Exact Validation of Genetic Evaluation Software

H. Leclerc and V. Ducrocq
INRA, Station de Génétique Quantitative et Appliquée, 78 352 Jouy-en-Josas Cedex, France

Introduction

Since 1995, Interbull provides international predicted breeding values on the scale of each participating country using national genetic evaluation results. Input data quality constitutes a crucial issue in international genetic evaluations, as results from complex genetic and statistical analyses depend on it. Therefore, Interbull confers a great importance to the monitoring and the validation of input data (Fikse, 2004): data included into international evaluations are assumed to fulfil a number of stringent criteria. The consistency of evaluations is assessed by the comparison of breeding values from consecutive evaluations to identify changes larger than excepted based on statistical properties of breeding values (Klei *et al.,* 2002). Genetic trends are estimated to check that national breeding values are unbiased (Boichard *et al.,* 2002). Diversity and complexity of methods used in the various countries to analyze different traits has led to a situation in which adoption of new methods of validation of national data and/or models is necessary. One research project identified by Interbull in 2002 is the development of a general simulation tool to validate national genetic evaluation systems, and especially the development of a simulation environment to test breeding value prediction software. With this aim, a program was developed from a strategy described by R. Thompson (1997) to simulate data with known breeding values and phenotypes for a single trait animal model (Täubert *et al*., 2002), in such a way that BLUP solutions for breeding values should be equal to the simulated ones. This strategy was extended to a multiple trait animal model (Wensch-Dorendorf *et al*., 2005). The latter algorithm requires that all animals, including males, have performance and that animals from the base generation belong to a "dummy" class of fixed effect.

National genetic evaluation models for dairy traits are nowadays more and more often based on test-day models (TDM) instead of 305-day lactation models. Various models have been proposed. They differ in the way the lactation curve is modelled as a function of days in milk (DIM) - with fixed classes, parametric or semi-parametric (splines) curves - in the way the genetic and permanent environment components are described (random regression using Legendre or other polynomials), and in the way heterogeneous residual variances are accounted for. Countries have developed their own TDM for routine genetic evaluations for their own population. Unfortunately, no general evaluation software is available for TDM for very large datasets. In several countries, a specific software had to be developed. But the software validation stage is made difficult when no reference software exists. Extension from the Thompson's strategy to random regression situations was not straightforward.

The aim of this paper is to present a general and flexible strategy to check the correctness of genetic evaluation software. This strategy differs from R. Thompson one but keeps the same basic idea : from simulated effects and residuals, performances are created in such a way that BLUP estimates from a correct evaluation software are equal to the simulated effects.

Material and Methods

Outline of the procedure

The starting point is a pedigree file and a data file containing for each record, the relevant levels and/or covariables of all effects, the animal's recoded number and permanent environment effect level and all other pertinent pieces of information (elements required to compute random regression coefficients, the record's weight, the genetic, permanent environment and residual (co)variance matrices, etc.). These files can be real data sets. Then, the procedure to check genetic evaluation software can be divided into three steps:

1) for each effect as well as for one residual per observation, simulated values are computed following the approach described below, leading to a simulated performance for each record in the data file.

2) these simulated performances are used as input data in the national genetic evaluation software. Estimates are obtained for all effects included in the model.

3) estimates of fixed effects and predicted random effects from the national genetic evaluation software are compared with the true (simulated) ones. If resulting breeding values, permanent environment effects and all estimable contrasts of fixed effects are identical to the true ones then the genetic evaluation software can be considered as correct.

Model

Consider, as an example, the following linear model for a single trait (the extension to multiple traits is straightforward and is not considered here):

$$\mathbf{y} = \mathbf{Xb} + \mathbf{ZQg} + \mathbf{Za}^* + \mathbf{Wp} + \mathbf{e}$$

