago realizado respecto al saldo del crédito (EXG_SALDO_T0)

Esta variable puede considerarse como una medida del uso del crédito y de la misma deuda. Valores cercanos a cero podrían indicar un comportamiento de pago estable, mientras que si la variable sobrepasa el valor de 1, implicaría que la deuda del acreditado se ha incrementado y esto podría originar un incumplimiento.

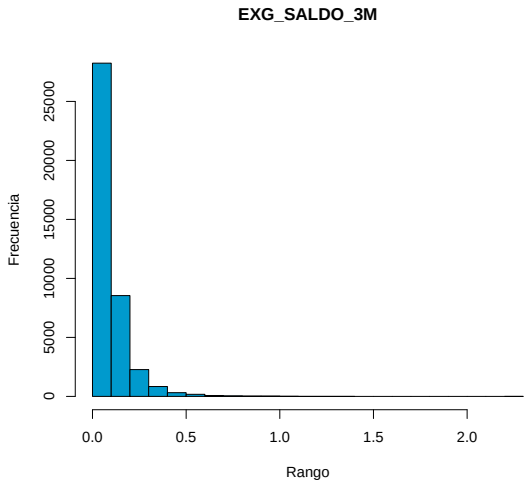
	Variable 9
Mínimo	0.008
Q_{25}	0.049
Mediana	0.744
Media	0.111
Q_{75}	0.121
Máximo	2.365



Pago realizado respecto al saldo del crédito durante los últimos 3 meses (EXG_SALDO_3M)

Esta variable tiene un comportamiento análogo al de la variable anterior, con la diferencia de que al tomar un periodo histórico podría capturarse con mayor precisión la inestabilidad de un crédito.

	Variable 10
Mínimo	0.008
Q_{25}	0.048
Mediana	0.071
Media	0.100
Q_{75}	0.113
Máximo	2.249



Puede observarse que todas las variables toman valores positivos, aunque en algunos casos no se aprecia la presencia de dos componentes, se pueden llegar a encontrar ambas poblaciones en caso de existir dos distribuciones implícitas. En los casos donde sea evidente la precencia de dos poblaciones distintas, el error de clasificación se espera sea sustancialmente menor, en comparación con aquellos casos donde no se puede apreciar claramente la diferencia.

4.2.2. Resultados del Algoritmo EM

Para ajustar el modelo de mezclas a cada variable, se utilizó el paquete *Mixtools* del Software R. Este paquete utiliza el algoritmo EM para estimar los parámetros y las probabilidades predictivas por categoría, además permite controlar los parámetros iniciales, las iteraciones y el criterio de convergencia. El detalle del funcionamiento del Algoritmo EM para el caso de mezclas se muestra en la sección de Anexos.
Dados los posibles valores que toma cada variable, se consideró apropiado ajustar componentes de distribuciones Gamma. Una vez ajustadas las mezclas, se compararán las categorías predichas por el algoritmo contra las categorías reales mediante tablas de contingencia. Mientras más grande sea el porcentaje de clasificación correcto, la variable tendrá un mejor poder predictivo de manera individual.

Pago respecto al monto exigible durante los últimos 3 meses (PAGO_EXG_3M)

El algoritmo convergió en 14 iteraciones para esta variable. Las componentes ajustadas y sus respectivos parámetros se muestran a continuación:

Variable 3	Mezcla 1	Mezcla 2
Proporción	0.24	0.75
α	0.61	69.99
β	0.88	0.014

Cuadro 4.14: Parámetros estimados

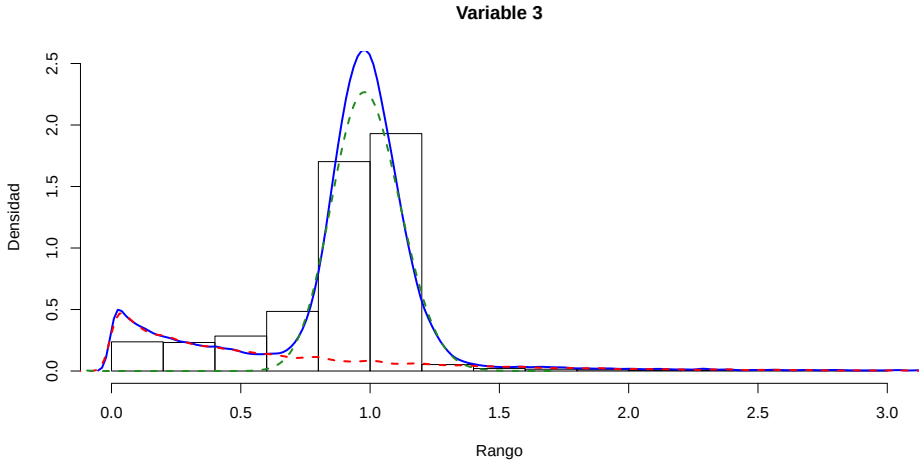


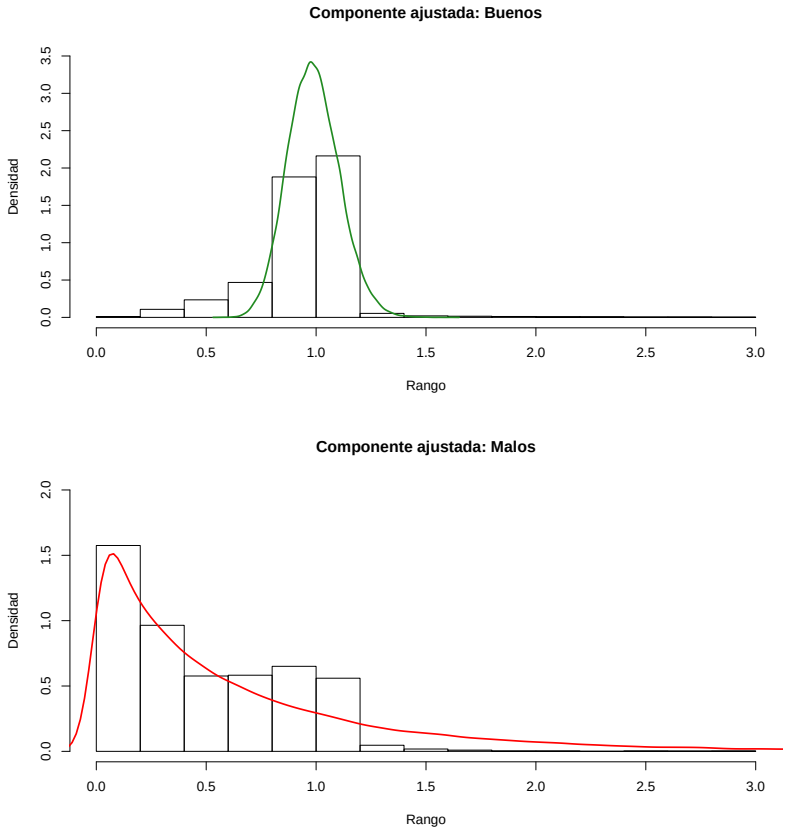
Figura 4.1: Componentes ajustadas

La tabla de contingencia muestra buenos resultados en cuanto a clasificación de buenos y malos, con una precisión del 83 %. En este caso, el valor esperado de ambas componentes es muy distinto, teniendo un valor cercano a 0 para la población de malos y mayor a 0.5 para los buenos. Los parámetros siguen la idea intuitiva que se planteó en el análisis descriptivo de la variable.

Variable 3	Buenos Reales	Malos Reales
Buenos Predicción	85 %	30 %
Malos Predicción	15 %	70 %

Cuadro 4.15: Tabla de clasificación

Si se extraen las poblaciones de buenos y malos reales respectivamente y se ajusta una mezcla simulada a partir de los parámetros obtenidos con el Algoritmo EM, se puede apreciar gráficamente la precisión del ajuste, además de que existe una diferencia marcada entre las distribuciones de ambas poblaciones:



Pago respecto al monto exigible durante los últimos 6 meses (PAGO_EXG_6M)

El algoritmo convergió en 17 iteraciones para esta variable. Las componentes ajustadas y sus respectivos parámetros se muestran a continuación:

Variable 4	Mezcla 1	Mezcla 2
Proporción	0.22	0.77
α	0.76	75.42
β	0.83	0.013

Cuadro 4.16: Parámetros estimados

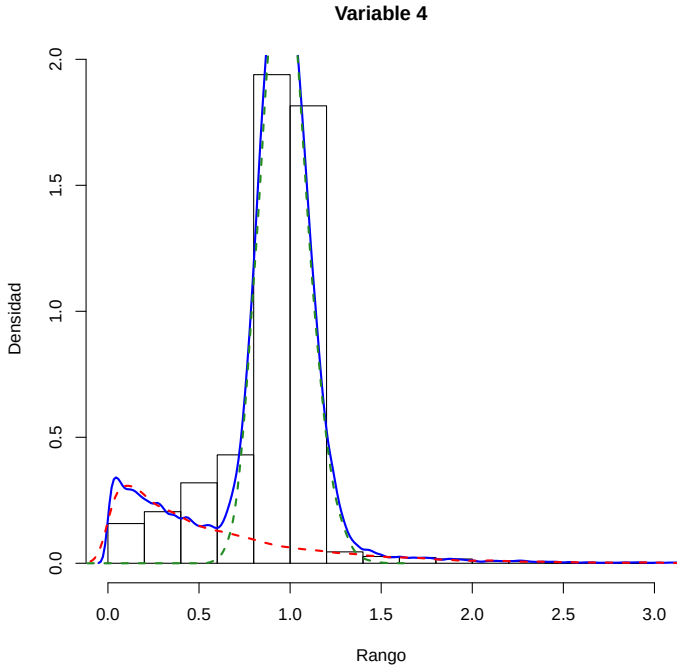


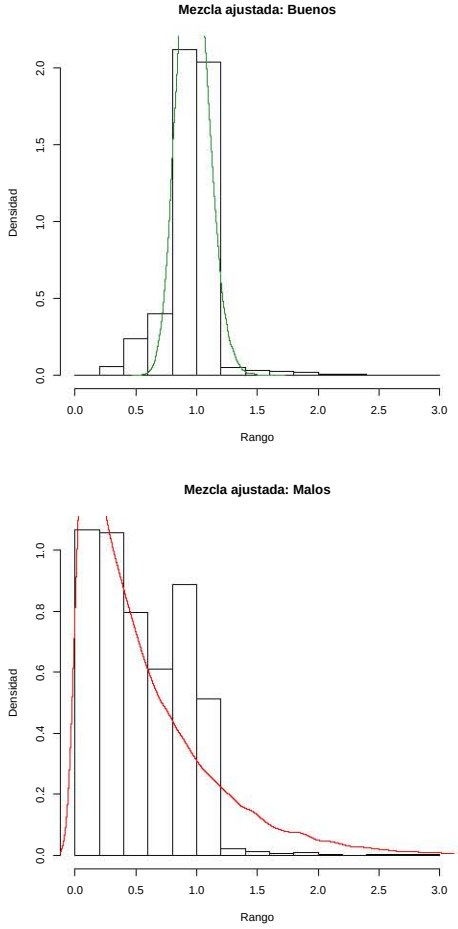
Figura 4.3: Componentes ajustadas

La tabla de contingencia, como en el caso anterior, muestra buenos resultados en cuanto a clasificación de buenos y malos, con una precisión del 84 %.

Variable 4	Buenos Reales	Malos Reales
Buenos Predicción	88 %	33 %
Malos Predicción	12 %	67 %

Cuadro 4.17: Tabla de clasificación

Si se extraen las poblaciones de buenos y malos reales respectivamente y se ajusta una mezcla simulada a partir de los parámetros obtenidos con el Algoritmo EM, se puede apreciar gráficamente la precisión del ajuste, además de que existe una diferencia marcada entre las distribuciones de ambas poblaciones:



Pago realizado respecto al saldo del crédito (EXG_SALDO_T0)

El algoritmo convergió en 56 iteraciones para esta variable. Las componentes ajustadas y sus respectivos parámetros se muestran a continuación:

