García, María del Carmen
Rapelli, Cecilia
Instituto de Investigaciones Teóricas y Aplicadas, Escuela de Estadística.

MODELO MIXTO CON UN ESTIMADOR SUAVIZADO DE LA DENSIDAD DE LOS EFECTOS ALEATORIOS. UNA APLICACIÓN [1]

Resumen: Una herramienta importante para el análisis de datos longitudinales son los modelos lineales mixtos. Estos modelos expresan los parámetros específicos de las unidades en función de efectos fijos y aleatorios y para considerar la correlación entre las mediciones repetidas se introducen errores intra unidad. Un supuesto usado habitualmente es el de distribución normal para los errores y los efectos aleatorios. El supuesto sobre estos últimos suele no ser acertado y su cumplimiento puede ser dificultoso de verificar con las herramientas estadísticas estándares. Debido a que la predicción de los efectos aleatorios depende tanto de los errores como de los efectos aleatorios, los gráficos usuales para comprobar el supuesto de normalidad no permiten diferenciar cual de los dos supuestos distribucionales es el incorrecto. Varios autores propusieron métodos que relajan el supuesto de normalidad de los efectos aleatorios y utilizan técnicas de suavizado para aproximar la distribución de los mismos. Este trabajo presenta una reseña de algunos de ellos y se utiliza el enfoque denominado modelo mixto con mezclas gaussianas penalizado para la aplicación.

Palabras claves: Datos longitudinales Modelos mixtos Efectos aleatorios Densidad suavizada.

Abstract: The linear mixed models are an important tool for the analysis of longitudinal data. These models express the specific parameters of the units in terms of fixed and random effects and intra-unit errors are introduced, to consider the correlation between the repeated measurements. A commonly used assumption is the normal distribution for errors and random effects. The assumption about the latter is usually not accurate and its compliance can be difficult to verify with the standard statistical tools. Because the prediction of random effects depends on both errors and random effects, the usual plots to check the assumption of normality do not allow us to differentiate which of the two distributional assumptions is the incorrect one. Several authors proposed methods that relax the assumption of normality of the random effects and use smoothing techniques to approximate the distribution of the same. This paper presents a review of some of them and uses the approach called mixed model with penalized Gaussian mixtures for the application.

Keywords: Longitudinal data Mixed models Random effects Smooth density

1. Introducción

[1] Este trabajo se elaboró en el marco del Proyecto ECO183 Titulado “Estrategias para la mode-

Los estudios longitudinales juegan un rol importante en la investigación, pues están diseñados para evaluar los cambios en las respuestas de una unidad a través del tiempo y relacionar estos cambios con covariables. Estos estudios no sólo permiten conocer los efectos poblacionales sino también comportamientos específicos individuales. Los modelos mixtos se utilizan para analizar este tipo de datos. Suponen que la forma del modelo que relaciona la respuesta con las covariables es común para todas las unidades, pero mediante la incorporación de efectos aleatorios permiten que algunos de los parámetros del mismo varíen entre los individuos.

Un supuesto usado habitualmente es el de distribución normal para los errores y los efectos aleatorios. El supuesto sobre estos últimos suele no ser acertado y su cumplimiento puede ser dificultoso de verificar con las herramientas estadísticas estándares.

Varios autores han comprobado que la inferencia sobre los efectos fijos es robusta a la falta de normalidad de los efectos aleatorios (Butler et al., 1992, Verbeke et al, 1997). Sin embargo una estimación eficiente de los mismos requiere una correcta especificación de su distribución. La inferencia relativa a los efectos aleatorios puede estar afectada cuando se supone erróneamente normalidad.

En los últimos años se propusieron varios enfoques que relajan este supuesto (Ghidey et al., 2010), asumiendo sólo que la distribución de efectos aleatorios tiene una densidad "suave" y representando la misma de diferentes maneras.

En este trabajo se presentan algunos enfoques para estimar los parámetros de un modelo mixto cuando la distribución de los efectos aleatorios no es normal. Para la aplicación se utiliza el método propuesto por Ghidey et al. (2004) que realiza un suavizado de la densidad de los efectos aleatorios.