where $\mathbf{y}$ is the vector of observations, $\mathbf{b}$ is a vector of fixed effects, $\mathbf{g}$ is the vector of genetic group effects (e.g., for phantom parents groups), $\mathbf{a}^*$ is the vector of breeding values corrected for genetic group effects, i.e. with $\mathrm{E}[\mathbf{a}^*] = 0$ or alternatively $\mathrm{E}[\mathbf{a}] = \mathbf{Qg}$ with $\mathbf{a} = \mathbf{Qg} + \mathbf{a}^*$ and $\mathrm{Var}[\mathbf{a}^*] = \mathbf{G} = \mathbf{G}_0 \otimes \mathbf{A}$, $\mathbf{p}$ is the vector of permanent environment effects with $\mathrm{Var}[\mathbf{p}] = \mathbf{P} = \mathbf{P}_0 \otimes \mathbf{I}$, $\mathbf{e}$ is the vector of random residual with $\mathrm{var}(\mathbf{e}) = \mathbf{R}$, $\mathbf{X}$, $\mathbf{Z}$ and $\mathbf{W}$ are matrices relating $\mathbf{y}$ to the appropriate fixed, genetic and permanent environment effects, possibly through continuous covariates and $\mathbf{Q}$ is the matrix assigning animals in $\mathbf{a}^*$ to groups in $\mathbf{g}$. In the case of random regressions, $\mathbf{G}_0$ and $\mathbf{P}_0$ are the covariance matrices for the genetic and permanent environment effects, respectively and $\mathbf{A}$ is the additive genetic relationship matrix. $\mathbf{R}$ is a diagonal matrix of residual variances. In case of heterogeneous variances, for example as a function of DIM (e.g., Druet *et al*., 2003), the diagonal terms vary from one record to the next.

The mixed model equations are:

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{ZQ} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{W} \\ \mathbf{Q}'\mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Q}'\mathbf{Z}'\mathbf{R}^{-1}\mathbf{ZQ} & \mathbf{Q}'\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} & \mathbf{Q}'\mathbf{Z}'\mathbf{R}^{-1}\mathbf{W} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{ZQ} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{W} \\ \mathbf{W}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{W}'\mathbf{R}^{-1}\mathbf{ZQ} & \mathbf{W}'\mathbf{R}^{-1}\mathbf{Z} & \mathbf{W}'\mathbf{R}^{-1}\mathbf{W}+\mathbf{P}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{b} \\ \mathbf{g} \\ \mathbf{a}^* \\ \mathbf{p} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Q}'\mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{W}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix} \quad [1]$$

We want $\mathbf{y}$ - and therefore $\mathbf{b}$, $\mathbf{g}$, $\mathbf{a}^*$, $\mathbf{p}$ and $\mathbf{e}$ needed to create $\mathbf{y}$ - to be simulated such that the solutions from the mixed model equations are exactly the simulated $\mathbf{b}$, $\mathbf{g}$, $\mathbf{a}^*$ and $\mathbf{p}$. Replacing $\mathbf{y}$ in the right hand side [1] by $\mathbf{Xb} + \mathbf{ZQg} + \mathbf{Za}^* + \mathbf{Wp} + \mathbf{e}$, e.g.:

$$\mathbf{X}'\mathbf{R}^{-1}\mathbf{y} = \mathbf{X}'\mathbf{R}^{-1}\mathbf{Xb} + \mathbf{X}'\mathbf{R}^{-1}\mathbf{ZQg} + \mathbf{X}'\mathbf{R}^{-1}\mathbf{Za}^* + \mathbf{X}'\mathbf{R}^{-1}\mathbf{Wp} + \mathbf{X}'\mathbf{R}^{-1}\mathbf{e}$$

and developing the left hand side, the initial requirement leads after some simplifications to four conditions:

$\mathbf{X}'\mathbf{R}^{-1}\mathbf{e} = \mathbf{0}$, $\mathbf{Q}'\mathbf{Z}'\mathbf{R}^{-1}\mathbf{e} = \mathbf{0}$, $\mathbf{G}^{-1}\mathbf{a}^* = \mathbf{Z}'\mathbf{R}^{-1}\mathbf{e}$ and $\mathbf{P}^{-1}\mathbf{p} = \mathbf{W}'\mathbf{R}^{-1}\mathbf{e}$.

To fulfil the first two conditions, a variable ε is first simulated for all observations with any underlying distribution, for example, $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$ with some arbitrary variance σ_ε^2. Then, two vectors $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are computed as solutions of:

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{ZQ} \\ \mathbf{Q}'\mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Q}'\mathbf{Z}'\mathbf{R}^{-1}\mathbf{ZQ} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\gamma} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\boldsymbol{\varepsilon} \\ \mathbf{Q}'\mathbf{Z}'\mathbf{R}^{-1}\boldsymbol{\varepsilon} \end{bmatrix}.$$

