

Des Données Longitudinales Et Des Mesures Répétées Dans Un Modèle A Effets Mixtes

[Study Of The Variance-Covariance Matrix For Longitudinal Data And Repeated Measures In A Mixed Effects Model]

A. Lanani

Labo. MAM, Département de mathématiques, Faculté des sciences exactes,
Université Frères Mentouri Constantine1.



Résumé -- Dans ce papier, nous nous sommes intéressés à l'estimation des différents paramètres d'un modèle linéaire mixte en particulier ceux de la matrice de variance-covariance. Le modèle utilisé, est appelé modèle à deux niveaux. C'est un modèle linéaire mixte à effets aléatoires. Il est intéressant car d'une part, les données ne sont pas forcément équilibrées; d'autre part, on a une modélisation explicite des effets intra et inter individus.

Mots clés -- EM algorithm; estimation; mesures répétées; modèles mixtes; vraisemblance.

Abstract -- In this paper, we are interested on the estimation of the various parameters of a mixed linear model, in particular those of the variance-covariance matrix. The used model is called two-step model. It is a linear mixed random effects model. We found that this model is interesting, firstly, because the data are not necessarily balanced; on the other hand, we have explicit expressions of the intra and inter individual effects.

I. INTRODUCTION

Le numéricien est souvent confronté à la résolution d'équations algébriques. Quand ces équations sont non linéaires, souvent elles n'admettent pas de solutions explicites. On essaye donc de trouver des solutions approximatives en appliquant des méthodes numériques. Ces solutions sont basées sur des procédés itératifs ou algorithmes. On peut citer la méthode de dichotomie; celle de Newton; du point fixe; etc... Nous allons considérer cette dernière, vue son importance; car il existe plus de cent variantes du théorème du point fixe. Dans ce papier, ce qui nous intéresse, ce n'est pas le côté analytique ou algébrique de ce théorème mais plutôt son application dans un contexte stochastique et qui est peu connu par rapport aux précédents. Dans la section II de ce papier, nous allons introduire le modèle utilisé. La section III est consacrée à l'estimation des différents paramètres du modèle. et ceux de la matrice de variance-covariance en appliquant l'algorithme EM. Les calculs sont développés dans la section IV. Enfin, on termine le papier par une conclusion.

II. MODÈLE

Dans un exemple d'application à une étude longitudinale, Laird et Ware [1] proposaient le modèle aléatoire à 2 niveaux suivant :

$$y_i = X_i \alpha + Z_i b_i + e_i \quad (2.1)$$

y_i vecteur des observations de dimensions n_i , α vecteur inconnu des paramètres de population de dimension p . X_i matrice de dimensions (n_i, p) du dispositif, connue liant y_i à α . Z_i matrice $(n_i; k)$ du dispositif, connue liant b_i à y_i . b_i vecteur inconnu des effets individus (aléatoires) de dimension k .

Etape 1: $\forall i, y_i = X_i \alpha + Z_i b_i + e_i$, e_i indépendantes $\rightarrow N(0, R_i)$; α, b_i fixes.

Etape 2: b_i indépendantes $\rightarrow N(0, D(k \times k))$, α fixe. Les y_i sont indépendantes et suivent $N(X_i \alpha, Z_i D Z_i^T + R_i)$. Soit θ , vecteur de dimension q et tel que : $R_i = R_i(\theta)$ et $D = D(\theta)$.

Pour l'estimation de α, b_i et θ , on a deux méthodes.

1^{ière} méthode: pour estimer α et θ , on a une approche basée sur le maximum de vraisemblance (MV), (Foulley [2]). Ayant $f(y/b; \alpha, \theta)$ dans la première étape et $f(b; \theta)$ dans la seconde; alors, on maximise la loi marginale de y , donnée par:

$$F(y; \theta, \alpha) = \int f(y/b; \alpha, \theta) f(b) db$$

Pour estimer b_i , on applique une méthode qui consiste à trouver un estimateur dont le MSE (Mean Squared Error) est le plus petit de la combinaison linéaire $\lambda_1' \alpha + \lambda_2' b$, avec $\lambda' \alpha$ supposé estimable (version du théorème de Gauss Markov étendue aux effets aléatoires, (Harville [3]; Harville [4])).

