Test du rapport de vraisemblance pour des composantes de la variance dans les modèles non linéaires mixtes

Charlotte Baey [1] & Estelle Kuhn [2] & Paul-Henry Cournède [1]

[1] *Laboratoire MICS, CentraleSupelec, Grande Voie des Vignes, 92290 Châtenay-Malabry, charlotte.baey@centralesupelec.fr, paul-henry.cournede@centralesupelec.fr*
[2] *INRA UNITÉ MaIAGE, Domaine de Vilvert, 78352 Jouy-en-Josas, estelle.kuhn@inra.fr*

Résumé. Les modèles non linéaires mixtes sont utilisés dans un grand nombre d'applications, afin de prendre en compte la variabilité inter- et intra- individuelle dans une population. L'une des questions qui se pose naturellement lorsque l'on ajuste un tel modèle paramétrique est celle de l'identification des paramètres pouvant être considérés comme constants dans la population (les "effets fixes") et ceux qui varient d'un individu à l'autre (les "effets aléatoires"). D'un point de vue statistique, ce problème peut se formuler sous la forme d'un test d'hypothèses, dans lequel on teste si les variances d'un sous ensemble d'effets aléatoires sont nulles, et peut se traiter par un test du rapport de vraisemblance (TRV). Le TRV peut être mis en œuvre mais les résultats standards sur ce test ne s'appliquent pas ici, car sous l'hypothèse nulle, la vraie valeur du paramètre se trouve sur la frontière de l'espace des paramètres. Cette question a été abordée par plusieurs auteurs dans le cas des modèles linéaires mixtes et dans quelques cas particuliers, et est liée plus généralement à l'inférence et aux tests d'hypothèses sous contraintes. Nous montrons que la distribution asymptotique de la statistique de test de rapport de vraisemblance est un mélange de lois du chi-deux, dont les poids dépendent de la matrice d'information de Fisher et du nombre de paramètres impliqués dans le test. Nous montrons en particulier que la loi limite dépend de la présence ou non de corrélations entre les effets aléatoires. Nous présentons des résultats sur données simulées et réelles.

Mots-clés. Loi du chi-bar-square, composantes de la variance, modèles non linéaires mixtes, test d'hypothèses

Abstract. Mixed effects models are widely used to describe inter and intra individual variabilities in a population. A fundamental question when adjusting such a model to the population consists in identifying the parameters carrying the different types of variabilities, i.e. those that can be considered constant in the population, referred to as fixed effects, and those that vary among individuals, referred to as random effects.

In this talk, we propose a test procedure based on the likelihood ratio one for testing if the variances of a subset of the random effects are equal to zero. The standard theoretical results on the asymptotic distribution of the likelihood ratio test can not be applied in

our context. Indeed the assumptions required are not fulfilled since the tested parameter values are on the boundary of the parameter space.

The issue of variance components testing has been addressed in the context of linear mixed effects models by several authors and in the particular case of testing the variance of one single random effect in nonlinear mixed effects models. We address the case of testing that the variances of a subset of the random effects are equal to zero. We proof that the asymptotic distribution of the test is a chi bar square distribution, indeed a mixture of chi square distributions, and identify the weights of the mixture. We highlight that the limit distribution depends on the presence or not of correlations between the random effects. We illustrate the finite sample size properties of the test procedure through simulation studies.

Keywords. Chi-bar-square distribution, hypothesis testing, nonlinear mixed models, variance components

1 Introduction

Les modèles à effets mixtes sont très largement utilisés dans un grand nombre de domaines d'applications, afin de prendre en compte et de décrire la variabilité inter-individuelle d'une population [3].

Du point de vue du modélisateur, l'une des questions fondamentales qui se pose dans ce contexte, est celle de l'identification des paramètres pouvant être considérés comme constants dans la population, et ceux qui varient d'un individu à l'autre, afin d'identifier les différentes sources de variabilité. Du point de vue du statisticien, cette question peut se formuler de la façon suivante : si l'on considère un ensemble d'effets aléatoires dans un modèle à effets mixtes, quels sont ceux dont la variance est nulle ? Autrement dit, on cherche à tester l'hypothèse nulle selon laquelle les variances d'un sous-ensemble d'effets aléatoires du modèle sont nulles.

