HETEROSKEDASTIC RANDOM COEFFICIENT MODELS

JL. Foulley and C. Robert-Granié[1]

INRA-SGQA, 78350 Jouy-en-Josas, [1]INRA-SAGA, BP 27, 31326 Castanet-Tolosan, France

INTRODUCTION

Random coefficient (RC) models have become very popular in quantitative genetics and animal breeding over the last decade for analysing longitudinal data (Schaeffer and Dekkers, 1994). These models use polynomials in time to describe mean profiles with random coefficients to generate a correlation structure among the repeated observations on each individual. Eventually such models are characterized by three parts: the mean profile component treated as fixed, the adjusted subject profile plus a within subject component (usually the error term) treated as random (Diggle *et al.*, 1994).

This paper is concerned by the random part i.e. with modelling the variance-covariance structure. We will extend RC models to a more general class of models termed heteroskedastic random coefficient (HRC) models. This class of models assumes that all variances of random effects can be heterogeneous. Inference is based on residual likelihood procedures (REML, Patterson and Thompson, 1971) and estimating equations derived from the Expectation-Maximization (EM, Dempster *et al.*, 1977) theory, more precisely the Expectation/Conditional Maximization (ECM) algorithm introduced by Meng and Rubin (1993). This procedure relies widely on previous works made on heterogeneous variances in an animal breeding context both at the theoretical (Foulley *et al.*, 1990; San Cristobal et al., 1993; Foulley and Quaas, 1995; Foulley, 1997; Foulley *et al.*, 1998) and applied levels (Gianola *et al.*, 1992; Robert-Granié *et al.*, 1999).

METHODS

A HRC model with K random coefficients can be written as follows :

$$y_{ijl} = x_{ijl}^{'}\boldsymbol{\beta} + \sum_{k=1}^{K}\sigma_{u_{ki}} z_{kijl} u_{kl}^{*} + e_{ijl}$$

where y_{ijl} is the j^{th} $(j=1,...,n_k)$ measurement recorded on the l^{th} $(l=1,...,q)$ individual in subclass i of the factor of heterogeneity $(i=1,...,p)$; $x_{ijl}^{'}\boldsymbol{\beta}$ represents the systematic component expressed as a linear combination of explanatory variables ($x_{ijl}^{'}$) with unknown linear coefficients ($\boldsymbol{\beta}$) ; $\sum_{k=1}^{K}\sigma_{u_{ki}} z_{kijl} u_{kl}^{*}$ represents the additive contribution of K random coefficient vectors (u_{kl}^{*}) based on covariate information (z_{kijl}) and which are specific to each l^{th} individual; $(\sigma_{u_{1i}},\ldots,\sigma_{u_{ki}},\ldots,\sigma_{u_{Ki}})$ are the K corresponding components of variance pertaining to stratum i. The K random effects $(u_{1l}^{*},\ldots,u_{kl}^{*},\ldots,u_{Kl}^{*})$ are correlated and the correlations are assumed

homogeneous over strata and equal to $\rho_{kk'}$ for k and $k'=1, ...,K$. The e_{ijl} represent independent errors.

A convenient and parsimonious procedure to handle heterogeneity of variances is to model them via a log-linear function. This approach has the advantage of maintaining parameter independence between the mean and covariance structure. Following Foulley *et al.* (1992), San Cristobal *et al.* (1993) among others, the residual variances were modelled as : $\ln\sigma^2_{e_i} = \mathbf{p}_i'\boldsymbol{\delta}$ where $\boldsymbol{\delta}$ is an unknown $(r \times 1)$ vector of parameters and $\mathbf{p}_i'$ is the corresponding $(1 \times r)$ row incidence vector of qualitative and/or continuous covariates. Just as with residual variances, the RC variances $\sigma^2_{u_{ki}}$ for $k=1,...,K$, are also described via a structural model : $\ln\sigma^2_{u_{ki}} = \mathbf{h}_{ki}'\boldsymbol{\eta}_k$ where $\boldsymbol{\eta}_k$ is an unknown vector of parameters and $\mathbf{h}_{ki}'$ is the corresponding row incidence vector of qualitative and/or continuous covariates.

For this sort of models, REML provides a natural approach for the estimation of fixed effects and all covariance components. To compute REML estimates, an "expectation-maximization" (EM) algorithm was applied (Dempster *et al.*, 1977; Foulley and Quaas, 1995; Foulley, 1997).

