JOURNAL DE LA SOCIÉTÉ FRANÇAISE DE STATISTIQUE

MAGALI SAN CRISTOBAL
CHRISTÈLE ROBERT-GRANIÉ
JEAN-LOUIS FOULLEY

Héteroscédasticité et modèles linéaires mixtes : théorie et applications en génétique quantitative

Journal de la société française de statistique, tome 143, nº 1-2 (2002), p. 155-165

<http://www.numdam.org/item?id=JSFS_2002__143_1-2_155_0>

© Société française de statistique, 2002, tous droits réservés.

L'accès aux archives de la revue « Journal de la société française de statistique » (http://publications-sfds.math.cnrs.fr/index.php/J-SFdS) implique l'accord avec les conditions générales d'utilisation (http://www.numdam.org/conditions). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
http://www.numdam.org/

HÉTEROSCÉDASTICITÉ ET MODÈLES LINÉAIRES MIXTES : THÉORIE ET APPLICATIONS EN GÉNÉTIQUE QUANTITATIVE

Magali SAN CRISTOBAL [1], Christèle ROBERT-GRANIÉ [2], Jean-Louis FOULLEY [3]

RÉSUMÉ

Cet article montre comment a été introduit et formalisé le concept de variances hétérogènes en génétique animale par le biais d'une modélisation à effets mixtes du logarithme des variances. Divers modèles sont présentés et discutés. L'ensemble est illustré à travers différentes applications ayant trait notamment à l'analyse de données subjectives (pointage), à l'évaluation génétique des reproducteurs (bovins laitiers), aux problèmes de l'interaction génotype x milieu et de la sélection canalisante, et à l'analyse de données longitudinales.

ABSTRACT

In this paper it is shown how the concept of heterogeneous variances was introduced and formalised in animal genetics, using a mixed model approach on logvariances. Several models are presented and discussed. The whole is illustrated through several applications, such as the analysis of subjective data (scores), the genetic evaluation of breeding animals (dairy cattle), the issues of genotype x environment interaction and canalising selection, and the analysis of longitudinal data.

1. Introduction

Le but principal de la génétique animale est l'amélioration des animaux domestiques par le biais de la sélection. L'évaluation du potentiel génétique de chacun des animaux candidats à la sélection est effectuée en routine à l'aide d'un modèle linéaire mixte du type :

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{Wp} + \boldsymbol{e} \tag{1}$$

1. Institut National de la Recherche Agronomique, Laboratoire de Génétique Cellulaire, 31326 Castanet Tolosan; E-mail : msc@toulouse.inra.fr

2. Institut National de la Recherche Agronomique, Station d'Amélioration Génétique des Animaux, 31326 Castanet Tolosan

3. Institut National de la Recherche Agronomique, Station de Génétique Quantitative et Appliquée, 78352 Jouy-en-Josas

Journal de la Société Française de Statistique, tome 143, n° 1-2, 2002

où $\boldsymbol{y}$ est le vecteur n x 1 des performances, $\boldsymbol{X}$, $\boldsymbol{Z}$ et $\boldsymbol{W}$ sont des matrices d'incidence de tailles respectives n x p, n x q et n x r, $\boldsymbol{\beta}$ est le vecteur des effets fixes, $\boldsymbol{u}$ est le vecteur des effets aléatoires représentant les valeurs génétiques des animaux : $\boldsymbol{u} \sim \mathcal{N}(\boldsymbol{0}, \sigma_u^2 \boldsymbol{A})$, avec $\boldsymbol{A}$ la matrice de parenté connue, $\boldsymbol{p}$ est le vecteur d'effets aléatoires représentant des effets d'environnement permanent : $\boldsymbol{p} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\sigma}_p^2 \boldsymbol{I}_n)$, et $\boldsymbol{e}$ est le vecteur des résidus : $\boldsymbol{e} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\sigma}_e^2 \boldsymbol{I}_n)$ avec $\boldsymbol{I}_n$ la matrice identité.

La notion d'héritabilité (la part de variance génétique sur la variance totale $\frac{\sigma_u^2}{\sigma_y^2}$) est importante pour les sélectionneurs, car elle permet de prédire l'efficacité de la sélection.

Cependant, l'hypothèse d'homogénéité des variances n'est pas toujours vérifiée (Garrick et VanVleck 1987, Hill 1984, Robert *et al.* 1995, Robert-Granié *et al.* 1997, SanCristobal *et al.* 1993, Visscher 1992, Visscher et Hill 1992, Weigel *et al.* 1993), et la non prise en compte de cette hétérogénéité des variances perturbe la prédiction des effets aléatoires et diminue donc l'efficacité de la sélection.

