多元長期追蹤資料分群方法與應用

逢甲大學統計學系　王婉倫

一、前言

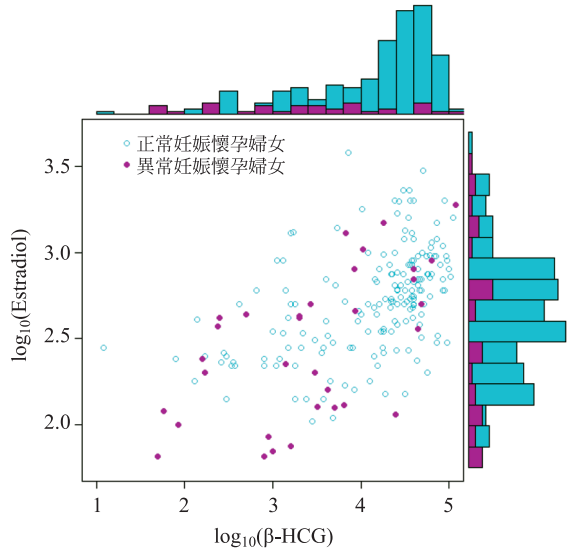
在資訊爆炸的大數據時代，面對資料蒐集與擷取快速膨脹的現象，分析涵蓋變數類型多元、重複量測數量龐大且結構複雜的長期追蹤資料(longitudinal data)是重要且具挑戰性的工作，尤其針對具異質性(heterogeneity)的長期追蹤資料進行基於模式(model-based)分群或分類，在近二十多年來已逐漸受到重視。有限混合式線性混合效應模型 (finite mixtures of linear mixed models [12])，其結合線性混合效應模型 (linear mixed-effects model [4]) 及高斯混合模型 (Gaussian mixture model [8]) 的優勢，已被廣泛地應用在長期追蹤資料之分群工作。此模型僅允許進行單一反應變數的長期追蹤曲線，然而在生物醫學和臨床試驗研究中，個體隨著時間被重複測量多個序列的反應變數是常見的，且這些受測個體可能來自異質的子群體而展現不同的反應成長曲線，同時資料本身亦可能存在潛在的離群值或厚尾白噪音。爲了處理來自數個異質子母體且具有潛在離群值的多元長期追蹤資料之分群工作，本文介紹一個新的統計模型，稱有限混合式多變量 T 非線性混合效應模型(finite mixtures of multivariate-*t* nonlinear mixed model; FM-MtNLMM [13])，此模型利用混合模型將變量分佈分解成若干個基於多變量 T 機率密度函數的概念，以描述母群體反應變量之異質性，同時結合多變量 T 非線性混合效應模型[14]，其能描述多個呈現非線性曲線形式且隱藏離群值的重複測量變數之優勢。因此，FM-MtNLMM 提供統計學家分析更廣泛類型的長期追蹤資料，特別是具異質性、離群值、厚尾分佈、以及呈現任意非線性軌跡的數據。

長期追蹤研究可能因爲實驗的限制、試驗者死亡或中途終止試驗等因素造成資料出現遺漏值。最常見的遺漏型態包含中途退出(dropout)及間歇式遺漏(intermittent missingness)，意即在同一時間點之下某些變數有測量記錄，但某些變數卻遺漏了，或者說觀測值在某遺漏值出現後的時間點亦可再次獲得[5]。一般而言，當分析不完整之單一反應變數，我們可以將資料視爲不均衡(unbalanced)，且設定每個個體有不同的設計矩陣即可輕易地完成模型架構。然而，當分析具間歇型遺失反應之多元長期追蹤資料，模型結構化和計算就變得更加複雜。本文也介紹如何運用 FM-MtNLMM 來分析不完整(incomplete)且具異質性的多元長期追蹤資料。

