Unravelling the ancestral regression: 1. comparative performance with known genome-wide relationships

S. Munilla1,2, C.A. García-Baccino1, N.S. Forneris1 & R.J.C. Cantet1,2

1 Departamento de Producción Animal, Facultad de Agronomía, Universidad de Buenos Aires, Av. San Martín 4453, C1417DSE, CABA, Argentina.
2 Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Godoy Cruz 2290, C1425FQG, CABA, Argentina.
munilla@agro.uba.ar (Corresponding Author)

Summary

The ancestral regression model generalizes the covariance structure assumed under the regular animal model by adding a set of extra path coefficients directly connecting an individual’s breeding value to his grandparents’ breeding values. The model was developed for performing genomic selection with a practical advantage related to the ease of computation, as the inverse of the genomic relationship matrix can be directly build-up by means of sequential contributions. To test the performance of the model in terms of prediction, we used data from a simulation experiment. The objective of this study was to compare the performance of the ancestral regression model when the true genome-wide relationships between an individual and his four grandparents are known. The model was compared to a regular animal model and we also fitted a single-step genomic selection model with the true genomic relationship matrix instead of the usual G matrix. The latter was considered as the reference model. The simulation experiment involved a simplified scenario of a pig nucleus breeding programme for an additive trait controlled by 250 segregating QTLs. After an historical population created to attain mutation-drift equilibrium with a realistic level of linkage disequilibrium, five new generations were developed by mating 20 boars to 200 sows and producing 2,000 offspring. The individuals in the last generation were considered candidates to selection and we focus on predicting their breeding values. Four statistics were computed: accuracy, bias, inflation and mean square error. All models behave very well in terms on bias and inflation. In turn, differences arose in terms of accuracy. The reference model showed a mean accuracy (across replicates) of 0.76, whereas the ancestral regression model outperformed the regular animal model (0.41 vs 0.32), showing that some knowledge was gained when including direct path coefficients from the grandparents.

Keywords: genomic selection, simulation, breeding value prediction, predictive performance

Introduction

The “ancestral regression” (AR) model generalizes the covariance structure assumed under the regular animal model, widely used to produce traditional genetic evaluations, with a set of new parameters that complement the information given by the pedigree with that provided by a dense panel of genomic markers (Cantet et al., 2017). The AR is formally interpreted as a recursive Gaussian system of equations and was derived after the idea of a causal relationship among breeding values (BV) in adjacent generations coupled with a set of extra (uncorrelated) path coefficients directly connecting grandparents to grand-progeny. In this
framework, marker data helps to differentiate between the expected inheritance process and the realized genetic process resulting from the mechanisms of recombination and segregation on a segmental background (Matsen & Evans, 2008).

The AR model was presented as a theoretical contribution and now the time has come to test its potential. The model was thought for performing genomic selection with two main practical advantages: 1. Ease of computation, as the inverse of the genomic relationship matrix can be directly build-up following rules similar to those by Henderson (1984) and thus no inversion is required; 2. A (potentially) reduced genotyping effort, as for each candidate to selection only parents and grand-parents should be genotyped to disentangle their differential contribution. The model also encompasses some challenges. First, the recursion stops at the grand-parental generation, whereas theoretically it could be performed backwards up to the founder population. Second, it assumes that realized genome-wide relationships (GWR) of an individual and his four grandparents are known, which in practice means all four grandparents must be genotyped and their phases known.

To tackle these challenges we used data from a simulation experiment to compare the performance of the AR model in estimating genomic breeding values against available methods. For ease of presentation we report our preliminary results in two conference papers, each focusing on a different challenge. The objective of this first paper was to compare the performance of the AR regression when the true GWR are known. The method was compared to a regular animal model, i.e., using the genealogical relationship matrix, and to a single-step genomic selection approach where the true genomic relationship matrix was used to model the covariance among breeding values.

