# MCMC based estimation of variance components in a very large dairy cattle data set

*Luc Janss[1] and Gerben de Jong[2]*
*[1]Institute for Animal Science and Health (ID-DLO) P.O.Box 65, 8200AB Lelystad, The Netherlands*
*[2]Dutch Cattle Syndicate (NRS),P.O. Box 454, 6800AL Arnhem, The Netherlands*

**Abstract**

Bayesian MCMC methods, like Gibbs sampling, are often considered specialised tools for small scale data analysis. However, due to certain computational differences with e.g. REML, these methods also apply well to large scale data analyses which are practically infeasible by REML. The main attraction for very large data sets is that the basic computational algorithm for Gibbs sampling looks like an Gauss-Seidel iterative BLUP scheme, which can indeed be applied well to very large data sets.

To illustrate this case, heritabilities for milk production and milk components were estimated in Dutch Friesian (DF) dairy cattle, (being) upgraded to Holstein Friesian (HF). The data contained 1 122 088 lactations of 585 785 cows and a total of 684 512 animals. Multiple lactations were analysed using a repeatability model. Fixed effects included were herd-year-season-parity (120 509 classes), month of calving (12 classes), permanent environmental effects for the 585 785 cows and regressions on fraction heterosis and fraction recombination loss to allow for the effects of crossing with HF.

As can be expected from using such large amounts of data, estimated genetic parameters have an astonishing high accuracy. Heritabilities for milk, fat and protein were 0.401 (min 0.394, max 0.407), 0.362 (0.350-0.370) and 0.356 (0.352-0.365). Repeatabilities found were between 0.60 and 0.64 and the regression on heterosis was 293 kg milk. Using only first lactations, heritability for milk was significantly higher (0.464) and a breed difference between HF and DF of 570 kg was found.

## 1. Introduction

Genetic evaluations, e.g. in dairy cattle, are nowadays obtained by large scale computation of BLUPs (Best Linear Unbiased Predictions) using all information available within a country. However, "true" BLUP is never performed, as BLUP requires to know the variance components (e.g. expressed as heritabilities and repeatabilities), whereas in practice only *estimates* of variance components will be available. Gianola et al. (1986) have shown that in this practical situation, the expected merit of selected animals is maximised when BLUP is performed based on REML estimates of variance components based on the same data. This ideal situation however is usually not found in dairy cattle evaluations: variance components used are not based on all data as available for BLUP, and genetic parameters usually are not continuously updated when more or different data becomes available. In fact, current computationally methodology does not allow to perform REML estimation of variance components using all data. Currently, in the Netherlands, variance components used for BLUP are mainly based on a study of Van der Werf (1990) based on first lactations. Genetic evaluations, however, are based on a repeatability model including up to three lactations.

The aim of this study was to show the feasibility of variance component estimation in much larger data sets than currently feasible with REML approaches. This alternative is based on Gibbs sampling, which can be implemented on an iterative BLUP like algorithm, by which scaling up to much larger analyses is indeed possible. Gibbs sampling can be applied in several inferential

procedures. Here, it will be used for a Bayesian inference that is taylored to closely mimic REML, by using a joint posterior distribution of variance components that is mathematically equal to the REML likelihood. Computational aspects will be explained, and especially the close relationship between iterative BLUP and Gibbs sampling.

## 2. Methods

*A general mixed model and iterative BLUP*

Consider a general mixed model:

$$y = X\beta + Wa + Zu + e \qquad (1)$$

where y is a vector with observations, X, W and Z are incidence matrices, $\beta$ is a vector with levels for a fixed effect, a is a "simple" random effect (with a diagonal variance structure), and u are random animal genetic effects. First, inclusion of only one fixed effect, one simple random effect and one animal genetic effect is considered. Variance components for a, u and e are denoted $\sigma_a^2$, $\sigma_u^2$ and $\sigma_e^2$. The mixed model equations can be expressed as the following set of equations:

$$X'X\beta + X'Wa + X'Zu = X'y \quad (2a)$$

$$W'X\beta + (W'W+kI)a + W'Zu = W'y \quad (2b)$$

$$Z'X\beta + Z'Wa + (Z'Z+\lambda A^{-1})u = Z'y \quad (2c)$$

where the simultaneous solutions for $\beta$, a and u are BLUEs and BLUPs. In (2b) $k=\sigma_e^2/\sigma_a^2$ and in (2c) $\lambda=\sigma_e^2/\sigma_u^2$ and $A^{-1}$ is the inverse of the numerator relationship matrix. The mixed model equations can be solved iteratively by a Gauss-Seidel scheme, which can be written using (1) as:

