Evaluation of antedependence model performance and genomic prediction for growth in Danish pigs

L. Wang, D. Edwards, L. Janss
Center for Quantitative Genetics and Genomics, Department of Molecular Biology and Genetics,
Aarhus University, Denmark

ABSTRACT: The widely used genomic prediction models such as GBLUP, BayesA/B models all assume marker effects independent. Bayesian antedependence models extend this by estimating correlated marker effects, arising from linkage disequilibrium between markers. In this study we compare model fit and complexity, as well as prediction accuracy between antedependence models and other models applied to Danish Duroc pig data, including 29,567 SNPs. The results showed that anteGBLUP model and other conventional models did equally well in prediction. DIC for the models showed that anteBayesA and double-anteBayesA models had better fit, but higher number of effective parameters, and lower accuracy. In conclusion, the simple antedependence model works well for genomic prediction, but more complex antedependence models may be interesting to estimate marker effects correcting for LD structure. The DIC appears a good indicator of prediction accuracy.
Keywords: antedependence model; genomic prediction; model complexity and fit

Introduction:

Whole genome regression models used in genomic prediction relate high-density genome-wide SNP genotypes to relatively small number of phenotypes (Meuwissen et al. (2001)). To date, there are mainly two types of models applied to genomic prediction in livestock. The first is GBLUP, which is a linear mixed model assuming SNP effects independently and identically follow a zero-mean Gaussian distribution, and equivalent to Bayesian ridge regression (BRR). The second type of models are Bayesian hierarchical models with different prior distributions on the SNP effect variance, such as BayesA model which assigns scaled inverse Chi-square prior for the SNP effect variances (Campos and Hickey (2013)). These widely used models can be implemented through Markov chain Monte Carlo (MCMC) sampling. Most studies only compare models in prediction ability, paying little attention to the accuracy of MCMC estimates of SNP effects, variance or hyperparameter for variance, neither in model fit and complexity.

The common feature of the above mentioned models is that SNP effects are assumed marginally independent. However, the existence of linkage disequilibrium (LD) between SNPs raised the problem of slow mixing in SNP effects. Due to LD, SNP effect estimates spread over regions in high LD with causal genes, and make it difficult to identify true QTL. To counter this problem, Yang and Tempelman (2012) proposed Bayesian antedependence models which model SNP effects as correlated by introducing an extra second level into the conventional BayesA/B models. It was reported in their paper that the Bayesian antedependence models were more accurate in prediction than their classical counterparts when applied to simulation studies and real rat data of small genome size.

In this study we present three versions of antedependence models: anteGBLUP, anteBayesA and double-anteBayesA models. The aims of this study are: evaluate antedependence model performance applied on a real pig genome data used in breeding industry; compare antedependence model in regard to MCMC mixing, model fit and prediction accuracy with GBLUP and BayesA/B models.

Materials and methods

Models: The GBLUP/BRR model can be written as:

$$y = \mu + Zg + e$$

where y is the phenotype vector, Z is the matrix for SNP genotypes with allele dosage coding as "0/1/2", g is a zero-mean vector of the random allele substitution effects with variance to be specified below, and $e \sim N(0, I\sigma_e^2)$ is the vector for residuals. GBLUP/BRR assumes equal $\sigma_{g_i}^2$ for each SNP effect g_i, whereas BayesA model allow heterogeneous $\sigma_{g_i}^2$ drawn from $\chi^{-2}(v_g, s_g^2)$, and BayesB further extends $\sigma_{g_i}^2$ to a mixture distribution as $\pi_g \cdot 0 + (1 - \pi_g) \cdot \chi^{-2}(v_g, s_g^2)$ where π_g represents the proportion of no-effect SNPs. (Meuwissen et al. (2001); Campos and Hickey (2013))

The antedependence models are constructed in two levels, with the first level based on GBLUP model. In the second level, g is modeled as correlated in such manner:

$$g_i = \begin{cases} \eta_1, & i = 1 \\ r_i g_{i-1} + \eta_i, & i > 1 \end{cases}$$

where r_i is the regression parameter, which is due to LD between markers or correlation between marker effects (Gianola et al. (2003)). The estimates of marker effects are interpreted as correcting for LD impact, i.e., potentially retrieving the true underlying effects. Three versions of antedependence models differ in prior distributions of variance of g and r.

