Modelling the growth curve of Maine-Anjou beef cattle using heteroskedastic random coefficients models

C. Robert-Granié, B. Heude, Jean-Louis Foulley

Genet. Sel. Evol. 34 (2002) 423–445

DOI: 10.1051/gse:2002016

Original article

Modelling the growth curve of Maine-Anjou beef cattle using heteroskedastic random coefficients models

Christèle ROBERT-GRANIÉ[a*], Barbara HEUDE[b], Jean-Louis FOULLEY[b]

[a] Station d'amélioration génétique des animaux, Institut national de la recherche agronomique, BP 27, 31326 Castanet-Tolosan, France
[b] Station de génétique quantitative et appliquée, Institut national de la recherche agronomique, 78352 Jouy-en-Josas Cedex, France

(Received 6 April 2001; accepted 7 January 2002)

Abstract – A heteroskedastic random coefficients model was described for analyzing weight performances between the 100th and the 650th days of age of Maine-Anjou beef cattle. This model contained both fixed effects, random linear regression and heterogeneous variance components. The objective of this study was to analyze the difference of growth curves between animals born as twin and single bull calves. The method was based on log-linear models for residual and individual variances expressed as functions of explanatory variables. An expectation-maximization (EM) algorithm was proposed for calculating restricted maximum likelihood (REML) estimates of the residual and individual components of variances and covariances. Likelihood ratio tests were used to assess hypotheses about parameters of this model. Growth of Maine-Anjou cattle was described by a third order regression on age for a mean growth curve, two correlated random effects for the individual variability and independent errors. Three sources of heterogeneity of residual variances were detected. The difference of weight performance between bulls born as single and twin bull calves was estimated to be equal to about 15 kg for the growth period considered.

heteroskedastic random coefficient model / EM-REML / robust estimators / growth curve / Maine-Anjou breed

* Correspondence and reprints
E-mail: robert@germinal.toulouse.inra.fr

1. INTRODUCTION

The weight performances of animals, recorded repeatedly during their lives, are a typical example of longitudinal data where the trait of interest is changing, gradually but continually, over time. Until recently in quantitative genetics, such records were frequently analysed fitting a so called "repeatability model", *i.e.* assuming all records were repeated measurements of a single trait with constant variances. Other approaches have been (i) to, somewhat arbitrarily, subdivide the range of ages and consider individual segments to represent different traits in a multivariate analysis or (ii) to fit a standard growth curve to the records and analyse the parameters of the growth curve as new traits.

Recently, there has been a great interest in random coefficient models [22] for the analysis of such data. 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. Instead of considering only the overall growth curve, we assume that there is a separate growth curve for each individual. These have by and large been ignored in animal breeding applications so far, although they are common in other areas (see, for example, [22] for a general exposition). Repeated measurements on the same animal are more closely correlated than two measurements on different animals, and the correlation between repeated measurements may decrease as the time between them increases. Therefore, the statistical analysis of repeated measures data must address the issue of covariation between measures on the same unit. Modeling the covariance structure of repeated measurements correctly is of importance for drawing correct inference from such data [5]. The main advantages of longitudinal studies are increased power and robustness to model selection [6]. In animal genetics, random regressions in a linear mixed model context have been considered by Schaeffer and Dekkers [36]. Moreover, the recently developped SAS procedure PROC MIXED greatly increases the popularity of linear mixed models [40].

In quantitative genetics and animal breeding, heteroskedasticity has recently generated much interest. In fact, the assumption of homogeneous variances in linear mixed models may not always be appropriate. There is now a large amount of experimental evidence of heterogeneous variances for most important livestock production traits [14,33,43,44]. Major theoretical and applied work has been carried out for estimating and testing sources of heterogeneous variances arising in univariate mixed models [4,9,11,12,15,30,31, 34,45].

In this paper, we extend the random regression model to a more general class of models termed the heteroskedastic random regression. This class of models assumes that all variances of random effects can be heterogeneous. Inference is based on likelihood procedures (REML, restricted maximum likelihood,

[29]) and estimating equations derived from the expectation-maximization (EM, [2]) theory, more precisely the expectation/conditional maximization (ECM) algorithm recently introduced by Meng and Rubin [23].

The selection of a global model requires the choice of fixed effects (model on phenotypic mean vector E) and the choice of random effects (model on variance-covariance matrix V). In fact, this choice is complex because the choice of fixed effects depends on variance-covariance structure of observations, and in particular on the number of random effects included in the model. In practice, the strategy adopted is as follows: a structure of variance-covariance matrix V is assumed and a model E is chosen (selection of significant fixed effects) and subsequently, with a model E fixed, different structures for V are tested. One alternative approach consists of obtaining an inference on fixed effects by robust estimators (so-called "sandwich estimator", [21]) with respect to the structure on V. In this paper, the theory of the "sandwich estimator" is presented and used to select significant fixed effects.

These procedures are illustrated and presented *via* an example in growth performance of beef cattle. The aim of this study was to compare the growth curve of animals born as singles or twins and to quantify the difference of weight at different ages. The data analyzed in this paper comprised 943 weight records of 127 animals of the Maine-Anjou breed and are presented in the section "Materials and methods". The methods section encompasses models, estimation procedures and tests of hypotheses. Then, the results of the beef cattle example are presented and discussed. The paper ends with concluding remarks on longitudinal data analysis *via* random coefficient models.

2. MATERIALS AND METHODS

2.1. Data

All animals were raised at the experimental Inra herd of "La Grêleraie" (Mayenne, France). This herd is part of a research project aimed at increasing the rate of natural twin calvings in cattle. From an economic point of view, breeders are also concerned with a comparison of growth performance of bull calves born as twins or single. Data consisted of 943 weight performances recorded between 100 and 650 days of age in 127 Maine-Anjou bulls (103 animals born as singles and 24 born as twins). There were on average 7 weight records per animal. The distribution of the number of records per animal and all characteristics of the data set analysed are presented in Table I.

The animals were grouped by year of birth and calving season. For each performance of an animal, the weight, the age at weighting, the calving parity of the mother and the birth status (single *vs.* twin) were recorded. These variables are presented in Table I.

Table I. Characteristics of the data set.

(a)

Number of records per animal	Number of animals born as single	Number of animals born as twins
4	15	6
5	10	3
6	12	2
7	18	1
8	8	3
9	17	3
10	14	2
11	7	2
12	2	0
13	0	2
	103	24

(b)

Season of birth	Number of animals born as single	Number of animals born as twins
1- Autumn	58	12
2- Spring	45	12
	103	24

(c)

Year of birth	Number of animals born as single	Number of animals born as twins
1990	4	0
1991	8	4
1992	14	2
1993	12	3
1994	11	0
1995	15	3
1996	21	5
1997	18	7
	103	24

(d)

Rank of calving of the mother	Number of animals born as single	Number of animals born as twins
1- Heifers	36	5
2- Ranks 2 and 3	38	10
3- Ranks ≥ 4	29	9
	103	24

2.2. Models

In this data set, animals can differ both in the number of records and in time intervals between them. One of the frequently used approaches is the linear mixed effects model [19] in which the repeated measurements are modeled using a linear regression model, with parameters allowed to vary over individuals and therefore called random effects.

2.2.1. Models for data

To characterize the effect of twinning on the growth curve between days 100 and 650, a mixed linear model including random effects and heterogeneous variances was used. The classical random coefficient model involves a random intercept and slope for each subject. The model considered here combines random regression with heteroskedastic variances; it can be written as follows:

$$y_{ijl} = \boldsymbol{x}'_{ijl}\boldsymbol{\beta} + \sigma_{u_{1i}} z_{1ijl} u^*_{1l} + \sigma_{u_{2i}} z_{2ijl} u^*_{2l} + e_{ijl} \tag{1}$$

where y_{ijl} is the jth ($j = 1, \ldots, n_k$) measurement recorded on the lth ($l = 1, \ldots, q$) individual at time t_{jl} in subclass i of the factor of heterogeneity ($i = 1, \ldots, p$); $\boldsymbol{x}'_{ijl}\boldsymbol{\beta}$ represents the systematic component expressed as a linear combination of explanatory variables ($\boldsymbol{x}'_{ijl}$) with unknown linear coefficients ($\boldsymbol{\beta}$); ($\sigma_{u_{1i}} z_{1ijl} u^*_{1l} + \sigma_{u_{2i}} z_{2ijl} u^*_{2l}$) represents the additive contribution of two random regression factors (u^*_{1l} is the intercept effect and u^*_{2l} is the slope effect) on covariable information (z_{1ijl} and z_{2ijl}) and which are specific to each lth individual; $\sigma_{u_{1i}}$ and $\sigma_{u_{2i}}$ are the corresponding components of variance pertaining to stratum i. The random effects u^*_{1l} and u^*_{2l} are correlated and this correlation is assumed homogeneous over strata and equal to ρ. The e_{ijl} represent independent errors.

