Journal of
Animal Breeding and Genetics

J. Anim. Breed. Genet. ISSN 0931-2668

ORIGINAL ARTICLE

Dimension reduction and variable selection for genomic selection: application to predicting milk yield in Holsteins

N. Long[1], D. Gianola[1,2,3], G.J.M. Rosa[1,3] & K.A. Weigel[2]

1 Department of Animal Sciences, University of Wisconsin, Madison, WI, USA
2 Department of Dairy Science, University of Wisconsin, Madison, WI, USA
3 Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison, WI, USA

Keywords
dimension reduction; genomic selection; variable selection.

Correspondence
N. Long, Department of Animal Sciences, University of Wisconsin, Madison, WI 53706, USA. Tel: +1 (608) 263-3499; Fax: +1 (608) 263-5157; E-mail:nlong@wisc.edu

Received: 3 September 2010; accepted: 5 January 2011

Summary

Genome-assisted prediction of genetic merit of individuals for a quantitative trait requires building statistical models that can handle data sets consisting of a massive number of markers and many fewer observations. Numerous regression models have been proposed in which marker effects are treated as random variables. Alternatively, multivariate dimension reduction techniques [such as principal component regression (PCR) and partial least-squares regression (PLS)] model a small number of latent components which are linear combinations of original variables, thereby reducing dimensionality. Further, marker selection has drawn increasing attention in genomic selection. This study evaluated two dimension reduction methods, namely, supervised PCR and sparse PLS, for predicting genomic breeding values (BV) of dairy bulls for milk yield using single-nucleotide polymorphisms (SNPs). These two methods perform variable selection in addition to reducing dimensionality. Supervised PCR preselects SNPs based on the strength of association of each SNP with the phenotype. Sparse PLS promotes sparsity by imposing some penalty on the coefficients of linear combinations of original SNP variables. Two types of supervised PCR (I and II) were examined. Method I was based on single-SNP analyses, whereas method II was based on multiple-SNP analyses. Supervised PCR II was clearly better than supervised PCR I in predictive ability when evaluated on SNP subsets of various sizes, and sparse PLS was in between. Supervised PCR II and sparse PLS attained similar predictive correlations when the size of the SNP subset was below 1000. Supervised PCR II with 300 and 500 SNPs achieved correlations of 0.54 and 0.59, respectively, corresponding to 80 and 87% of the correlation (0.68) obtained with all 32 518 SNPs in a PCR model. The predictive correlation of supervised PCR II reached a plateau of 0.68 when the number of SNPs increased to 3500. Our results demonstrate the potential of combining dimension reduction and variable selection for accurate and cost-effective prediction of genomic BV.

Introduction

In regression models for quantitative traits using a large number of genetic markers [e.g. single nucleotide polymorphisms (SNPs)] as predictor variables, multicollinearity occurs because markers are intercorrelated. This is attributable to linkage disequilibrium, i.e. the nonrandom association of alleles at two or more loci. A well-known consequence of multicollinearity in least-squares regression is

  doi:10.1111/j.1439-0388.2011.00917.x

unstable estimates. It inflates variances of estimates of regression coefficients (e.g. marker effects) owing to near-singularity of the incidence matrix. Various methods have been proposed to handle multicollinearity, such as ridge regression, principal component regression (PCR)(Massy 1965) and partial least squares (PLS) regression (Wold 1985). PCR and PLS are well-documented methods in chemometrics (Frank & Friedman 1993), and have also found applications in bioinformatics (e.g. Boulesteix & Strimmer 2007), neuroimage analysis (e.g. McIntosh *et al.* 1996) and recently in genome-assisted prediction of breeding values (BV) (Solberg *et al.* 2009;Macciotta *et al.* 2010). Besides multicollinearity, another challenging problem is that data typically contain many more variables (*p*, e.g. genes or markers) than observations (*n*, e.g. individuals), disabling the use of least-squares methodology. For instance, it is common to have genotypes for hundreds of thousands of SNP markers on only hundreds of animals in genomic selection.

To reduce model dimension and to overcome the multicollinearity problem, PCR and PLS transform the large number of original variables into a relatively small number of orthogonal latent components and then regress the response variable on those latent components. When applied to genome-wide analysis with n<<p, a small number (relative to *n*) of latent components is usually sufficient for the model to achieve a competitive predictive performance. Therefore, the computational burden can be reduced with PCR and PLS, compared to methods that estimate all marker effects by treating them as random variables, such as Bayes A (Meuwissen *et al.* 2001) or the Bayesian least absolute shrinkage and selection operator (Lasso) (Park & Casella 2008;de los Campos *et al.* 2009).

In practical applications of genome-enabled selection, it is often desired to genotype a subset of whole-genome SNP markers to reduce genotyping cost, especially for initial screening of animals. For commonly used SNP-based regression models, it has been shown (Long *et al.* 2007;Habier *et al.* 2009;Usai *et al.* 2009;Weigel *et al.* 2009) that, with a carefully chosen SNP subset, one can achieve a reasonably high reliability of predicted BV. For example, Weigel *et al.* (2009) found that using the 300 SNPs that had the largest estimated effects from a full Bayesian Lasso regression with all 32 518 SNPs, the prediction reliability was half of that obtained with all SNPs. Usai *et al.* (2009) found in simulations that the accuracy of the estimated BV reached a maximum when 169 markers were used, which were selected from 6000 SNPs using a Lasso least angle regression algorithm.

Although PCR and PLS may efficiently reduce model dimension, they do not offer the possibility of selecting important variables, because each latent component is a linear combination of all original variables. It is desirable to have both model dimension and number of variables reduced. A simple way is to ignore variables with small-magnitude coefficients in the latent components, but this often leads to misinterpretation of the original variables (Cadima & Jolliffe 1995). In the context of gene expression analysis, Bair *et al.* (2006) proposed the notion of supervised principal component analysis (PCA) for PCR, i.e. preselecting a subset of genes prior to forming latent components, with the selection based on the strength of association of each gene with the outcome. Because of this 'supervised' (i.e. guided by the outcome) preselection procedure, the resulting latent components had better prediction accuracy for the outcome than that attained with all genes. For PLS, Chun & Keleş (2010) developed a novel methodology (the 'sparse PLS') to perform dimension reduction and variable selection simultaneously. Briefly, sparse PLS promotes sparsity by imposing some penalty on the coefficients of the linear combinations of original variables. Usefulness of these methods has been demonstrated in applications to microarray data (Chen *et al.* 2008;Chun & Keleş 2009).

