MODELOS DE EFECTOS MIXTOS EN ANÁLISIS DE CONGLOMERADOS

Tesis para optar al grado de Doctor en Estadística

Luis A. Villarroel del Pino

11 de Junio de 2007

Contents

Abstract

En este trabajo se aborda una situación que se presenta frecuentemente en el manejo de información en varias disciplinas y particularmente en el área de la Medicina y las Ciencias Biológicas: contar con información para una o más variables numéricas medidas en diferentes tiempos para un conjunto de individuos, las que a su vez pueden depender de un conjunto de variables poblacionales o individuales medidas en los mismos tiempos, observándose una variable de resultado (outcome) al final del estudio (como la muerte o sobrevida de los individuos en estudio). Se asume que el outcome final está relacionado con la evolución y la relación entre las variables medidas longitudinalmente

El caso general de esta situación ocurre cuando la información longitudinal es medida en tiempos distintos para cada sujeto (no en tiempos predefinidos, los que suelen ser equiespaciados) y no se cuenta con información completa para todos los individuos en estudio (es decir, no se tiene información para todas las variables medidas durante el seguimiento).

El objetivo de esta tesis es construir modelos para predecir la clasificación final de los individuos en seguimiento en estudios longitudinales. Debido a la naturaleza de los datos en estudio, se usarán modelos de efectos mixtos lineales y no lineales para describir el comportamiento de los datos medidos longitudinalmente. En ambas situaciones se considerarán modelos univariados y multivariados para modelar el comportamiento de una o más variables en el tiempo en función de covariables poblacionales e individuales. Se asume que la variable respuesta de interés, medida al final del período de seguimiento, es desconocida y corresponde a la clasificación del estado final de los individuos en dos o más grupos. Se usa un método de análisis discriminante semi-bayesiano para clasificar los individuos en conglomerados y se determina la capacidad predictiva de los conglomerados resultantes en la clasificación final de los sujetos, comparando la clasificación entregada por el mecanismo de cluster con el verdadero estado del sujeto usado como regla de oro (o gold standard).

Aunque en la literatura existen varias propuestas de métodos de conglomerados basados en modelos, e incluso algunas basadas en modelos lineales de efectos mixtos, las metodologías propuestas de esta tesis permiten la estimación iterativa de los parámetros fijos y aleatorios independientemente para cada conglomerado, el uso de modelos no lineales de efectos mixtos en la clasificación, el uso de múltiples variables respuesta (clasificación multivariada) y la estimación de parámetros para el caso desbalanceado.

1 Introducción

El uso de información longitudinal donde un conjunto de individuos es seguido por un período de tiempo, siendo de interés el estado final de estos individuos, ha sido fuente de desarrollo metodológico por muchos autores en el área de la estadística, y estas metodologías han sido ampliamente utilizadas en el área de la Medicina y las Ciencias Biológicas. Un ejemplo de estudio longitudinal donde estas técnicas son aplicables es el estudio de Anderson et al [2], basado en el estudio de Framingham, en el cual se mide el nivel de colesterol sérico en 4374 sujetos durante un período de 5 años y se determina el efecto del cambio en el nivel de colesterol sobre la mortalidad por causas cardiovasculares a 30 años de seguimiento. Otro ejemplo puede verse en Ellard et al [14], donde se mide el consumo diario de nicotina durante el embarazo en 338 mujeres, medido en un test de orina, y se analiza la relación entre el nivel de nicotina consumido y el déficit de peso de nacimiento del recién nacido. Con frecuencia, estos diseños resultan en conjuntos de datos muy desbalanceados debido a que los individuos podrían ser medidos en tiempos distintos, además de tener un número variable de mediciones para cada individuo.

Aunque el modelo de efectos mixtos permite formular y modelar los problemas descritos, no hay aplicaciones en las que se analice la relación de un conjunto de variables en el tiempo y se use esta relación para clasificar a los sujetos en grupos, de acuerdo a variables que podran ser desconocidas al momento de medirse los datos longitudinales.

El objetivo de esta tesis es construir modelos para predecir la clasificación final de los individuos en seguimiento en estudios longitudinales, usando modelos de efectos mixtos lineales y no lineales para describir el comportamiento de los datos medidos longitudinalmente. En ambas situaciones se considerarán modelos univariados y multivariados para modelar el comportamiento de una o más variables en el tiempo en función de covariables poblacionales e individuales. En este proceso, se asumirá que existe una variable respuesta de interés, medida al final del período de seguimiento, que es desconocida y que corresponde a la clasificación del estado final de los individuos en dos o más grupos. Se usa un método de análisis de conglomerados semi-bayesiano para clasificar los individuos en grupos afines de acuerdo al modelo mixto ajustado y se determina la capacidad predictiva de los conglomerados resultantes en la clasificación final de los sujetos, comparando la clasificación entregada por el mecanismo de cluster con el verdadero estado del sujeto usado como gold standard.

En la subsección siguiente se muestra el desarrollo y estado actual del modelo de efectos mixtos y la estimación de componentes de la varianza. Posteriormente se describe el algoritmo de clustering utilizado en esta tesis y otros algoritmos que podrían ser explorados en esta misma línea. Finalmente, se describen trabajos en los que se ha abordado el uso de modelos de efectos mixtos en conjunto con análisis de conglomerados para resolver problemas similares al abordado en esta tesis.

1.1 Desarrollo del modelo de efectos mixtos y la estimación de componentes de la varianza

Respecto al desarrollo de las metodologías de análisis de efectos mixtos, Henderson en 1953 [24] aporta las primeras ideas para la estimación de componentes de la varianza (la varianza de los efectos aleatorios en un modelo de efectos mixtos) para datos no balanceados, en situaciones más complejas que el análisis de la varianza con solo un factor fijo. Sin embargo, como lo muestra Searle en 1991 [48], la metodología propuesta por Henderson tenía la desventaja de no proveer restricciones para prevenir la obtencion de estimadores de componentes de la varianza negativos. Antes de Henderson, la estimación de componentes de la varianza se centraba principalmente en el problema de datos balanceados y en modelos de análisis de la varianza con un factor fijo, como en el trabajo de Daniels de 1939 [12].

La estimación de componentes de la varianza en situaciones en que se conocen las variables de seguimiento y la variable de resultado (outcome) para cada individuo, con observaciones medidas en los mismos tiempos y sin valores ausentes, ha sido considerada por varios autores después de Henderson. En particular, el modelo de análisis de la varianza multivariado de Potthoff y Roy de 1964 [40] era usado habitualmente, siendo de la forma

$$E(y_{n\times t}) = X_{n\times q}\beta_{q\times p}P_{p\times t} \tag{1}$$

donde Y es una matriz con filas independientes representando las t observaciones medidas en el tiempo para cada uno de los n sujetos, X es matriz de diseño conocida de efectos entre individuos, P es matriz de diseño conocida de efectos intra-individuos y β es una matriz de parámetros desconocidos, con $p < t$. En este esquema, el número t de tiempos en que se realizan las mediciones es un número fijo.

Al igual que Potthoff y Roy, otros autores centraban su atención en el desarrollo de modelos para curvas de crecimiento. Ver por ejemplo el artículo de Rao de 1959 [43], donde se considera el problema del análisis de curvas de crecimiento polinomiales de mediciones seriales. Cuando todos los individuos son observados en las mismas p ocasiones, el modelo de Rao es de la forma

$$E(y_i) = A\beta \tag{2}$$

donde las columnas de A son potencias de t (tiempo) o polinomios ortogonales definidos por los tiempos de observación. Rao da una solución máximo verosímil para el modelo de curvas de crecimiento polinomiales con estructura de error multivariada. Por otra parte, Grizzle y Allen en 1969 [19], introducen covariables al diseño balanceado de Rao, definiendo el valor esperado como

$$E(y_i) = A\beta x_i \tag{3}$$

donde x_i es el vector de valores de una covariable para el i-ésimo individuo. En este caso también se dan soluciones máximo verosímiles para los parámetros del modelo. Una característica común de estos enfoques es que se basan en una premisa muy poco

práctica: las observaciones longitudinales son medidas en un número fijo de tiempos (diseño balanceado) y completo (no hay observaciones faltantes).

El análisis de datos longitudinales con datos incompletos es abordado por Kleinbaum en 1973 [29], también con aplicación en la estimacion de parámetros en curvas de crecimiento. En el modelo de Kleinbaum, un conjunto de n individuos en estudio son divididos en s particiones disjuntas $S_1, \ldots, S_s$, de modo que la partición S_j contiene los n_j sujetos con observaciones en los tiempos específicos t_j. Por ejemplo, si el estudio contempla observar a los individuos en $t = 5$ tiempos, podrían existir $S = 2$ particiones: una compuesta por n_1 sujetos medidos en los 5 tiempos y otra compuesta por n_2 sujetos medidos en los tiempos $1, 3$ y 5 ($n_1 + n_2 = n$). Luego, el modelo (1) de Potthoff y Roy se puede reescribir de la siguiente forma para considerar vectores de respuesta incompletos

$$E(y_{n_j \times t_j}) = X_{n_j \times q} \beta_{q \times t} P_{t \times t_j}, \qquad j = 1, \ldots, s \tag{4}$$

donde $y_{n_j \times t_j}$ representa los n_j individuos medidos en el tiempo t_j, X es matriz de diseño conocida de los n_j sujetos, P es matriz de diseño conocida de efectos intra-individuos medida para un total de t tiempos distintos y β es una matriz de parámetros desconocidos, con $p < t$. Nótese que el modelo (4) sigue siendo un esquema en que los tiempos son fijos, sólo que esta vez es al interior de cada partición. La estimación de parámetros se hace mediante mínimos cuadrados, asumiendo que se conoce la matriz de varianzas covarianzas de Y al interior de cada partición o para el conjunto completo de datos. En general, el modelo de Kleinbaum falla cuando el número de tiempos de observación es grande relativo al número de individuos. Una variante del modelo de Kleinbaum, con aplicación a estudios cross-sectional, puede verse en Woolson et al de 1978 [60].

Rao en 1965 [44] y 1975 [45] describe una familia de modelos de efectos aleatorios en 2 etapas para estimación en una situación con datos balanceados. En la primera etapa, Rao plantea un modelo lineal condicional en los parámetros de la curva de crecimiento individual β_i, en la forma

$$Etapa\ 1: \qquad Y_i|\beta_i = Z_i\beta_i + \epsilon_i \tag{5}$$

donde $\epsilon_i \sim MVN(0, R_i)$, donde R_i es matriz de covarianza. En la etapa 2, se plantea que el parámetro aleatorio β_i tiene distribución

$$Etapa\ 2: \qquad \beta_i \sim MVN(W_i\alpha, \Lambda) \tag{6}$$

Nótese que el modelo plantea curvas de crecimiento para cada individuo, y que los parámetros de las curvas de crecimiento individuales 'oscilan' en torno a una curva de crecimiento poblacional de media $W_i\alpha$. Rao da una solución máximo verosimil para el modelo de efectos aleatorios planteado en (5) y (6). Una vez más, la solución está limitada a observaciones longitudinales medidas en un número fijo de tiempos y completo.

Beale y Little en 1975 [6] presentan un método de estimación en análisis multivariado con datos incompletos. Este trabajo muestra el uso de un algoritmo tipo estimación-maximización para la obtención de estimadores máximo verosímiles. Dempster, Laird y Rubin en 1977 [13] presentan el algoritmo EM para la computación iterativa de estimadores MV para datos incompletos.

Harville, en 1977 [22], enfoca el problema de estimar componentes de la varianza como un caso especial de estimación del modelo lineal general, que combina efectos fijos y aleatorios en la forma

$$y = X\alpha + Zb + e \tag{7}$$

donde y es un vector de dimensión $n \times 1$ de valores observados, X y Z son matrices de 'regresores', α es un vector de parámetros de efectos fijos y b es un vector de efectos aleatorios. En su trabajo, Harville explora la estimación de componentes de la varianza del modelo mediante máxima verosimilitud (ML) y máxima verosimilitud restringida (REML), además de algunos procedimientos numéricos para la estimación máximo verosimil y la relación entre ML y REML con otros métodos de estimación.

Siguiendo la notación de Harville (que será utilizada en general en el transcurso de este trabajo), los supuestos del modelo (7) son que $E(b) = 0$, $E(e) = 0$ y $cov(b, e) = 0$. Si llamamos $D = var(b)$, $R = var(e)$ y $V = R + ZDZ'$, entonces $var(y) = V$. Se asume que X es una matriz conocida, pero los elementos de D y R pueden estar en función de un vector de parámetros desconocidos $\theta = (\theta_1, \ldots, \theta_m)'$. El espacio de parámetros de α y θ es $\{(\alpha, \theta) : \theta \in \Omega\}$, donde Ω es subconjunto del espacio euclidiano m-dimensional tal que R y V son matrices no singulares. Las matrices R y D pueden tener distintas parametrizaciones, las que permiten dar cuenta de heterocedasticidad en los errores o acomodar situaciones especiales como el ajuste de series de tiempo, al asumir que los errores están correlacionados.

Harville hace notar además que el modelo (7) permite también ajustar modelos lineales multivariados, ya que no hay nada en la formulación del modelo que excluya la situación donde diferentes tipos de medidas sean incluidas en el vector de respuestas y.

Por otra parte, Beal y Sheiner en 1979 [4] [5] y Sheiner y Beal en 1980 [50] comienzan el desarrollo de métodos máximo verosímiles para la estimación de parámetros poblacionales en modelos de efectos mixtos no lineales. Los autores presentan además un programa computacional escrito en lenguaje FORTRAN para el análisis de datos longitudinales usando modelos mixtos no lineales, con aplicación en el área de la farmacología. Una restricción importante del programa era que, aunque los efectos fijos podían ser considerados lineales o no lineales, no podían incluirse en el modelo efectos aleatorios no lineales. Por otra parte, un enfoque bayesiano del modelo de efectos mixtos no lineales puede verse en Racine-Poon 1985 [41]. Un resumen de la bibliografía disponible hasta el año 1994 sobre modelos no lineales de efectos mixtos puede verse en el artículo de Yuh et al [61].

1.2 El modelo de efectos mixtos de Laird y Ware y desarrollos posteriores

Laird y Ware en 1982 [31] obtienen una generalización del modelo de Harville cuando las observaciones son medidas en tiempos arbitrarios. En este caso, se plantea un modelo en dos etapas dado por

$$Etapa\ 1: \qquad y_i|b_i = X_i\alpha + Z_ib_i + \epsilon_i \tag{8}$$

donde $\epsilon_i \sim MVN(0, R_i)$, y la etapa 2 es

$$Etapa\ 2: \qquad b_i \sim MVN(0, D) \tag{9}$$

donde y_i es el vector de respuestas de dimensión $n_i \times 1$ que contiene n_i observaciones medidas longitudinalmente para el i-ésimo sujeto, $i = 1 \dots\ n$. El vector α es de dimensión $p \times 1$ de coeficientes de regresión poblacionales y b_i es un vector $q \times 1$ de coeficientes de regresión individuales, específicos para cada sujeto, los cuales describen la desviación del i-ésimo individuo en relación a la evolución promedio de la población. Las matrices X_i y Z_i son matrices de diseño conocidas de dimensión $n_i \times p$ y $n_i \times q$, repectivamente. Los errores ϵ_i se asumen independientes, con distribución N(0,R_i) donde R_i es matriz de covarianzas que depende de i solo a través de su dimensión n_i. Una simplificación habitual es asumir que $R_i = \sigma^2 I_{n_i \times n_i}$. Los vectores b_i tienen distribución N(0,D) y son independientes entre sí y de los ϵ_i. D es una matriz de covarianza de dimensión $k \times k$. Con esta formulación, la distribución marginal de y_i es normal, con media $X_i\alpha$ y matriz de covarianza $V_i = Z_iDZ_i^T + R_i$.

La estimación de parámetros en Laird y Ware se hace usando máximo verosimilitud y a través de una formulación bayesiana empírica. En ambos casos, se utiliza el algoritmo EM para encontrar los EMV.

Reisnel en 1984 [46] propone un modelo lineal de efectos mixtos para el caso multivariado, en el cual y_{ijk} representa la medición de la i-ésima característica en la ocasión j para el individuo k, donde $i = 1, \dots, m$, $j = 1, \dots, p$, $k = 1, \dots, N$. Aunque mucha de la notación matricial del trabajo de Reisnel es posible de incorporar en esta tesis para el planteamiento del modelo de efectos mixtos multivariado, su limitante es que Reisnel considera el caso en que las observaciones son medidas en una cantidad fija p de ocasiones, por lo que no puede considerarse una extensión multivariada del modelo de Laird y Ware de 1982. Además, Reisnel tampoco considera en su estudio el caso de observaciones faltantes, el cual sí será considerado en esta tesis.

Un enfoque multivariado en la línea de Laird y Ware para el caso de datos longitudinales es planteado por Ware en 1985 [58], el cual es aplicable cuando el diseño es incompleto y no balanceado. Ware plantea un modelo genérico para el vector de respuestas individuales Y_i, $i = 1, \dots, n$, el cual es de dimensión p_i, denotando una cantidad variable de ocasiones en que se observa cada individuo. Se asume que Y_i sigue el modelo lineal

$$Y_i = X_i\beta + e_i \tag{10}$$

donde X_i es la matriz de diseño para el i-ésimo individuo y e_i es un vector de desviaciones con distribución normal multivariada y matriz de covarianza Σ_i. Ware plantea que si $\Sigma_i = cov(e_i)$ es conocida, el vector de parámetros β se puede estimar mediante el estimador de mínimos cuadrados generalizados de Aitken

$$\tilde{\beta} = (\sum_{i=1}^{n} X_i' \Sigma_i^{-1} X_i)^{-1} (\sum_{i=1}^{n} X_i' \Sigma_i^{-1} Y_i) \tag{11}$$

Cuando Σ_i es desconocida, entonces β será función de un estimador $\tilde{\Sigma}_i$ de Σ_i. Ware plantea formas de estimación de Σ_i para el modelo de efectos aleatorios, el modelo para datos longitudinales multivariado y para modelos AR.

Una variante al modelo de Lair y Ware es la propuesta por Jennrich y Schluchter en 1986 [28], aunque ésta también puede verse como una extensión del trabajo de Ware de 1985 [58]. Con esta formulación, es posible obtener estimadores máximo verosímiles no restringidos y restringidos (ML y REML) mediante un modelo lineal general con estructura de la matriz de covarianza intra-sujetos arbitraria. Dada la flexibilidad de la estructura de covarianzas intra-sujeto, los modelos posibles de ajustar pueden ser univariados o multivariados con datos incompletos, además de estructuras AR. Si y_i es el vector de dimensión $t_i \times 1$ que contiene las respuesta del sujeto i-ésimo, donde $i = 1, \ldots, n$, el modelo planteado es de la forma

$$y_i = X_i \alpha + e_i \tag{12}$$

donde X_i es una matriz de diseño conocida de dimensión $t_i \times p$, α es un vector $p \times 1$ de parámetros desconocidos y e_i son errores aleatorios N(0,Σ_i). Se asume que la matriz de covarianzas Σ_i contiene elementos que son función de un vector θ de parámetros desconocidos ($\Sigma_i = \Sigma_i(\theta)$). Asumiendo diferentes estructuras para Σ_i es lo que hace posible ajustar los modelos indicados por Jennrich y Schluchter. En particular, el caso $\Sigma_i = Z_i D Z_i^T + R_i$ surge asumiendo que $e_i = Z_i b_i + u_i$, donde Z_i y b_i son como en Laird y Ware, y u_i tiene distribución N(0,R_i).

Jennrich and Schluchter muestran que cuando se asume el modelo (12) para y_i, la obtención de $\widetilde{\alpha}$ y $\widetilde{\theta}$, los estimadores de α y el vector de componentes de la varianza θ respectivamente, se obtienen resolviendo ecuaciones separadas (algo ya observado por Ware en 1985), y describen la forma de usar los procedimientos iterativos de Newton-Raphson y Fisher's scoring para la estimación, además de un algoritmo que mezcla scoring y algoritmo EM generalizado (GEM) para el modelo de datos incompletos.