本研究的動機起因於懷孕婦女的臨床試驗研究，此爲期二年的試驗是由智利聖地牙哥一私立婦產科診所，其追蹤 124 位被診斷爲能順利分娩的正常妊娠懷孕婦女，以及 37 位被診斷爲自然流產或產生其他不良併發症之異常妊娠懷孕婦女。這 161 位婦女在產前檢查時重複測量β亞基人絨毛膜促性腺激素(β-subunit human chorionic gonadotropin; β-HCG) 和雌二醇(estradiol)濃度。醫學研究指出婦女在懷孕早期β-HCG 和雌二醇會急遽變化[15]，此血漿血清濃度是檢測婦女併發症或高失去胎兒風險的重要反應指標。在數據中，正常婦女的 β-HCG 和雌二醇遺失率分別爲 2.6%和 26.5%；異常婦女的兩指標遺失率分別爲 0%和 57.5%。圖一顯示兩組婦女的反應指標(經$\log_{10}$轉換)之變異量存在差異，且分佈似乎展現厚尾現象，此爲引發我們發展更穩健模型的動機。以下章節將介紹所提出的模型、參數估計演算法、分群技術及實證分析結果。

二、資料結構與統計模型

假設資料包含來自G個異質性子母體的n個受測個體，每個個體在可能不同的時間點下被重複測量r個特徵量(反應變數)。令$\boldsymbol{Y}_i = [\boldsymbol{y}_{i1} : \boldsymbol{y}_{i2} : \cdots : \boldsymbol{y}_{ir}]$爲第$i$個個體$(i = 1, \ldots, n)$的反



圖一　124 位正常妊娠及 37 位異常妊娠懷孕婦女之$\log_{10}(\beta-\mathrm{HCG})$與$\log_{10}(\mathrm{Estradiol})$觀測值的散佈圖及直方圖

應矩陣，其中$\boldsymbol{y}_{ij}=(y_{ij,1},\ldots,y_{ij,s_i})^{\mathrm{T}}$是第$j$個$(j=1,\ldots,r)$的反應向量；$\boldsymbol{X}_i$為感興趣的共變量(解釋變數)；$\boldsymbol{E}_{ig}=[\boldsymbol{e}_{i1,g}:\cdots:\boldsymbol{e}_{ir,g}]$為了第$g$群相對應於$\boldsymbol{Y}_i$的$s_i\times r$個體內誤差矩陣。為了模式化的便利，我們取$\boldsymbol{y}_i=\mathrm{vec}(\boldsymbol{Y}_i)$及$\boldsymbol{e}_i=\mathrm{vec}(\boldsymbol{E}_i)$，有限混合式多變量 T 非線性混合效應模型(FM-MtNLMM [13])，其結合了g個在混合效應上呈非線性數值向量且可微分的函數，以描述任意型態的反應軌跡。模型定義如下：

$$\boldsymbol{y}_i=\boldsymbol{\mu}_g(\boldsymbol{\eta}_{ig},\boldsymbol{X}_i)+\boldsymbol{e}_{ig}\ ,$$
機率爲$w_g, g=1,\ldots,G$，　(1)

其中$\boldsymbol{\mu}_g$是數值向量混合效應參數$\boldsymbol{\eta}_{ig}$和共變量$\boldsymbol{X}_i$的非線性數值向量可微分函數，成份混合效應$\boldsymbol{\beta}_g$和$\boldsymbol{b}_{ig}$可透過方程式$\boldsymbol{\eta}_{ig}=\boldsymbol{A}_i\boldsymbol{\beta}_g+\boldsymbol{B}_i\boldsymbol{b}_{ig}$被結合於模型中，而$\boldsymbol{A}_i$和$\boldsymbol{B}_i$分別爲固定效應與隨機效應之設計矩陣。假設隨機項$(\boldsymbol{b}_{ig}^{\mathrm{T}},\boldsymbol{e}_{ig}^{\mathrm{T}})^{\mathrm{T}}$滿足

$$\begin{bmatrix}\mathbf{b}_{ig}\\ \mathbf{e}_{ig}\end{bmatrix}\sim T_{(q+s_i r)}\left(\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix},\begin{bmatrix}\mathbf{D}_g & \mathbf{0}\\ \mathbf{0} & \mathbf{R}_{ig}\end{bmatrix},\nu_g\right),\qquad(2)$$

