Impact of Using Reduced Rank Random Regression Test-Day Model on Genetic Evaluation

H. Leclerc[1], I. Nagy[2] and V. Ducrocq[2]
[1] Institut de l'Elevage, Département Génétique, Bât 211, 78 352 Jouy-en-Josas, France
[2] UMR1313 INRA, Génétique Animale et Biologie Intégrative, 78 352 Jouy-en-Josas, France
helene.leclerc@inst-elevage.asso.fr

Abstract

The development of genetic evaluations on dairy traits based on individual test-day records represents a major computing challenge due to the number of parameters in the model and the number of records to analyse. To reduce computer requirements, we proposed to use reduced rank test-day models where the smallest eigenvalues of the covariance matrix of random effects (genetic, permanent environment and herd-year) are set to zero. Different levels of reduction were tested. The model with 4 genetic, 4 permanent environment and 2 herd-year effects including heterogeneous herd-year residual variance led to only minor changes in estimated breeding values compared with the full rank model.

Introduction

Dairy traits have always held a major place in the choice of males and females to generate the next generation. Until the 90s, genetic evaluations on dairy traits were based on cumulative performance on 305-day. During the last decade, most dairy countries implemented genetic evaluations based on test-day models (TDM), which present numerous advantages, particularly in terms of modelling as mentioned in several reviews (in particular Swalve, 2000 and Jensen, 2001). However, TDM depend on large number of parameters and usually require to analyse 10 times more data than with a lactation model. Memory and CPU time requirements constitute a major stake in the development of TDM.

To get over these limits, some authors like Wiggans and Goddard (1997) or Van der Werf *et al*. (1998) proposed to reduce the dimensionality of random regression test-day models by using a principal component approach where the smallest eigenvalues and the corresponding covariates are eliminated. Two advantages arise for TDM: reduced rank models reduce the number of levels to be estimated per random effect but also reduce computational requirements substantially This approach is used (or has been studied) by countries such as The Netherlands (De Roos *et al*., 2002) or Finland (Lidauer *et al*., 2003).

In a preliminary study, Leclerc and Ducrocq (2009) showed that the rank reduction approach should be based on the canonical decomposition of the covariance matrix rather than of the correlation matrix. In fact, the reduction based on the covariance matrix leads to results more consistent with the full rank matrix. It leads to higher correlations between random estimates for both genetic, permanent environment and herd-year effects.

The aim of this study was to evaluate the possibility of rank reduction of a random regression test-day model based on the stability of random effects estimates and reduction in computational requirements.

Material & Methods

Data

Data were test-day (TD) yields of milk, fat and protein, fat and protein percentages for the first three lactations of French Montbéliarde cows. Analyses were based on data collected between September 1988 and December 2007. Days in milk (DIM) ranged from 7 to 335 days. The dataset included more than 24 million of TD from 1.36 million cows. The pedigree file contained about 1.75 million animals. To be included in the analysis, cows were required to have a first lactation record and known sire and

dam. Only cows with at least three TD were considered. At least five records were required to define a herd by test-date effect (HTD).

Model

Test-day records used in the single-trait random regression test-day model were pre-adjusted in a first-step for time-independent fixed effects related to the shape of the lactation curve which varies as a function of calving month, calving age, length of dry period, gestation and parity (Leclerc *et al*., 2008). In the second step of the genetic evaluation, test-day records are described as:

$$y_{iklmnopqtt'} = HTD_i + \text{year-dep fixed eff}_{klmno} + \sum_{a=1}^{A}(\text{genetic}_{pa} \cdot \upsilon_{at}) + \sum_{b=1}^{B}(\text{perm\_env}_{pb} \cdot \zeta_{bt}) + \sum_{c=1}^{C}((\text{herd}_q \times \text{year}_k)_c \cdot \tau_{ct}) + e_{iklmnopqtt'}$$

where $y_{iklmnopqtt'}$ is the record of animal p in herd q on DIM t and days carried calf (DCC) t'. HTD_i is the i^{th} herd-test-date effect. The year-dep fixed eff_{klmno} is the sum of year-dependent fixed effects which are constant over the lactation obtained for year k, parity l, length of dry period m, calving month n and calving age o ;