Sous l'hypothèse nulle, la vraie valeur du paramètre est sur la frontière de l'espace des paramètres, et les résultats standards de la théorie des tests du rapport de vraisemblance ne s'appliquent pas. Plusieurs auteurs se sont penchés sur cette question, que ce soit dans le cadre des modèles linéaires mixtes [8], ou plus généralement dans le cadre des tests du rapport de vraisemblance sous contraintes (voir par exemple [1, 5], ou [7] pour une revue plus récente).

Dans cet article, nous proposons un test sur les composantes de la variance pour tester si un sous-ensemble d'effets aléatoires peuvent être considérés comme des effets fixes. Nous étudions la loi asymptotique de la statistique de test, et nous montrons en particulier que celle-ci est un mélange de lois du chi-deux, les poids intervenant dans le mélange dépendant à la fois de la matrice d'information de Fisher et du nombre de paramètres impliqués dans le test. Nous montrons également que la loi limite dépend de la présence ou non de corrélations entre les effets aléatoires. Nous présentons des outils

numériques pour calculer les quantiles de la loi limite, et nous illustrons les propriétés du test à distance finie sur des données simulées et réelles.

2 Modèles non linéaires mixtes

Nous considérons le modèle non linéaire à effets mixtes suivant :

$$\begin{aligned} y_{ij} &= g(\phi_i, x_{ij}) + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \\ \phi_i &\sim \mathcal{N}_p(\beta, \Gamma). \end{aligned}$$

où y_{ij} est l'observation j de l'individu i, g une fonction non linéaire, ϕ_i le vecteur des effets aléatoires de l'individu i, et x_{ij} est un vecteur de covariables. Les vecteurs (ϕ_i) et $(\varepsilon_{ij})_{1\leq i\leq N,\ 1\leq j\leq n_i}$ sont supposés indépendants, et les suites (ε_{ij}) et (ϕ_i) sont supposées mutuellements indépendantes.

Le vecteur de paramètres du modèle est $\theta = (\beta, \Gamma, \sigma)$, et l'espace des paramètres associé est $\Theta = \mathbb{R}^p \times \mathbb{S}_+^p \times \mathbb{R}_+$, où $\mathbb{S}_+^p$ est l'ensemble des matrices symétriques semi-définies positives de taille $p \times p$.

3 Test sur les composantes de la variance

3.1 Test du rapport de vraisemblance

On note q le nombre total de paramètres dans le modèle, et r le nombre de variances dont on souhaite tester si elles sont nulles. Nous considérons alors le test suivant :

$$H_0 : \theta \in \Theta_0 \quad \text{vs.} \quad \theta \in \Theta_1, \tag{1}$$

où, en notant $\Gamma = \left[\begin{array}{c|c} \Gamma_1 & \Gamma_{12} \\ \hline \Gamma_{12}^t & \Gamma_2 \end{array}\right] \begin{matrix} \} \ p-r \\ \} \ r \end{matrix}$, on a :

$$\begin{aligned} \Theta_0 &= \{\theta \in \mathbb{R}^q \mid \beta \in \mathbb{R}^p, \Gamma_1 \in \mathbb{S}_+^{p-r}(\mathbb{R}), \Gamma_{12} = 0, \Gamma_2 = 0, \sigma \geq 0\} \\ \Theta_1 &= \{\theta \in \mathbb{R}^q \mid \beta \in \mathbb{R}^p, \Gamma \in \mathbb{S}_+^p(\mathbb{R}), \sigma \geq 0\}. \end{aligned}$$

En notant $\ell_N(\theta)$ la vraisemblance des observations, on définit la statistique du test du rapport de vraisemblance par :

$$LRT_N := -2 \ \log\left(\frac{\sup_{\theta\in\Theta_0} \ell_N(\theta)}{\sup_{\theta\in\Theta_1} \ell_N(\theta)}\right). \tag{2}$$

