Use of Principal Component and Factor Analysis to reduce the number of independent variables in the prediction of Genomic Breeding Values

Nicolò Pietro Paolo Macciotta, Giustino Gaspa

Dipartimento di Scienze Zootecniche, Università di Sassari, Italy

Corresponding author: Nicolò P.P. Macciotta. Dipartimento di Scienze Zootecniche, Università di Sassari. Via De Nicola 9, 07100 Sassari, Italy - Tel. +39 079 229298 – Fax: +39 079 229302 – Email: macciott@uniss.it

ABSTRACT - On a simulated population of 2,500 individuals, Principal Component Analysis and Factor Analysis were used to reduce the number of independent variables for the prediction of GEBVs. A genome of 100 cM with 300 bialleic SNPs and 20 multiallelic QTLs was considered. Two heritabilities (0.2 and 0.5) were tested. Multivariate reduction methods performed better than the traditional BLUP with all the SNPs, either on generations with phenotypes available or on those without phenotypes, especially in the low heritability scenario (about 0.70 *vs.* 0.45 in generations without phenotypes). The use of multivariate reduction techniques on the considered data set resulted in a simplification of calculations (reduction of about 90% of predictors) and in an improvement of GEBV accuracies.

Key words: Principal component analysis, Factor analysis, Genome wide selection.

Introduction - In recent years, the use of marker assisted selection programs in livestock has been constrained by poor knowledge on causal mutations affecting the expression of traits of economic interest (Dekkers, 2004). Dense SNP maps allowed the prediction of Genomic Breeding Values (GEBVs) based on the estimation of SNP genotype effects on the considered trait (Meuwissen *et al.*, 2001). Possible advantages of Genome Wide Selection (GWS) over the conventional selection are the reduction of the generation interval, the increase of accuracy, particularly for females (Schaeffer, 2006) and the reduction of costs (Konig *et al.*, 2009). In the basic idea of GWS, marker effects are estimated in a training population where both phenotypes and SNP genotypes are measured. Estimates are then used to calculate GEBVs in a prediction population (i.e. juvenile animals) where only SNP genotypes are available. One major issue in GEBV estimation is represented by the large number of predictors (for example 50K SNPs for cattle) and the relatively small number of records available. Several approaches have been proposed to select relevant markers. Among these, multivariate methods (Woolaston *et al.*, 2007) are of particular interest. In the present paper, two multivariate dimension reduction techniques, Principal Component Analysis and Factor Analysis were used to select a reduced number of independent variables for the prediction of GEBVs.

Material and methods - A population of 2,500 individuals, belonging to five overlapping generations was simulated. A one chromosome genome (2n=2) with length set to 100 cM was created. A total of 20 multiallelic (n. alleles≤5) QTL and 301 bialleic SNP markers were generated. Two heritabilities were considered: 0.2 and 0.5. Initial frequencies of SNP and QTL alleles were set at 0.5 and 1, respectively. Founder population consisted of 50 males and 50 females. One thousand random mating generations were performed, with mutation rates of 25*10-4 and 25*10-5 for SNPs and QTLs, respectively. Starting from generation 1,001, the population was expanded and phenotypes were created adding random noise to true breeding values (TBV). Individuals of generations 1,001 and 1,002 were used as training data (TRAIN) whereas those of generations 1,003-1,005 were considered as prediction gen-

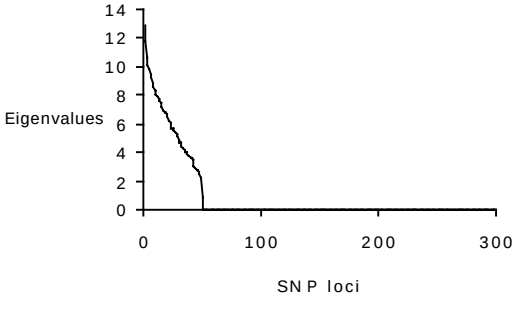
erations (PRED). A SNP data matrix M with m rows (m=2,500, the number of individuals of the whole data set) and n columns (n=301, the number of SNP markers) was created. Each element corresponded to the genotype of a specific marker. Principal Component and Factor Analysis were carried out on M. The number of retained principal components and factors retained for further analysis was based on of the sum of the eigenvalues of the correlation matrix of M, corresponding to the amount of the total variance explained. Single marker effects were estimated in the TRAIN data using three different predictors: all the 301 markers (SNP301), the Principal Component (PC) or the Factor (FACT) scores. The estimation was carried out with the following mixed linear model:

$$y = 1\mu + Zu + e$$

where y is the vector of phenotypes, μ is the overall mean, Z is the incidence matrix of random effects (SNP genotypes, PC, or FAC scores), u is the vector of solutions for random effects, e is the random residual. The (co)variance matrix G of the random effect was assumed to be diagonal $I\sigma^2/n$, where σ^2 is the genetic variance (assumed known, equal to 20 or 50 in the two scenarios), and n is the number of SNPs, PC, or FACT used. Such an assumption is a strong simplification when SNPs are directly modeled, meaning absence of interaction between different loci, whereas it is correct when PC or FACT are used, being their scores orthogonal. Effects estimated in the TRAIN generations ($\hat{u}$) were then used to predict GEBVs both for TRAIN and PRED individuals as

$$GEBV = \mu + \sum z'\hat{u}$$

Figure 1. Pattern of eigenvalues of the correlation matrix of SNP genotypes.