Il existe de nombreux domaines d'application en génétique animale, pour lesquels l'hétérogénéité des variances est requise dans un modèle, soit pour améliorer l'estimation de paramètres d'intérêt du modèle traditionnel (1), soit parce que c'est le but même des généticiens. Le lien entre les divers domaines d'application réside dans leur modélisation. Pour cette raison, nous allons tout d'abord présenter notre approche modélisatrice, puis dans un second temps, diverses problématiques rencontrées en génétique quantitive et ayant recours à la modélisation des hétérogénéités de variances.

2. Modélisations

Un premier pas vers la modélisation des variances hétérogènes peut être décrit par une matrice de variance-covariance résiduelle diagonale par blocs, chaque bloc correspondant à un niveau d'un facteur de variation :

$$Var(\boldsymbol{e}) = \oplus_{i=1}^{k} \sigma_{e,i}^2 \boldsymbol{I}_{n_i}. \tag{2}$$

Quand plusieurs facteurs doivent être pris en compte, il devient nécessaire de structurer les paramètres de dispersion $\sigma_{e,i}^2$ pour $i = 1, ...k$. Il est intéressant de rester dans le cadre du modèle linéaire, aussi le modèle dit « structural » (Foulley *et al.* 1990)

$$\ln \sigma_{e,i}^2 = \boldsymbol{p}_{e,i} \boldsymbol{\delta}_e \qquad i = 1, ...k \tag{3}$$

est-il couramment employé. Des paramètres à estimer (vecteur $\boldsymbol{\delta}_e$ d'effets fixes pouvant comporter des facteurs ou des covariables) sont introduits, les $\boldsymbol{p}_{e,i}$ étant des vecteurs d'incidence correspondant aux covariables incriminées dans l'hétéroscédasticité. La fonction de lien logarithmique est intéressante à plus d'un titre. Elle conduit tout d'abord aux estimations des variances $\sigma_{e,i}^2$ positives, sans pour autant restreindre l'espace des paramètres $\boldsymbol{\delta}_e$. Elle

permet de travailler indifféremment sur les variances ou leurs ratios (Foulley 1997). D'autre part, le lien log est le lien canonique dans le modèle linéaire généralisé sur $d_i = (y_i - \mu_i)^2$, dont la loi est gamma et l'espérance $\sigma^2_{e,i}$, où μ_i représente l'estimée de la moyenne de la cellule i sous le modèle homoscédastique (1). Enfin, la transformation logarithmique sur les variances a un effet normalisateur, et permettra, comme nous le verrons plus loin dans le traitement de la sélection canalisante, de se placer dans le cadre génétique conceptuel du modèle infinitésimal.

L'analogie entre modélisation des moyennes et celle des variances résiduelles peut être poussée plus avant par l'introduction de facteurs à effets aléatoires dans (3) :

$$\ln\sigma^2_{e,i} = \boldsymbol{p}_{e,i}\boldsymbol{\delta}_e + \boldsymbol{q}_{e,i}\boldsymbol{v}_e \qquad i = 1, ...k, \tag{4}$$

où $\boldsymbol{v}_e$ représente un effet aléatoire ayant une certaine distribution. Quelle est-elle ? D'un point de vue bayésien, il est tentant de choisir une loi a priori conjuguée. Comme la loi a posteriori de $\sigma^2_{e,i}$ est une Gamma inverse, nous pouvons proposer que la loi a priori de $\boldsymbol{v}_e$ soit une log gamma inverse. Or cette loi peut être approchée avec une bonne précision par une loi normale (Foulley *et al.* 1992) :

$$\boldsymbol{v}_e \sim \mathcal{N}(\boldsymbol{0}, \sigma^2_v \boldsymbol{I}). \tag{5}$$

Cette formulation peut conduire à une interprétation génétique de la variation de la variabilité environnementale (SanCristobal *et al.* 1998) :

$$\boldsymbol{v}_e \sim \mathcal{N}(\boldsymbol{0}, \sigma^2_v \boldsymbol{A}), \tag{6}$$

où $\boldsymbol{A}$ est la matrice de parenté entre les animaux présents dans l'analyse, si $\boldsymbol{v}_e$ représente ceux-ci. Dans ce cas, $\boldsymbol{v}_e$ pourra être interprété comme un vecteur de valeurs génétiques additives. Le modèle génétique sous-jacent met en oeuvre un ensemble de polygènes agissant sur le niveau moyen du caractère ($\boldsymbol{u}$ étant la résultante des effets de ces gènes), et un autre ensemble de polygènes contrôlant la sensibilité aux variations du milieu (résultante $\boldsymbol{v}_e$).