For genome-assisted BV prediction, the problem of SNP subset selection has not been investigated in the framework of latent variable methods for dimension reduction. Therefore, the aim of this study was to exploit advances in methods that combine dimension reduction and variable selection for analysing high-dimensional genomic data as presented by modern SNP chips. Two variants of supervised PCR and the sparse PLS regression model of Chun & Keleş (2010) were examined using data from dairy cattle. The objective was to predict sires' predicted transmitting ability (PTA, half of the BV) for milk yield derived from progeny testing, using SNP markers. This study evaluated predictive performance delivered by each of the three models fitted to SNP subsets of various sizes in order to judge the effectiveness of using SNP subsets for genomic selection.

Materials and methods

Data

High-density SNP genotypes and PTAs for milk yield (derived from performance records of female offspring

via progeny testing) of Holstein bulls were obtained from the Bovine Functional Genomics Laboratory and Animal Improvement Programs Laboratory, respectively, at the USDA-ARS Beltsville Agricultural Research Center (Beltsville, MD). After editing, the final data used in the analysis consisted of 32 518 SNPs and 4703 sires; 3305 sires born from 1952 to 98 were used as training set to build models, and the remaining 1398 sires born from 1999 to 2002 were designated as the testing set for model evaluation. August 2003 progeny-testing PTAs and April 2008 progeny-testing PTAs were available for bulls in the training and testing sets, respectively. Weigel *et al.* (2009) used the same genotypic data with a different trait (lifetime net merit), and a detailed description of data preprocessing can be found therein.

Methods

Supervised PCR and sparse PLS models were fitted to the data of the 3305 training bulls using different numbers of selected SNPs. Comparison between methods was based on their predictive ability, assessed by Pearson's correlations between realized progeny-testing PTAs and PTAs predicted by supervised PCR or sparse PLS on the 1398 testing bulls ('predictive correlation' is used hereinafter).

Ordinary PCR and PLS using all 32 518 SNPs were also fitted, such that the potential of the SNP selection approaches being examined could be evaluated. In this section, ordinary PCR and PLS are reviewed first. Then, essentials of the supervised PCR and sparse PLS are given. Lastly, Bayesian Lasso regression is described briefly, as it was used to estimate regression coefficients in all PCR models in this study.

Following conventions in the literature, in what follows, the latent components in PCR will be called 'principal components', or PCs; for PLS, the term 'latent components' will be used. The strategy for finding PCs/latent components differs between PCR and PLS. PCR considers maximizing the variances of the PCs, whereas PLS seeks latent components that have large variances as well as high correlations with the response variable.

In the context of genetic marker data, define an $n \times p$ matrix, $\mathbf{X}$, which consists of genotypes of p SNP markers for each of the n individuals. Each element of $\mathbf{X}$ takes the value 0, 1 or 2, depending on the number of allele copies observed. The responses (realized progeny-testing PTAs) are stored in an $n \times 1$ column vector $\mathbf{y}$. Let r denote the column rank of $\mathbf{X}$.

Derivation of principal components in PCR

The first step of PCR is to apply PCA to $\mathbf{X}$ (assuming that it has been column-centred in advance) to extract a number (K,$K < r$) of PCs, each of them being a linear combination of the columns of $\mathbf{X}$. The matrix $\mathbf{T}(n\times K)$, which contains these PCs, is formed as $\mathbf{T} = \mathbf{XP}$, where $\mathbf{P}(p \times K)$ is the loading matrix and can be obtained from either the spectral decomposition of $\mathbf{X}'\mathbf{X}/(n-1)$ (sample covariance of $\mathbf{X}$) or singular value decomposition of $\mathbf{X}$. The former approach gives $\mathbf{X}'\mathbf{X}/(n-1) = \mathbf{P}_r\Lambda\mathbf{P}_r'$, where columns of $\mathbf{P}_r$ consist of orthonormal eigenvectors, and $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1,\ldots,\lambda_r)$ contains the corresponding eigenvalues. $\mathbf{P}$ is then obtained by taking the first K columns of $\mathbf{P}_r$. From an optimization perspective, each column of $\mathbf{P}$, $\mathbf{p}_k$ ($1 \le k \le K$), satisfies

$$\mathbf{p}_k = \mathrm{argmax}_{\mathbf{p}}\mathrm{var}(\mathbf{Xp}), \quad \text{subject to } \mathbf{p}'\mathbf{p} = 1, \{\mathbf{p}'\mathbf{p}_l = 0\}_{l=1}^{k-1}, \tag{1}$$

where $\mathrm{var}(\mathbf{Xp}_k)$ denotes the sample variance of the kth PC ($k = 1,\ldots,K$). From the criterion and the two constraints, it follows that the PCs have successive maximal variances and are mutually orthogonal. Specifically, the variance of the kth PC is equal to the kth eigenvalue, because $\mathrm{var}(\mathbf{Xp}_k) = \mathbf{p}_k'\mathbf{X}'\mathbf{X}\mathbf{p}_k/(n-1) = \mathbf{p}_k'(\mathbf{P}_r\Lambda\mathbf{P}_r')\mathbf{p}_k = \lambda_k$.

Derivation of latent components in PLS

PLS is based on the assumptions that $\mathbf{X} = \mathbf{TP}'+\mathbf{E}$ and $\mathbf{y} = \mathbf{Tm} + \mathbf{f}$, assuming both $\mathbf{X}$ and $\mathbf{y}$ are column-centered. Here, $\mathbf{T}(n \times K)$ is the latent component matrix; $\mathbf{P}(p \times K)$ and $\mathbf{m}(K \times 1)$ are often denoted as 'X-loadings' and 'Y-loadings', respectively; $\mathbf{E}(n \times p)$ and $\mathbf{f}(n \times 1)$ are random errors. This decomposition indicates that the relationship between $\mathbf{X}$ and $\mathbf{y}$ is conveyed through $\mathbf{T}$. Frequently used PLS approaches include PLS1 (Manne 1987;Helland 1988) and SIMPLS (de Jong 1993). For a univariate response variable (as in the current study), the two are equivalent (de Jong 1993), with the goal of finding a weight matrix $\mathbf{W}(p \times K)$ such that its k-th column $\mathbf{w}_k(k = 1,2,\ldots,K)$ satisfies

$$\begin{aligned}\mathbf{w}_k = \mathrm{argmax}_{\mathbf{w}}\mathrm{corr}^2(\mathbf{y},\mathbf{Xw})\mathrm{var}(\mathbf{Xw}),\\ \text{subject to } \mathbf{w}'\mathbf{w} = 1, \{\mathbf{w}'\mathbf{X}'\mathbf{X}\mathbf{w}_l = 0\}_{l=1}^{k-1}.\end{aligned} \tag{2}$$