In matrix notation, the model can be expressed as:

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \sigma_{u_{1i}} \boldsymbol{Z}_{1i}\boldsymbol{u}^*_1 + \sigma_{u_{2i}} \boldsymbol{Z}_{2i}\boldsymbol{u}^*_2 + \boldsymbol{e}_i \tag{2}$$

where $\boldsymbol{u}^*_1 = (u^*_{11}, \ldots, u^*_{1l}, \ldots, u^*_{1q})'$ is the vector of normally distributed standardized intercept values $N(\boldsymbol{0}, \boldsymbol{I}_q)$, $\boldsymbol{u}^*_2 = (u^*_{21}, \ldots, u^*_{2l}, \ldots, u^*_{2q})'$ is the vector of normally distributed standardized slope effects $N(\boldsymbol{0}, \boldsymbol{I}_q)$, and $\boldsymbol{e}_i$ is the vector of normally distributed residuals for stratum i $N(\boldsymbol{0}, \boldsymbol{I}\sigma^2_{e_i})$. The regression components ($\boldsymbol{u}^*_1$ and $\boldsymbol{u}^*_2$) and environmental effects $\boldsymbol{e}_i$ are assumed to be independent. Then,

$$var \begin{pmatrix} \boldsymbol{u}^*_1 \\ \boldsymbol{u}^*_2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{I}_q & \rho\boldsymbol{I}_q \\ \rho\boldsymbol{I}_q & \boldsymbol{I}_q \end{pmatrix}$$

with the correlation coefficient ρ defined as previously.

It would have been possible to introduce additional levels of random coefficients without any difficulty. But for the sake of simplicity, this example only

considers two random regression components: the standard random intercept-slope model. However, the equations shown in the appendix apply to both general ($k = 1, \ldots, K$) and particular ($K = 2$) cases.

More generally, a heteroskedastic random coefficient model with K random coefficient components can be written as follows:

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \sum_{k=1}^{K} \sigma_{u_{ki}} \boldsymbol{Z}_{ki}\boldsymbol{u}_k^* + \boldsymbol{e}_i.$$

2.2.2. Models for variances

There are situations where variances are heterogeneous, *i.e.*, variances are assumed to vary according to several factors. A convenient and parsimonious procedure to handle heterogeneity of variances is to model them *via* a log-linear function [20,27]. This approach has the advantage of maintaining parameter independence between the mean and covariance structure. As compared to transformations, it also avoids "to destroy a simple linear mean relationship making the interpretation and estimation of the mean and covariance parameters more difficult..." [46].

In the heteroskedastic model, residual variances ($\sigma^2_{e_i}$), for example, were assumed to vary according to several factors such as twinning, season of birth, rank of calving of the mother, age at weight. The idea was to find a model for the variance that describes the heterogeneity among p different subclasses (usually a large number in animal breeding) in terms of a few parameters. Following Foulley *et al.* [10] and San Cristobal *et al.* [34] among others, the residual variances were modeled as:

$$\ln \sigma^2_{e_i} = \boldsymbol{p}_i'\boldsymbol{\delta}$$

where $\boldsymbol{\delta}$ is an unknown ($r \times 1$) vector of parameters, and $\boldsymbol{p}_i'$ is the corresponding ($1 \times r$) row incidence vector of qualitative (*e.g.*, twinning, rank of calving of the mother) or continuous covariates (*e.g.*, age at weight).

Just as was done with the residual variances, the individual variances $\sigma^2_{u_{1i}}$ and $\sigma^2_{u_{2i}}$ can be heteroskedastic and are also modeled with a structural model [11]:

$$\ln \sigma^2_{u_{1i}} = \boldsymbol{h}'_{1i}\boldsymbol{\eta}_1$$
$$\ln \sigma^2_{u_{2i}} = \boldsymbol{h}'_{2i}\boldsymbol{\eta}_2$$

where $\boldsymbol{\eta}_j$ with $j = (1, 2)$ is an unknown vector of parameters and $\boldsymbol{h}'_{ji}$ is the corresponding row incidence vector of qualitative or continuous covariates.

2.3. Estimation of dispersion parameters

For the model developed in this paper, REML (restricted maximum likelihood, [29]) provides a natural approach for the estimation of fixed effects and all (co)variance components. To compute REML estimates, a generalized expectation-maximization (EM) algorithm was applied [7,8,11]. The theory of this method is described by Dempster *et al.* [2].

Let $\boldsymbol{\gamma} = (\boldsymbol{\delta}', \boldsymbol{\eta}_1', \boldsymbol{\eta}_2', \rho)'$ denote the vector of parameters. The application of the generalized EM algorithm is based on the definition of a vector of complete data $\boldsymbol{x}$ (where $\boldsymbol{x}$ includes the data vector and the vector of fixed and random effects of the model, except the residual effect) and on the definition of the corresponding likelihood function $L(\boldsymbol{\gamma}; \boldsymbol{x}) = \ln p(\boldsymbol{x}|\boldsymbol{\gamma})$. $L(\boldsymbol{\gamma}; \boldsymbol{x})$ can be decomposed as the sum of the log-likelihood Q_u of $\boldsymbol{u}^*$ as a function of ρ and of the log-likelihood Q_e of $\boldsymbol{e}$ as a function of $\boldsymbol{\delta}$, $\boldsymbol{\eta}_1$, $\boldsymbol{\eta}_2$. The E step consists of computing the function $Q(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{[t]}) = E[L(\boldsymbol{\gamma}; \boldsymbol{x})|\boldsymbol{y}, \boldsymbol{\gamma}^{[t]}]$ where $\boldsymbol{\gamma}^{[t]}$ is the current estimate of $\boldsymbol{\gamma}$ at iteration $[t]$ and $E[.]$ is the conditional expectation of $L(\boldsymbol{\gamma}; \boldsymbol{x})$ given the data $\boldsymbol{y}$, $\boldsymbol{\delta} = \boldsymbol{\delta}^{[t]}$, $\boldsymbol{\eta}_1 = \boldsymbol{\eta}_1^{[t]}$, $\boldsymbol{\eta}_2 = \boldsymbol{\eta}_2^{[t]}$, $\rho = \rho^{[t]}$. The M step consists of selecting the next value $\boldsymbol{\gamma}^{[t+1]}$ of $\boldsymbol{\gamma}$ by maximizing $Q(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{[t]})$ with respect to $\boldsymbol{\gamma}$. The function to be maximized could be written as:

$$Q(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{[t]}) = C - \frac{1}{2}\sum_{i=1}^{p} n_i \ln(\sigma_{e_i}^2) - \frac{1}{2}\sum_{i=1}^{p} \sigma_{e_i}^{-2} E_c^{[t]}[\boldsymbol{e}_i'\boldsymbol{e}_i] - \frac{1}{2}\ln|\boldsymbol{G}| - \frac{1}{2}E_c^{[t]}[\boldsymbol{u}^{*\prime}\boldsymbol{G}^{-1}\boldsymbol{u}^*] \quad (3)$$

where $\boldsymbol{e}_i = \boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \sigma_{u_{1i}}\boldsymbol{Z}_{1i}\boldsymbol{u}_1^* - \sigma_{u_{2i}}\boldsymbol{Z}_{2i}\boldsymbol{u}_2^*$, 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 $\boldsymbol{x}$ given the observation $\boldsymbol{y}$ and the parameter $\boldsymbol{\gamma}$ set at their current value $\boldsymbol{\gamma}^{[t]}$.

For example,

$$E_c^{[t]}[\boldsymbol{e}_i] = \boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \sigma_{u_{1i}}\boldsymbol{Z}_{1i}E[\boldsymbol{u}_1^*|\boldsymbol{y}, \gamma = \gamma^{[t]}] - \sigma_{u_{2i}}\boldsymbol{Z}_{2i}E[\boldsymbol{u}_2^*|\boldsymbol{y}, \gamma = \gamma^{[t]}].$$

For more complex functions, the same rules apply as shown in the appendix.

And $\boldsymbol{G} = var(\boldsymbol{u}^*) = var\begin{pmatrix}\boldsymbol{u}_1^* \\ \boldsymbol{u}_2^*\end{pmatrix} = \begin{pmatrix}\boldsymbol{I}_q & \rho\boldsymbol{I}_q \\ \rho\boldsymbol{I}_q & \boldsymbol{I}_q\end{pmatrix}$.

$Q(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{[t]})$ can be decomposed into two parts:

$$Q(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{[t]}) = C + Q_e + Q_u \quad (4)$$

where

$$-2Q_e = \sum_{i=1}^{p} n_i \ln(\sigma_{e_i}^2) + \sum_{i=1}^{p} \sigma_{e_i}^{-2} E_c^{[t]}[\boldsymbol{e}_i'\boldsymbol{e}_i]$$

and

$$-2Q_u = \ln|\boldsymbol{G}| + E_c^{[t]}[\boldsymbol{u}^{*\prime}\boldsymbol{G}^{-1}\boldsymbol{u}^*].$$

Note that Q_u depends only on ρ. Thus, the maximisation of $Q(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{[t]})$ with respect to ρ is reduced to the maximisation of Q_u with respect to ρ.