Variable 9	Mezcla 1	Mezcla 2
Proporción	0.50	0.50
α	11.83	1.85
β	0.005	0.088

Cuadro 4.18: Parámetros estimados

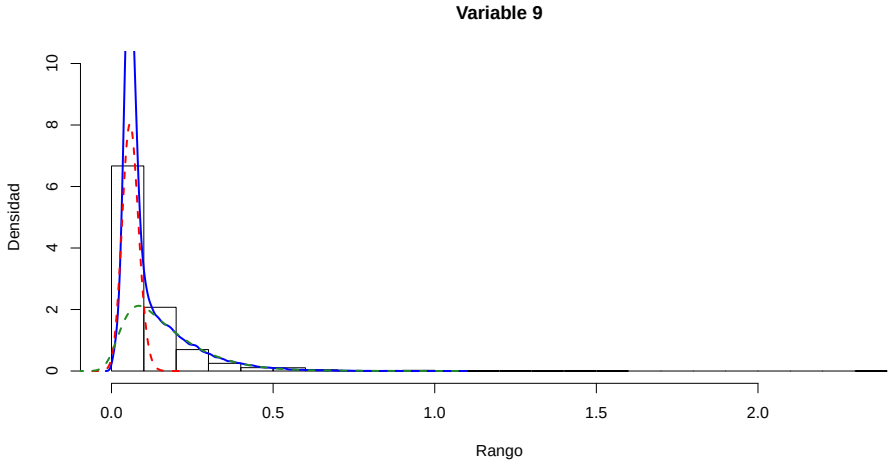


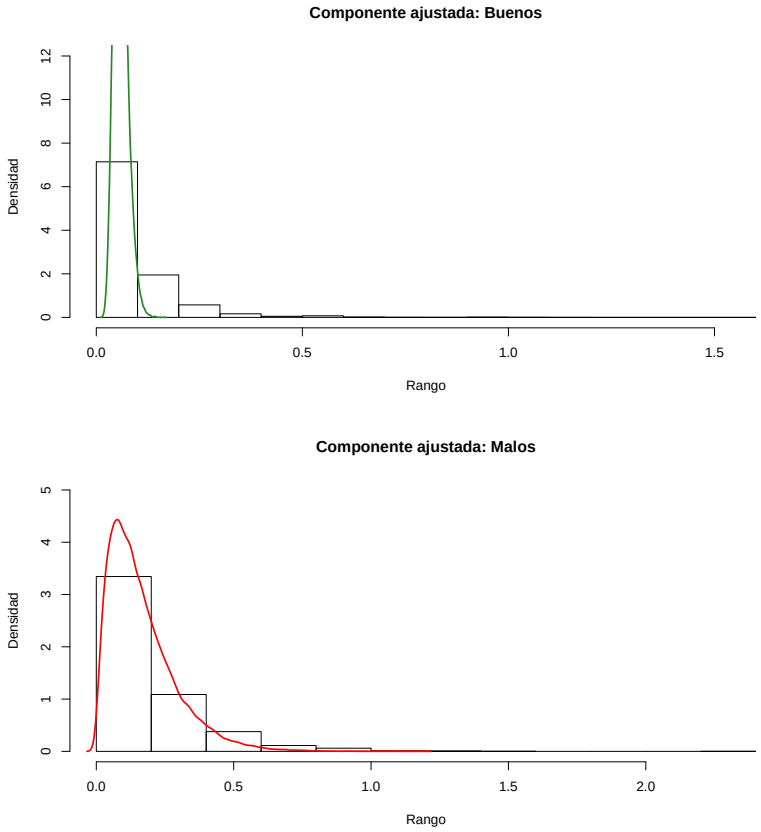
Figura 4.5: Componentes ajustadas

El modelo tiene una precisión del 65 %. Esta precisión es baja ya que la proporción real entre buenos y malos es 0.85 y 0.15 aproximadamente mientras que la proporción estimada es de 0.5 para ambas poblaciones.

Variable 9	Buenos Reales	Malos Reales
Buenos Predicción	65 %	34 %
Malos Predicción	35 %	66 %

Cuadro 4.19: Tabla de clasificación

Las componentes ajustadas se muestran a continuación:



Pago realizado respecto al saldo del crédito durante los últimos 3 meses (EXG_SALDO_3M)

El algoritmo convergió en 29 iteraciones para esta variable. Las componentes ajustadas y sus respectivos parámetros se muestran a continuación:

Variable 10	Componente 1	Componente 2
Proporción	0.45	0.55
α	16.08	2.50
β	0.003	0.054

Cuadro 4.20: Parámetros estimados

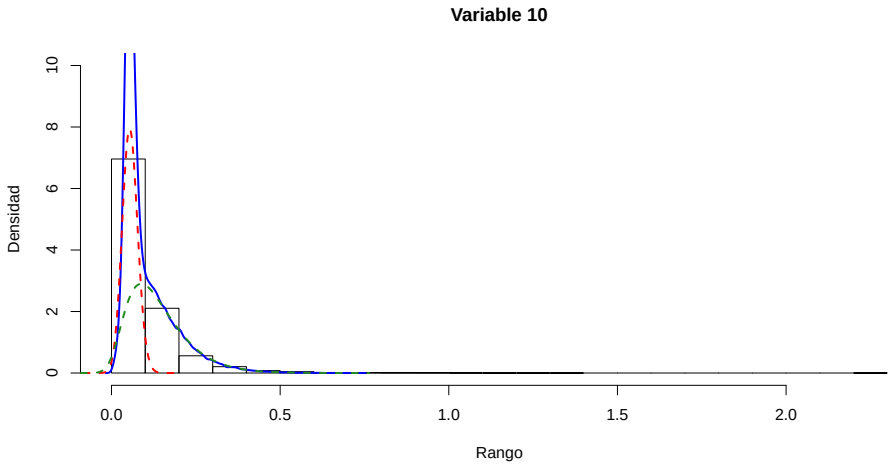


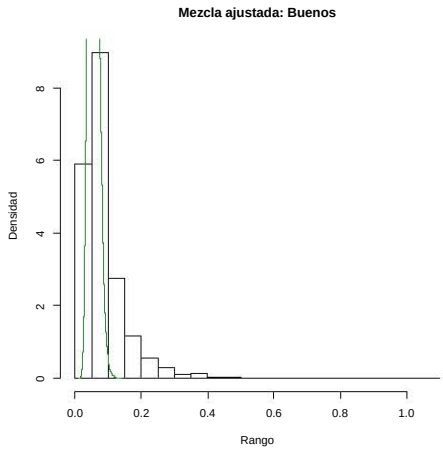
Figura 4.7: Componentes ajustadas

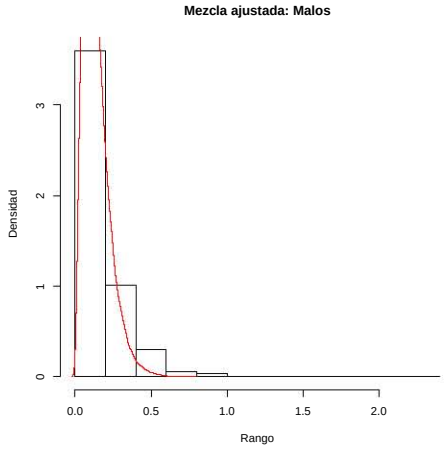
Similar a la variable anterior, el modelo tiene una precisión baja, que es del 60 %.