Jusqu'ici, une similarité de modélisation a été décrite, entre la moyenne et la variance résiduelle en introduisant des effets fixes, des effets aléatoires et des effets aléatoires génétiques. Une approche identique peut être menée sur la variance génétique σ^2_u du modèle de base (1). Ainsi des variances hétérogènes peuvent apparaître à deux niveaux : soit dans la partie environnementale, ce qui est le plus souvent envisagé, soit dans la partie génétique, soit dans les deux :

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \sigma_{u,i}\boldsymbol{Z}_i\boldsymbol{u}^* + \boldsymbol{e}_i \tag{7}$$

où $\boldsymbol{u}^* \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{A})$ est le vecteur des valeurs génétiques standardisées, et $\boldsymbol{e}_i \sim \mathcal{N}(\boldsymbol{0}, \sigma^2_{e,i}\boldsymbol{I}_{n_i})$.

Un modèle linéaire généralisé est établi :

$$\log\sigma^2_{u,i} = \boldsymbol{p}'_{u,i}\boldsymbol{\delta}_u, \tag{8}$$

voire

$$\log\sigma^2_{u,i} = \boldsymbol{p}'_{u,i}\boldsymbol{\delta}_u + \boldsymbol{q}'_{u,i}\boldsymbol{v}_u, \tag{9}$$

où $\boldsymbol{p}_{u,i}$ est un vecteur d'incidence, $\boldsymbol{\delta}_u$ un vecteur d'effets fixes et $\boldsymbol{v}_u$ un vecteur d'effets aléatoires tel que $\boldsymbol{v}_u \sim \mathcal{N}(\boldsymbol{0}, \sigma^2_{v,u}\boldsymbol{I})$.

Diverses contraintes liant ces paramètres ont aussi été proposées, telles que (Foulley *et al.* 1998) :

$$\frac{\sigma_{u,i}}{\sigma^b_{e,i}} = const. \tag{10}$$

où le paramètre b est un réel et permet notamment de tester l'hypothèse d'héritabilité constante ($H_0 : b = 1$).

De tels modèles s'avèrent très utiles en génétique quantitative. Ils permettent par exemple de tester l'homogénéité de certains paramètres génétiques selon les milieux (Robert *et al.* 1995) et de rendre compte aisément des phénomènes d'échelle dans les interactions génotype x milieu. On a pu utiliser les modèles (4) et (6) pour rendre compte de la sensibilité génétique aux micro variations de milieu (SanCristobal *et al.* 1998).

Les modèles (7), (9) et (10) sont aussi à la base de l'évaluation génétique française des bovins sur la production laitière intégrant différents facteurs d'hétérogénéité de variance (Robert-Granié *et al.* 1999). Enfin, ces modèles peuvent s'intégrer aisément dans l'analyse de données longitudinales, un des facteurs d'hétérogénéité pouvant être la covariable temps (Robert-Granié *et al.* 2002).

3. Extensions

Lorsque des données catégorielles ordonnées sont analysées, les généticiens quantitatifs ont pris l'habitude d'utiliser le modèle dit « modèle à seuils » qui relie la performance à une variable aléatoire gaussienne sous-jacente. Ainsi, une valeur j sera observée si la variable sous-jacente est comprise entre les seuils τ_{j-1} et τ_j, avec $\tau_0 = -\infty < \tau_1 < ... < \tau_{J-1} < \tau_J = +\infty$, où J est le nombre de catégories. Notons que le cas particulier de données binaires ordonnées modélisées de cette façon rentre dans le cadre d'un modèle linéaire généralisé, de distribution binomiale, avec un lien probit. Sur l'échelle sous-jacente, les composantes de la variance peuvent être modélisées comme dans le paragraphe précédent, toutes les gammes étant a priori possibles (Foulley et Gianola 1996, SanCristobal *et al.* 2001).

4. Inférence

Le maximum de vraisemblance restreinte (REML, Patterson et Thompson 1971) est la méthode statistique de choix pour l'estimation des composantes de la variance (Foulley *et al.* 2002). Cependant le calcul des estimateurs REML nécessite de recourir à des procédures itératives. L'algorithme EM (Dempster *et al.* 1977) s'avère particulièrement bien adapté au cas du modèle mixte à variances hétérogènes. Par exemple, dans le cas du modèle (7), on va définir le vecteur des données complètes $\boldsymbol{x}$ par $\boldsymbol{x}' = (\boldsymbol{\beta}', \boldsymbol{u}^{*'}, \boldsymbol{e}')$ à la phase M, la

formule donnant la valeur actualisée de $\sigma_{u,i}$ à l'itération $(t+1)$ a maintenant la forme d'un coefficient de régression :

$$\sigma_{u,i}^{(t+1)} = \frac{E\left[\boldsymbol{u}^{*'}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})|\boldsymbol{y}, \sigma_{u,i}^{(t)}, \sigma_{e,i}^{(t)}\right]}{E\left[\boldsymbol{u}^{*'}\boldsymbol{Z}_i'\boldsymbol{Z}_i\boldsymbol{u}^{*}|\boldsymbol{y}, \sigma_{u,i}^{(t)}, \sigma_{e,i}^{(t)}\right]}. \tag{11}$$

Dans le cas homogène cet algorithme dit « scaled EM » du fait de la normalisation des effets aléatoires (cf Foulley et Quaas 1995) se différencie nettement de l'algorithme quadratique classique et s'avère beaucoup plus performant dans une certaine gamme de valeurs de paramètres. Il préfigure certaines nouvelles versions de l'algorithme EM basées sur l'introduction de paramètres de travail (Meng et VanDyk 1998) telle que celle dite PX-EM (Foulley et VanDyk 2002).