The REML estimates can be obtained efficiently *via* the Newton-Raphson algorithm for $\boldsymbol{\delta}$, $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ estimates and *via* the Fisher scoring algorithm for the parameter ρ. The corresponding systems of equations and their necessary inputs are shown in the appendix.

2.4. Tests of hypotheses

Tests of hypotheses involving fixed effects are more complex in mixed than in fixed effects models. The intuitive reason is clear: the fixed effects model has only one variance component and all fixed effects are tested against the error variance; a mixed model, however, contains different variance components and a particular fixed effects hypothesis must be tested against the appropriate background variability which can be expressed in terms of variance components present in a model.

Fitting linear mixed models implies that an appropriate mean structure as well as covariance structure needs to be specified. They are not independent of each other. Adequate covariance modeling is not only useful for the interpretation of the variation in the data, it is essential to obtaining valid inferences for the parameters in the mean structure. An incorrect covariance structure also affects predictions [1]. On the contrary, since the covariance structure models all variability in the data which is not explained by systematic trends, it highly depends on the specified mean structure.

2.4.1. Testing fixed effects

An approach based on robust estimators ("sandwich estimators", [21]) was chosen to select significant fixed effects. This method is defined as follows:

Let $\boldsymbol{\alpha}$ denote the vector of all variance and covariance parameters found in $\boldsymbol{V}$. If $\boldsymbol{y} \sim N(\boldsymbol{\mu}, \boldsymbol{V})$ with $\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ is known, the maximum likelihood estimator of $\boldsymbol{\beta}$, obtained by maximizing the likelihood function of $\boldsymbol{y}$ conditional on $\boldsymbol{\alpha}$, is given by [19]:

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{N} \boldsymbol{X}_i'\boldsymbol{W}_i\boldsymbol{X}_i\right)^{-1} \left(\sum_{i=1}^{N} \boldsymbol{X}_i'\boldsymbol{W}_i\boldsymbol{y}_i\right) \tag{5}$$

and its variance-covariance matrix equals:

$$Var(\hat{\boldsymbol{\beta}}) = \left(\sum_{i=1}^{N} \boldsymbol{X}_i' \boldsymbol{W}_i \boldsymbol{X}_i\right)^{-1} \left(\sum_{i=1}^{N} \boldsymbol{X}_i' \boldsymbol{W}_i Var(\boldsymbol{y}_i) \boldsymbol{W}_i \boldsymbol{X}_i\right) \left(\sum_{i=1}^{N} \boldsymbol{X}_i' \boldsymbol{W}_i \boldsymbol{X}_i\right)^{-1} \quad (6)$$

$$Var(\hat{\boldsymbol{\beta}}) = \left(\sum_{i=1}^{N} \boldsymbol{X}_i' \boldsymbol{W}_i \boldsymbol{X}_i\right)^{-1} \quad (7)$$

where $\boldsymbol{W}_i$ equals $\boldsymbol{V}_i^{-1}$.

Note that a sufficient condition for (5) to be unbiased is that the mean $E(\boldsymbol{y}_i)$ is correctly specified as $\boldsymbol{X}_i\boldsymbol{\beta}$. However, the equivalence of (6) and (7) holds under the assumption that the covariance matrix is correctly specified. Thus, an analysis based on (7) will not be robust with respect to model deviations in the covariance structure. Therefore Liang and Zeger [21] propose inferential procedures based on the so-called "sandwich estimator" for $Var(\hat{\boldsymbol{\beta}})$, obtained by replacing $Var(\boldsymbol{y}_i)$ by $(\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}})(\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}})'$. Liang and Zeger [21] showed that the resulting estimator of $\boldsymbol{\beta}$ is consistent, as long as the mean is correctly specified in the model. To that respect the simplest choice consists of $\hat{\boldsymbol{\beta}}$ in (5) fitted by ordinary least squares, *i.e.*, $\hat{\boldsymbol{\beta}} = (\sum_{i=1}^{N} \boldsymbol{X}_i'\boldsymbol{X}_i)^{-1}(\sum_{i=1}^{N} \boldsymbol{X}_i'\boldsymbol{y}_i)$. However, it might be worthwhile to consider more complex structures for the working dispersion matrix $\boldsymbol{W}_i$, or generalized least squares estimation.

When $\boldsymbol{\alpha}$ is not known but an estimate $\hat{\boldsymbol{\alpha}}$ is available, we can set $\hat{\boldsymbol{V}}_{\mathbf{i}} = \boldsymbol{V}_i(\hat{\boldsymbol{\alpha}}) = \hat{\boldsymbol{W}}_{\mathbf{i}}^{-\mathbf{1}}$ and estimate $\boldsymbol{\beta}$ by using the expression (5) in which $\boldsymbol{W}_i$ is replaced by $\hat{\boldsymbol{W}}_{\mathbf{i}}$. Estimates of the standard errors of $\hat{\boldsymbol{\beta}}$ can then be obtained by replacing $\boldsymbol{\alpha}$ by $\hat{\boldsymbol{\alpha}}$ in (6) and in (7) respectively, which are both available in the SAS MIXED procedure [35]. However, as noted by Dempster *et al.* [3], they underestimate the variability introduced by estimating $\boldsymbol{\alpha}$. The SAS MIXED procedure accounts to some extend for this downward bias by providing approximate t- and F-statistics for testing about $\boldsymbol{\beta}$ [18].

Practically, the resulting standard errors can be requested in the SAS MIXED procedure by adding the option "empirical" in the proc mixed statement. Note that this option does not affect the standard errors reported for the variance component in the model. For some fixed effects, however, the robust standard errors tend to be somewhat smaller than the model-based standard errors, leading to less conservative inferences for the fixed effects in the final model, but for others, there are larger with opposite effects on the real size of the test [41]. In any case, this procedure relies on asymptotic properties and therefore should be applied with at least a minimum number of individuals (about 100).

In this study, comparisons between robust and standard estimators will be presented for different homogeneous models: (0) a fixed effect model with

independent errors, (1) a classical mixed model with one random effect and independent errors, (2) a fixed effect model with errors following a first order autoregressive process and (3) a random coefficient model with two correlated random effects (intercept and slope effects) and independent errors.

After selection of fixed effects in the model, random effects and factors of heterogeneity can be tested.

2.4.2. Testing random effects

Although the estimation of the parameters in the model is generally the main interest in an analysis, tests of hypotheses are usually required in assessing the significance of effects and in model selection. Tests of significance of random effects usually involve testing whether a single variance component is 0. For example, testing the significance of a random-intercept effect involves testing whether $\sigma^2_{u_1} = 0$. These tests are carried out by using residual maximum likelihood ratio tests. However, the null hypothesis places the parameter on the boundary of the parameter space and the non-regular likelihood ratio theory is required [37]. Stram and Lee [38] considered the specific issue of tests concerning variance components and random coefficients.

For a single variance component, the asymptotic distribution of the likelihood ratio test is a mixture of a Dirac mass at zero and of a chi-square with a single degree of freedom with mixing probabilities equal to 0.5 [38]. The approximate *P*-value for the residual likelihood ratio statistic $\delta = -2\log(\Lambda)$ is easily calculated as $0.5Pr(X > d)$ where $X \sim \chi^2_1$ under the null hypothesis and d is the observed value of δ. The residual maximum likelihood ratio test for the test that p variance components are 0 involves a mixture of χ^2-variates from 0 to p degrees of freedom. The mixing probabilities depend on the geometry of the situation [37]. Stram and Lee [38] found that the likelihood ratio test is conservative and for the residual maximum likelihood ratio test this was confirmed in a limited simulation study reported in Verbyla *et al.* [42]. A similar application was presented in Robert-Granié *et al.* [32].