2^{ième} Méthode: Vu que les estimateurs des composantes de la variance sont biaisés dans la méthode du MV; on propose une alternative basée sur une formulation bayésienne. On estime θ en maximisant la vraisemblance marginale de y après avoir intégré sur α et b_i , $i = 1, \dots, m$. On se donne en plus une loi a priori pour α . On maximise donc :

$$f(y; \theta) = \iiint f(y/b, \alpha; \theta) f(b; \theta) f(\alpha) db d\alpha$$

(Approche considérée par Dempster et al, [5]). Cette méthode donne un estimateur de θ du MV restreinte (REML), on rappelle l'équivalence qu'on a entre l'estimateur REML et celui de Bayes (Foulley [2]). Les estimateurs de Bayes empiriques de α et de b_i sont les moyennes estimées des distributions conditionnelles.

III. ESTIMATION

3.1 Estimation de α et de b_i en supposant la variance connue

$$\text{Var } y_i = V_i = R_i + Z_i D Z_i^T, \quad (3.1)$$

$$\text{Soit } W = V_i^{-1}; \alpha = (\sum X_i^T W_i X_i)^{-1} (\sum X_i^T W_i y_i) \quad (3.2)$$

L'estimateur de α est celui qui maximise la vraisemblance basée sur les distributions marginales et il est à variance minimum. Pour sa part, b_i est donné par :

$$b_i = E(b_i / y_i, \alpha, \theta) \quad (3.3)$$

b_i n'est pas l'estimateur du MV mais l'estimateur de Bayes empirique :

$$b_i = DZ_i^T W_i (y_i - X_i \alpha) \quad (3.4)$$

3.2 Estimation de α et de b_i avec la variance inconnue

On estime la variance ou le θ qui la génère. Nous avons 2 estimateurs pour θ .

L'estimateur du MV noté θ_M et l'estimateur du MV restreinte noté θ_R . Pour calculer ces 2 estimateurs, on applique l'algorithme EM (Denoew [6] ; Dempster et al. [7]).

3.2.1 Utilisation du EM algorithme pour calculer l'estimateur du MV, θ_M

Les estimateurs des composantes de θ sont basés sur des formes quadratiques de b_i et e_i . Par exemple si $R_i = \sigma^2 I$ (I matrice identité); D matrice arbitraire (k, k) définie non négative, on utilisera la statistique $\sum e_i^T e_i$ pour estimer σ^2 et la statistique $\sum b_i b_i^T$ pour estimer D . On obtient les résultats suivants:

$$\sigma^2 = \sum e_i^T e_i / \sum n_i = t_1 / \sum n_i \quad (3.5)$$

$$D = \sum b_i b_i^T / m = t_2 / m \quad (3.6)$$

Soit θ un estimateur de θ , on calcule :

$$t_1 = E \{ \sum e_i^T e_i / y_i, \alpha(\theta), \theta \} = \sum [e_i(\theta)^T e_i(\theta) + \text{traceVar}(e_i / y_i; \alpha(\theta), \theta)] \quad (3.7)$$

$$t_2 = \sum [b_i(\theta) b_i(\theta)^T + \text{Var}(b_i / y_i; \alpha(\theta), \theta)] \quad (3.8)$$

$$\text{Où } e_i(\theta) = E(e_i / y, \alpha(\theta), \theta) = y_i - X_i \alpha(\theta) - Z_i b_i(\theta)$$

Pour obtenir l'estimateur du MV, θ_M , on commence par une valeur initiale convenable de θ et donc de $\alpha(\theta)$, ensuite on fait des itérations entre (3.7) et (3.8) définissant le E-Step et (3.5) et (3.6) définissant le M-Step. A la convergence, on a non seulement θ mais aussi $\alpha(\theta_M)$ et $b(\theta_M)$ du calcul du dernier E-Step.

3.2.2 Utilisation du EM algorithme pour calculer l'estimateur du MV restreinte REML

L'estimateur REML est noté par θ_R

Etape1 : $y_i = X_i \alpha + Z_i b_i + e_i$; α, b_i fixes.