The solution, $\widehat{\mathbf{W}}$, is then used to construct $\mathbf{T}(\mathbf{T} = \mathbf{X}\widehat{\mathbf{W}})$, which has mutually orthogonal columns as indicated by the second constraint. The loading vector $\mathbf{m}$ is estimated via solving the

regression problem $\mathbf{y} = \mathbf{Tm}+\mathbf{f}$ via least-squares, giving $\hat{\mathbf{m}} = (\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}'\mathbf{y}$. This leads to the final regression model: $\hat{\mathbf{y}} = \mathbf{T}\hat{\mathbf{m}} = \mathbf{X}\widehat{\mathbf{W}}\hat{\mathbf{m}} = \mathbf{X}\hat{\beta}^{PLS}$. This means that $(\widehat{\mathbf{W}}\hat{\mathbf{m}})$ can be regarded as a PLS estimate of β, the regression coefficients with respect to the original $\mathbf{X}$ variables.

When solving the optimization problem (2), it is useful to obtain explicit an expression of the objective function. Noting that $\mathbf{y}$ and hence var($\mathbf{y}$) are constants given the data, the original maximization problem, $\text{argmax}_{\mathbf{w}}\text{corr}^2(\mathbf{y},\mathbf{Xw})\text{var}(\mathbf{Xw})$, is equivalent to

$$\text{argmax}_{\mathbf{w}}\text{corr}^2(\mathbf{y},\mathbf{Xw})\text{var}(\mathbf{Xw})\text{var}(\mathbf{y}) = \text{cov}^2(\mathbf{Xw},\mathbf{y}),$$

and, further, to $\text{argmax}_{\mathbf{w}}\mathbf{w}'\mathbf{X}'\mathbf{y}\mathbf{y}'\mathbf{X}\mathbf{w}$. This explicit form of objective function will be used later for describing sparse PLS. In a nutshell, in PLS the latent components are not redundant and have maximal covariance with the response. As such, they are expected to have high predictive power.

Prediction.

The main application of PCR and PLS is to use the final regression model for prediction purposes. The second step of PCR (after the first PCA step) consists of replacing the original $\mathbf{X}$ by $\mathbf{T}$ and estimating coefficients associated with it. The model is

$$\mathbf{y} = \mathbf{1}_n\mu + \mathbf{Tg} + \mathbf{e} = \mathbf{1}_n\mu + \mathbf{X}(\mathbf{Pg}) + \mathbf{e}, \qquad (3)$$

where μ is the overall mean, $\mathbf{g}(K \times 1)$ is a coefficient vector to be estimated and $\mathbf{e}$ is an error vector. Here, $(\mathbf{Pg})$ can be regarded as $\boldsymbol{\beta}^{PCR}$, the regression coefficients with respect to the original $\mathbf{X}$ variables. Given a new observation $\mathbf{x}^*(p \times 1)$, its predicted response is $\hat{\mathbf{y}}^* = \hat{\mu} + (\mathbf{x}^* - \bar{\mathbf{x}})'\mathbf{P}\hat{\mathbf{g}}$, where $\bar{\mathbf{x}}$ is a p-dimensional column mean vector averaged over the training samples $\mathbf{X}$ and $\hat{\mu}$ is the average of $\mathbf{y}$ owing to zero column means of $\mathbf{X}$. The coefficients $\mathbf{g}$ in (3) may be estimated by ordinary least-squares or Bayesian shrinkage methods. In our analysis, the Bayesian Lasso was adopted because it produced models with better predictive performance than those from least-squares estimation (data not shown).

Prediction for PLS is given by $\hat{\mathbf{y}}^* = \hat{\mu} + (\mathbf{x}^* - \bar{\mathbf{x}})'\widehat{\mathbf{W}}\hat{\mathbf{m}}$, where $\hat{\mu}$ is the average of $\mathbf{y}$ values; $\widehat{\mathbf{W}}$ and $\hat{\mathbf{m}}$ are readily available after applying a PLS algorithm to the training data.

Supervised PCR

The SNP selection step in the supervised PCR, as in Bair *et al.* (2006) and Chen *et al.* (2008), is based upon association with the phenotype of each SNP. The association could be measured by the *t*-statistic $(\hat{\beta}/s.e.(\hat{\beta}))$ from, for example, single-marker regression. Then, a threshold on the absolute value of the *t*-statistic is chosen so that only SNP markers whose $|t|$ values are above the threshold are retained. If one is interested in obtaining a single subset (without caring about its size), the threshold must be chosen carefully to maximize the best predictive power of the final subset. Bair *et al.* (2006) and Chen *et al.* (2008) proposed methods for threshold selection.

For the current study, the sizes of SNP subsets were predetermined, making the choice of an optimal threshold unnecessary. Hence, our procedure was to use an association measure to rank all SNPs, and a specified number of top-ranked SNPs were selected. Two measures were considered.

1. The first strategy used the aforementioned $|t|$ from least-squares single-marker regression and was called supervised PCR I.

2. The second strategy for ranking SNPs was based on a full PCR model involving all SNPs. Specifically, 3000 PCs were extracted from the $3305 \times 32\,518$ $\mathbf{X}$ (SNP genotype) matrix. The estimated regression coefficients (via the Bayesian Lasso) of these PCs were transformed back to coefficients for the original SNPs. The magnitudes of these SNP coefficients were then used to rank and select SNPs. This procedure was termed supervised PCR II.

Given a subset of selected SNPs, a PCR model with the Bayesian Lasso as the method for estimating regression coefficients of PCs was fitted and used to predict the testing data. SNP subsets of different sizes (300,500,1000,1500,2000,…,8000) were tested for the two supervised PCR methods. For a subset of p SNPs, the number of PCs fitted was increased gradually from a small value to a large value [but no larger than min(3305,p)], and predictions were made accordingly.

Sparse PLS

The sparse PLS developed by Chun & Keleş (2010) aims to produce sparsity on the original variables by imposing an L_1 penalty to the weight vectors of PLS. Given the number (K) of latent components to be extracted, the K weight vectors are calculated sequentially by optimizing some objective functions. The following describes the derivation of the first weight vector. The full procedure is outlined in Appendix I.

