Training population selection for (breeding value) prediction

Deniz Akdemir
Department of Plant Breeding & Genetics
Cornell University
Ithaca, NY

January 31, 2014

Abstract

Training population selection for genomic selection has captured a great deal of interest in animal and plant breeding. In this article we derive a computationally efficient statistic to evaluate the quality of a training set for a given test dataset. We adopt a genetic algorithm scheme to find a plausible training set for a given test dataset. Our statistic is related to the reliability measures from the mixed model. Finally, we implement our algorithm on two datasets, namely, the FHB-Barley CAP dataset and Arabidopsis dataset.

1 Introduction

Genomic selection (GS) in animal or plant breeding is based on estimates of genetic breeding values (GEBVs). Prediction of GEBV’s involves implementing a whole-genome regression model where phenotype is regressed on the markers. In GS, first a set of genotypes to be phenotyped (selecting a training population) are identified and a regression model is trained to predict GEBVs for individuals which were not phenotyped. Various regression models have been successfully used for predicting the breeding values in plants and animals. In both simulation studies and in empirical studies of dairy cattle, mice and in bi-parental populations of maize, barley and Arabidopsis marker based GEBVs have been quite accurate. However, it has also been shown that as the training and testing population diverge the accuracies of the GEBVs decrease. As the breeding populations tend to change over time, the result is that the accuracies of the GEBVs obtained from the training population decrease over time. Similarly, in the existence of strong population structure, the
GEBVs obtained by using sub-populations are usually not accurate for individuals in other sub-populations.

In this article we concentrate on the first step of GS, i.e., the selection of training population, to address the accuracy of the GS models. We imagine a scenario in which we are given two sets of individuals and their markers. The first set includes the candidate individuals from which a training set is to be selected for phenotyping to predict the GEBVs of the individuals in the second test set. It will be shown that dynamic model building process which uses genotypes of the individuals in the test sample into account while selecting the training individuals improves the performance of GS models.

In breeding the same issue has captured a great deal of interest. For example, the reliability measure of VanRaden ([6]) is expressed as

\[ K_{21}(K_{11} + \delta I)^{-1}K'_{21} \]  

where \( K_{21} \) is the matrix of genomic relationships between the individuals in the test set to each of the individuals in the training set and \( K_{11} \) measures the genomic relationships in the training set and finally the parameter \( \delta \) is related to the heritability (\( h \)) of the trait by \( \delta = (1 - h^2)/h^2 \). This reliability measure is related to Henderson’s prediction error variance (PEV) ([2]) and the more recent coefficient of determination (CD) of Laloe ([3]) which were both utilized in ([5]) for the training population selection problem.

These measures are definitely related to the situation at hand but they all require expensive evaluations (inversion of large matrices) at each iteration so therefore computationally not feasible for large applications. In the next sections, we derive a computationally efficient approximation to the PEV and use this measure for the training population selection. Another novelty in our method compared to the optimization scheme in ([1]) is that in our case we calculate the prediction error variance for the individuals in the test set instead of evaluating it within the candidate set. We use domain information about the test data while building the estimation model by selecting in the individuals to the training set such that they minimize the PEV in the test set. The methods developed here will be used for dynamic the model building, in other words, different test sets will amount to different individuals be selected from the candidate set and hence different estimation models. It will be shown by examples that this approach can improve accuracies of prediction models.
2 Training Population Selection via Mean Prediction Error Variance Minimization

Traditionally, the breeder is interested in the total additive genetic effects as opposed to the total genetic value. Therefore, a linear model is assumed between the markers and the phenotypes. This is expressed as writing

\[ y = \beta_0 + m^t \beta + e \]  

where \( y \) stands for the phenotype, \( \beta_0 \) is the mean parameter, \( m \) is the \( p \)-vector of marker values, \( \beta \) is the \( p \)-vector of marker effects and \( e \), the difference between the observed and the fitted linear relationship, has a normal distribution with zero mean and variance \( \sigma_e^2 \).

In order to estimate the parameters of this model, we will acquire \( n_{Train} \) individuals from a larger candidate population. The model the will be used to estimate a fixed set of \( n_{Test} \) individuals.