$$X'X\beta^{k+1} = X'(e + X\beta^k) \qquad (3a)$$

$$(W'W+kI)a^{k+1} = W'(e + Wa^k) \qquad (3b)$$

$$(Z'Z+\lambda A^{-1})u^{k+1} = Z'(e + Z'u^k) \qquad (3c)$$

Equations (3abc) show a similar structure for solving each block of equations, involving only the design matrix for the respective effect, the vector of errors and the old solutions. For (one) fixed and (one) simple random effect, the left-hand side involves a diagonal matrix, which allows simple solving. The equations for animal genetic effects (3c), however, also contain off-diagonal elements in the left-hand side matrix; solving these equations by Gauss-Seidel requires to store list of progeny by parent (e.g. by sorting progeny codes on parents) in order to be able to construct all required off-diagonal terms. Solving mixed model equations in this manner allows to build flexible software that allows to fit any number of fixed and (simple) random effects by simply repeating the blocks. The block of animal genetic effects, however, can not simply be repeated (e.g. to allow for multiple trait analyses), because covariances between blocks then arise and result in a variable number of different terms to be added to the equations (3c).

*Gibbs sampling*

A Gibbs sampling scheme that allows estimation of variance components is implemented based on the equations for iterative BLUP (3abc). When, for each single parameter, the equation to be solved is expressed as $d_i\beta_i = r_i$ then Gibbs samples are generated by sampling:

$$\beta_i^* \sim N(r_i/d_i , \sigma_e^2/d_i )$$

and similarly for all $a_i$ and $u_i$ (see e.g. Wang et al., 1993). The Gibbs sampling scheme is completed by sampling variance components from inverted chi-square distributions based on obtained sampled vectors for a, u and e. Here, flat priors are used which is implemented by using "-2" degrees of freedom for the prior (see Wang et al. 1994). The Gibbs samples generated, once the Gibbs chain has convcerged, will show the unconditional uncertainty about each parameter, i.e. information which otherwise could only be obtained by (partly) inverting the mixed model equations. The simple quadratic $u'A^{-1}u$ can thus be used for estimation of genetic variance.

*Comparison to REML*

The above scheme that results in a Bayesian inference on variance components will give estimates very similar to REML. In fact, the mode of joint posterior distribution $f(\sigma_a^2, \sigma_u^2, \sigma_e^2 \mid y)$ corresponds exactly to REML, as, with flat priors, this joint distribution takes the same mathematical form as the REML likelihood. In the application presented here, not this joint mode was computed from the Gibbs samples, but the means of the marginal posterior distributions of variance components. In large analyses, as presented here, there will be virtually no difference between these marginal means and the joint mode. Anyhow, these Bayesian inference will have the same desirable property as REML of accounting for the estimation of fixed effects; in Bayesian terms this is accomplished by marginalising with respect to fixed effects.

## 3. Material

Data of up to three lactions available from a period of 18 years (1978-1995) were collected, provided that per farm at least 42 lactations per year and 500 lactations in total were present. Data consisted of 305 day yields for milk, fat and protein. Only data from cows that were Dutch Friesian (DF) or Holstein Friesian (HF), or crosses between these two, was included. In the period studied, (most of) the DF cows were being crossed with HF, resulting in various crosses. For each animal, the HF blood-share and the expected fractions heterosis and recombination loss could be computed. In total, 1 122 088 lactations were available from 585 758 cows. About 42% of the lactations were first lactations, 33% were second lactations and 25% were third lactations.

*Models:* The main fixed effects in the model were Herd-Year-Season-Parity (HYSP) classes, calving months and regressions on heterosis fraction, fraction recombination loss and (optionally) HF blood-share. As random effects, apart from random error, were included animal genetic effects, and permanant environmental effects to account for repeated lactations. The effects and their levels for the main model are:

| Effect | Levels |
|---|---|
| Animal Genetic (random with relationships) | 684 512 |
| Herd-Year-Season-Parity (fixed) | 120 509 |
| Permanent Environment (random) | 585 758 |
| Calving Month (fixed) | 216 |
| Heterosis% (regression) | 1 |
| Recombination Loss% (regression) | 1 |
| Total | 1 390 997 |

Some variations on the main analysis were used: a seperate analysis was performed using only first lactations. In this case, permanent environmental effects could be dropped from the model. Finally, also an analysis was performed that included a regression on HF blood-share to estimate the difference between the DF and HF breed. All models were applied to the analysis of milk, fat and protein yield.

The Gibbs sampler was run for 15 000 cycles for the main model and for 12 000 cycles for variations on the main model. From the Gibbs samples, posterior mean, minimum and maximum were reported. The minimum and maximum supply a rough, probably quite stringent, confidence interval.

## 4. Results

Results for the main analysis, combining information from three lactations, are presented in Table 1. Heritability for milk yield was estimated as 0.40; heritabilities for fat and protein yield were 0.36. The rough confidence intervals, obtained as the minumum and maximum value in the Gibbs samples, was <0.02, which shows that these heritabilities were estimated with large precision. Repeatabilities were all above 0.60 with rough confidence intervals <0.01. Heterosis was estimated as 293 kg for milk yield, 14.8 kg for fat yield and 12.2 kg for protein yield. The values for recombination loss are all about half of the values for heterosis and estimates are considerably less precise than the estimates for heterosis.