**Material and methods**

**Theory**

The AR model extends the linear regression of the breeding value of an individual on his parents’ breeding values by including all four grandparental BVs weighed with coefficients uncorrelated to the contribution of the parents. The AR is better defined by the following equation:

\[
\alpha_X = 0.5 \alpha_S + 0.5 \alpha_D + \beta_S [\alpha_{PGS} - \alpha_{PGD}] + \beta_D [\alpha_{MGS} - \alpha_{MGD}] + \phi_X,
\]  

(1)

where \(\alpha_X\) stands for the BV of the individual and the set \{S, D, PGS, PGD, MGS, MGD\} to his most recent ancestors; i.e., parents and grandparents. Notice that under the AR model coefficients \(\beta_S\) and \(\beta_D\) are related to the difference between the BVs of the paternal grandparents and the maternal grandparents, respectively. This is to ensure the identifiability of the parameters and is easily interpretable as deviations of the realized relationships from the expected ones. Thus, positive values of \(\beta\)’s mean that grand-sires had a greater contribution in the individual’s genome whereas negative values indicate that the grand-dams contribution is in excess.

The matrix representation of the Gaussian recursive linear system (1) is:

\[
a = Ba + \phi
\]  

(2)

Matrix \(B\) is lower-triangular with nonzero elements being the path coefficients relating the BV of ancestors to descendants (one-half for parents and the \(\beta\)’s for the grandparents).
Solution to equation (2) is:

\[ a = (I - B)^{-1} \varphi \]  \hspace{1cm} (3)

from where the following covariance structure arises:

\[ \Sigma = (I - B)^{-1} D (I - B^T)^{-1} \]  \hspace{1cm} (4)

where \( D \) is a diagonal matrix with entries corresponding to the variance of Mendelian residuals. Cantet et al. (2017) presented an algorithm to directly compute the inverse of this matrix by sequential contributions.

**Simulation experiment**

Data were obtained from the simulation by Forneris et al. (2016) and important details are presented here for the sake of self-containment. The experiment involved a simplified scenario for a pig nucleus breeding programme. By means of the QMSim software (Sargolzaei & Schenkel, 2009), a genome of 5 autosomal chromosomes of 160 cM each was generated and 250 segregating QTL were randomly distributed across the genome under an infinite allele model with a mutation rate of \( 2 \times 10^{-4} \) per locus per generation. Bi-allelic markers were also generated and distributed randomly across the genome, but they were not used in this study. Following a gene-dropping approach, an historical population was simulated by considering an equal number of males and females, discrete generations, random mating, no selection and no migration. Recombination was modelled at a rate of 1 cM/Mb assuming a Poisson distribution. After 2,500 generations with an effective size of 500 individuals, a severe bottleneck was induced for 30 generations by reducing the effective size to 75 so as to establish a population in mutation-drift equilibrium with a realistic level of linkage disequilibrium. After that, 20 males and 200 females were generated by random choice of two gametes from the male and female gametic pools. These animals constituted the founders of the recent population. At this point, additive effects were generated from a Gamma distribution and assigned to the extant segregating QTL’s alleles. The sizes of the effects were scaled so as to produce a genetic variance of 0.25 whereas phenotypic variance was set to one. Then, the following selection scheme was followed for five generations. In each generation, 20 boars were mated with 200 sows to produce 2,000 offspring. Mating design was optimized to minimize inbreeding. The 20 boars were selected based on their EBVs, whereas the 200 sows were randomly selected. Pedigree and phenotypic values were available for the 5 recent generations (10,000 animals), but we deleted phenotypic records from the last one to emulate a target population of candidates to selection. A total of 20 replicates were used to report our results here.

