Genomic Selection with Linear Models and Rank Aggregation

Marco Scutari

m.scutari@ucl.ac.uk
Genetics Institute
University College London

March 5th, 2012

Genomic Selection

Genomic Selection: an Overview

Genomic selection (GS) is a form of marker-assisted selection in which genetic markers covering the whole genome are used, with the aim of having quantitative trait loci (QTL) for a given trait of interest in linkage disequilibrium with at least one marker.

This is in contrast with:

- pedigree-based selection, which uses kinship and repeated crossings to introduce and fix the desired trait.
- QTL-based selection, which uses only those markers that display a strong association with the trait.

Implementing Genomic Selection

The fundamental steps in genomic selection:

1. set up one or more designed experiments to measure the traits of interest controlling for extraneous (confounding) environmental and population effects;
2. collect the marker profiles of the varieties involved in the experiments;
3. use marker profiles, which should provide as good a coverage as possible of the genome, to model the trait of interest;
4. use the genetic effects estimated in the model to predict the performance of new varieties based on their marker profiles.

Selection of new varieties is then performed on the basis of the predicted traits.

Implementing Genomic Selection

Some important points:

- the number of varieties, the number of experimental units for each variety and the marker density jointly affect the precision of the predictions;
- experimental design is further complicated by the fact that environmental effects are much stronger than most marker effects on the trait, so great care must be taken to avoid confounding;
- some care must be taken in choosing a diverse set of varieties to ensure that different alleles are well represented and, therefore, that their effects are estimated with sufficient accuracy.

Linear Modelling

Linear Modelling: an Overview

In the context of genomic selection, linear modelling is usually denoted as

$$\mathbf{y} = \mu \mathbf{1}_n + \mathbf{X}\mathbf{g} + \boldsymbol{\varepsilon}$$

where

- $\mathbf{y}$ is the trait of interest;
- $\mu \mathbf{1}_n$ is the intercept of the model, with $\mu = \bar{y}$;
- $\mathbf{X}$ is the matrix containing the (coded) marker profiles;
- $\mathbf{g}$ is the vector of the genetic effects;
- $\boldsymbol{\varepsilon}$ is the error term, usually assumed to be normally distributed.

Assumptions

- The model only accounts for additive effects (i.e. no epistasis and no dominance).
- All the environmental effects are assumed to have been removed beforehand, as the model is of the form

  $$\text{TRAIT} \sim \text{GENETIC EFFECTS}$$

  or, at most,

  $$\text{TRAIT} \sim \text{GENETIC EFFECTS} \times \text{TREATMENT}.$$

- Residuals are usually assumed to be independent, so if the varieties whose profiles are used in the model are related, all kinship effects are in turn assumed to be modelled through the markers.

Ridge Regression

Ridge Regression shrinks the genetic effects by imposing a quadratic penalty on their size, which amounts to the penalised least squares

$$\hat{\mathbf{g}}_{\mathrm{ridge}} = \operatorname*{argmin}_{\mathbf{g}} \left\{ \sum_{i=1}^{n} (y_i - \mu - \sum_{j=1}^{p} x_{ij} g_j)^2 + \lambda \sum_{j=1}^{p} g_j^2 \right\}.$$

It is equivalent to a best linear unbiased predictor (BLUP) when the genetic covariance between lines is proportional to their similarity in genotype space, which is why it is sometimes called Ridge Regression-BLUP (RR-BLUP).

LASSO Regression

LASSO is similar to Ridge Regression, but with a different penalty (L_1 vs L_2):

$$\hat{\mathbf{g}}_{\mathrm{lasso}} = \operatorname*{argmin}_{\mathbf{g}} \left\{ \sum_{i=1}^{n} (y_i - \mu - \sum_{j=1}^{p} x_{ij} g_j)^2 + \lambda \sum_{j=1}^{p} |g_j| \right\}.$$

The main difference with Ridge Regression is that LASSO can force some of the genetic effects to be exactly zero, which is consistent with the relative sizes of the profile and the sample ($n \ll p$).