The objective function for the first weight vector is

$$\min_{\mathbf{w},\mathbf{c}}\{-\kappa\mathbf{w}'\mathbf{M}\mathbf{w}+(1-\kappa)(\mathbf{c}-\mathbf{w})'\mathbf{M}(\mathbf{c}-\mathbf{w}) + \lambda_1||\mathbf{c}||_1+\lambda_2||\mathbf{c}||_2^2\}, \quad \text{subject to } \mathbf{w}'\mathbf{w}=1, \tag{4}$$

where $||\mathbf{c}||_1=\sum_{i=1}^{p}|c_i|$,$||\mathbf{c}||_2^2=\sum_{i=1}^{p}c_i^2$;$\mathbf{M}=\mathbf{X}'\mathbf{y}\mathbf{y}'\mathbf{X}$; κ ($0<\kappa<1$), λ_1 ($\lambda_1\geq 0$) and λ_2 ($\lambda_2\geq 0$) are three tuning parameters. This formulation imposes the L_1 penalty not on $\mathbf{w}$, but on its surrogate vector $\mathbf{c}$, while keeping $\mathbf{c}$ and $\mathbf{w}$ close to each other. The L_1 penalty (λ_1) promotes sparsity on $\mathbf{c}$ (setting some components to zero), whereas the L_2 penalty (λ_2) overcomes potential singularity of $\mathbf{M}$ when solving for $\mathbf{c}$. As can be seen from equation (4), when $\kappa=1$ the sparse PLS reduces to the ordinary PLS problem, that is, $\text{argmax}_{\mathbf{w}}\mathbf{w}'\mathbf{X}'\mathbf{y}\mathbf{y}'\mathbf{X}\mathbf{w}$ (see ''Derivation of latent components in PLS'').

As pointed out by Chun & Keleş (2010), the motivation of using (4) instead of a simpler criterion such as

$$\max_{\mathbf{w}}(\mathbf{w}'\mathbf{M}\mathbf{w}), \quad \text{subject to } \mathbf{w}'\mathbf{w}=1, ||\mathbf{w}||^1\leq\lambda,$$

is that the solution of $\mathbf{w}$ in the above formulation tends not to be sparse enough. This observation was found when Jolliffe *et al.* (2003) proposed a sparse-loadings technique for PCA (the SCoTLASS).

In (4), the solution $\hat{\mathbf{c}}$ is used as the first estimated sparse PLS weight vector. In the case of univariate $\mathbf{y}$, Chun & Keleş (2010) showed that $\hat{\mathbf{c}}$ depends only on a new tuning parameter η ($0\leq\eta\leq 1$) and takes the form of a soft threshold estimator:

$$\hat{c_i}=\left(|u_i|-\eta\max_{1\leq j\leq p}|u_j|\right)I\left(|u_i|\geq\eta\max_{1\leq j\leq p}|u_j|\right)\text{sign}(u_i). \tag{5}$$

Above, u_i is the ith element of vector $\mathbf{u}$, with $\mathbf{u}=\mathbf{X}'\mathbf{y}/\|\mathbf{X}'\mathbf{y}\|$; $I()$ is an indicator function, taking value of 1 if its argument is true and 0 otherwise. This estimator retains weight components that are larger than some fraction of the maximum component value, leading to a sparse weight vector with zeros indicating that the corresponding SNP variables are discarded. Chun & Keleş (2010) provided an algorithm for implementing sparse PLS, and its univariate version is given in Appendix I. After fitting the sparse PLS model, one can obtain a vector of regression coefficients with respect to the original variables, which can be directly used for prediction. As a result of variable selection, some of the estimated coefficients are exactly zero.

The sparse PLS was implemented using R package spls (http://www.stat.wisc.edu/ chungdon/spls/). The entire procedure requires specification of two tuning parameters, the aforementioned η and the number of latent components desired, K. Different values for these parameters lead to selection of different numbers of SNPs, and it is easier to predefine an upper bound than an exact value for the size of the subset. Therefore, the 17 subset sizes chosen previously for supervised PCR were used as the upper bounds here. Given an upper bound, different combinations of η and K that generated subsets of sizes smaller than this bound were compared, and the one producing the highest cross-validation (CV) correlation was used to build a sparse PLS model for final prediction on the testing data.

Bayesian Lasso for estimating regression coefficients

Consider a linear regression model $\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\mathbf{e}$, where $\mathbf{X}$ is an $n\times p$ incidence matrix, $\boldsymbol{\beta}$ is a $p\times 1$ vector of unknown coefficients and the errors in $\mathbf{e}$ are independent and identically distributed as $N(\mathbf{0},\mathbf{I}\sigma_e^2)$. Bayesian Lasso assigns the same double exponential prior distribution to each element of $\boldsymbol{\beta}$, $\beta_j(j=1,\ldots,p)$. This is equivalent to the following two steps (Park & Casella 2008):

$$p(\beta_j|\tau_j)\sim N(0,\tau_j^2), j=1,\ldots,p$$
$$p(\tau_j^2|\lambda)\sim\text{Exponential}(\lambda), j=1,\ldots,p$$

Usually, λ is assigned a Gamma prior with its hyperparameters (shape and rate) tuned by the user. Details on Gibbs sampling for implementing Bayesian Lasso can be found in Park & Casella (2008) and de los Campos *et al.* (2009). As noted, the Bayesian Lasso is one form of attaining differential shrinkage of coefficients ($\boldsymbol{\beta}$). Relative to a normal prior, the double exponential distribution produces stronger shrinkage of coefficients that are close to zero and less shrinkage of those with large absolute values (de los Campos *et al.* 2009). In our Bayesian analysis, the prior for the error variance, σ_e^2, was a scaled inverted chi-squared distribution with df = 0.002 and scale = 1; for λ, a vague Gamma distribution with shape = 1 and rate = 0.0001 was used as prior distribution. The Markov chain Monte Carlo sampling was run for 50 000 iterations with the first 30 000 as burn-in. The rest of the iterations were thinned at a rate of 20. The posterior mean (after burn-in and thinning) of each parameter was used as its estimate.