3.2 Loi du chi-bar-square

Soit $\mathcal{C}$ un cône convexe fermé de $\mathbb{R}^p$. Soit V une matrice semi-définie positive de taille $p \times p$ et soit $Z \sim \mathcal{N}(0, V)$. La loi de la variable aléatoire $\bar{\chi}^2(V, \mathcal{C})$, définie par :

$$\bar{\chi}^2(V, \mathcal{C}) := Z'V^{-1}Z - \min_{\theta \in \mathcal{C}}(Z - \theta)'V^{-1}(Z - \theta). \tag{3}$$

est appelée **loi du chi-bar-square**. C'est un mélange de lois du chi-deux à différents degrés de liberté, avec :

$$\forall t \geq 0 \; P(\bar{\chi}^2(V, \mathcal{C}) \leq t) = \sum_{i=0}^{p} w_i(p, V, \mathcal{C}) P(\chi_i^2 \leq t),$$

où les $w_i(p, V, \mathcal{C})$ sont des réels positifs dont la somme vaut 1, et où χ_i^2 représente la loi du chi-deux à i degrés de libertés, avec la convention $\chi_0^2 \equiv 0$.

3.3 Loi limite et cône d'approximation

On suppose que l'hypothèse nulle est vérifiée, et donc que que la vraie valeur des paramètres θ_0 est dans Θ_0, et on pose $\theta_0 = (\beta_0, \Gamma_0, \sigma_0)$, avec $\Gamma_0 = \left[\begin{array}{c|c} \Gamma_{0,1} & 0 \\ \hline 0 & 0 \end{array}\right]$. On suppose de plus que $\Gamma_{0,1} > 0$.

Soit I_0 la matrice d'information de Fisher évaluée en θ_0. Alors, on a [7] :

$$LRT_N = \xrightarrow[n \to \infty]{d} D_T(Z),$$

$$D_T(z) = \|z - T(\Theta_0, \theta_0)\|^2_{I(\theta_0)^{-1}} - \|z - T(\Theta_1, \theta_0)\|^2_{I(\theta_0)^{-1}}$$

où $Z \sim \mathcal{N}(0, I^{-1}(\theta_0))$, $T(\Theta, \theta)$ est le cône tangent de Θ en θ, et $\| \cdot \|_V$ est la norme induite par le produit scalaire $\langle x, y \rangle_V = x^t V^{-1} y$.

Alors, en supposant que la matrice de covariance des effets aléatoires est non diagonale, on peut montrer que l'on a :

$$T(\Theta_0, \theta_0) = \mathbb{R}^p \times \mathbb{R}^{\frac{(p-r)(p-r+1)}{2}} \times \{0\}^{\frac{r(r+1)}{2} + r(p-r)} \times \mathbb{R} \tag{4}$$

$$T(\Theta_1, \theta_0) = \mathbb{R}^p \times \mathbb{R}^{\frac{(p-r)(p-r+1)}{2}} \times \mathbb{R}^{r(p-r)} \times \mathbb{S}_+(\mathbb{R})^r \times \mathbb{R}. \tag{5}$$

Comme $T(\Theta_0, \theta_0)$ est un espace vectoriel, et $T(\Theta_1, \theta_0)$ est un cône convexe fermé, alors la loi de $D(Z)$ est une loi du chi-bar-square $\bar{\chi}^2(I^{-1}(\theta_0), T(\Theta_1, \theta_0) \cap T(\Theta_0, \theta_0)^{\perp})$, c'est-à-dire un mélange de lois du chi-deux dont les degrés de libertés varient entre 0 et q.

En utilisant les propriétés de la loi du chi-bar-square (voir par exemple [6]), on peut identifier certains poids comme nuls, et on obtient finalement un mélange de $\frac{r(r+1)}{2} + 1$ lois du chi-deux dont les degrés de libertés sont compris entre $r(p-r)$ et $r(p-r) + \frac{r(r+1)}{2}$.

On remarque donc que la loi limite dépend de la matrice d'information de Fisher en θ_0 à travers les poids intervenant dans le mélange de chi-deux, mais également de la présence ou non de corrélations entre les effets aléatoires. En effet, les cônes d'approximation définis en (4) ont été obtenus en supposant une matrice de covariance Γ non diagonale.