3. RESULTS AND DISCUSSION

3.1. Plot of data

With longitudinal data, an obvious first graph to consider is the scatterplot of the weight of animals against time. Figure 1 displays the data on weight of bulls in relation to age at weight. This simple graph reveals several important patterns. All bulls gained weight. The spread among all animals was substantially smaller at the beginning of the study than at the end. This pattern of increasing variance over time could be explained in terms of variation in the growth rates of the individual animals. In the case of the beef cattle data,

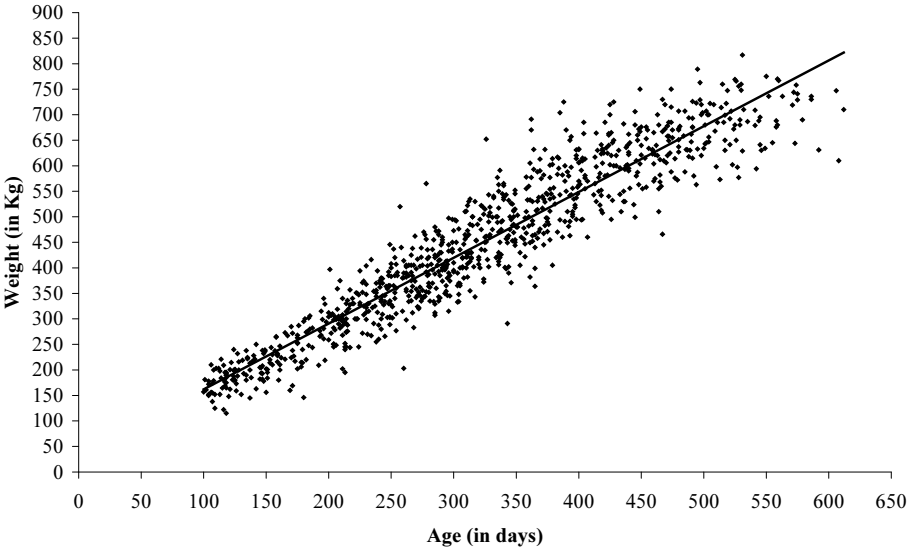

Figure 1. Growth curve of Maine Anjou beef cattle.

the choice of a linear function between the 100th and the 650th days seemed appropriate for fitting the mean growth curve.

3.2. Model selection

As explained in the section "Tests of hypotheses", fixed effects were selected using robust estimators [21]. Comparisons between robust and standard estimators are presented for four homogeneous models with different structures of the variance-covariance matrix. The four models chosen are traditional models in longitudinal data analysis [13]:

(0) a fixed effects model: $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{e}$, with independent errors, with $\boldsymbol{y}$ normally distributed, and with a variance-covariance matrix equal to $\boldsymbol{I}\sigma_e^2$;

(1) a classical homogeneous mixed model: $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{e}$, with $\boldsymbol{u} \sim N(\boldsymbol{0}, \boldsymbol{I}\sigma_u^2)$ and $\boldsymbol{e} \sim N(\boldsymbol{0}, \boldsymbol{I}\sigma_e^2)$;

(2) a fixed effect model: $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{e}$, with first order autoregressive errors, $\boldsymbol{e} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma})$ where $\Sigma_{ij} = \sigma_e^2 \rho^{|t_i - t_j|}$, ρ is a real positive number, and $|t_i - t_j|$ representing the distance between measurements i and j of the same animal. The error term corresponds to the contribution of a stationary Gaussian time process, where the correlation between repeated measurements decreases as the time between them increases;

(3) a random coefficient model: $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Z}_1\boldsymbol{u}_1 + \boldsymbol{Z}_2\boldsymbol{u}_2 + \boldsymbol{e}$, with $\boldsymbol{u}_1 \sim N(\boldsymbol{0}, \boldsymbol{I}\sigma_{u_1}^2)$, $\boldsymbol{u}_2 \sim N(\boldsymbol{0}, \boldsymbol{I}\sigma_{u_2}^2)$, $Cov(\boldsymbol{u}_1, \boldsymbol{u}_2) = \sigma_{12}$, and $\boldsymbol{e} \sim N(\boldsymbol{0}, \boldsymbol{I}\sigma_e^2)$.

For each model, all fixed effects were tested. Table II presents the value of the F-test and the *P*-value associated with each fixed effect and each model

Table II. Selection of fixed effects.

Fixed effects	Model (0)		Model (1)		Model (2)		Model (3)	
	F[a]	*P*-value	F[a]	*P*-value	F[a]	*P*-value	F[a]	*P*-value
Twins	[b]5.57	0.0209	1.86	0.1726	1.54	0.2192	2.67	0.1027
	8.11	0.0057	4.52	0.0337	5.78	0.0187	6.10	0.0138
Rank of calving	[b]3.90	0.0246	1.40	0.2469	1.28	0.2846	1.17	0.3111
	3.75	0.0282	4.06	0.0176	4.86	0.0104	4.28	0.0142
Period of birth	[b]5.34	0.0001	3.40	0.0001	1.47	0.1410	4.90	0.0001
	7.35	0.0001	8.34	0.0001	6.31	0.0001	14.74	0.0001
Age	[b]17.94	0.0001	57.86	0.0001	60.17	0.0001	104.82	0.0001
	22.55	0.0001	26.25	0.0001	62.41	0.0001	48.37	0.0001
Age	[b]9.47	0.0001	24.71	0.0001	6.71	0.0001	10.09	0.0001
* period of birth	10.16	0.0001	10.87	0.0001	10.96	0.0001	14.89	0.0001
Age * twins	[b]0.05	0.8247	0.84	0.3589	0.30	0.5842	0.08	0.7809
	0.10	0.7505	0.60	0.4398	0.63	0.4283	0.16	0.6868
Age	[b]0.07	0.9366	0.02	0.9810	0.01	0.9869	0.04	0.9652
* rank of calving	0.08	0.9210	0.01	0.9919	0.02	0.9828	0.05	0.9535
Twins	[b]0.51	0.6052	0.23	0.7941	0.40	0.6722	0.24	0.7874
* rank of calving	0.25	0.7807	0.65	0.5209	1.01	0.3683	0.49	0.6158
Twins	[b]6.52	0.0001	0.96	0.4732	1.00	0.4462	1.04	0.4045
* period of birth	8.96	0.0001	7.93	0.0001	21.33	0.0001	2.53	0.0073
Rank of calving	[b]11.61	0.0001	1.48	0.0678	1.24	0.2433	1.80	0.0126
* period of birth	149.42	0.0001	147.80	0.0001	143.37	0.0001	44.62	0.0001
Age^2	[b]10.84	0.0010	20.40	0.0001	10.11	0.0015	10.38	0.0017
	12.53	0.0004	8.72	0.0032	10.81	0.0011	4.38	0.0388
Age^3	[b]14.12	0.0002	25.86	0.0001	12.30	0.0005	13.20	0.0004
	14.76	0.0001	10.19	0.0015	12.72	0.0004	5.01	0.0272

Twins: variable representing bulls born as single or twins.
Period of birth: variable combining year and season of birth.
(a) Value of F-test.
(b) **First line**: standard estimator and **second line**: robust estimator.
Model (0): $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{e}$ with errors independent and normally distributed, with variance-covariance structure equal to $\boldsymbol{I}\sigma_e^2$;
Model (1): $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{e}$ with $\boldsymbol{u} \sim N(\boldsymbol{0}, \boldsymbol{I}\sigma_u^2)$ and $\boldsymbol{e} \sim N(\boldsymbol{0}, \boldsymbol{I}\sigma_e^2)$;
Model (2): $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{e}$ with first order autoregressive errors, $\boldsymbol{e} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma})$ where $\Sigma_{ij} = \sigma_e^2 \rho^{|t_i - t_j|}$;
Model (3): $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Z}_1\boldsymbol{u}_1 + \boldsymbol{Z}_2\boldsymbol{u}_2 + \boldsymbol{e}$ with $\boldsymbol{u}_1 \sim N(\boldsymbol{0}, \boldsymbol{I}\sigma_{u_1}^2)$, $\boldsymbol{u}_2 \sim N(\boldsymbol{0}, \boldsymbol{I}\sigma_{u_2}^2)$, $Cov(\boldsymbol{u}_1, \boldsymbol{u}_2) = \sigma_{12}$ and $\boldsymbol{e} \sim N(\boldsymbol{0}, \boldsymbol{I}\sigma_e^2)$.

Table III. Selection of random effects.

Models	$-2L$ [d]	Test	δ [e]	Degree of freedom [f]	Conclusion
(a) Fixed	9127.26				
(b) Random intercept	8462.80	(b) against (a)	664.46	0:1	Significant
(c) Random intercept and slope	8297.44	(c) against (b)	165.36	1:2	Significant

Model (c): model where intercept and slope are assumed correlated.
[d]: -2 log-likelihood.
[e]: Likelihood ratio statistic.
[f]: Asymptotic distribution of the likelihood ratio under the null hypothesis: Chi-square or mixture of Chi-square distributions.

considered. In each case, standard and robust estimators are given. Whatever the method considered, the interactions "age*twins", "age*rank of calving" and "twins*rank of calving" were not significant at the 5% level. The robust method led to the same conclusions whatever models were considered with respect to fixed effects. In contrast, using the standard approach, interactions "rank of calving*period of birth" and "twins*period of birth" were either significant or not significant depending on the structure of the variance-covariance matrix.

Despite the linear trend shown in Figure 1 for the mean growth curve of the animals, age^2 and age^3 were statistically significant, and thus, were kept in the model.

Finally, the list of the fixed effect retained in the model was: age, age^2, age^3, twins, rank of calving, period of birth, age*period of birth, rank of calving*period of birth and twins*period of birth; the non significant interaction age*twins was included in the model because this parameter is of primary interest to evaluate the difference in growth rate between single and twin born bulls.