If we choose $\mathbf{e}=\boldsymbol{\varepsilon}-\mathbf{X}\boldsymbol{\beta}-\mathbf{ZQ}\boldsymbol{\gamma}$, it can be checked that $\mathbf{X}'\mathbf{R}^{-1}\mathbf{e}=\mathbf{0}$ and $\mathbf{Q}'\mathbf{Z}'\mathbf{R}^{-1}\mathbf{e}=\mathbf{0}$. Note that **b** and **g** can be simulated using any underlying distribution: they do not influence **e**. The next step is to derive $\mathbf{a}^*$ such that $\mathbf{G}^{-1}\mathbf{a}^*=\left(\mathbf{G}_0^{-1}\otimes\mathbf{A}^{-1}\right)\mathbf{a}^*=\left(\mathbf{I}\otimes\mathbf{A}^{-1}\right)\left(\mathbf{G}_0^{-1}\otimes\mathbf{I}\right)\mathbf{a}^*=\mathbf{Z}'\mathbf{R}^{-1}\mathbf{e}$. First, let $\mathbf{u}=\left(\mathbf{G}_0^{-1}\otimes\mathbf{I}\right)\mathbf{a}^*$. For example, if the random regression model includes three genetic terms (3 genetic values per animal, $\mathbf{a}^*_{1j}$, $\mathbf{a}^*_{2j}$ and $\mathbf{a}^*_{3j}$), $\mathbf{u}=\{\mathbf{u}_i\}$ also includes three terms $\mathbf{u}_{1j}$, $\mathbf{u}_{2j}$, $\mathbf{u}_{3j}$ for each animal *j* which are linear combinations of the terms $\mathbf{a}^*_{1j}$, $\mathbf{a}^*_{2j}$, $\mathbf{a}^*_{3j}$. For each *i*, we need to solve $\mathbf{A}^{-1}\mathbf{u}_i=\mathbf{Z}'\mathbf{R}^{-1}\mathbf{e}_i$, This is easily done using the decomposition $\mathbf{A}^{-1}=\mathbf{T}^{-1}\mathbf{D}^{-1}\mathbf{T}^{-\mathrm{T}}$: first solve $\left(\mathbf{T}^{-1}\mathbf{D}^{-1}\right)\mathbf{v}_i=\mathbf{Z}'\mathbf{R}^{-1}\mathbf{e}_i$ for $\mathbf{v}_i$ and then solve $\mathbf{T}^{-\mathrm{T}}\mathbf{u}_i=\mathbf{v}_i$. These are 2 simple triangular systems in which at most two non-diagonal elements are nonzero. Finally, compute $\mathbf{a}^*=\mathbf{G}_0\otimes\mathbf{u}$.

We obtain **p** enforcing $\mathbf{P}^{-1}\mathbf{p}=\left(\mathbf{P}_0^{-1}\otimes\mathbf{I}\right)\mathbf{p}=\mathbf{W}'\mathbf{R}^{-1}\mathbf{e}$ by choosing $\mathbf{p}=\left(\mathbf{P}_0\otimes\mathbf{I}\right)\mathbf{W}'\mathbf{R}^{-1}\mathbf{e}$. Finally, **y** is constructed as $\mathbf{y}=\mathbf{Xb}+\mathbf{ZQg}+\mathbf{Za}^*+\mathbf{Wp}+\mathbf{e}$.

Application to a test-day model

The data available were 47,492 first lactations of Montbéliarde cows from a sample of herds in Jura, i.e. 377,080 test-day observations recorded between September 1995 and August 2005 in 28,020 herd test-day (HTD) combinations. The average number of test-day records per lactation was 8, with a minimum of 5. On average 115 test-days records were known per herd and per year, with a minimum of 50 and a maximum of 431. 118 classes of calving month and 96 classes of calving age were considered. Days in milk had to be between 5 and 335. Information about gestation (days carried calf, DCC) was also available. The pedigree file included 102,339 animals with 8 groups of unknown parents. The approach described above was tested on the following simplified model, where for animal *i,* test-day *j* and DCC *k,* the performance recorded after *t* days of lactation is:

$$y_{ijkrs(t)}=\mathrm{HTD}_j+\beta\cdot\mathrm{DCC}_k+\sum_{l=1}^{Nb}\delta_{l(t)}\cdot b_{rl}+\sum_{p=0}^{Nc}\xi_{p(t)}\cdot c_{sp}+\sum_{m=0}^{Na}\alpha_{m(t)}\cdot a_{im}+\sum_{n=0}^{Np}\psi_{n(t)}\cdot p_{in}+e_{ijkrs(t)}$$