Concrètement, à chaque itération de l'algorithme, tous les paramètres sont réactualisés $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{u}^*, \boldsymbol{\delta}_e, \boldsymbol{v}_e, \boldsymbol{\delta}_u, \boldsymbol{v}_u)$ d'une part et $\boldsymbol{\sigma}^2 = (\boldsymbol{\sigma}_e^2, \boldsymbol{\sigma}_u^2)$ d'autre part. À la convergence, des estimateurs REML $\hat{\boldsymbol{\sigma}}^2$ des composantes de la variance $\boldsymbol{\sigma}^2$ sont obtenus, avec, comme sous-produit, les maxima de vraisemblance de $\boldsymbol{\theta}$ conditionnellement à $\boldsymbol{\sigma}^2 = \hat{\boldsymbol{\sigma}}^2$.

Notons aussi l'utilisation du calcul bayésien, notamment l'algorithme MCMC (Ros *et al.* 2002), ce qui paraît bien naturel dans un domaine où le paradigme bayésien est bien implanté.

En règle générale, le choix de la méthode d'estimation dépend des convictions de chacun et de la facilité d'implémentation.

5. Applications

5.1. Pointages

Tout étudiant s'est trouvé confronté à divers types d'examinateurs, ayant chacun sa propre échelle de notation. Un 12/20 avec un professeur notant habituellement entre 8 et 12 représente une plus belle performance qu'un 12/20 avec un autre professeur utilisant tout l'intervalle [0, 20]. Dans les deux cas, la moyenne est égale à 10.

Le contrôle de morphologie de la descendance des taureaux est réalisé par des agents agréés qui utilisent un document ou table de pointage regroupant une trentaine de postes relatifs à la morphologie de l'animal. Pour chacun de ces postes, les « pointeurs » mesurent à l'œil ou avec un instrument des longueurs ou des angles sur une échelle de 1 à 9. L'ensemble de ces informations de description morphologique concourt à l'évaluation génétique des taureaux. Certains facteurs tels que le pointeur (en raison du caractère très subjectif de certaines notes) ou le type de logement (en raison du stress différent qu'il peut engendrer sur les membres de l'animal) étaient suspectés induire des hétérogénéités de variance au niveau de certains caractères (les aplombs ou angle de jarret). C'est pourquoi, afin d'améliorer la fiabilité de l'évaluation génétique, il nous a paru intéressant et nécessaire d'utiliser les méthodes du modèle linéaire mixte hétéroscédastique (SanCristobal *et al.*

1993, Robert *et al.* 1997). Deux types d'hétérogénéité ont été mis en évidence pour l'analyse de ces caractères : pour certaines combinaisons caractère x source d'hétérogénéité, l'hypothèse d'héritabilité constante est retenue ($b = 1$ dans l'équation (10)). C'est le cas par exemple pour les caractères de mamelle avec une hétérogénéité liée au stade de lactation. Pour d'autres combinaisons (par exemple aplombs x pointeurs), les écarts d'héritabilité et de corrélations génétiques entre pointeurs traduisent des problèmes importants dans la définition même du caractère (une redéfinition plus précise de certains postes élémentaires est peut-être à envisager). L'utilisation de ces résultats est sans doute un moyen d'aider les pointeurs à homogénéiser leurs techniques de pointage. L'analyse de telles données conduit d'un point de vue statistique à la mise en évidence de la variabilité entre milieux mais elle permet aussi de juger de la qualité des données, de leur recueil, et par conséquent d'apprécier la cohérence du travail des pointeurs. L'analyse de ces données et la modélisation retenue est décrite en détail dans Robert-Granié *et al.* (1997).

