Une approche par vraisemblance hiérarchique pour les modèles dynamiques appliqués au VIH

Danaëlle Jolly, Daniel Commenges

Une approche par vraisemblance hiérarchique pour les modèles dynamiques appliqués au VIH

Danaëlle JOLLY & Daniel COMMENGES

INSERM U897, Bordeaux F-33076, France
IMB, Université de Bordeaux, Talence F-33405, France

Résumé : L'évolution dynamique du VIH se modélise par des systèmes non linéaires d'équations différentielles ordinaires n'ayant pas de solution analytique. Dès lors que l'on insére des effets aléatoires, l'estimation de tels modèles non linéaires à effets mixtes par maximum de vraisemblance peut s'avérer, en raison du calcul numérique d'intégrales multiples, très coûteuse en terme de temps.
Nous proposons d'adapter la vraisemblance hiérarchique, développée par Lee & Nelder (1996), à ce type de problèmes. L'idée principale est d'estimer directement les paramètres individuels plutôt que les variances des effets aléatoires dont l'ordre de grandeur est supposé connu : un grand nombre de paramètres sont donc à estimer. Pour calculer les estimateurs du maximum de vraisemblance hiérarchique, un algorithme performant basé sur l'algorithme de Marquardt a été développé. Les effets fixes ainsi estimés ont la particularité d'être biaisés. Une méthode a été proposé pour réduire ce biais et le rendre négligeable. Enfin cette approche par vraisemblance hiérarchique a été évaluée grâce à plusieurs simulations et a également été appliquée sur un essai clinique.

Mots clés : modèle VIH, équations différentielles ordinaires, modèle non linéaire à effets mixtes, vraisemblance hiérarchique, algorithme, biais.

Abstract : HIV dynamical models are based on nonlinear systems of ordinary differential equations, which do not have analytical solution. Introducing random effects in such models leads to non-linear mixed-effects models whose the estimations is very time-consuming owing to the numerical computation of multiple integrals.
We propose to adapt the hierarchical likelihood developed by Lee & Nelder (1996) to these problems. This approach is focused on the estimations of individual parameters rather than random effect variances which are supposed to be approximately known : we have to estimate a large number of parameters. To compute the maximum hierarchical likelihood estimators, we developed a special and efficient algorithm based on the Marquardt algorithm. The fixed effects estimated by this method have a bias which can be corrected. We apply our hierarchical likelihood approach, first to simulations in order to validate it, second to a clinical trial.

Keywords : HIV model, differential equations, nonlinear mixed effects model, hierarchical likelihood, algorithm, bias.

1 Introduction

De nombreux efforts ont été faits pour construire des modèles mathématiques traduisant le plus fidèlement les principaux mécanismes de l'interaction entre le virus du VIH et le système immunitaire.
Plusieurs auteurs (Ho & Al. (1995), Perelson & Al. (1996)) ont utilisé des équations différentielles ordinaires (EDO) pour décrire l'évolution de plusieurs populations : les lymphocytes $T\ CD_4^+$ (infectés ou non) et le virus. Des estimations pour chaque individu ont été obtenues en utilisant des méthodes simples de régression non linéaire. Leurs travaux ont permis de démontrer que la réplication virale ainsi que la destruction des $T\ CD_4^+$ étaient des phénomènes très rapides. Cependant, comme le soulignait Le Corfec & Al. (2000), les modèles utilisés étaient basés sur une hypothèse très forte : l'efficacité parfaite du traitement alors qu'il est très intéressant d'avoir une estimation de ce paramètre.
D'autres auteurs ont également utilisé des méthodes plus complexes d'estimation dans l'esprit des modèles non linéaires à effets mixtes. En particulier, Guedj & Al. (2007) ont proposé un modèle mathématique basé sur un système de cinq EDO. En utilisant des données d'un essai clinique de mise sous traitement et par maximisation de la vraisemblance, les auteurs ont estimé certains paramètres dont l'efficacité des traitements. Cependant, il est nécessaire de calculer une intégrale multiple dont la dimension est égale au nombre d'effets aléatoires. De plus, le système d'EDO n'ayant pas de solution explicite, cette méthode peut s'avérer très lourde.

