An application of stochastic variable selection for the genome-wide estimation of genetic effects and their complexity

Dörte Wittenburg and Norbert Reinsch*

Leibniz Institute for Farm Animal Biology, Genetics & Biometry
Wilhelm-Stahl-Allee 2, 18196 Dummerstorf, Germany

Introduction: Quantitative traits are typically influenced by a myriad of genes, whereat their number and kind of interplay are hardly known. In breeding applications, it is of special interest to identify the genomic regions with additive genetic effect, but the impact of non-additive genetic effects (dominance, epistasis) is important, for instance, for mate allocation or for studying heterosis effects in cross-breeding schemes. Thus, to understand the genetic architecture of a complex trait, it is desired to distinguish the kind of genetic effect and to separate the non-zero effects from unimportant ones. A variety of stochastic variable selection (SVS) approaches exists, that allow for shrinkage of potential zero effects. We adapted a previously published spike and slab approach, which enables the direct estimation of complexity parameters representing the proportion of non-zero effects for each kind of effect. With aid of the complexity parameters, an empirical selection procedure was appended to determine the significance of the non-zero effects a posteriori. The suitability of this approach is verified with simulations.

Methods: We model the genetic effects in a Bayesian framework using the spike and slab approach of Ishwaran & Rao (2005, model 4). The trait of interest $\boldsymbol{y} = (y_1, \ldots, y_n)'$ is fitted by

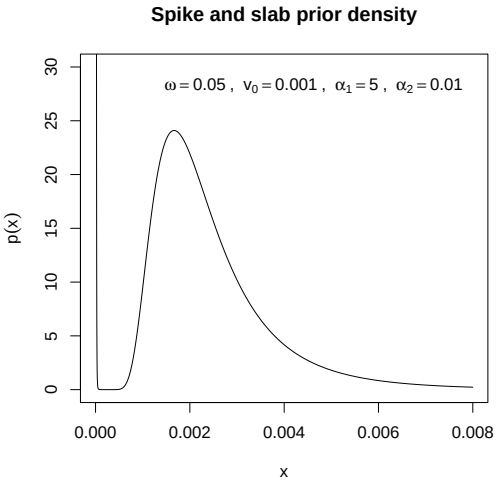
$$\boldsymbol{y} \;=\; \mathbf{1}\mu + \boldsymbol{X}\boldsymbol{g}_a + \boldsymbol{D}\boldsymbol{g}_d + \boldsymbol{e}\,.$$

Let μ be a population mean and $\mathbf{1}$ a vector of ones. In total m loci are studied on the genome. The $\boldsymbol{X}$ and $\boldsymbol{D}$ are the design matrices for the additive $\boldsymbol{g}_a = (g_{a,1}, \ldots, g_{a,m})'$ and dominance $\boldsymbol{g}_d = (g_{d,1}, \ldots, g_{d,m})'$ genetic effects, respectively. 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 $D_{i,j} = 1$. In order to obtain uncorrelated additive and dominance genetic values, i.e. $\mathsf{Cov}(X_{i,j}g_{a,j}, D_{i,j}g_{d,j}) = 0$, the columns of the design matrices are re-parametrised according to the theory of Álvarez-Castro & Carlborg (2007). Furthermore, it is assumed that genotypic effects at different loci are independently distributed. In order to omit the estimation of μ, the vector of observations is shifted by their mean value. In contrast to the originally published work (Ishwaran & Rao, 2005), we distinguish different kinds of effects and use effect-specific hyper-parameters. We apply the following prior distributions to the model

$$\begin{aligned}
\boldsymbol{y}|\boldsymbol{g}_a, \boldsymbol{g}_d, \sigma_e^2 &\sim N(\boldsymbol{X}\boldsymbol{g}_a + \boldsymbol{D}\boldsymbol{g}_d, \boldsymbol{I}\sigma_e^2)\,, \\
\sigma_e^{-2}|\beta_1, \beta_2 &\sim \Gamma(\beta_1, \beta_2)\,, \\
g_{s,j}|\sigma_{s,j}^2 &\sim N(0, \sigma_{s,j}^2)\,, \quad s \in \{a, d\},\ j \in \{1, \ldots, m\}\,.
\end{aligned}$$