Laird, Lange y Stram en 1987 [32] muestran el uso del algoritmo EM para la estimación máximo verosimil no restringida y restringida (ML y REML) en un modelo lineal de efectos mixtos generalizado que incluye varias de las formulaciones propuestas anteriormente (basado en el trabajo de Laird y Ware). Los autores hacen notar, por ejemplo, que el modelo en 2 etapas de Rao en (5) y (6) es una versión restringida del modelo de Laird y Ware en (8) y (9). En efecto, en el modelo de Rao se tiene que $E(Y_i) = Z_i W_i \alpha$ y $Var(Y_i) = Z_i \Lambda Z_i' + R_i$, los que corresponden a $E(y_i)$ y Σ_i del modelo de Laird y Ware (definiendo en el modelo de Rao: $Z_i W_i = X_i$, lo cual impone una restricción a la matriz de diseño X que no presenta en el modelo de Laird y Ware). Por

otra parte, muestran que cuando el diseño es completo y balanceado, las estimaciones en el modelo de Laird y Ware son equivalentes a las obtenidas en el modelo de Potthoff y Roy mostrado en (1), el modelo de Grizzle y Allen mostrado en (3) o el modelo de Reisnel para el caso multivariado. Finalmente, los autores presentan valores iniciales para los parámetros para el caso de estimación iterativa de componentes de la varianza en el modelo de Laird y Ware.

Lindstrom y Bates en 1990 [33] proponen un modelo no lineal de efectos mixtos para datos longitudinales con errores normales. En este modelo, y_{ij}, la j-ésima observación del i-ésimo individuo, es representado como

$$y_{ij} = f(\phi_i, x_{ij}) + \epsilon_{ij} \tag{13}$$

donde x_{ij} es un vector de predictores para la j-ésima observación del i-ésimo individuo, f es una función nolineal de x_{ij} y ϕ_i es un vector de parámetros de dimensión $r \times 1$. Se asume además que el error aleatorio e_{ij} tiene distribución normal. Para capturar la variabilidad entre e intra individuos, se asume que $\phi_i = A_i\beta + B_i b_i$, donde β es vector $p \times 1$ de parámetros poblacionales fijos y b_i es vector $q \times 1$ de efectos aleatorios asociados el individuo iésimo, con $b_i \sim$ N(0,σ^2D). Derivadas parciales de f con respecto a β y b_i permiten obtener expresiones para las matrices de diseño X_i y Z_i, respectivamente, y la subsecuente estimación de los parámetros del modelo por máxima verisimilitud (ML) o máxima verisimilitud restringida (REML).

Una versión menos generalizada de modelo de regresión no lineal de efectos mixtos es propuesta por Vonesh y Carter en 1992 [56]. En este caso, el modelo es de la forma

$$y_i = f(X_i, \alpha_0) + Z_i b_i + \epsilon_i, , \qquad i = 1, \ldots, n \tag{14}$$

donde $y_i = (y_{i1}, \ldots, y_{ir_i})$ es el vector de observaciones del i-ésimo individuo, X_i es matriz de variables explicatorias conocida, α_0 es vector de parámetros desconocidos, $f(X_i, \alpha_0)$ es una función (posiblemente) no lineal, Z_i es matriz de contantes conocidas y b_i es vector de coeficientes de regresión aleatorios. Se asume que $\epsilon_i \sim N(0, \sigma^2 I_{n_i \times n_i})$ y $b_i \sim N(0, \Psi)$, con b_i independientes entre sí y de los ϵ_i. Este modelo incluye al modelo de Laird y Ware como caso particular (cuando $f(X_i, \alpha_0) = X_i\alpha_0$), al modelo de Reinsel de 1984 y puede ser modificado para incluir al modelo no lineal de Beal y Sheiner de 1979.

Otros desarrollos recientes se relacionan principalmente con el uso del algoritmo EM para la estimación máximo verosimil de parámetros en distintos diseños. Por ejemplo, Walker en 1996 [57] muestra el uso de algoritmo EM para estimación en modelos mixtos no lineales, Shah et al en 1997 [49] el uso de EM en modelos mixtos multivariados y Hogan y Laird en 1997 [25] lo utilizan en modelos para la distribución conjunta de medidas repetidas y tiempos de evento sujetos a censura.

Finalmente, para una revisión de la bibliografía disponible hasta el año 1994 sobre modelos de efectos mixtos puede verse el ya mencionado artículo de Yuh et al [61]. Para una introducción al uso del modelo lineal general de efectos mixtos en el análisis de medidas repetidas no balanceadas y datos longitudinales puede verse en el artículo

de Cnaan et al de 1997 [9], el capítulo 1 del libro de Pinheiro y Bates del año 2000 [39] o el libro de McCulloch y Searle del año 2001 [36].

1.3 Análisis de conglomerados y la descomposición de mezcla (Mixture Decomposition Scheme)

El análisis de conglomerados es un procedimiento que permite la clasificación de individuos en grupos (clusters o conglomerados) usando un vector x de variables conocidas, de modo que sujetos clasificados en un mismo conglomerado son homogéneos (similares en algún sentido respecto a sus valores en x) y estos sujetos son, a su vez, tan heterogéneos como sea posible respecto a los sujetos clasificados en los otros conglomerados.

El análisis de conglomerados pertenece a la clase de métodos de clasificación "no supervisada", en la cual los individuos no pertenecen a ningún gupo predefinido y la pertenencia a un grupo queda finalmente determinada por el conglomerado en el cual son clasificados, de acuerdo a sus características individuales en el vector x. Esta se diferencia de la clasificación "supervisada", en la cual los individuos están preclasificados en un grupo y el interés se centra en determinar, para un nuevo individuo, en qué medida su vector de características x permite anticipar su patrón de clasificación. Una revisión de diferentes métodos de clasificación supervisados y no supervisados puede verse en Jain et al [26].

El mecanismo mediante el cual un individuo es clasificado en un determinado conglomerado (mecanismo de clustering) depende de varios factores, entre los que destaca el tipo de variables usadas para la clasificación (características cuantitativas o cualitativas), la medida de proximidad o distancia utilizada (como distancia Euclideana, norma de Manhattan, distancia de Tanimoto, etc.) y el algoritmo de clustering. Respecto a estos últimos, hay varias formas de hacer una taxonomía de los algoritmos; por ejemplo, Theodoridis y Koutroumbas [53] los dividen en secuenciales, jerárquicos (divisivo y aglomerativo), esquemas basados en funciones de optimización de costos (esquema de descomposición de mezcla, algoritmos fuzzy, posibilísticos, etc.) y técnicas especiales (basados en teoría de grafos, algoritmo de detección de límites, etc.). Una taxonomía distinta, además de una extensa revisión bibliográfica sobre análisis de conglomerados, puede verse en la revisión de Jain et al [27]. Una descripción en detalle de todos los métodos indicados puede verse en el libro de Theodoridis y Koutroumbas [53].

Los algoritmos de clustering de interés en esta tesis son los basados en funciones de optimización de costos, en general, y el "mixture decomposition scheme" en particular, ya que en este esquema (también llamado mixture-resolving scheme) se asume que los individuos a ser agrupados pertenecen a una de varias distribuciones de probabilidad y el objetivo es identificar los parámetros de cada una de estas distribuciones. En la mayoría de los trabajos en esta área se asume que los componentes individuales de la mezcla son distribuciones normales, y los parámetros deben ser estimados por el procedimiento.

Una descripción de la metodología de estimación de parámetros en mezclas de distribuciones que pertenecen a la familia exponencial puede verse en el artículo de

Hasselblad de 1969 [23]. Wolfe, en 1970 [59] muestra una metodología general de estimación en mezclas multivariadas (y en particular la mezcla de distribuciones normales) usando el método de scoring de Fisher. La aparición del algoritmo EM en 1977 permitió adoptar una nueva forma de estimación de parámetros de las distribuciones de la mezcla. Esta metodología se describe a continuación en términos generales, sin asumir una función densidad determinada para los datos.

1.3.1 Mixture Decomposition Scheme.

Sea y_i el vector de respuestas de dimensión $n_i \times 1$ que almacena un conjunto de n_i observaciones medidas longitudinalmente para el i-ésimo sujeto, $i = 1 \dots n$. Si se asume una distribución de probabilidad para los y_i, interesa clasificar los n sujetos en alguno de m conglomerados, donde m es un número fijo, de modo que la clasificación permita simultáneamente la estimación de los parámetros del modelo asumido para y_i, al interior de cada conglomerado.

Sea C_i una variable aleatoria con valores posibles $1, 2, \dots, m$, que indica el conglomerado al que pertenece el i-ésimo individuo. En este esquema, la probabilidad a priori de que el vector de datos y_i medido longitudinalmente pertenezca al k-ésimo conglomerado es $P(C_i = k)$, lo cual se denotará como π_k. El vector $\pi = (\pi_1, \pi_2, \dots, \pi_m)$ es desconocido y debe cumplir con las restricciones

$$\pi_k \geq 0 \quad k = 1, \dots, m, \quad \sum_{j=1}^{m} \pi_j = 1 \tag{15}$$

La probabilidad a posteriori de que el sujeto i-ésimo pertenezca al conglomerado k is $P(C_i = k|y_i)$, lo cual se denotará como $\pi_{k|y_i}$. Luego, la regla de decisión es clasificar el i-ésimo sujeto en el conglomerado k si

$$P(C_i = k|y_i) > P(C_i = j|y_i) \quad j, k = 1, 2, \dots, m, \quad j \neq k \tag{16}$$

Asumiendo que C_i es un dato no observado, definimos $y_i^* = (y_i, C_i)$ como el vector de datos completos, donde y_i es la parte observada de y_i^* y C_i es su parte no observada. La distribución conjunta de y_i^* se puede escribir como

$$P(y_i^*) = P(y_i, C_i) = P(y_i|C_i = k)P(C_i = k) = \pi_{y_i|k}\pi_k \tag{17}$$

donde $P(y_i|C_i = k)$ es la distribución del vector y_i al interior del conglomerado k. Esta probabilidad será denotada como $\pi_{y_i|k}$. Se asume que esta distribución está indexada por parámetros desconocidos que deben ser estimados. Asumiendo que los y_i^* son independientes, la distribución conjunta de $y = (y_1^*, y_2^*, \dots, y_n^*)$ es

$$P(y^*) = \prod_{i=1}^{n} P(y_i|C_i = k)P(C_i = k) = \prod_{i=1}^{n} \pi_{y_i|k}\pi_k \tag{18}$$

y la log-verosimilitud es

$$\ell(y^*) = \sum_{i=1}^{n} \log \{P(y_i|C_i = k)P(C_i = k)\} = \sum_{i=1}^{n} \log \{\pi_{y_i|k}\pi_k\} \tag{19}$$

Entonces, la maximización de $\ell(y^*)$ permitirá obtener los parámetros del modelo asumido para y_i en el conglomerado k y las probabilidades a priori $\pi_1, \pi_2, \ldots, \pi_m$.

La estructura de datos incompletos asumida nos permite usar el algoritmo EM (Dempster, Laird and Rubin [13]) para la estimación de los parámetros desconocidos de la verosimilitud de y_i^*. Sea

$$Q(\theta, \theta^{(t)}) = E_{\theta^{(t)}}\{\ell(y_i^*)|y_i\} \tag{20}$$

donde $E()$ es el valor esperado de la log-verosimilitud de los datos completos, condicional en los datos observados en y_i^*. Esta esperanza es evaluada en $\theta^{(t)}$, el vector de parámetros en la iteración t.

El E-Step del algoritmo es

$$\begin{aligned} Q(\theta, \theta^{(t)}) &= E_{\theta^{(t)}}(\sum_{i=1}^{n} \log \{\pi_{y_i|k}\pi_k\}|y_i) \\ &= \sum_{i=1}^{n} E_{\theta^{(t)}}(log\{\pi_{y_i|k}\pi_k\}|y_i) \\ &= \sum_{i=1}^{n}\sum_{k=1}^{m} \pi_{k|y_i} \log \{\pi_{y_i|k}\pi_k\} \end{aligned} \tag{21}$$

donde

$$\pi_{k|y_i} = \frac{\pi_{y_i|k}\pi_k}{\sum_{k=1}^{m} \pi_{y_i|k}\pi_k} \tag{22}$$

El M-Step del algoritmo consiste en la obtención del parámetro $\theta^{(t+1)}$ que maximiza $\frac{\partial Q(\theta,\theta^{(t)})}{\partial \theta} = 0$. En este caso, si los θ_k son funcionalmente independendientes (es decir, para todo par de parámetros θ_i, θ_j se cumple que θ_i $(i \neq j)$ no aporta ninguna información para la estimación de θ_j), la maximización es equivalente a

$$\sum_{i=1}^{n}\sum_{k=1}^{m} \pi_{k|y_i} \frac{\partial}{\partial \theta_k} \log \pi_{y_i|k} = 0 \tag{23}$$

Usando Lagrange para la restricción en (15), es posible obtener una expresión para las probabilidades a priori π_k dada por

$$\pi_k = \frac{1}{n}\sum_{i=1}^{n} \pi_{k|y_i} \tag{24}$$

Este procedimiento general de estimación de parámetros será utilizado asumiendo que y_i sigue un modelo de efectos mixtos lineal o no lineal, en un esquema univariado y multivariado y considerando que el diseño puede estar sujeto a la ausencia de datos.

1.4 Algunos diseños de análisis de conglomerados usando mezclas de distribuciones y modelos de efectos mixtos

Existe amplia literatura respecto a clasificación supervisada usando modelos de efectos mixtos. Por ejemplo, en McLachlan and Gordon [37] y McLachlan and Basford [38] se usa "mixture decomposition scheme" para la clasificación en grupos predefinidos de observaciones que siguen un modelo lineal de efectos mixtos, donde se usa algoritmo EM para la clasificación de algunas observaciones asumiendo que se conoce la clasificación de parte de los datos. Marshall y Barón [35] hacen análisis discriminante de datos no balanceados medidos longitudinalmente que siguen un modelo de efectos mixtos lineal o no lineal.

La clasificación en conglomerados asumiendo que los datos en la mezcla provienen de distribuciones de probabilidad (normal o no normal) es introducida por Banfield y Raftery en 1993 [3]. Un enfoque de clasificación en conglomerados usando Gibbs sampling es propuesto por Bensmail et al en 1997 [8].

Verbeke y Lesaffre [55] analizan la clasificación en conglomerados en modelos lineales de efectos mixtos, asumiendo que los efectos aleatorios del modelo son muestreados a partir de una mezcla de g distribuciones normales. En este trabajo los autores usan Bayes empírico para la estimación de los parámetros. Una rutina para SAS que implementa el modelo de Verbeke y Lesaffre para modelos mixtos lineales y no lineales puede verse en Spiessens et al 2002 [51].

Trabajos más recientes incluyen el artículo de Fraley y Raftery de 2002 [17], donde se propone un método de clustering basado en mezclas de modelos, con aplicaciones en análisis discriminante y estimación de densidades multivariadas. Un trabajo posterior de Raftery y Dean de 2004 [42] aborda el problema de selección de variables del modelo propuesto en 2002.

Finalmente, de la Cruz-Mesía [10] y de la Cruz-Mesía, Quintana y Marshall [11] abordan la clasificación en conglomerados y análisis discriminante para datos longitudinales mediante un método de estimación y clasificación semi-bayesiano.

En la revisión bibliográfica hecha para esta tesis no se encontraron artículos que analizaran la clasificación en conglomerados asumiendo modelos mixtos lineales y no lineales, uni y multivariados, con todos los parámetros del modelo dependientes del conglomerado al que pertenece cada observación, como los que se describen a partir de la sección siguiente.

2 Análisis de conglomerados y el modelo de efectos mixtos

En esta sección se describe el uso de modelos de efectos mixtos lineal y no lineal, en diseños univariado y multivariado, para la clasificación en conglomerados de observaciones medidas longitudinalmente, posiblemente sujetas a observaciones faltantes. En la primera subsección se introduce el modelo lineal y no lineal mixto univariado y en la segunda se generaliza al modelo mixto no lineal multivariado. Finalmente, se describe la metodología utilizada para determinar el número de conglomerados subyacentes en los datos y algunos detalles computacionales.

2.1 Clasificación usando modelos de efectos mixtos univariados

Sea y_i el vector de respuestas de dimensión $n_i \times 1$ que almacena un conjunto de n_i observaciones medidas longitudinalmente para el i-ésimo sujeto, $i = 1 \dots n$. Luego, interesa clasificar los n sujetos en alguno de m conglomerados, donde m se considerará un número fijo. Siguiendo la notación en la sección (1.3.1), sea C_i una variable aleatoria con valores posibles $1, 2, \dots, m$, que indica el conglomerado al que pertenece el i-ésimo individuo. Si se asume el modelo lineal de efectos mixtos (12) de Jennrich y Schluchter para y_i, condicional en que $C_i = k$, entonces

$$y_i = X_i\alpha_k + e_{ik} \quad i = 1, 2, \dots, n \quad k = 1, 2, \dots, m \tag{25}$$

Entonces, $y_i|C_i = k \sim N(X_i\alpha_k, V_{ik})$. Esto es, y_i tiene distribución normal, donde α_k depende del conglomerado k y la varianza de y_i en $C_i = k$ es $V_{ik} = Z_iD_kZ_i^T + R_{ik}$. La matriz de covarianzas Σ_{ik} resulta de asumir que $e_{ik} = Z_ib_{ik} + u_{ik}$, donde $b_{ik} \sim N(0, D_k)$ y $u_{ik} \sim N(0, R_{ik})$. Se asume que los errores u_{ik} no están correlacionados al interior del i-ésimo sujeto y que tienen varianza constante. En otras palabras, $R_{ik} = \sigma_k^2 I_{n_i \times n_i}$. La matriz de covarianzas D_k y la varianza σ_k^2 también dependen del conglomerado al que pertenece y_i.

En notación de Laird y Ware, el modelo lineal de efectos mixtos para el i-ésimo sujeto al interior del conglomerado $C_i = k$ puede ser escrito como

$$y_i = X_i\alpha_k + Z_ib_{ik} + e_{ik} \quad i = 1, 2, \dots, n \quad k = 1, 2, \dots, m \tag{26}$$

Por otra parte, si se asume el modelo Lindstrom and Bates en (13) para y_i, entonces condicional en $C_i = k$ el modelo para y_{ij}, la j-ésima observación del sujeto i-ésimo, es de la forma

$$y_{ij} = f(\phi_{ik}, x_{ij}) + \epsilon_{ijk} \tag{27}$$

donde x_{ij} es como en (13) y ϕ_{ik} es el vector de parámetros desconocidos para el i-ésimo individuo al interior del conglomerado k. Usando expansión de Taylor en torno a un valor inicial $\phi_{ik}^0 = A_i\alpha_k^0 + B_ib_{ik}^0$, el modelo en (27) se puede aproximar por

$$y_{ij} - f(\phi_{ik}^0, x_{ij}) + \dot{f}(\phi_{ik}^0, x_{ij})'\phi_{ik}^0 = \dot{f}(\phi_{ik}^0, x_{ij})'\phi_{ik} + \epsilon_{ijk} \tag{28}$$

donde $\dot{f}$ es la derivada parcial de f con respecto a ϕ_{ik}. Al reemplazar ϕ_{ik}^0 en (28) por su expresión inicial $A_i\alpha_k^0 + B_i b_{ik}^0$, el modelo puede ser reescrito como

$$\tilde{y}_{ijk} = \tilde{x}_{ijk}\alpha_k + \tilde{z}_{ijk}b_{ik} + \epsilon_{ijk} \tag{29}$$

donde

$$\tilde{y}_{ijk} = y_{ijk} - f(\phi_{ik}^0, x_{ij}) + \dot{f}(\phi_{ik}^0, x_{ij})'\phi_{ik}^0 \tag{30}$$

$$\tilde{x}_{ijk} = \dot{f}(\phi_{ik}^0, x_{ij})'A_i \tag{31}$$

$$\tilde{z}_{ijk} = \dot{f}(\phi_{ik}^0, x_{ij})'B_i \tag{32}$$

Por lo tanto, el modelo para el i-ésimo sujeto al interior del conglomerado k es

$$\tilde{y}_{ik} = \tilde{x}_{ik}\alpha_k + \tilde{z}_{ik}b_{ik} + \epsilon_{ik} \tag{33}$$

2.1.1 Estimación de parámetros via Algoritmo EM

Interesa estimar, para un número fijo m de conglomerados, las probabilidades de clasificación a priori π_k, el vector de parámetros poblacionales α_k, la varianza σ_k^2 y los componentes de la varianza en D_k, para el caso lineal y el no lineal.