2 La vraisemblance hiérarchique

Nous proposons d'éviter le calcul numérique de cette intégrale multiple en utilisant la vraisemblance hiérarchique développée par Lee & Nelder (1996). L'idée principale est que, plutôt que d'estimer la variance des effets aléatoires, les paramètres individuels sont directement estimés. Ainsi, plutôt que de calculer une intégrale multiple, nous devons estimer un très grand nombre de paramètres.
Considérons le modèle suivant :
La j-ième observation de l'individu i : $Y_{ij} = g_{ij}(a_i, \beta) + \varepsilon_{ij}$,
ε_{ij} représentant les erreurs de mesure supposées iid et suivant une loi normale de moyenne 0 et de variance σ^2. Le vecteur de dimension R (R étant le nombre d'effets aléatoires) $a_i = (a_{i1}, ..., a_{iR})^T$ représente les paramètres de l'individu i. Nous supposons que les a_{ir} sont normalement distribués de moyenne α_r et de variance τ_r^2. Deux types d'effets fixes sont donc à traiter : $\alpha = (\alpha_1, ..., \alpha_R)^T$ où α_r représente l'espérance des paramètres individuels a_{ir} et $\beta = (\beta_1, ..., \beta_{q-R})^T$. Appelons $\theta = [\alpha, \beta]^T$ le vecteur de dimension q regroupant

tous les effets fixes.
Conditionnellement à a_i, le vecteur d'observations de l'individu i, Y_i, a une densité $f(.; a_i, \theta)$. De plus, si l'on suppose que a_i a une densité $f_a(.; \tau)$ où τ est un vecteur connu de paramètres, alors la vraisemblance hiérarchique est donnée par :

$$hL(\theta, a) = \prod_{i=1}^{n} f(Y_i; a_i, \theta) f_a(a_i; \tau)$$

Le principal inconvénient de cette méthode est de supposer la variance des effets aléatoires τ^2 connue mais nous verrons comment contourner ce problème. La fonction à maximiser, qui correspond au logarithme de la fonction précédente est donc donnée par :

$$hl(\theta, a) = -\sum_{i,j} \frac{(Y_{ij} - g_{ij}(a_i, \beta))^2}{2\sigma^2} - \sum_{i,r} \frac{(a_{ir} - \alpha_r)^2}{2\tau_r^2}$$

Nous remarquons que la fonction hl comporte une partie moindre carré et une partie pénalisation qui est d'autant plus grande que les paramètres individuels sont dispersés.

3 L'algorithme hybride

Pour maximiser cette fonction un algorithme de type Newton-raphson, l'algorithme de Marquardt est utilisé. Il nécessite le calcul numérique des scores et de la Hessienne, matrice qui peut être très grande. Par exemple, si le nombre de sujets $n = 100$, si le nombre d'effets aléatoires $R = 3$ et si le nombre d'effets fixes $q = 6$ alors 306 paramètres doivent être estimés, la Hessienne étant de taille 306×306. Cependant, une étude plus approfondie de sa structure montre qu'elle est aussi très creuse et qu'elle est proche d'une matrice diagonale par blocs :

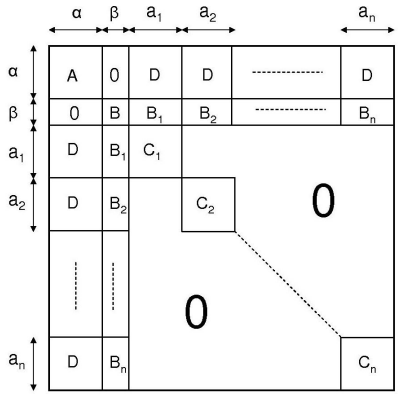

FIGURE 1 – Matrice Hessienne de la hl.

Les blocs A et D sont diagonaux. On appelle les blocs C_i « bloc individuel du sujet i » car

ils correspondent à la dérivée seconde de la hl par rapport aux paramètres individuels a_i du sujet i. Si l'un de ces blocs n'est pas défini positif alors la hessienne n'est pas définie positive. L'algorithme de Marquardt s'adapte à cette situation en gonflant la diagonale de la matrice alors qu'il suffirait de gonfler la diagonale des sous blocs non définis positifs. Ces remarques nous ont amenés à développer un algorithme « hybride »qui tient compte de la structure très particulière de la hessienne. Cet algorithme procède en deux temps : une première étape dite « patient par patient »qui a pour but de rendre la hessienne définie positive en traitant les blocs individuels C_i et les blocs A et B séparemment. Lorsque tous les blocs sont définis positifs, la seconde étape consiste à utiliser la matrice globale.