Accuracies of prediction were evaluated as correlations between TBV and GEBVs. Each scenario was replicated 30 times.

Results and conclusions - The pattern of eigenvalues of the correlation matrix of SNP genotypes (Figure 1) showed a sudden drop after the extraction of about 50 PC. Assuming 0.80 as a reasonable threshold for the explained variance, a total of 32 PC and FAC were retained for further analyses.

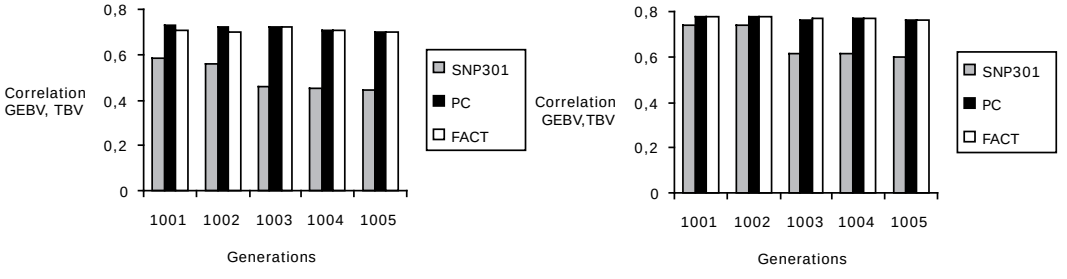
Higher accuracies of GEBVs estimation in the TRAIN generations (Table 1) were obtained for the scenario with high heritability, in agreement with previous results reported in literature (Kolbedhari *et al.*, 2007).

Table 1. Correlations between Genomic and True Breeding Values in the TRAIN and PRED generations with the different methods and for two levels of heritability.

	TRAIN				PRED			
	h^2=0.2		h^2=0.5		h^2=0.2		h^2=0.5	
Method	Mean	s.d.	Mean	s.d.	Mean	s.d.	Mean	s.d.
SNP301	0.57	0.07	0.74	0.06	0.45	0.06	0.61	0.07
PC	0.72	0.05	0.78	0.05	0.71	0.05	0.76	0.05
FACT	0.70	0.09	0.78	0.05	0.69	0.11	0.76	0.05

In spite of the large reduction in the number of predictors used (about 90%), both PC and FACT gave better results than those obtained using all SNPs, especially with the low heritable trait. This advantage is maintained in both TRAIN generations (Figure 2a). As expected, a drop in GEBV accuracy is observed in the PRED generations (Table 1), but only for the SNP301 method. Accuracies obtained with PC and FACT are comparable with those reported for BLUP estimation methodology (Meuwissen *et al.*, 2001; Van Raden, 2008) and tend to remain constant in the three prediction generations (Figure 2b).

Figure 2. Correlations between Genomic (GEBV) and True (TBV) Breeding Values in the different generations in the scenario with h^2=0.2 (a) and h^2=0.5 (b).



In the simulated data set analysed, the use of principal component and factor analysis to extract latent variables from SNP genotypes had a relevant impact on GEBV calculations, with a reduction of about 90% of the predictor variables. In spite of such a reduction, accuracies of GEBVs were always higher in comparison with the model that used all markers available, especially for the scenario with h^2=0.2. Moreover, no differences in accuracies were observed between TRAIN and PRED. A possible interpretation could be found in the ability of multivariate methods, that were carried out on SNP data of both TRAIN and PRED generation simultaneously, to extract latent variables that are able to summarize the genetic relationships (i.e., the resemblance in the specific combinations of SNPs) between individuals of different generations. These results, if confirmed in larger genomes and actual data, may be of great interest for routine use of GWS.

Authors wish to acknowledge Dr. Ezequiel Nicolazzi for his contribution.
Research funded by the Italian Ministry of Agricultural and Forestry Policies (grant SELMOL).

REFERENCES - Dekkers, J.C.M., 2004. Commercial application of marker- and gene-assisted selection in livestock: strategies and lessons. **J. Anim. Sci. 82(E. Suppl.):E313-E328. Kolbehdari,** D., Schaeffer, L.R., Robinson, J.A., **2007. Estimation of genome-wide haplotype effects in half sibs designs. J. Anim. Breed. Genet.** 124:356-361. **Konig,** S., Simianer, H., Willam, A., 2009. Economic evaluation of genomic breeding programs. J. Dairy Sci. 92:382-391. **Meuwissen,** T.H.E., Hayes, B.J., Goddard, M.E., 2001. Prediction of total genetic values using genome-wide dense marker maps. Genetics. 157:1819-1829. **Schaeffer,** L.R., 2006. Strategy for applying genome-wide selection in dairy cattle. J. Anim. Breed. Genet. 123:218-223. **Woolaston,** A.F., Tier, B., Murison, R.D., 2007. Proc. XI QTL-MAS Workshop, Tolouse, 80. **Van Raden,** P.M., 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91:4414-4423.