Sea $\theta_k = (\alpha_k, \tau_k, \pi_k)$ para $k = 1, 2, \ldots, m$, el vector de parámetros desconocidos a estimar. El vector τ_k contiene los componentes de la varianza, $\tau_k = (\sigma_k^2, D_k)$.

Asumiendo que $y_i|C_i = k \sim N(X_i\alpha_k, V_{ik})$ para $k = 1, 2, \ldots, m$, según el modelo lineal (25), la probabilidad (22) de clasificación a posteriori en el conglomerado k es de la forma

$$\hat{\pi}_{k|y_i} = \frac{|V_{ik}|^{-1/2} exp\{-\frac{1}{2}(y_i - X_i\alpha_k)^T V_{ik}^{-1}(y_i - X_i\alpha_k)\}\pi_k}{\sum_{k=1}^m |V_{ik}|^{-1/2} exp\{-\frac{1}{2}(y_i - X_i\alpha_k)^T V_{ik}^{-1}(y_i - X_i\alpha_k)\}\pi_k} \tag{34}$$

donde $\alpha_k = \alpha_k^{(\nu)}$ y $V_{ik} = V_{ik}^{(\nu)}$ son los valores de α_k y V_{ik} en la iteración ν.

Cuando en la función (23) a ser maximizada se reemplaza $\pi_{y_i|k}$ por su forma normal al interior del conglomerado k, ésta queda expresada como

$$\sum_{i=1}^n \sum_{k=1}^m \pi_{k|y_i} \frac{\partial}{\partial \theta_k}\{-\frac{1}{2}\log|V_{ik}| - \frac{1}{2}(y_i - X_i\alpha_k)^T V_{ik}^{-1}(y_i - X_i\alpha_k)\} = 0 \tag{35}$$

sujeta a la restricción $\sum_{j=1}^m \pi_j = 1$ mostrada en (15).

La expresión (35) requiere independencia funcional de los elementos en θ_k. Sin embargo, esto es cierto sólo para α_k, π_k y $\tau_k = (\sigma_k^2, D_k)$, ya que los elementos en τ_k no necesariamente cumplen con esta condición. Por esta razón, se usarán distintos caminos para la estimación de α_k y las varianzas en τ_k, sujeto a la restricción que hace posible la obtención de π_k.

Estimación de α_k y b_{ik}. Cuando se deriva la expresión (35) con respecto a α_k, se obtiene el estimador máximo verosimil $\hat{\alpha}_k$, dado por

$$\hat{\alpha}_k = (\sum_{i=1}^{n} \pi_{k|y_i} X_i^T V_{ik}^{-1} X_i)^{-1} (\sum_{i=1}^{n} \pi_{k|y_i} X_i^T V_{ik}^{-1} y_i) \tag{36}$$

Se observa que $\hat{\alpha}_k$ es el estimador habitual del vector de parámetros de efectos fijos en un modelo lineal de efectos mixtos, pero con cada sumando del numerador y denominador ponderados por la probabilidad a posteriori de que el sujeto pertenezca al conglomerado k.

El estimador de α_k se puede expresar matricialmente generando vectores y, b y ϵ y una matriz de diseño X superponiendo los n vectores y_i, b_i y ϵ_i y las matrices de diseño X_i, respectivamente. Luego, marginalmente el super vector de datos y es normal con media $X\alpha$ y matriz de covarianzas V_k, la cual es diagonal en bloques, con bloques iguales a $V_{ik} = Z_i D_k Z_i^T + R_{ik}$. Modificando V_k de modo que incorpore las probabilidades de pertenencia a posteriori $\pi_{k|y_i}$ en la forma $V_k^* = diag(\frac{1}{\pi_{k|y_1}} V_{1k}, \ldots, \frac{1}{\pi_{k|y_n}} V_{nk})$, el estimador de α_k expresado matricialmente es

$$\hat{\alpha}_k = (X^T V_k^{*-1} X)^{-1} X^T V_k^{*-1} y \tag{37}$$

En esta representación matricial, al interior del conglomerado $C_i = k$ los vectores y y ϵ son de dimensión $N \times 1$, la matriz X es de dimensión $N \times p$ y el vector b_k es de dimensión $Nq \times 1$, donde $N = \sum_{i=1}^{n} n_i$. La matriz V_k (o V_k^*) es de dimensión $N \times N$.

Nótese que en la expresión matricial (37) se puede reemplazar V_k^* por $\Pi_{k|y} V_k^*$, donde $\Pi_{k|y} = diag(\pi_{k|y_1}, \ldots, \pi_{k|y_n})$, con cada $\pi_{k|y_i}$ repetido n_i veces en la diagonal de $\Pi_{k|y}$.

Un método de estimación alternativo es una adaptación de las ecuaciones de modelo-mixto (mixed-model equations, MME), que consisten en un sistema lineal de ecuaciones de estimación que permiten obtener simultáneamente α_k y b_{ik}. Las ecuaciones originales pueden verse en Harville (ver Teorema 2 en [22]). Para esto, se define $Z = diag(Z_1, \ldots, Z_n)$, $D_k^* = diag(D_k, \ldots, D_k)$ y $R_k = diag(R_{1k}, \ldots, R_{nk})$. Luego, el sistema lineal de ecuaciones es

$$\begin{bmatrix} X^T \Pi_{k|y}^{-1} R_k^{-1} X & X^T \Pi_{k|y}^{-1} R_k^{-1} Z \\ Z^T R_k^{-1} X & D_k^{*-1} + Z^T R_k^{-1} Z \end{bmatrix} \begin{bmatrix} \hat{\alpha}_k \\ \hat{b}_k \end{bmatrix} = \begin{bmatrix} X^T \Pi_{k|y}^{-1} R_k^{-1} y \\ Z^T R_k^{-1} y \end{bmatrix}$$

Este sistema de ecuaciones resulta particularmente útil cuando el número n_i de observaciones individuales es muy grande, lo que implica invertir matrices V_{ik} de gran dimensión. La diferencia entre obtener la inversa de V_{ik} o de las matrices R_{ik} y D_k es que si estas últimas son matrices diagonales sus inversas son triviales de obtener. En nuestro caso, $R_{ik} = \sigma_k^2 I_{n_i \times n_i}$ y D_k es de dimensión $q \times q$, donde q, el número de parámetros aleatorios del modelo, generalmente es bajo. Por otra parte, la matriz de la izquierda en el sistema lineal anterior, la cual debe ser invertida para obtener $\hat{\alpha}_k$ y

$\hat{b}_k$, es de dimensión $(p+q)\times(p+q)$, lo que usualmente es considerablemente menor que la dimensión de V_{ik} aun cuando n_i no sea muy grande.

Nótese que el estimador de b_k obtenido del sistema de ecuaciones es

$$\hat{b}_k = (D_k^{*-1} + Z^T R_k^{-1} Z)^{-1} Z^T R_k^{-1}(y - X\hat{\alpha}_k) \tag{38}$$

el cual, usando la identidad $Z^T V_k^{-1} \equiv (I + Z^T R_k^{-1} Z D_k^*)^{-1} Z^T R_k^{-1}$ (ver Harville[22] página 323), puede escribirse como

$$\hat{b}_k = D_k^* Z^T V_k^{-1}(y - X\hat{\alpha}_k) \tag{39}$$

La expresión anterior corresponde a la media a posteriori de b_k condicional en los datos observados y, el estimador $\hat{\alpha}_k$ y las varianzas y componentes de la varianza en $\tau_k = (\sigma_k^2, D_k)$.

Para obtener los estimadores $\hat{\alpha}_k$ y $\hat{b}_k$ se requiere tener estimaciones de $\tau_k = (\sigma_k^2, D_k)$, de modo que $\hat{\alpha}_k = \hat{\alpha}_k(\hat{\tau})$ y $\hat{b}_k = \hat{b}_k(\hat{\tau})$. A continuación se muestra una forma de obtener estimaciones máximo verosímiles de la varianza σ_k^2 y de los componentes de la varianza en D_k, usando algoritmo EM.

Estimación de σ_k y D_k. Usando ideas introducidas por Laird y Ware[31], los componentes de la varianza en $\tau_k = (\sigma_k^2, D_k)$ serán estimados usando formas cuadráticas de b_{ik} y e_{ik} y los estimadores "actuales" de α_k y θ_k. Dado que $e_{ik}|C_i = k \sim N(0, \sigma_k^2 I_{n_i})$ y $b_{ik}|C_i = k \sim N(0, D_k)$, las ecuaciones de estimación para σ_k^2 y D_k son

$$\sum_{i=1|C_i=k}^{n} \{e_i^T e_i - n_i\sigma_k^2\} = 0 \tag{40}$$

y

$$\sum_{i=1|C_i=k}^{n} \{b_i b_i^T - nD_k\} = 0 \tag{41}$$

Luego, el estimador de σ_k^2 es

$$\hat{\sigma}_k^2 = \frac{\sum_{i=1}^n I(C_i = k) e_i^T e_i}{\sum_{i=1}^n I(C_i = k) n_i} \tag{42}$$

y el estimador de D_k es

$$\hat{D}_k = \frac{\sum_{i=1}^n I(C_i = k) b_i b_i^T}{n} \tag{43}$$

Las ecuaciones (42) y (43) constituyen el M-Step de la estimación de σ_k^2 y D_k. Si denominamos t_1 y t_2 a los numeradores de $\hat{\sigma}_k^2$ y $\hat{D}_k$, respectivamente, la forma de obtener nuevos estimadores de t_1 y t_2 es igualándolos a su esperanza, condicional en los

datos observados y_i y los estimadores actuales de α_k y θ_k. Sea $\hat{\tau}_k$ y $\hat{\alpha}_k$ los estimadores de τ_k y α_k, respectivamente. Entonces, el E-Step del algoritmo de estimación es

$$\hat{t}_1 = E(\sum_{i=1}^{n} I(C_i = k) e_i^T e_i | y_i, \hat{\alpha}_k(\hat{\tau}_k), \hat{\tau}_k) \tag{44}$$

$$\hat{t}_2 = E(\sum_{i=1}^{n} I(C_i = k) b_i b_i^T | y_i, \hat{\alpha}_k(\hat{\tau}_k), \hat{\tau}_k) \tag{45}$$

donde el valor esperado de (44) es obtenido considerando que $e_i|y_i, C_i = k \sim N((y_i - X_i\alpha_k)\Sigma_{ik}^{-1}R_k, R_k - R_k\Sigma_{ik}^{-1}R_k)$ donde $R_k = \sigma_k^2 I_{n_i}$. Entonces,

$$\begin{aligned} E(e_i^T e_i I(C_i = k)|y_i) &= \sum_{k=1}^{m} E_{e_i|y_i, C_i=k}(e_i^T e_i | y_i, C_i = k) I(C_i = k) \pi_{k|y_i} \\ &= E(e_i^T e_i | y_i, C_i = k) \pi_{k|y_i} \end{aligned} \tag{46}$$

Por lo tanto

$$\begin{aligned} \hat{t}_1 &= \sum_{i=1}^{n} E(e_i^T e_i | y_i, C_i = k) \pi_{k|y_i} \\ &= \sum_{i=1}^{n} \{\tilde{e}_{ik}^T(\tilde{\tau}_k)\tilde{e}_{ik}(\tilde{\tau}_k) + trVar(e_i|y_i, C_i = k)\}\pi_{k|y_i} \end{aligned} \tag{47}$$

donde $\tilde{e}_{ik} = E(e_i|y_i, C_i = k, \hat{\tau}_k, \hat{\alpha}_k(\hat{\tau}_k))$. Dado que $y_i \sim N(X_i\alpha, V_i)$ y $e_i \sim N(0, \sigma^2 I_{n_i})$, se tiene que $\tilde{e}_{ik}$, la esperanza de e_{ik} al interior del conglomerado $C_i = k$, condicional en los datos y_i y los valores "actuales" de τ_k y α_k está dada por

$$\tilde{e}_{ik} = \sigma_k^2 I_{n_i} V_{ik}^{-1}(y_i - X_i\alpha_k) \tag{48}$$

Por otra parte, la varianza condicional de e_i dado y_i y $C_i = k$ está dada por

$$Var(e_i|y_i, C_i = k) = \sigma_k^2(I_{n_i} - \sigma_k^2 V_{ik}^{-1}) \tag{49}$$

Similarmente, para t_2 se obtiene

$$\hat{t}_2 = \sum_{i=1}^{n} \{\tilde{b}_{ik}(\tilde{\tau}_k)\tilde{b}_{ik}^T(\tilde{\tau}_k) + Var(b_i|y_i, C_i = k)\}\pi_{k|y_i} \tag{50}$$

donde $\tilde{b}_{ik} = E(b_i|y_i, C_i = k, \tilde{\tau}_k, \tilde{\alpha}_k(\tilde{\tau}_k))$. Dado que $b_i \sim N(0, D)$, se tiene que $\tilde{b}_{ik}$ está dada por

$$\tilde{b}_{ik} = D_k Z_i^T V_{ik}^{-1}(y_i - X_i\hat{\alpha}_k) \tag{51}$$

Por otra parte, la varianza condicional de b_i dado y_i y $C_i = k$ está dada por

$$Var(b_i|y_i, C_i = k) = D_k - D_k Z_i^T V_{ik}^{-1} Z_i D \tag{52}$$

Dado que la expresión (52) subestima la varianza de $\tilde{b}_{ik} - b_{ik}$ al usar el estimador $\hat{\tau}_k$ en vez de τ_k, se usará la expresión

$$\begin{aligned} Var(b_i|y_i, C_i = k) &= D_k - D_k Z_i^T V_{ik}^{-1} Z_i D + D_k Z_i^T V_{ik}^{-1} X_i (\sum_{i=1}^{n} X_i^T V_{ik}^{-1} X_i)^{-1} X_i^T V_{ik}^{-1} Z_i D \\ &= D_k - D_k Z_i^T \{V_{ik}^{-1} + V_{ik}^{-1} X_i (\sum_{i=1}^{n} X_i^T V_{ik}^{-1} X_i)^{-1} X_i^T V_{ik}^{-1}\} Z_i D \end{aligned} \tag{53}$$

Entonces, los pasos (42),(43) representan el M-Step y (47),(50) representan el E-Step en la estimación de σ_k^2 y D_k. En los pasos de este algoritmo, $\tilde{\alpha}_k$ es considerado un valor fijo, como en el algoritmo de Jennrich y Schluchter. Una vez obtenidos nuevos estimadores de σ_k^2 y D_k, es necesario volver a la estimación de α_k; en este caso, $\hat{\sigma}_k^2$ y $\hat{D}_k$ son considerados valores fijos.

Estimación de parámetros en modelo mixto no lineal univariado. Cuando se asume el modelo no lineal (27), el procedimiento de estimación de parámetros es muy similar al descrito para el caso lineal, pero usando las versiones linealizadas de y_{ijk}, x_{ijk} y z_{ijk} dadas por las expresiones (30), (31) y (32).

En la sección (2.2) siguiente se muestra la estimación de parámetros para el caso multivariado, del cual la estimación univariada mostrada en esta sección es un caso particular.

2.2 Clasificación usando modelos de efectos mixtos multivariados

El desarrollo teórico de los modelos de efectos mixtos lineal y no lineal para clasificar individuos en conglomerados en un esquema de múltiples variables explicadas es similar al caso univariado en sección (2.1). Por este motivo, en esta sección se muestra sólo la resolución para el caso no lineal.

La estimación de parámetros de los modelos mixtos multivariados, tanto para el caso balanceado como no balanceado, es análoga a la utilizada por Marshall y otros[34] en un artículo actualmente en proceso de publicación. En ese trabajo, los modelos se utilizan para clasificación de individuos mediante análisis discriminante.

Para introducir el vector de respuestas multivariado, sea Y_i una matriz de dimensión $n_i \times p$ que almacena un conjunto de p variables respuestas para el i-ésimo sujeto en n_i tiempos diferentes. Es decir, Y_i es de la forma

$$Y_i = \begin{bmatrix} y_{i11} & y_{i12} & \cdots & y_{i1p} \\ y_{i21} & y_{i22} & \cdots & y_{i2p} \\ \ddots & \ddots & \ddots & \ddots \\ y_{in_i1} & y_{in_i2} & \cdots & y_{in_ip} \end{bmatrix}$$

Sea E_i la matriz $n_i \times p$ de términos de error asociados con Y_i. Introduciendo el operador vec que genera un vector vertical con las columnas de Y_i, obtenemos $y_i = vec(Y_i) = (y'_{i1}, y'_{i2}, \ldots, y'_{ip})$, donde y_i es el vector $pn_i \times 1$ de respuestas y $\epsilon_i = vec(E_i)$ es el vector $pn_i \times 1$ de errores asociados a y_i.

Si se supone que los individuos $y_i = vec(Y_i)$ pertenencen a un total fijo de m conglomerados, entonces se asume que y_{ijkg}, el individuo i, en el tiempo j, para la k-ésima respuesta, al interior del conglomerado $C_i = g$, sigue un modelo de efectos mixtos no lineal de la forma

$$y_{ijkg} = f_k(\eta_{ig}, x_{ij}) + \epsilon_{ijkg} \tag{54}$$

donde f_k es una función nolineal del vector de párametros η_{ig}, donde el subíndice k indica que la función cambia para cada respuesta $k = 1, \ldots, p$, y el vector de covariables x_{ij}. El vector ϵ_{ijkg} es de errores aleatorios asociado a y_{ijkg}. El vector de parámetros η_{ig} puede ser incorporado al modelo como $\eta_{ig} = A_i\beta_g + B_ib_{ig}$, donde β_g es un vector $q \times 1$ de parámetros poblacionales fijos al interior del conglomerado g, b_{ig} es un vector $r \times 1$ de efectos aleatorios individuales que también varían según el conglomerado y las matrices A_i y B_i son matrices de diseño de dimensión $s \times q$.

Se asume que $b_{ig} = b_i|C_i = g \sim MVN(0, D_g)$, con D_g de dimensión $r \times r$, y $\epsilon_{ig} = \epsilon_i|C_i = g \sim MVN(0, R_{ig})$ donde R_{ig} es matriz de varianzas-covarianzas de dimensión $pn_i \times pn_i$. Se asume además que para los individuos y_i y $y_{i'}$ se cumple que $cov(e_{ig}, e_{i'g}) = 0$ y que para un mismo individuo se cumple que $cov(\epsilon_{ig}, b_{ig}) = 0$.

Para darle una estructura a la matriz de errores intra-sujeto R_{ig}, consideremos $E_{i[j]g}$, la j-ésima fila de E_{ig}, la matriz de errores aleatorios del i-ésimo sujeto al interior del conglomerado g. Luego, $E_{i[j]g}$ tiene distribución

$$E_{i[j]g} = E_{i[j]}|C_i = g \sim MVN(0, \Sigma_g) \tag{55}$$

con Σ_g de dimensión $p \times p$, y $cov(E_{i[j]g}, E_{i[j']g}) = 0$ para todo $i = 1, \ldots, n$ y $j = 1, \ldots, n_i$, excepto cuando $i = i'$ y $j = j'$. Bajo estos supuestos, la matriz R_{ig} se puede escribir como

$$R_{ig} = \Sigma_g \otimes I_i \tag{56}$$

donde I_i es la matriz identidad de dimensión $n_i \times n_i$ y $\otimes$ indica el producto kronecker entre Σ_g e I_i.