5.2. Évaluation génétique des bovins laitiers

L'existence de variances hétérogènes et leur non prise en compte dans l'évaluation génétique en routine des bovins laitiers peuvent engendrer des biais importants. C'est particulièrement le cas pour le choix des mères à taureaux lorsque l'intensité de sélection est forte. Les facteurs de variation (l'année de production, la région, le troupeau, le niveau de production, le numéro de lactation) sont en général bien connus mais leur prise en compte reste délicate. En effet, l'hypothèse d'hétéroscédasticité implique de réaliser une analyse simultanée des paramètres de dispersion (variances génétiques et résiduelles) et de position (effets de milieu, valeurs génétiques) sur un grand nombre de données. Cette analyse est importante pour affiner le choix des mères à taureaux, pour obtenir une bonne appréciation historique des résultats de l'évaluation, une bonne précision entre unités de sélection et aussi une plus grande clarté dans les comparaisons internationales. Une prise en compte rigoureuse de l'hétérogénéité de la variance dans le modèle d'évaluation laitier était donc souhaitable et nécessaire. Les premières études réalisées ont consisté à modéliser les variances génétiques et résiduelles en situation d'hétéroscédasticité sous l'hypothèse d'héritabilité et de répétabilité constantes, en utilisant le modèle (4) et $b = 1$ dans (10) (Meuwissen *et al.* 1996, Robert-Granié *et al.* 1999).

Les conséquences de la prise en compte rigoureuse de l'hétérogénéité de la variance ont été importantes sur les index femelle. En effet, les femelles ne produisant généralement que dans un seul troupeau voient leur index plus affectés que ceux des mâles ; le reclassement des femelles est alors important (seules 600 femelles du top 1000 sous le modèle homogène restent dans le top 1000 sous le modèle hétérogène). Le mode de recrutement de l'élite est changé puisqu'avec le nouveau modèle la contribution des différents milieux est plus équilibrée, alors qu'avec le modèle homogène seuls les troupeaux à forte variabilité contribuaient à l'élite (Robert-Granié *et al.*, 1999).

5.3. Intéraction génotype x milieu

Au départ et pour simplifier, l'hypothèse d'additivité des effets génétiques et du milieu est supposée, les corrélations génétiques entre répétitions d'un caractère dans différents milieux sont supposées alors égales à 1. Or, cette hypothèse n'est pas toujours réaliste. Lorsqu'on dispose d'observations de plusieurs génotypes (performances des descendants de taureaux) présents dans des milieux différents, on peut aboutir à des résultats très variés, donnant lieu notamment à des phénomènes d'interaction génotype x milieu. En effet, les divers génotypes issus de schémas de sélection sont confrontés à une gamme de milieux très large et le risque de se trouver en présence d'interaction génotype x milieu va croissant. Le modélisation de ce phénomène peut être appréhendé soit par une approche multicaractères (Falconer, 1952), soit par une approche unidimensionnelle hétéroscédastique beaucoup moins coûteuse en paramètres (Robert *et al.* 1995). L'approche multicaractère de Falconer est basé sur un modèle linéaire mixte multidimensionnel avec des matrices de composantes de variance-covariance entre caractères. Ce modèle considère l'expression d'un caractère mesuré dans des milieux différents comme celle d'autant de caractères génétiquement liés. La deuxième approche est basée sur l'écriture d'un modèle linéaire à deux facteurs croisés : génotype (effet aléatoire), milieu (effet fixe) avec interaction. L'hypothèse d'hétéroscédasticité complète entre les différents milieux conduit à supposer l'existence de variances génétiques et résiduelles différentes d'un milieu à l'autre. Ce modèle peut s'écrire sous la forme suivante :

$$y_{ijk} = \mu + h_i + \sigma_{si} s_j^* + \sigma_{hsi} hs_{ij}^* + e_{ijk} \tag{12}$$

avec h_i effet du milieu i, $s_j^* \sim NID(0,1)$ effet aléatoire du père j, $hs_{ij}^* \sim NID(0,1)$ effet aléatoire d'interaction génotype x milieu, $e_{ijk} \sim NID(0, \sigma_{ei}^2)$ effet résiduel, $Cov(y_{ijk}, y_{ijk'}) = \sigma_{si}^2 + \sigma_{hsi}^2$ (même père, même milieu), $Cov(y_{ijk}, y_{i'jk'}) = \sigma_{si}\sigma_{si'}$ (même père, milieux différents). La valeur génétique u_{ij} de l'individu j dans le milieu i s'exprime donc comme la somme de deux composantes : $\sigma_{si} s_j^*$ représente une composante d'aptitude générale qui traduit la perennité des valeurs génétiques (s_j^*) à un facteur d'échelle près (σ_{si}) et $\sigma_{hsi} hs_{ij}^*$ représente simplement l'écart entre la valeur vraie u_{ij} et la composante d'aptitude générale $\sigma_{si} s_j^*$. L'expression d'un caractère mesuré dans deux milieux différents peut être considérée comme celle de deux caractères différents mais génétiquement liés. La corrélation génétique qui mesure leur liaison reflète ainsi l'importance de l'interaction. À partir de ce modèle général, plusieurs hypothèses peuvent être formulées. On peut en particulier s'intéresser à la question de l'invariance de certains paramètres génétiques tels que la corrélation génétique entre milieux $\rho_{ii'} = \dfrac{\sigma_{si}\sigma_{si'}}{\sqrt{(\sigma_{si}^2 + \sigma_{hsi}^2)(\sigma_{si'}^2 + \sigma_{hsi'}^2)}}$ et l'héritabilité d'un caractère dans un milieu donné $h_i^2 = \dfrac{\sigma_{si}^2 + \sigma_{hsi}^2}{\sigma_{si}^2 + \sigma_{hsi}^2 + \sigma_{ei}^2}$. Un exemple est fourni par le test de l'homogénéité des composantes de (co)variance génétique entre milieux que nous avons étudié en analysant les

effets d'un stress environnemental dans le cadre de l'amélioration génétique des plantes (Foulley *et al.* 1994, Robert *et al.* 1995).