Let $\boldsymbol{\gamma} = (\boldsymbol{\delta}', \boldsymbol{\eta}_k', \boldsymbol{\rho}', \boldsymbol{\beta}')'$ denote the vector of parameter with $\boldsymbol{\rho} = \{\rho_{kk'}\}$. The application of the EM algorithm involves the definition of a vector of complete data $\mathbf{x}$ (where $\mathbf{x}$ includes the data vector and the vector of random effects of the model except the residual effect) and on the definition of the corresponding likelihood function $L(\boldsymbol{\gamma};\mathbf{x}) = \ln p(\mathbf{x} \mid \boldsymbol{\gamma})$. $L(\boldsymbol{\gamma};\mathbf{x})$ can be decomposed as the sum of the log-likelihood of $\mathbf{u}^*$ as a function of $\mathbf{G} = \mathrm{Var}(\mathbf{u}^*)$ with $\mathbf{u}^* = \{\mathbf{u}_k^*\}$ and $\mathbf{u}_k^* = \{u_{kl}^*\}$, and of the log-likelihood of $\mathbf{e}_i$ as a function of $\mathbf{R} = \mathbf{I}\sigma^2_{e_i}$. The E step consists of evaluating the function $Q(\boldsymbol{\gamma} \mid \boldsymbol{\gamma}^{[t]}) = E[L(\boldsymbol{\gamma};\mathbf{x}) \mid \mathbf{y}, \boldsymbol{\gamma}^{[t]}]$ where $\boldsymbol{\gamma}^{[t]}$ is the current estimate of $\boldsymbol{\gamma}$ at iteration [t] and E[.] designates the conditional expectation of $L(\boldsymbol{\gamma};\mathbf{x})$ given the data ***y***, and the current values of the parameters: $\boldsymbol{\delta} = \boldsymbol{\delta}^{[t]}$, $\boldsymbol{\eta}_k = \boldsymbol{\eta}_k^{[t]}$, $\boldsymbol{\rho} = \boldsymbol{\rho}^{[t]}$ and $\boldsymbol{\beta} = \boldsymbol{\beta}^{[t]}$. The M step consists of updating $\boldsymbol{\gamma}$ (i.e., compute $\boldsymbol{\gamma}^{[t+1]}$) by maximizing $Q(\boldsymbol{\gamma} \mid \boldsymbol{\gamma}^{[t]})$ with respect to $\boldsymbol{\gamma}$. The function to be maximized can be written as :

$$Q(\boldsymbol{\gamma} \mid \boldsymbol{\gamma}^{[t]}) = C - \frac{1}{2}\sum_{i=1}^{p} n_i \ln(\sigma^2_{e_i}) - \frac{1}{2}\sum_{i=1}^{p}\sigma^{-2}_{e_i} E_c^{[t]}[\mathbf{e}_i'\mathbf{e}_i] - \frac{1}{2}\ln|\mathbf{G}| - \frac{1}{2}E_c^{[t]}[\mathbf{u}^{*'}\mathbf{G}^{-1}\mathbf{u}^*]$$

where $\mathbf{e}_i = \mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \sum_{k=1}^{K}\sigma_{u_{ki}}\mathbf{Z}_{ki}\mathbf{u}_k^*$, C is a constant, n_i is the number of records in subclass i, $E_c^{[t]}[.]$ is a condensed notation for a conditional expectation taken with respect to the distribution of the complete data ***x*** given the observation ***y*** and the parameter $\boldsymbol{\gamma}$ set at their current value $\boldsymbol{\gamma}^{[t]}$, and $\mathbf{G} = \mathrm{Var}(\mathbf{u}^*) = \mathbf{G}_0 \otimes \mathbf{I}_q$ where $\mathbf{G}_0$ is a correlation matrix with (k,k') element : $g_{0,kk'} = \rho_{kk'}$ with k and $k' = 1,2,...,K$.

Now, we can decompose $\mathbf{Q}(\boldsymbol{\gamma} \mid \boldsymbol{\gamma}^{[t]})$ into two parts which can be written as:

$$-2\,Q_e(\mathbf{R} \mid \boldsymbol{\gamma}^{[t]}) = \sum_{i=1}^{p} n_i \ln(\sigma^2_{e_i}) + \sum_{i=1}^{p}\sigma^{-2}_{e_i} E_c^{[t]}[\mathbf{e}_i'\mathbf{e}_i], \text{and } -2\,Q_u(\mathbf{G} \mid \boldsymbol{\gamma}^{[t]}) = \ln|\mathbf{G}| + E_c^{[t]}[\mathbf{u}^{*'}\mathbf{G}^{-1}\mathbf{u}^*].$$

Note that $\mathbf{Q}_u$ depends only on $\boldsymbol{\rho}$. Thus, the maximisation of $\mathbf{Q}(\boldsymbol{\gamma} \mid \boldsymbol{\gamma}^{[t]})$ with respect to $\boldsymbol{\rho}$ is reduced to the maximisation of Q_u with respect to $\boldsymbol{\rho}$. The REML estimates can be obtained efficiently via the Newton-Raphson algorithm for $\boldsymbol{\delta}$ and $\boldsymbol{\eta}_k$ (for $k=1,...,K$) estimates and via the Fisher scoring algorithm for parameter $\boldsymbol{\rho}$ (vector of correlations $\rho_{kk'}$).

