Proc. Assoc. Advmt. Anim. Breed. Genet. 22:101-104

POSSIBILITIES OF BINOMIAL PROBABILISTIC PRINCIPAL COMPONENT MODELS TO IDENTIFY GROUPS IN GENOTYPED POPULATIONS

J.B. Holmes[1], K.G. Dodds[2], and M.A. Lee[1]

[1] Department of Mathematics and Statistics, University of Otago, Dunedin, New Zealand
[2] AgResearch, Invermay Agricultural Centre, Mosgiel, New Zealand

SUMMARY

There has been extensive research, particularly in human genetics, devoted to the development of methods that use genotype data for the identification of distinct genetic sub-populations within the population of interest. Some of these methods have also been incorporated in the field of animal breeding in order to improve the accuracy of predicted breeding values through their use as genetic group effects. In this paper, we compared a method of finding sub-populations based on a decomposition of a normalised matrix derived from genotype data, to a modified probabilistic PCA model that took into account the non-normal nature of the genotype data. In an initial study, where we used a dataset from the New Zealand sheep industry with a known breed composition, we found that the modified probabilistic PCA model gave equivalent separation between breeds to EIGENSTRAT.

INTRODUCTION

Livestock programs aim to optimise long-term genetic gain. To do this the ideal is for breeding values to be as accurate as possible. One method of improving breeding value accuracy is through the fitting of genetic groups. However in practice, genetic groups often prove difficult to define (Kuehn *et al.* 2007).

With the increased availability of genotype data, there has been a move towards replacing pedigree records with genotype data for the construction of the relationship matrix to improve breeding value accuracy. In addition there have been attempts to use genotype data to define structure within the population of interest, which is then fitted in the model, usually as a fixed effect. An example of this is EIGENSTRAT (Patterson *et al.* 2006), which in practice is very similar to the eigen-decomposition of the second genomic relationship matrix proposed in VanRaden (2008). This method ignores the non-normal nature of the genotype data and has been shown to reduce across breed accuracy when used as a genetic group (Daetwyler *et al.* 2012).

To deal with the issues outlined, we propose a probabilistic PCA model that explicitly takes into account the ideal conditions of binomially distributed genotypes. We then compared the two methods, focusing on their respective ability to distinguish between genetic groups, which we took to correspond to the recorded breed.

MATERIALS AND METHODS

Data. The genotype data (5K Illumina SNP Chip) available was from 8,902 animals born from 2000 to 2014, each with up to 5,283 markers recorded. Genotypes which were missing for more than 1 % of animals or monomorphic for all animals were omitted from analysis. The removal of animals with any missing genotypes reduced the dataset to 1,672 animals with 5,170 markers recorded. Breed composition data was obtained from Sheep Improvement Limited (SIL). The distribution of breeds in the dataset is indicated on Table 1.

EIGENSTRAT. This method of identifying population structure was introduced in Patterson *et al.* (2006). It assumes a $n \times m$ matrix of genotypes $\boldsymbol{Z}$ with rows corresponding to individuals and columns to markers and coded 0, 1, 2 where the numbers correspond to the number of copies of the A allele. Each column j of $\boldsymbol{Z}$ was then normalised by subtracting by twice the allele

frequency p_j and dividing the result by $\sqrt{p_j(1-p_j)}$ to form the matrix $\boldsymbol{M}$. Eigen-decomposition (Principal Component analysis) was then performed on the matrix $\frac{1}{m}\boldsymbol{MM}'$. Determination of population structure was then made using the resulting eigenvectors (Principal components).

Table 1. Breed distribution of genotyped animals as recorded in SIL

Breed distribution of animals			
Breed	Number of animals	Breed	Number of animals
Unknown	11	Perendale	133
Romney	495	Highlander	31
Coopworth	67	Composite	2
Overall distribution of breeds where known			
Breed	% in population	Breed	% in population
Romney	48.13	Poll Dorset	1.28
Coopworth	14.87	East Friesian	1.04
Perendale	13.88	Highlander	3.37
Finnish Landrace	1.12	Composite	3.57
Texel	6.70	Other Breeds	2.53
Suffolk	3.51	(less than 1 % of population)	

Binomial probabilistic principal component analysis (BPPCA). Under ideal conditions of Hardy-Weinberg equilibrium and no linkage disequilibrium, each of the markers *j* observed from individual *i* can be regarded as realisations of a binomial random variable.

$$\boldsymbol{Z}_{ij} \sim Bin(2, p_{ij}) \quad [1]$$

BPPCA assumes that the individual-marker specific allele frequency p_{ij} can be modelled using the link function $\boldsymbol{\theta}_{ij} = \log(p_{ij}/(1-p_{ij}))$ as a function of a marker specific intercept μ_j, f principal components, where *f* was pre-determined, and an error term. This results in the following model for the observed genotype pattern, where $\boldsymbol{\theta}$ is a $n \times m$ matrix of link functions, $\boldsymbol{L}$ a $n \times f$ matrix of components, $\boldsymbol{F}$ a $f \times m$ matrix of scores, and $\boldsymbol{e}$ is a $n \times m$ matrix of residuals.