Usando expansión en serie de Taylor de primer orden en torno a un valor inicial $\eta_{ig}^{(0)} = A_i\beta_g^{(0)} + B_i b_{ig}^{(0)}$ al interior del conglomerado g, el modelo (54) queda como

$$y_{ijkg} - f_k(\eta_{ig}^{(0)}, x_{ij}) + \dot{f}_k(\eta_{ig}^{(0)}, x_{ij})'\eta_{ig}^{(0)} = \dot{f}_k(\eta_{ig}^{(0)}, x_{ij})\eta_{ig} + \epsilon_{ijkg} \tag{57}$$

donde $\dot{f}$ es la primera derivada de f con respecto a η_g. El modelo (57) puede escribirse como

$$\tilde{y}_{ijkg} = \tilde{x}_{ijk}\beta_g + \tilde{z}_{ijk}b_{ig} + \epsilon_{ijkg} \tag{58}$$

donde $\tilde{y}_{ijkg} = y_{ijkg} - f_k(\eta_{ig}^{(0)}, x_{ij}) + \dot{f}_k(\eta_{ig}^{(0)}, x_{ij})'\eta_{ig}^{(0)}$, $\tilde{x}_{ijk} = \dot{f}_k(\eta_{ig}^{(0)}, x_{ij})'A_i$ es un vector de dimension $1 \times q$ y $\tilde{z}_{ijk} = \dot{f}_k(\eta_{ig}^{(0)}, x_{ij})'B_i$ es un vector de dimensión $1 \times r$. El modelo para el vector columna con las p pseudo respuestas del individuo i-ésimo, al interior del conglomerado g, es

$$\tilde{y}_{ig} = \tilde{X}_i\beta_g + \tilde{Z}_i b_{ig} + \epsilon_{ig} \tag{59}$$

donde $\tilde{X}_i$ es una matriz de dimensión $pn_i \times q$ y $\tilde{Z}_i$ es una matriz de dimensión $pn_i \times r$, las cuales se construyen con las filas de $\tilde{x}_{ijk}$ y $\tilde{z}_{ijk}$, respectivamente. Nótese que las matrices $\tilde{X}_i$ y $\tilde{Z}_i$ tambien son dependientes del conglomerado en que se calculen, ya que ambas son funciones de η_{ig}.

2.2.1 Estimación de parámetros via Algoritmo EM

Los supuestos distribucionales $b_i|C_i = g \sim MVN(0, D_g)$ y $\epsilon_i|C_i = g \sim MVN(0, R_{ig})$, con $R_{ig} = \Sigma_g \otimes I_i$, implican que $\tilde{y}_{ig} = \tilde{y}_i|C_i = g \sim MVN(\tilde{X}_i\beta_g, \tilde{Z}_i D_g \tilde{Z}_i' + \Sigma_g \otimes I_i)$. Luego, definiendo $V_{ig} = \tilde{Z}_i D_g \tilde{Z}_i' + \Sigma_g \otimes I_i$ y $W_{ig} = V_{ig}^{-1}$, la probabilidad en la expresión (22) de clasificación a posteriori en el conglomerado g (ver esquema de descomposición mixta en sección 1.3.1), es

$$\hat{\pi}_{g|\tilde{y}_i} = \frac{|V_{ig}|^{-1/2} exp\{-\frac{1}{2}(\tilde{y}_i - \tilde{X}_i\beta_g)^T W_{ig}(\tilde{y}_i - \tilde{X}_i\beta_g)\}\pi_g}{\sum_{k=1}^{m} |V_{ig}|^{-1/2} exp\{-\frac{1}{2}(\tilde{y}_i - \tilde{X}_i\beta_g)^T W_{ig}(\tilde{y}_i - \tilde{X}_i\beta_g)\}\pi_g} \tag{60}$$

donde $\beta_g = \beta_g^{(\nu)}$ and $V_{ig} = V_{ig}^{(\nu)}$ son los valores de β_g and V_{ig} en la iteración ν.

Cuando se reemplaza $\pi_{\tilde{y}_i|g}$ por su forma normal, la función (23) a ser maximizada es

$$\sum_{i=1}^{n}\sum_{k=1}^{m}\pi_{g|\tilde{y}_i}\frac{\partial}{\partial\theta_g}\{-\frac{1}{2}\log|V_{ig}|-\frac{1}{2}(\tilde{y}_i-\tilde{X}_i\beta_g)^T W_{ig}(\tilde{y}_i-\tilde{X}_i\beta_g)\}=0 \tag{61}$$

sujeto a la restricción $\sum_{j=1}^{m}\pi_j=1$ mostrada en (15).

Al igual que en el caso univariado, se usarán caminos diferentes para la estimación de los parámetros fijos y aleatorios β_g y b_{ig} y para las matrices de covarianza Σ_g y D_g.

Estimación de β_g y b_{ig}. Derivando la expresión (61) con respecto a β_g se obtiene el estimador máximo verosimil $\hat{\beta}_g$, dado por

$$\hat{\beta}_g=(\sum_{i=1}^{n}\pi_{g|\tilde{y}_i}\tilde{X}_i^T W_{ig}\tilde{X}_i)^{-1}(\sum_{i=1}^{n}\pi_{g|\tilde{y}_i}\tilde{X}_i^T W_{ig}\tilde{y}_i) \tag{62}$$

Al igual que en el caso univariado, el estimador de β_g es una versión ponderada del estimador de mínimos cuadrados generalizados de Aitken, con ponderación igual a la probabilidad a posteriori de pertenecer al conglomerado g.

La estimación de los errores entre-sujetos b_{ig} está dada por

$$\tilde{b}_{ig}=D_g\tilde{Z}_i^T W_{ig}(\tilde{y}_i-\tilde{X}_i\hat{\beta}_g) \tag{63}$$

La expresión anterior corresponde a la media a posteriori de b_{ig}, condicional en los datos observados $\tilde{y}_i$, el estimador $\hat{\beta}_g$ y las matrices de covarianzas Σ_g y D_g (la distribución de $b_i|\tilde{y}_i, C_i=g$ se muestra en la expresión (64)).

Estimación de Σ_g y D_g. Para la estimación de las matrices de covarianzas se usará el algoritmo EM, asumiendo que los datos completos son $\tilde{y}_{ig}$, b_{ig}, ϵ_{ig} y C_i, el conglomerado al que pertenece $\tilde{y}_{ig}$. La parte no observada de los datos corresponde a b_{ig}, ϵ_{ig} y C_i. Sucesivas estimaciones de D_g y Σ_g permitirán a su vez actualizar las estimaciones de β_g.

El estadístico suficiente para Σ_g es $\sum_{i=1}^{n}E_{ig}'E_{ig}$, donde E_{ig} es la matriz $n_i\times p$ de errores aleatorios de Y_i al interior del conglomerado g. El estadístico suficiente para D_g es $\sum_{i=1}^{n}b_{ig}b_{ig}'$. Usando la distribución conjunta de los datos completos, podemos obtener la distribución condicional de los estadísticos suficientes como

$$b_i|\tilde{y}_i, C_i=g\sim N(D_g\tilde{Z}_i'W_{ig}(\tilde{y}_i-\tilde{X}_i\beta_g), D_g-D_g\tilde{Z}_i'W_{ig}\tilde{Z}_iD_g) \tag{64}$$

y

$$\epsilon_i|\tilde{y}_i, C_i=g\sim N(\tilde{y}_i-\tilde{X}_i\beta_g-\tilde{Z}_ib_{ig}, R_{ig}-R_{ig}W_{ig}R_{ig}) \tag{65}$$

Basados en estos resultados, podemos encontrar los primeros dos momentos de la distribución condicional de b_i dados los datos observados como

$$\tilde{b}_{ig} = E\{b_i|\tilde{y}_i, C_i = g, \beta_g^{(\nu)}, D_g^{(\nu)}, \Sigma_g^{(\nu)}\} = D_g^{(\nu)}\tilde{Z}_i' W_{ig}^{(\nu)}(\tilde{y}_i - \tilde{X}_i\beta_g^{(\nu)}) \tag{66}$$

y

$$E\{b_i b_i'|\tilde{y}_i, C_i = g, \beta_g^{(\nu)}, D_g^{(\nu)}, \Sigma_g^{(\nu)}\} = \tilde{b}_i\tilde{b}_i' + D_g^{(\nu)} - D_g^{(\nu)}\tilde{Z}_i' W_{ig}^{(\nu)}\tilde{Z}_i D_g^{(\nu)} \tag{67}$$

y los dos primeros momentos de la distribución condicional de ϵ_i dados los datos observados son

$$\tilde{\epsilon}_{ig} = E\{\epsilon_i|\tilde{y}_i, C_i = g, \beta_g^{(\nu)}, D_g^{(\nu)}, \Sigma_g^{(\nu)}\} = \tilde{y}_i - \tilde{X}_i\beta_g^{(\nu)} - \tilde{Z}_i\tilde{b}_{ig} \tag{68}$$

y

$$E\{\epsilon_i\epsilon_i'|\tilde{y}_i, C_i = g, \beta_g^{(\nu)}, D_g^{(\nu)}, \Sigma_g^{(\nu)}\} = \tilde{\epsilon}_i\tilde{\epsilon}_i' + R_{ig}^{(\nu)} - R_{ig}^{(\nu)}W_{ig}^{(\nu)}R_{ig}^{(\nu)} \tag{69}$$

Para la estimación de Σ_g, se puede usar alternativamente la esperanza condicional de las filas de la matriz E_{ig}, la cual denominamos $E_{i[j]g}$ (ver expresión (55)). En este caso, se debe calcular

$$\tilde{E}_{i[j]g}' = E\{E_{i[j]g}|\tilde{y}_i, C_i = g, \beta_g^{(\nu)}, D_g^{(\nu)}, \Sigma_g^{(\nu)}\} \tag{70}$$

y

$$E\{E_{i[j]g}' E_{i[j]g}|\tilde{y}_i, C_i = g, \beta_g^{(\nu)}, D_g^{(\nu)}, \Sigma_g^{(\nu)}\} = \tilde{E}_{i[j]g}\tilde{E}_{i[j]g}' + \Sigma_g^{(\nu)} - \Sigma_g^{(\nu)}W_{i[j,j]g}^{(\nu)}\Sigma_g^{(\nu)} \tag{71}$$

donde $W_{i[j,j]g}^{(\nu)}$ es una matriz $p \times p$ con los elementos de $W_{ig}^{(\nu)}$ correspondientes a la observación en el tiempo j. El cálculo de estas esperanzas condicionales en los datos observados constituyen el E-Step del algoritmo. El M-Step esta dado por

$$D_g^{(\nu+1)} = \frac{1}{n}\sum_{i=1}^{n}\pi_{g|\tilde{y}_i}\{\tilde{b}_i\tilde{b}_i' + D_g^{(\nu)} - D_g^{(\nu)}\tilde{Z}_i' W_{ig}^{(\nu)}\tilde{Z}_i D_g^{(\nu)}\} \tag{72}$$

y

$$\Sigma_g^{(\nu+1)} = \frac{1}{\sum_{i=1}^{n}\pi_{g|\tilde{y}_i}n_i}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\pi_{g|\tilde{y}_i}\{\tilde{E}_{i[j]g}\tilde{E}_{i[j]g}' + \Sigma_g^{(\nu)} - \Sigma_g^{(\nu)}W_{i[j,j]g}^{(\nu)}\Sigma_g^{(\nu)}\} \tag{73}$$

Si se define $H_{ij} = I_p \otimes a_{ij}$, donde a_{ij} es la j-ésima fila de la matriz identidad $I_{n_i \times n_i}$, entonces la expresión (73) para $\Sigma_g^{(\nu+1)}$ puede escribirse como

$$\Sigma_g^{(\nu+1)} = \frac{1}{\sum_{i=1}^{n}\pi_{g|\tilde{y}_i}n_i}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\pi_{g|\tilde{y}_i}\{\tilde{E}_{i[j]g}\tilde{E}_{i[j]g}' + \Sigma_g^{(\nu)} - \Sigma_g^{(\nu)}H_{ij}W_{ig}^{(\nu)}H_{ij}'\Sigma_g^{(\nu)}\} \tag{74}$$

Para ilustrar el uso de $H_{ij} = I_p \otimes a_{ij}$, supongamos un modelo con $p = 2$ variables explicadas y un individuo con $n_i = 2$ observaciones longitudinales. Luego, para el tiempo $j = 1$ se tiene

$$H_{ij} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \otimes \begin{bmatrix} 1 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}$$

de modo que, para $j = 1$, la expresión $H_{ij} W_{ig}^{(\nu)} H_{ij}'$ considera el primer elemento de la primera variable explicada y el primer elemento de la segunda variable explicada en $W_{ig}^{(\nu)}$, de la misma forma en que lo hace $W_{i[j,j]g}^{(\nu)}$ en la expresión (73).

2.2.2 Estimación de parámetros para el caso desbalanceado

Sea O_i una matriz con solo *unos* y *ceros*, generada a partir de la matriz identidad de dimensión $pn_i \times pn_i$ (asociada al vector de datos y_i), a la cual se eliminan las filas correspondientes a las observaciones ausentes. Usando O_i se puede construir un nuevo vector de datos $y_i^0 = O_i y_i$, el cual contiene solo los datos realmente observados para las p variables contenidas en el super vector y_i. Un caso particular ocurre cuando un individuo no tiene observaciones ausentes, en cuyo caso $O_i = I_{pn_i \times pn_i}$ y $y_i^0 = y_i$.

Al premultiplicar por O_i el modelo linealizado en (59), se obtiene

$$\tilde{y}_{ig}^0 = \tilde{X}_i^0 \beta_g + \tilde{Z}_i^0 b_{ig} + \epsilon_{ig}^0 \tag{75}$$

donde $\tilde{X}_i^0 = O_i \tilde{X}_i$, $\tilde{Z}_i^0 = O_i \tilde{Z}_i$ y $\epsilon_{ig}^0 = O_i \epsilon_{ig}$.

Con esta modificación, se tiene que $\tilde{y}_{ig}^0 \sim MVN(\tilde{X}_i^0 \beta_g, V_{ig}^0)$, con $V_{ig}^0 = \tilde{Z}_i^0 D_g^{(\nu)} \tilde{Z}_i^{0\prime} + O_i R_{ig}^{(\nu)} O_i'$ y R_{ig} es como se define en (56).

La estimación del vector de parámetros β al interior del conglomerado g para el caso no balanceado queda como

$$\hat{\beta}_g = (\sum_{i=1}^{n} \pi_{g|\tilde{y}_i} \tilde{X}_i^{0\prime} W_{ig}^0 \tilde{X}_i^0)^{-1} (\sum_{i=1}^{n} \pi_{g|\tilde{y}_i} \tilde{X}_i^{0\prime} W_{ig}^0 \tilde{y}_i^0) \tag{76}$$

donde W_{ig}^0 es la matriz inversa de V_{ig}^0.

Usando la distribución conjunta de los datos completos $(\tilde{y}_i^0, C_i, \beta_g, D_g, \Sigma_g)$, podemos obtener la distribución condicional de los estadísticos suficientes para el caso no balanceado,

$$b_i | \tilde{y}_i^0, C_i = g \sim N(D_g \tilde{Z}_i^{0\prime} W_{ig}^0 (\tilde{y}_i^0 - \tilde{X}_i^0 \beta_g), D_g - D_g \tilde{Z}_i^{0\prime} W_{ig}^0 \tilde{Z}_i^0 D_g) \tag{77}$$

y

$$\epsilon_i|\tilde{y}_i^0, C_i = g \sim N(R_{ig}O_i'W_i^0(\tilde{y}_i^0 - \tilde{X}_i^0\beta_g), R_{ig} - R_{ig}O_i'W_{ig}^0O_iR_{ig}) \tag{78}$$

En forma análoga al caso balanceado, podemos encontrar los primeros dos momentos de la distribución condicional de b_{ig} dados los datos observados como

$$\tilde{b}_{ig} = E\{b_i|\tilde{y}_i^0, C_i = g, \beta_g^{(\nu)}, D_g^{(\nu)}, \Sigma_g^{(\nu)}\} = D_g^{(\nu)}\tilde{Z}_i^{0\prime}W_{ig}^{0(\nu)}(\tilde{y}_i^0 - \tilde{X}_i^0\beta_g^{(\nu)}) \tag{79}$$

y

$$E\{b_ib_i'|\tilde{y}_i^0, C_i = g, \beta_g^{(\nu)}, D_g^{(\nu)}, \Sigma_g^{(\nu)}\} = \tilde{b}_i\tilde{b}_i' + D_g^{(\nu)} - D_g^{(\nu)}\tilde{Z}_i^{0\prime}W_{ig}^{0(\nu)}\tilde{Z}_i^0D_g^{(\nu)} \tag{80}$$

y los dos primeros momentos de la distribución condicional de ϵ_i dados los datos observados son

$$\tilde{\epsilon}_{ig} = E\{\epsilon_i|\tilde{y}_i^0, C_i = g, \beta_g^{(\nu)}, D_g^{(\nu)}, \Sigma_g^{(\nu)}\} = R_{ig}^{(\nu)}O_i'W_i^{0(\nu)}(\tilde{y}_i^0 - \tilde{X}_i^0\beta_g^{(\nu)}) \tag{81}$$

y

$$E\{\epsilon_i\epsilon_i'|\tilde{y}_i^0, C_i = g, \beta_g^{(\nu)}, D_g^{(\nu)}, \Sigma_g^{(\nu)}\} = \tilde{\epsilon}_i\tilde{\epsilon}_i' + R_{ig}^{(\nu)} - R_{ig}^{(\nu)}O_i'W_{ig}^{0(\nu)}O_iR_{ig}^{(\nu)} \tag{82}$$

Si se considera la esperanza condicional de las filas de E_{ig} como en (70) y (71), se tiene que

$$\tilde{E}_{i[j]g}' = E\{E_{i[j]g}|\tilde{y}_i^0, C_i = g, \beta_g^{(\nu)}, D_g^{(\nu)}, \Sigma_g^{(\nu)}\} \tag{83}$$

y

$$E\{E_{i[j]g}'E_{i[j]g}|\tilde{y}_i^0, C_i = g, \beta_g^{(\nu)}, D_g^{(\nu)}, \Sigma_g^{(\nu)}\} = \tilde{E}_{i[j]g}\tilde{E}_{i[j]g}' + \Sigma_g^{(\nu)} - \Sigma_g^{(\nu)}H_{ij}O_i'W_{i[j,j]g}^{0(\nu)}O_iH_{ij}\Sigma_g^{(\nu)} \tag{84}$$

El cálculo de estas esperanzas condicionales en los datos observados constituyen el E-Step del algoritmo para el caso no balanceado. El M-Step está dado por la estimación de D_g y Σ_g como

$$D_g^{(\nu+1)} = \frac{1}{n}\sum_{i=1}^{n}\pi_{g|\tilde{y}_i^0}\{\tilde{b}_i\tilde{b}_i' + D_g^{(\nu)} - D_g^{(\nu)}\tilde{Z}_i^{0\prime}W_{ig}^{0(\nu)}\tilde{Z}_i^0D_g^{(\nu)}\} \tag{85}$$

y

$$\Sigma_g^{(\nu+1)} = \frac{1}{\sum_{i=1}^{n}\pi_{g|\tilde{y}_i^0}n_i}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\pi_{g|\tilde{y}_i^0}\{\tilde{E}_{i[j]g}\tilde{E}_{i[j]g}' + \Sigma_g^{(\nu)} - \Sigma_g^{(\nu)}H_{ij}O_i'W_{ig}^{0(\nu)}O_iH_{ij}'\Sigma_g^{(\nu)}\} \tag{86}$$

2.3 Análisis de conglomerados incorporando una variable explicada fija al modelo longitudinal de efectos mixtos

En esta sección se describe una metodología que permite incluir dos tipos de variables explicadas en un mismo modelo de efectos mixtos: una variable medida longitudinalmente en uno o más tiempos arbitrarios y una variable que es considerada fija. Se considera la posibilidad de que esta última variable sea medida en forma independiente del tiempo de seguimiento o que sea medida en un solo tiempo durante el transcurso del seguimiento (y por lo tanto se conoce el tiempo en que se hace esta medición).

La hipótesis que justifica un modelo de estas características es que algunas variables explicadas fijas, al ser incorporadas al análisis, pueden hacer un aporte a la identificación de los conglomerados subyacentes en los datos.

El modelo mixto lineal considerado para la variable longitudinal (Modelo-1), es de la forma

$$Modelo\,1: \quad y_i = X_i\alpha + Z_i b_i + \epsilon_i \qquad i = 1, \ldots, n \tag{87}$$

donde y_i es el vector longitudinal de respuestas $n_i \times 1$ para el i-ésimo sujeto. El vector α es $p \times 1$ de parámetros poblacionales fijos y b_i es un vector $q \times 1$ de parámetros aleatorios individuales. X_i y Z_i son matrices de diseño conocidas de dimensión $n_i \times p$ y $n_i \times q$, repectivamente. Los errores ϵ_i son independientes con distribución $\mathrm{N}(0,\sigma^2 I_{n_i \times n_i})$ y los vectores $b_i \sim \mathrm{N}(0,\mathrm{D})$ y son independientes entre sí y de los ϵ_i, con D matriz de covarianzas de dimensión $q \times q$. Con esta formulación, la distribución marginal de y_i es normal, con media $X_i\alpha$ y matriz de covarianzas $V_i = Z_i D Z_i^T + R_i$.