Si on suppose que Γ est diagonale, les deux espaces de paramètres Θ_0 et Θ_1 se simplifient, et alors leurs cônes d'approximation respectifs deviennent :

$$T(\Theta_0, \theta_0) = \mathbb{R}^p \times \mathbb{R}^{p-r} \times \{0\}^r \times \mathbb{R}$$
$$T(\Theta_1, \theta_0) = \mathbb{R}^p \times \mathbb{R}^{p-r} \times \mathbb{R}_+^r \times \mathbb{R}.$$

On obtient alors pour la loi limite un mélange de $r+1$ lois du chi-deux dont les degrés de liberté varient entre 0 et r.

4 Implémentation

D'un point de vue pratique, le test proposé nécessite le calcul de la statistique de test, c'est-à-dire de la vraisemblance des observations sous les hypothèses nulle et alternative, de la matrice d'information de Fisher, et des quantiles de la loi du chi-bar-square associée.

Dans le cas des modèles non linéaires mixtes, la vraisemblance n'est pas calculable explicitement, et on a recours tout d'abord à un algorithme de type MCMC-SAEM pour obtenir les estimateurs du maximum de vraisemblance sous les deux hypothèses $\hat{\theta}_0$ et $\hat{\theta}_1$, puis à un échantillonnage d'importance pour calculer la valeur de la statistique de test. La matrice d'information de Fisher peut être approchée par la matrice d'information de Fisher observée [2] et le principe de Louis [4].

Les quantiles de la loi du chi-bar-square sont quant à eux approchés par des sommes de Monte Carlo.

5 Simulations

Pour illustrer les résultats obtenus dans les sections précédentes, nous proposerons plusieurs exemples sur des données simulées, en distinguant les cas avec effets indépendants ou corrélés :

- $r = 1$ et $p = 2$
- $r = 1$ et $p = 3$
- $r = 2$ et $p = 4$

On utilisera pour cela le modèle suivant :

$$y_{ij} = \varphi_{i4} + \frac{\varphi_{1i}}{1 + \exp\left(\frac{t_{ij} - \varphi_{2i}}{\varphi_{3i}}\right)} + \varepsilon_{ij} \ , \tag{6}$$

où y_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots J$ sont les observations, $\phi_i = (\phi_{i1}, \phi_{i2}, \phi_{i3}, \phi_{i4})$ les effets aléatoires, où $\varphi_i \sim \mathcal{N}(m, \Gamma)$, et où $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ sont indépendants.

Pour évaluer le niveau empirique du test, R jeux de données sont générés selon le modèle (6) sous l'hypothèse nulle, et pour chacune de ces répétitions, on évalue la statistique de test correspondante. Les résultats seront présentés pour différentes valeurs de n, afin de mettre en évidence l'asymptotique du test.

Références

[1] H. Chernoff. On the Distribution of the Likelihood Ratio. *The Annals of Mathematical Statistics*, 25(3) :573–578, 1954.

[2] B. Efron and D. V. Hinkley. Assessing the accuracy of the maximum likelihood estimator : Observed versus Expected Fisher information. *Biometrika*, 65(3) :457 – 487, 1978.

[3] M. Lavielle. *Mixed effects models for the population approach : models, tasks, methods and tools*. CRC Press, 2014.

[4] T. A. Louis. Finding the Observed Information Matrix when using the EM-algorithm. *Journal of the Royal Statistical Society*, 44(2) :226–233, 1982.

[5] S. G. Self and K.-Y. Liang. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, 82(398) :605–610, 1987.

[6] A. Shapiro. Asymptotic distribution of test statistics in the analysis of moment structures under inequality constraints. *Biometrika*, 72 :133–144, 1985.

[7] M. J. Silvapulle and P. K. Sen. *Constrained statistical inference : Order, inequality, and shape constraints*, volume 912. John Wiley & Sons, 2011.

[8] D. O. Stram and J. W. Lee. Variance components testing in the longitudinal mixed effects model. *Biometrics*, 50(4) :1171–1177, 1994.