Elastic Net Regression

Elastic Net combines Ridge Regression and LASSO by weighting their penalties as follows:

$$\hat{\mathbf{g}}_{\text{enet}} = \operatorname*{argmin}_{\mathbf{g}} \left\{ \sum_{i=1}^{n} (y_i - \mu - \sum_{j=1}^{p} x_{ij} g_j)^2 + \right.$$
$$\left. + \lambda \sum_{j=1}^{p} (\alpha g_j^2 + (1-\alpha)|g_j|) \right\}.$$

The Elastic Net selects variables like the LASSO, and shrinks together the coefficients of correlated predictors like Ridge Regression.

Partial Least Squares Regression

Partial Least Squares (PLS) Regression models the trait of interest using the k principal components $\mathbf{z}_1, \ldots, \mathbf{z}_k$ of $\mathbf{X}$ that are most strongly associated with the trait. The fundamental idea behind this model is

$$\hat{\mathbf{b}}_{\mathrm{pls}} \approx \operatorname*{argmin}_{\mathbf{b}} \left\{ \sum_{i=1}^{n} (y_i - \mu - \sum_{j=1}^{k} z_{ij} b_j)^2 \right\}.$$

Because of that, the dimension of the problem is greatly reduced but the model does not provide explicit estimates of the genetic effects $\mathbf{g}$.

BayesB Bayesian Regression

BayesB is a Bayesian linear regression in which the genetic effects g_i have a normal prior distribution with variance

$$\begin{cases} \sigma^2_{g_i} & = & 0 & \text{with probability } \pi \\ \sigma^2_{g_i} & \sim & \chi^{-2}(\nu, S) & \text{with probability } 1 - \pi \end{cases}.$$

The probability mass at 0 forces many genetic effects to zero. The posterior distribution for $\mathbf{g}$ is not in closed form, so genetics effects are estimated with a (not so fast) combination of Gibbs sampling and Metropolis-Hastings MCMC.

The Feature Selection Problem

It is not possible for all markers in the profile to be relevant for the trait we are selecting for, both because they usually outnumber the varieties ($n \ll p$) and because some provide essentially the same information due to linkage disequilibrium.

Therefore, genomic selection is a feature selection problem. We aim to find the subset of markers $\mathbf{S} \subset \mathbf{X}$ such that

$$\mathrm{P}\left(\mathbf{y} \,|\, \mathbf{X}\right) = \mathrm{P}\left(\mathbf{y} \,|\, \mathbf{S}, \mathbf{X} \setminus \mathbf{S}\right) = \mathrm{P}\left(\mathbf{y} \,|\, \mathbf{S}\right),$$

that is, the subset of markers ($\mathbf{S}$) that makes all other markers ($\mathbf{X} \setminus \mathbf{S}$) redundant as far as the trait we are selecting for is concerned.

Markov Blankets & Feature Selection

There are several ways to identify $\mathbf{S}$; some of the models above do that implicitly (i.e. LASSO). A probabilistic approach that does that explicitly is Markov blanket learning. A Markov blanket (MB) is a minimal set $\mathcal{B}(\mathbf{y})$ that satisfies

$$\mathbf{y} \perp\!\!\!\perp X \setminus \mathcal{B}(\mathbf{y}) \,|\, \mathcal{B}(\mathbf{y})$$

and is unique under very mild conditions. It can be learned from the data in polynomial time using a sequence of conditional independence tests involving small subsets of markers. The markers in $\mathcal{B}(\mathbf{y})$ can then be used for genomic selection with one of the linear models illustrated above.