*wittenburg@fbn-dummerstorf.de

The main idea is to shrink the posterior expectation of zero genetic effects. This is done by adjusting the value of the corresponding hyper-variance, which is obtained from two components, i.e. $\sigma^2_{s,j} := \mathcal{J}_{s,j}\tau^2_{s,j}$. A latent variable $\mathcal{J}_{s,j}$ is required to classify zero and non-zero genetic effects of kind $s \in \{a, d\}$. The classifier depends on a complexity parameter. Because we assume that the proportion of non-zero effects is different for each kind of effect, we consider effect-specific complexity parameters: ω_a reflects the proportion of the genome causing additive genetic variance and ω_d denotes the proportion responsible for dominance variation. The remaining prior distributions are



$$\begin{aligned}
\mathcal{J}_{s,j}|\omega_s, v_0 &\sim (1-\omega_s)\delta_{v_0} + \omega_s\delta_1\,, \quad s \in \{a,d\},\ j \in \{1,\ldots,m\}\,,\\
\omega_s &\sim U[0,1]\,,\\
\tau^{-2}_{s,j}|\alpha_1,\alpha_2 &\sim \Gamma(\alpha_1,\alpha_2)\,.
\end{aligned}$$

The symbol δ_x denotes the Dirac delta at point x. Thus, the hyper-variance of a genetic effect follows the bimodal distribution

$$\sigma^2_{s,j}|\omega_s,\alpha_1,\alpha_2,v_0 \sim (1-\omega_s)\Gamma^{-1}(\alpha_1,\alpha_2 v_0) + \omega_s\Gamma^{-1}(\alpha_1,\alpha_2)\,.$$

Owing to the spike at v_0 and the tail on the right, a genetic effect is either shrunk or enlarged. Because $v_0 > 0$ (but small), we achieve continuity of the prior distribution, which enables the implementation of a Gibbs sampler for this hierarchical model; the details can be found in Ishwaran & Rao (2005). The parameters $\alpha_1, \alpha_2, \beta_1, \beta_2, v_0$ have to be specified a priori.

The significance test of a genetic effect $g_{s,j}$ of kind $s \in \{a, d\}$ at locus $j \in \{1, \ldots, m\}$ is carried out via a conditional test. The hypothesis testing problem is

$$H_0:\ g_{s,j} = 0 \quad \text{vs.} \quad H_A:\ g_{s,j} \neq 0 \quad \text{for} \quad j \in \{1,\ldots,m\}, s \in \{a,d\}\,.$$

In order to fulfil the condition that $\omega_s \cdot 100\,\%$ of the genetic effects in $\boldsymbol{g}_s$ are different from zero, we apply an empirical selection procedure. First, we evaluate how often a genetic effect was classified as non-zero. On this account we sum over the corresponding indicator when the burn-in phase (iteration $k = 1, \ldots, b$) has been completed

$$\mathbf{1}_{s,j} = \sum_{k=b+1}^{B} \mathbb{1}^{(k)}_{s,j} \quad \text{with} \quad \mathbb{1}^{(k)}_{s,j} = \begin{cases} 1 & \text{if } \mathcal{J}^{(k)}_{s,j} = 1\,,\\ 0 & \text{if } \mathcal{J}^{(k)}_{s,j} = v_0\,. \end{cases}$$

We obtain a vector $(\mathbf{1}_{s,1}, \ldots, \mathbf{1}_{s,m})'$ reflecting the importance of all genetic effects. Second, we characterise the condition for each kind of effect. Let $\mathcal{B}_s$, $s \in \{a, d\}$, be a set containing those indices of effects of which the importance is larger than a certain threshold C_s. The cut-off C_s has to be chosen depending on the (estimated) complexity parameter ω_s to fulfil the condition

$$|\mathcal{B}_s| = |\{j : \mathbf{1}_{s,j} > C_s\}| = \omega_s \cdot m\,.$$

Thus, the cut-off is determined as the empirical ω_s-quantile of $\mathbf{1}_{s,1}, \ldots, \mathbf{1}_{s,m}$. Eventually, the conditional test ψ can also be written as