$$\boldsymbol{Z}_{ij} \sim Bin(2, (1+\mathrm{e}^{-\theta_{ij}})^{-1})$$
$$\boldsymbol{\theta}_{ij} = \mu_j + \sum_k \boldsymbol{L}_{ik}\boldsymbol{F}_{kj} + \boldsymbol{e}_{ij},\ \boldsymbol{F}_{kj} \sim N(0,1),\ \boldsymbol{e}_{ij} \sim N(0, \sigma^2) \quad [2]$$

To fit the model, we used Pólya-gamma data augmentation as outlined in Polson et.al (2013) and previously implemented for a similar model in Klami (2014). This allowed closed form conditional posteriors to be obtained for all model parameters. Based on the eigenvalue scree plot obtained from implementing the EIGENSTRAT method, the number of components to fit was fixed at five. Estimates were obtained from the posterior means found by using a blocked Gibbs sampler based on the conditional posteriors. The Gibbs sampler was stopped once the relative change in $\bar{\boldsymbol{\theta}}$ dropped below 1×10^{-5}. Spectral value decomposition was then applied to the initial estimates to ensure orthogonal components. This ensured comparability of components to those extracted using EIGENSTRAT.

RESULTS AND DISCUSSION

Ability to separate breeds based on principal components. Figure 1 plots the first two principal components obtained from EIGENSTRAT and BPPCA with pure breed animals highlighted. Both methods were able to distinguish between different pure breed populations.

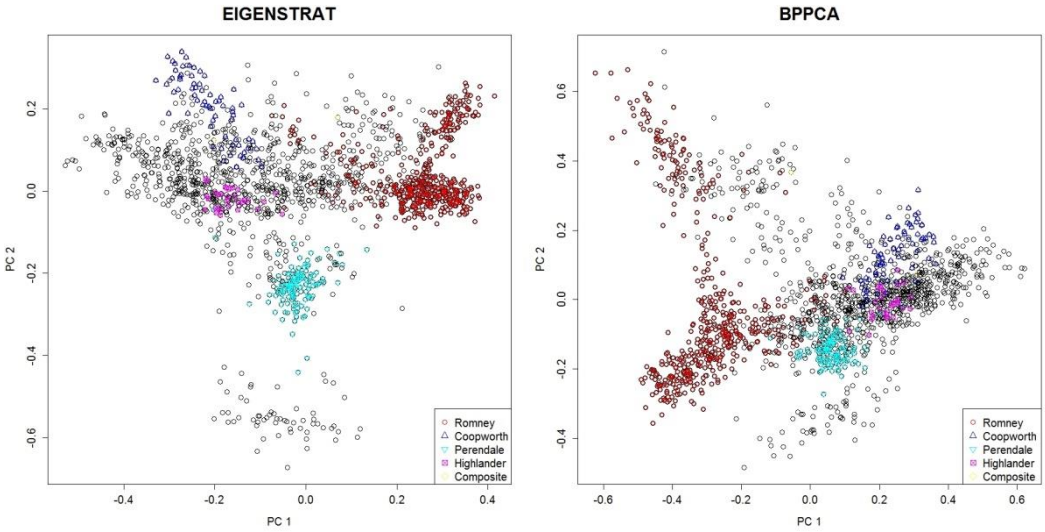

Figure 1. First two principal components obtained from EIGENSTRAT and BPPCA.

Possible uses of the principal components (PC) to represent population structure. Since it is established that principal component analysis on normalised genotype data can distinguish between sub-populations, the fitting of PC has been used extensively to account for population structure in models. The PC are usually fitted as fixed effects. Since EIGENSTRAT extracts PC from the decomposition of the genomic relationship matrix, we suggest that it is more appropriate to fit the PC as random effects. In addition, fitting a decomposition of the genomic relationship matrix in addition to the genomic relationship matrix could be regarded as over-fitting.

In BPPCA, PC are constructed at the link function level, not directly from the observed data. This means that the relationship between the PC and the genomic relationship matrix is indirect. This can be demonstrated by the law of total variance and noting that $E(\boldsymbol{p})$ and $Var(\boldsymbol{p})$, where $\boldsymbol{p}$ is the vector of latent probabilities for each animal, are both functions of the BPPCA PC. It may also mean representing population structure using PC from the BPPCA model is less prone to the reduction of across breed accuracy seen in Daetwyler *et al.* (2012).

$$Var(\boldsymbol{Z}) = E\big(Var(\boldsymbol{Z}|\boldsymbol{p})\big) + Var\big(E(\boldsymbol{Z}|\boldsymbol{p})\big) \quad [3]$$
$$= diag\big\{E\big(2\boldsymbol{p}(1-\boldsymbol{p})\big)\big\} + Var(2\boldsymbol{p}) = 2diag\{E(\boldsymbol{p}) - E(\boldsymbol{p})^2 - Var(\boldsymbol{p})\} + 4Var(\boldsymbol{p})$$

If the genotype data can be represented by a low rank matrix factorisation at the link function level, the correlations between animals implied by the PC would be higher (if correlation is positive) or lower (if correlation is negative) than the corresponding correlations in the genomic relationship matrix. However EIGENSTRAT extracts a reduced number of PC, which contain more information about covariance than variance elements. Therefore the implied correlation between random structure effects of different animals is similar between the two methods. This is shown in Figure 2, which shows heat maps of the implied between animal correlation.