In addition, as the QMSim software stores the location of the cross-overs, we were able to trace the true inheritance path of each segment from the founder population to the target population (i.e., the individuals from the fifth generation). As a consequence, we had available the true proportion of the genome shared IBD (Hill & Weir, 2011) for each pair of individuals \( i \) and \( j \), which was computed by integrating along the genome the point-wise realized coancestry. Let \( r_{X,Y} \) denote such realized relationship for individuals \( X \) and \( Y \), then \( \beta_S \) and \( \beta_D \) coefficients were computed as (Cantet et al., 2017):

\[ \beta_S = 0.5 \left[ r_{X,PGS} - r_{X,PGD} \right] \]  \hspace{1cm} and \hspace{1cm} \[ \beta_D = 0.5 \left[ r_{X,MGS} - r_{X,MGD} \right] \]  \hspace{1cm} (5)
Comparative performance

Genomic prediction methods

We fitted the simulated data with three models: a regular animal model (AM-BLUP) and two Single-Step genomic selection models, one with a covariance matrix based on the true GWR for each pair of individuals (ssGBLUP-G) and one with a covariance structure as defined by equation (4) (ssGBLUP-AR). All computations were done by means of the blupf90 package (Misztal et al., 2002), with the option “user_file” to indicate the appropriate relationship matrix. We focused on the prediction of BVs (EBV) for the individuals in the target population (i.e., last generation) and compared the results to the true breeding values (TBV).

Prediction performance statistics

Prediction performance was assessed by the following statistics. First, the accuracy of the methods was measured as the Pearson correlation between TBV and EBV. Second, average bias and inflation were computed as the intercept and slope of the linear regression of TBV on EBV, as suggested by the Interbull validation method (Mantysaari et al., 2010). Notice that a desirable method should have maximum accuracy, zero bias and a regression coefficient of one. Finally, we calculated the mean squared error (MSE) for each model. Means and standard deviations were computed across replicates.

Results and Discussion

The results of our analyses are displayed in Table 1. Overall, means and standard deviations across replicates for all four statistics employed were similar to those reported by García-Baccino et al. (2017) under a similar simulation approach. All models performed well in terms of mean bias and inflation, although the standard deviation of bias across replicates was quite large. In turn, differences emerged in terms of accuracy. The reference model, based on a covariance matrix with entries corresponding to the true realized relationship between every pair of individuals in the pedigree, resulted in the maximum accuracy. In comparison, the AR model outperformed the regular animal model, indicating that some knowledge was accrued when including direct and uncorrelated path coefficients from the grandparents.

In this particular study, genome-wide relationships were assumed known. However, this is never the case. In real cases, the practitioner is armed with marker data and faced to the challenge of eliciting correlations between breeding values. In some scenarios, the amount of information available makes the problem of setting-up and inverting a genomic relationship matrix computationally unfeasible. We believe the approach derived from the ancestral regression, which parsimoniously extends the robust animal model to include uncorrelated grandparental contributions, could help to tackle this problem, as much less computational effort is required (Cantet et al., 2017). Our results indicated that some gain in accuracy was possible with the AR model, but still not enough to approach to the reference model. Of course, the estimation of the β’s is a real challenge and we explored the issue in our next communication.

Table 1. Predictive performance of the models fitted to the simulated data.

<table>
<thead>
<tr>
<th></th>
<th>ssGBLUP-G</th>
<th>AM-BLUP</th>
<th>ssGBLUP-AR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correlation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Statistic</td>
<td>Mean</td>
<td>Std</td>
<td>Mean</td>
</tr>
<tr>
<td>-----------</td>
<td>------</td>
<td>-----</td>
<td>------</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.76</td>
<td>0.04</td>
<td>0.32</td>
</tr>
<tr>
<td>Inflation</td>
<td>0.97</td>
<td>0.07</td>
<td>0.98</td>
</tr>
<tr>
<td>Bias</td>
<td>&lt;0.01</td>
<td>0.07</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>MSE</td>
<td>0.09</td>
<td>0.02</td>
<td>0.19</td>
</tr>
</tbody>
</table>

1 Single-step genomic BLUP with true realized relationship matrix G.
2 Regular animal model with relationship matrix A.
3 Single-step genomic BLUP with covariance matrix Σ.

**List of References**