Variable 10	Buenos Reales	Malos Reales
Buenos Predicción	58 %	27 %
Malos Predicción	41 %	72 %

Cuadro 4.21: Tabla de clasificación

Gráficamente las componentes se ajustan de la siguiente forma:





Para las siguientes dos variables se consideró conveniente ajustar componentes con distribución Normal, ya que presentan una elevada concentración entre los valores 0 y 1 y una varianza pequeña. Además, se puede apreciar que la concentración de valores extremos en las colas es baja.

Exigible respecto al ingreso durante los últimos 3 meses (EXG_ING_3M)

El algoritmo convergió en 34 iteraciones para esta variable. Las componentes ajustadas y sus respectivos parámetros se muestran a continuación:

Variable 1	Mezcla 1	Mezcla 2
Proporción	0.12	0.88
μ	0.59	0.16
σ	0.38	0.09

Cuadro 4.22: Parámetros estimados

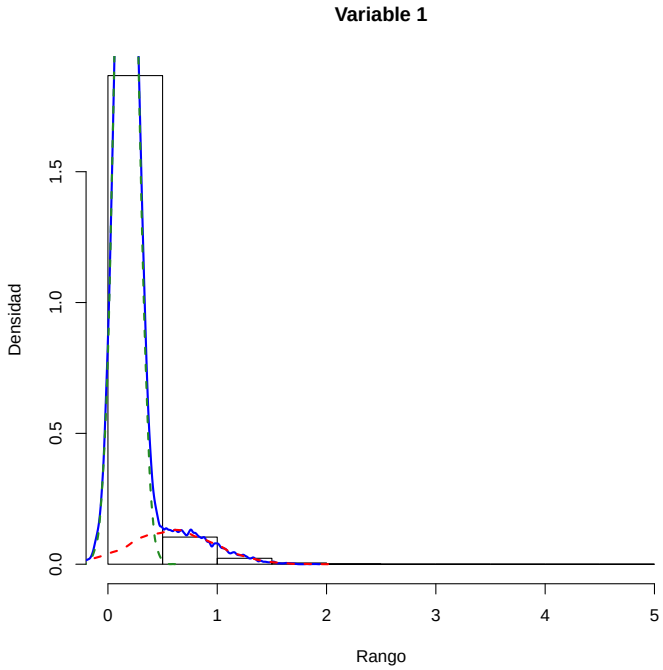


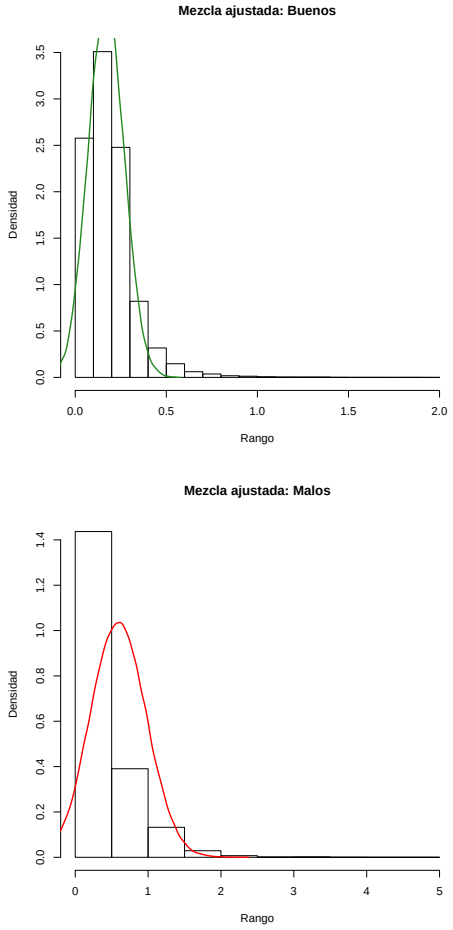
Figura 4.9: Componentes ajustadas

La tabla de contingencia muestra aparentemente una buena precisión, indicando altos porcentajes de clasificación correcta para la población de buenos, mas no así para la población de malos. La precisión del modelo es del 86 %. Con este resultado, podemos concluir que la variable tiene un poder de clasificación intermedio.

Variable 1	Buenos Reales	Malos Reales
Buenos Predicción	95 %	66 %
Malos Predicción	5 %	34 %

Cuadro 4.23: Tabla de clasificación

A continuación se puede apreciar gráficamente la precisión del ajuste:



Exigible respecto al ingreso durante los últimos 6 meses (EXG_ING_6M)

El algoritmo convergió en 32 iteraciones para esta variable. Las componentes ajustadas y sus respectivos parámetros se muestran a continuación:

Variable 2	Mezcla 1	Mezcla 2
Proporción	0.13	0.87
α	0.53	0.17
β	0.32	0.10

Cuadro 4.24: Parámetros estimados

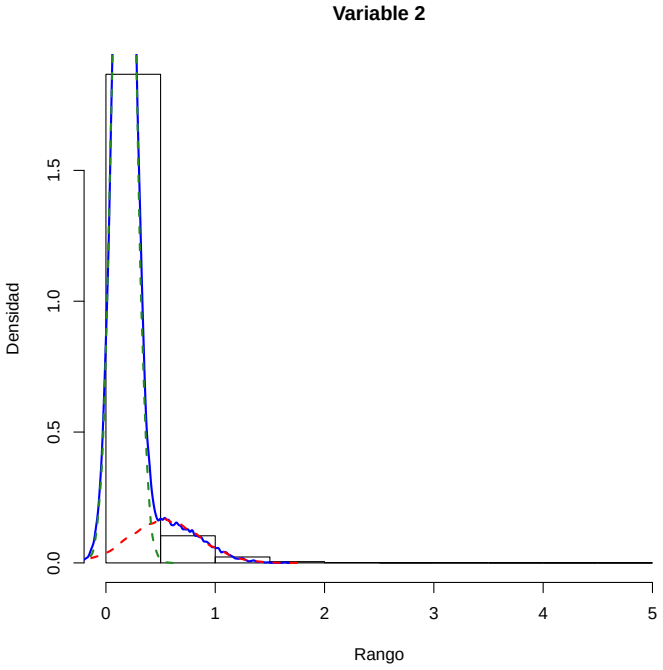


Figura 4.11: Componentes ajustadas

Como en el caso de la variable 1, la tabla de contingencia muestra una alta precisión para clasificar al grupo de buenos, siendo la precisión total del modelo del 85 %.