其中$T_d(\boldsymbol{\mu},\boldsymbol{\Omega},\nu)$表示$d$維度、位置向量$\boldsymbol{\mu}$、尺度共變異數矩陣$\boldsymbol{\Omega}$且自由度爲$\nu$的多變量 T 分佈，且$\mathbf{R}_{ig}=\boldsymbol{\Sigma}_g\otimes\mathbf{I}_{s_i}$。根據模型(1)及分佈假設(2)得知反應向量$\mathbf{y}_i$給定隨機效應$\mathbf{b}_i=\{\mathbf{b}_{i1},\ldots,\mathbf{b}_{iG}\}$的條件分佈爲：

$$f(\mathbf{y}_i|\mathbf{b}_i)=\sum_{g=1}^{G}w_g t_{s_i r}\left(\mathbf{y}_i\middle|\boldsymbol{\mu}_g(\boldsymbol{\eta}_{ig},\mathbf{X}_i),\mathbf{R}_{ig},\nu_g\right),$$

其中$t_d(\cdot\,|\boldsymbol{\mu},\boldsymbol{\Omega},\nu)$爲$T_d(\boldsymbol{\mu},\boldsymbol{\Omega},\nu)$之機率密度函數。

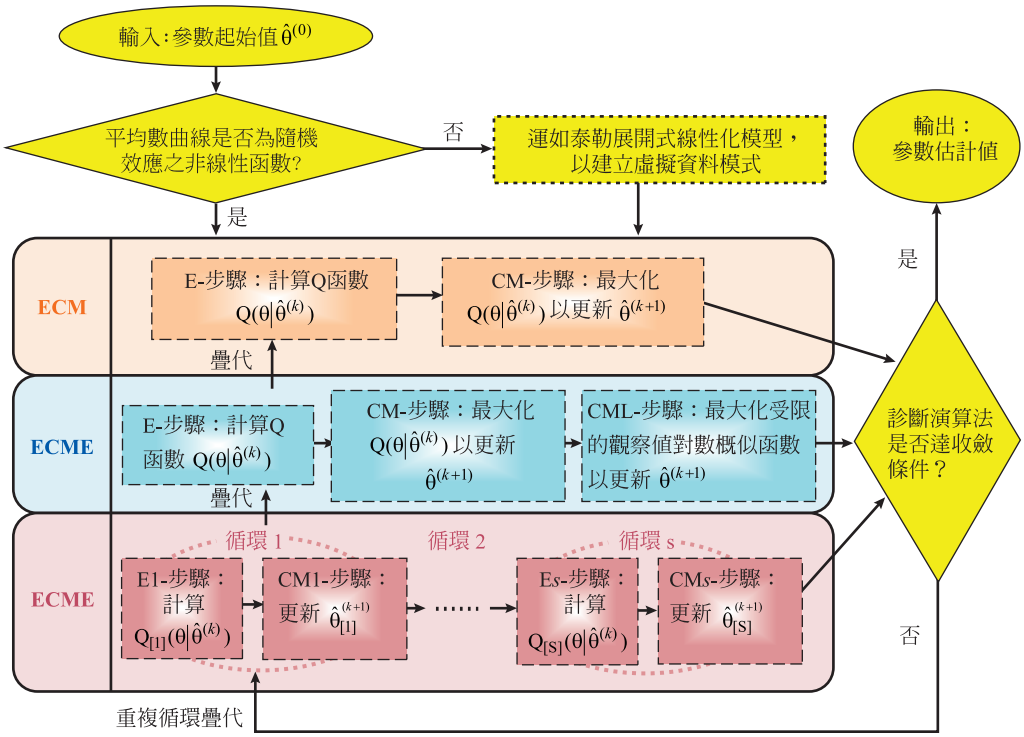
由於受測個體錯過就診時間、從研究中移除、或者後續追蹤遺漏等因素，在不規則觀測時間下紀錄的多元反應變數有遺失資料是無法避免的問題。因此，我們將$\mathbf{y}_i$切割成觀測值向量$\mathbf{y}_i^{\mathrm{o}}$和遺失值向量$\mathbf{y}_i^{\mathrm{m}}$以建構具遺失訊息的 FM-MtNLMM。爲方便計算，我們引入輔助排列矩陣$\mathbf{O}_i$和$\mathbf{M}_i$使得$\mathbf{y}_i^{\mathrm{o}}=\mathbf{O}_i\mathbf{y}_i$及$\mathbf{y}_i^{\mathrm{m}}=\mathbf{M}_i\mathbf{y}_i$，其它相關設計矩陣可透過此輔助矩陣加以定義，進一步推演不完整資料結構下的階層模式，以便發展可行的參數估計程序。