Results and discussion

PCR and PLS using all SNPs

Vázquez *et al.* (2010) reported a predictive correlation of 0.69 from fitting a Bayesian Lasso model with

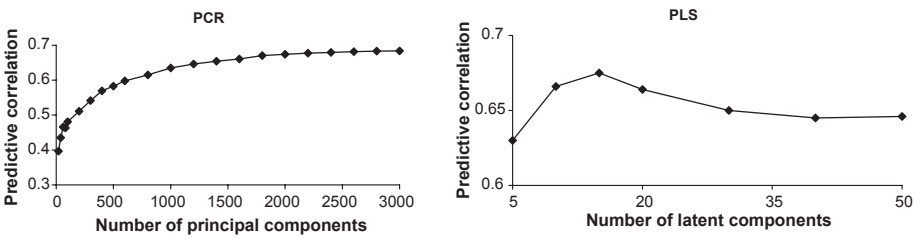

Figure 1 Predictive correlation (correlation between progeny-testing predicted transmitting ability (PTA) and predicted PTA on testing bulls) against the number of principal components/latent components for PCR/PLS when all 32 518 SNPs were used.

all SNPs using the same training and testing data sets as in the present study. Figure 1 shows predictive correlations of PCR and PLS using all 32 518 SNPs. Different numbers of PCs/latent components were fitted. For PCR, the correlation kept increasing with the number of PCs until a plateau was reached somewhere between 2000 and 2500 PCs, where the correlation was 0.68. For PLS, the highest correlation (0.67) was obtained with only 15 latent components. As more latent components were added to the model, the correlation began to decrease. Relative to the 32 518 SNPs, the 2000 PCs and the 15 latent components represented a great reduction in the number of model parameters. Therefore, with PCR/PLS, one is able to reduce model dimension without impairing predictive ability. This conclusion was also supported by Macciotta *et al.* (2010). On the other hand, to achieve a similar predictive correlation, the number of PCs needed for PCR was much larger than the number of latent components needed for PLS, suggesting that the latent components had much greater predictive power than the PCs. This was expected, because construction of PCs is based solely on information regarding predictors whereas the construction of latent components also takes into account the response variable.

Supervised PCR (I and II)

Similar to PCR using all SNPs, predictive correlations from the supervised PCR (I and II) procedure reached a plateau after an increasing phase, as the number of PCs was increased. This is shown in Figure 2, using results from fitting a subset of 3000 SNPs as an example. The plateau was reached at approximately 1000 for supervised PCR I and II; method II was clearly better than method I from 200 to 2800 PCs. For SNP subsets of other sizes, similar patterns were found. Therefore, the reported predictive correlations for supervised PCR I/II (shown later in Figure 4) were their plateau values (i.e. the maximums).

Sparse PLS

As noted, η controls sparsity in the weight vector (5), and a larger value of η results in fewer selected SNPs. On the other hand, the number of selected SNPs increases with the number of latent components K (Figure 6 in Appendix II).

A fivefold CV on the training data (repeated three times to reduce uncertainty) was carried out for sparse PLS with different values of (η,K); η was varied over 0.5,0.55,…,0.9 and K was varied over 1, 3,

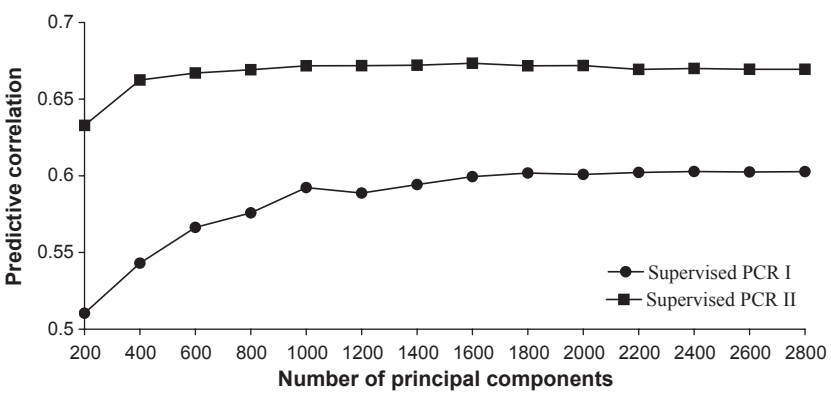

Figure 2 Predictive correlation (correlation between progeny-testing predicted transmitting ability (PTA) and predicted PTA on testing bulls) against the number of principal components for supervised PCR I and II when the number of selected single nucleotide polymorphisms was 3000 for both methods.

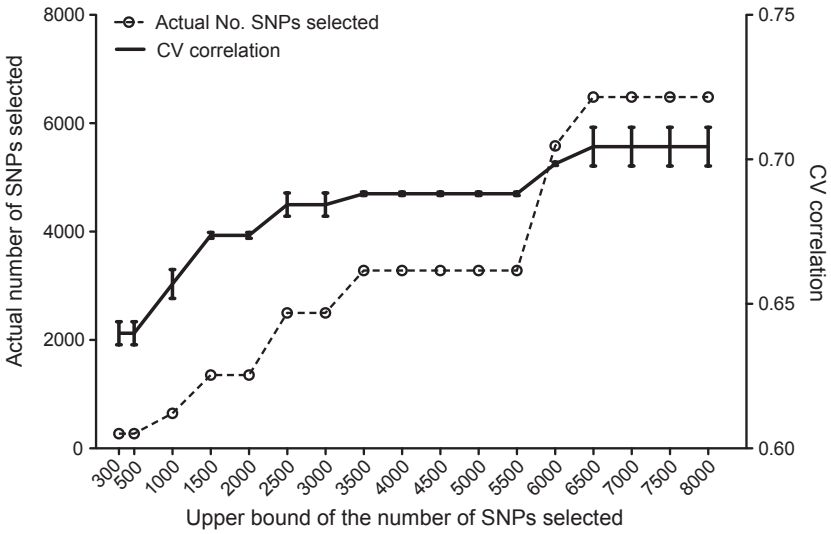

Figure 3 The actual number of single nucleotide polymorphisms (SNPs) selected and corresponding cross-validation (CV) correlation for each upper bound on the number of SNPs selected in sparse partial least squares. The correlation shown (between observed and predicted predicted transmitting abilities) was the average of three replicates of a fivefold CV. Each error bar represents mean ± standard error. The standard errors of some points are too small to be visible.

5, 10, 15, 20, 25, 30, 35, 40, resulting in a total of 90 combinations. Further, the maximal subset size was restricted to 8000 to be consistent with that of supervised PCR. The endpoint was the CV correlation, that is, the correlation between observed and predicted PTA values of the 1/5 hold-out data after averaging over folds and replicates. For each upper bound on the SNP subset size, the actual number of SNPs selected was determined by the best (η,K) values, which had the highest CV correlation. As the upper bound was increased, the actual number of selected SNPs varied (Figure 3). For example, when the upper bound was between 3500 and 5500, the actual size of the SNP subset maintained a value of 3284, because larger SNP subsets did not render better CV performance. The largest subset contained 6484 SNPs, considerably smaller than the largest allowable size (8000). This also meant that only seven SNP subsets (rather than 17 in the supervised PCR) were evaluated for sparse PLS in the final prediction on the testing data.