Para la variable respuesta fija r_i se asume el modelo

$$Modelo\,2: \quad r_i = f(t)\beta + b_i' \qquad i = 1, \ldots, n \tag{88}$$

donde β es un parámetro fijo y b_i' es una perturbación aleatoria del i-ésimo sujeto en torno a la media poblacional $f(t)\beta$, con $b_i' \sim N(0, u)$. No hay varianza intra-sujeto ya que r_i es una variable fija. La función $f(t)$ puede tomar uno de dos posibles valores:

- $f(t) = 1$. Cuando la variable r_i es fija y no depende de ningún tiempo t para su medición. Por ejemplo, en un estudio de seguimiento de embarazadas, la variable respuesta "peso pre-embarazo" no depende de ningún tiempo de seguimiento.

- $f(t) = t$. Cuando la medición de r_i se hace una sola vez en un tiempo específico t. Por ejemplo, una variable respuesta medida en algún momento en el transcurso del embarazo podría depender del tiempo t en que ésta es medida.

El modelo conjunto que considera como respuesta el vector $y_i^* = (y_i, r_i)^T$ puede escribirse de la forma

$$y_i^* = \begin{bmatrix} y_i \\ r_i \end{bmatrix} = \begin{bmatrix} X_i & 0 \\ 0 & f(t) \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} + \begin{bmatrix} Z_i & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} b_i \\ b_i' \end{bmatrix} + \begin{bmatrix} \epsilon_i \\ 0 \end{bmatrix}$$

o bien

$$y_i^* = X_i^* \beta^* + Z_i^* b_i^* + \epsilon_i^* \qquad i = 1, \ldots, n \tag{89}$$

donde $y_i^* = (y_i, r_i)^T$ es $(n_i + 1) \times 1$, la matriz X_i^* es de dimensión $(n_i + 1) \times (p + 1)$ y Z_i^* es de dimensión $(n_i + 1) \times (q + 1)$. El vector de parámetros fijos β^* es $(p + 1) \times 1$ y el de parámetros aleatorios b_i^* es $(q + 1) \times 1$. Finalmente, el vector de errores aleatorios ϵ_i^* es de dimensión $(n_i + 1) \times 1$.

Si se asume que $cov(\epsilon_i, b_i') = 0$ y $cov(\epsilon_i', b_i) = 0$, para todo $i = 1, \ldots, n$, entonces, la matriz de covarianzas de ϵ_i^* está dada por

$$R_i^* = \begin{bmatrix} \sigma^2 & 0 & \ldots & 0 & 0 \\ 0 & \sigma^2 & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & 0 & 0 \\ 0 & 0 & \ldots & \sigma^2 & 0 \\ 0 & 0 & \ldots & 0 & 0 \end{bmatrix} = \begin{bmatrix} \sigma^2 I_{n_i \times n_i} & 0_{n_i \times 1} \\ 0_{1 \times n_i} & 0 \end{bmatrix}$$

La matriz de covarianzas R_i^* es de dimensión $(n_i + 1) \times (n_i + 1)$. Se asume entonces que el vector $\epsilon_i^* \sim N(0, R_i^*)$. Por simplicidad, se usará la notación $R_i^* = \sigma^2 I_i^*$, donde I_i^* es similar a una matriz identidad de dimensión $(n_i + 1) \times (n_i + 1)$, pero con un 0 como último valor en la diagonal en vez de un 1.

Respecto al vector de parámetros aleatorios $b_i^* = (b_i, b_i')^T$, se asume que tiene distribución $N(0, D^*)$, con D^* matriz covarianzas de la forma

$$D^* = \begin{bmatrix} D_{q \times q} & C_{q \times 1} \\ C_{1 \times q}^T & u \end{bmatrix}$$

donde el vector C almacena la covarianza del parámetro aleatorio b_i' con cada uno de los q parámetros aleatorios en el vector b_i. Es decir, $C = (cov(b_{i1}, b_i'), \ldots, cov(b_{iq}, b_i'))^T$.

En el modelo asumido podemos obtener la esperanza y varianza de y_i^*, las cuales están dadas por $E(y_i^*) = X_i^* \beta^*$ y $Var(y_i^*) = V_i^* = Z_i^* D^* Z^{*T} + R_i^*$. Al escribir en forma matricial la matriz de covarianzas V_i^*, se tiene

$$V_i^* = \begin{bmatrix} Z_i & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} D_{q \times q} & C_{q \times 1} \\ C_{1 \times q}^T & u \end{bmatrix} \begin{bmatrix} Z_i^T & 0 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} \sigma^2 I_{n_i \times n_i} & 0_{n_i \times 1} \\ 0_{1 \times n_i} & 0 \end{bmatrix}$$

luego,

$$V_i^* = \begin{bmatrix} Z_i D Z_i^T + \sigma^2 I_{n_i} & Z_i C \\ C^T Z_i^T & u \end{bmatrix}$$

Finalmente,

$$y_i^* \sim N(X_i^* \beta^*, V_i^*) \tag{90}$$

Estimación de parámetros. La estimación del vector $\theta_k^* = (\beta_k^*, \sigma_k^2, D_k^*, \pi_k)$ para $k = 1, \ldots, m$, donde m es un número fijo de conglomerados subyacentes en los datos, es análoga a la estimación en el modelo mixto univariado, descrito en sección (2.1.1). En este caso, al función a maximizar es de la forma

$$\sum_{i=1}^{n}\sum_{k=1}^{m}\pi_{k|y_i^*}\frac{\partial}{\partial\theta_k^*}\{-\frac{1}{2}\log|V_{ik}^*| - \frac{1}{2}(y_i^* - X_i^*\beta_k^*)^T V_{ik}^{*-1}(y_i^* - X_i^*\beta_k^*)\} = 0 \tag{91}$$

la cual al ser derivada con respecto a β_k^* se obtiene el estimador máximo verosímil del vector de parámetros fijos del modelo, dado por

$$\hat{\beta}_k^* = (\sum_{i=1}^{n}\pi_{k|y_i^*}X_i^{*T}V_{ik}^{*-1}X_i^*)^{-1}(\sum_{i=1}^{n}\pi_{k|y_i}X_i^{*T}V_{ik}^{*-1}y_i^*) \tag{92}$$

Para la estimación de σ_k^2 y D_k^* utilizamos algoritmo EM, donde el M-step está dado por estimar

$$\hat{\sigma}_k^2 = \frac{t_1}{\sum_{i=1}^{n} I(C_i = k)n_i} = \frac{\sum_{i=1}^{n} I(C_i = k))e_i^{*T}e_i^*}{\sum_{i=1}^{n} I(C_i = k)n_i} \tag{93}$$

y

$$\hat{D}_k^* = \frac{t_2}{n} = \frac{\sum_{i=1}^{n} I(C_i = k)b_i^*b_i^{*T}}{n} \tag{94}$$

y el E-Step está dado por la actualización de los estadísticos suficientes t_1 y t_2. La actualización de t_1 está dada por

$$\hat{t}_1 = \sum_{i=1}^{n}\{\tilde{e}_{ik}^{*T}\tilde{e}_{ik}^* + trVar(e_i^*|y_i^*, C_i = k)\}\pi_{k|y_i} \tag{95}$$

donde $\tilde{e}_{ik}^* = \sigma^2 I_i^* V_{ik}^{*-1}(y_i^* - X_i^*\beta_k^*)$ y $Var(e_i^*|y_i^*, C_i = k) = \sigma^2(I_i^* - \sigma^2 V_{ik}^{*-1})$. La actualización de t_2 es

$$\hat{t}_2 = \sum_{i=1}^{n}\{\tilde{b}_{ik}^*\tilde{b}_{ik}^{*T} + Var(b_i^*|y_i^*, C_i = k)\}\pi_{k|y_i} \tag{96}$$

donde $\tilde{b}_{ik}^* = D_k^* Z_i^{*T} V_{ik}^{*-1}(y_i^* - X_i^*\beta_k^*)$ y $Var(b_i^*|y_i^*, C_i = k) = D_k^* - D_k^* Z_i^{*T} V_{ik}^{*-1} Z_i^* D^*$

2.4 Número de Conglomerados

Para determinar el número máximo de conglomerados subyacentes en los datos medidos longitudinalmente se optó por utilizar tres métodos: Test de razón de verosimilitud, el Criterio de Información de Akaike (AIC) [1] y el Criterio de Información Bayesiana (BIC) [47].

Si la log-verosimilitud marginal de y en un esquema que permite identificar m conglomerados está dada por

$$\ell(y) = \sum_{i=1}^{n} \log \sum_{k=1}^{m} f(y_i|C_i = k) f(C_i = k) \tag{97}$$

Entonces es necesario ajustar modelos para determinar un número fijo de $m = 1, 2, 3, \ldots$ conglomerados y luego comparar estas verosimilitudes. El número de conglomerados a considerar será el que significativamente incremente la log-verosimilitud de y, más allá del número de parámetros adicionales incluidos en el modelo en cada paso.

El Criterio de Información de Akaike se define como

$$AIC = -2\ell(y) + 2n_{par} \tag{98}$$

donde n_{par} es el número de parámetros del modelo ajustado. Por otra parte, el Criterio de Información de Bayesiana se define como

$$BIC = -2\ell(y) + n_{par} log(N) \tag{99}$$

donde N es el número de observaciones usadas para ajustar el modelo. Tanto para el criterio AIC como BIC se prefiere el modelo que tenga el menor valor.

Algunos métodos alternativos descritos en la literatura para identificar el número de conglomerados, y su posible aplicación a la detección de conglomerados usando modelos mixtos, se describen en la Discusión de esta tesis, en sección (4).

3 Ejemplos.

En esta sección se muestran tres ejemplos. El primero es un ejemplo de determinación de conglomerados usando modelos mixtos lineal y no lineal univariado, en un esquema con datos completos, en el cual se quiere estimar los parámetros de las distribuciones componentes para un total de g conglomerados, donde se quiere determinar también el número de conglomerados subyacentes en los datos. El segundo ejemplo muestra el uso de una variable respuesta adicional, también medida longitudinalmente, al esquema planteado en el ejemplo 1, de modo que se cuenta con $p = 2$ variables medidas longitudinalmente, con datos sujetos a observaciones faltantes, en el cual se quiere identificar un total de $g = 2$ conglomerados. Finalmente, el tercer ejemplo muestra el ajuste conjunto de una variable explicada medida longitudinalmente y una variable explicada fija.

3.1 Ejemplo 1. Modelo longitudinal mixto univariado

Los datos a usar en este ejemplo corresponden a un estudio de seguimiento de 161 mujeres embarazadas en un período de dos años en una clínica privada de Santiago. Para estas mujeres se midieron las variables beta-subunit human chorionic gonadotropin (β-HCG) y estradiol durante los primeros 80 días de gestación. En el estudio, las mujeres que tuvieron un embarazo y parto normal fueron clasificadas como normales; si presentaban cualquier complicación con resultado de embarazo interrumpido con pérdida del feto fueron clasificadas como patológicas. De las 161 mujeres consideradas en el estudio, 124 tuvieron parto normal (77%) y 37 resultaron en embarazo interrumpido (23%).

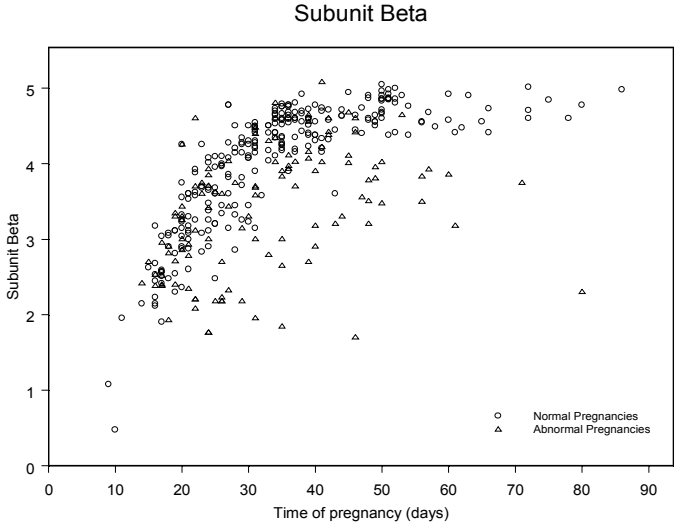

Figure 1: β-HCG en los primeros 80 días de gestación

Para este ejercicio, el resultado del embarazo (parto normal o anormal) fue omitido de los datos, con la finalidad de determinar la capacidad predictiva de los mecanismos de cluster usando diferentes modelos de efectos mixtos, en los cuales la variable respuesta es β-HCG (en el ejemplo univariado) y agregando como respuesta la variable estradiol (en el ejemplo multivariado). En ambos casos los modelos son en función de la edad gestacional en que se midieron las variables.

3.1.1 Modelos y datos

En este ejemplo se analiza el ajuste de modelos mixtos lineal y no lineal univariados para la variable β-HCG como predictor de embarazo interrumpido. La figura 1 muestra los datos analizados para el total de mujeres en estudio.

Tres modelos fueron ajustados a los datos: un modelo lineal por partes, con cambio de pendiente en $t = 30$, que aparece como el momento (aproximado) de inflexión de los datos de β-HCG, un modelo lineal usando el logaritmo de las semanas de gestación como variable explicatoria de β-HCG y el modelo no lineal curva de respuesta logística. Estos modelos se describen en la Tabla 1 para la i-ésima observación en tiempo j.

Los dos modelos lineales descritos en Tabla 1 (modelo lineal por partes y transformación log) fueron ajustados incluyendo efectos aleatorios para todos los parámetros de interés. Para el ajuste del modelo logístico se incluyó un efecto aleatorio para el parámetro β_1. De esta forma, en el conglomerado g el modelo para y_{ij} es

$$y_{ijg} = \frac{\beta_{1g} + b_{ig}}{1 + \beta_{2g} \exp\{-\beta_{3g} t_{ij}\}} + e_{ijg} \tag{100}$$

Para los 3 modelos mostrados en Tabla 1 se asume que $b_{ig} \sim MVN(0, D_g)$ y $e_{ijg} \sim MVN(0, R_{ig})$, con $R_{ig} = \sigma_g^2 I_{n_i}$. Dado que los modelos lineal por partes, lineal con $log(t)$ y logístico tendrán 3, 2 y 1 efectos aleatorios, respectivamente, la matriz D_g será de dimensión 3×3, 2×2 y 1×1, respectivamente.

Como el modelo logístico planteados en (100) es no lineal, es necesario obtener una versión linealizada de la forma $\tilde{y}_{ig} = \tilde{X}_i \beta_g + \tilde{Z}_i b_{ig} + \epsilon_{ig}$, donde la pseudo variable respuesta es $\tilde{y}_{ijkg} = \tilde{x}_{ijk} \beta_g + \tilde{z}_{ijk} b_{ig} + \epsilon_{ijkg}$. La matriz de diseño $\tilde{X}_i$ de dimensión 1×3 está dada por

$$\tilde{X}_i' = \begin{pmatrix} w_{ijg} \\ -f_1(\beta_g, t_{ij}) exp\{\{-\beta_{3g} t_{ij}\}\}/w_{ijg} \\ f_1(\beta_g, t_{ij}) \beta_{2g} t_{ij} exp\{\{-\beta_{3g} t_{ij}\}\}/w_{ijg} \end{pmatrix}$$

donde $w_{ijg} = 1/(1 + \beta_{2g} exp(-\beta_{3g} t_{ij}))$. La matriz de diseño $\tilde{Z}_i$ de dimensión 1×1 está dada por

$$\tilde{Z}_i = \begin{pmatrix} w_{ijg} \end{pmatrix}$$

Dada la estructura de datos completos de y_i, la estimación de los parámetros de los modelos se hará usando la metodología descrita para el modelo mixto lineal y no lineal en la sección 2.2.

Table 1: Modelos Ajustados

Modelo	Forma
Lineal por partes	$(\beta_1 + b_{i1}) + (\beta_2 + b_{i2})t_{ij} + (\beta_3 + b_{i3})(t_{ij} - 30)I(t_{ij} > 30) + e_{ij}$
Lineal transformación log	$y_{ij} = (\beta_1 + b_{i1}) + (\beta_2 + b_{i2}) \ln t_{ij} + e_{ij}$
Modelo Logístico	$y_{ij} = \frac{\beta_1 + b_{i1}}{1 + \beta_2 \exp\{-\beta_3 t_{ij}\}} + e_{ij}$

Como se explicó antes, los dos modelos lineales fueron ajustados incluyendo efectos aleatorios para todos los parámetros, mientras que para el ajuste del modelo logístico se incluyó un efecto aleatorio para el parámetro β_1.

3.1.2 Resultados

De los tres modelos mixtos propuestos para identificar los componentes de la mezcla (dos lineales y uno no lineal), el que presentó el mejor ajuste fue el modelo mixto no lineal, es decir, la curva de respuesta logística. Por otra parte, la identificación del número de componentes en la mezcla (número de conglomerados) indicó que el mejor modelo es el que identifica 2 conglomerados. La Tabla 2 muestra la log-verosimilitud marginal de y para un número creciente de conglomerados (hasta 4) usando el modelo logístico.

Table 2: Log-verosimilitud

Número de clusters	Log-verosimilitud	Número de parámetros	AIC
1	47.5	5	-85.0
2	113.7	10	-207.4
3	115.4	15	-200.8
4	116.1	20	-192.2

Usando el test de razón de verosimilitud para comparar los modelos en la Tabla 2, se observan diferencias significativas entre el modelo que identifica sólo 1 conglomerado y el que identifica 2 conglomerados. No se observa una ganancia significativa en la

log-verosimilitud del modelo 3 respecto al modelo 2 ni del modelo 4 respecto al 3. Por lo tanto, se eligió el modelo que identifica 2 componentes en la mezcla.

Por otra parte, si se usa el Criterio de Información de Akaike (AIC), se observa que también se elige el modelo que identifica 2 componentes en la mezcla (es el que presenta el menor valor de AIC). El mismo resultado se obtiene cuando se utiliza el Criterio de Información Bayesiana (BIC).

Los parámetros estimados de los tres modelos (para identificar 2 conglomerados) se muestran en Tabla 3 y Tabla 4. En cada columna de la Tabla 3, los parámetros $\beta_1^{'} = (\beta_{11}, \beta_{12}, \beta_{13})$ y $\beta_2^{'} = (\beta_{21}, \beta_{22}, \beta_{23})$ corresponden a los parámetros estimados de los efectos fijos de los modelos para los 2 conglomerados identificados por el algoritmo.

Table 3: Estimación de Efectos Fijos

Pará-metro	Lineal por partes	Lineal con $\log t$	Logístico
β_{11}	1.3640	-0.4801	3.4242
β_{12}	0.0318	0.9413	2.2944
β_{13}	-0.1246		-0.1037
β_{21}	2.4345	-1.8192	4.6503
β_{22}	0.0472	1.6881	9.1918
β_{23}	0.00139		-0.1450

En Tabla 4 se muestran las varianzas y componentes de la varianza estimados para los 3 modelos. Los parámetros σ_1^2 y σ_2^2 corresponden a la varianza del error (varianza intra-sujeto) en cada conglomerado y D_1 y D_2 corresponden a la varianza de los parámetros aleatorios del modelo (varianza entre-sujetos) al interior de cada conglomerado. La figura 2 muestra el ajuste de la curva de respuesta logística, el cual fue el mejor modelo obtenido mediante el procedimiento iterativo.

Table 4: Estimación de Varianzas y Componentes de la varianza

Pará-metro	Lineal por partes	Lineal con $\log t$	Logístico
σ_1^2	0.0550	0.5181	0.7216
σ_2^2	0.2014	0.1609	0.1194
D_1	$\begin{bmatrix} 0.549 & -0.017 & -0.043 \\ -0.017 & 0.00054 & 0.00135 \\ -0.043 & 0.00135 & 0.0034 \end{bmatrix}$	$\begin{bmatrix} 1.701 & -0.626 \\ -0.626 & 0.236 \end{bmatrix}$	0.2499
D_2	$\begin{bmatrix} 0.688 & -0.019 & 0.00021 \\ -0.019 & 0.00056 & -0.000014 \\ 0.00021 & -0.000014 & 0.000027 \end{bmatrix}$	$\begin{bmatrix} 2.337 & -0.701 \\ -0.701 & 0.210 \end{bmatrix}$	0.0317

Las probabilidades estimadas $\hat{\pi}_{g|\tilde{y}_i}$ de clasificación a posteriori en el conglomerado g fueron utilizadas para determinar la capacidad predictiva de los 3 modelos propuestos para identificar el verdadero estado de las mujeres al final del embarazo (parto normal y anormal). Usando $\hat{\pi}_{1|\tilde{y}_i},\ \ i = 1, \ldots, n$ (las probabilidades a posteriori estimadas para el conglomerado $g = 1$), se calculó el área bajo la curva ROC para los 3 modelos usando como gold standard el verdadero resultado final del parto.

