Including Non-additive Effects In Bayesian Methods For The Prediction Of Genetic Values From Genome-wide SNP Data

D. Wittenburg[*], N. Melzer[*] and N. Reinsch[*]

Introduction

It is a challenging task to predict the genetic values for traits in dairy cattle on the basis of genome-wide SNP markers. Methods including additive genetic effects have already been studied, but the importance of non-additive effects for the genetic variation is not fully understood. From the biological perspective it is especially of interest to know which loci interact. Furthermore, a better prediction of genetic values is intended, when additive effects, dominance and pairwise epistatic effects are jointly involved in fitting a model to a trait. In order to avoid the estimation of additional covariance components, the genotypic effects have to be appropriately re-parameterised in advance. Three methods of orthogonalisation, which are proposed in the literature, are applied to simulated datasets. Different scenarios are simulated and analysed to verify whether the non-additive effects are precisely estimated.

Material and methods

Theory. For statistical analysis in a Bayesian framework a hierarchical model is set up similarly to Meuwissen et al. (2009). First, only main genetic effects are included. The vector of phenotypes $\mathbf{y} = (y_1, \ldots, y_n)'$ is modelled as

$$\mathbf{y} \quad = \quad \mathbf{1}\mu + \mathbf{X}\mathbf{g}_a + \mathbf{D}\mathbf{g}_d + \mathbf{e}\,. \tag{1}$$

The $\mathbf{X}$ and $\mathbf{D}$ are the design matrices for the additive $\mathbf{g}_a = (g_{a,1}, \ldots, g_{a,m})'$ and dominance $\mathbf{g}_d = (g_{d,1}, \ldots, g_{d,m})'$ genetic effects, respectively. In total m loci are studied on the genome. It is $X_{i,j} = \pm 1$ and $D_{i,j} = 0$ for a homozygous genotype at locus j of animal i (the positive effect is assigned to the more frequent allele). For a heterozygous genotype it is $X_{i,j} = 0$ and $D_{i,j} = 1$. Let μ be a population mean. We apply the following prior distributions

$$\begin{aligned} e_i &\sim N(0, \sigma_e^2)\,, \quad i = 1, \ldots, n\,, \\ g_{s,j} &\sim L^*(\gamma, \lambda)\,, \quad s \in \{a, d\},\ j = 1, \ldots, m\,. \end{aligned}$$

$L^*(\gamma, \lambda)$ denotes a mixture of a Laplace distribution with zero expectation and the point mass at zero. It is $\Pr(g_{s,j} = 0) = 1 - \gamma$ and $\mathrm{Var}(g_{s,j}) = \gamma \frac{2}{\lambda^2}$. The parameters γ and λ are treated as hyper-parameters at the current stage of study.

[*]Leibniz Institute for Farm Animal Biology, Wilhelm-Stahl-Allee 2, 18196 Dummerstorf, Germany

It is assumed that genotypic effects at different loci are independently distributed. In order to obtain uncorrelated effects at one locus, i.e. $\mathrm{Cov}(X_{i,j}g_{a,j}, D_{i,j}g_{d,j}) = 0\ \forall j$, the corresponding coefficients of the marker genotypes have to be re-parameterised in advance. Three methods of orthogonalisation are taken from the literature. The first and second one, Cockerham (1954, A) and Zeng et al. (2005, B), are similar to each other and both require a Hardy-Weinberg equilibrium (HWE) of allele frequencies. The third approach (Álvarez-Castro and Carlborg, 2007, C) relaxes this assumption and applies genotype probabilities.

In a second step, the pairwise epistatic effects are modelled. The genetic effect caused by an interaction between locus j and k is denoted as $g_{s,j,k}$ with $s \in \{aa, ad, da, dd\}$. The effect is called to be additive $\times$ additive (aa), when it occurs only if the individual i is homozygous at the loci j and k. It is called additive $\times$ dominance (ad), when is appears at a homozygous locus j and a heterozygous locus k ($j < k$) and so on. Model (1) can be extended to epistatic effects as suggested by Kao and Zeng (2002)

$$\mathbf{y} \quad = \quad \mathbf{1}\mu + \mathbf{X}\mathbf{g}_a + \mathbf{D}\mathbf{g}_d + \sum_{s\in\{aa,ad,da,dd\}} \mathbf{W}_s\mathbf{g}_s + \mathbf{e}\,. \tag{2}$$