三、最大概似估計

模型的未知參數包含$\boldsymbol{\theta}=\{w_g,\nu_g,\boldsymbol{\beta}_g,\boldsymbol{D}_g,\boldsymbol{\Sigma}_g\}_{g=1}^{G}$，對於觀察資料$\mathbf{y}=\{\mathbf{y}_i\}_{i=1}^{n}$，參數$\boldsymbol{\theta}$的對數概似函數爲：

$$\ell(\boldsymbol{\theta}|\mathbf{y})=\Sigma_{i=1}^{n}\log f(\mathbf{y}_i)$$
$$=\Sigma_{i=1}^{n}\log\int f(\mathbf{y}_i|\mathbf{b}_i)f(\mathbf{b}_i)d\mathbf{b}_i$$

然而此對數概似函數涉及「對加總符號取對數」，且在非線性函數中涉及對高維度的$\mathbf{b}_i$積分，導致無法提供參數的明確公式解。爲了獲得模型參數之最大概似估計並評估觀察資料的對數概似函數，我們採用泰勒序列展開式建構虛擬資料模型(pseudo-data model)，並引入指標變數與隱藏權重變數，再運用期望值最大化型態(expectation maximization; EM [3])型態演算法來估計參數。EM 演算法通常被用來處理不完整資料的問題。因混合模型可視爲不完整資料之模型結構，故可以利用 EM 演算法求取模型的最大概似估計值。依照模型特性，有三種 EM 型態演算法可供選擇：期望值條件最大化(expectation conditional maximization; ECM [9])、選擇期望值條件最大化(expectation conditional maximization either; ECME [6])、以及交替期望值條件最大化(alternating expectation conditional maximization; AECM [10])演算法。演算程序如圖二所示。



圖二　期望值最大化(EM)型態演算法示意圖

四、追蹤曲線分群與判別分析

當完成模型配適並獲得參數估計值後，我們可利用個體的成份歸屬(component membership)之後驗機率來進行追蹤曲線分群(clustering)。

假設有一組訓練資料(training data)，其中個體的預後歸屬(prognostic membership)可事先獲得，我們欲判斷測驗資料(test data) $\mathbf{y}_{\text{new}}^{\text{o}}$是屬於哪一群，因此發展一判別分析法。首先，假設$\mathbf{y}_{\text{new}}^{\text{o}}$滿足模型(1)之設定，即

$$\mathbf{y}_{\text{new}}^{\text{o}} = \mathbf{O}_{\text{new}}\boldsymbol{\mu}_g\left(\boldsymbol{\eta}_{\text{new}}^{g}, \mathbf{X}_{\text{new}}\right) + \mathbf{e}_{\text{new}g}^{\text{o}} \ ,$$