The same parameters, but based on first lactations only are in Table 2. For first lactations, heritabilities are a little higher (0.03 to 0.06), heterosis is somewhat lower and recombination loss is higher than the average of all three. Due to the large precision of these estimates, these differences are statistically significant, except the difference in heterosis for milk yield.

In a third analysis, a regression on HF blood-share was also included in the model. Estimated variance components and regressions on heterosis and recombination loss were very similar to the values given in Table 1 (not presented). The estimated regression on fraction HF, which represents an estimate of the difference between DF and HF was 570 kg for milk yield, 22.4 kg for fat yield and 17.6 kg for protein yield.

*Table 1.* Results for the main analysis: heritability, repeatability and estimated regression on fraction heterosis and recombination loss. Presented figures are the posterior means and (in parenthesis below) the minimum and maximum value found in the Gibbs samples, supplying a confidence interval.

| | Milk | Fat | Protein |
|---|---|---|---|
| Heritability | 0.401<br>(0.394 - 0.407) | 0.362<br>(0.350 - 0.370) | 0.360<br>(0.352 – 0.365) |
| Repeatability | 0.635<br>(0.632 - 0.639) | 0.604<br>(0.601 - 0.607) | 0.629<br>(0.626 – 0.631) |
| Heterosis [kg] | 293<br>(282 – 305) | 14.8<br>(14.2 – 15.3) | 12.2<br>(11.8 – 12.6) |
| Recomb loss [kg] | -140<br>(-176 - -105) | -5.37<br>(-6.55 - -3.33) | -5.15<br>(-6.06 - -3.41) |

*Table 2.* Results for analysis of first lactations only: heritability and estimated regression on fraction heterosis and recombination loss. Presented figures are the posterior means and (in parenthesis below) the minimum and maximum value found in the Gibbs samples, supplying a confidence interval.

| | Milk | Fat | Protein |
|---|---|---|---|
| Heritability | 0.464<br>(0.451 - 0.480) | 0.417<br>(0.406 - 0.432) | 0.399<br>(0.385 – 0.411) |
| Heterosis [kg] | 272<br>(256 – 288) | 12.8<br>(12.0 – 13.5) | 10.3<br>(9.77 – 10.9) |
| Recomb loss [kg] | -204<br>(-253 - -155) | -7.18<br>(-8.84 - -5.54) | -6.41<br>(-7.69 - -3.31) |

## 5. Discussion

Variance components were estimated with a model that is also used for the Dutch genetic evaluations, i.e. a repeatability model with at maximum 3 lactations. All data from the larger farms in a period of 18 years was used, amounting to more than 1 million lactation records. Effects of heterosis and recombination loss which are applied as pre-corrections in the genetic evaluations were included here as a part of the model. Results showed higher values for heritabilities, repeatabilities and regressions on heterosis and recombination loss than currently in use in the genetic evaluations, based on Van der Werf (1990). The values currently in used for milk yield are a heritability of 0.35 (here 0.40), repeatability of 0.55 (here 0.64), heterosis effect 120 kg (here 293) and recombination loss −60 kg (here −140). Heritability of milk yield based on first lactations was found 0.46. With an average heritability for milk yield over all lactations of 0.40, heritability in later lactations should be below 0.40, while heritability in first lactations is clearly above 0.40.

Gibbs sampling proved a useful tool to obtain variance components in very large data sets. Although computations are still intensive, i.e. days of computing are required, these analyses are feasible, whereas traditional REML estimation would have been impossible. Typically, memory and computing requirements increases quadratically for REML, whereas for Gibbs sampling memory requirements and computing time (per Gibbs cycle) increase linearly. Gibbs sampling, when implemented on an iterative BLUP scheme as done here, requires to store the data, levels of fixed and random effects, and several working vectors; for the analysis performed here this amounted to about 100 Mb.

The feasibility to perform variance component estimation on large data sets, may allow to put into practice the ideal described by Gianola et al. (1986) to base BLUP on REML estimates from the same data. The close computational resemblence between (iterative) BLUP and MCMC based estimation of variance components can be very useful in this respect. For instance, parallel to computing BLUP, one could compute an MCMC chain, sharing the

building of the mixed model equations required for both applications.

## References


Gianola D, Foulley JL, Fernando RL (1986) Prediction of breeding values when variances are not known. Genet Sel Evol 18: 485-498

Van der Werf (1990) Models to estimate genetic parameters in crossbred cattle populations under selection. Doctoral thesis, Wageningen Agricultural University, Wageningen, The Netherlands.

Wang CS, Rutledge JJ, Gianola D (1993) Marginal inferences about variance components in a mixed linear model using Gibbs sampling. Genet Sel Evol 25:41-62

Wang CS, Rutledge JJ, Gianola D (1994) Bayesian analysis of mixed linear models via Gibbs sampling with an application to litter size in Iberian pigs. Genet Sel Evol 26:91-115