As an example, it is $W_{aa,j,k} = X_j \cdot X_k$, where the symbol $\cdot$ denotes the element-wise multiplication of column j and k of $\mathbf{X}$. Furthermore, $W_{ad,j,k} = X_j \cdot D_k$ is calculated to obtain the coefficients for the effect $g_{ad,j,k}$. This way, in total four times $m(m-1)/2$ epistatic effects are modelled. The priors remain the same as used for (1), but the probability of having a non-zero epistatic effects should be smaller than the γ for main effects. The genetic value of an individual is obtained as the sum over all genotypic effects.

When working with model (2), it is necessary to additionally standardise the coefficients of the main genetic effects to reach numerical stability. Assuming $\mathbf{X}$ and $\mathbf{D}$ are already orthogonal and p_j is the (estimated) allele frequency at locus j, then one possible normalisation is

$$X_j \;\longmapsto\; \frac{X_j}{\sqrt{2p_j(1-p_j)}} \quad \text{and} \quad D_j \;\longmapsto\; \frac{D_j}{2p_j(1-p_j)}\,.$$

Note that this normalisation step also assumes HWE. As a consequence, the genetic variance components are estimated as $\widehat{\sigma}^2_s = \sum_{j=1}^{m} \widehat{g}^2_{s,j}$ for $s \in \{a, d\}$ and $\widehat{\sigma}^2_s = \sum_{j=1}^{m-1}\sum_{k=j+1}^{m} \widehat{g}^2_{s,j,k}$ for $s \in \{aa, ad, da, dd\}$, where the latter is an approximation under linkage disequilibrium.

Solver. So far the Bayesian methods used for marker assisted prediction applied Gibbs sampling steps which require a lot of computing time. These sampling methods collapse for 50K (or more) SNP markers if further non-additive effects are included. For some time an approximate Bayesian approach is available which applies the analytically derived posterior density for a marker effect rather than samples thereof (Meuwissen et al., 2009). This approach is known to be slightly less accurate, because only a single SNP effect is studied at a time while the vector of phenotypes is corrected for all other previously estimated effects. In an iterative procedure, one wriggles through all genetic effects and starts again until the estimates of genetic effects converge. This iterative procedure is much faster than the conventional Bayesian methods (e.g. BayesB, Meuwissen et al., 2001).
We adapted this fast Bayesian method (fBayesB) to non-additive effects as described above. It

is used to estimate the genetic effects on the basis of simulated datasets. Different scenarios are simulated to study the loss of accuracy of prediction if epistatic effects are not simulated but modelled and vice versa. The accuracy is determined as the correlation between estimated and simulated genetic values in a test set. When studying only the main genetic effects via model (1), the results of fBayesB are compared with the "gold standard" BayesB. An implementation of BayesB of Berry and Stranden is available on www.genomicselection.net (obtained Sep 4, 2009). This version was also extended to non-additive effects. In principal, it would be possible to estimate epistatic effects using BayesB, but this tool would probably require a few month to finish 50 000 Gibbs sampling rounds for a simulated dataset.

Simulation study. We applied a mutation-drift model and simulated a population with effective population size of 100 animals and 52 273 SNP markers on a 30 Morgan genome (in style of the Illumina Chip BovineSNP50). Starting with homozygous loci, a mutation rate of $2.5 \cdot 10^{-3}$ was chosen for each SNP locus and 400 generations of random mating involving recombination events on the genome were carried out. About 8 % of the loci were fixed due to drift and the linkage disequilibrium was obtained as $r^2 = 0.13$. The training generations 401 and 402 consisted each of 50 half-sib families with 20 offspring. These individuals were genotyped and phenotyped ($n = 2\,000$). The test generations 403 and 404 were built up the same way but without phenotyping the animals. To begin with, we used every 10th marker in this study. Loci deviating from HWE (in average one locus per dataset) were not excluded from statistical analyses. In total 23 SNP loci were randomly chosen to be the QTL. Additive effects were drawn from a gamma distribution (similar to Meuwissen et al., 2001). Dominance coefficients were drawn from a normal distribution (Bennewitz, EAAP 2009). Epistatic effects were drawn from normal distributions with arbitrary parameters chosen to such an extend that epistasis explains about 20 % of the total genetic variance. For each source of epistatic variation different parameters were used. The distance among QTL had no impact on epistatic effects, because all interactions are assumed to be equally important in this Bayesian approach. Altogether 24 gene interactions were simulated. To obtain residual error terms, which should be comparable between simulations with and without epistasis, the residual variance component was determined depending on the chosen broad-sense heritability of $H^2 = 0.50$. This results in a narrow-sense heritability of $h^2 = 0.48$ ($h^2 = 0.40$) without (with) simulated epistasis.