.

2. Modelos mixtos

La expresión del modelo lineal mixto es,

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\ \mathbf{b}_i + \mathbf{e}_i, \qquad i=1,\ldots,N, \tag{1}$$

siendo, $\mathbf{Y_i} = (Y_{i1},\ldots,Y_{ini})'$ el vector (n_i x 1) de las repuestas de la i-ésima unidad, i=1,…,N, $\mathbf{X}_i$ una matriz de diseño de dimensión (n_i x p), $\boldsymbol{\beta}$ un vector de dimensión (p x 1) de parámetros denominados efectos fijos, $\mathbf{Z_i}$ una matriz de diseño de dimensión (n_i x q) que caracteriza la parte aleatoria del modelo, $\mathbf{b_i}$ un vector de dimensión (q x 1) de efectos aleatorios, $\mathbf{e}_i = (e_{i1}, e_{i2}, ..., e_{in_i})'$ un vector de dimensión (n_i x 1) de los errores aleatorios. Los supuestos que se realizan son,

$$\mathbf{e}_i \sim N_{n_i}(\mathbf{0};\mathbf{R}_i) \quad \text{y} \quad \mathbf{b}_i \sim N_q(\mathbf{0},\mathbf{D}).$$

La distribución marginal del vector $\mathbf{Y_i}$ es, $\mathbf{Y}_i \sim N(\mathbf{X}_i\boldsymbol{\beta}; \mathbf{Z}_i'\mathbf{D}\mathbf{Z}_i + \mathbf{R}_i)$ $\mathbf{V}_i = \mathbf{Z}_i'\mathbf{D}\mathbf{Z}_i + \mathbf{R}_i$.

---

lación estadística y el análisis de datos longitudinales", dirigido por María del Carmen Garcia

La estimación de los parámetros se obtiene maximizando la verosimilitud marginal mediante el método de máxima verosimilitud. El estimador máximo verosímil de $\boldsymbol{\beta}$ es,

$$\tilde{\boldsymbol{\beta}} = \left(\sum_{i=1}^{N} \mathbf{X}_i' \hat{\mathbf{V}}_i^{-1}(\boldsymbol{\theta}) \mathbf{X}_i \right)^{-1} \sum_{i=1}^{N} \mathbf{X}_i' \hat{\mathbf{V}}_i^{-1}(\boldsymbol{\theta}) \mathbf{Y}_i .$$

La predicción del vector de efectos aleatorios, denominada EBLUP o estimador empírico de Bayes, viene dada `por,

$$\hat{\mathbf{b}} = \hat{\mathbf{D}} \mathbf{Z}_i' \hat{\mathbf{V}}_i^{-1} (\mathbf{Y}_i - \mathbf{X}_i \hat{\boldsymbol{\beta}}) .$$

Los estimadores de los parámetros en el modelo líneal mixto son insesgados cuando no se cumple el supuesto de normalidad (Verbeke y Lesaffre, 1997), sin embargo, la predicción de los efectos aleatorios ($\hat{\mathbf{b}}_i$) puede estar muy influenciada por el mismo. A su vez, esta distribución es dificultosa de comprobar mediante el gráfico probabilístico normal ya que la predicción del efecto aleatorio depende tanto del efecto aleatorio ($\mathbf{b}_i$) como del error aleatorio ($\mathbf{e}_i$) y no se puede distinguir cuál es el supuesto erróneo (Verbeke y Lesaffre, 1996).

En los últimos años surgieron métodos que relajan el supuesto de normalidad y estiman la densidad de los efectos aleatorios mediante una distribución más general y flexible. Entre estos métodos se pueden mencionar,

1.- el enfoque semi-no paramétrico de Zhang y Davidian (SPN). Plantea un modelo lineal mixto semiparamétrico donde los efectos aleatorios se suponen pertenecer a una clase de densidades suavizadas. La forma de la densidad es representada por una expansión en serie truncada semi no paramétrica. Esto permite obtener una forma cerrada para la verosimilitud marginal de los datos. Esta representación admite el modelo normal como un caso particular.