Numerically, the current estimates $\boldsymbol{\delta}^{[t+1]}$ and $\boldsymbol{\eta}_k^{[t+1]}$ for $k=1,...,K$ of $\boldsymbol{\delta}$ and $\boldsymbol{\eta}_k$ are computed with the following iterative system : $\left(\frac{\partial^2 Q_e}{\partial \boldsymbol{\gamma}^2} \right)^{[t]} (\boldsymbol{\gamma}^{[t+1]} - \boldsymbol{\gamma}^{[t]}) = \left(-\frac{\partial Q_e}{\partial \boldsymbol{\gamma}} \right)^{[t]}$ (Robert-Granié *et al.*, 2002). In the general case, $-2\,\mathrm{Q}_u(\mathbf{G} \mid \boldsymbol{\gamma}^{[t]}) = \ln \mid \mathbf{G} \mid + \mathrm{E}_c^{[t]}[\mathbf{u}^{*'}\mathbf{G}^{-1}\mathbf{u}^*] = q \ln \mid \mathbf{G}_0 \mid + \mathrm{tr}[\mathbf{G}_0^{-1}\mathbf{D}]$ with $\mathbf{D} = \mathrm{E}_c\left(\mathbf{u}^{*'}\mathbf{u}^*\right)$ and the current estimate of $\boldsymbol{\rho}$ is computed from the following equation : $\mathrm{E}\left(\frac{\partial^2 \mathrm{Q}_u}{\partial \boldsymbol{\rho}' \partial \boldsymbol{\rho}} \right)^{[t]} (\boldsymbol{\rho}^{[t+1]} - \boldsymbol{\rho}^{[t]}) = \left(-\frac{\partial \mathrm{Q}_u}{\partial \boldsymbol{\rho}} \right)^{[t]}$ where $\frac{\partial(-2\,\mathrm{Q}_u)}{\partial \rho_{kl}} = q\,\mathrm{tr}\left[(\mathbf{G}_0^{-1} - \mathbf{G}_0^{-1}\mathbf{D}^*\mathbf{G}_0^{-1}) \frac{\partial \mathbf{G}_0}{\partial \rho_{kl}} \right]$

with $\mathbf{D}^* = \{d_{kl}^* = \frac{1}{q} E_c[\mathbf{u}^{*'}\mathbf{u}^*]\}$ and $\mathrm{E}\left(\frac{\partial^2(-2Q_u)}{\partial \rho_{kl}' \partial \rho_{kl}} \right) = q\,\mathrm{tr}\left(\frac{\partial \mathbf{G}_0}{\partial \rho_{kl}} \mathbf{G}_0^{-1} \frac{\partial \mathbf{G}_0}{\partial \rho_{k'l'}} \mathbf{G}_0^{-1} \right)$.

Calculations have been made easier by taking advantage of the simple expression of the Fisher information matrix since $\mathrm{E}\left(\mathbf{D}^*\right) = \mathbf{G}_0$. For $K = 2$, this system reduces to a third degree polynomial equation, i.e. $\rho^3 - d_{12}\rho^2 + (d_{11} + d_{22} - 1)\rho - d_{12} = 0$. If individuals are not independent, one has to replace $\mathbf{G}$ by $\mathbf{G}_0 \otimes \mathbf{A}$ where $\mathbf{A}$ is a symmetric, positive definite matrix of known coefficients. Notice that the M step for the correlation matrix $\boldsymbol{\rho}$ does not reduce to the usual $\mathbf{G}_0$ formula for variance covariance parameters but requires a special treatment. For this kind of models and variance covariance structures, one may also envision to implement the EM algorithm under its PX ("parameter expansion") version (Foulley and Van Dyk, 2000).

DISCUSSION AND CONCLUSION

These procedures have been illustrated via an example in growth performance of beef cattle (Robert-Granié *et al.*, 2002). The aim of this study was to compare the growth curve of animals born singles or twins and to quantify the difference of weight at different ages. Growth of Maine Anjou cattle was described by a third order regression on age for a mean growth curve, using a linear regression with two correlated random effects for the individual profiles plus independent errors. Three sources of heterogeneity of residual variances have been detected.