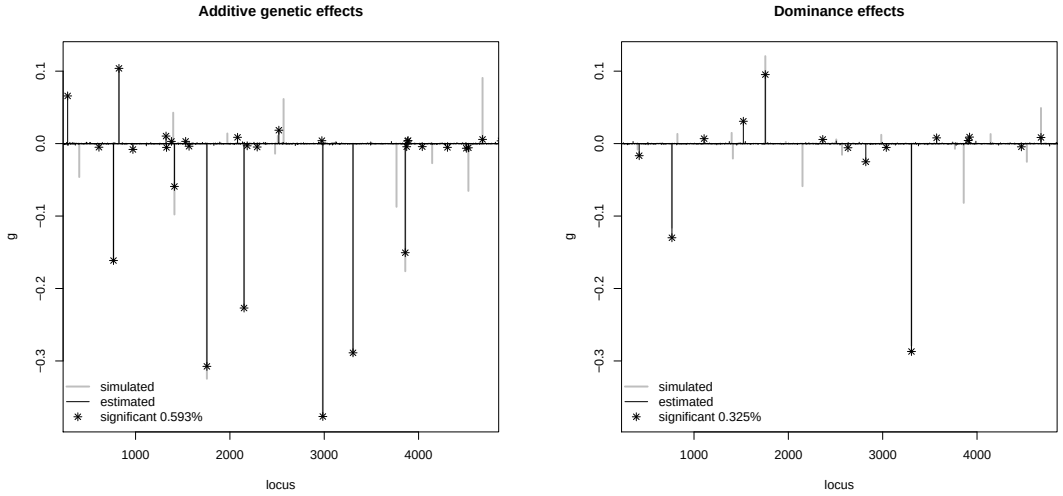
$$\psi(g_{s,j} \mid |\mathcal{B}_s| = \omega_s \cdot m) \quad = \quad \begin{cases} 1 & \text{if } \mathbf{1}_{s,j} > C_s\,, \\ 0 & \text{else}\,. \end{cases}$$

That means, H_0 is rejected for $g_{s,j}$ if this effect was classified as non-zero more than C_s times.

Simulation study: The simulated scheme resembles a dairy cattle population. On the genetic level, the information of single nucleotide polymorphisms (SNP) is employed. 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). The details of simulation can be found in Melzer *et al.* (2011). Two main scenarios were set up which differed in the number of QTL. Either 23 or 230 SNP loci were randomly chosen to be the QTL. Allele substitution effects were drawn from a gamma distribution with fixed shape parameter and varying scale parameter depending on the number of QTL similar to Meuwissen *et al.* (2001). The sign of an allele substitution effect was drawn at random with equal chance. Dominance coefficients were drawn from a normal distribution (Bennewitz & Meuwissen, 2010). The residual variance component was determined depending on the broad-sense heritability of $H^2 = 0.5$. The two training generations consisted each of 50 half-sib families with 20 offspring. These individuals were genotyped and phenotyped ($n = 2,000$). The test generations were built up the same way. The scenarios were repeated 100 times. We carried out 50,000 Gibbs sampling rounds of SVS, but 40 % were omitted as burn-in. We set the prior parameters $\beta_1 = \beta_2 = 0.0001$, $\alpha_1 = 5$, $v_0 = 0.001$ and studied two choices of α_2: $\alpha_2 = 0.1$ (called P1) and $\alpha_2 = 0.01$ (called P2). The prior density with parameter constellation P1 is similar to the BayesB prior, whereas the prior density with P2 is more alike the BayesA prior (Meuwissen *et al.*, 2001).

The estimated components were compared with results obtained from BayesB. This MCMC approach is similar to the presented SVS method in assuming a bimodal prior for the hyper-variance, but this prior is not continuous. For this reason, BayesB invokes a Metropolis-Hastings (MH) algorithm. Furthermore, prior knowledge about the proportion π of non-zero genetic effects in total is required; we set $\pi = 0.005$ in the 23-QTL scenario and $\pi = 0.05$ in case of 230 QTL. We carried out 50,000 MCMC rounds (40 % burn-in) and within each iteration 1,000 MH steps were executed.

Results and discussion: For a simulation example, Figure 1 shows the estimated additive and dominance genetic effects. Large effects were estimated well and their significance (indicated by



(a) (b)

Figure 1: Simulation example; estimates of (a) additive and (b) dominance effects in the 23-QTL scenario (P2). Accuracy of genetic value prediction was 0.947 in this dataset.

a star) could be proved in most cases. Moderate to small effects, especially dominance effects, could hardly be identified. In order to improve parameter estimation, one could retain the significant effects and repeat the statistical analysis.