Let \( M \) be the matrix of markers partitioned as

\[ M = \begin{bmatrix} M_{Candidate} \\ M_{Test} \end{bmatrix} \]  

where \( M_{Candidate} \) is the matrix of markers for the individuals in the candidate set and \( M_{Test} \) is the matrix of markers for the individuals in the test set. We would like to identify \( n_{Train} \) training set individuals from the candidate set (and therefore a matrix \( M_{Train} \)) for which the average prediction variance for the individuals in the test set needs to be minimized. Given we have determined \( M_{Train} \) and observed their phenotypes \( y_{Train} \), we can write

\[ y_{Train} = (1, M_{Train})(\beta_0, \beta')' + e. \]  

Under the assumptions of this model the uniformly minimum variance estimators for the phenotypes in the test data is expressed as

\[ \hat{y}_{Test} = M_{Test}((1, M_{Train})' (1, M_{Train}))^{-1} M_{Train}' y_{Train}. \]  

The covariance matrix (Prediction Error Variance (PEV)) for \( \hat{y}_{Test} \) is

\[ PEV(M_{Test}) = (1, M_{Test})((1, M_{Train})' (1, M_{Train}))^{-1}(1, M_{Test})' \]  

where the \( - \) denotes the pseudo inverse of a matrix.

With the emergence of modern genotyping technologies the number of markers can vastly exceed the number of individuals. To overcome the problems emerging
in these large \( p \) with small \( n \) regressions, estimation procedures performing variable selection, shrinkage of estimates, or a combination of both are commonly used while estimating the effects of markers. These methods trade the decreasing variance to increasing bias due to shrinkage of individual marker effects to obtain a better overall prediction performance. Since the variance of these selection-shrinkage methods will be smaller than the least squares estimators, the \( \text{PEV}(M_{\text{Test}}) \) is an upper bound on the covariance matrix of the PEV of these models. To see this consider the PEV from the ridge regression:

\[
\text{PEV}^{\text{Ridge}}(M_{\text{Test}}) = (1, M_{\text{Test}})'(1, M_{\text{Train}})'(1, M_{\text{Train}}) + \lambda I)^{-1}(1, M_{\text{Test}})' \\
\]

Clearly, \( \text{PEV}^{\text{Ridge}}(M_{\text{Test}}) \leq \text{PEV}(M_{\text{Test}}) \) for any \( \lambda \geq 0 \).

We would like to obtain minimum variance for our predictions in the test data set. Therefore, we recommend minimizing

\[
\text{TotalPEV}(M_{\text{Test}}) = \text{tr}(\text{PEV}(M_{\text{Test}}))
\]

with respect to \( M_{\text{Train}} \) when selecting individuals to the training set.

The training data evaluation criteria \( \text{TotalPEV} \) is related to the integrated average prediction variance (IV), where

\[
IV = \frac{1}{A} \int_{\chi} x'(X_{\text{Train}}'X_{\text{Train}})^{-1}x \, dx
\]

where \( A \) is the volume of the space of interest \( \chi \). See Box and Draper ([1]) for a detailed discussion of this criterion. A design that minimizes IV is referred to as IV-optimal. The task is then to choose a subset of \( n_{\text{Train}} \) individuals from the training set for which \( \text{TotalPEV}(M_{\text{Test}}) \) is minimized. Since this is a combinatorial problem we have adopted genetic algorithm where a population of candidate solutions that are represented as binary strings of 0s and 1s is evolved toward better solutions. However, since we are dealing with a large number of markers and any optimization scheme would involve numerous evaluation of this objective function the formula for the \( \text{TotalPEV}(M_{\text{Test}}) \) is not practically applicable.