4 La correction du biais

A τ fixé et en maximisant la fonction hl grâce à cet algorithme hybride nous obtenons les estimateurs du maximum de la vraisemblance hiérarchique $\hat{\theta}^\tau, \hat{a}^\tau$. Nous nous sommes intéressés aux propriétés des estimateurs des effets fixes $\hat{\theta}^\tau$ et nous avons montré que asymptotiquement $\hat{\theta}^\tau$ tend vers θ_0^τ. Cependant, θ_0^τ est en général différent de θ^*, le vecteur correspondant aux vraies valeurs des effets fixes. Il existe donc un biais asymptotique $\theta_0^\tau - \theta^*$ qui doit être corrigé. Nous avons proposé une méthode pour la correction du biais basée sur le boostrap paramétrique. Après avoir estimé θ^* par $\hat{\theta}^\tau$, l'idée consiste à simuler pour $s = 1, .., S$ des données à partir de $\hat{\theta}^\tau$ et à en obtenir des estimateurs $\hat{\hat{\theta}}_s^\tau$ par la procédure habituelle. Un estimateur du biais peut être calculé par :

$$\widehat{biais} = S^{-1} \sum_{s=1}^{S} (\hat{\hat{\theta}}_s^\tau - \hat{\theta}^\tau)$$

et l'estimateur corrigé est alors donné par $\hat{\theta}_c^\tau = \hat{\theta}^\tau - \widehat{biais}$. Il est à noter que cette correction augmente la variance de l'estimateur puisque $Var\ \hat{\theta}_c^\tau \simeq (1 + S^{-1}) Var\ \hat{\theta}^\tau$. L'efficacité de l'algorithme hybride et de la correction du biais ont été vérifiées grâce à un certain nombre de simulations.

5 Applications

Le modèle le plus classique (Perelson & Al. (1996), Nowak & May (2001)) a été utilisé pour l'application. Trois populations sont modélisées : les $T\ CD_4^+$ non infectés (T), les $T\ CD_4^+$ infectés (I) et le virus (V). Les $T\ CD_4^+$ non infectés sont produits par le thymus à un taux constant λ. Ils peuvent être infectés par le virus à un taux γ qui peut être diminué par la présence d'un traitement dont l'efficacité est notée η. Ces $T\ CD_4^+$ infectés produisent du virus à un taux π. Enfin ces trois populations ont des taux de décès : $\mu_T,\ \mu_I,\ \mu_V$. Ces phénomènes peuvent être traduits en terme d'EDO pour chaque individu i :

$$\frac{dT^i}{dt} = \lambda^i - (1 - \eta^i)\gamma^i V^i T^i - \mu_T^i T^i$$

$$\begin{aligned}\frac{dI^i}{dt} &= (1-\eta^i)\gamma^i V^i T^i - \mu_I^i I^i \\ \frac{dV^i}{dt} &= \pi^i I^i - \mu_V^i V^i\end{aligned}$$

Ce modèle est complété par un modèle statistique et par un modèle d'observations décrits ci-dessous.

5.1 Simulations

Pour les simulations nous avons supposé que les $T\ CD_4^+$ totaux (T+I), les $T\ CD_4^+$ infectés (I) ainsi que le virus (V) étaient mesurés à dix temps. Le modèle non linéaire mixte $Y_{ijk} = g_{ijk}(a_i, \beta) + \varepsilon_{ijk}$ s'écrit :

$$\left\{\begin{array}{ll} Y_{ij1} & = [T^i(t_{ij}) + I^i(t_{ij})]^{0.25} + \varepsilon_{ij1} \\ Y_{ij2} & = [I^i(t_{ij})]^{0.25} + \varepsilon_{ij2} \\ Y_{ij3} & = \log_{10}[V^i(t_{ij})] + \varepsilon_{ij3} \end{array}\right.$$