Comparison of predictive ability between methods

Figure 4 summarizes predictive correlations (on the 1398 testing bulls) using supervised PCR (I and II) and sparse PLS. Supervised PCR II was clearly better than supervised PCR I, and sparse PLS was in between. Despite the notable differences between methods when evaluated on small SNP subsets, all methods approached a similar level of predictive correlation (0.65–0.68) when more SNPs were included in the model. Supervised PCR II plateaued at 3500 SNPs, sparse PLS at approximately 5600 SNPs and supervised PCR I at 6500 SNPs. When all 32 518 SNPs were used, the correlation was 0.68 for PCR and 0.67 for PLS.

Prediction using a small subset of SNPs showed promising results with supervised PCR II and sparse PLS. For example, supervised PCR II achieved predictive correlations of 0.54 and 0.59 with 300 and 500 SNPs, respectively. These corresponded to 80 and 87% of the correlation obtained with all 32 518

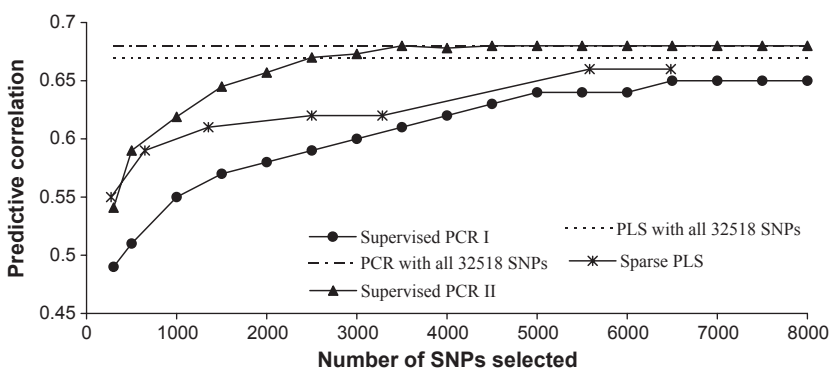

Figure 4 Predictive correlations (correlation between progeny-testing predicted transmitting ability (PTA) and predicted PTA on testing bulls) for the two supervised principal component regression (PCR) methods (I and II) and sparse partial least squares (PLS), evaluated at different numbers of SNPs. Results from ordinary PCR and PLS with all SNPs fitted are also displayed.

SNPs in a PCR model, respectively. Likewise, sparse PLS produced correlations of 0.55 with 272 SNPs and 0.59 with 648 SNPs, corresponding to 82 and 88% of that from fitting all SNPs in a PLS model. Supervised PCR I, however, compared unfavourably with the other two methods in the efficiency of the selected SNP subsets, especially for small subsets.

Discussion

Besides the feature of variable selection manifested in Figure 4, the supervised PCR and sparse PLS also reduced model dimension substantially. As stated earlier, the number of PCs that can be fitted in PCR has a limit, which is the lesser of sample size and SNP subset size. A conservative estimate from our analysis indicated that the gain in predictive correlation was minimal when the number of PCs increased beyond 60% of the limit. For the sparse PLS, the dimension reduction was even stronger. In the seven SNP subsets generated, the first five (ordered by size) required only five latent components, and the last two required ten latent components.

A major difference between the two supervised PCR methods was that method I evaluated SNPs one at a time, while method II considered SNPs jointly when assessing their importance. Here, simultaneous estimation of all SNP effects was accomplished by two steps: regression on SNP PCs and back-transformation of regression coefficients. Advantages of doing so, compared to regression on individual SNPs, were dimension reduction (from 32 518 SNPs to 3000 PCs) and computational savings. It has been noted that SNPs selected by single-SNP analysis may produce more false positives than those selected by multiple-SNP analysis, because the signal at a SNP when analysed individually is often weakened by the inclusion of other correlated SNPs (Hoggart *et al.* 2008). In line with this, it was found that the top-ranked SNPs in method I did not present equivalent (actually lower) ranks when scored in method II. This is shown in Figure 5, where the 300 top-ranked SNPs selected by method I covered almost the whole range of ranks when evaluated by method II, indicating the inconsistency in SNP evaluation between the two methods. Because multi-SNP based analysis tends to give more reliable estimates of SNP effects, this may explain why method II was consistently better than method I when evaluated on SNP subsets of various sizes.

In PCR modelling, the variance structure of the random PC effects is important for prediction accuracy. The ordinary least-squares PCR reflects assumption of equality of variance of predictors and was shown to be inferior to PCR with heterogenous PC variances (Macciotta *et al.* 2009). There, the authors suggested using $\lambda_j \times \sigma_a^2/K$ as the prior variance for PC j, (λ_j was the eigenvalue associated with PC j, K was the number of PCs, and σ_a^2 was the total additive genetic variance) as it relied only on data and did not require prior assumptions about the distributions of PC effects. With regard to

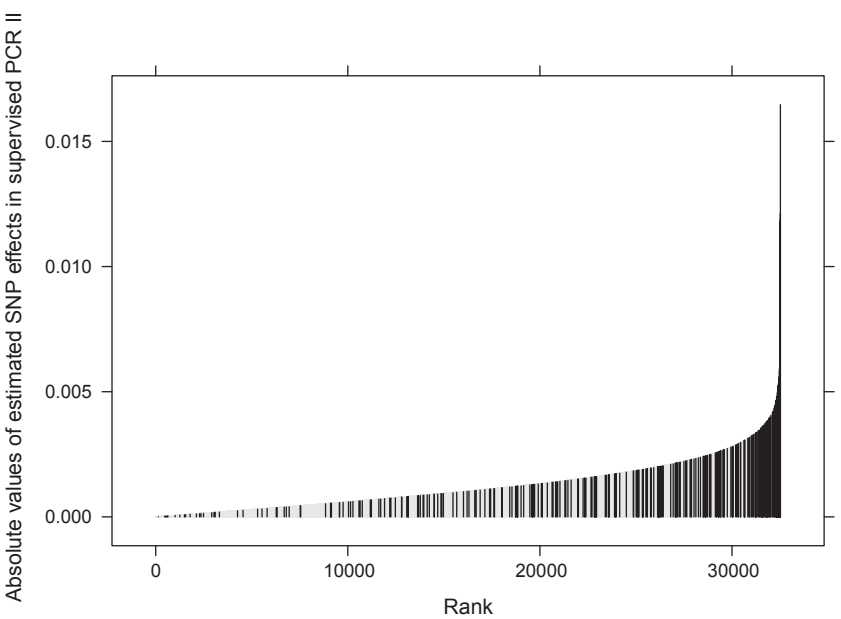

Figure 5 Absolute values of the effects of all 32 518 SNPs estimated by supervised PCR II. The top 300 single nucleotide polymorphisms (SNPs) selected by supervised principal component regression (PCR) I are highlighted by solid bars, with their heights corresponding to the sizes of effects and their positions on the *x*-axis corresponding to the ranks. The 300 SNPs do not cluster in the high-rank region, indicating discrepancy in SNP evaluation between the two methods (I and II).