$genetic_{pa}$ is the a^{th} genetic effect of animal p, υ_{at} is the value of the a^{th} eigenvector of genetic variance-covariance matrix **G** of rank A at DIM t. In the full rank model used as reference, the genetic variance-covariance matrix had a rank of 6 which was later reduced to 4 in reduced rank models ;

$perm\_env_{pb}$ is the b^{th} permanent environment effect of cow p, ξ_{bt} is the value of the b^{th} eigenvector of permanent environment variance-covariance matrix **P** of rank B at DIM t. As for the genetic effects, the variance-covariance matrix had a rank of 6 for the reference model which was later reduced to 4 in the reduced rank models ;

$(herd_q \times year_k)_c$ is the c^{th} herd×year effect of herd q, for year k, τ_{ct} is the value of the c^{th} eigenvector of herd×year variance-covariance matrix **H** of rank C at DIM t. In the full rank model, the herd-year variance-covariance matrix had a rank of 9 which was later reduced to 6, 4, 2 or 0 in reduced rank models, respectively. These drastic reductions in the dimensionality of the herd-year of calving effect could be envisaged due to its low contribution of the total variance (between 2 and 5% on average over the 3 lactations).

$e_{iklmnopqtt'}$ is the residual value whose variance matrix **R** is expressed as a function of a 12-knot regression splines of DIM separately for the 3 lactations (Druet *et al*., 2005). The phenotypic variance was ensured to remain the same before and after rank reduction by adding to **R**, the loss of variance in **G**, **P** and **H** due to rank reduction. The effects of region x parity and region x calving year were taken into account as sources of residual variance heterogeneity. The effect of herd-year of calving (HY) heterogeneity on residual variance was accounted for, as described by Robert-Granié *et al.* (1999) in one version of the tested models.

As described in table 1, 10 different models with various levels of rank reduction (where the effect of herd-year of calving heterogeneity on residual variance was either included or omitted) were compared for milk yield. The dark grey line corresponding to the full rank model including heterogeneity variance on HY was used as the reference model. The comparison were made on each of the first three lactations and on an average index of the three lactations with weight of 0.5 for the first, 0.3 for the second and 0.2 for the third one, hereafter called "global" index.

Druet *et al*. (2005) showed that the 6 eigenvectors of the genetic (co)variance matrix had a biological interpretation. They represent the average production level and a persistency measurement for each lactation, respectively. This last trait can be interesting for selection. However, the use of reduced rank model leads to a change in the matrix structure and then, information about production level and persistency per parity is no longer available. Using the approach developed by Tarrés *et al.* (2008), it was possible to back-transform the reduced rank random effect estimates to the interpretable ones.

Table 1. Characteristics of 10 models for milk yield with variable ranks for genetic effect (G), permanent environment effect (P) and herd-year effect (H) with or without heterogeneous herd-year residual variance.

G rank	P rank	H rank	Heterog. var on HY	Model name
6	6	9	No	$G_6P_6H_9$
4	4	6	No	$G_4P_4H_6$
4	4	4	No	$G_4P_4H_4$
4	4	2	No	$G_4P_4H_2$
4	4	0	No	$G_4P_4H_0$
6	**6**	**9**	**Yes**	**$G_6P_6H_9h$**
4	4	6	Yes	$G_4P_4H_6h$
4	4	4	Yes	$G_4P_4H_4h$
4	4	2	Yes	$G_4P_4H_2h$
4	4	0	Yes	$G_4P_4H_0h$

Results and Discussion

For the five studied traits, 4 eigenvectors explained between 98.5 and 99.9% of the total genetic variance in the three lactations, and between 91.3 and 94.4% for the permanent environment variance. For the herd-year variance, 6 eigenvectors out of 9 explained from 96.6 to 99.7%, 4 eigenvectors explained from 86.9 to 94.5% and the 2 largest eigenvectors explained from 56.3 to 79.5% of the initial variance.