機率爲w_g，其中$\boldsymbol{\eta}_{\text{new}}^{g} = \mathbf{A}_{\text{new}}\boldsymbol{\beta}^{g} + \mathbf{B}_{\text{new}}\mathbf{b}_{\text{new}}^{g}$，且$\left(\mathbf{b}_{\text{new}}^{g\text{T}}, \mathbf{e}_{\text{new}g}^{\text{o T}}\right)^{\text{T}}$滿足模型分佈假設(2)。將訓練資料配適 FM-MtNLMM 求得參數估計值$\widehat{\boldsymbol{\theta}} = \{\widehat{\boldsymbol{\theta}}^{g}\}_{g=1}^{G}$，接著給定此參數估計值，基於 0-1 損失函數(zero-one loss function)，我們計算出$\mathbf{y}_{\text{new}}^{\text{o}}$屬於第$g$群的後驗機率：

$$\hat{\mathcal{P}}_{g,\text{new}} = \hat{w}^{g}\hat{p}\left(\mathbf{y}_{\text{new}}^{\text{o}}|\widehat{\boldsymbol{\theta}}^{g}\right)/\hat{p}\left(\mathbf{y}_{\text{new}}^{\text{o}}|\widehat{\boldsymbol{\theta}}\right), \quad g = 1, \ldots, G \ ,$$

其中$\hat{p}\left(\mathbf{y}_{\text{new}}^{\text{o}}|\widehat{\boldsymbol{\theta}}\right) = \sum_{g=1}^{G}\hat{w}_g\hat{p}\left(\mathbf{y}_{\text{new}}^{\text{o}}|\widehat{\boldsymbol{\theta}}^{g}\right)$爲$\mathbf{y}_{\text{new}}^{\text{o}}$的預測機率密度函數，而$\hat{p}\left(\mathbf{y}_{\text{new}}^{\text{o}}|\widehat{\boldsymbol{\theta}}^{g}\right)$爲$\mathbf{y}_{\text{new}}^{\text{o}}$屬於第$g$群的預測密度之估計。根據鑑別原則(discriminant rule [2])，若$\hat{\mathcal{P}}_{s,\text{new}} > \hat{\mathcal{P}}_{g,\text{new}}$，其中$s \neq g = 1, \ldots, G$，或者當$\hat{w}^{s}\hat{p}\left(\mathbf{y}_{\text{new}}^{\text{o}}|\widehat{\boldsymbol{\theta}}^{s}\right) > \hat{w}^{g}\hat{p}\left(\mathbf{y}_{\text{new}}^{\text{o}}|\widehat{\boldsymbol{\theta}}^{g}\right)$，則$\mathbf{y}_{\text{new}}^{\text{o}}$將被歸納到第$s$群。