A numerically efficient approximation to \( \text{PEV}(M_{\text{Test}}) \) can be obtained by using the first few principal components (PCs) of the markers matrix \( M \) instead of \( M \) in the training population selection stage. Let \( P \) be the matrix of PCs partitioned as

\[
P = \begin{bmatrix} P_{\text{Candidate}} \\ P_{\text{Test}} \end{bmatrix}
\]

where \( P_{\text{Candidate}} \) is the matrix of PCs for the individuals in the candidate set and \( P_{\text{Test}} \) is the matrix of PC’s for the individuals in the test set. Now, \( \text{PEV}^{\text{Ridge}}(M_{\text{Test}}) \)
can be approximated by

\[ PEV(M_{\text{Test}}) \approx (1, P_{\text{Test}})((1, P_{\text{Train}})'(1, P_{\text{Train}}) + \lambda I)^{-1}(1, P_{\text{Test}})' \]  

(11)

Finally, we would like to note that the \( PEV(M_{\text{Test}}) \) is related to the reliability measure in (1). To see this, write

\[ (M'_{\text{Train}}M_{\text{Train}} + \lambda I)^{-1} = \frac{1}{\lambda}(I - M'_{\text{Train}}(M_{\text{Train}}M'_{\text{Train}} + \lambda I)^{-1}M_{\text{Train}}) \]  

(12)

If we let \( \delta = m\lambda, K_{21} = M_{\text{Test}}M'_{\text{Train}}/m, K_{11} = M_{\text{Train}}M'_{\text{Train}}/m \) and \( K_{22} = M_{\text{Test}}M'_{\text{Test}}/m \) then we have

\[
\begin{align*}
PEV(M_{\text{Test}}) &= M_{\text{Test}}(M_{\text{Train}}M_{\text{Train}} + \lambda I)^{-1}M_{\text{Test}}' \\
&= M_{\text{Test}}(\lambda(M'_{\text{Train}}M_{\text{Train}}/\lambda) + I)^{-1}M_{\text{Test}}' \\
&= \frac{1}{\lambda}M_{\text{Test}}(I - M'_{\text{Train}}(M_{\text{Train}}M'_{\text{Train}} + \lambda I)^{-1}M_{\text{Train}})M_{\text{Test}}' \\
&= \frac{1}{\lambda}M_{\text{Test}}M'_{\text{Test}} - M_{\text{Test}}M'_{\text{Train}}(M_{\text{Train}}M'_{\text{Train}} + \lambda I)^{-1}M_{\text{Train}}M'_{\text{Test}} \\
&\propto K_{22} - K_{21}(K_{11} + m\lambda I)^{-1}K_{21}'.
\end{align*}
\]

Therefore, maximizing average reliability is equivalent to minimizing the total \( PEV_{\text{Ridge}} \) in (7), however since we would like to be evaluate many candidate training sets in the course of optimization we prefer the computationally efficient approximation in (11). Note that a scalar measure of reliability can be obtained by taking the trace of (11) and will be used subsequently.

### 3 Optimization using genetic algorithm

The training selection optimization involves identification of a set of \( n_{\text{Train}} \) individuals from \( n_{\text{Candidate}} \) individuals in the candidate set and is therefore a combinatorial optimization problem. Genetic algorithms are particularly suitable for optimization of combinatorial problems, therefore its our choice here.

Genetic algorithms work with a population of candidate solutions which are represented as binary strings. At each iteration of the algorithm a fitness function is used to evaluate and select the elite individuals and subsequently the next population is formed from the elites by genetically motivated operations like crossover, mutation.
Figure 1: FHB dataset: The genotypes are grouped into 2 clusters.

4 Applications

The accuracies of the genomic selection models tend to decrease as the training and test populations diverge. We claim that this can be partially fixed by training different models for each test set. In order to evaluate the performance of the selection algorithm for situations like this, we have devised the following illustrations.

Two datasets of different origins were used in our illustrations. The FHB dataset is from the Barley Coordinated Agricultural Project (2011) and it is available from the author in request. A detailed explanation of this data set is given in [4]. The Arabidopsis dataset was published by Atwell et al. (2010) and is available at https://cynin.gmi.oeaw.ac.at/home/resources/atpolydb/

Example 4.1. The FHB data set included FHB and DON measurements along with 2251 single nucleotide polymorphisms (SNP) on 622 elite North American barley lines. FHB is a plant disease caused by the fungus Fusarium Graminearum and results in tremendous losses by reducing grain yield and quality. In addition to the decrease in grain yield and quality, another damage due to FHB is the contamination of the crop with mycotoxins. Therefore, breeding for improved FHB resistance is an important breeding goal.