2.- el modelo de heterogeneidad de Verbeke y Lesaffre. Es un método alternativo al enfoque anterior en el cual se representa la densidad de los efectos aleatorios por una mezcla de normales. Para la implementación de este enfoque se utiliza el algoritmo E-M considerando el número de normales en la mezcla como un parámetro de ajuste e imponiendo restricciones sobre las probabilidades de la mezcla para poder utilizar las técnicas de optimización usuales.

3.- el suavizado de Shen y Louis (SBR). Partiendo de un modelo con un solo efecto aleatorio en la ordenada, este método recursivo fue sugerido para obtener un estimador suavizado de la densidad univariada de los efectos aleatorios, pues supone la presencia de un sólo efecto aleatorio. El método comienza con un cálculo aproximado del estimador suavizado de la densidad y lo va refinando hasta que se estabiliza, después de pocas iteraciones. Cuando se logra la convergencia la distribución estimada se convierte en el estimador máximo verosímil no paramétrico propuesto por Laird (1978).

4.- el modelo lineal mixto con mezclas Gaussianas penalizado (Ghidey et al, 2004). El suavizado de la densidad de los efectos aleatorios está basado sobre un enfoque similar al suavizado P-spline, reemplazando las funciones base (B-spline) por sus densidades Gaussianas aproximadas.

En la siguiente sección se presenta este último método en forma más ampliada, ya que se utilizará en la aplicación.

3. Modelo lineal mixto con mezclas gaussianas penalizado (PGM)

La distribución normal puede ser demasiado restrictiva en la práctica para representar la distribución de la variabilidad entre unidades, por lo que existe la necesidad de un modelo con un supuesto distribucional más flexible para los efectos aleatorios.

Para el modelo (1) se supone que la distribución de $\mathbf{b}_i = (b_{0i}, b_{1i})$ no es normal y que su densidad se puede aproximar por una mezcla de densidades Gaussianas,

$$f(\mathbf{b}_i) = \sum_{j=1}^{J}\sum_{l=1}^{L} c_{jl}\, N(\boldsymbol{\mu}_{jl}, \mathbf{D}_b),$$

donde,

$\boldsymbol{\mu}_{jl} = (\mu_{1j}, \mu_{2l})'.\ j = 1,...,J.\ l = 1,...,L$, es el vector de medias de la densidad Gaussiana obtenidos de una grilla de dimensión JxL,

$\mathbf{D_b}$ es una matriz de covariancias de la densidad Gaussiana bivariada,

$c_{jl} = \dfrac{\exp(a_{jl})}{\sum_{k=1}^{J}\sum_{m=1}^{L}\exp(a_{km})}$ son elementos de la matriz de JxL de coeficientes, tal que $\sum_{j=1}^{J}\sum_{l=1}^{L} c_{jl} = 1.$

Esta expresión permite maximizar con respecto a los parámetros de suavizado a_{jl} y garantiza que todos los coeficientes de la mezcla c_{jl} sean positivos.

Tanto los parámetros del modelo como la distribución de los efectos aleatorios (determinada por los a_{jl}) son conjuntamente estimados maximizando la verosimilitud marginal de $\mathbf{Y}_i$. Los efectos fijos $\boldsymbol{\beta}$, la variancia de error σ^2 y los parámetros de suavizado $\mathbf{a} = (a_{11}, a_{12}, ..., a_{JL})'$ se incluyen dentro de un vector $\boldsymbol{\theta}$. La densidad marginal, dado $\boldsymbol{\theta}$, que es mezcla de densidades Gaussianas, viene dada por,

$$f(\mathbf{Y}_i; \boldsymbol{\theta}) = \sum_{j=1}^{J}\sum_{l=1}^{L} c_{jl}\, N(\mathbf{X}_i\,\boldsymbol{\beta} + \mathbf{Z}_i\,\boldsymbol{\mu}_{jl}, \mathbf{Z}_i\,\mathbf{D}_b\mathbf{Z}_i' + \sigma^2\,\mathbf{I}_{n_i}).$$