Etape 2 : $\alpha \rightarrow N(0, \Gamma)$; $b_i \rightarrow N(0, D)$; $D = D(\theta)$; $\text{Cov}(\alpha, b_i) = 0$

$$y_i \rightarrow N(0, X_i \Gamma X_i^T + Z_i D Z_i^T + R_i); R_i = R_i(\theta)$$

Puisque généralement Γ ne peut pas être estimé car on n'a aucune information sur α , prenons $\Gamma^{-1} = 0$. Un estimateur de θ est obtenu en maximisant la vraisemblance marginale limite de θ quand $\Gamma^{-1} \rightarrow 0$ sachant y ; c'est-à-dire, on cherche le θ qui maximise :

$$f(y; \theta) = \lim \left\{ \int \int f(y/b, \alpha; \theta) f(b; \theta) db f(\alpha) d\alpha \right\} = \lim \left\{ \int f(y; \theta, \alpha) f(\alpha) d\alpha \right\} =$$

$$\lim \int f(y; \theta, \alpha) \exp \left\{ -(1/2) \alpha^T \Gamma^{-1} \alpha \right\} d\alpha$$

Harville [3], a montré que cette vraisemblance limite est équivalente à la vraisemblance restreinte.

Les estimateurs de Bayes empiriques sont:

$$E(\alpha / y, \Gamma^{-1} = 0, \theta_R) = \alpha(\theta_R)$$

$$E(b_i / y_i, \Gamma^{-1} = 0, \theta_R) = E(b_i / y_i; \alpha(\theta_R), \theta_R) = b_i(\theta_R)$$

Calcul de θ_R : application de l'algorithme EM.

M-Step: le même que dans a).

E-Step: analogue à a) avec α éliminé du calcul par intégration.

$$t_1 = E \left\{ \sum e_i^T e_i / y_i, \theta_R \right\} = \sum [e_i(\theta_R)^T e_i(\theta_R)] + \text{trace Var}(e_i / y_i; \theta_R)$$

$$t_2 = \sum [b_i(\theta_R) b_i(\theta_R)^T] + \text{Var}(b_i / y_i; \theta_R)$$

Où $e_i(\theta_R) = y_i - X_i \alpha(\theta_R) - Z_i b_i(\theta_R)$, (Stiratelli et al, [8]; Laird et al, [9]).

IV. CALCULS

En utilisant les notations de Laird et al. [9], on a:

$$\alpha^{(w)} = \left(\sum X_i^T W_i^{(w)} X_i \right)^{-1} \left(\sum X_i^T W_i^{(w)} y_i \right)$$

$$\text{Où } W_i^{(w)} = \left[\Sigma_i^{(w)} \right]^{-1} \text{ et } \Sigma_i^{(w)} = \sigma^{2(w)} I + Z_i D^{(w)} Z_i^T$$

$$b_i = D^{(w)} Z_i^T W_i^{(w)} r_i^{(w)}, \text{ où } r_i^{(w)} = y_i - X_i \alpha^{(w)}$$

$$\text{E-Step: } t_1^{(w)} = E \{ \sum e_i^T e_i / y_i, \alpha(\theta^{(w)}), (\theta^{(w)}) \}$$

$$= \sum [e_i(\theta^{(w)})^T e_i(\theta^{(w)})] + \text{traceVar}(e_i / y_i ; \alpha(\theta^{(w)}), \theta^{(w)})$$

$$\text{Où } e_i(\theta^{(w)}) = E(e_i / y, \alpha(\theta^{(w)}), \theta^{(w)}) = y_i - X_i \alpha(\theta^{(w)}) - Z_i b_i(\theta^{(w)})$$

$$t_2^{(w)} = E \{ \sum b_i b_i^T / y_i, \alpha(\theta^{(w)}), \theta^{(w)} \}$$

$$= \sum [b_i(\theta^{(w)}) b_i(\theta^{(w)})^T + \text{Var}(b_i / y_i ; \alpha(\theta^{(w)}), \theta^{(w)})$$

$$\text{M-Step: } \sum E [e_i^T e_i / y_i, \alpha(\theta^{(w+1)}), \theta^{(w+1)}] = t_1$$

$$= \sum [e_i(\theta^{(w+1)})^T e_i(\theta^{(w+1)})] + \text{traceVar}(e_i / y_i ; \alpha(\theta^{(w+1)}), \theta^{(w+1)})$$

$$= \sum n_i (\sigma^2)^{(w+1)} = t_1^{(w)} \Rightarrow (\sigma^2)^{(w+1)} = t_1^{(w)} / \sum n_i = t_1^{(w)} / N; \text{ où } N = \sum n_i$$