Pros & Cons of the Different Models for GS

- Finding the optimal value for the λ parameter of Ridge Regression and LASSO is nontrivial, because cross-validated estimates based on predictive correlation and predictive log-likelihood often do not agree.
- Tuning the Elastic Net is very time consuming, because cross-validation must be performed over a grid (α, λ) of parameters. However, once tuned Elastic Net outperforms both Ridge Regression and LASSO.
- PLS and MB Feature Selection are the easiest to tune, as they have a single parameter (k and the type I error for the tests, respectively) and both predictive correlation and predictive log-likelihood usually are unimodal in that parameter.
- Choosing the π in BayesB greatly benefits from some prior knowledge on the genetics of the trait we are selecting for.

Genomic Selection in Barley

Spring Barley Data

We applied the models described in the previous section to perform genomic selection in spring barley. The training set comprises:

- 133K yield measurements for 1189 varieties, collected from 769 of trials in the UK, France and Germany from 2006 to 2010;
- both treated (with pesticides) and untreated data;
- a marker profile of 6318 SNPs for each variety.

Varieties in this set are (closely) related, as they are the result of repeated selections performed over the years.

Estimating the Expected Yield for Each Variety

To separate the genetic components from other effects, we used the following mixed linear model:

$$\mathrm{YIELD} \sim \underbrace{\mathrm{TREATMENT}}_{experimental} + \mathrm{VARIETY} \times \mathrm{TREATMENT} + \mathrm{VARIETY} \times \mathrm{TRIAL} + \underbrace{\mathrm{TRIAL}}_{environmental}.$$

The expected yield for each variety, known as the expected breeding value (EBV), was then computed as

$$\mathrm{EBV}(\mathrm{VARIETY}, \mathrm{TREATMENT}) = \\ = \mu + \mathrm{VARIETY} \times \mathrm{TREATMENT} + \\ + \sum_{\mathrm{TRIAL}} w_{\mathrm{TRIAL}} \cdot \mathrm{VARIETY} \times \mathrm{TRIAL}.$$

Pre-Processing the Marker Profiles

Marker profiles were screened prior to genomic selection as follows:

- 105 were dropped because monomorphic;
- 46 were dropped because missing for more than 20% of the varieties;
- when a pair of SNPS were found to be highly correlated ($> 99\%$), one of them was removed to increase the numerical stability of the genomic selection models. Higher thresholds (i.e. $> 99.5\%$, $> 99.9\%$) can be used to make the marker set even more regular.

The remaining missing marker data were imputed using a k-nearest neighbour with $k = 2$ (i.e. the closest two varieties) with an estimated imputation error of 5%.

Performance of the Models for Genomic Selection

	with Treatment		Treated only	
Model	COR	COR_{xval}	COR	COR_{xval}
Ridge Regression	0.6842	0.5227	0.7177	0.4164
LASSO Regression	0.7221	0.5122	0.6456	0.3566
Elastic Net	0.7438	0.5236	0.7388	0.4172
PLS Regression	0.7358	0.5071	0.6359	0.3572
BayesB	–	–	0.7203	0.3900
MB Feature Selection	0.7279	0.6658	0.5791	0.5139

COR = Pearson's correlation between observed and predicted EBVs.

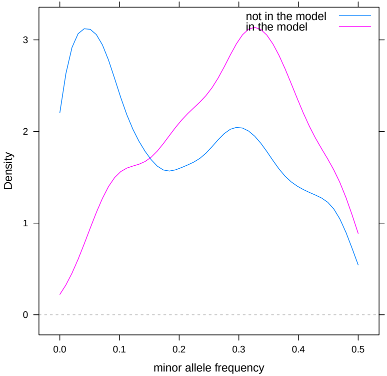
COR_{xval} = same as above, but computed using cross-validation to avoid unrealistically optimistic estimates.

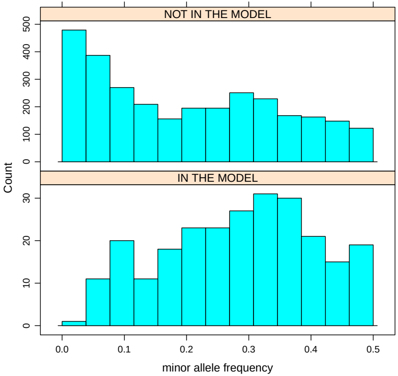
A Note on Rare Alleles

Even when they have comparable predictive power, different models can provide different insights on the genetic effects involved controlling a particular trait.