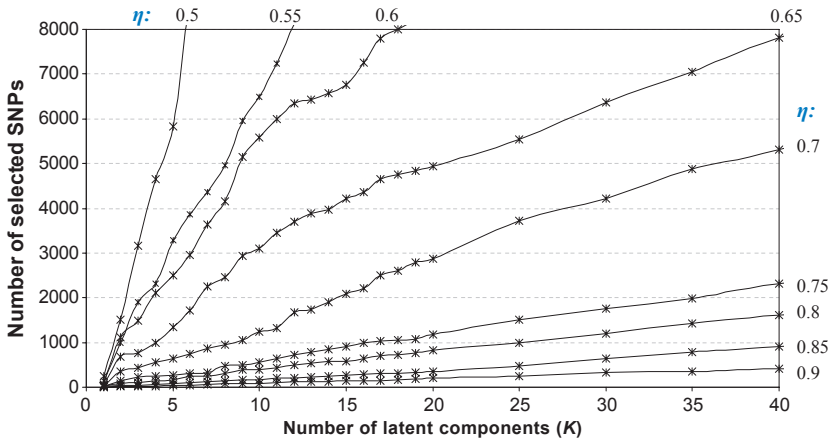

Figure 6 Two tuning parameters (η,K) in the sparse partial least squares control the number of single nucleotide polymorphisms selected. η is between 0 and 1; K is the number of latent components to be extracted.

this, one concern is that the eigenvalues contain information about the SNP correlation structure only, and their magnitudes do not necessarily reflect the association between SNP genotypes and the phenotype in question. In our study, the regressions on PCs were treated as common random effects with differential shrinkage as imposed by double exponential priors, which is a more general setting for variance structure.

The mechanism of SNP selection for sparse PLS is different from that for supervised PCR. The former is controlled by two tuning parameters (η and K) whose values determine the number of selected SNPs, whereas the latter can produce any number of SNPs by adjusting the cut-off value of SNP effects. Although in sparse PLS the number of selected SNPs cannot be directly controlled by user, one can try different combinations of (η, k) so as to find one that produces the subset of SNPs whose size is close to the desired value or below some predefined value. In this study, different upper bounds rather than exact values of the subset size were used for sparse PLS. For each upper bound, the best (η, k) value among all qualified candidates was chosen by CV. Hence, it is necessary to evaluate a sufficient number of (η, k) values, although that would be computationally intensive. Here, η ranged from 0.5 to 0.9 at a step size of 0.05; K ranged from 1 to 40 at a step size of 2 or 5. It is possible that, by trying more values for η and K (e.g. by reducing the step size), the predictive performance of sparse PLS can be enhanced compared to that attained in the current study.

For genomic selection, the primary goal of selecting a subset of whole-genome markers is to reduce genotyping cost while maintaining predictive ability obtained with the selected markers at a reasonable level. Hence, it is natural to base the choice on associations between markers and the trait of interest, and the selected marker set will be trait specific. Our studies presented herein followed this logic. In an alternative line of research (Weigel *et al.* 2010a,b), one starts by choosing a small number of equally spaced markers along the genome and then imputing genotypes for the rest of the markers using a reference population that contains genotypes at all markers. One may expect that predictive ability of these markers would compare unfavourably to that of markers selected based on effects on the trait, especially when the number of markers is small, e.g. <750 in Weigel *et al.* (2010b).

Regarding the choice of response variable in a genomic prediction model, several simulation studies (e.g. Pimentel *et al.* 2009;Macciotta *et al.* 2010) have demonstrated that accuracy (correlation between true BV and predicted BV) of predicted genomic BVs from using raw phenotypes as response variable is even higher than that from using BLUP-estimated BVs (EBV). This may be attributable to low reliability of EBVs or to over-smoothing by the infinitesimal model. Unlike in a simulation setting, the true accuracy (as opposed to model-based accuracy) of the predicted genomic PTAs in a real-data study cannot be assessed, because true BVs are unknown. However, we do not expect an obvious superiority of using raw phenotypes as opposed to using PTAs, because the reliability of the PTAs of bulls in our data was quite high (84% of the sires had PTA reliability greater than 0.8). On the other hand, a better choice is to use weighted versions of PTAs by incorporating reliability information (Garrick *et al.* 2009), owing to that the data comprised bulls with varying reliabilities.

Conclusions

Compared to regression on a large number of SNPs, model dimension in supervised PCR and sparse PLS was greatly reduced by using a smaller number of PCs or latent components as predictors. Supervised PCR II was clearly better than supervised PCR I in predictive ability when evaluated on SNP subsets of various sizes, and sparse PLS was in between. Increasing the number of SNPs improved predictive ability for all methods, and supervised PCR II showed the fastest increase as more SNPs were included. When the SNP subset size was below 1000, sparse PLS presented similar predictive performance to that of supervised PCR II. However, with more SNPs used, the ability of sparse PLS in increasing prediction correlation was inferior to that of the supervised PCR II. Taken together, this study demonstrated the potential of combining dimension reduction and variable selection for accurate and cost-effective prediction of genomic BV.

Acknowledgements

Support by the Wisconsin Agriculture Experiment Station and by grants NRICGP/USDA 2003-35205-12833, NSF DEB-0089742 and NSF DMS-044371 is acknowledged. A. I. Vázquez is thanked for data cleaning and edition.

References

Bair E., Hastie T., Paul D., Tibshirani R. (2006) Prediction by supervised principal components. *J. Am. Stat. Assoc.*, **101**, 119–137.

Boulesteix A.-L., Strimmer K. (2007) Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Briefings Bioinformatics*, **8**, 32–44.

Cadima J., Jolliffe I.T. (1995) Loadings and correlations in the interpretation of principal components. *J. Appl. Stat.*, **22**, 203–214.

de los Campos G., Naya H., Gianola D., Crossa J., Legarra A., Manfredi E., Weigel K. A., Cotes J. (2009) Predicting quantitative traits with regression models for dense molecular markers and pedigrees. *Genetics*, **182**, 375–385.