In a second step, a set of random effects was chosen for the covariance model. A selection of random effects is summarized in Table III. The choice of random effects was based on the set of fixed effects selected with the robust procedure presented in the first step. Likelihood ratio tests (REML version) were used for comparisons among the following models:

(a) a fixed effect model: $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{e}$, with $\boldsymbol{e} \sim N(\boldsymbol{0}, \boldsymbol{I}\sigma_e^2)$;

(b) a classical homogeneous mixed model with a random intercept for each subject: $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{e}$, with $\boldsymbol{u} \sim N(\boldsymbol{0}, \boldsymbol{I}\sigma_u^2)$ and $\boldsymbol{e} \sim N(\boldsymbol{0}, \boldsymbol{I}\sigma_e^2)$;

(c) a homogenous random coefficient model with a random intercept and slope for each animal with two random effects assumed correlated: $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Z}_1\boldsymbol{u}_1 + \boldsymbol{Z}_2\boldsymbol{u}_2 + \boldsymbol{e}$, with $\boldsymbol{u}_1 \sim N(\boldsymbol{0}, \boldsymbol{I}\sigma_{u_1}^2)$, $\boldsymbol{u}_2 \sim N(\boldsymbol{0}, \boldsymbol{I}\sigma_{u_2}^2)$, $Cov(\boldsymbol{u}_1, \boldsymbol{u}_2) = \sigma_{u_{12}}$ and $\boldsymbol{e} \sim N(\boldsymbol{0}, \boldsymbol{I}\sigma_e^2)$.

The results in Table III show large values for the likelihood ratio statistics. The model finally accepted is a homogeneous mixed model with two correlated random effects and independent errors. This model includes a three degree polynomial function in time to describe the mean growth curve; an intercept and a slope for each animal.

From the model defined above (model including as fixed effects age, age^2, age^3, twins, rank of calving, period of birth, age*period of birth, rank of calving*period of birth, age*twins and twins*period of birth and as random effects an intercept and a slope for each animal), sources of heterogeneity (*e.g.*, rank of calving, season of birth, twins or age at weight) were tested on different variances (intercept, slope or residual variances) of the model. Only residual variances were found to be heterogeneous according to rank of calving, season of birth and age at weight. No heterogeneity of variances was observed for individual intercepts and slopes. Final estimates of variance-covariance parameters are presented in Table IV. The correlation between the two random effects is negative and equal to -0.34; *i.e.*, if an animal's intercept is larger than the others, its slope will tend to be smaller as well. The individual variability for the intercept is very large and equal to 827.65. The variance of the slope is equal to 0.012 which corresponds to a value of the coefficient of variation of 12% indicating a rather substantial variability in the growth rate of bulls. The results about the heterogeneity of variances suggest an increasing variance of weight records in time and a larger variability for bulls born in the spring and out of heifers.

3.3. Results for a heteroskedastic random coefficient model

Figure 2 presents the graph of mean growth curves estimated from the last model (heteroskedastic random coefficient model presented in Tab. IV) for bulls born as twins or single. It shows that single born bulls were larger at birth than twins and the weights of both of them increased linearly; the growth difference between single and twin bulls was approximately constant and equal to about 15 kg during the period of growth considered.

Figure 3 shows the differences under two models between the mean growth curves of single born bulls or twins: (a) fixed effects model and (d) heteroskedastic random coefficient model with the same fixed effects as in model (a). The difference between singles and twins shows two opposite patterns: increasing under model (a) and decreasing under model (d). For instance, at 550 days, the difference is estimated to be 11 kg under model (a) and 17 kg under model (d). How can this be explained given that both estimators are *a priori* unbiased. Actually they are not unbiased. The downward pattern seen under OLS (Ordinary Least Squares) can be explained as follows: usually heavier bulls are going to be slaughtered earlier resulting in an apparent decrease of growth rate with time. These missing data do not arise completely at random.

Table IV. Estimation of variance-covariance parameters.

Variances	Estimates
Intercept variance $\sigma^2_{u_1}$	827.65
Slope variance $\sigma^2_{u_2}$	0.012
Correlation between random intercept and slope ρ	−0.34
Residual variances*	
$\ln \sigma^2_{e_i} = \boldsymbol{p}'_i\boldsymbol{\delta}$	
Intercept	5.08
Calving effects	
(1–3)	0.29
(2–3)	−0.25
Season of birth effects	
(1–2)	−0.24
Age	0.002

Model (d): $\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \sigma_{u_1}\boldsymbol{Z}_{1i}\boldsymbol{u}^*_1 + \sigma_{u_2}\boldsymbol{Z}_{2i}\boldsymbol{u}^*_2 + \boldsymbol{e}_i$ with $\boldsymbol{e}_i \sim N(\boldsymbol{0}, \boldsymbol{I}\sigma^2_{e_i})$ and $\ln \sigma^2_{e_i} = \boldsymbol{p}'_i\boldsymbol{\delta}$.
$\boldsymbol{\beta}$ = { Age, Age^2, Age^3, Twins, Rank of calving, Period of birth, Age*Period of birth, Rank of calving*Period of birth, Age*Twins, Twins*Period of birth }.
$\boldsymbol{\delta}$ = { Rank of calving, Season of birth, Age }.
* Model selected: intercept + parity (3 levels; 1, 2, 3) + season of birth (2 levels; 1, 2) + age at weight (in days).

This missingness process is not taken into account under OLS which leads to an apparent smaller difference between bulls born as single and twins (heavier bulls being in general single born).

3.4. Concluding remarks

This study illustrates a way to analyze repeated measurements with models that use variance-covariance structures for the observations modeled as functions of time. Random coefficient models are convenient tools for modeling such data. They not only reduce the number of parameters, as compared to multiple traits but they can also easily cope with irregular recording patterns in time. They are easily interpretable and manageable under mixed model methodology. For instance, they are of great interest in practice, since they allow for easy calculation of trait performance at typical ages (*e.g.*, here weights at 100, 200, 400 days). They can also be very useful in genetic evaluation for breeding purposes [36].

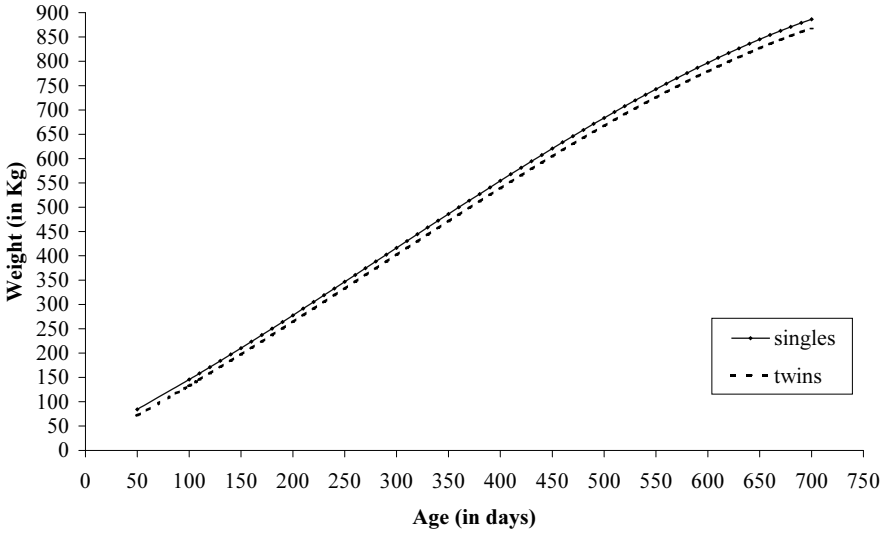

Figure 2. Comparison of mean growth curves between bulls born as single or twins.

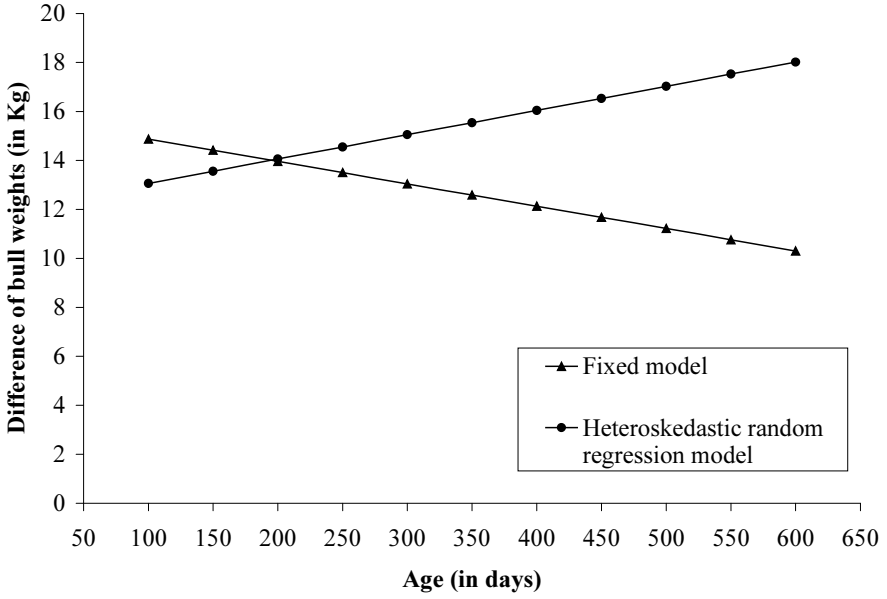

Figure 3. Difference of weights between bulls born as single or twins estimated under two models.