5.4. Sélection canalisante

La sélection animale a eu pour objectif, depuis l'après-guerre, d'augmenter la production (quantité de viande, de lait, ...). Son efficacité n'est plus à prouver, les pays industrialisés rencontrant de nos jours des problèmes de surproduction. L'attente des consommateurs est maintenant autre, la qualité des produits étant devenu un souci majeur, cette qualité ne correspondant pas en général à une valeur maximale d'un caractère, mais plutôt à une valeur intermédiaire. Par exemple, un foie gras ne devra être ni trop petit (car pas assez gras), ni trop gros car alors il perdra trop de graisse à la cuisson. Pour la qualité de la viande de porc, le pH post-mortem doit être égal à 7.5 pour obtenir de bonnes qualités organoleptiques (recherchées par le consommateur), mais aussi de bons rendements technologiques nécessaires à la fabrication du jambon cuit (qui assure donc une bonne compétitivité à la filière professionnelle). Dans ce cadre, une homogénéité des lots autour d'un niveau moyen optimal sera privilégiée par la sélection, appelée sélection canalisante. Un autre exemple concerne la production d'escargots, qui requiert de lourds moyens en main d'oeuvre pour le tri quotidien d'animaux aptes à la vente. Des lots d'animaux qui arriveraient à maturité au même moment est le rêve de tout héliciculteur. Enfin, la préoccupation croissante de l'opinion publique sur la question du bien-être animal implique des conduites d'élevage moins contrôlées, telle que celle des poulets qui se fera de plus en plus en plein air. Si la sensibilité aux variations de l'environnement est sous contrôle génétique, la sélection canalisante pourra limiter les pertes de l'éleveur et fournir malgré tout des produits de qualité constante.

Les conditions d'élevage étant donc variables et non caractérisables (contrairement au paragraphe précédent où les milieux sont bien définis), l'homogénéité de la production soumise à des environnements variables i peut être mesurée par la variabilité environnementale $\sigma^2_{e,i}$, dans un modèle mixte gaussien. L'ajustement des modèles (4) (6) pourra quantifier le contrôle génétique σ^2_v de la sensibilité aux variations du milieu. La réponse espérée à une sélection canalisante peut être calculée (SanCristobal *et al.* 1998), et ce type de sélection envisagé ou pas.

Quand les données sont ordinales, un index de sélection peut être défini sur l'échelle sous jacente (ce qui nous ramène au cadre gaussien) ou observée (l'index peut être la proportion de 2 par exemple). Dans tous les cas, il a été montré à l'aide de simulations qu'un progrès génétique est effectif pour une sélection canalisante de données ordinales, mais que la traduction sur l'échelle observée de ce progrès génétique sous jacent est longtemps masqué par la liaison moyenne-variance propre à ce type de données (SanCristobal *et al.* 2001).

5.5. Données longitudinales

Les données répétées dans le temps, données généralement disponibles quand on analyse une croissance, une production laitière ou une carrière, sont l'expression d'un nombre potentiellement infini de caractères dont les covariances varient de façon continue avec le temps. La caractéristique principale des mesures répétées est leur corrélation dont la structure temporelle fait l'objet de la modélisation.

Une des façons de modéliser de telles données consiste à faire appel à un modèle linéaire mixte dit à coefficients aléatoires (Foulley *et al.* 2000) qui pour l'unité expérimentale i (l'individu) va s'écrire :

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \sum_{k=0}^{K} \boldsymbol{Z}_{ik}\boldsymbol{u}_{ik} + \boldsymbol{e}_i \tag{13}$$