Scenario: Cluster genotypes into 2 groups (See Figure 4.1), estimate the trait FHB for the first (second) cluster using n=25, 50 training observations in the rest of the clusters.

The accuracies of the optimized sample compared to a random sample are compared with the box-plots in Figures 2, 3, and 4.
Figure 2: Trait: FHB. The box-plots summarize the accuracies 30 replications of the experiment. In each sub-figure ((a)-(c)) a different setting is tested and the box-plots to the left correspond to the optimized set whereas the ones on the right correspond to the random sample.

**Example 4.2.** Arabidopsis dataset consisted of genotypes of 199 inbred lines along with observations on 107 traits. Here we will analyze only two of these traits: SDV and FT10. The results from the following scenarios are reported below.

**Scenario 1:** Cluster genotypes (only those that had SDV measured) into 9 groups (See Figure 4.2), estimate the trait SDV for the first (sixth) cluster using n=25, 50 training observations in the rest of the clusters.

**Scenario 2:** Cluster genotypes (only those that had FT10 measured) into 4 groups (See Figure 4.2), estimate the trait FT10 for the second cluster using n=25, 50 training observations in the rest of the clusters.

The accuracies (correlation between the estimated and test data response) of the optimized sample compared to a random sample are compared with the box-plots in Figures 7 and 8.

### 5 Conclusions

In this article we have taken on the training selection problem and have shown by examples that incorporating information about the test set when available can improve the accuracies of prediction models. The approach we developed here is particularly suitable when the size of the candidate and training data sets are large.
since our approach is computationally efficient. Our findings suggests dynamic model building should be used when possible, when the domain of the test data is known. In fact, an approach we would like to explore in our further studies is to build a separate model for each point in the dataset by identifying the best individuals in a candidate set for estimating that point or by eliminating the irrelevant, outlier or influential individuals to enter into the model.

Acknowledgments

This research was supported by the USDA-NIFA-AFRI Triticeae Coordinated Agricultural Project, award number 2011-68002-30029.

References

[1] George EP Box and Norman R Draper. A basis for the selection of a response surface design. *Journal of the American Statistical Association*, 54(287):622–654, 1959.
Figure 4: Trait: DON. The box-plots summarize the accuracies 30 replications of the experiment. In each sub-figure ((a)-(c)) a different setting is tested and the box-plots to the left correspond to the optimized set where as the ones on the right correspond to the random sample.

[2] Charles R Henderson. Best linear unbiased estimation and prediction under a selection model. *Biometrics*, pages 423–447, 1975.

[3] D Laloë, F Phocas, and F Ménissier. Considerations on measures of precision and connectedness in mixed linear models of genetic evaluation. *Genetics Selection Evolution*, 28(4):359–378, 1996.

[4] AJ Lorenz, KP Smith, and J-L Jannink. Potential and optimization of genomic selection for fusarium head blight resistance in six-row barley. *Crop Science*, 52(4):1609–1621, 2012.

[5] Renaud Rincent, Denis Laloë, Stéphane Nicolas, Thomas Altmann, Dominique Brunel, Pedro Revilla, Victor M Rodriguez, J Moreno-Gonzalez, A Melchinger, Eva Bauer, et al. Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals: Comparison of methods in two diverse groups of maize inbreds (zea mays l.). *Genetics*, 192(2):715–728, 2012.

[6] PM VanRaden. Efficient methods to compute genomic predictions. *Journal of dairy science*, 91(11):4414–4423, 2008.
Figure 5: Arabidopsis dataset: The genotypes are grouped into 9 clusters.
Figure 6: Arabidopsis dataset: The genotypes are grouped into 4 clusters.
Figure 7: Scenario 1: The box-plots summarize the accuracies 30 replications of the experiment. In each sub-figure ((a)-(d)) a different setting is tested and the box-plots to the left correspond to the optimized set whereas the ones on the right correspond to the random sample.

Figure 8: Scenario 2: The box-plots summarize the accuracies 30 replications of the experiment. In each sub-figure ((a)-(b)) a different setting is tested and the box-plots to the left correspond to the optimized set whereas the ones on the right correspond to the random sample.