Les ε_{ijk} sont supposés normalement distribués de moyenne nulle et de variance $\sigma_k^2 = 0.5^2$. Nous avons simulé les données de 100 patients, la moitié suivant le traitement A (d'efficacité η_A), l'autre moitié suivant le traitement B (d'efficacité η_B). Nous avons ensuite choisi d'estimer le logarithme (désigné par le tilde) de certains paramètres : $\theta = (\tilde{\lambda},\ \tilde{\mu}_I,\ \tilde{\pi},\ \tilde{\gamma},\ \tilde{\eta}_A,\ \tilde{\eta}_B)$, les autres étant fixés à des valeurs de la littérature. Les effets aléatoires portent sur les trois premiers paramètres ($a_i = (\tilde{\lambda}^i,\ \tilde{\mu}_I^i,\ \tilde{\pi}^i)$).
Une comparaison entre l'algorithme hybride et l'algorithme global (qui consiste à utiliser la hessienne dans son ensemble dès le départ) prouve que le premier est beaucoup plus performant que le second. En considérant comme échec toute simulation n'ayant pas convergé après 150 itérations, l'algorithme global ne réussit que dans 49% des simulations et converge en moyenne en 71 itérations tandis que l'algorithme hybride réussit dans 94% des simulations et met en moyenne 25 itérations. L'algorithme hybride est donc plus rapide et plus efficace.
Nous avons également étudié l'efficacité de la correction du biais. Le biais des estimateurs non corrigés des effets fixes est de l'ordre de 10^{-2} pour tous les paramètres et la correction réduit ce biais qui devient de l'ordre 10^{-3} et qui semble alors négligeable.
Fixer la variance des effets aléatoires τ^2 est un inconvénient majeur. En effet, le vecteur des vraies variances τ^* n'est en général pas connu. Cependant, en prenant une borne supérieure « raisonnable », ce problème peut être contourné. Nous avons étudié deux cas : le premier pour lequel $\tau_\lambda^* = \tau_{\mu_T}^* = \tau_{\mu_I}^* = 0.2$ et le second pour lequel $\tau_\lambda^* = 0.1,\ \tau_{\mu_T}^* = 0.2,\ \tau_{\mu_I}^* = 0.3$. En calculant le RMSE (root mean square error) ainsi que les taux de couverture à .95 des estimateurs des effets fixes, les résultats basés sur ces deux critères étaient très satisfaisants en prenant $\tau = 0.3$.

5.2 L'essai clinique ALBI ANRS 070

Nous avons appliqué cette méthode sur les données d'un essai clinique comparant sur 24 semaines deux traitements anti-rétroviraux : ALBI ANRS 070. Chaque groupe de traitement comprenait 51 sujets ayant en moyenne 7 données de charge virale et de $T\,CD_4^+$ totaux. Les données étant moins riches que dans l'étude des simulations, $\tilde{\gamma}$ n'a pas pu être estimé. Nous avons estimé les autres paramètres en prenant $\tau = 0.3$. Les résultats de nos estimations ont montré que l'un des deux traitements est plus efficace que l'autre, résultats en accord avec Molina & Al (1999).

Bibliographie

[1] Le Corfec, E., Rouzioux, C., and Costagliola, D. (2000) *Dynamique Quantitative du VIH- 1 in Vivo : Revue des Modèles Mathématiques.* Revue d'épidémiologie et de santé publique, 48(2) :168–181.

[2] Guedj, J., Thiebaut, R., and Commenges, D. (2007) *Maximum Likelihood Estimation in Dynamical Models of HIV.* Biometrics, 63 : 1198–1206.

[3] Ho, D.D., Neumann, A.U., Perelson, A.S., Chen, W., Leonard, J.M., and Markowitz M. (1995) *Rapid Turnover of Plasma Virions and CD4 Lymphocytes in HIV-1 Infection.* Nature, 373 :123–126.

[4] Lee, Y. and Nelder, J.A. (1996) *Hierarchical Generalized Linear Models.* Journal of the Royal Statistical Society. Series B (Methodological), 58 : 619–678.

[5] Molina, J., Chêne, G., Ferchal, F., Journot, V., Pellegrin, I., Sombardier, M. N., Rancinan, C., Cotte, L., Madelaine, I., Debord, T., and Decazes, J. M. (1999) *The ALBI Trial : A randomized controlled trial comparing stavudine plus didanosine with zidovudine plus lamivudine and a regimen alternating both combinations in previously untreated patients infected with human immunodeficiency virus.* The Journal of Infectious Diseases 180, 351–358.

[6] Nowak, M. and Nowak, MA and May, R. (2001) *Virus Dynamics : Mathematical Principles of Immunology and Virology* Oxford.

[7] Perelson, A. S., Neumann, A. U., Markowitz, M., Leonard, J. M., and Ho, D. (1996) *Viral dynamics in human immunodeficiency virus type 1 infection.* Science 271, 1582–1586.