RC models provide a valuable tool for modelling repeated records in animal breeding adequately, especially if traits measured change gradually over time (e.g., analysis of lactation curves in dairy cattle, of feed intake or growth curves in beef cattle, etc). They not only reduce the number of parameters, as compared to multiple traits but they can easily cope with irregular recording patterns in time.

However, there are critical issues to be aware of in order to use these models properly and efficiently. With respect to fixed effects, a critical question lies in the order of the polynomials used to model response. In many studies especially in animal breeding, the authors assume the same regression structure on the fixed and random effects. This is neither mandatory in theory

nor desirable in practice, since variation between populations and between subjects within populations do not necessarily follow the same pattern. In practice, the order of polynomials for fitting the random part of the model (adjusted profiles) is usually lower than that for the fixed part (population trend). Selecting the polynomial degree at both fixed and random levels is not an easy task and eigenfunctions and eigenvalues of covariance (Kirkpatrick and Heckman, 1989) might be useful tools to do that for the random part (Meyer, 1998). One may also question the relevance of using conventional polynomials vs other types e.g. the fractional polynomials (Petim-Batista *et al.*, 2002; Robert-Granié *et al.*, 2002) which may provide a better adjustment at a lower cost in terms of numbers of parameters. With respect to the random part, dispersion models can also be improved significantly by the application of stochastic time processes to take into account the existing correlations between successive measurements (Diggle *et al.*, 1994 ; Foulley *et al.*, 2000 ; Verbeke and Molenberghs, 2000). This can be easily accommodated in HRC models allowing for instance autoregressive (or exponential) heteroskedastic time processes with variances depending on time and/or population strata (Wolfinger, 1996 ; Meyer, 2001).

REFERENCES

Dempster A.P., Laird N.M. and Rubin D.B. (1977) *J. Royal Stat. Assoc. B.* **39** : 1-38.

Diggle P.J., Liang K.Y. and Zeger S.L. (1994) "Analysis of longitudinal data". Clarendon Press, Oxford.

Foulley J.L. (1997) *Genet. Sel. Evol*. **29** : 297-318.

Foulley J.L. and Quaas R.L. (1995) *Genet. Sel. Evol*. **27**: 211-228.

Foulley J.L. and Van Dyk D.A. (2000) *Genet. Sel. Evol*. **32** : 143-163.

Foulley J.L., Gianola D., San Cristobal M. and Im S. (1990) *J. Dairy Sci.* **73** : 1612-1624.

Foulley J.L., San Cristobal M., Gianola D. and Im S. (1992) *Comput. Stat. Data Anal*. **13** : 291-305.

Foulley J.L., Quaas R.L. and Thaon d'Arnoldi C. (1998) *Genet. Sel. Evol*., **30**: 27-43.

Foulley J.L., Jaffrézic F.and Robert-Granié C. (2000) *Genet. Sel. Evol*. **32** : 129-141.

Gianola D., Foulley J.L., Fernando R.L., *et al*. (1992) *J. Dairy Sci*. **75** : 2805-2823.

Kirkpatrick M. and Heckman N. (1989) *J. Math. Biol.* **27** : 429-450.

Meng X.L. and Rubin D.B. (1993) *Biometrika*, **80** : 267-278.

Meyer K. (1998) *Genet. Sel. Evol*. **30** : 221-240.

Meyer K. (2001) *Genet. Sel. Evol*. **33** : 557-585.

Patterson H.D. and Thompson R. (1971) *Biometrika* **58** : 545-554.

Petim-Batista F., Foulley J.L., Robert-Granié C., Silvestre A. and Colaco J. (2002) *7th WCGALP, Montpellier.*

Robert-Granié C., Bonaiti B., Boichard D. and Barbat A. (1999) *Livest. Prod. Sci*. **60** :343-357.

Robert-Granié C., Heude B. and Foulley J.L. (2002) *Genet. Sel. Evol*, in press.

Robert-Granié C., Maza E., Rupp R. and Foulley J.L. (2002) *7th WCGALP, Montpellier.*

San Cristobal M., Foulley J.L. and Manfredi E. (1993) *Genet. Sel. Evol.* **25** : 3-30.

Schaeffer L.R. and Dekkers J.C.M. (1994) *Proc. 5th WCGALP* **18** : 443-446.

Verbeke G. and Molenberghs G. (2000) "Linear mixed models for longitudinal data". Springer Verlag, New-York.

Wolfinger R.D. (1996*) J. Agricult. Biol. Environ. Statist.* **1** : 205-230.