1. anteGBLUP: $g \sim N(0, I\sigma_g^2)$, and $r \sim N(0, I\sigma_r^2)$, with flat prior for σ_g^2 and σ_r^2 to be estimated.
2. anteBayesA: allow g heterogeneous, $g_i \sim N(0, \sigma_{g_i}^2)$, with $\sigma_{g_i}^2 \sim \chi^{-2}(v_g, s_g^2)$, of which d.f. ($v_g$) is fixed to 5, and scale (s_g^2) will be estimated; keep $r \sim N(0, I\sigma_r^2)$, and fix σ_r^2=1.
3. double-anteBayesA: $\sigma_{g_i}^2$ is kept in line with anteBayesA model; further allow r heterogeneous,

$r_i \sim N(0, \sigma^2_{r_i})$, with $\sigma^2_{r_i} \sim \chi^{-2}(v_r, s^2_r)$, and fix v_r=5, and s^2_r=1.

We fix some parameters, such as d.f., because to estimate these parameters jointly from the data accurately may not be possible due to LD (Campos and Hickey (2013)). Polygenic effect (u_j) for each animal is also included later in comparing different models, where $u_j \sim N(0, G\sigma^2_u)$ and G is the marker-derived realized relationship matrix (VanRaden (2008)).

Monte Carlo error, model fit, and prediction accuracy: Antedependence models are implemented through MCMC sampling, and the mixing of both marker effects (g) and the variance component (σ^2_g) are assessed. The Monte Carlo error of the estimates of marker effects was evaluated as (1-*c*), where *c* is the correlation between estimates from two independent chains. Time-series standard error is also calculated to check the accuracy of posterior mean estimates of marker effect variance σ^2_g or scale s^2_g, using R Coda package (Plummer et al. (2006)). Model fit and complexity are compared using deviance information criterion (DIC, Spiegelhalter et al. (2002)). DIC is a Bayesian information criterion combining model fit and a penalty of 2pD, where pD is an estimate for the effective number of model parameters. Genomic estimated breeding values (GEBV) are calculated by multiplying estimated marker effects ($\hat{g}$) obtained from training animals with testing animal genotypes. Prediction accuracy (p_a) is assessed by the correlation of GEBV with phenotypes in testing animals (Pérez-Cabal et al. (2012)). The standard error for prediction accuracy was computed on the correlation as$SE = \sqrt{(1 - p^2_a)/(n-2)}$, where n is the number of animals used for testing.

Data: A dataset of Danish Duroc pigs is used to evaluate antedependence model performance (Ostersen et al. (2011)). There are 4,244 pigs in total, with daily gain (g/day) as the phenotype. All the pigs were genotyped using Illumina PorcineSNP60 BeadChip. A total of 29,567 SNPs from 18 autosomal chromosomes are included in the analysis. The criterion applied here for data editing is the same as Wang et al. (2013). The average linkage disequilibrium (r^2) between adjacent SNPs is 0.55, which is relatively high among livestock species (de Roos et al. (2008); Uimari and Tapio (2011)).

Cross-validation to evaluate prediction accuracy was performed as follows: first the whole population was divided into training and testing sets by animal birthdate before or after October 1, 2008; to assess the effects of relationship between training and testing animals on prediction accuracy, the testing set was further divided into two subsets: one with no parents in training (test0), and the other with at least one parent in training (test12). The sizes of training, test0 and test12 are 2430, 1392 and 422, respectively.

Results and discussion:

All models were run 500,000 MCMC samples, with 10% of the sample as burn-in. Table 1 shows that Monte Carlo errors for marker effects in different models were about 1%, except double-anteBayesA model (~4.5%). The Monte Carlo errors for GEBV were under 0.1% in all models and testing groups (not presented in Table 1). The time-series standard errors for the posterior mean of $\sigma^2_g / \sigma^2_1 / s^2_g$ were considered acceptable. The anteGBLUP and BayesA models had the lowest DIC, and highest prediction accuracy for test0 (Table 2). For the two anteBayesA models, DIC indicates that these models had better fit, if we retrieve the penalty of effective number of model parameters (2pD) from DIC, as pD for these models were so high. This implies that the anteBayesA models were too complex for this data to achieve accurate parameter estimates. The models with the worst DICs had the lowest prediction accuracies.

Table 1. Monte Carlo error (MCE) for marker effect estimates, deviance information criterion (DIC), effective number of parameters in a model (pD), posterior mean (PM) of different variance related parameters and Time-series standard error (TSSE) of posterior mean, for GBLUP (GBP), BayesA (BA), BayesB (BB), anteGBLUP (aGBP), anteBayesA (aBA) and double-anteBayesA (d-aBA) models.

Models	MCE	DIC	pD	PM	TSSE
GBP	0.007	22305	453	0.15[1]	0.0004
BA	0.007	22304	455	0.08[2]	0.0005
BB	0.002	22346	404	47.7[3]	0.2176
aGBP	0.015	22305	453	0.08[1]	0.0054
aBA	0.018	22564	1192	0.43[2]	0.0149
d-aBA	0.048	22614	1236	0.34[2]	0.0162

[1]posterior mean for marker effect variance, σ^2_g.
[2]posterior mean for scale parameter s^2_g of $\sigma^2_{g_i}$.
[3]posterior mean for non-zero effect marker variance σ^2_g.