Puede observarse que tanto las medias como las proporciones de las componentes son significativamente distintas y siguen la lógica que se comentó en

la sección de análisis. Se esperaba que aquellos créditos que tuvieran valores superiores al 20 % serían más riesgosos.

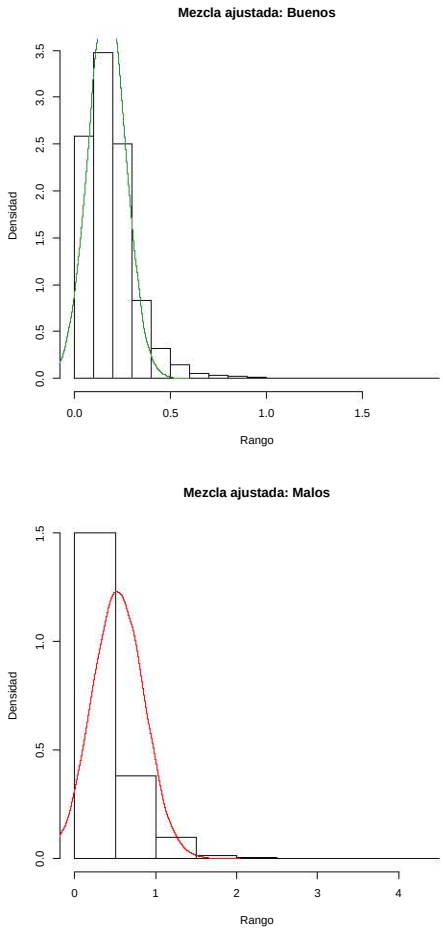
Ambas variables presentan parámetros muy similares, por lo que elegir alguna de ellas como la que mejor clasifica, dependerá de la longitud del periodo histórico que sea más conveniente.

En general, este tipo de créditos son de corta duración, por lo que resultaría más apropiado considerar una historia de 3 meses y no una de 6, evitando así posibles pérdidas de información.

Variable 3	Buenos Reales	Malos Reales
Buenos Predicción	94 %	68 %
Malos Predicción	6 %	32 %

Cuadro 4.25: Tabla de clasificación

Los ajustes a las poblaciones de buenos y malos se muestran en los siguientes gráficos, donde puede apreciarse de una manera más clara el buen resultado del ajuste a la población de buenos y el ajuste regular para la población de malos:



Las variables restantes no se incluyen en los resultados ya que no se pudo ajustar ningun modelo. Cabe destacar que las varibles con mejores resultados presentan una proporción cercana a 14.48 % en alguna de las componentes. Este porcentaje representa la tasa real de incumplimiento de la cartera, o dicho de otra manera, la proporción de malos respecto al total de créditos.

Conclusiones

El algoritmo EM resultó ser una herramienta práctica y eficiente para la estimación de parámetros cuando se tienen datos faltantes. En el caso donde los datos faltantes fueron explícitos se encontró que el algoritmo EM fue el mejor método de estimación, con los menores errores y una gran facilidad de implementación. La desventaja del algoritmo es que no es un método de imputación directa, es decir, no se pueden obtener valores individuales para cada registro, sin embargo se puede implementar a la par de métodos estocásticos para generar resultados más precisos. Algunas paqueterías utilizan combinaciones del algoritmo EM con la imputación múltiple para generar conjuntos de datos y realizar estimaciones, en donde los parámetros iniciales para el método de imputación múltiple se generan a través del algoritmo EM.

En muchas situaciones resulta más práctico recurrir a métodos de eliminación o de imputación de la media, por los volúmenes de información que se manejan y que no implican esfuerzos computacionales, sin embargo se están adecuando métodos que permitan estimar los datos faltantes en poblaciones de gran volumen a través de métodos estadísticos más confiables, sobre todo, para garantizar la validez y evitar sesgos en las estimaciones generadas a través de los datos. En esta aplicación los resultados fueron óptimos ya que los datos provenían de una distribución normal multivariada y los datos faltantes fueron generados de manera aleatoria, sin embargo, es difícil encontrar este tipo de distribución en cualquier conjunto de datos, además de que pueden presentarse patrones de datos faltantes no aleatorios que dificultarían las estimaciones. En general, se asume que el patrón de datos faltantes es aleatorio, y que la población proviene de una distribución normal multivariada para poder implementar el algoritmo.

En el caso donde los datos faltantes fueron latentes, se pudo observar que en algunos casos se pudieron ajustar acertadamente algunas distribuciones,

pero en otros esto no sucedió así. Esto tiene diversas explicaciones, como por ejemplo, la calidad de la información, la proporción real de las mezclas. En este caso, el insumo que presentaba una calidad dudosa de reporte fue la del Ingreso del Acreditado, por lo que a las variables que involucraban este insumo no se les pudo ajustar un modelo (no hubo convergencia). Además, se tiene que hacer la minería de datos previa para identificar valores atípicos, puesto que estos podrían ocasionar sesgos en las estimaciones del algoritmo o incluso podrían llegar a ser un factor que dificulte la convergencia del mismo. Es por esto que fue necesario aplicar diversos filtros a las variables.

Para las variables a las que sí se les logró ajustar un modelo tenemos dos casos, las que tuvieron una alta precisión y las que no la tuvieron. Para las de alta precisión, se observa que los criterios intuitivos son congruentes con los parámetros estimados por el modelo, tanto la proporción de buenos y malos como sus respectivos parámetros fueron estimados acorde con lo que se esperaba. Sin embargo, para las variables donde la proporción fue ajustada alrededor del 50 % para cada mezcla no se obtuvo un resultado adecuado o válido, puesto que se aleja bastante de la proporción real de buenos y malos. Se puede concluir que aquellas variables donde las mezclas son claramente diferenciables presentan un poder discriminatorio mayor que aquellas en donde no se observa este fenómeno. Estos resultados pueden ser de gran interés en el ámbito de la segmentación de variables, donde se busca distribuir a los acreditados en 2 o más categorías de acuerdo a sus características crediticias. En este ejercicio se decidió ajustar únicamente 2 mezclas, pero se puede ajustar un mayor número de distribuciones para cada variable y generar, por ejemplo, una tarjeta de puntuación donde se tendrán justificados con precisión cada uno de los segmentos o grupos de calificación generados.