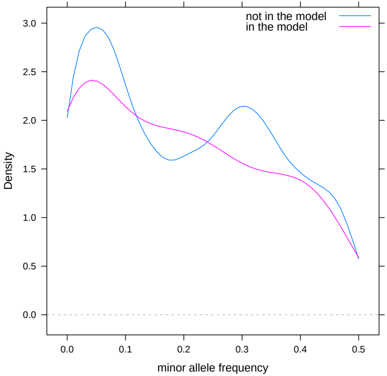
Consider for example, the LASSO and MB Feature Selection. Both include a subset of markers in the respective genomic selection models, while assigning null effects to the others. While the dimensions of those subsets are comparable, the influence of minor allele frequency on the probability of inclusion is completely different.

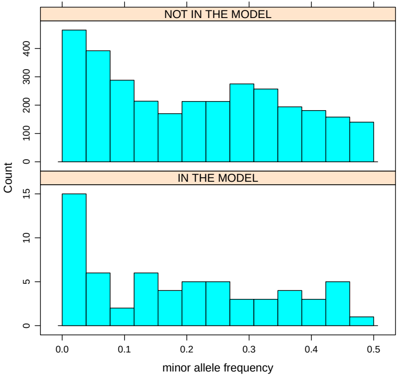
LASSO Regression and Rare Alleles



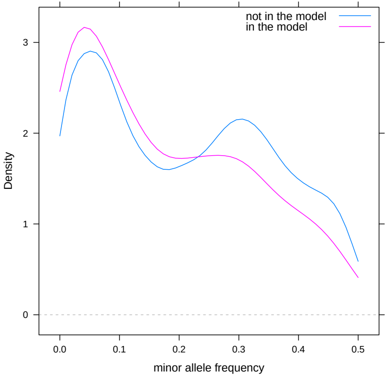


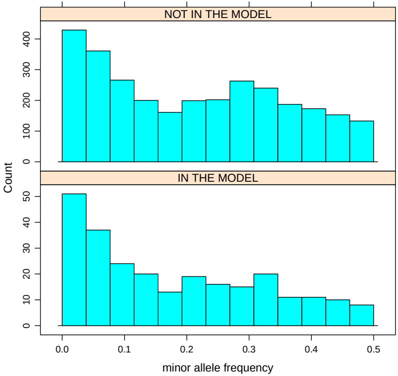
Feature Selection and Rare Alleles





LASSO (with standardised SNPs) and Rare Alleles





Ranking and Model Averaging

Ranking and Genomic Selection

The main goal of genomic selection is to select new varieties with better values for the trait of interest. Therefore, the value of the trait of a particular variety is less important than how it compares with the values of other, competing varieties.

For this reason, it is natural to order new varieties according to their predicted EBVs and focus on their rank:

- ranks are more robust than the EBVs they are derived from;
- and they still contain all the information needed to perform the selection.

Measuring Distance Between Ranks

Having different genomic selection models, it is useful to compare the rankings that they produce for new varieties. The most common distance measure to do that is Kendall's τ:

$$\tau = \frac{\text{(concordant pairs)} - \text{(discordant pairs)}}{\frac{1}{2}(n)(n-1)}$$

where concordant pairs are pairs of EBVs whose ranks agree (the highest ranked EBV of the pair is the same in both rankings) and discordant pairs are pairs whose ranks do not agree (each EBVs is ranked higher than the other in one ranking and lower in the other ranking).

Model Averaging

In addition, having different genomic selection models for the same varieties makes the use of model averaging possible. Combining the predicted ranks from different models:

- makes the prediction errors made by any one model irrelevant as long as the other models behave correctly;
- allows the combination of the predictions based on different information, because different models are better at capturing different kinds of genetic effects;
- averaged models are "smoother" than the original ones, and have been proved to have better predictive power for many classes of statistical models.