La log verosimilitud marginal es, entonces,

$$\ell(\boldsymbol{\theta}; \mathbf{Y}) = \sum_{i=1}^{N} \log\{f(\mathbf{Y}_i; \boldsymbol{\theta})\}.$$

La cantidad (J x L) de las densidades componentes caracteriza el nivel de suavizado y bondad de ajuste de la distribución estimada. Para evitar el sobreajuste que surge al considerar una grilla amplia de (J x L) densidades se penaliza la log verosimilitud.

Siguiendo la idea de Eilers y Marx (1996), Ghidey et al. (2004) toman un número relativamente grande de densidades Gaussianas base y penalizan la log verosimilitud con una penalidad basada en las diferencias de los coeficientes adyacentes produciendo la función de log verosimilitud penalizada,

$$\ell_p(\boldsymbol{\theta};\mathbf{Y}\mid\boldsymbol{\lambda}) = \ell(\boldsymbol{\theta};\mathbf{Y}) - \left[\frac{\lambda_1}{2}\sum_j\sum_k(\Delta_1^k a_{jl})^2 + \frac{\lambda_2}{2}\sum_j\sum_k(\Delta_2^k a_{jl})^2\right],$$

donde. Δ_1^k es un operador diferencia de orden k para la i-ésima dimensión y $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)$ es un vector de parámetros de penalidad para cada una de las dimensiones.

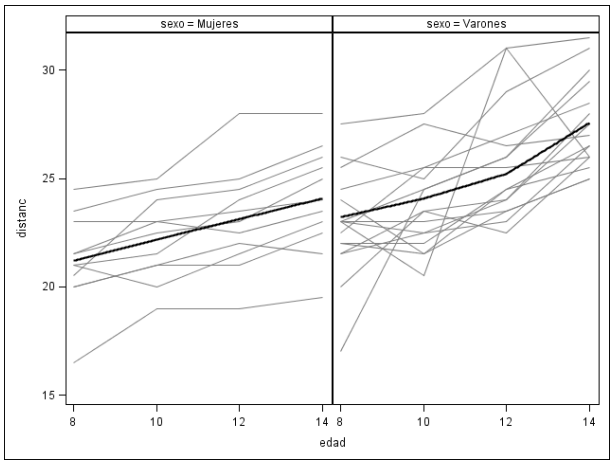
Esta función, para un $\boldsymbol{\lambda}$ dado, se puede maximizar usando Newton-Raphson o el algoritmo E-M y para elegir el $\boldsymbol{\lambda}$ óptimo se usa el criterio de Akaike.

Este método está implementado en la macro PGM del programa estadístico SAS.

4. Aplicación

La metodología descrita se aplica a un conjunto de datos, clásico en el análisis de datos longitudinales, que fueron recolectados en la Universidad de Carolina del Norte y analizados por Pothoff y Roy en 1964. El estudio considera 27 niños, 16 varones y 11 mujeres. Se midió la distancia (en mm) desde el centro de la glándula pituitaria hasta la fisura pterygomaxilar de cada uno de los niños involucrados en el estudio, a las edades de 8, 10, 12 y 14 años, con el objetivo de determinar si la tasa de cambio de la distancia de los maxilares a través del tiempo es similar para varones y mujeres.

Gráfico 1 Perfiles individuales y perfil promedio por sexo



La inspección de los perfiles individuales de la distancia maxilar en función de la edad muestra que dos de las trayectorias de los varones están más alejadas que las del resto. El gráfico del perfil promedio permite inferir que la trayectoria promedio es lineal.