where $y_{ijkrs(t)}$ is the performance recorded, HTD_j is the j^{th} fixed herd by test-date effect applied to all animals in the herd on test-date *j*, β is a regression coefficient related to the effect of the calf after k days of gestation on performance, the b_{rl}'s and the c_{sp}'s are fixed regression coefficients specific to calving month *r* and calving age class *s*, the a_{im}'s are the genetic values for cow *i*, the p_{in}'s are the permanent environment effects for cow *i*. $\delta_{l(t)}$, $\xi_{p(t)}$, $\alpha_{m(t)}$ and $\psi_{n(t)}$ are continuous covariates depending on DIM. For this model, random regression curves and fixed regression curves for calving age were modeled using Legendre polynomials of order two (*Nc* = *Na* = *Np* = 2). Fixed regression curves for calving month were modeled using 6 knots splines at DIM 5, 20, 50, 135, 245 and 335. $e_{ijkrs(t)}$ is the residual effect for each observation, with an heterogeneous variance continuously changing with DIM.

Results and Discussion

Correlations between simulated effects and estimates obtained with the national genetic evaluation software from the 377,080 TD records were 1.00000. Standard deviations for all effects included in the model were identical. For contrast analysis, the largest relative difference observed between "estimates" (estimable contrasts or predicted random effects) and true values (i.e. 478,468 estimated effects) was 4E-04 when convergence criteria used for BLUP evaluation (i.e. average solution change between two iterations) was 0.1E-05.

The software implemented for national evaluation is considered as correct for the tested model.

In contrast with other approaches (Thompson, 1997, Täubert *et al*., 2002, Wensch-Dorendorf *et al*., 2005), the proposed strategy to generate breeding values and performances can be applied to actual data sets by simply replacing performance values by simulated ones. Its main specificity is that appropriate residuals are created when the other approaches do not simulate any residual.

Moreover, the method can help to investigate relevance of different iterative algorithms and convergence criteria. Misztal *et al*. (1988) showed that the real accuracy of solutions could be far from the one suggested by some convergence criteria. Here, true solutions of the system are directly available. Relative average difference between current and true solutions as proposed by Misztal *et al*. (1988) can be computed at each iteration for a particular data set.

For Interbull's needs, two main advantages for such an approach can be highlighted: it enables any participating country to check the correctness of their national genetic evaluation software and to verify that the number of iteration rounds or the convergence criteria are adequate.

Acknowledgement

The research was funded by a French Ministry of Agriculture through an "Actions Innovantes" grant called UTILEG .

References

Boichard, D., Bonaiti, B., Barbat, A. & Mattalia, S. 1995. Three Methods to Validate the Estimation of Genetic Trend for Dairy Cattle. *J. Dairy. Sci. 78,* 431-437.

Druet, T., Jaffrézic, F., Boichard, D. & Ducrocq, V. 2003. Modeling Lactation Curves and Estimations of Genetic Parameters for First Lactation Test-Day Records of French Holstein Cows. *J. Dairy. Sci*. *86*, 2480-2490.

Fikse, W.F. 2004. Interbull guides through the labyrinth of national genetic evaluations. *Proc. 55th EAAP*. Comm HG5.1.

Klei, B., Mark, T., Fikse, W.F. & Lawlor, T. 2002. A method for verifying genetic evaluation results. *Interbull Bulletin 29,* 178-182.

Misztal, I., Gianola, D. & Schaeffer, L.R. 1988.Convergence rates in animal model solutions. *J. Dairy. Sci. 70,* 2577-2584.

Täubert, H., Swalve, H.H. & Simianer, H. 2002. The Interbull Audit Project Part II: Development of a Program for Auditing Breeding Value Estimation Programs. *Interbull Bulletin 29,* 165-167.

Thompson, R. 1997. Generating data to check Mixed Model Equations *Personal communication*.

Wensch-Dorendorf, M., Swalve, H.H. & Wensch, J. 2005. Simulation of multiple trait data for testing breeding value estimation programs. *Proc. 56th EAAP*. *11*:207.

White, I.M.S., Thompson, R. & Brotherstone, S. 1999. Genetic and environmental smoothing of lactation curves with cubic splines. *J. Dairy. Sci. 82,* 632-638.