Chen X., Wang L., Smith J.D., Zhang B. (2008). Supervised principal component analysis for gene set enrichment of microarray data with continuous or survival outcomes. *Bioinformatics*, **24**, 2474–2481.

Chun H., Keleş S. (2009) Expression quantitative trait loci mapping with partial least squares regression. *Genetics*, **182**, 79–90.

Chun H. and Keleş S. (2010) Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *J. R. Stat. Soc. B*, **72**, 3–25.

Frank I.E., Friedman J.H. (1993) A statistical view of some chemometrics regression tools. *Technometrics*, **35**, 109–135.

Garrick D.J., Taylor J.F., Fernando R.L. (2009) Deregressing estimated breeding values and weighted information fro genomic regression analyses. *Genet. Sel. Evol.*, **41**, 1.

Habier D., Fernando R.L., Dekkers J.C.M. (2009) Genomic selection using low-density marker panels. *Genetics*, **182**, 343–353.

Helland, I. S. (1988). On the structure of partial least squares regression. *Commun. Stat.*, **17**, 581–607.

Hoggart C.J., Whittaker J.C., lorio M.D., Balding D.J. (2008) Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. *PLoS Genet.*, **4**, e1000130.

Jolliffe I.T., Trendafilov N.T., Uddin M. (2003) A modified principal component technique based on the lasso. *J. Comput. Graph. Stat.* **12**, 531–547.

de Jong S. (1993). SIMPLS: an alternative approach to partial least squares regression. *Chemom. Intell. Lab. Syst.*, **18**, 251–263.

Long N., Gianola D., Rosa G.J.M., Weigel K.A., Avendano S. (2007) Machine learning classification procedure for selecting SNPs in genomic selection: application to early mortality in broilers. *J. Anim. Breed. Genet.*, **124**, 377–389.

Macciotta N., Gaspa G., Steri R., Pieramati C., Carnier P., Dimauro C. (2009) Preselection of most significant SNPs for the estimation of genomic breeding values. *BMC Proc.*, **3**(Suppl 1), S14.

Macciotta N., Gaspa G., Steri R., Nicolazzi E., Dimauro C., Pieramati C., Cappio-Borlino A. (2010) Using eigenvalues as variance priors in the prediction of genomic breeding values by principal component analysis. *J. Dairy Sci.*, **93**, 2765–2774.

Manne R. (1987) Analysis of two partial-least-squares algorithms for multivariate calibration. *Chemom. Intelligent Lab. Syst.*, **2**, 187–197.

Massy W.F. (1965) Principal components regression in exploratory statistical research. *J. Am. Stat. Assoc.*, **60**, 234–256.

McIntosh A.R., Bookstein F.L., Haxby J.V., Grady C.L. (1996) Spatial pattern analysis of functional brain images using partial least squares. *Neuroimage*, **3**, 143–157.

Meuwissen T.H.E., Hayes B.J., Goddard M.E. (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, **157**, 1819–1829.

Park T., Casella G. (2008). The Bayesian lasso. *J. Am. Stat. Assoc.*, **103**, 681–686.

Pimentel E.C.G., Konig S., Schenkel F.S., Simianer H. (2009) Comparison of statistical procedures for estimat-

ing polygenic effects using dense genome-wide marker data. *BMC Proc.*, **3**(Suppl 1), S12.
Solberg T., Sonesson A., Woolliams J., Meuwissen T. (2009) Reducing dimensionality for prediction of genome-wide breeding values. *Genet. Sel. Evol.*, **41**, 29.
Usai M.G., Goddard M.E., Hayes B.J. (2009) LASSO with cross-validation for genomic selection. *Genet. Res.*, **91**, 427–436.
Vázquez A.I., Rosa G.J.M., Weigel K.A., de los Campos G., Gianola D., Allison D.B. (2010) Predictive ability of subsets of single nucleotide polymorphisms with and without parent average in US Holsteins. *J. Dairy Sci.*, **93**, 5942–5949.
Weigel K.A., de los Campos G., González-Recio O., Naya H., Wu X.L., Long N., Rosa G.J.M., Gianola, D. (2009) Predictive ability of direct genomic values for lifetime net merit of Holstein sires using selected subsets of single nucleotide polymorphism markers. *J. Dairy Sci.*, **92**, 5248–5257.
Weigel K., de los Campos G., Vázquez A., Rosa G., Gianola D., Tassell C.V. (2010a) Accuracy of direct genomic values derived from imputed single nucleotidepolymorphism genotypes in Jersey cattle. *J. Dairy Sci.*, **93**, 5423–5435.
Weigel K., Tassell C. V., O'Connell J., VanRaden P., Wiggans G. (2010b) Prediction of unobserved single nucleotide polymorphism genotypes of Jersey cattle using reference panels and population-based imputation algorithms. *J. Dairy Sci.*, **93**, 2229–2238.
Wold H. (1985) Partial least squares. In: S. Kotz, N. L. Johnson (eds), Encyclopedia of Statistical Sciences, Volume 6. Wiley, New York, NY, pp. 581–591.

Appendices

I. Implementation of sparse PLS with univariate response.

Given values of η and K (number of latent components), sparse PLS can be used for variable selection as follows (Chun & Keleş 2010).

Define $\mathscr{A}$ to be an index set for selected variables and $\mathbf{X}_{\mathscr{A}}$ to be a sub-matrix of $\mathbf{X}(n \times p)$ containing variables indexed in $\mathscr{A}$.

1. Set $\hat{\beta}^{PLS} = \mathbf{0}$, $\mathscr{A} = \{\cdot\}$, $k = 1$, $\mathbf{y}_1 = \mathbf{y}$.
2. While $(k \leq K)$
   2.1. Find $\hat{\mathbf{c}}$ [given by (5)], using $\mathbf{y}_1$ as response variable.
   2.2. Update $\mathscr{A}$ as
   $\{i : \hat{c}_i \neq 0\} \cup \{i : \hat{\beta}_i^{PLS} \neq 0\}, i = 1, \ldots, p$.
   2.3. Fit PLS on $\mathbf{X}_{\mathscr{A}}$ using k latent components $\mathbf{T} = \{\mathbf{t}_1, \ldots, \mathbf{t}_k\}$; update $\hat{\boldsymbol{\beta}}^{PLS}$ via least-squares regression of $\mathbf{y}_1$ on $\mathbf{T}$ ('Overview of PCR and PLS').
   2.4. Update $\boldsymbol{\beta}_1 = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{PLS}$.
   2.5. Update k to $k + 1$.

II. Number of SNPs selected controlled by the two tuning parameters in sparse PLS. Please refer to Figure 6.