Se propuso el siguiente modelo lineal mixto con dos efectos aleatorios,

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 \delta_i + \beta_3 \delta_i t_{ij} + b_{0i} + b_{1i} t_{ij} + e_{ij} \quad i = 1,2,...,27 \quad j = 1,2,...,4$$

$$Var(\mathbf{b}_i) = Var\begin{pmatrix} b_{0i} \\ b_{1i} \end{pmatrix} = \mathbf{D}$$

$$Var(\mathbf{e}_i) = Var\begin{pmatrix} e_{i1} \\ . \\ e_{in_i} \end{pmatrix} = \mathbf{R} = \sigma^2 \mathbf{I}$$

donde, t_{ij} representa la edad y δ_i =0 si la unidad i es mujer y δ_i =1 si la unidad es varón i.

La estimación de los parámetros se realiza utilizando el procedimiento “mixed” del programa estadístico SAS (denominado de ahora en adelante MIXTO), considerando que la distribución de los efectos aleatorios es normal, y la macro PGM, que relaja el supuesto de normalidad y estima los parámetros del modelo conjuntamente con la densidad de los efectos aleatorios

Tabla 1 Estimación de los efectos fijos y sus errores estándares mediante los enfoques MIXTO y PGM

Parámetro	MIXTO		PGM	
	Estimación	Error Estándar	Estimación	Error Estándar
Ordenada	17.3727	1.1820	17.4441	0.49122
Edad	0.4795	0.09980	0.4838	0.03639
Sexo	-1.0321	1.5355	-0.9441	0.82661
Edad *Sexo	0.3048	0.1296	0.2666	0.07147

La Tabla 1 presenta los resultados de las estimaciones y los errores estándares. La estimación de los efectos fijos resultan similares para los dos ajustes, sin embargo los errores estándares con el método PGM son más chicos que con el procedimiento Mixto.

Las variancias de los efectos aleatorios presentan una magnitud levemente mayor cuando se estiman mediante el enfoque PGM, en cambio, la variancia del error resulta menor.

MIXTO

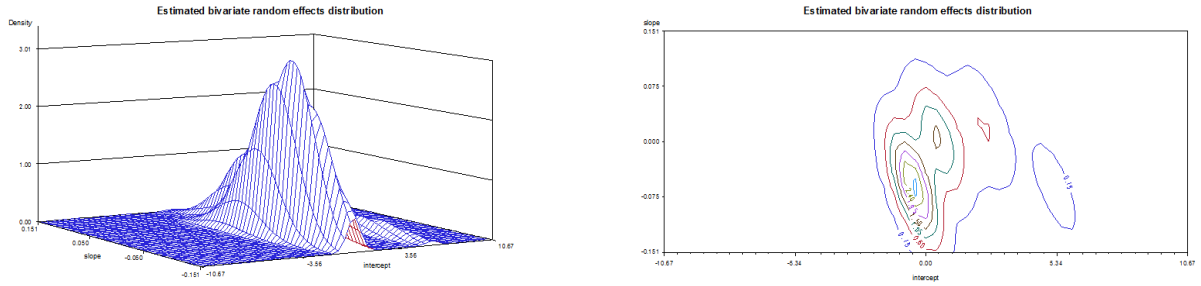
$$\mathbf{D} = \begin{bmatrix} 4.5569 & -0.1983 \\ -0.1983 & 0.02376 \end{bmatrix} \qquad \sigma^2 = 17162$$

PGM

$$\mathbf{D} = \begin{bmatrix} 5.60631 & -0.27224 \\ -0.27224 & 0.03089 \end{bmatrix} \qquad \sigma^2 = 149628$$

Gráfico 2 Distribución estimada de los efectos aleatorios y gráfico de contornos

(a) (b)

