Crossing traditional and contemporary genomic prediction techniques.

UNIVERSITY OF CAMBRIDGE
Chemical Engineering and Biotechnology

Summary

Two of the challenges that statisticians face when performing genomic prediction include accounting for epistatic interactions and population structure. Each problem favours a different approach. In our work, we combine these approaches aiming to improve genomic prediction algorithms.

Background

Epistasis is a circumstance where the expression of one gene is affected by the expression of one or more independently inherited genes [1].

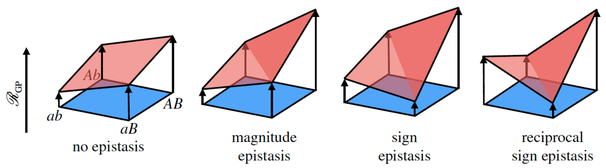


Figure 1. Forms of pairwise epistasis. The fitness (height above the plane) effects of different mutations, modelled as flipping a single bit of the genotype, can differ depending on the genetic background. These forms of epistasis can inhibit evolutionary trajectories between two genotypes.[2]

Population structure (also called genetic structure and population stratification) is the presence of a systematic difference in allele frequencies between subpopulations.[3]

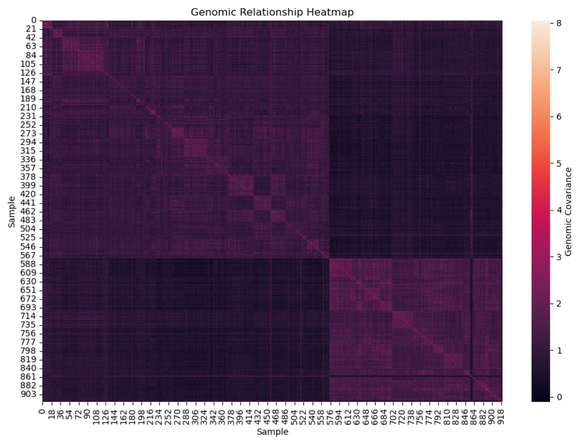


Figure 2. Population structure is often measured using the genetic covariance between individuals in the population. Here we have plotted the covariance as a heatmap. The covariance plotted above comes from a Lolium dataset [4] using one of the methods defined by Henderson et al [5].

Current approaches

Regression trees partition a data set into smaller groups and then fit a simple model (constant) for each subgroup. Unfortunately, a single tree model tends to be highly unstable and a poor predictor. [6]

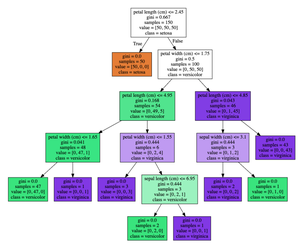


Figure 3. Example decision tree.[5]

Regression trees are capable of dealing with epistatic interactions. This is due to their propensity to split a dataset into subgroups.

By bootstrap aggregating (bagging) regression trees, this technique can become quite powerful and effective. Random Forest is one such example.

Linear mixed-effect models extend simple linear models. They allow for both fixed and random effects. They are used when data is not independently distributed, such as arises from a hierarchical structure [7].

$$\underbrace{y}_{\text{Phenotype}} = \underbrace{X\alpha}_{\text{Fixed Effects}} + \underbrace{u}_{\text{Random Effects}} + \underbrace{\epsilon}_{\text{Noise}}$$

Where $u \in \mathcal{N}(0, \mu G)$

$\epsilon \in \mathcal{N}(0, \sigma I)$

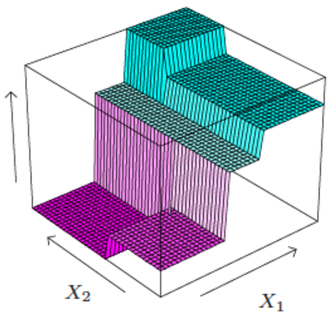
Hence $y \in \mathcal{N}(X\alpha, \mu G + \sigma I)$

The G represents the genomic relationship matrix - an example of this is displayed in Figure 2. The I is an identity matrix. The sample genomes are represented by X. The effect of each SNP on the phenotype is represented by α. The inclusion of the Random Effects Component improve the optimal solution of the regressor. Each SNP can be seen to act independently of any other SNP.

