Phenotype Prediction from Human Whole Genome Profile

CSE847 COURSE PROJECT

SUN, MENGYING; TONG, XIAORAN

Intro: Problem

Genomic Prediction

❑ Many phenotypes are highly heritable, e.g., height, IQ, and diseases, which can be well predicted by pedigree trees.

❑ Yet, the prediction from genomic profile are not optimal.

❑ Assumption: sub-optimal performance was due to neglected nonlinear association.

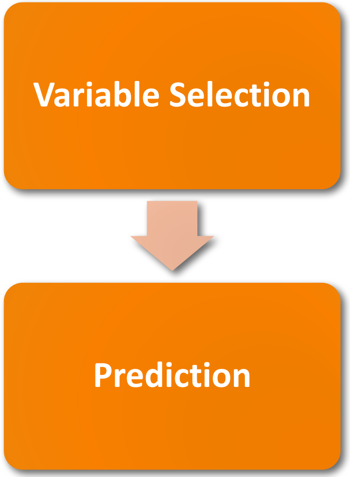
❑ Our goal: build non-linear predictive models for a typical phenotype - body height, that uses markers across the whole genome.

Genomic Profile

- SNP takes value from {0, 1, 2}
- Dimensionality (million)
- **UK biobank, Height**
  - 589,028 SNPs (~600K)
  - 102,221 Observations (~100K)
- Training Testing Splitting
  - TRN: 80000
  - TST: 22221

Strategy: Two Stage Modeling

- Final Goal: Y = f(G)
- **Variable Selection**
  - Whole genome --> LD Blocks
  - Select SNPs
- **Prediction**
  - Generative models
  - Neural Networks



Variable Selection

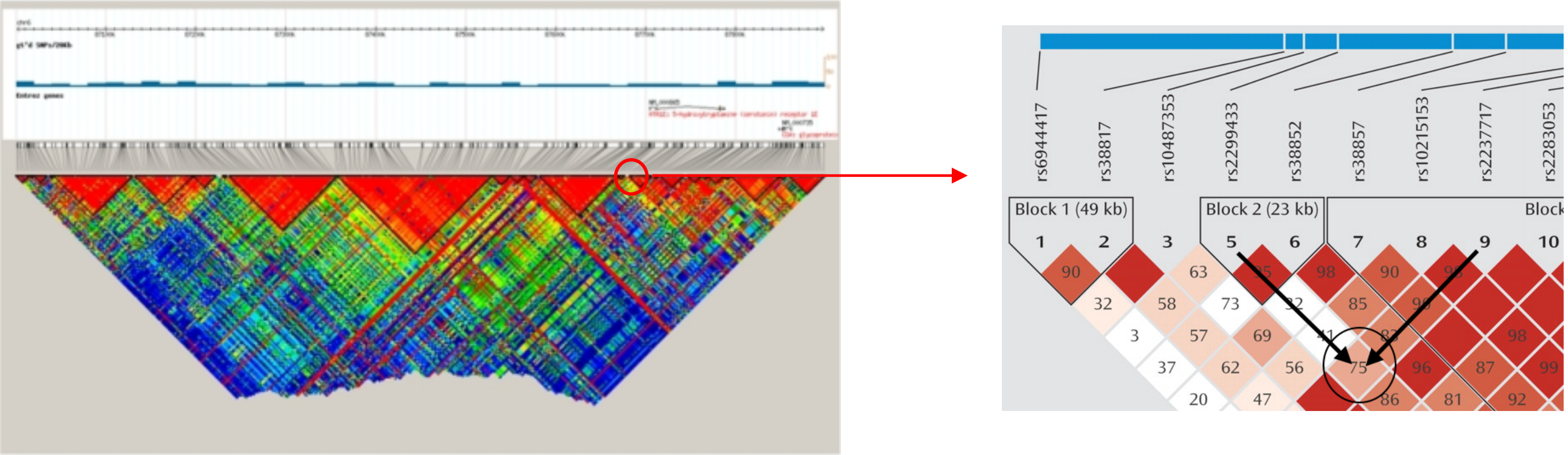
Prediction

Methods

LD Blocks

- Linkage Disequilibrium (LD)