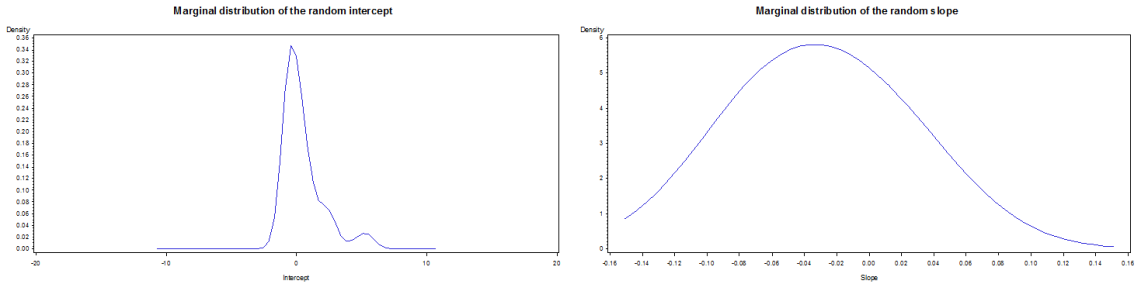


El gráfico 2a representa la densidad bivariada estimada de los efectos aleatorios b_i para el ajuste PGM. En el mismo se observa que la distribución es diferente a la de una normal ya que muestra algo de asimetría y la presencia de otro pequeño modo.

El gráfico 2b proporciona otra perspectiva de esa distribución, los contornos de la densidad. Se puede observar que las estimaciones empíricas de Bayes de los $\mathbf{b}_i$ se agrupan en dos lugares distintos.

Gráfico 3 Distribuciones marginales estimadas de los efectos aleatorios de ordenada y pendiente

(a) (b)



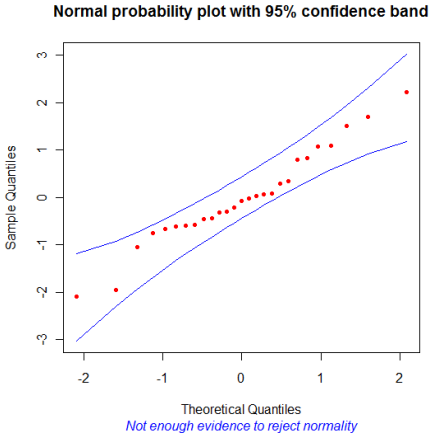
El gráfico 3a y 3b muestra, respectivamente, las densidades marginales estimadas de b_{0i} y b_{1i}. La distribución de las pendientes se asemeja a una normal, mostrando algo de asimetría. La densidad estimada para las ordenadas muestra evidencia de asimetría y sugiere un patrón aparente de no normalidad que se podría deber a causas no contempladas en el estudio, como por ejemplo, que no se haya considerado en el modelo alguna covariable importante.

Para evaluar la normalidad de los efectos aleatorios se utiliza un gráfico probabilístico normal (QQ-plot), conjuntamente con una banda de confianza simulada, llamada un "envelope" o sobre, del 95%, de manera que si los puntos caen fuera de la banda de confianza el supuesto de normalidad para los residuos no es válido.

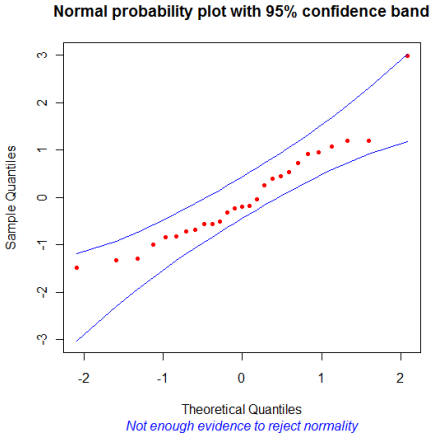
Gráfico 4 Gráfico probabilístico normal con "envelope" simulado para los efectos aleatorios