La ventaja más importante del algoritmo EM es que siempre se obtendrán los mismos parámetros, iniciando en puntos iniciales diferentes y aunque se mostró un ejemplo de la literatura en donde esto no ocurre por la existencia de un punto silla, esta situación no suele ser común. En estos casos, siempre es conveniente probar con distintos puntos iniciales, o incluir la inferencia a priori que se tenga de ellos para garantizar una convergencia más rápida.
El uso de este algoritmo tanto para tratar datos faltantes explícitos como latentes es ampliamente recomendado y utilizado en diversos ámbitos, tanto por su flexibilidad como por su simpleza computacional, además de ser una garantía para encontrar siempre las mejores estimaciones para los parámetros en cuestión. Cabe destacar que con el algoritmo EM se puede abarcar

una gran rama de problemas para resolver, en donde incluso se han hecho extensiones más complejas del mismo (como la estocástica y la bayesiana) y que han generado importantes resultados a pesar de ser métodos relativamente nuevos.

Bibliografía

[1] A. P. Dempster; N. M. Laird; D. B. Rubin, *Maximum Likelihood from Incomplete Data via the EM Algorithm*. Journal of the Royal Statistical Society. Series B (Methodological), Vol.39, No.1 , (1977), pp. 1-38.

[2] Anirban DasGupta, *The Exponential Family and Statistical Applications*. Series: » Springer Texts in Statistics 2011, Probability for Statistics and Machine Learning.

[3] C.F. Jeff Wu, *On the Convergence Properties of the EM Algorithm*. The Annals of Statistics, Vol. 11, No. 1 (1983), pp. 95-103.

[4] Rubin D.B., *Inference and Missing Data*. Journal of the Royal Statistical Society. Biometrika, Vol. 63, No. 3, (1976).

[5] Little R.J.A., Rubin D.B., *Statistical Analysis with Missing Data*. Wiley Series in Probability and Statistics. Primera Edición, (1987).

[6] Geoffrey J. Mc Lachlan; T. Krishnan, *The EM Algorithm and Extensions*. Wiley Series in Probability and Statistics. Segunda Edición, (2008).

[7] Guobing Lu, John B. Copas, *Missing at random, likelihood ignorability and model completeness*. The Annals of Statistics, Vol. 32, No. 2 (2004), pp. 754–765.

[8] Alan C. Acock, *Working with missing values*. Journal of Marriage and Family 67 (November 2005): 1012–1028.

Anexos

Modelos de Mezclas Finitas con el Algoritmo EM

Cuando se desea clasificar una población en diferentes categorías de acuerdo a determinadas características descriptivas, se puede recurrir a diferentes métodos de segmentación.

Existen diversos métodos para estimar los parámetros de las mezclas en cuestión, algunos de ellos son el método de k-medias, Gibbs Sampler y el Algoritmo EM.

Para la estimación de estos parámetros el Algoritmo EM trabajará en dos pasos iterativos:

1. Clasificación: En este paso se asume que se tienen los estimadores de los parámetros que describen la mezcla de k-densidades en donde k es el número de componentes en cuestión. Este modelo proporciona la verosimilitud de cada sujeto de la población dentro de la mezcla, lo que permite su clasificación en una de las k categorías.

2. Modelo de Mezclas: Utilizando la clasificación del paso anterior, se modela la distribucion que describe a cada categoría. Para efectuar esta modelación se calculan los estimadores por máxima verosimilitud de los parámetros que describen cada categoría y se obtienen los respectivos pesos de cada componente de la mezcla.

En términos formales, supongamos que se tienen n individuos que pertenecen a una mezcla M de k densidades f. El objetivo es encontrar el conjunto de parámetros que describen a cada densidad para poder realizar la asignación de los individuos en k grupos.

$$M(x) = \sum_{i=1}^{k} \pi_i f_i(x), donde \sum_{i=1}^{k} \pi_i = 1 \qquad (4.2.2.1)$$

Para modelar este problema se introduce la variable $C = (c_1, \ldots, c_n)$, que indica de qué componente de la mezcla proviene cada individuo, de tal manera que cada uno de ellos tiene k posibles densidades a las cuales pertenecer.

$$C = \left\{ \begin{array}{ll} 1 & \text{si } x \in c_i \\ 0 & \text{en otro caso} \end{array} \right\} \qquad (4.2.2.2)$$

Dado que en este caso los datos faltantes son las k categorías, el paso-E consiste en calcular la probabilidad posterior que determina la pertenencia a cada categoría:

$$P[c_i \mid x] = \frac{\pi_i f_i(x)}{\sum_{i=1}^{k} \pi_i f_i(x)} \qquad (4.2.2.3)$$

En el paso-M se buscará encontrar los parámetros del modelo de mezclas dadas las verosimilitudes para cada clase de acuerdo a lo obtenido en el paso-E. De este modo, los parámetros de las componentes son re-estimados según la clasificación obtenida en el paso-E. Dependiendo del modelo de densidades que se ajuste a la situación, se buscará obtener los parámetros θ y π_i por máxima verosimilitud resolviendo:

$$\frac{\partial}{\partial \theta} \log L(\theta) = \frac{\partial}{\partial \theta} \log \prod_{i=1}^{n} \left(\sum_{i=1}^{k} \pi_i f_i(x) \right) = 0 \qquad (4.2.2.4)$$

y

$$\frac{\partial}{\partial \pi} \log L(\theta) = \frac{\partial}{\partial \pi} \log \prod_{i=1}^{n} \left(\sum_{i=1}^{k} \pi_i f_i(x) \right) = 0 \qquad (4.2.2.5)$$

Dado que se tiene la restricción $\sum_{i=1}^{n} \pi_i = 1$, se requiere el método de Lagrange para resolver el sistema de ecuaciones generado por (4.2.2.3).