Results and discussion

The following preliminary results are based on the analysis of ten simulated datasets with epistatic effects as well as ten sets without epistasis. In total 100 repetitions of the simulation with varying broad-sense heritability are under way.

First of all, the re-parameterisation methods A–C were evaluated. The choice of the hyperparameter λ had impact on successfully running the iterations of fBayesB and it strongly affected the accuracy of genetic value prediction. Since the particular genetic variance components have different impact on the total genetic variance, the λ should be different for each class of genetic parameters. The method C was quite robust in the specific choice of λ and yielded highest correlation between estimated and simulated genetic values (results not shown). The correlations of C were only slightly better than B. This was expected, because if the assumption of HWE holds, method B and C are identical. All approaches were sensitive in

terms of γ, i.e. the probability of a non-zero genetic effect. Because only 23 QTL were simulated, $\gamma \in \{0.005, 0.01\}$ worked sufficient for main effects. Running model (2) was solely feasible with method C.

Second, we compared the differences in accuracy of prediction between fBayesB and BayesB based on model (1). Table 1 shows the average correlation between estimated and simulated genetic values and the contrast in computing time. Though fBayesB only required a small fraction of time compared with BayesB, a lack of accuracy was not observed.

Table 1: Average correlation ρ between estimated and simulated genetic values

		Simulation		Computing time
		without epistasis	with epistasis	for single dataset
Model (1)	fBayesB	$\rho = 0.966$	$\rho = 0.912$	1 second
	BayesB	$\rho = 0.961$	$\rho = 0.903$	4 hours
Model (2)	fBayesB	$\rho = 0.953$	$\rho = 0.902$	7 hours

Third, we studied the impact of including or not including pairwise epistatic effects on the accuracy of predicting the genetic values in the test generations. With default values for the hyper-parameters the estimation of epistatic effects was insufficient, but the loss of accuracy was rather small, when epistasis was modelled and not simulated. On the basis of model (2), the estimates of additive genetic parameters (breeding values; results not shown) and the variance components (Table 2) were improved.

Table 2: Average estimated variance components based on model (2)

	σ^2_a	σ^2_d	σ^2_{aa}	σ^2_{ad}	σ^2_{da}	σ^2_{dd}
Epistasis was not simulated	0.826	0.059	0.004	0.009	0.008	0.011
Epistasis was simulated	0.860	0.067	0.032	0.018	0.017	0.028
Simulated variance components	1.000	0.067	0.150	0.038	0.023	0.045

Conclusion

This simulation study showed that the fast Bayesian method is convenient for genetic value prediction, but it is rather sensitive concerning variation in default values for the hyper-parameters involved. If pairwise epistatic effects are included in fitting a model to a trait, the re-parameterisation method of Álvarez-Castro and Carlborg (2007) leads to numerically stable iterations.
Our version of fBayesB will be published on www.genomicselection.net in due course.

References

Álvarez-Castro, J. and Carlborg, Ö. (2007). *Genetics*, 176:1151–1167.

Cockerham, C. (1954). *Genetics*, 39:859–882.

Kao, C.-H. and Zeng, Z.-B. (2002). *Genetics*, 160:1243–1261.

Meuwissen, T., Hayes, B., and Goddard, M. (2001). *Genetics*, 157:1819–1829.

Meuwissen, T., Solberg, T., Shepherd, R., and Woolliams, J. (2009). *Genet. Sel. Evol.*, 41:2.

Zeng, Z.-B., Wang, T., and Zou, W. (2005). *Genetics*, 169:1711–1725.