Table 2. Prediction accuracy (standard error) of the correlation between phenotypes and genomic estimated breeding values (GEBV) for testing group with no parent in training (test0) and testing group with at least one parent in training (test12), for GBLUP (GBP), BayesA (BA), BayesB (BB), anteGBLUP (aGBP), anteBayesA (aBA) and double-anteBayesA (d-aBA) models.

	test0[3]		test12[4]	
Models	-[1]	+ G[2]	-[1]	+ G[2]
GBP	0.179	-	0.387	-
BA	0.177	0.178	0.387	0.387
BB	0.165	0.177	0.388	0.391
aGBP	0.179	-	0.389	-
aBA	0.160	0.161	0.326	0.326
d-aBA	0.154	0.157	0.318	0.318

[1]models without polygenic effects.
[2]models with polygenic effects, G is marker-derived realized relationship matrix, as the variance-covariance matrix for animals.
[3]standard error for prediction accuracy is about 0.026~0.027 for all the models validated by test0 group.
[4]standard error for prediction accuracy is about 0.045~0.047 for all the models validated by test12 group.

By comparing the prediction accuracy (Table 2), the GBLUP and anteGBLUP models obtained accuracy as high as BayesA/B models, which could be explained by close relationship within pig population or many small-effect genes regulating the trait (daily gain). It should be noticed that BayesB achieved highest prediction accuracy (0.391) after including polygenic effects for test12 group, which have closer relationship with training animals. The two anteBayesA models had 6% lower accuracy compared with other models in test12 group, but only 0.5~1% decrease compared to other models in test0 group. It indicated that anteBayesA models performed well by exploiting LD to maintain equal prediction accuracy for less related animals to the reference population, even with high model complexity.

Figure 1 plots the estimates of marker effects from GBLUP and double-anteBayesA (y-axis) against the estimates from BayesB (x-axis). The shrinkage of marker effects for double-anteBayesA was intermediate between the other two models. We found double-anteBayesA picked out some markers with large effects which were not identified by BayesB, an interesting result which should be investigated in future work.

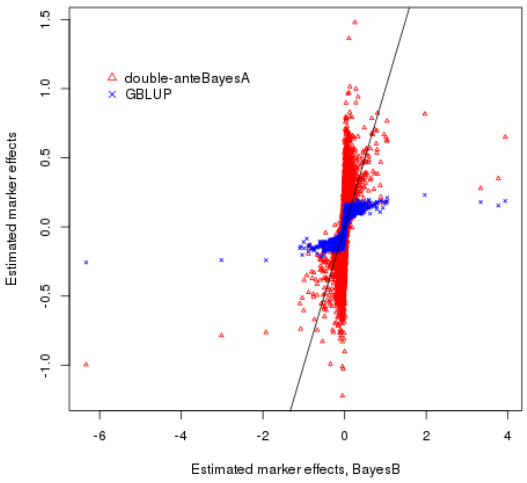

Figure 1. Estimated marker effects from double-anteBayesA and GBLUP models (y-axis) against estimated marker effects from BayesB model for daily gain in pigs

Conclusion:

The anteGBLUP model with simple parameter settings performed equally well as GBLUP and BayesA/B in prediction applied in the pig data. The anteBayesA and double-anteBayesA models showed lower accuracy compared to other models. These models had higher number of effective parameters, which suggests the accuracy to estimate marker effects is poor in current dataset due to model complexity. We recommend that the simple anteGBLUP model be used for genomic prediction, and the more complex anteBayesA models may be interesting to estimate marker effects corrected for LD structure.

Literature Cited

Campos, G. de los, and Hickey, J. (2013). *Genetics*. 193:327–345.
Gianola, D., et al. (2003). *Genetics*. 163:347–65.
Meuwissen, T.H., et al. (2001). *Genetics*. 157:1819–29.
Ostersen, T., et al. (2011). *Genet. Sel. Evol.* 43:38.
Pérez-Cabal, M.A., et al. (2012). *Genet.* 3:27.
Plummer, M., et al. (2006). *R News*. 6:7–11.
De Roos, a P.W., et al. (2008). *Genetics*. 179:1503–12.
Spiegelhalter, D.J., et al. (2002). *J. R. Stat. Soc. Ser. B (Statistical Methodol)*. 64:583–639.
Uimari, P., et al. (2011). *J. Anim. Sci.* 89:609–614.
VanRaden, P. (2008). *J. Dairy Sci.* 91:4414–4423.
Wang, L., et al. (2013). *BMC Genet.* 14:115.
Yang, W., and Tempelman., R.J. (2012). *Genetics*. 190:1491–501.