All simulated QTL caused both additive and dominance effects, but the proportion of non-zero dominance effects was seriously underestimated in all scenarios. The choice of the parameter α_2 affected the amount of bias, see Table 1. Best results in terms of estimated complexity parameters were obtained with parameter constellation P2. The estimated proportion of non-zero additive effects coincided roughly with the simulated number of QTL, whereas only 0.2 % of the loci were estimated to contribute to dominance variation ignoring the real number of QTL. In any case, the accuracy of prediction was at a high level depending on the number of QTL. The genetic variance components estimated with P1 were less biased compared with P2 and rather similar to the analysis with BayesB. In the simulated scenarios, where the dominance effects of 23 or 230 loci caused approximately 5 % of the genetic variation, dominance plays a subordinate role in genomic selection and may be neglected. A more interesting scenario appears if a few loci contribute to a larger extent of dominance variation. In this case, methods for genetic value prediction may be improved by including some additional parameters for those loci with significant dominance effect.

The results obtained with the SVS approach may also be utilised in one-step approaches in the field of genomic selection. Therein the genetic value is determined e.g. via BLUP, where a genomic relationship matrix is required. This matrix is calculated on the basis of genome-wide marker information and at this point it would be possible to provide loci with a larger weight to emphasise the significant loci. The amount of extra weight could also be concluded, in some way, from the vector of importance.

Conclusions: The presented SVS approach with appended selection procedure is useful to estimate additive and non-additive genetic effects and to test the significance thereof. A com-

Table 1: Average estimated variance components and complexity parameters (standard deviation in parenthesis) in training set and average correlation ρ of predicted genetic values in test set.

	Method	σ_a^2	σ_d^2	σ_e^2	ω_a	ω_d	ρ
23-QTL scenario	SVS P1	0.727	0.037	0.653	0.002	0.001	0.952
		(0.569)	(0.037)	(0.611)	(0.001)	(0.000)	
	SVS P2	0.699	0.027	0.768	0.006	0.002	0.977
		(0.555)	(0.034)	(0.621)	(0.002)	(0.001)	
	BayesB	0.743	0.035	0.775	–	–	0.980
		(0.578)	(0.039)	(0.605)			
	Simulated	*0.757*	*0.040*	*0.798*	*0.004*	*0.004*	–
230-QTL scenario	SVS P1	0.541	0.022	0.666	0.007	0.001	0.872
		(0.180)	(0.023)	(0.228)	(0.001)	(0.000)	
	SVS P2	0.468	0.015	0.753	0.023	0.002	0.881
		(0.145)	(0.021)	(0.225)	(0.004)	(0.001)	
	BayesB	0.631	0.056	0.652	–	–	0.860
		(0.204)	(0.035)	(0.180)			
	Simulated	*0.709*	*0.043*	*0.754*	*0.044*	*0.044*	–

variance components: σ_a^2 additive, σ_d^2 dominance; σ_e^2 residual; complexity parameter: ω_a additive, ω_d dominance

plexity parameter is evaluated for each kind of effect, which enables the study of the genetic architecture of the underlying trait. In simulations, where most of the genetic variation was caused by additive effects, it was also possible to assess the contribution of dominance effects. In general, the inclusion of epistatic effects is possible as well, but this issue remains a question of computational effort.

References

Álvarez-Castro, J. M. & Carlborg, Ö. (2007), A unified model for functional and statistical epistasis and its application in quantitative trait loci analysis. *Genetics* **176**(2), 1151–1167.

Bennewitz, J. & Meuwissen, T. H. E. (2010), The distribution of QTL additive and dominance effects in porcine F2 crosses. *Journal of Animal Breeding and Genetics* **127**(3), 171–179.

Ishwaran, H. & Rao, J. S. (2005), Spike and slab variable selection: Frequentist and Bayesian strategies. *The Annals of Statistics* **33**(2), 730–773, URL `http://www.jstor.org/stable/3448605`.

Melzer, N., Wittenburg, D. & Repsilber, D. (2011), Simulating a more realistic genotype-phenotype map for development of methods to predict phenotypes based on genome-wide marker data - the livestock breeding scenario. *(submitted)* .

Meuwissen, T. H., Hayes, B. J. & Goddard, M. E. (2001), Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**(4), 1819–1829.