For ranks, model averaging takes the name of rank aggregation.

Rank Aggregation

```
* top 20 lines by averaged rank:
   INDIVIDUAL   ridge   lasso elastic     pls feature
1   xxxx-yyyy 77.2705 78.7776 78.0880 76.0533 77.2841
2   xxxx-yyyy 76.8105 78.2329 77.8659 75.8181 80.2320
3   xxxx-yyyy 76.8467 77.9358 77.0641 75.8988 79.1587
4   xxxx-yyyy 76.5639 77.7688 77.3653 76.0560 77.4509
5   xxxx-yyyy 76.6305 77.4622 77.4581 76.1455 75.2964
* bottom 20 lines by averaged rank (from bottom up):
   INDIVIDUAL   ridge   lasso elastic     pls feature
1   xxxx-yyyy 73.5585 73.0527 73.2776 74.8116 73.1224
2   xxxx-yyyy 73.4462 73.0858 73.1713 75.1587 73.9776
3   xxxx-yyyy 73.6860 72.1180 72.9532 75.4189 72.5972
4   xxxx-yyyy 73.8401 73.4667 73.3646 75.2319 73.5014
5   xxxx-yyyy 73.5797 73.4756 73.4463 74.9363 74.8271
```

Conclusions

Conclusions

- Many different models have been proposed in literature for genomic selection, each with its own strength and weaknesses.
- Different models are better at capturing different information; for instance, some make better use of rare alleles than others.
- For the purpose of genomic selection, using ranks instead of the predicted EBVs provides a more robust alternative.
- Rank aggregation provides the means of combining information from different models and at the same time, to offset their weak points.

References I

G. Claeskens and N. L. Hjort.
Model Selection and Model Averaging.
Cambridge University Press, 2008.

J. B. Endelman.
Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP.
The Plant Genome, (4):250–255, 2011.

J. J. Goeman.
Penalized R package, 2012.
R package version 0.9-38.

T. Hastie, R. Tibshirani, and J. Friedman.
The Elements of Statistical Learning: Data Mining, Inference, and Prediction.
Springer, 2nd edition, 2009.

E. L. Heffner, M. E. Sorrells, and J. Jannink.
Genomic Selection for Crop Improvement.
Crop Science, (49):1–12, 2009.

References II

J. M. Hickey and B. Tier.
AlphaBayes: Software for Polygenic and Whole Genome Analysis, 2009.
University of New England, Armidale, Australia.

I. T. Jolliffe.
Principal Component Analysis.
Springer, 2nd edition, 2002.

D. Koller and M. Sahami.
Toward Optimal Feature Selection.
In *Proceeding of the 13th International Conference on Machine Learning*, pages 284–292, 1996.

T. H. E. Meuwissen, B. J. Hayes, and M. E. Goddard.
Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps.
Genetics, 157:1819–1829, 2001.

B. Mevik, R. Wehrens, and K. H. Liland.
pls: Partial Least Squares and Principal Component regression, 2011.
R package version 2.3-0.

References III

V. Pihur, S. Datta, and S. Datta.
Weighted Rank Aggregation of Cluster Validation Measures: a Monte Carlo Cross-Entropy Approach.
Bioinformatics, 23(13):1607–1615, 2007.

V. Pihur, S. Datta, and S. Datta.
RankAggreg: Weighted Rank Aggregation, 2011.
R package version 0.4-2.

M. Scutari.
Learning Bayesian Networks with the bnlearn R Package.
Journal of Statistical Software, 35(3):1–22, 2010.

J. C. Whittaker, R. Thompson, and M. C. Denham.
Marker-assisted selection using ridge regression.
Genetical Research, (75):249–252, 2000.

H. Zou and T. Hastie.
Regularization and Variable Selection via the Elastic Net.
Journal of the Royal Statistical Society (Series B), 67(2):301–320, 2005.