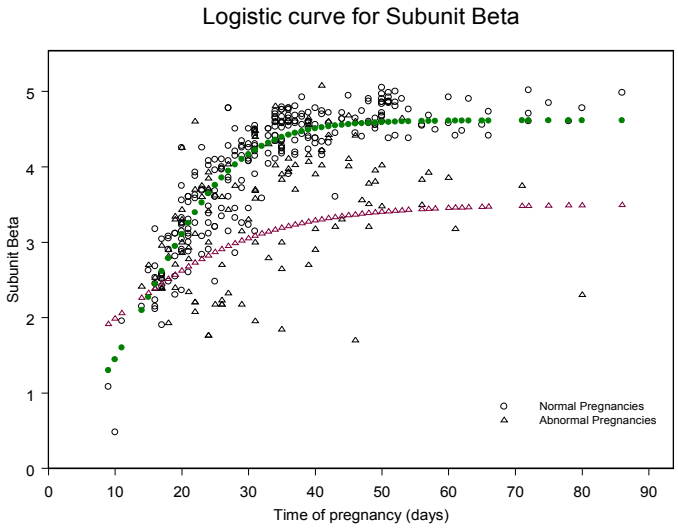

Figure 2: Ajuste de Modelo Logístico

Para los modelos lineal por partes, lineal con $log(t)$ y logístico, el área bajo la curva ROC fue 0.689, 0.837 y 0.849, respectivamente. La Tabla 5 resume éste y otros resultados obtenidos del análisis de sensibilidad-especificidad de los 3 modelos.

Table 5: Análisis de Sensibilidad-Especificidad

Indicador	Lineal por partes	Lineal con $\log t$	Logístico
AUC	0.689±0.057	0.837±0.043	0.849±0.042
Punto de corte para $\hat{\pi}_{1\|\tilde{y}_i}$	0.980	0.963	0.931
Sensibilidad (S)	16/37 (43.2%)	24/37 (64.9%)	27/37 (73.0%)
Especificidad (E)	115/124 (92.7%)	115/124 (92.7%)	112/124 (90.3%)
Exactitud (S+E)	131/161 (81.4%)	139/161 (86.3%)	139/161 (86.3%)

Se observa en la Tabla 5 que el mejor modelo es la curva de crecimiento logístico (mayor área bajo la curva ROC). Sin embargo, al comparar las áreas usando el procedimiento de Hanley y McNeil [20, 21] no se encontró una diferencia significativa entre los modelos lineal con $log(t)$ y la curva de respuesta logística, aunque ambos son signi-

ficativamente mejores que el modelo lineal por partes. La figura 3 muestra las curvas ROC de los 3 modelos ajustados.

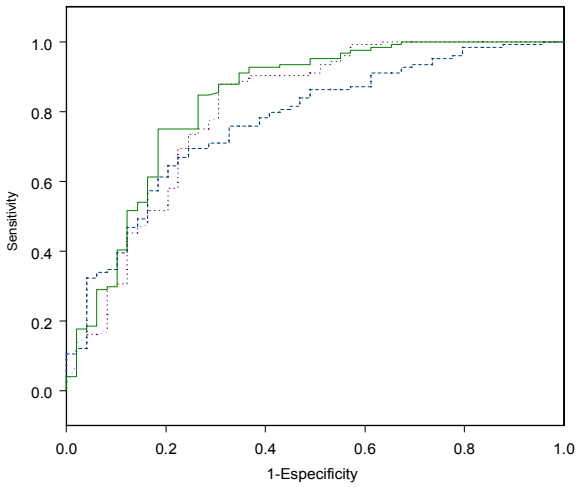

Figure 3: Curvas ROC para los tres modelos ajustados

Para evaluar el comportamiento del mejor modelo ajustado (curva de crecimiento logístico), se usó validación cruzada. Una descripción general de este método puede verse en Friedman, Hastie y Tibshirani [18].

Aunque inicialmente se usó ten-fold para estimar el error de predicción, el relativamente bajo número de observaciones disponibles (161 casos) derivó en una sobrestimación del error, al ajustar modelos con sólo el 90% de los datos. Por este motivo se optó por el método leave-one-out, de modo de maximizar el número de observaciones usados para ajustar el modelo. La Tabla 6 resume los resultados de la validación cruzada.

Table 6: Validación Cruzada del Modelo Logístico

Indicador	Resultado
AUC	0.830±0.039
Punto de corte para $\hat{\pi}_{1\|\tilde{y}_i}$	0.927
Sensibilidad (S)	29/37 (78.4%)
Especificidad (E)	108/124 (87.1%)
Exactitud (S+E)	137/161 (85.1%)
Valor Predictivo (+) con $\hat{\pi}_{1\|\tilde{y}_i} > 0.5$	16/19 (84.2%)
Valor Predictivo (-) con $\hat{\pi}_{1\|\tilde{y}_i} > 0.5$	121/142 (85.2%)

Se observa en Tabla 6 que el modelo logístico tiene un ajuste muy satisfactorio

($AUC = 0.830 \pm 0.039$) y confirma en gran medida los resultados del ajuste mostrado en la última columna de la Tabla 5.

Si se usara $\hat{\pi}_{1|\tilde{y}_i} > 0.5$ como punto de corte en la probabilidad a posteriori para clasificar el estado de las mujeres al final del embarazo (como parto normal y anormal), el valor predictivo positivo sería 84.2% y el valor predictivo negativo sería 85.2% .

3.2 Ejemplo 2. Modelo longitudinal mixto multivariado

3.2.1 Modelos y datos

En este segundo ejemplo se analiza el efecto de incluir la variable estradiol, medida longitudinalmente, como respuesta adicional al modelo mixto para β-HCG, descrito en (3.1). De esta forma, el modelo es multivariado con $p = 2$ variables respuesta. El mecanismo de cluster será utilizado para detectar un número fijo de $m = 2$ conglomerados.

Para las 161 mujeres en estudio se hicieron un total de 348 mediciones de β-HCG y estradiol, con un promedio de 2.2 mediciones por mujer. La tasa de valores ausentes fue 2% en β-HCG (7 missing values) y 33.6% en estradiol (117 missing values). Esta estructura de datos incompletos será considerada en el modelo multivariado a ajustar.

El comportamiento de β-HCG es el mostrado en la Figura 1 del Ejemplo 1. Para esta variable respuesta se ajustará el modelo de efectos mixtos no lineal de curva de respuesta logística, que fue el que mostró el mejor ajuste univariado.

La Figura 4 muestra el comportamiento de estradiol (transformación *log*) en relación a los días de gestación. En este caso, se asumirá que la relación es lineal, aunque claramente es más difícil identificar la forma de la relación entre estradiol y días de gestación.

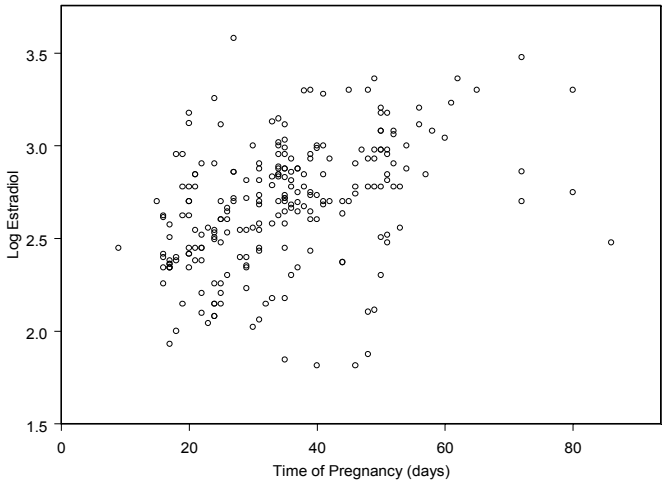

Figure 4: Log estradiol durante los primeros 80 días de gestación

La información individual disponible es como se muestra en la matriz siguiente para la mujer número 56 de la base de datos, la cual tiene un máximo de $n_i = 3$ observaciones para las $p = 2$ variables en estudio

$$Y_{i=56} = \begin{bmatrix} y_{i11} & y_{i12} \\ y_{i21} & y_{i22} \\ y_{i31} & y_{i32} \end{bmatrix} = \begin{bmatrix} NA & 2.15 \\ 2.90 & NA \\ 4.19 & 2.34 \end{bmatrix}$$

El uso del operador $y_i = vec(Y_i)$ generará un vector de largo pn_i. En el caso de la observación número 56, el vector es $y_{56} = (NA, 2.90, 4.19, 2.15, NA, 2.34)^t$.

Definiendo a log β-HCG como la respuesta $k = 1$ y log estradiol la respuesta $k = 2$, los modelos para la observación $y_{ijk}|C_i = g$, con 1 efecto aleatorio en cada modelo, son los siguientes

$$y_{ij1g} = \frac{\beta_{1g} + b_{i1g}}{1 + \beta_{2g} \exp\{-\beta_{3g} t_{ij}\}} + e_{ij1g} \tag{101}$$

y

$$y_{ij2g} = \beta_{4g} + \beta_{5g} t_{ij} + b_{i2g} + e_{ij2g} \tag{102}$$

Se asume que $b_{ikg} \sim MVN(0, D_g)$ y $e_{ikg} \sim MVN(0, R_{ig})$, con $R_{ig} = \Sigma_g \otimes I_{n_i}$. Los modelos planteados en (101) y (102) son de la forma $y_{ijkg} = f_g(\eta_{ig}, x_{ij}) + \epsilon_{ijkg}$, según se muestra en (54). Para obtener una versión linealizada de la forma $\tilde{y}_{ig} = \tilde{X}_i \beta_g + \tilde{Z}_i b_{ig} + \epsilon_{ig}$, la nueva variable respuesta es $\tilde{y}_{ijkg} = \tilde{x}_{ijk} \beta_g + \tilde{z}_{ijk} b_{ig} + \epsilon_{ijkg}$, donde $\beta'_g = (\beta_{1g}, \beta_{2g}, \beta_{3g}, \beta_{4g}, \beta_{5g})$ y $b'_{ig} = (b_{i1g}, b_{i2g})$.

La matriz de diseño $\tilde{X}_i$ de dimensión 2×5 ($p = 2$ variables respuestas $\times$ $q = 5$ parámetros en el vector β_g), está dada por

$$\tilde{X}'_i = \begin{pmatrix} w_{ijg} & 0 \\ -f_1(\beta_g, t_{ij}) exp\{\{-\beta_{3g} t_{ij}\}\}/w_{ijg} & 0 \\ f_1(\beta_g, t_{ij}) \beta_{2g} t_{ij} exp\{\{-\beta_{3g} t_{ij}\}\}/w_{ijg} & 0 \\ 0 & 1 \\ 0 & t_{ij} \end{pmatrix}$$

donde $w_{ijg} = 1/(1 + \beta_{2g} exp(-\beta_{3g} t_{ij}))$. La matriz de diseño $\tilde{Z}_i$ de dimensión 2×2 ($p = 2$ variables respuestas $\times$ $r = 2$ parámetros en el vector $\tilde{b}_{ig}$), esta dada por

$$\tilde{Z}_i = \begin{pmatrix} w_{ijg} & 0 \\ 0 & 1 \end{pmatrix}$$

Nótese que, como el modelo elegido para log estradiol es lineal (response $k = 2$), se cumple que $\tilde{X}_i = \beta_4 + \beta_5 t_i$ y $\tilde{Z}_i = b_{i2}$, y por lo tanto $\tilde{y}_{i2} = y_{i2}$.

Dada la estructura de datos incompletos de y_i, es necesario usar la matriz O_i, generada a partir de la matriz identidad de dimensión $pn_i \times pn_i$ asociada al vector y_i, a la cual se eliminan las filas correspondientes a las observaciones ausentes. Luego, el modelo para la i-ésima observación en el conglomerado g es de la forma

$$\tilde{O}_i y_{ig} = O_i \tilde{X}_i \beta_g + O_i \tilde{Z}_i b_{ig} + O_i \epsilon_{ig} \tag{103}$$

Por lo tanto, la estimación de los parámetros del modelo se hará usando la metodología descrita en la sección 2.2.2.

3.2.2 Resultados

La tabla 7 muestra los parámetros β_g estimados para los conglomerados $g = 1$ y $g = 2$, obtenidos después de 32 iteraciones del algoritmo escrito en S-Plus

Table 7: Parámetros Estimados

Parámetro	Cluster $g = 1$	Cluster $g = 2$
β_1	4.7397	4.1830
β_2	10.2381	5.9878
β_3	-0.1493	-0.1280
β_4	2.2382	2.3337
β_5	0.0148	0.0063

Las matrices de covarianza estimadas $\hat{\Sigma}_g$ para los conglomerados $g = 1$ y $g = 2$, que estiman la variabilidad intra-sujeto en cada conglomerado, son

$$\hat{\Sigma}_1 = \begin{pmatrix} 0.0227 & 0.0018 \\ 0.0018 & 0.0212 \end{pmatrix}, \hat{\Sigma}_2 = \begin{pmatrix} 0.2814 & 0.0439 \\ 0.0439 & 0.0446 \end{pmatrix}$$

Las matrices de covarianza $\hat{D}_g$ para los conglomerados $g = 1$ y $g = 2$, que estiman la variabilidad entre-sujetos en cada conglomerado, son

$$\hat{D}_1 = \begin{pmatrix} 0.0352 & 0.00018 \\ 0.00018 & 0.0265 \end{pmatrix}, \hat{D}_2 = \begin{pmatrix} 0.3624 & 0.1203 \\ 0.1203 & 0.0982 \end{pmatrix}$$

La figura 5 muestra el modelo ajustado para log β-HCG en conglomerados $g = 1$ y $g = 2$.

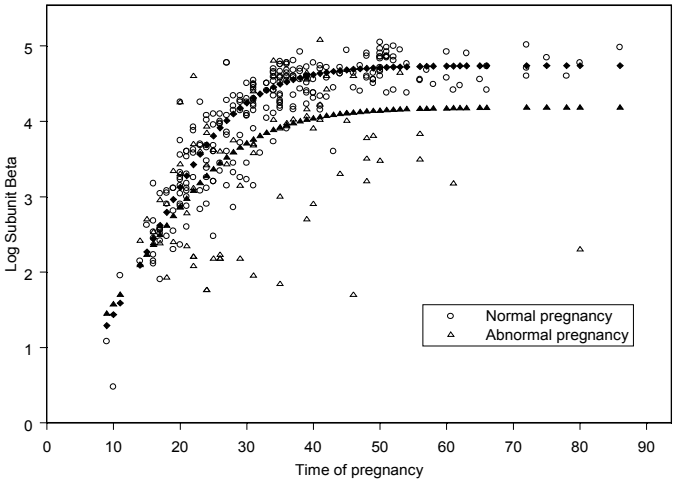

Figure 5: Modelo Logístico ajustado para Log β-HCG

La figura 6 muestra el modelo ajustado para log estradiol en conglomerados $g = 1$ y $g = 2$.

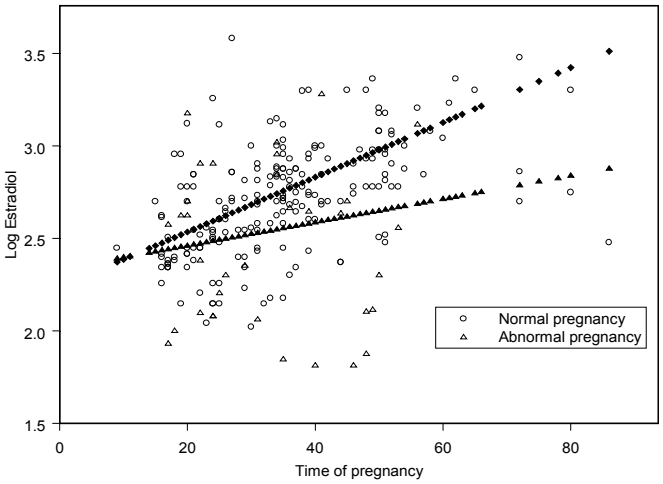

Figure 6: Modelo Lineal ajustado para Log Estradiol

Como en el ejemplo univariado, las probabilidades estimadas $\hat{\pi}_{g|\tilde{y}_i}$ de clasificación a posteriori en el conglomerado g fueron utilizadas para determinar la capacidad pre-

dictiva del método propuesto para identificar mujeres con parto normal y anormal. Usando $\hat{\pi}_{1|\tilde{y}_i},\ \ i = 1, \ldots, n$ (las probabilidades a posteriori estimadas para el conglomerado $g = 1$), se calculó el área bajo la curva ROC usando como gold standard el verdadero resultado final del parto. El área bajo la curva fue igual a 0.816 con un error estandard 0.042. La figura 7 muestra la curva ROC resultante. La tabla 8 muestra los resultados más relevantes del análisis de sensibilidad y especificidad.

Table 8: Análisis de Sensibilidad-Especificidad

Indicador	Resultado
AUC	0.816 ± 0.042
Punto de corte para $\hat{\pi}_{1\|\tilde{y}_i}$	0.65
Sensibilidad (S)	31/37 (83.8%)
Especificidad (E)	86/124 (69.4%)
Exactitud (S+E)	117/161 (72.7%)

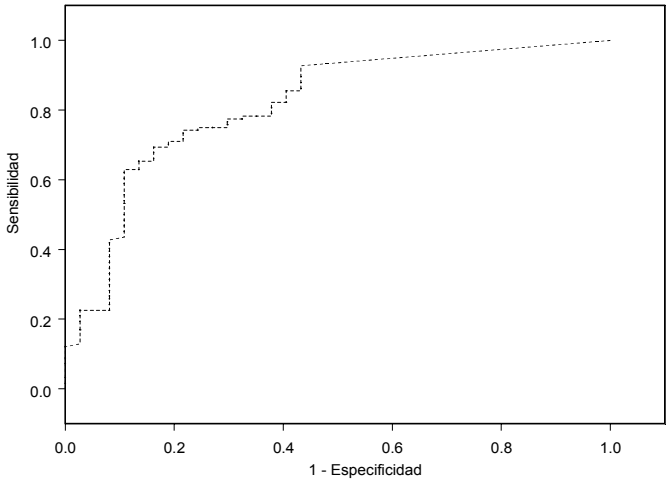

Figure 7: Curva ROC para el modelo multivariado

3.3 Ejemplo 3. Modelo mixto con variables explicadas longitudinal y fija

3.3.1 Modelos y datos

Para ilustrar la metodología descrita en la sección (2.3), consideremos una muestra de 240 individuos portadores del virus de inmunodeficiencia humana (VIH) que acceden por primera vez a un tratamiento antiretroviral (TAR). Para estos sujetos se tiene el nivel de carga viral (CV) antes de iniciar el tratamiento y el nivel de linfocitos CD4, que da cuenta del estado de las defensas inmunológicas de un individuo, medido longitudinalmente desde el inicio del tratamiento hasta el último control o muerte del paciente.

Estos 240 casos corresponden a una submuestra tomada del Estudio de Evaluación del Impacto de las TAR en Pacientes que viven con VIH/SIDA beneficiarias del Sistema Público de Salud de Chile [7], e incluye a los 60 pacientes fallecidos en la cohorte en estudio y una submuestra aleatoria de 180 pacientes vivos hasta la fecha.

En este ejemplo se muestra el uso de un modelo de efectos mixtos que incluye una variable explicada longitudinal (linfocitos CD4) y una variable explicada fija (carga viral antes de TAR). El mecanismo de cluster será utilizado para detectar un número fijo de $m = 2$ conglomerados, y observar el grado de concordancia de la clasificación con el estado final de los pacientes (vivo o fallecido).

