Random and implications for genomic relationship matrices

J.B. Holmes & M.A. Lee

University of Otago, Department of Mathematics and Statistics, Cumberland St, Dunedin 9016, New Zealand
jholmes@maths.otago.ac.nz (Corresponding Author)

Summary

Many models have been proposed to estimate stratification within populations, of which some treat allele frequency either explicitly or implicitly as a random variable. Through comparison of the implied genotype distributions, we demonstrate random allele frequency is inconsistent with the theory justifying genomic relationship matrix construction. In particular, random allele frequency introduces over-dispersion, which in the context of relationship estimation implies inbreeding, and also will lead to a highly constrained range of allowable relationships between animals. Using a model that assumes random allele frequency, we will show these differences lead to bias in estimates of close relationships and inbreeding coefficients.

Keywords: genomic relationship matrix, principal components, population stratification

Introduction

An ongoing goal in plant and animal breeding is the improvement of the accuracy of genetic evaluation. To achieve this, the inclusion of genetic group effects (Westell et al. 1988) has been proposed. While normally used to account for unknown parents, we would argue fitting groups is equivalent to controlling for population stratification. For genotyped populations, EIGENSTRAT (Patterson et al. 2006) is a popular technique to find such stratification. However as EIGENSTRAT is the principal component decomposition of a genomic relationship matrix (GRM), it has been argued in Janss et al. (2012) and others, that such an approach of defining group effects in BLUP leads to overfitting.

To avoid this overfitting, we considered models that explicitly treated single nucleotide polymorphism data as binomial. In Holmes et al. (2017), we examined the performance of such a model, binomial probabilistic principal components, in identifying strata within the population under analysis. In this paper, we will consider the implied GRM obtained from such a latent variable model and determine how it differs from a traditional GRM.

Data and models

Data

The genotype data (5K Illumina SNP Chip), previously used in Holmes et al. (2017), was obtained from 8,902 animals born from 2000 to 2014, each with genotypes recorded on up to 5,283 markers. Genotypes that were missing for more than 1% of animals or monomorphic for all animals were omitted from analysis. The removal of all other animals with any missing
genotypes reduced the dataset to 1,626 animals with 5,170 markers recorded.

Theoretical justification of the genomic relationship matrix construction

The expectation of elements of the GRM ($G$) in method one and two of VanRaden (2008) are,

\[(1)\]

\[(2)\]

where is the coefficient of inbreeding for animal $i$ and is the coefficient of relationship between animal $i$ and $j$. This implies the mean, variance and covariance of an observed genotype from marker $k$ are,

\[(3)\]

\[(4)\]

\[(5)\]

where is the allele frequency of marker $k$. While these results are well-known, we wanted to draw attention to the distributional implication in the context of binomial data. It implies is the sum of two identical but not independent Bernoulli trials with fixed.

Binomial probabilistic principal components (BPPCA)

Our choice of population structure model, BPPCA, was first presented in Holmes et al. (2017). This assumed marker from animal lies in Hardy-Weinberg equilibrium and thus can be viewed as a realisation of a binomial random variable,

\[(6)\]

where the allele frequency was a realisation of the random variable, . We then model using a logit link function, with taking the representation of probabilistic principal components with $f$ components,

\[(7)\]

where was the marker specific intercept, a loading matrix containing information on stratification with a flat prior, a score matrix with prior, and an error term. To fit the model, we used variational Bayes approximation (Klami, 2014) and the data augmentation approach of Polson et al. (2013) to ensure closed form approximate posteriors could be found for all parameters. Iterations were stopped when , where was fixed at .

Implied genomic relationship matrix from a population structure model

The elements, of the GRM given in VanRaden (2008) are equivalent to (method one) or (method two), where $m$ is the number of markers. As BPPCA treats as a random variable, can be determined using the laws of total variance,
This indicates BPPCA introduces over-dispersion in the genotype data, or equivalently assumes all animals are in-bred, since, such that the inbreeding coefficient or }, depending on method. This is despite BPPCA assuming Hardy-Weinberg equilibrium for all animals, which in traditional GRM would imply inbreeding is not present. From the properties of variance and covariance, the permissible range for the values of off-diagonal elements can be obtained,

\[
\text{(9)}
\]

which implies the coefficient of relationship must be within for both methods.

**Method of comparison**

To demonstrate the properties described for a GRM that assumes random, we fitted BPPCA with the number of factors fixed to 5, 10, 20, 50, 100 and 200. The implied GRM was constructed from the output of BPPCA assuming normalisation over all genotypes, and compared to the result of a GRM constructed using the first method of VanRaden (2008).

**Results**

In Figure 1, the elements of the GRM built using BPPCA when either 5 or 200 components were fitted were plotted against the corresponding elements of the GRM constructed using method one of VanRaden (2008). As expected from the theoretical results, the diagonal elements of the BPPCA GRM were consistently above one. For off-diagonal elements, the relationships between close relatives were under-estimated, particularly when the number of components fitted was small. The improvement in relationship estimation as more components were fitted is consistent with the behaviour of low rank matrix approximations.

**Table 1. Mean diagonal element of BPPCA GRM and correlation of off-diagonal elements of BPPCA compared against the VanRaden GRM off-diagonal elements at differing values of \(f\).**

<table>
<thead>
<tr>
<th>Number of components ((f))</th>
<th>5</th>
<th>10</th>
<th>20</th>
<th>50</th>
<th>100</th>
<th>200</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean of diagonal</td>
<td>1.033</td>
<td>1.046</td>
<td>1.065</td>
<td>1.086</td>
<td>1.092</td>
<td>1.102</td>
</tr>
<tr>
<td>Correlation off-diagonal</td>
<td>0.767</td>
<td>0.822</td>
<td>0.878</td>
<td>0.914</td>
<td>0.917</td>
<td>0.920</td>
</tr>
</tbody>
</table>

Table 1 gives the mean diagonal element of the BPPCA GRM. As the number of components fitted increased, the mean diagonal element increased. Earlier we showed there was a constraint in the BPPCA GRM that restricted the range of an off-diagonal element to within . In order to estimate close relationships, the diagonal elements of the BPPCA derived GRM must be inflated and increasing the number of components allows for this possibility. However, as Figure 1 shows, even when 200 components were fitted, we still could not estimate parent-child or sibling relationships within the data. This reflects that if genotype data is consistent with the assumptions implied in VanRaden (2008), the difference between the true level of inbreeding and the required level of inbreeding needed to estimate the true relationship between animals \(i\) and \(j\) under the assumptions of BPPCA is too great to correctly estimate either the inbreeding coefficient or the coefficient of relationship.
Figure 1. The elements of the BPPCA GRM against the VanRaden GRM for $f = 5$ or 200.

Conclusions

We suggest GRM construction requires stronger assumptions than commonly realised. We have demonstrated through theory and application that relaxing the assumption of two identical but not independent Bernoulli trials used to justify the construction of GRM by allowing random allele frequency results in negatively biased estimates of close relationships and positive biased estimates of inbreeding. This indicates that models that assume random are only appropriate for populations containing distantly related individuals.

List of References