MIXTO

Ordenada

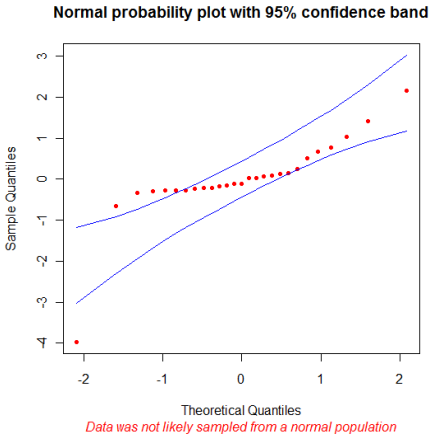


Pendiente

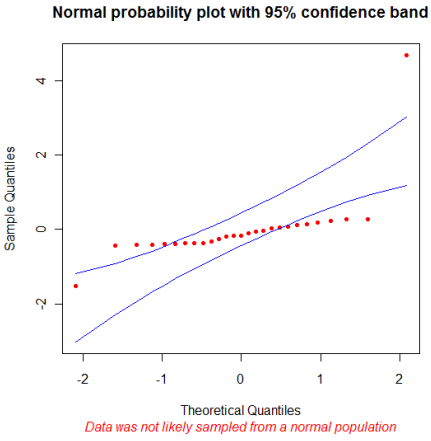


PGM

Ordenada



Pendiente



El grafico PGM sugiere no normalidad de los efectos aleatorios, mientras que en el MIXTO no se identifican valores fuera del intervalo simulado, indicando erróneamente que son normales.

5. Consideraciones finales

En este trabajo se presenta una introducción al análisis de datos longitudinales mediante un modelo mixto que relaja el supuesto de normalidad de los efectos aleatorios.

Los cuatro métodos presentados estiman la densidad de los efectos aleatorios mediante una distribución más flexible. Para la aplicación se utilizaron en forma comparativa el análisis mediante el modelo mixto tradicional que supone normalidad de los efectos aleatorios y el enfoque PGM.

De la aplicación se destaca que:

- La estimación de los efectos fijos resultó similar con ambos enfoques.
- En la estimación de algunos de los parámetros de covariancia se gana algo de eficiencia cuando se usa PGM.
- Como existen dos efectos (q=2) es posible representar la densidad conjunta permitiendo visualizar la bimodalidad de la estimación.
- La inspección de las densidades marginales de los efectos aleatorios permite observar la aparente desviación de la normalidad.
- El uso de PGM permitió corroborar la falta de normalidad de los efectos aleatorios.

La forma de la estimación sugiere que la inferencia sobre los efectos individuales bajo el supuesto de normalidad habitual podría ser engañosa. Sin embargo, la posibilidad de estimar la densidad de efectos aleatorios brinda un mejor conocimiento del problema y plantea una futura investigación.

REFERENCIAS BIBLIOGRÁFICAS

Butler SM, Louis TA. 1992. Random effects models with nonparametric priors. *Statistics in Medicine*; **11:** 1981–2000.

Eilers PHC, Marx BD. 1996. Flexible smoothing with *B*-splines and penalties (with discussions). *Statistical Science*; **11:** 89–121.

Ghidey W, Lesaffre E, Eilers P. 2004 Smooth random effects distribution in a linear mixed model. *Biometrics*; 60: 945–53.

Harville DA. 1977. Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*; 72: 320–40.

Laird NM. 1978. Nonparametric maximum likelihood estimation of a mixing distribution.. *Journal of the American Statistical Association*, **73:** 805–11.

Magder LS, Zeger SL. 1996. A smooth nonparametric estimate of a mixing distribution using mixtures of Gaussians. *Journal of the American Statistical Association*; **91:** 1141–51.

Diggle PJ, Liang KY, Zeger SL. 1994. *Analysis of Longitudinal Data*. Oxford: Clarendon press.

Pothff R F and Roy SN. 1964. A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika*, **51**, 313-326.

Shen W, Louis TA. 1999. Empirical Bayes estimation via the smoothing by roughening approach. *Journal of Computational and Graphical Statistics*; **8:** 800–23.

Verbeke G, Lesaffre E. 1996. A linear mixed model with heterogeneity in the random-effects population. *Journal of the American Statistical Association*; **91:** 217–21.

Verbeke G, Lesaffre E. 1997. The effect of misspecifying the random effects distribution in linear mixed models for longitudinal data. *Computational Statistics and Data Analysis*; **23:** 541–56.

Zhang D, Davidian M. 2001. Linear mixed models with flexible distributions of random effects for longitudinal data. *Biometrics*; 57: 795–802.