La Figura 9 muestra la evolución de los linfocitos CD4 (transformación log)para algunos pacientes vivos y fallecidos incluidos en el ejemplo. La carga viral basal (transformación $log10$) tuvo un promedio $\pm$ DS igual a 5.0 ± 0.7 en el grupo de pacientes vivos y 5.2 ± 0.8 en el grupo de los fallecidos. No se observan diferencias significativas entre los promedios de los logaritmos de CV (p=0.16).

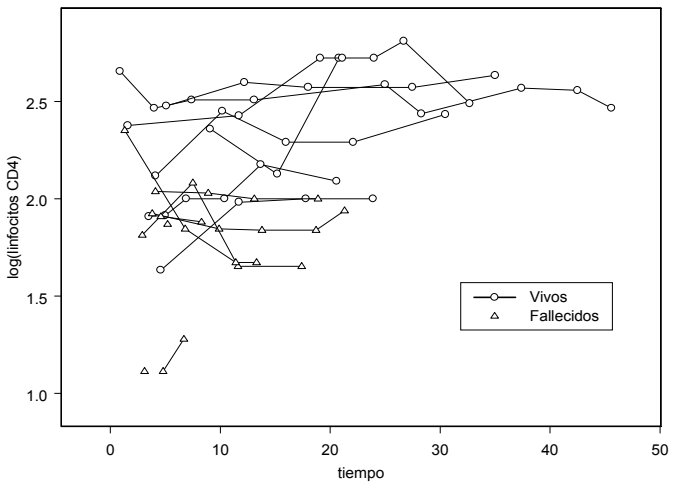

Figure 8: Linfocitos CD4 en individuos VIH(+) vivos y fallecidos

Si el vector de variables respuestas para el i-ésimo sujeto es $y_i^* = (y_i, r_i)^T$, donde y_i es un vector $n_i \times 1$ que almacena los valores de linfocitos CD4 y r_i es la carga viral antes de TAR, el modelo a ajustar será

$$y_i^* = \begin{bmatrix} y_i \\ r_i \end{bmatrix} = \begin{bmatrix} X_i & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} + \begin{bmatrix} Z_i & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} b_i \\ b_i' \end{bmatrix} + \begin{bmatrix} \epsilon_i \\ 0 \end{bmatrix}$$

donde el nivel de linfocitos en tiempo j para el i-ésimo sujeto sigue un modelo lineal de efectos mixtos de la forma $y_{ij} = (\alpha_1 + b_{i1}) + (\alpha_2 + b_{i2})t_{ij} + \epsilon_{ij}$. Los supuestos del modelo son los descritos en la sección (2.3).

Adicionalmente, se ajustará un modelo sólo para los linfocitos en y_i usando la metodología descrita en (2.1), para observar el efecto sobre la clasificación de incluir la variable explicada fija r_i.

3.3.2 Resultados

La tabla 9 muestra los parámetros estimados para los conglomerados $g = 1$ y $g = 2$ obtenidos al ajustar el modelo sin incluir la variable explicada fija. Este modelo se obtuvo con 71 iteraciones del algoritmo escrito para este efecto en S-Plus

Table 9: Parámetros Estimados sin incluir variable fija en el modelo

Parámetro	Cluster $g = 1$	Cluster $g = 2$
α_1	5.0973	3.9941
α_2	0.1522	0.2064
σ^2	0.1501	0.9874
D	$\begin{bmatrix} 0.084 & -0.032 \\ -0.032 & 0.012 \end{bmatrix}$	$\begin{bmatrix} 0.291 & 0.026 \\ 0.026 & 0.046 \end{bmatrix}$

La tabla 10 muestra los parámetros estimados obtenidos al incluir la variable fija en el modelo. El algoritmo de estimación converge con 60 iteraciones.

Table 10: Parámetros Estimados incluyendo variable fija en el modelo

Parámetro	Cluster $g = 1$	Cluster $g = 2$
α_1	4.8791	3.4308
α_2	0.1833	0.1190
β	5.0007	5.3075
σ^2	0.1388	0.3021
D	$\begin{bmatrix} 0.323 & -0.034 & -0.059 \\ -0.034 & 0.011 & 0.004 \\ -0.059 & 0.004 & 0.510 \end{bmatrix}$	$\begin{bmatrix} 1.641 & -0.939 & -0.090 \\ -0.939 & 0.687 & 0.257 \\ -0.090 & 0.257 & 0.371 \end{bmatrix}$

Al utilizar las probabilidades estimadas $\hat{\pi}_{g|\tilde{y}_i}$ de clasificación a posteriori en el conglomerado g para determinar la capacidad predictiva de los modelos para identificar a los pacientes vivos y fallecidos, se obtienen los resultados mostrados en tabla 11.

Table 11: Análisis de Sensibilidad-Especificidad de los modelos

Indicador	Modelo sin variable fija	Modelo con variable fija
AUC	0.779 ± 0.032	0.821 ± 0.032
Punto de corte para $\hat{\pi}_{1|\tilde{y}_i}$	0.88	0.99
Sensibilidad (S)	52/60 (86.7%)	49/60 (81.7%)
Especificidad (E)	105/180 (58.3%)	126/180 (70.0%)
Exactitud (S+E)	157/240 (65.4%)	175/240 (72.9%)

Se observa una mejor clasificación de los pacientes vivos y fallecidos al usar el modelo que incluye la variable fija. Sin embargo, no hay diferencias significativas entre ambos ajustes (p=0.114 usando procedimiento de Hanley y McNeil [20, 21]). El gráfico siguiente muestra la curva ROC para los modelos ajustados.

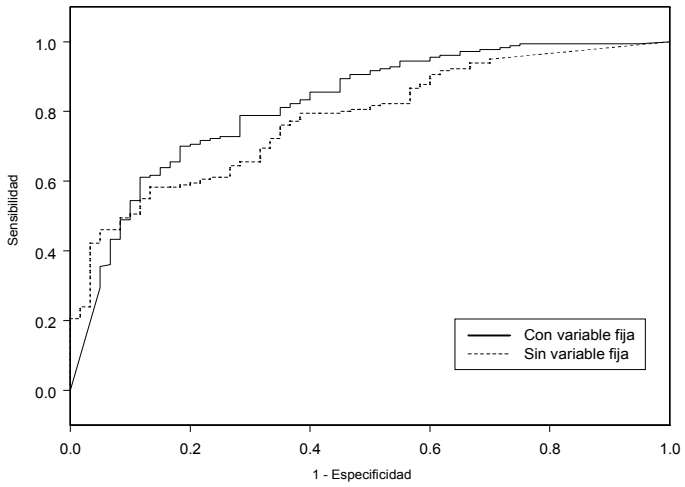

Figure 9: Curva ROC para modelos con y sin variable explicada fija

3.4 Detalles Computacionales

Los algoritmos usados en los ejemplos de (3.1), (3.2) y (3.3) para ajustar los modelos mixtos fueron implementados como macros en S-Plus versión 6.

Dado que los parámetros debían ser estimados al interior de cada conglomerado $g = 1, 2$, tanto en el caso univariado como en el multivariado, fue necesario obtener parámetros iniciales para β_g, Σ_g y D_g en forma separada para cada conglomerado, para lo cual se usó la función NLME de Pinheiro y Bates [39].

Para facilitar la convergencia del algoritmo EM, usado para estimar Σ_g y D_g, se consideró que una mujer tenía probabilidad a priori 0.7 de tener un buen resultado en su embarazo (probabilidad 0.3 de presentar un embarazo interrumpido). Estas probabilidades corresponden a las probabilidades iniciales *apriori* de pertenencia a cada conglomerado. Una alternativa es usar probabilidades de clasificación a priori 0.5 para cada resultado, lo cual es recomendable cuando se desconoce la proporción de casos que se obtendrá en cada conglomerado. Además, para tener una diferenciación inicial entre los conglomerados, se asumió que el grupo con menor probabilidad de clasificación a priori tendría una mayor variabilidad (lo cual es esperable por ser el grupo de menor tamaño), por lo que se decidió usar valores iniciales $\Sigma_g^{(0)}$ y $D_g^{(0)}$ mayores para el conglomerado $g = 1$.

El código fuente en S-Plus para el ajuste del modelo mixto no lineal multivariado puede verse en el Apéndice.

4 Discusión

La metodología propuesta en esta tesis mezcla dos procedimientos estadísticos de amplio uso en el ámbito de las ciencias biológicas: modelos de efectos mixtos, usados para describir el comportamiento de dos o más variables medidas longitudinalmente o en combinación con variables fijas, y análisis de conglomerados, usado para identificar grupos de individuos similares respecto su comportamiento longitudinal.

Desde el punto de vista del análisis de conglomerados, la metodología descrita en esta tesis es un mecanismo general de identificación de componentes subyacentes en un conjunto de datos longitudinales. Es una metodología general porque otros métodos resultan ser casos particulares de éste, como el propuesto por McLachlan y Gordon [37] (modelos lineal de efectos mixtos con conocimiento parcial de la clasificación de los individuos), Banfield y Raftery [3] (de identificación de mezclas normales y no normales), Verbeke y Lesaffre [55] (mezclas de los efectos aleatorios de modelo mixto lineal) o Spiessens et al [51] (mezclas de los efectos aleatorios de modelo mixto lineal o no lineal).

Por otra parte, si los valores de las variables medidas longitudinalmente corresponden a los valores tempranos de un proceso en el tiempo, que culminará finalmente en una nueva variable respuesta aun desconocida (como el tipo de parto normal o anormal de los ejemplos 1 y 2 o la muerte por SIDA del ejemplo 3), entonces la clasificación de los individuos en conglomerados, de acuerdo a su comportamiento longitudinal, puede ser visto como un método de screening, que permite identificar precozmente a aquellos individuos que presentarán un mal pronóstico respecto a su respuesta a largo plazo.

A continuación se discuten otros aspectos relevantes sobre la metodología descrita y una propuesta para futuros desarrollos en la línea del análisis de conglomerados usando modelos de efectos mixtos.

4.1 Limitaciones del Algoritmo EM

Un problema conocido del algoritmo EM, y observado también en esta tesis, es la baja tasa de convergencia del algoritmo. Sin embargo, esto no debiera ser un problema cuando las distribuciones subyacentes en los datos están bien separadas o definidas, con valores iniciales razonables para la obtención de parámetros estimados.

En el problema particular de identificación de parámetros en modelos de efectos mixtos en análisis de conglomerados, otro elemento que mejora la tasa de convergencia es la selección de las variables que serán utilizadas en la identificación de las distribuciones.

Una limitación del Algoritmo EM no relacionada con la tasa de convergencia se relaciona con el método de clustering elegido (mixture decomposition scheme), ya que el número de probabilidades condicionales asociadas con cada observación es igual al número de componentes en la mezcla. Por lo tanto, el algoritmo EM para clustering podría ser poco práctico para modelos con un alto número de componentes.

4.2 Selección de parámetros iniciales

Un aspecto crítico del análisis de conglomerados usando modelos de efectos mixtos es la elección de los parámetros iniciales de los modelos. En esta tesis se consideraron dos posibles soluciones para este problema:

Método 1 Ajustar un modelo de efectos mixtos para el total de datos disponibles y usar los parámetros estimados de este modelo como parámetros iniciales para los conglomerados a ser construidos. En este caso, es necesario inicializar las varianzas σ_1^2, σ_2^2, D_1 and D_2 con valores diferentes para cada conglomerado, o de lo contrario no será posible obtener una diferenciación entre los modelos.

Método 2 Ajustar un modelo de efectos mixtos con un efecto aleatorio para el total de datos disponibles y hacer un análisis de conglomerados con los efectos aleatorios estimados por el modelo mixto de modo que dos (o más) conglomerados sean identificados. Luego, ajustar un modelo de efectos mixtos con los datos identificados en cada conglomerado por separado. Finalmente, usar los parámetros estimados de estos modelos como valores iniciales. En este caso, se tendrán valores iniciales para cada $beta_i$, σ_1^2, σ_2^2, D_1 and D_2.

Los dos métodos descritos para obtener parámetros iniciales mostraron resultados similares en nuestras pruebas, tanto para modelos lineales como no lineales (de hecho, el primer método fue usado en los ejemplos de esta tesis). Sin embargo, ya que el procedimiento iterativo es muy sensible a los valores iniciales escogidos, el segundo método es preferible, ya que éste no requiere especular acerca de los valores iniciales para las varianzas y componentes de la varianza para cada modelo.

Respecto a las probabilidades de clasificación a priori $\pi_k,\ k = 1, \ldots, g$, éstas pueden inicializarse en $1/g$ si se desconocen las probabilidades de clasificación del resultado final (o cuando no se pretende hacer coincidir los conglomerados con algún resultado final). Por el contrario, si se quiere determinar la concordancia de los conglomerados con algún outcome predefinido, entonces se pueden inicializar los π_k como las probabilidades poblacionales en que se presentan estos hallazgos.

4.3 Determinación de la incertidumbre de la clasificación

A diferencia de los ejemplos usados en la sección 3, en problemas reales de análisis de conglomerados no se conoce el resultado final de los individuos a clasificar, como en una clasificación supervisada. Por lo tanto, no se tiene un "gold standard" contra el cual contrastar la clasificación hecha por el algoritmo de cluster.

El único método encontrado en la literatura que permite estimar el grado de incertidumbre de la clasificación es el descrito por Bensmail et al en 1997 [8], que consiste en calcular

$$U_i = min_{k=1,\ldots,g}\{1 - \tilde{\pi}_{k|y_i}\}, \quad i = 1, \ldots, n \tag{104}$$

donde $\tilde{\pi}_{k|y_i}$ es la probabilidad estimada de clasificación a posteriori en el conglomerado k. Se observa que cuando es clara la pertenencia de y_i al conglomerado k, el valor de $1 - \tilde{\pi}_{k|y_i}$ es pequeño y por tanto el valor de U_i es pequeño. Fraley y Raftery [15] muestran la distribución de U_i en cuartiles para determinar la calidad de la clasificación.

Aunque no es posible establecer la incertidumbre de la clasificación sin contar con un gold standard, el método de Bensmail indica que si hay muchos valores de $\tilde{\pi}_{k|y_i}$ cercanos a $1/g$ podría ser indicador de incertidumbre en la clasificación. Por ejemplo, para $g = 2$ conglomerados, un individuo con probabilidad a posteriori $\tilde{\pi}_{k|y_i} \approx 0.5$ indicaría incertidumbre de pertenencia del individuo i a cualquiera de los 2 conglomerados.

Al aplicar el método de Bensmail a las probabilidades $\tilde{\pi}_{k|y_i}$ del modelo mixto con curva de crecimiento logístico del ejemplo univariado en sección 3.1, se obtiene que los percentiles 75%, 90% y 95% de los U_i son 0.056, 0.1254 y 0.2341, respectivamente, resultados coherentes con el alto poder predictivo del modelo como predictor de embarazo interrumpido (exactitud 86.3%). Por otra parte, la aplicación del método al modelo multivariado en sección 3.2 arroja percentiles 75%, 90% y 95% de U_i iguales a 0.1675, 0.3388 y 0.4196, resultado también concordante con la menor exactitud en la clasificación de embarazo interrumpido de este ejemplo (72.7%).

4.4 Determinación del número de conglomerados

La mayoría de los métodos descritos en la literatura para determinar el número de conglomerados no son aplicables al análisis de conglomerados usando modelos mixtos. Por ejemplo, el método de Sugar y James [52] se basa en el uso de K-means para detectar distintos números de conglomerados K, se calcula la distancia d_k de los vectores de datos a cada centroide (Mahalanobis) y se elige el número de conglomerados que maximiza la distancia $d_k - d_{k-1}$. Alternativamente, existen varios métodos basados en la matriz combinada W de covarianzas intra-grupo de las p variables usadas en la clasificación, para cualquier partición de la muestra. Si en una división en k conglomerados se utiliza, por ejemplo, la traza S_k o el determinante D_k de la matriz W, valores pequeños de S_k o D_k serían indicadores de una buena partición de los datos. Por otra parte, una pequeña diferencia entre S_k y S_{k+1}, la traza de W en un esquema con k conglomerados y otro con $k+1$ conglomerados, serí a indicador de "convergencia" respecto al número de conglomerados. Una revisión de algunos de estos métodos puede verse en Krzanowski y Lai [30]. Ninguno de los métodos descritos es aplicable al esquema de conglomerados usado en esta tesis, ya que en nuestro caso se cuenta con información de varibles medidas longitudinalmente.

La identificación del número de conglomerados en esta tesis se hizo mediante test de razón de verosimilitud (TRV), Criterio de Información de Akaike (AIC) [1] y Criterio de Información Bayesiana (BIC) [47]. El uso de estos métodos ha sido recomendado en la literatura, principalmente el Criterio de Información Bayesiana, ya que permite comparar más de dos modelos simultáneamente y no tiene la restricción de que los modelos comparados sean anidados [16]. Finalmente, estos tres métodos pueden ser utilizados con información medida longitudinalmente.

4.5 Posibles desarrollos en el análisis de conglomerados mediante modelos de efectos mixtos

En este trabajo se describe una metodología de detección de conglomerados en el cual las variables explicadas en los modelos mixtos son real valoradas, con errores aleatorios y errores específicos por sujeto distribuidos normalmente.

Una posible extensión de la metodología es considerar variables explicadas que tomen solo valores positivos, como en el caso de conteo de eventos (por ejemplo, mortalidad en distintas zonas geográficas). En este caso, se podría asumir que los errores aleatorios y específicos por sujeto tienen distribución Poisson. Un ejemplo del uso de modelo de efectos mixtos con errores Poisson aplicado a la estimación de tasas de mortalidad puede verse en Tsutakawa [54].

Por otra parte, la independencia entre las observaciones y_i, asociado a una matriz de covarianzas intra-sujetos de la forma $\sigma^2 I_{n_i \times n_i}$, podría modificarse para considerar, por ejemplo, modelos autoregresivos de primer orden ($\sigma_{ij} = \sigma^2 \rho^{|i-j|}$) u otros modelos como los descritos por Jennrich y Schluchter [28].

Finalmente, la metodología de clustering utilizada (Mixture Decomposition Scheme) podría ser reemplazada por esquemas alternativos, como el Fuzzy Clustering Algorithm, en el cual cada observación y_i pertenece simultáneamente a todos los conglomerados y la estimación de parámetros se basa en la minimización de una función de costos [53].

En cualquier caso, el análisis de conglomerados usando modelos de efectos mixtos es un campo abierto en muchas direcciones y con múltiples posibilidades de desarrollo futuro.