où $\boldsymbol{y}_i = \{y_{ij}\}$ est le vecteur des n_i mesures (indice j) faites sur l'individu i ; $\boldsymbol{X}_i\boldsymbol{\beta}$ représente la tendance moyenne ; $\boldsymbol{Z}_{ik}$ est le vecteur des n_i éléments relatifs à la $k^{\text{ème}}$ covariable temporelle de coefficient aléatoire $\boldsymbol{u}_{ik}$; $\boldsymbol{e}_i$ est le vecteur (n_i x 1) des erreurs, supposé $\mathcal{N}(\boldsymbol{0}, \boldsymbol{R}_i)$ avec dans le cas usuel le plus simple $\boldsymbol{R}_i = \sigma_e^2 \boldsymbol{I}_{n_i}$. Soit $\boldsymbol{u}_i = (\boldsymbol{u}_{i0}, ..., \boldsymbol{u}_{ik}, ..., \boldsymbol{u}_{iK})'$ le vecteur des K coefficients aléatoires relatifs à l'individu i, on suppose $\boldsymbol{u}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{G}_0)$ où $\boldsymbol{G}_0 = \{g_{kl}\}$ est la matrice de variance covariance entre coefficients. En fait on va aisément s'affranchir de cette hypothèse d'homogénéité de la composante résiduelle σ_e^2 et de $\boldsymbol{G}_0$ en écrivant le modèle suivant pour la strate m de la population

$$\boldsymbol{y}_{im} = \boldsymbol{X}_{im}\boldsymbol{\beta} + \sum_{k=0}^{K} \boldsymbol{Z}_{imk}\sigma_{u_m}\boldsymbol{u}_{ik}^* + \boldsymbol{e}_{im} \tag{14}$$

où les $\boldsymbol{u}_{ik}^*$ ont la même signification que précédemment mais sont ici standardisés si bien que le vecteur $\boldsymbol{u}_i^* = (\boldsymbol{u}_{i0}^*, ..., \boldsymbol{u}_{ik}^*, ..., \boldsymbol{u}_{iK}^*)'$ est maintenant distribué selon $\mathcal{N}(\boldsymbol{0}, \boldsymbol{\Omega}_0)$ où $\boldsymbol{\Omega}_0 = \{\rho_{kl}\}$ est une matrice de corrélation entre coefficients aléatoires. De la même façon qu'en (4) et (9) on écrira

$$\log \sigma_{e,im} = \boldsymbol{p}'_{e,im}\boldsymbol{\delta}_e + \boldsymbol{q}'_{e,im}\boldsymbol{v}_e \tag{15}$$

$$\log \sigma_{u,im} = \boldsymbol{p}'_{u,im}\boldsymbol{\delta}_u + \boldsymbol{q}'_{u,im}\boldsymbol{v}_u. \tag{16}$$

Cela permet de rendre compte de diverses sources d'hétérogénéité structurelle. Ainsi avons nous pu modéliser les courbes de croissance pondérale de veaux mâles nés simples et jumeaux, et mettre en évidence une hétérogénéité de variance résiduelle selon les facteurs suivants : rang de vélage, saison de naissance du veau et âge à la pesée (Robert-Granié *et al.* 2002).

6. Conclusion

Le concept de variance hétérogène s'introduit aisément aussi bien dans le modèle linéaire mixte de la variance des effets résiduels qu'à celui des effets

aléatoires. La modélisation structurelle $\log \sigma_i = \boldsymbol{p}_i'\boldsymbol{\delta} + \boldsymbol{q}_i'\boldsymbol{v}$ incluant à son tour une partie fixe ($\boldsymbol{p}_i'\boldsymbol{\delta}$) et une partie aléatoire ($\boldsymbol{q}_i'\boldsymbol{v}$) donne beaucoup de généralité et de flexibilité. Elle permet notamment de faire le lien entre l'approche classique et l'approche bayésienne. À cet égard, elle rend compte aussi bien d'un modèle à bruit de fond indifférencié ou de mélange du type $\log \sigma_i = \delta_i + v_i$ avec $v_i \sim \mathcal{N}(0, \sigma_v^2)$ que d'une approche structurale fine (cf interaction génotype x milieu, sélection canalisante). La grande latitude qui existe dans le choix des covariables explicatives (discrètes, continues) ouvre sur une grande gamme d'applications (cf données longitudinales) et fait que ce type de modélisation est partie intégrante de la théorie du modèle mixte.

RÉFÉRENCES

DEMPSTER AP., LAIRD NM., RUBIN DB. (1977), Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statis. Soc B*, 39, 1-38.

FALCONER DS. (1952), The problem of environment and selection, *Am. Nat.*, 86, 293-298.

FOULLEY JL. (1997), ECM approaches to heteroskedastic mixed models with constant variance ratios, *Genet. Sel. Evol.*, 29, 297-318.

FOULLEY JL., DELMAS C., ROBERT-GRANIÉ C. (2002), Méthodes du maximum de vraisemblance en modèle linéaire mixte, *Journal de la Société Française de Statistique*, 143, 1-2, 5-52.

FOULLEY JL., GIANOLA D. (1996), Statistical analysis of ordered categorical data via a structural heteroskedastic threshold model, *Genet. Sel. Evol.*, 28, 217-320.

