Application Of PLS And Sparse PLS Regression In Genomic Selection

C. Colombani[*], A. Legarra[*], P. Croiseau[†], F. Guillaume[‡], S. Fritz[§], V. Ducrocq[†] and C. Robert-Granié[*]

Introduction

Genomic selection is based on the estimation of genomic estimated breeding values (GEBV). In this way, a prediction equation based on the estimation of a large number of DNA marker effects, such as SNP (Single Nucleotide Polymorphism) markers and a limited number of genotyped animals must be established. The most significant difficulty is to find performing methods when the number of explanatory variables is much higher than the number of individuals. In this study, we compare two such approaches. The first one is the Partial Least Squares (PLS) regression, which reduces the number of variables in the final model, by projecting independent variables onto latent structures. The second method is the Sparse PLS regression which derives from PLS regression: a step of penalization is added to PLS in order to select the most important variables.

Material and methods

PLS: The PLS regression (Wold 1966) is a data analysis method, built to deal with the p>>n problem (i.e. the number of independent variables p is larger than the number of observations n). When facing collinear matrices, the stability of PLS is clearly advantageous over multiple linear regression, ridge regression or other regression techniques. Its goal is to predict Y (n x 1 matrix, dependent variable) from X (n x p matrix, independent variables or predictors) by extracting from X a set of orthogonal factors (latent variables) which maximizes the covariance between X and Y. These latent components $(\xi_1, ..., \xi_H)$ and associated loadings $(u_1, ..., u_H)$ are estimated in order to solve the following optimization problem:

$$\max_{\|u_h\|=1} \operatorname{cov}(\xi_h, Y_{h-1})$$

where Y_{h-1} is the residual matrix in the regression of Y on ξ_1 ,…, ξ_{h-1} for each dimension h=1,…,H. The main idea is to perform successive regressions by projections onto latent structures to highlight hidden or latent underlying biological effects.

Sparse PLS: The Sparse PLS regression, developed by Lê Cao *et al.* (2008) in an integrating "omics" data context, aims at combining variable selection and modelling in a one-step procedure. This sparse PLS is based on a Lasso penalization (Tibshirani 1996) and is obtained by penalizing a sparse Singular Value Decomposition (SVD), as proposed by Shen and Huang

* INRA, UR 631, Station d'Amélioration Génétique des Animaux, 31326 Castanet-Tolosan, France
† INRA, UMR1313, Génétique Animale et Biologie Intégrative, 78352 Jouy en Josas, France
‡ Institut de l'Elevage, 149 rue de Bercy, 75595 Paris, France
§ UNCEIA, 149 rue de Bercy, 75595 Paris, France
This work has been financed by ANR project AMASGEN.

(2008), by using a PLS variant with SVD (Lorber *et al.* 1987). In this decomposition, the singular vectors correspond to the PLS loading vectors.

In the PLS and the Sparse PLS regression, the adequate number of dimensions H has to be determined. The number of dimensions leading to the highest correlation between estimated values and observed values is kept. There is an additional parameter in the Sparse PLS regression: the number of selected variables in each dimension.

Cross-validation: In Sparse PLS regression, the number of selected variables in each dimension of the model has to be fixed. This choice can be made by examining the Root Mean Squared Error of Prediction (RMSEP) by K-fold cross-validation in the learning data set (Mevik and Wehrens 2007):

$$RMSEP = \sqrt{\frac{1}{K}\sum_{k=1}^{K}(\hat{y}_k - y_k)}$$

The adequate number of selected variables is the one minimizing RMSEP.

Variable Importance in the Projection (VIP): To give a better interpretation of the model, coefficients which represent the explanatory power of X_j on Y must be defined. The VIP coefficients measure the contribution of X_j in the construction of Y through latent variables t_h and is defined by:

$$VIP_{Hj} = \sqrt{\frac{p}{\sum_{h=1}^{H} cor^2(Y,t_h)} \sum_{h=1}^{H} cor^2(Y,t_h) w_{hj}^2} \quad \text{with} \quad \sum_{j=1}^{p} VIP_{Hj}^2 = p \cdot$$

The contribution of X_j in the construction of t_h is measured by its weight w_{hj}^2 .The VIP coefficients allow to classify the variables X_j according to their explanatory power of Y.

PLS and Sparse PLS were performed with the R package named "integrOmics" (http://cran.r-project.org/web/packages/integrOmics/index.html) (Lê Cao *et al.* 2009).

Data: We studied Montbéliarde bulls genotyped with the Illumina Bovine SNP50 BeadChip (described in Croiseau *et al*. this congress). The population was made up of 678 Montbéliarde bulls and was split into a learning data set and a validation data set consisting of the youngest bulls. First, the prediction equation was estimated with the learning data set, composed of 451 Montbéliarde bulls, genotyped and phenotyped. 48660 polymorphic SNP were used as independent variables. Five traits were considered independently as response variable, as response variable: milk yield, fat yield, fat percent, protein yield, and protein percent. The bulls' phenotypes were DYD (Daughter Yield Deviation) (Mrode and Swanson 2004) that is the corrected average of their daughters' performances, weighted by their Effective Daughter Contribution (EDC). Then, phenotypes were predicted on the validation data set, composed by 227 Montbéliarde bulls. In order to test the accuracy of the model, the EDC-weighted correlation between estimated DYD (DYD_{est}) and observed DYD (DYD_{obs}) were computed.

Results and discussion

PLS regression and several Sparse PLS regressions according to the part of SNP to select by dimension were applied to the learning data set. Then, models were compared on the validation

data set. PLS and Sparse PLS were tested until dimension 100 but the correlations obtained after about thirty dimensions no longer increased and stayed constant. So, the number of dimensions which led to the highest correlation was kept (Table 1). For milk yield, in Sparse PLS regression, the best correlation (ρ=0.343) was obtained making a selection of 3% of the 48660 SNP set in 3 dimensions, that is to say making a selection of 4375 SNP. In PLS regression, for a model with 12 dimensions, the correlation was lower (ρ=0.312).