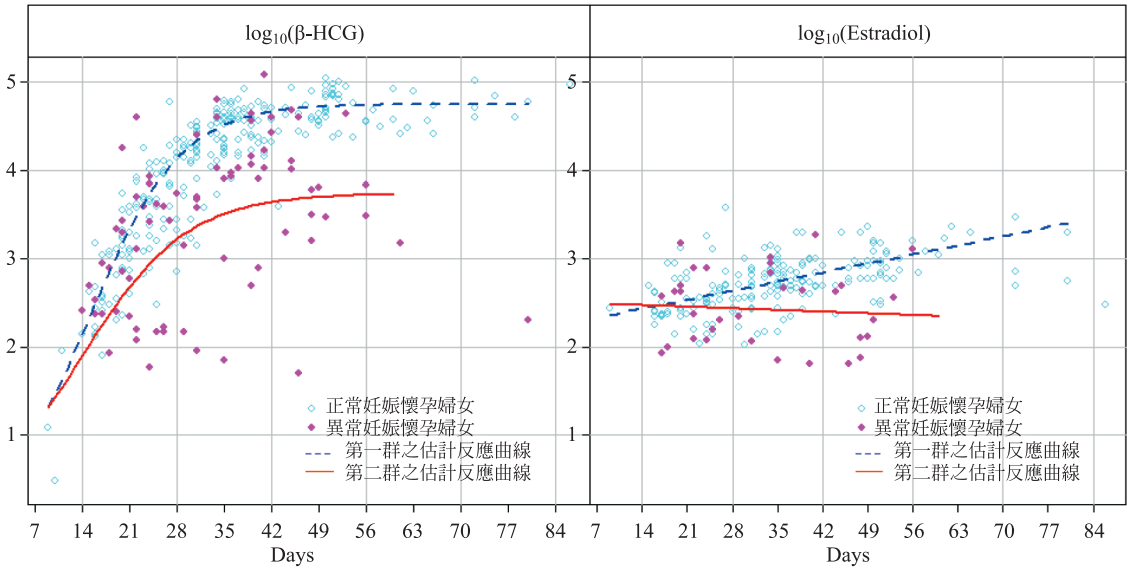
五、臨床試驗研究

我們將所提方法應於分析懷孕婦女的臨床試驗。定義y_{i1k}及y_{i2k}分別爲第i位婦女在第t_{ik}次產檢時所量測之$\log_{10}(\beta-\text{HCG})$及$\log_{10}(\text{Estradiol})$，其中$i = 1, \ldots, 161$，$k = 1, \ldots, s_i$。根據非監督式學習(unsupervised learning)的想法，我們配適$G = 1-4$群的 FM-MtNLMM 及其常態類似模型，簡稱 FM-MNLMM，其中y_{i1k}及y_{i2k}的平均反應曲線[7]設定如下：

$$y_{i1k} = \frac{\beta_{g1} + b_{i1,g}}{1 + \exp\{(\beta_{g2} + t_{ik})/\beta_{g3}\}} + e_{i1k,g} \ ,$$

$$y_{i2k} = \beta_{g4} + \beta_{g5}t_{ik} + b_{i2,g} + e_{i2k,g} \ ,$$

其中$\boldsymbol{\beta}_g = \left(\beta_{g1}, \beta_{g2}, \beta_{g3}, \beta_{g4}, \beta_{g5}\right)^{\text{T}}$爲固定效應，用以描述成份$g$的二維反應變量平均曲線；



圖三　正常與異常妊娠懷孕婦女之$\log_{10}(\beta - HCG)$及 $\log_{10}(\text{Estradiol})$觀測值及最佳模型下分群估計反應曲線

$(b_{i1,g}, b_{i2,g})^{\mathrm{T}} \sim t_2(\mathbf{0}, \mathbf{D}_g, \nu_g)$爲隨機效應，用以描述個體之追蹤曲線於成份$g$內與整體平均曲線的變異(variation)；而$(e_{i11,g}, \ldots, e_{i1s_i,g}, e_{i21,g}, \ldots, e_{i2s_i,g})^{\mathrm{T}} \sim t_{2s_i}(\mathbf{0}, \boldsymbol{\Sigma}_g \otimes \mathbf{I}_{s_i}, \nu_g)$表示第$g$個成份之個體內誤差項，其與隨機效應不相關。

爲了選擇最佳模型及最適群數，我們採用赤池訊息準則(Akaike information criterion; AIC [1])與貝氏訊息準則(Bayesian information criterion; BIC [11])。由表一所列 8 個候選模型之最大概似函數、AIC 及 BIC 結果得知基于 T 分佈之模型皆較常態模型佳，且 FM-MtNLMM (G = 2)爲最佳模型。進而，利用最佳模型之參數估計結果估計兩群的平均反應曲線，如圖三所示，正常與異常妊娠婦女之兩反應變數成長曲線確實存在差異，另在此模型下能達到 86.3%的分群正確率。