Problems & Solutions

Regression trees do not make use of the various forms of covariance between samples to improve their predictions.

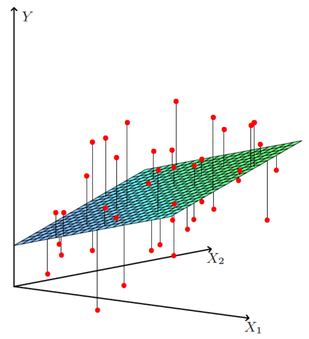
Figure 4. When we fit a regression tree we partition the sample space based on its features. This figure shows how a tree partitions a feature space. Where, X1 and X2 represent features and the prediction for each group is plotted along the Z axis. In our current work we are looking at changing the positions of these partitions to account for population structure.[6]



We have been looking at ways of augmenting decision trees to account for the genomic covariance inherent in genomic datasets. Our first attempt involved using the genomic relationship matrix - used in linear mixed effect models - to change the propensity to split on SNPs aligned with population structure.

Linear mixed-effect models in high-dimensional spaces can be seen as a form of Gaussian Process. Where the random effects and noise terms are equivalent to a noisy anisotropic linear kernel.

Figure 5. When we fit a linear mixed-effects model we fit a hyperplane to the samples. The objective is to maximise the likelihood of the training data given the hyperparameters. The covariance between samples in the model is used to calculate this likelihood. The covariance does not take into account epistatic interactions. We aim to produce a kernel that does.[6]



The Linear mixed effect models are not capable of accounting for epistasis. This is due to the linear kernel used by these algorithms. Alternative kernels that work well with high dimensional data are lacking. Thus, we are implementing a kernel that allows these models to account for epistasis. This kernel is based on how decision trees partition the input space.

Results

Our customised trees have a positive effect on the error margin of the trees applied to the NIAB MAGIC dataset [8]. The result loses significance when applied to the Lollium dataset.

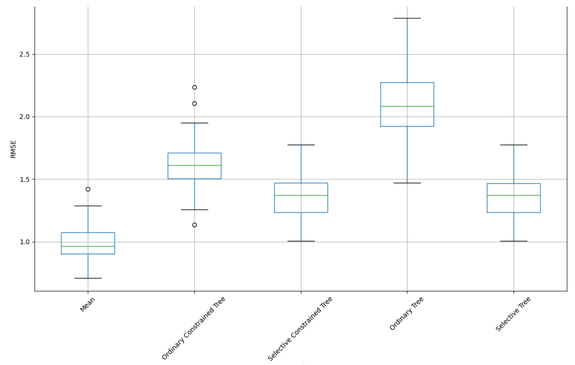


Figure 6. When we fit a decision tree we partition the sample space based on its features. This figure shows how a tree partitions a feature space. In our current work, we are looking at changing the positions of these partitions to account for population structure.

This improvement in tree performance does not appear to carry through to an improvement once the trees are ensembled.

Moving Forward

We have more work to do to assess the possibility of ensembling our customised decision trees in different ways. Our aim in doing this is to allow our improvements to instantiate themselves in ensembles.

Further to this we are working to complete and implement our decision tree kernel and assess its performance across various datasets.

This work was produced by Harry Blakiston Houston, under the supervision of Dr Nastasiya Grinberg and Professor Ross King. Harry is PhD Student in the Department of Biotechnology and Chemical Engineering at the University of Cambridge. The author declares no conflict of interest.

References

[1] Allaby. Dictionary of Plant Sciences. 2006.
[2] Nichol D, Robertson-Tessi M, Anderson ARA & Jeavons P. 2019.
[3] Ubbens, Feldmann, Stavness & Sharpe. 2022.
[4] Blackmore et al. 2015.
[5] VanRaden. 2008
[6] Hastie, Friedman & Robert. 2001.
[7] Pedregosa et al. Scikit-Learn. 2011.
[8] Scott et al. 2020.