The correlations between the "global" indices (combining the first 3 lactations) obtained with the reference model $G_6P_6H_9h$ and the 9 reduced models are reported in table 2. The general trend shows a decrease in correlation with the reduction of the matrix rank. However, the magnitude of the decrease was low. Interestingly, the inclusion of heterogeneous hear-year variance contributed to maintain high correlations. Correlations obtained with the model $G_4P_4H_2h$ (0.998 for genetic effect, 0.997 for permanent environment effect and 0.965 for herd-year effect) seemed to be a good compromise between number of effects to estimate and precision of the index, so it was chosen to study the impact of rank reduction on the index of the 4 other traits (fat and protein yields and percentages) (Table 3).

Table 2. Correlations between genetic effect (G), permanent environment effect (P) and herd-year effect (H) estimated with full rank model $G_6P_6H_9h$ and various rank reduced models including or not heterogeneous hear-year (HY) variance for the "global" index on milk yield.

Model	Heterog. var on HY	G	P	H
$G_6P_6H_9$	Without	0.995	0.990	0.991
$G_4P_4H_6$		0.994	0.989	0.982
$G_4P_4H_4$		0.993	0.989	0.974
$G_4P_4H_2$		0.993	0.987	0.961
$G_4P_4H_0$		0.989	0.979	
$G_4P_4H_6$**h**	With	0.999	0.999	0.988
$G_4P_4H_4$**h**		0.999	0.998	0.980
$G_4P_4H_2$**h**		0.998	0.997	0.965
$G_4P_4H_0$**h**		0.994	0.988	

Table 3. Correlations between genetic effect (G), permanent environment effect (P) and herd-year effect (H) estimated with full rank model $G_6P_6H_9h$ and with a reduced rank model $G_4P_4H_2h$ for the average index on the first three lactations on milk, fat and protein yield, and fat and protein %.

Trait	**G**	**P**	**H**
Milk	0.998	0.997	0.965
Fat	0.996	0.995	0.806
Protein	0.997	0.995	0.895
Fat%	0.999	0.997	0.904
Protein %	0.995	0.986	0.563

Correlations on the "global" index for the 4 other traits were similar to those of milk yield for genetic and permanent environment effects. The correlations within parity (not shown) were similar to those obtained with the "global' index (over 0.99). Only the correlation for the 3rd lactation permanent environment effects obtained for fat and protein % were lower: 0.945 and 0.968 respectively. However, the situation was different for herd-year effect. The correlations obtained for these 4 traits on "global" index were lower than for milk (between 0.563 and 0.904). These correlations varied a lot with parity. For instance, the correlation between full rank and reduced rank breeding values for 1st and 2nd lactation were

0.449 and 0.912 respectively for fat yield. The worst situations were for protein % with 0.088 and 0.731 for 1st and 2nd lactation respectively. This situation was due to the lower part of the variance explained by the 2 largest eigenvectors. The selected eigenvectors did not include information about differences between parities.

The correlations were analysed for different sub-populations (Table 4). The well known bulls (with more than 200 daughters) had correlations between full rank index and reduced rank index slightly higher than young bulls (with 20 daughter or less) or cows, but in all cases, correlations were at least 0.993.

Table 4. Correlations on the "global" index between genetic effect estimated with full rank model $G_6P_6H_9h$ and with a reduced rank model $G_4P_4H_2h$ on milk, fat and protein yield, and fat and protein % for different sub-populations (bulls with 20 daughters or less, bulls with more than 200 daughters and cows born since 1998).