Para comenzar con las iteraciones del Algoritmo EM se requieren valores iniciales para las proporciones π_i y para el conjunto de parámetros θ. Si no se tiene conocimiento a priori sobre estos parámetros, se pueden seleccionar de manera aleatoria, sin embargo, la convergencia del algoritmo puede llegar a ser mucho más lenta. Cabe destacar que puede tenerse o no conocimiento del

número de componentes que tiene la mezcla. En este caso existen criterios de selección como el Criterio de Información Bayesiana (BIC) y el Criterio de Información de Akaike (AIC). Ambos criterios balancean el ajuste y parsimonía del modelo que se está ajustando, es decir, penalizan la complejidad de los modelos y castigan un sobre-ajuste.

En el caso de mezclas, ajustar un modelo demasiado grande de componentes, implicaría calcular un elevado número de parámetros, lo que originaría una pérdida en la precisión del modelo. Ambos criterios se encuentran en función de la verosimilitud, la cual decrece conforme la complejidad del modelo se incrementa. La forma general de los criterios de información está dada por:

$$-2\log L(\hat{\theta}) + C$$

En donde el primer término es el negativo de la verosimilitud, y el segundo término C, penaliza modelos complejos y se incrementa conforme el número de parámetros a estimar aumenta. Así, aplicar alguno de los dos criterios implica minimizar el factor C.