More generally, random coefficient models provide a valuable tool for modeling 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.). 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 [24–26,28,39]. This is neither mandatory in theory nor desirable in practice, since the variation between populations and between subjects within populations does not necessarily follow the same pattern. In practice, the order of polynomials for fitting the random part of the model (adjusted individual profiles) is usually lower than that for the fixed part (population trend), as was the case here.

In addition, semiparametric methods (*e.g.*, splines or kernel methods) can be applied at the fixed effect level, while the between subject variation is fitted *via* random regression [16,42,47,48].

With respect to the random part, dispersion models can be improved significantly (i) by the application of stochastic time processes to take into account the existing correlations between successive measurements, *e.g.* autoregressive processes [6,13,41] and (ii) by allowing for heterogeneity of variances, as was done here (see also [46]).

ACKNOWLEDGEMENTS

The authors wish to thank P. Gillard (Inra, Domaine de la Grêleraie) and P. Maugrion (Inra) for providing the data set and their valuable comments on this application.

REFERENCES

[1] Chi E.M., Reinsel G.C., Models for longitudinal data with random effects and AR(1) errors, J. Am. Stat. Assoc. 84 (1989) 452–459.

[2] Dempster A.P., Laird N.M., Rubin D.B., Maximum likelihood from incomplete data *via* the EM algorithm, J. Royal Stat. Soc. B. 39 (1977) 1–38.

[3] Dempster A.P., Rubin D.B., Tsutakawa R.K., Estimation in covariance components models, J. Am. Stat. Assoc. 76 (1981) 341–353.

[4] DeStefano A.L., Identifying and quantifying sources of heterogeneous residual and sire variances in dairy production data, Ph.D. thesis, Cornell University, Ithaca, New York, 1994.

[5] Diggle P.J., An approach to the analysis of repeated measurements, Biometrics 44 (1988) 959–971.

[6] Diggle P.J., Liang K.Y., Zeger S.L., Analysis of longitudinal data, Oxford Science Publications, Clarendon Press, Oxford, 1994.

[7] Foulley J.L., ECM approach to heteroskedastic mixed models with constant variance ratios, Genet. Sel. Evol. 29 (1997) 297–318.

[8] Foulley J.L., Gianola D., Im S., A simple algorithm for computing marginal maximum likelihood estimates of variance components and its relation to EM, 47th Session of ISI, August 29 to September 6, 1989, Paris, France.
[9] Foulley J.L., Gianola D., San Cristobal M., Im S., A method for assessing extent and sources of heterogeneity of residual variances in mixed linear models, J. Dairy Sci. 73 (1990) 1612–1624.
[10] Foulley J.L., San Cristobal M., Gianola D., Im S., Marginal likelihood and Bayesian approaches to the analysis of heterogeneous residual variances in mixed linear Gaussian models, Comput. Stat. Data Anal. 13 (1992) 291–305.
[11] Foulley J.L., Quaas R.L., Heterogeneous variances in Gaussian linear mixed models, Genet. Sel. Evol. 26 (1995) 117–136.
[12] Foulley J.L., Quaas R.L., Thaon d'Arnoldi C., A Link function approach to heterogeneous variance components, Genet. Sel. Evol. 30 (1998) 27–43.
[13] Foulley J.L., Jaffrézic F., Robert-Granié C., EM-REML estimation of covariance parameters in Gaussian mixed models for longitudinal data analysis, Genet. Sel. Evol. 32 (2000) 129–141.
[14] Garrick D.J., Pollack E.J., Quaas R.L., Van Vleck L.D., Variance heterogeneity in direct and maternal weight by sex and percent purebred for Simmental-sired calves, J. Anim. Sci. 67 (1989) 2513–2528.
[15] Gianola D., Foulley J.L., Fernando R.L., Henderson C.R., Weigel K.A., Estimation of heterogeneous variances using empirical Bayes methods: theoretical considerations, J. Dairy Sci. 75 (1992) 2805–2823.
[16] Green P.J., Silverman B.W., Nonparametric regression and generalized linear models, Chapman and Hall, London, 1994.
[17] Henderson C.R., Applications of Linear Models in Animal Breeding, University of Guelph, Guelph, Ontario, 1984.
[18] Kenward M.G., Roger J.H., Small sample inference for fixed effects from restricted maximum likelihood, Biometrics 53 (1997) 983–997.
[19] Laird N.M., Ware J.H., Random effects models for longitudinal data, Biometrics 38 (1982) 963–974.
[20] Leonard T.A., A bayesian approach to the linear model with unequal variances, Technometrics 17 (1975) 95–102.
[21] Liang K.Y., Zeger S.L., Longitudinal data analysis using generalized linear models, Biometrika 73 (1986) 13–22.
[22] Longford N.T., Random coefficient models, Clarendon Press, Oxford, 1993.
[23] Meng X.L., Rubin D.B., Maximum likelihood estimation *via* the ECM algorithm: a general framework, Biometrika 80 (1993) 267–278.
[24] Meyer K., Hill W.G., Estimation of genetic and phenotypic covariance functions for longitudinal or repeated records by restricted maximum likelihood, Livest. Prod. Sci. 47 (1997) 185–200.
[25] Meyer K., Estimating covariance functions for longitudinal data using a random regression model, Genet. Sel. Evol. 30 (1998) 221–240.
[26] Meyer K., Estimates of genetic and phenotypic covariance functions for postweaning growth and mature weight of beef cows, J. Anim. Breed. Genet. 116 (1999) 181–205.
[27] Nair V.N., Pregibon D., Analyzing dispersion effects from replicated factorial experiments, Technometrics 30 (1988) 247–257.

[28] Olori V.E., Hill W.G., McGuirk B.J., Brotherstone S., Estimating variance components for test day milk records by restricted maximum likelihood with a random animal model, Livest. Prod. Sci. 61 (1999) 53–63.
[29] Patterson H.D., Thompson R., Recovery of interblock information when block sizes are unequal, Biometrika 58 (1971) 545–554.
[30] Robert C., Foulley J.L., Ducrocq V., Genetic variation of traits measured in several environments. I. Estimation and testing of homogeneous genetic and intra-class correlations between environments, Genet. Sel. Evol. 27 (1995a) 111–123.
[31] Robert C., Foulley J.L., Ducrocq V., Genetic variation of traits measured in several environments. II. Inference on between-environment homogeneity of intra-class correlations, Genet. Sel. Evol. 27 (1995b) 125–134.
[32] Robert-Granié C., Ducrocq V., Foulley J.L., Heterogeneity of variance for type traits in the Montbeliarde cattle breed, Genet. Sel. Evol. 29 (1997) 545–570.
[33] Robert-Granié C., Bonaïti B., Boichard D., Barbat A., Accounting for variance heterogeneity in French dairy cattle genetic evaluation, Livest. Prod. Sci. 60 (1999) 343–357.
[34] San Cristobal M., Foulley J.L., Manfredi E., Inference about multiplicative heteroskedastic components of variance in a mixed linear Gaussian model with an application to beef cattle breeding, Genet. Sel. Evol. 25 (1993) 3–30.
[35] SAS® Institute Inc., Cary NC: SAS® institute Inc., SAS/STAT Software: changes and enhancements through release 6.11, 1996.
[36] Schaeffer L.R., Dekkers J.C.M., Random regressions in animal models for test day production in dairy cattle, in: Proceedings of the 5th World Congress on Genetics Applied to Livestock Production vol. 18, August 7–12, 1994, University of Guelph, Guelph, Ontario.
[37] Self S.G., Liang K.Y., Asymptotic properties of maximum likelihood estimation and likelihood ratio test under nonstandard conditions, J. Am. Stat. Assoc. 82 (1987) 605–610.
[38] Stram D.O., Lee J.W., Variance components testing in the longitudinal mixed effect model, Biometrics 50 (1994) 257–267.
[39] Veerkamp R.F., Goddard M.E., Covariance functions across herd production levels for test day records on milk, fat and protein yields, J. Dairy Sci. 81 (1998) 1690–1701.
[40] Verbeke G., Molenberghs G., Linear mixed models in practice, Springer-Verlag, New York, 1997.
[41] Verbeke G., Molenberghs G., Linear mixed models for longitudinal data, Springer-Verlag, New York, 2000.
[42] Verbyla A.P., Cullis B.R., Kenward M.G., Welham S.J., The analysis of designed experiments and longitudinal data by using smoothing splines, Appl. Stat. 48 (1999) 269-311.
[43] Visscher P.M., Thompson R., Hill W.G., Estimation of genetic and environmental variances for fat yield in individual herds and an investigation into heterogeneity of variance between herds, Livest. Prod. Sci. 28 (1991) 273–290.
[44] Visscher P.M., Hill W.G., Heterogeneity of variance and dairy cattle breeding, Anim. Prod. 55 (1992) 321–329.