D'où :

$$(\sigma^2)^{(w+1)} = \{ \sum \{ [r_i^{(w)} - Z_i b_i^{(w)}]^T [r_i^{(w)} - Z_i b_i^{(w)}]$$

$$+ (\sigma^2)^{(w)} \text{trace}(I - (\sigma^2)^{(w)} W_i^{(w)}) \} / N$$

$$\text{et } D^{(w+1)} = \sum [b_i^{(w)} b_i^{(w)T} + \text{Var}(b_i / y_i)$$

$$= \sum [b_i^{(w)} b_i^{(w)T} + D^{(w)} (I - Z_i^T W_i^{(w)} Z_i D^{(w)})] / m$$

Pour l'estimateur REML, on a :

$$(\sigma^2)^{(w+1)} = \{ \sum \{ [r_i - Z_i b_i^{(w)}]^T [r_i - Z_i b_i^{(w)}]$$

$$+ (\sigma^2)^{(w)} \text{trace}(I - (\sigma^2)^{(w)} p_i^{(w)}) \} / N.$$

$$\text{et } D^{(w+1)} = \sum [b_i^{(w)} b_i^{(w)T} + D^{(w)} (I - Z_i^T p_i^{(w)} Z_i D^{(w)})] / m.$$

Où :

$$p_i = W_i^{(w)} (I - X_i (\sum X_i^T W_i^{(w)} X_i)^{-1} X_i^T W_i^{(w)})$$

Remarque: Suivant Cook, [10], des valeurs initiales acceptables sont les estimateurs des moindres carrés ordinaires qui sont données par :

$$\alpha_0 = (\sum X_i^T X_i)^{-1} (\sum X_i^T y_i)$$

$$b_i = (Z_i^T Z_i)^{-1} Z_i^T (y_i - X_i \alpha_0)$$

V. CONCLUSION

Dans ce travail, nous avons calculé les estimateurs des coefficients de régression ainsi que ceux générant la matrice de variance-covariance pour un modèle linéaire mixte en appliquant l'algorithme EM; cependant tous ces calculs ont été faits moyennant l'hypothèse de normalité des observations. La question que nous nous posons est : quelles alternatives faut-il proposer et à quel(s) prix, si on suppose que les données ne sont pas gaussiennes?

RÉFÉRENCES

- [1] Laird, N.M. & Ware, J.H. : " Random effects models for longitudinal data " . Biometrics. 1982, 38, pp. 963-974.
- [2] Foulley, J.L. : " Estimation des composantes de la variance en modèle linéaire " DEA de stat. 1991/1992 Paris 11 INSERM.INRA.78352 Jouy en JOSAS.
- [3] Harville, D.A.: "Extension of the Gauss-Markov theorem to include the estimation of random effects ". Ann. Stat.1976, V4, pp. 384-395.
- [4] Harville, D.A. : " Maximum likelihood approaches to variances components estimation and to related problems ". JASA, jun. (1977), 72, N358, pp. 320-329.
- [5] Dempster, A.P; Rubin, D.B. & Tsutakawa, R.K. " Estimation covariance component model " JASA (1981), 76, pp. 341-353.
- [6] Thierry Denoeux. Maximum likelihood estimation from fuzzy data using the EM algorithm. Fuzzy Sets and Systems, Elsevier, 2011, 183 (1), pp.72-91. fhal-00654118f
- [7] Dempster, A.P; Laird, N.M. & Rubin, D.B. " Maximum likelihood with incomplète data via the EM algorithme ". JRSS, B39, pp. 1-38.
- [8] Stiratelli, R.; Laird, N.M. & Ware, J.H. : " Random effects models for serial observations ". Biometrics, Dec.1984, 40, pp. 961-971.
- [9] Laird, N.M. ; Lange, N. & Stram, D. : " Maximum likelihood computations with repeated measures : application of the EM algorithme ". JASA, MAR. 1987, V82, N 397, pp. 97-105.
- [10] Cook, N : " A fortran programm for random effects models ". Technical Report. Harvard School of public health; dept of biostat. 1982.