Códigos de los Ejemplos

```
############################
##  Ejemplo Multinomial  ##
############################

####### Algoritmo EM ##########

EM<-function(n,vinicial,error){
t=0
x1=0
x2=0
e=0
e2=0
for (i in 1:n)  {
t[1]<-vinicial
x2[i]=125*(0.25*t[i])/(0.5+0.25*t[i])
x1[i]=125*(0.5)/(0.5+0.25*t[i])
t[i+1]=(x2[i]+34)/(x2[i]+34+18+20)
e[i]=t[i+1]-t[i]
if(e[i]<error){break}
min<-as.numeric(which.min(e))
est<-c(x1[min],x2[min])
resultados<-list()
resultados$Iteraciones<-which.max(t)-1
resultados$ML<-max(t)
resultados$Pasos_ML<-t
resultados$Errores<-e
```

```
resultados$EstimadoresML<-est
}
resultados
}
EM(200,0.5,1e-17)



################################
## Ejemplo Normal Univariada ##
################################

######### Generación de datos ###################
set.seed(4213)
datos<-rnorm(100,5,2)
sum(datos)
sum(datos*datos)
mean(datos)
var(datos)
miss<-runif(100,0,1)
datos2<-cbind(datos,miss)
datos3<-ifelse(datos2[,2]<0.15,NA,datos2[,1])
sumyobs<-sum(datos3,na.rm=TRUE)
mean(datos3,na.rm=TRUE)
var(datos3,na.rm=TRUE)
y2<-datos3*datos3
sumyobs2<-sum(y2,na.rm=TRUE)

####### Algoritmo EM #######
EMNORM<-function(m,muinicial,sigmainicial,error){
mu=0
sigma=0
sumy=0
sumy2=0
ey=0
ey2=0

### PASO E ###
sumyobs<-sum(datos3,na.rm=TRUE)
sumyobs2<-sum(datos3^2,na.rm=TRUE)
for (k in 1:m) {
mu[1]<-muinicial
sigma[1]<-sigmainicial

### PASO M ###
sumy[k+1]<-sumyobs+13*mu[k]
sumy2[k+1]<-sumyobs2+13*(mu[k]^2+sigma[k])
mu[k+1]=(sumy[k+1])/100
sigma[k+1]=(sumy2[k+1])/100-(mu[k+1])^2

ey[k]=mu[k+1]-mu[k]
ey2[k]=sigma[k+1]-sigma[k]

if(ey[k]<error & ey2[k]<error){break}

min<-as.numeric(which.min(abs(ey)))
```

```
min2<-as.numeric(which.min(abs(ey2)))

est<-setNames(c(sumy[min],mu[min],sumy2[min2],sigma[min]),c("Sum y","Mu","Sum y2","Var"))
resultados<-list()
resultados$ML_mu<-max(mu)
resultados$Pasos_ML_mu<-mu
resultados$Pasos_ML_sigma<-sigma
resultados$Errores_y<-ey
resultados$Errores_y2<-ey2
resultados$Iteraciones_Mu<-which.min(abs(ey))
resultados$Iteraciones_Sigma<-which.min(abs(ey2))
resultados$EstimadorML<-est
resultados$sumy<-sumy
resultados$sumy2<-sumy2
}
resultados
}
EMNORM(50,4.986371,4,1e-10)


######################
## Normal Bivariada ##
## Caso 1           ##
######################

###################### Datos ###################################
a_<-c(8,11,16,18,6,4,20,25,9,13)
b_<-c(10,14,16,15,20,4,18,22,NA,NA)
e_<-cbind(a_,b_)
d<-e_[1:8,]

################ Algoritmo EM ###################################

EMBIVAR2<-function(m,mu1ini,mu2ini,sigma1ini,sigma2ini,rhoini,error){
mu1=0
mu2=0
sigma1=0
sigma2=0
rho=1
n=10
err11=0
err12=0
err21=0
err22=0
err3=0
E11=0
E12=0
E13=0
E21=0
E22=0
##########
# Paso-E #
##########
#Parte observada
O11<-sum(b_,na.rm=TRUE)
O12<-sum(b_^2,na.rm=TRUE)
```

```
O21<-sum(a_,na.rm=TRUE)
O22<-sum(a_^2,na.rm=TRUE)
O3<-sum(d[,1]%*%d[,2])

#Parte no observada
b_a<-e_[9:10,1]
a_b<-a_

for (k in 1:m) {
mu1[1]=mu1ini
mu2[1]=mu2ini
sigma1[1]=sigma1ini
sigma2[1]=sigma2ini
rho[1]=rhoini
#########################################################################################################
E11[k]<-O11+ 2*( mu1[k]+( rho[k]*sqrt(sigma1[k]) )/ sqrt(sigma2[k] )*(sum(b_a)/2-mu2[k])  )
E21[k]<-O21
#########################################################################################################
E12[k]<-O12 + 2*mu1[k]^2 + 2*mu1[k]*( rho[k]*sqrt(sigma1[k]) )/ sqrt(sigma2[k])*( sum(b_a)-2*mu2[k] )+
       ((rho[k]*sqrt(sigma1[k]))^2/sqrt(sigma2[k])^2)*(sum(b_a^2)-2*sum(b_a)*mu2[k]+2*mu2[k]^2) +
       2*sigma1[k]*(1-rho[k]^2)
E22[k]<-O22
#########################################################################################################
E13[k]<-sum(b_a)*(mu1[k])+ sum(b_a^2)*(rho[k]*sqrt(sigma1[k])/sqrt(sigma2[k]))-
        (rho[k]*sqrt(sigma1[k])/sqrt(sigma2[k]))*sum(b_a)*mu2[k]
#########################################################################################################


##########
# Paso-M #
##########
mu1[k+1]<-E11[k]/n
mu2[k+1]<-E21[k]/n
sigma1[k+1]<-E12[k]/n - mu1[k+1]^2
sigma2[k+1]<-E22[k]/n - mu2[k+1]^2
rho[k+1]<-((E13[k] + O3)/n - mu1[k+1]*mu2[k+1]) /sqrt(sigma1[k+1]*sigma2[k+1])
err11[k]=mu1[k+1]-mu1[k]
err12[k]=sigma1[k+1]-sigma1[k]
err21[k]=mu2[k+1]-mu2[k]
err22[k]=sigma2[k+1]-sigma2[k]
err3[k]=rho[k+1]-rho[k]
if(err11[k]<error & err12[k]<error & err21[k]<error & err22[k]<error & err3[k]<error){break}
}
estimados<-list()
estimados$mu1_est<-mu1
estimados$mu2_est<-mu2
estimados$sigma1_est<-sigma1
estimados$sigma2_est<-sigma2
estimados$rho_est<-rho
estimados$cov_est<-rho*sqrt(sigma1*sigma2)
estimados
}

EMBIVAR2(30,
mean(b_,na.rm=TRUE),
mean(a_),
```

```
var(b_,na.rm=TRUE),
var(a_),
cor(d[,1],d[,2]),
0)

######################
## Normal Bivariada ##
## Caso 2           ##
######################

###################### Datos ####################################
a_<-c(1,1,-1,-1,2,2,-2,-2,NA,NA,NA,NA)#=y2
b_<-c(1,-1,1,-1,NA,NA,NA,NA,2,2,-2,-2)#=y1
e_<-cbind(a_,b_)
d<-e_[1:4,]

################ Algoritmo EM ###################################

EMBIVAR<-function(m,mu1ini,mu2ini,sigma1ini,sigma2ini,rhoini,error){
mu1=0
mu2=0
sigma1=0
sigma2=0
rho=1
n=12
err11=0
err12=0
err21=0
err22=0
err3=0
E11=0
E12=0
E13=0
E21=0
E22=0
E23=0
##########
# Paso-E #
##########
#Parte observada
O11<-sum(b_,na.rm=TRUE)
O12<-sum(b_^2,na.rm=TRUE)
O21<-sum(a_,na.rm=TRUE)
O22<-sum(a_^2,na.rm=TRUE)
O3<-sum(d[,1]%*%d[,2])

#Parte no observada
b_a<-e_[5:8,1]
a_b<-e_[9:12,2]

for (k in 1:m) {
mu1[1]=mu1ini
mu2[1]=mu2ini
sigma1[1]=sigma1ini
sigma2[1]=sigma2ini
rho[1]=rhoini
```

```
#########################################################################################################
E11[k]<-O11+ 4*( mu1[k]+( rho[k]*sqrt(sigma1[k]) )/ sqrt(sigma2[k] )*(sum(b_a)/4-mu2[k])  )
E21[k]<-O21+ 4*( mu2[k]+( rho[k]*sqrt(sigma2[k]) )/ sqrt(sigma1[k] )*(sum(a_b)/4-mu1[k])  )
#########################################################################################################
E12[k]<-O12+ 4*mu1[k]^2 + 2*mu1[k]*( rho[k]*sqrt(sigma1[k]) )/ sqrt(sigma2[k])*( sum(b_a)-4*mu2[k] )+
       ((rho[k]*sqrt(sigma1[k]))^2/sqrt(sigma2[k])^2)*(sum(b_a^2)-2*sum(b_a)*mu2[k]+4*mu2[k]^2) +
       4*sigma1[k]*(1-rho[k]^2)
E22[k]<-O22+ 4*mu2[k]^2 + 2*mu2[k]*( rho[k]*sqrt(sigma2[k]) )/ sqrt(sigma1[k])*( sum(a_b)-4*mu1[k] )+
       ((rho[k]*sqrt(sigma2[k]))^2/sqrt(sigma1[k])^2)*(sum(a_b^2)-2*sum(a_b)*mu1[k]+4*mu1[k]^2) +
       4*sigma2[k]*(1-rho[k]^2)
#########################################################################################################
E13[k]<-sum(b_a)*(mu1[k])+ sum(b_a^2)*(rho[k]*sqrt(sigma1[k])/sqrt(sigma2[k]))
        -(rho[k]*sqrt(sigma1[k])/sqrt(sigma2[k]))*sum(b_a)*mu2[k]
E23[k]<-sum(a_b)*(mu2[k])+ sum(a_b^2)*(rho[k]*sqrt(sigma2[k])/sqrt(sigma1[k]))
        -(rho[k]*sqrt(sigma2[k])/sqrt(sigma1[k]))*sum(a_b)*mu1[k]
#########################################################################################################


##########
# Paso-M #
##########
mu1[k+1]<-E11[k]/n
mu2[k+1]<-E21[k]/n
sigma1[k+1]<-E12[k]/n - mu1[k+1]^2
sigma2[k+1]<-E22[k]/n - mu2[k+1]^2
rho[k+1]<-((E13[k] + E23[k] + O3)/n - mu1[k+1]*mu2[k+1]) /sqrt(sigma1[k+1]*sigma2[k+1])
err11[k]=mu1[k+1]-mu1[k]
err12[k]=sigma1[k+1]-sigma1[k]
err21[k]=mu2[k+1]-mu2[k]
err22[k]=sigma2[k+1]-sigma2[k]
err3[k]=rho[k+1]-rho[k]
if(err11[k]<error & err12[k]<error & err21[k]<error & err22[k]<error & err3[k]<error){break}
}
estimados<-list()  #(sigma2,sigma1,rho)
estimados$mu1_est<-mu1
estimados$mu2_est<-mu2
estimados$sigma1_est<-sigma1
estimados$sigma2_est<-sigma2
estimados$rho_est<-rho
estimados
}

EMBIVAR(50,0,0,0.5,0.5,0.5,1e-10000)
EMBIVAR(50,0,0,1,2,-0.1,1e-10000)
EMBIVAR(50,0,0,2,1,0,1e-10000)
```