The Root Mean Squared Error of Prediction was calculated for models with 3 dimensions in order to fix the number of selected SNP in each dimension. This number of dimensions led to the best correlations whatever the Sparse PLS tested. The model which gave the lowest error of prediction kept 2% of the 48660 SNP set in each dimension (2919 SNP were selected) for a correlation equal to 0.338, higher than the correlation obtained with PLS. This model (good correlation, minimum RMSEP) was retained in the following analyses.

Table 1: Correlations (ρ) between DYD_{obs} and DYD_{est} and RMSEP as a function of the percentage of SNP kept in each dimension of the model, for milk yield

	Sparse PLS								PLS
% SNP	0.5	1	2	3	4	5	10	30	100
ρ	0.317	0.324	0.338	**0.343**	0.342	0.341	0.332	0.307	0.312
RMSEP	0.250	0.269	**0.246**	0.255	0.265	0.258	0.296	0.258	0.267

The Sparse PLS regression gave similar or better results than PLS regression with a good selection of explanatory variables. This selection enabled to have a better interpretation of VIP coefficients (Figure 1).

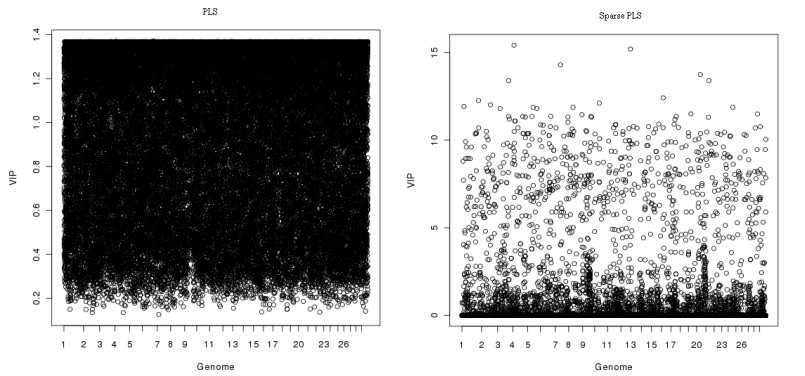

Figure 1: VIP coefficients as a function of SNP position along the genome.

All the 48660 VIP coefficients are drawn on Figure 1, both in PLS and Sparse PLS. PLS is compared to the Sparse PLS which leads to a satisfactory correlation, a minimum RMSEP and a good selection of SNP (Sparse PLS 2% of selected SNP in 3 dimensions for milk yield). In particular, for the milk yield, the VIP coefficients on PLS do not permit to rank the variables X_j according to their explanatory power of Y while the VIP coefficients on Sparse PLS show SNP with large effects on the 4^{th} ,12^{th} ,7^{th} and 20^{th} chromosome.

Unfortunately, most of these results can not be generalized to the other traits. Depending on the considered trait, Sparse PLS gave better or worse results than PLS (Table 2). But the advantage of Sparse PLS is to select the most important SNP in a one-step procedure. This makes models more interpretable. The estimation of the RMSEP to fix the number of SNP selected in each dimension seems to be an efficient approach, for the most of studied traits.

Table 2: Correlations between DYD_{obs} and DYD_{est} by pedigree-based BLUP (BLUP), PLS, and Sparse PLS (sPLS)

	MilkYield	FatYield	Protein Yield	Fat %	Protein %
BLUP	0.273	0.355	0.276	**0.372**	0.214
PLS	0.312	**0.450**	0.376	0.363	**0.391**
sPLS	**0.338**	0.431	**0.382**	0.356	0.367

Pedigree-based BLUP was applied to the data for comparison with PLS and Sparse PLS methods (Table 2). For Fat percent, pedigree-based BLUP led to higher correlations but not far from correlations reached by PLS or Sparse PLS, while, for the other traits, PLS or Sparse PLS gave much better results.

Conclusion

The Sparse PLS regression is not always more efficient than the PLS regression. It depends on the trait. In the sparse version, variable selection is included with a Lasso penalization in order to bring more information about the underlying biological features. This approach can help to identify relevant variables (SNP) linked to phenotype.
Similar results (not shown) were obtained with another breed (Holstein breed with 1216 bulls in the learning data set and 550 bulls in the validation data set). According to the study presented by Croiseau *et al*. (this congress), a SNP pre-selection based on a QTL detection was tested with these two methods. Results were improved whatever the method and whatever the trait.
Moreover, the PLS and Sparse PLS algorithm are fast to compute, even when the size of the reference population is large.

References

Croiseau, P., Colombani, C., Legarra, A. *et al*., *Proc. 9th WCGALP*, 2010.

Lê Cao, KA., Rossouw, D., Robert-Granié, C. *et al*., *Stat Appl Genet Mo B*, 7: 32, 2008.

Lê Cao, KA., González, I. and Déjean, S., *Bioinformatics*, 25:2855-2856, 2009.

Lorber, A., Wangen, L. and Kowalski, B., *J Chemometr*, 1:19-31, 1987.

Mevik B.H. and Wehrens, R., *J Stat Softw*, 18 (2), 2007.

Mrode, R. A. and Swanson, G. j. T., *Livest Prod Sci*, 86:253-260, 2004.

Shen, H and Huang, J.Z., *J Multivariate Ana*l, 99:1015-1034, 2008.

Tibshirani, R., *J R Stat Soc, Series B.* 58(1):267-288, 1996.

Wold, H., *Academic Press*, 1966.