A APENDICE. Rutina S-Plus para Ajuste de Modelo Multivariado

```
# Cluster Analysis usando Multivariate Mixed Models
# Parametros Iniciales con "nlme" de Pinheiro-Bates
# Version enero/05
#
library(matrix)
n <- as.vector(table(DS4$sujeto))
m <- length(n)

alfa1   <- c(4.5,8.55,-0.142,2.23,0.05)
alfa2   <- c(4.5,8.55,-0.142,2.23,0.05)
priori  <- c(0.7,0.3)
poster1 <- rep(0.7,m)
poster2 <- 1-poster1
sigma1  <- Matrix(c(0.1340, 0.0200, 0.0200, 0.3000),2,2)
sigma2  <- Matrix(c(0.3400, 0.0200, 0.0200, 0.6000),2,2)
D1      <- Matrix(c(0.2610, 0.0200, 0.0200, 0.0100),2,2)
D2      <- Matrix(c(0.3610, 0.0200, 0.0200, 0.0200),2,2)
b1i     <- rep(0.5,2*m)
b2i     <- rep(0.5,2*m)

#Construccion de X, Y, Z
#
matrices_function(DS4,n,m,beta,bi)
{ sujeto  <- DS4$sujeto
  tiempo  <- DS4$tiempo
  uno     <- DS4$uno
  x       <- Matrix(0,nrow=2*length(uno),ncol=5)
  z       <- Matrix(0,nrow=2*length(uno),ncol=2)
  y       <- rep(0,2*length(uno))
  a       <- rep(0,2*length(uno))
  subject <- rep(0,2*length(uno))
  mu      <- rep(0,2*length(uno))
  k       <- 1
  for (i in 1:m)
  {
   l1 _ k+n[i]-1;   l2 _ k+n[i];   l3 _ k+2*n[i]-1;   l4 _ 2*i-1;   l5 _ 2*i

   y[k:l1]        <- DS4$subunit[sujeto==i]
   y[l2:l3]       <- DS4$estrad[sujeto==i]
   subject[k:l1]  <- DS4$sujeto[sujeto==i]
   subject[l2:l3] <- DS4$sujeto[sujeto==i]
   tiemp          <- DS4$tiempo[sujeto==i]
   uno            <- DS4$uno[sujeto==i]
   den            <- uno + beta[2]*exp(beta[3]*tiemp)
   mu[k:l1]     <- (beta[1] + bi[l4])/den
   mu[l2:l3]    <- beta[4] + beta[5]*tiemp + bi[l5]
   x[k:l1,1]    <- 1/den
   x[k:l1,2]    <- ( - mu[k:l1]*exp(beta[3]*tiemp) )/den
   x[k:l1,3]    <- ( - mu[k:l1]*beta[2]*tiemp*exp(beta[3]*tiemp) )/den
   x[l2:l3,4]   <- 1
   x[l2:l3,5]   <- tiemp
   z[k:l1,1]    <- 1/den
   z[l2:l3,2]   <- 1
   a[k:l3]      <- y[k:l3] - mu[k:l3] + x[k:l3,]%*%beta + z[k:l3,]%*%as.Matrix(bi[l4:l5])
   k            <- k+2*n[i]
  }
  list(y=a,x=x,z=z,sujeto=subject)
}
```

```
# Construccion de matriz Oi para eliminacion de missing values
#
misout_function(yi)
{ k_length(yi)
  l_length(na.exclude(yi))
  idreal  <- diag(k)
  idmodif <- Matrix(0,nrow=l,ncol=k)
  linea_1
  for (j in 1:k)
  { if( !is.na(yi[j]) )
     {
      idmodif[linea,] _ idreal[j,]
      linea _ linea+1
     }
  }
  if (!is.Matrix(idmodif)) idmodif <- as.Matrix(idmodif)
  Oi _ idmodif
}

# Calculo de probabilidades a posteriori y verosimilitudes
#
posteriori_function(datos1,datos2,n,m,priori,poster1,sigma1,sigma2,alfa1,alfa2,D1,D2)
{ verotot1_0; verotot2_0; logvero_0
  for (i in 1:m)
  {  Oi_misout(datos1$y[datos1$sujeto==i])
     y1i <- na.exclude(datos1$y[datos1$sujeto==i])
     z1i <- Oi%*%datos1$z[datos1$sujeto==i,]
     x1i <- Oi%*%datos1$x[datos1$sujeto==i,]
     y2i <- na.exclude(datos2$y[datos2$sujeto==i])
     z2i <- Oi%*%datos2$z[datos2$sujeto==i,]
     x2i <- Oi%*%datos2$x[datos2$sujeto==i,]
     if (!is.Matrix(x1i))x1i<- t(as.Matrix(x1i));if (!is.Matrix(x2i))+x2i <- t(as.Matrix(x2i))
     if (!is.Matrix(z1i))z1i <-t(as.Matrix(z1i));if (!is.Matrix(z2i))+z2i <- t(as.Matrix(z2i))

     var1 <- Oi%*%kronecker(sigma1,diag(rep(1,n[i])))%*%t(Oi) + z1i%*%D1%*%t(z1i)
     var2 <- Oi%*%kronecker(sigma2,diag(rep(1,n[i])))%*%t(Oi) + z2i%*%D2%*%t(z2i)
     det1 <- exp(as.double(det(var1)$modulus))
     det2 <- exp(as.double(det(var2)$modulus))

     vero1 <- det1^(-1/2)*exp(-0.5*t(y1i-x1i%*%alfa1)%*%solve(var1)%*%(y1i-x1i%*%alfa1))
     vero2 <- det2^(-1/2)*exp(-0.5*t(y2i-x2i%*%alfa2)%*%solve(var2)%*%(y2i-x2i%*%alfa2))
     v <- vero1*priori[1]+vero2*priori[2]
     logvero <- logvero + log(v)

     poster1[i] <- vero1*priori[1]/(vero1*priori[1] + vero2*priori[2])
     verotot1 <-verotot1 + vero1
     verotot2 <-verotot2 + vero2
   }
   list(vero1=verotot1,vero2=verotot2,poster1=poster1,logvero=logvero)
}

# Estimacion de Alfa.k
#
alfa_function(datos,n,m,poster,sigma,D)
{   sum1 <- Matrix(0,nrow=5,ncol=5)
    sum2 <- Matrix(0,nrow=5,ncol=1)
    for (i in 1:m)
    { Oi_misout(datos$y[datos$sujeto==i])
      yi <- na.exclude(datos$y[datos$sujeto==i])
      zi <- Oi%*%datos$z[datos$sujeto==i,]
      xi <- Oi%*%datos$x[datos$sujeto==i,]
```

```
      if (!is.Matrix(xi))   xi <- t(as.Matrix(xi))
      if (!is.Matrix(zi))   zi <- t(as.Matrix(zi))

      var  <- Oi%*%kronecker(sigma,diag(rep(1,n[i])))%*%t(Oi) + zi%*%D%*%t(zi)
      Wi   <- solve(var)
      sum1 <- sum1 + poster[i]*t(xi)%*%Wi%*%xi
      sum2 <- sum2 + poster[i]*t(xi)%*%Wi%*%yi
    }
    alfa <- solve(sum1)%*%sum2
}

# Mini-EM para estimar D.k y sigma.k
#
miniem_function(datos1,datos2,n,m,poster1,poster2,sigma1,sigma2,alfa1,alfa2,D1,D2)
{   gr1t1  <- Matrix(0,nrow=2,ncol=2);     gr1t2  <- Matrix(0,nrow=2,ncol=2)
    gr2t1  <- Matrix(0,nrow=2,ncol=2);     gr2t2  <- Matrix(0,nrow=2,ncol=2)
    Ip_diag(2)

    for (i in 1:m)
    { y1i <- datos1$y[datos1$sujeto==i]
      z1i <- datos1$z[datos1$sujeto==i,]
      x1i <- datos1$x[datos1$sujeto==i,]
      y2i <- datos2$y[datos2$sujeto==i]
      z2i <- datos2$z[datos2$sujeto==i,]
      x2i <- datos2$x[datos2$sujeto==i,]

      if (!is.Matrix(x1i)) x1i <- t(as.Matrix(x1i));  if (!is.Matrix(x2i)) x2i <- t(as.Matrix(x2i))
      if (!is.Matrix(z1i)) z1i <- t(as.Matrix(z1i));  if (!is.Matrix(z2i)) z2i <- t(as.Matrix(z2i))

      Oi_misout(datos1$y[datos1$sujeto==i])

      R1i_kronecker(sigma1,diag(rep(1,n[i])))
      R2i_kronecker(sigma2,diag(rep(1,n[i])))

      k_2*i; l_k-1

      y1Oi     <- na.exclude(datos1$y[datos1$sujeto==i])
      var1     <- R1i + z1i%*%D1%*%t(z1i)
      W1i      <- solve(var1)
      var1Oi   <- Oi%*%R1i%*%t(Oi) + Oi%*%z1i%*%D1%*%t(Oi%*%z1i)
      W1iOi    <- solve(var1Oi)
      b1i[l:k] <- D1%*%t(Oi%*%z1i)%*%W1iOi%*%(y1Oi-Oi%*%x1i%*%alfa1)
      e1i      <- R1i%*%t(Oi)%*%W1iOi%*%(y1Oi-Oi%*%x1i%*%alfa1)

      y2Oi     <- na.exclude(datos2$y[datos2$sujeto==i])
      var2     <- R2i + z2i%*%D2%*%t(z2i)
      W2i      <- solve(var2)
      var2Oi   <- Oi%*%R2i%*%t(Oi) + Oi%*%z2i%*%D2%*%t(Oi%*%z2i)
      W2iOi    <- solve(var2Oi)
      b2i[l:k] <- D2%*%t(Oi%*%z2i)%*%W2iOi%*%(y2Oi-Oi%*%x2i%*%alfa2)
      e2i      <- R2i%*%t(Oi)%*%W2iOi%*%(y2Oi-Oi%*%x2i%*%alfa2)

# Calculos para estimar SIGMA
     Ini_diag(n[i])
     for (j in 1:n[i])
     { j2 <- j+n[i]
       H_Oi%*%kronecker(Ip,Ini[j,])
       e1ij  <- c(e1i[j],e1i[j2])
       gr1t1 <- gr1t1 + poster1[i]*(e1ij%*%t(e1ij)+sigma1-sigma1%*%t(H)%*%W1iOi%*%H%*%sigma1)
       e2ij  <- c(e2i[j],e2i[j2])
       gr2t1 <- gr2t1 + poster2[i]*(e2ij%*%t(e2ij)+sigma2-sigma2%*%t(H)%*%W2iOi%*%H%*%sigma2)
     }
```

```
# Calculos para estimar matriz covarianzas D
     gr1t2  <- gr1t2 + poster1[i]*(b1i[l:k]%*%t(b1i[l:k]) +
D1 - D1%*%t(Oi%*%z1i)%*%W1iOi%*%Oi%*%z1i%*%D1)
     gr2t2  <- gr2t2 + poster2[i]*(b2i[l:k]%*%t(b2i[l:k]) +
D2 - D2%*%t(Oi%*%z2i)%*%W2iOi%*%Oi%*%z2i%*%D2)
    }
    sigma1 <- gr1t1/sum(poster1*n)
    D1     <- gr1t2/sum(poster1)
    sigma2 <- gr2t1/sum(poster2*n)
    D2     <- gr2t2/sum(poster2)

    list(sigma1=sigma1,sigma2=sigma2,D1=D1,D2=D2,b1i=b1i,b2i=b2i)
 }

# Pasos del Algoritmo
#
veroini1 <- 5000; veroini2 <- 5000
delta <- 1
iter  <- 0
while(delta > 0.0001)
{ iter_iter+1
  datos1  <- matrices(DS4,n,m,alfa1,b1i)
  datos2  <- matrices(DS4,n,m,alfa2,b2i)
  temp_posteriori(datos1,datos2,n,m,priori,poster1,sigma1,sigma2,alfa1,alfa2,D1,D2)
  poster1 <- temp$poster1
  poster2 <- 1-poster1

  verofin1_temp$vero1
  verofin2_temp$vero2
  logvero_temp$logvero

  alfa1 <- alfa(datos1,n,m,poster1,sigma1,D1)
  alfa2 <- alfa(datos2,n,m,poster2,sigma2,D2)

  temp_miniem(datos1,datos2,n,m,poster1,poster2,sigma1,sigma2,alfa1,alfa2,D1,D2)
  sigma1 <- temp$sigma1;  sigma2 <- temp$sigma2
  D1     <- temp$D1;      D2    <- temp$D2
  b1i    <- temp$b1i;     b2i   <- temp$b2i

  priori[1] <- sum(poster1)/m
  priori[2] <- 1-priori[1]

  delta_max( abs(log(veroini1)-log(verofin1))/abs(log(veroini1)), +
abs(log(veroini2)-log(verofin2))/abs(log(veroini2)) )
  veroini1_verofin1
  veroini2_verofin2

  print(c("Veros: ",round(veroini1,5),round(veroini2,5),"DELTA: ",round(delta,7)))
  print(c("LOG(Vero) TOTAL          ",as.double(logvero)))
  print(c("Probabilidades a Priori ",round(priori,3)))
  print(c("Parametro Alfa1   ",round(alfa1,4)))
  print(c("Parametro Alfa2   ",round(alfa2,4)))
  print("Varianzas del error 1 y 2")
  print.matrix(sigma1)
  print.matrix(sigma2)
  print("Varianzas D1 y D2")
  print.matrix(D1)
  print.matrix(D2)
  print("Probabilidades a Posteriori")
  print(round(poster1,3))
}
```

References

[1] Akaike H. A new look at the statistical identification model. *IEEE transactions on Automatic Control* 1974; 19:716-723.

[2] Anderson KM, Castelli WP, Levy D. Cholesterol and mortality. 30 years of follow-up from the Framingham study. *Journal of the American Medical Association* 1987; 257: 2156-2180.

[3] Banfield JD, Raftery AE. Model-based gaussian and non gaussian clustering. *Biometrics* 1993; 49:803-821.

[4] Beal SL, Sheiner LB. NONMEM User's Guide Part I. San Francisco: Division of Clinical Pharmacology. University of California. 1979.

[5] Beal SL, Sheiner LB. The NONMEM system. *The American Statistician* 1980; 34:118-119.

[6] Beale EM, Little RJ. Missing values in multivariate analysis. *Journal of the Royal Statistical Society, Series B* 1975; 37:129-145.

[7] Beltrán C, Wolff M. Estudio de evaluación del impacto de las TAR en PVVIH beneficiarias del Sistema Público de Salud de Chile. *Informe Técnico SIDACHILE* 2005.

[8] Bensmail H, Celeux G, Raftery AE, Robert C. Inference in model-based cluster analysis. *Statistics and Computing* 1997; 7:1-10.

[9] Cnaan A, Laird NM, Slasor P. Using the general linear mixed model to analyse unbalanced repeated measures and longitudinal data. *Statistics in Medicine* 1997; 16:2349-2380.

[10] de la Cruz-Mesía R. Discriminant and cluster analysis for longitudinal data. *Tesis de grado*. Pontificia Universidad Católica de Chile 2005; 16:2349-2380.

[11] de la Cruz-Mesía R, Quintana F, Marshall G. Model based clustering for longitudinal data. *En vías de publicación*. Pontificia Universidad Católica de Chile 2006; 16:2349-2380.

[12] Daniels HE. The estimation of components of variance. *Journal of the Royal Statistical Society, Supplement* 1939; 6(2):186-197.

[13] Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* 1977; 39:1-38.

[14] Ellard GA, Johnstone FD, Prescott RJ, Ji-Xian W, Jian-Hua M.Smoking during pregnancy: the dose dependence of birthweight deficits. *British Journal of Obstetrics and Gynaecology* 1996;103:806-813.

[15] Fraley C, Raftery AE. MCLUST: Software for model-based cluster and discriminant analysis. *Technical Report No. 342l* Department of Statistics. University of Washington 1999.

[16] Fraley C, Raftery AE. How many clusters? which clustering method? answers via model-based cluster analysis. *The Computer Journal* 1998; 41:578-588.

[17] Fraley C, Raftery AE. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association* 2002; 97:611-631.

[18] Friedman J, Hastie T, Tibshirani R. The elements of statistical learning: data mining, inference, and prediction. New York: Springer 2001.

[19] Grizzle JE, Allen DM. Analysis of growth and dose-response curves. *Biometrics* 1969; 25:357-381.

[20] Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982; 143:29-36.

[21] Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology* 1983; 148:839-843.

[22] Harville DA. Maximum likelihood approaches to variance components estimation and to related problems. *Journal of the American Statistical Association* 1977; 72(358):320-338.

[23] Hasselblad V. Estimation of finite mixtures of distributions from the exponential family. *Journal of the American Statistical Association* 1969; 64(328):1459-1471.

[24] Henderson CR. Estimation of variance and covariance components. *Biometrics* 1953; 9:226-252.

[25] Hogan JW, Laird NM. Mixture models for the joint distribution of repeated measures and event times. *Statistics in Medicine* 1997; 16:239-257.

[26] Jain AK, Duin RP, Jianchang M. Statistical pattern recognition: A review. *IEEE Transaction on Pattern Analysis and Machine Inteligence* 2000; 22(1):4-37.

[27] Jain AK, Murty MN, Flynn PJ. Data clustering: A review. *ACM Computing Surveys* 1999; 31(3):264-323.

[28] Jennrich RI, Schluchter MD. Unbalanced repeated measures models with structural covariance matrices. *Biometrics* 1986; 42:805-820.

[29] Kleinbaum DG. A generalization of the growth curve model which allows missing data. *J. Multiv. Anal.* 1973; 3:117-124.

[30] Krzanowski WJ, Lai YT. A criterion for determining the number of groups in a data set using sum-of-squares clustering. *Biometrics* 1988; 44:23-34.

[31] Laird NM, Ware JH. Random-effects models for longitudinal data. *Biometrics* 1982; 42:121-130.

[32] Laird N, Lange N, Stram D. Maximum likelihood computations with repeated measures: applications of the EM algorithm. *Journal of the American Statistical Association* 1987; 82:97-105.

[33] Lindstrom MJ, Bates DM. Nonlinear random effects models for repeated measures data. *Biometrics* 1990; 46:673-687.

[34] Marshall G, de la Cruz-Mesía R, Barón AE, Rutledge JH, Zerbe, GO. Nonlinear random effects model for multivariate response with missing data. *Statistics in Medicine* (En publicación).

[35] Marshall G., Barón A.E. Linear discriminant models for unbalanced longitudinal data. *Statistics in Medicine* 2000; 19:1969-1981.

[36] McCulloch CE, Searle SR. Generalized, Linear and Mixed Models. New York: John Wiley & Sons, 2001.

[37] McLachlan GJ, Gordon RD. Mixture models for partially unclassified data: a case study of renal venous renin in hypertension. *Statistics in Medicine* 1989; 8:1291-1300.

[38] McLachlan GJ, Basford KE. Mixture models: inference and applications to clustering. Marcel Dekker, New York, 1988.

[39] Pinheiro JC, Bates DM. Mixed-effects models in S-PLUS. New York: Springer, 2000.

[40] Potthoff RF, Roy SN. A generalized analysis of variance model useful especially for growth curves. *Biometrika* 1964; 51:313-326.

[41] Racine-Poon A. A Bayesian approach to nonlinear random effects models. *Biometrics* 1985; 41:1015-1023.

[42] Raftery AE, Dean N. Variable selection for model-based clustering. *Technical Report no. 452* Department of Statistics, University of Washington 2004.

[43] Rao CR. Some problems involving linear hipotheses in multivariate analysis. *Biometrika* 1959; 46:49-58.

[44] Rao CR. The theory of least squares when the parameters are stochastic ans its application to the analysis of growth curves. *Biometrics* 1965; 52:447-468.

[45] Rao CR. Simultaneous estimation of parameters in different linear models and applications to biometrics problems. *Biometrics* 1975; 31:545-554.

[46] Reinsel G. Estimation and prediction in a multivariate random-effects generalized linear model. *Journal of the American Statistical Association* 1984; 79:406-414.

[47] Schwarz G. Estimating the dimension of a model. *The Annals of Statistics* 1978; 6(2):461-464.

[48] Searle SR. CR Henderson, the statistician; and his contribution to variance components estimation. *J Dairy Sci* 1990; 74:4035-4044.

[49] Shah A, Laird N, Schoenfeld D. A random-effects model for multiple characteristics with possibly missing data. *Journal of the American Statistical Association* 1997; 92:775-779.

[50] Sheiner LB, Beal SL. Evaluation of methods for estimating population pharmacokinetic parameters. I.Michaelis-Menten model: routine clinical pharmacokinetic data. *Journal of Pharmacokinetics and Biopharmaceutics* 1980; 8:553-571.

[51] Spiessens B, Verbeke G, Komarek A. A SAS-macro for the classification of longitudinal profiles using mixtures of normal distributions in nonlinear and generalised linear mixed model. *Technical Report* 2002.

[52] Sugar CA, James GM. Finding the number of clusters in a data set: an information theoretic approach. *Technical Report* 2003.

[53] Theodoridis S, Kostroumbas K. Pattern recognition. Academic Press, San Diego, 1999.

[54] Tsutakawa RK. Mixed model for analyzing geographic variability in mortality rates. *Journal of the American Statistical Association* 1988; 83(401):37-42.

[55] Verbeke G, Lesaffre E. A linear mixed effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association* 1996; 443(91):217-221.

[56] Vonesh EF, Carter RL. Mixed-effects nonlinear regression for unbalanced repeated measures. *Biometrics* 1992; 48:1-17.

[57] Walker S. An EM Algorithm for nonlinear random effects models. *Biometrics* 1996; 52:934-944.

[58] Ware JH. Linear models for the analysis of longitudinal studies. *The American Statistician* 1985; 39(2):95-101.

[59] Wolfe JH. Pattern clustering by multivariate mixture analysis. *Multivariate Behavioral Research* 1970; 5:329-350.

[60] Woolson RF, Leeper JD, Clarke WR. Analysis of incomplete data from longitudinal and mixed longitudinal studies. *Journal of the Royal Statistical Society, Series A* 1978; 141:242-252.

[61] Yuh L et al. Population pharmacokinetic/pharmacodynamic methodology and applications: a bibliography. *Biometrics* 1994; 50:566-575.

Index