表一　懷孕婦女臨床試驗研究之模型配適選模準則結果

模型	準則	G=1	G=2	G=3	G=4
FM-MNLMM	$\ell_{\max}$	-274.64	-221.63	-208.85	-191.87
	AIC	571.27	489.27	487.69	477.73
	BIC	605.17	560.14	595.54	622.56
FM-MtNLMM	$\ell_{\max}$	-240.42	-206.59	-199.81	-191.46
	AIC	504.84	**463.19**	475.62	484.91
	BIC	541.81	**540.22**	592.72	642.06

備註：粗體字標示最佳模型

資料來源：取自 Wang (2019)[13]中的 Table 1

六、結語

本文所介紹的混合式多變量 T 非線性混合效應模型[13]及其模型參數估計方法能提供來處理分群或分類多元長期追蹤剖面成數個同質性的子群體，描述多元重複觀測值之「演變的相關性」與「相關性的演變」，並捕捉存在於資料中的厚尾現象等課題。同時，如何填補存在於多元長期追蹤資料中的遺失反應值及判別未來個體的群集等問題，亦能透過本文介紹的統計方法加以處理。然而，由於檢測儀器讀數的限制，反應變數可能產生左設限或右設限的效應，加上遺失資料可能呈現更複雜且不可忽略的非單調型態。因此，我們將持續發展新的統計方法來處理複雜長期資料的分析與推論工作。

參考文獻

[1] H. Akaike. Information theory and an extension of the maximum likelihood principle. In: Petrov BN and Csaki F (Eds.). Proc. 2nd Inter. Symp. Information Theory, Akademiai, Kiado, Budapest 267-281 (1973).

[2] T.W. Anderson. An Introduction to Multivariate Statistical Analysis, 3rd edition. Wiley and Sons, New York (2003).

[3] A.P. Dempster, N.M. Laird and D.B. Rubin. Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). J Roy Statist Soc Ser B 39, 1-38 (1977).

[4] N.M. Laird and J.H. Ware. Random effects models for longitudinal data. Biometrics 38, 963-974 (1982).

[5] R.J.A. Little and D.B. Rubin. Statistical Analysis with Missing Data, 2nd edition. Wiley, New York (2002).

[6] C. Liu and D.B. Rubin. The ECME algorithm: a simple extension of EM and ECM with faster monotone convergence. Biometrika 81, 633-648 (1994).

[7] G. Marshall, R. De la Cruz-Mesia, A.E. Baron, J.H. Rutledge and G.O. Zerbe. Non-linear random effects model for multivariate responses with missing data. Stat Med 25, 2817-2830 (2006).

[8] G.J. McLachlan and D. Peel. Finite Mixture Models. Wiley, New York (2000).

[9] X.L. Meng and D.B. Rubin. Maximum likelihood estimation via the ECM algorithm: a general framework. Biometrika 80, 267-278 (1993).

[10] X.L. Meng and D. van Dyk. The EM algorithm – an old folk-song sung to a fast new tune. J Roy Statist Soc Ser B 59, 511-567 (1997).

[11] G. Schwarz. Estimating the dimension of a model. Ann Statist 6, 461-464 (1978).

[12] G. Verbeke and E. Lesaffre. A linear mixed-effects model with heterogeneity in the random-effects population. J Am Stat Assoc 91, 217-221 (1996).

[13] W.L. Wang. Mixture of multivariate *t* nonlinear mixed models for multiple longitudinal data with heterogeneity and missing values. TEST 28, 196-222 (2019).

[14] W.L. Wang and T.I. Lin. Multivariate *t* nonlinear mixed-effects models for multi-outcome longitudinal data with missing values. Stat Med 33(17), 3029-3046 (2014).

[15] T. Yamashita, S. Okamoto, A. Thomas, V. MacLachlan and D.L. Healy. Predicting pregnancy outcome after in vitro fertilization and embryo transfer using estradiol, progesterone and human chorionic gonadotrophin β-subunit. Fertility and Sterility 51, 304-309 (1989).