[45] Weigel K.A., Gianola D., Yandel B.S., Keown J.F., Identifications of factors causing heterogeneous within-herd variance components using structural model for variances, J. Dairy Sci. 76 (1993) 1466–1478.
[46] Wolfinger R.D., Heterogeneous variance covariance structures for repeated measures, J. Agric. Biol. Env. Stat. 1 (1996) 205–230.
[47] Zeger S.L., Diggle P.J., Semiparamatric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters, Biometrics 50 (1994) 689–699.
[48] Zhang D., Lin X., Raz J., Sowers M., Semiparametric stochastic mixed models for longitudinal data, J. Am. Stat. Assoc. 93 (1998) 710–719.

APPENDIX

Estimation of dispersion parameters

The function to be maximized is:

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

where 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 $\boldsymbol{x}|\boldsymbol{y}$, $\boldsymbol{\delta} = \boldsymbol{\delta}^{[t]}$, $\boldsymbol{\eta}_1 = \boldsymbol{\eta}_1^{[t]}$, $\boldsymbol{\eta}_2 = \boldsymbol{\eta}_2^{[t]}$, $\rho = \rho^{[t]}$,

$$\boldsymbol{e}_i = \boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \sigma_{u_{1i}}\boldsymbol{Z}_{1i}\boldsymbol{u}_1^* - \sigma_{u_{2i}}\boldsymbol{Z}_{2i}\boldsymbol{u}_2^*$$

and

$$\boldsymbol{G} = var(\boldsymbol{u}^*) = var\begin{pmatrix}\boldsymbol{u}_1^* \\ \boldsymbol{u}_2^*\end{pmatrix} = \begin{pmatrix}\boldsymbol{I}_q & \rho\boldsymbol{I}_q \\ \rho\boldsymbol{I}_q & \boldsymbol{I}_q\end{pmatrix}.$$

More generally, for K random regression components, $\boldsymbol{G} = \boldsymbol{G}_0 \otimes \boldsymbol{I}_q$ where $\boldsymbol{G}_0$ is a correlation matrix with (k, l) element: $g_{0,kl} = \rho_{kl}$ with $k = 1, \ldots, K$ and $l = 1, \ldots, K$.

In the general case, the random coefficient model can be written as follows:

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \sum_{k=}^{K} \sigma_{u_{ki}}\boldsymbol{Z}_{ki}\boldsymbol{u}_k^* + \boldsymbol{e}_i.$$

The REML estimates can be obtained efficiently *via* the Newton-Raphson algorithm for $\boldsymbol{\delta}, \boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \ldots, \boldsymbol{\eta}_K$ estimates and *via* the Fisher scoring algorithm for the parameter $\boldsymbol{\rho} = \{\rho_{12}, \rho_{13}, \ldots, \rho_{1K}, \rho_{23}, \ldots, \rho_{2K}, \ldots, \rho_{K-1,K}\}$, vector of correlations ρ_{kl}.

Numerically, the current estimates $\boldsymbol{\delta}^{[t+1]}, \boldsymbol{\eta}_1^{[t+1]}, \boldsymbol{\eta}_2^{[t+1]}, \ldots, \boldsymbol{\eta}_K^{[t+1]}$ of $\boldsymbol{\delta}$, $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \ldots, \boldsymbol{\eta}_K$ are computed with the following iterative system:

$$\left(\frac{\partial^2 Q_e}{\partial \gamma^2}\right)^{[t]} \left(\gamma^{[t+1]} - \gamma^{[t]}\right) = \left(-\frac{\partial Q_e}{\partial \gamma}\right)^{[t]}$$

$$\Longleftrightarrow$$

$$\begin{pmatrix} \boldsymbol{P}'\boldsymbol{W}_{\delta\delta}\boldsymbol{P} & \boldsymbol{P}'\boldsymbol{W}_{\delta\eta_1}\boldsymbol{H}_1 & \cdots & \boldsymbol{P}'\boldsymbol{W}_{\delta\eta_K}\boldsymbol{H}_K \\ \boldsymbol{H}_1'\boldsymbol{W}_{\eta_1\delta}\boldsymbol{P} & \boldsymbol{H}_1'\boldsymbol{W}_{\eta_1\eta_1}\boldsymbol{H}_1 & \cdots & \boldsymbol{H}_1'\boldsymbol{W}_{\eta_1\eta_K}\boldsymbol{H}_K \\ \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{H}_K'\boldsymbol{W}_{\eta_K\delta}\boldsymbol{P} & \boldsymbol{H}_K'\boldsymbol{W}_{\eta_K\eta_1}\boldsymbol{H}_1 & \cdots & \boldsymbol{H}_K'\boldsymbol{W}_{\eta_K\eta_K}\boldsymbol{H}_K \end{pmatrix}^{[t]} \begin{pmatrix} \Delta\boldsymbol{\delta} \\ \Delta\boldsymbol{\eta}_1 \\ \vdots \\ \Delta\boldsymbol{\eta}_K \end{pmatrix}^{[t+1]} = - \begin{pmatrix} \boldsymbol{P}'\boldsymbol{v}_\delta \\ \boldsymbol{H}_1'\boldsymbol{v}_{\eta_1} \\ \vdots \\ \boldsymbol{H}_K'\boldsymbol{v}_{\eta_K} \end{pmatrix}^{[t]}.$$

In the general case, $-2Q_u = \ln|\boldsymbol{G}| + E_c[\boldsymbol{u}^{*\prime}\boldsymbol{G}^{-1}\boldsymbol{u}^*] = q(\ln|\boldsymbol{G}_0| + tr[\boldsymbol{G}_0^{-1}\boldsymbol{D}^*])$ with $\boldsymbol{D}^* = \left\{d_{kl}^* = \frac{1}{q}E_c(\boldsymbol{u}_k^{*\prime}\boldsymbol{u}_l^*)\right\}$.

And the current estimate of $\boldsymbol{\rho}$ is computed from the following equation:

$$E\left(\frac{\partial^2 Q_u}{\partial\boldsymbol{\rho}\partial\boldsymbol{\rho}'}\right)^{[t]}(\boldsymbol{\rho}^{[t+1]} - \boldsymbol{\rho}^{[t]}) = \left(-\frac{\partial Q_u}{\partial\boldsymbol{\rho}}\right)^{[t]}$$

where

$$\begin{aligned}\frac{\partial(-2Q_u)}{\partial\rho_{kl}} &= qtr\left[\boldsymbol{G}_0^{-1}\frac{\partial\boldsymbol{G}_0}{\partial\rho_{kl}}\right] - tr\left[\boldsymbol{G}_0^{-1}\frac{\partial\boldsymbol{G}_0}{\partial\rho_{kl}}\boldsymbol{G}_0^{-1}\boldsymbol{D}\right] \\ &= qtr\left[(\boldsymbol{G}_0^{-1} - \boldsymbol{G}_0^{-1}\boldsymbol{D}^*\boldsymbol{G}_0^{-1})\frac{\partial\boldsymbol{G}_0}{\partial\rho_{kl}}\right]\end{aligned}$$

and

$$E\left[\frac{\partial^2(-2Q_u)}{\partial\rho_{kl}\partial\rho'_{k'l'}}\right] = qtr\left[\frac{\partial\boldsymbol{G}_0}{\partial\rho_{kl}}\boldsymbol{G}_0^{-1}\frac{\partial\boldsymbol{G}_0}{\partial\rho_{k'l'}}\boldsymbol{G}_0^{-1}\right].$$

Calculations have been made easier by taking advantage of the simple expression of the Fisher information matrix since $E[\boldsymbol{D}^*] = \boldsymbol{G}_0$. This system reduces to a third degree polynomial equation, *i.e.*

$$\rho^3 - d_{12}^*\rho^2 + (d_{11}^* + d_{22}^* - 1)\rho - d_{12}^* = 0.$$

This equation can be solved either analytically or numerically.

If individuals are not independent, one has to replace $\boldsymbol{G}$ by $\boldsymbol{G}_0 \bigotimes \boldsymbol{A}$, where $\boldsymbol{A}$ is a symmetric, positive definite matrix of known coefficients.