Trait	**Bulls** (<20 daughters)	**Bulls** (>200 daughters)	**Cows**
Milk	0.998	0.998	0.997
Fat	0.997	0.999	0.995
Protein	0.993	0.996	0.993
Fat %	0.998	1.000	0.999
Protein %	0.994	0.997	0.996

When looking at the two underlying genetic traits, i.e. production level and persistency, the reduced rank indices need to be back-transformed to make them comparable to the indices estimated with the full rank model. The correlations obtained between these 2 models for production level on the 5 studied traits were higher or equal to 0.996 (Table 5). For persistency, correlations were on average lower (between 0.976 and 0.999) but largely high enough for our needs.

Table 5. Correlations between production level (and persistency) of genetic effect (G) estimated with full rank model $G_6P_6H_9h$ and with a reduced rank model $G_4P_4H_2h$ for the average index on first three lactations on milk, fat and protein yield, and fat and protein %.

Trait	Production level	Persistency
Milk	0.997	0.985
Fat	0.996	0.976
Protein	0.997	0.977
Fat%	0.999	0.999
Protein %	0.996	0.995

Conclusion

The high correlations obtained between the full rank model $G_6P_6H_9h$ and the reduced rank model $G_4P_4H_2h$ led to negligible changes in estimated breeding values. So, it makes it possible to use the reduced rank model for genetic evaluation. Only the moderate correlations of herd-year effects can be worrying. However, the contribution of herd-year to the total variance is minor, this effect being included in the model only to account for differences in the lactation curves across herd. The index on herd-year is not intended to be released and used.

Two advantages can be put forward about the use of reduced rank model $G_4P_4H_2h$: it requires less memory due to the reduction by 1/3 of the number of equations - from 21.4 to 14.2 million - and also significantly less computing time (-54%).

Acknowledgements

This study was performed within the framework of the Unité Mixte Technologique INRA – Institut de l'Elevage on cattle genetic evaluations.

References

De Roos, A.P.W., Koenen, E.P.C., Harbers, A.G.F. & de Jong, G. 2002. Model validation and rank reduction of covariance matrices in the random regression Test-Day model in the Netherlands. *Interbull Bulletin 29,* 91-94.

Druet, T., Jaffrézic, F. & Ducrocq, V. 2005. Estimation of genetic parameters for test day records of dairy traits in the first three lactations. *Genet. Sel. Evol. 37,* 257-271.

Jensen, J. 2001. Genetic evaluation of dairy cattle using test-day models. *J. Dairy. Sci. 84,* 2803-2812.

Leclerc, H., Duclos, D., Barbat, A., Druet, T. & Ducrocq, V. 2008. Environmental effects on lactation curves included in a test-day model genetic evaluation. *Animal 2,* 344-353.

Leclerc, H. & Ducrocq, V. 2009. Correlation *vs* Covariance matrices for rank reduction in test-day models. *Proc. 60th EAAP*. 15:38.

Lidauer, M., Mäntysaari, E.A. & Strandén, I. 2003. Comparison of test-day models for genetic evaluation of production traits in dairy cattle. *Livest. Prod. Sci. 79,* 73-86.

Robert-Granié, C., Bonaïti, B., Boichard, D. & Barbat, A. 1999. Accounting for variance heterogeneity in French dairy cattle genetic evaluation. *Livest. Prod. Sci. 60,* 343-357.

Swalve, H.H. 2000. Theoretical basis and computational methods for different test-day genetic evaluation methods. *J. Dairy. Sci. 83,* 1115-1124.

Tarres, J., Liu, Z., Ducrocq, V., Reinhardt, F. & Reents, R. 2008. Data transformation for rank reduction in multi-trait MACE model for international bull comparison. *Genet. Sel. Evol. 40,* 295-308.

Van der Werf, J.H.J., Goddard, M.E. & Meyer, K. 1998. The use of covariance functions and random regressions for genetic evaluation of milk production based on test day records. *J. Dairy Sci. 81,* 3300-3308.

Wiggans, G.R. & Goddard, M.E. 1997. A computationally feasible test day model for genetic evaluation of yields traits in the United States. *J. Dairy Sci. 80,* 1785-1800.