Figure 2 shows if PC are used as a classification tool to distinguish between breeds, similar results were obtained from EIGENSTRAT and BPPCA. In our dataset, both clearly identify each pure breed population, sub-groups within the Romneys and classify the animals of unknown breed as Perendale. Corresponding PC extracted by the two methods were highly correlated, except for component 2 and 3, as seen in Table 2. The high negative correlation seen in component 1 and 5 is due to the sign invariance property of estimated loadings in latent factor models.

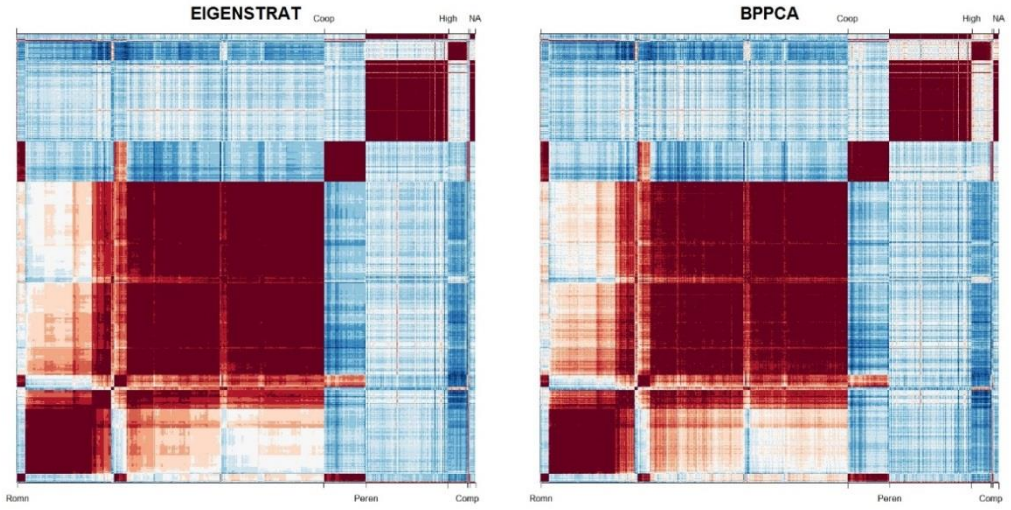

Figure 2. Heat maps of implied correlations between animals that were either of pure or unknown breed. (Dark Red: High positive correlation, Dark Blue: High negative correlation)

Table 2. Correlation between EIGENSTRAT and BPPCA principal components

EIGENSTRAT component	BPPCA component 1	2	3	4	5
1	-0.9922	-0.0342	-0.0062	0.0194	0.0537
2	-0.0299	0.7297	0.6536	0.1353	-0.0483
3	0.0159	-0.6636	0.7352	0.0157	-0.0345
4	0.0130	-0.0895	-0.1086	0.9677	-0.1509
5	-0.0509	-0.0023	-0.0406	-0.1534	-0.9683

Conclusions. BPPCA can be shown to successfully distinguish between different breeds and identify the breed of unknown animals but we did not find substantial differences to EIGENSTRAT for either property. Currently BPPCA is much slower to implement and the challenge will be to determine if the method has advantages in populations with different sub-structure than the example given. In the future, the fitting of principal components from EIGENSTRAT and BPPCA as random effects in a BLUP model can be compared for their efficacy in the prediction of breeding values with respect to accuracy and bias.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the generous support of Beef + Lamb New Zealand Genetics, the Ministry of Business, Innovation and Employment and NZ Sheep breeders.

REFERENCES

Daetwyler H.D., Kemper K.E., van der Werf J.H.J., and Hayes, B.J. (2012) *J. Anim. Sci.* **90**:3375.
Klami A. (2014) *J. Mach. Learn. Res.* **39**: 112.
Kuehn L.A., Lewis, R.M. and Notter, D.R. (2007) *Genet. Sel. Evol.* **39**:225
Patterson N., Price A.L. and Reich D. (2006) *Plos. Genet.* **2(12)**: 2074.
Polson N.G., Scott J.G., and Windle J. (2013) *J. Am. Stat. Assoc.* **108**: 1339.
VanRaden P.M. (2008) *J. Dairy Sci.* **91**: 4414.