After deleting $[t]$ for reasons of simplicity, we have:

$$
\begin{aligned}
&\boldsymbol{P}' = (\boldsymbol{p}_1, \ldots, \boldsymbol{p}_i, \ldots, \boldsymbol{p}_p);\\
&\boldsymbol{H}_1' = (\boldsymbol{h}_{11}, \ldots, \boldsymbol{h}_{1i}, \ldots, \boldsymbol{h}_{1p}), \boldsymbol{H}_2' = (\boldsymbol{h}_{21}, \ldots, \boldsymbol{h}_{2i}, \ldots, \boldsymbol{h}_{2p}), \ldots,\\
&\boldsymbol{H}_K' = (\boldsymbol{h}_{K1}, \ldots, \boldsymbol{h}_{Ki}, \ldots, \boldsymbol{h}_{Kp}).\\
&\boldsymbol{W}_{\delta\delta} = Diag\{w_{\delta\delta,ii}\}\\
&\boldsymbol{W}_{\delta\eta_1} = Diag\{w_{\delta\eta_1,ii}\}\\
&\vdots\\
&\boldsymbol{W}_{\delta\eta_K} = Diag\{w_{\delta\eta_K,ii}\}\\
&\boldsymbol{W}_{\eta_1\eta_1} = Diag\{w_{\eta_1\eta_1,ii}\}\\
&\vdots\\
&\boldsymbol{W}_{\eta_1\eta_K} = Diag\{w_{\eta_1\eta_K,ii}\}\\
&\vdots\\
&\boldsymbol{W}_{\eta_K\eta_K} = Diag\{w_{\eta_K\eta_K,ii}\}
\end{aligned}
$$

with

$$
\begin{aligned}
&w_{\delta\delta,ii} = \sigma_{e_i}^{-2} E_c[\boldsymbol{e}_i'\boldsymbol{e}_i]\\
&w_{\delta\eta_1,ii} = \sigma_{u_{1i}}\sigma_{e_i}^{-2} E_c(\boldsymbol{u}_1^{*\prime}\boldsymbol{Z}_{1i}'\boldsymbol{e}_i)\\
&\vdots\\
&w_{\delta\eta_K,ii} = \sigma_{u_{Ki}}\sigma_{e_i}^{-2} E_c(\boldsymbol{u}_K^{*\prime}\boldsymbol{Z}_{Ki}'\boldsymbol{e}_i)\\
&w_{\eta_k\eta_k,ii} = 0.5\sigma_{u_{ki}}\sigma_{e_i}^{-2}[-E_c(\boldsymbol{u}_k^{*\prime}\boldsymbol{Z}_{ki}'\boldsymbol{e}_i) + \sigma_{u_{ki}} E_c(\boldsymbol{u}_k^{*\prime}\boldsymbol{Z}_{ki}'\boldsymbol{Z}_{ki}\boldsymbol{u}_k^*)], \quad \forall k = 1, \ldots, K\\
&w_{\eta_k\eta_l,ii} = 0.5\sigma_{u_{ki}}\sigma_{u_{li}}\sigma_{e_i}^{-2} E_c(\boldsymbol{u}_k^{*\prime}\boldsymbol{Z}_{ki}'\boldsymbol{Z}_{li}\boldsymbol{u}_l^*), \quad \forall k \neq l
\end{aligned}
$$

and,

$$
\begin{aligned}
&\boldsymbol{\nu}_\delta = \{\nu_{\delta,i}\}\\
&\boldsymbol{\nu}_{\eta_1} = \{\nu_{\eta_1,i}\}\\
&\vdots\\
&\boldsymbol{\nu}_{\eta_K} = \{\nu_{\eta_K,i}\}
\end{aligned}
$$

with

$$
\begin{aligned}
\nu_{\delta,i} &= n_i - \{\sigma_{e_i}^{-2} E_c[\boldsymbol{e}_i\boldsymbol{e}_i]\}\\
\nu_{\eta_k,i} &= \sigma_{u_{ki}}\sigma_{e_i}^{-2} E_c(\boldsymbol{u}_k^{*\prime}\boldsymbol{Z}_{ki}'\boldsymbol{e}_i), \quad \forall k = 1, \ldots, K
\end{aligned}
$$

based on $\boldsymbol{e}_i = \boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \sum_{k=1}^{K} \sigma_{u_{ki}}\boldsymbol{Z}_{ki}\boldsymbol{u}_k^*$; $\sigma_{e_i}^2 = \exp(\boldsymbol{p}_i'\boldsymbol{\delta})$; $\sigma_{u_{ki}}^2 = \exp(\boldsymbol{h}_{ki}'\boldsymbol{\eta}_k)$, $\forall k = 1, \ldots, K$.

Finally, the expectation step of the EM algorithm consists of determining all conditional expectations at each iteration. In the EM-REML algorithm and

after deleting $[t]$ for reasons of simplicity, $E_c^{[t]}(.)$ can be expressed as follows:

$$
\begin{aligned}
E_c[(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})'(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})] &= (\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}})'(\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}) + tr[\boldsymbol{X}_i'\boldsymbol{X}_i\boldsymbol{C}^{\beta\beta}] \\
E_c[(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})'\boldsymbol{Z}_{ki}\boldsymbol{u}_k^*] &= (\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}})'\boldsymbol{Z}_{ki}\hat{\boldsymbol{u}_k}^* - tr[\boldsymbol{X}_i'\boldsymbol{Z}_{ki}\boldsymbol{C}^{\beta u_k}], \\
&\qquad \forall k = 1, \ldots, K \\
E_c[\boldsymbol{u}_k'\boldsymbol{Z}_{ki}'\boldsymbol{Z}_{li}\boldsymbol{u}_l^*] &= \hat{\boldsymbol{u}_k}'\boldsymbol{Z}_{ki}'\boldsymbol{Z}_{li}\hat{\boldsymbol{u}_l}^* + tr[\boldsymbol{Z}_{ki}'\boldsymbol{Z}_{li}\boldsymbol{C}^{u_l u_k}], \\
&\qquad \forall k = 1, \ldots, K \quad \text{and} \quad \forall l = 1, \ldots, K
\end{aligned}
$$

where $\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{u}_1}^*, \hat{\boldsymbol{u}_2}^*, \ldots, \hat{\boldsymbol{u}_K}^*$ are solutions of the mixed model equations [17] $\boldsymbol{C}\boldsymbol{s} = \boldsymbol{r}$ where the coefficient matrix $\boldsymbol{C}$ is equal to:

$$
\begin{pmatrix}
[l]\sum_{i=1}^{p}\boldsymbol{X}_i'\boldsymbol{X}_i\sigma_{e_i}^{-2} & \sum_{i=1}^{p}\boldsymbol{X}_i'\boldsymbol{Z}_{1i}\sigma_{u_{1i}}\sigma_{e_i}^{-2} & \cdots & \sum_{i=1}^{p}\boldsymbol{X}_i'\boldsymbol{Z}_{Ki}\sigma_{u_{Ki}}\sigma_{e_i}^{-2} \\
\sum_{i=1}^{p}\boldsymbol{Z}_{1i}'\boldsymbol{X}_i\sigma_{u_{1i}}\sigma_{e_i}^{-2} & \sum_{i=1}^{p}\boldsymbol{Z}_{1i}'\boldsymbol{Z}_{1i}\sigma_{u_{1i}}^2\sigma_{e_i}^{-2} + g_0^{11} & \cdots & \sum_{i=1}^{p}\boldsymbol{Z}_{1i}'\boldsymbol{Z}_{Ki}\sigma_{u_{1i}}\sigma_{u_{Ki}}\sigma_{e_i}^{-2} + g_0^{1K} \\
\vdots & \vdots & \vdots & \vdots \\
\sum_{i=1}^{p}\boldsymbol{Z}_{Ki}'\boldsymbol{X}_i\sigma_{u_{Ki}}\sigma_{e_i}^{-2} & \sum_{i=1}^{p}\boldsymbol{Z}_{Ki}'\boldsymbol{Z}_{1i}\sigma_{u_{1i}}\sigma_{u_{Ki}}\sigma_{e_i}^{-2} + g_0^{K1} & \cdots & \sum_{i=1}^{p}\boldsymbol{Z}_{Ki}'\boldsymbol{Z}_{Ki}\sigma_{u_{Ki}}^2\sigma_{e_i}^{-2} + g_0^{KK}
\end{pmatrix}
$$

$$
\boldsymbol{r} = \begin{pmatrix}
[l]\sum_{i=1}^{p}\boldsymbol{X}_i'\boldsymbol{y}_i\sigma_{e_i}^{-2} \\
\sum_{i=1}^{p}\boldsymbol{Z}_{1i}'\boldsymbol{y}_i\sigma_{u_{1i}}\sigma_{e_i}^{-2} \\
\vdots \\
\sum_{i=1}^{p}\boldsymbol{Z}_{Ki}'\boldsymbol{y}_i\sigma_{u_{Ki}}\sigma_{e_i}^{-2}
\end{pmatrix} \quad \text{and} \quad \boldsymbol{s} = \begin{pmatrix}
\hat{\boldsymbol{\beta}} \\
\hat{\boldsymbol{u}_1^*} \\
\vdots \\
\hat{\boldsymbol{u}_K^*}
\end{pmatrix}
$$

where g_0^{kl} is element (k, l) of $\boldsymbol{G}_0^{-1}$.