FOULLEY JL., GIANOLA D., SANCRISTOBAL M., IM S. (1990), A method for assessing extent and sources of heterogeneity of residual variances in mixed linear models, *J. Dairy Sci.*, 73, 1612-1624.

FOULLEY JL., HÉBERT D., QUAAS R.L. (1994), Inferences on homogeneity of between-family components of variance and covariance among environments in balanced cross-classified designs, *Genet. Sel. Evol.*, 26, 117-136.

FOULLEY JL., JAFFREZIC F., ROBERT-GRANIÉ C. (2000), EM-REML estimation of covariance parameters in Gaussian mixed models for longitudinal data analysis, *Genet. Sel. Evol.*, 32, 129-141.

FOULLEY JL., QUAAS RL. (1995), Heterogeneous variances in Gaussian linear mixed models, *Genet. Sel. Evol.*, 27, 211-228.

FOULLEY JL., QUAAS RL., THAON D'ARNOLDI C. (1998), A link function approach to heterogeneous variance components, *Genet. Sel. Evol.*, 30, 27-43.

FOULLEY JL., SANCRISTOBAL M., GIANOLA D., IM S. (1992), Marginal likelihood and Bayesian approaches to the analysis of heterogeneous residual variances in mixed linear Gaussian models, *Comput. Stat. Data Anal.*, 13, 291-305.

FOULLEY JL., VANDYK DA. (2002), The PX-EM algorithm for fast fitting of Henderson's mixed model, *Genet. Sel. Evol.*, 32, 143-163.

GARRICK D.J., VAN VLECK L.D. (1987), Aspects of selection for performance in several environments with heterogeneous variances, *J. Anim. Sci.*, 65, 409-421.

HILL W.G. (1984), On selection among groups with heterogeneous variance, *Anim. Prod.*, 39, 473-477.

MENG XL., VANDYK DA. (1998), Fast EM-type implementation for mixed effects models, *J. R. Statis. Soc. B*, 60, 559-578.

MEUWISSEN T.H.E., DE JONG G., ENGEL B. (1996), Joint estimation of breeding values and heterogeneous variances of large data files, *J. Dairy Sci.*, 79, 310-316.

PATERSON HD., THOMPSON R. (1971), Recovery of inter-block information when block sizes are unequal, *Biometrika*, 58, 545-554.

ROBERT C., DUCROCQ V., FOULLEY JL. (1997), Heterogeneity of variance for type traits in the Montbéliarde cattle, *Genet. Sel. Evol.*, 29, 545-570.

ROBERT C., FOULLEY JL., DUCROCQ V. (1995), Genetic variation of traits measured in several environments. I Estimation and testing of homogeneous genetic and intra-class correlations. II Inference on between environment homogeneity of intra-class correlations using heteroskedastic models, *Genet. Sel. Evol.*, 27, 111-134.

ROBERT-GRANIÉ C., BONAITI B., BOICHARD D., BARBAT A. (1999), Accounting for variance heterogeneity in French dairy cattle genetic evaluation, *Livestock Prod. Sci.*, 60, 343-357.

ROBERT-GRANIÉ C., HEUDE B., FOULLEY JL. (2002), Modelling the growth curve of Maine-Anjou beef cattle using heteroskedastic random regression models, *Genet. Sel. Evol.*, 34, 423-445.

ROS M., SANCRISTOBAL M., DUPONT-NIVET M., MALLARD J. (2002), Comparison of maximum likelihood, bootstrap and MCMC methods in a structural model for heterogeneous variances, *Proc. 7th WCGALP, 19-23 August 2002, Montpellier, France.*

SANCRISTOBAL M., FOULLEY JL., MANFREDI E. (1993), Inference about multiplicative heteroskedastic components of variance in a mixed linear Gaussian model with an application to beef cattle breeding, *Genet. Sel. Evol.*, 25, 3-30.

SANCRISTOBAL-GAUDY M., BODIN L., ELSEN JM., CHEVALET C. (2001), Genetic components of litter size variability in sheep, *Genet. Sel. Evol.*, 33, 249-271.

SANCRISTOBAL-GAUDY M., ELSEN JM., BODIN L., CHEVALET C. (1998), Prediction of the response to a selection for canalisation of a continuous trait in animal breeding, *Genet. Sel. Evol.*, 30, 423-451.

VISSCHER P.M. (1992), On the power of likelihood ratio tests for detecting heterogeneity of intra-class correlations and variances in balanced half-sib designs, *J. Dairy Sci.*, 75, 1320-1330.

VISSCHER P.M., HILL W.G. (1992), Heterogeneity of variance and dairy cattle breeding, *Anim. Prod.*, 55, 312-329.

WEIGEL K.A., GIANOLA D., YANDEL B.S., KEOWN J.F. (1993), Identification of factors causing heterogeneous within-herd variance components using a structural model for variances, *J. Dairy Sci.*, 76, 1466-1478.