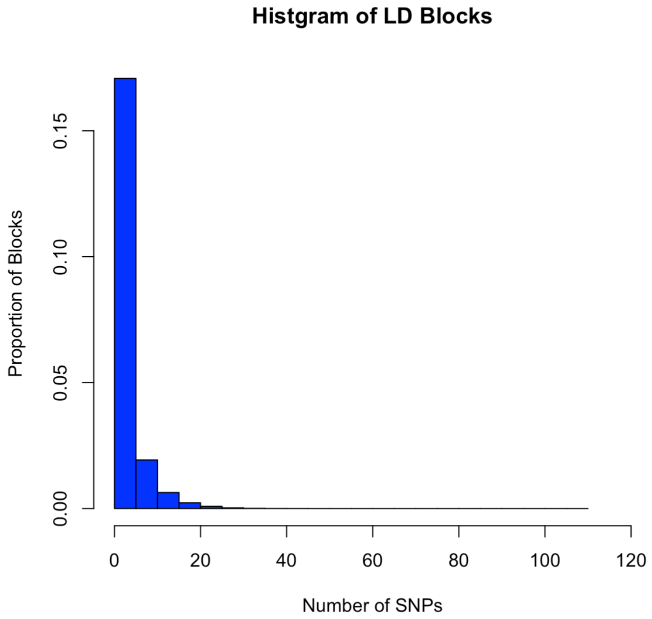
  Variations at close by locations are not independent, due to the molecule bounds.



LD Blocks

- LD blocks identified by PLINK
- End up with 157K LD blocks

Statistics	Value
Min	1
Max	107
Mean	3.097
SD	3.686



Variable Selection

- Stepwise selection (BIC/AIC)
- LASSO

Variable Selection

- **BIC**
  - Only keep selected SNPs in a LD-block
  - Completely discard blocks with no selected SNPs
  - A total of 6K features selected by BIC
- **LASSO**
  - Merge consecutive LD blocks to form superblocks (300 per block, ~2k blocks)
  - Split training into sub-training & sub-validation, calculate solution path for b w.r.t. Lambda (>200k features left)
  - Calculate one risk score for every superblock, a total of 2k scores
- **Control**
  - Random 6K
  - Top p-value 6K from GWAS study

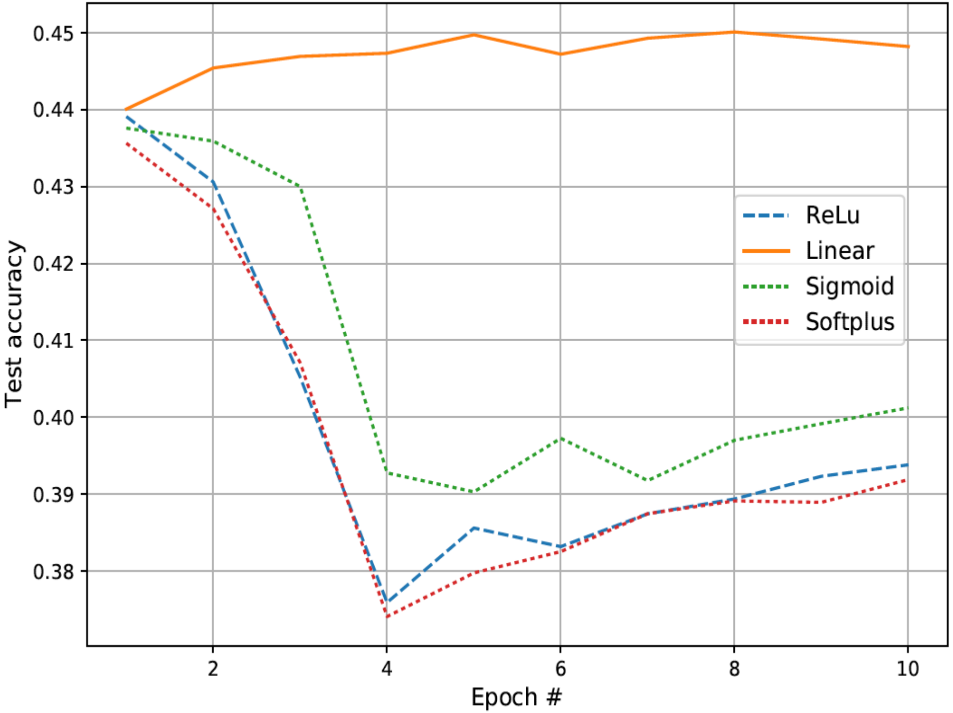
Prediction

- BGLM (BGLR)
- NN (Keras)

Results

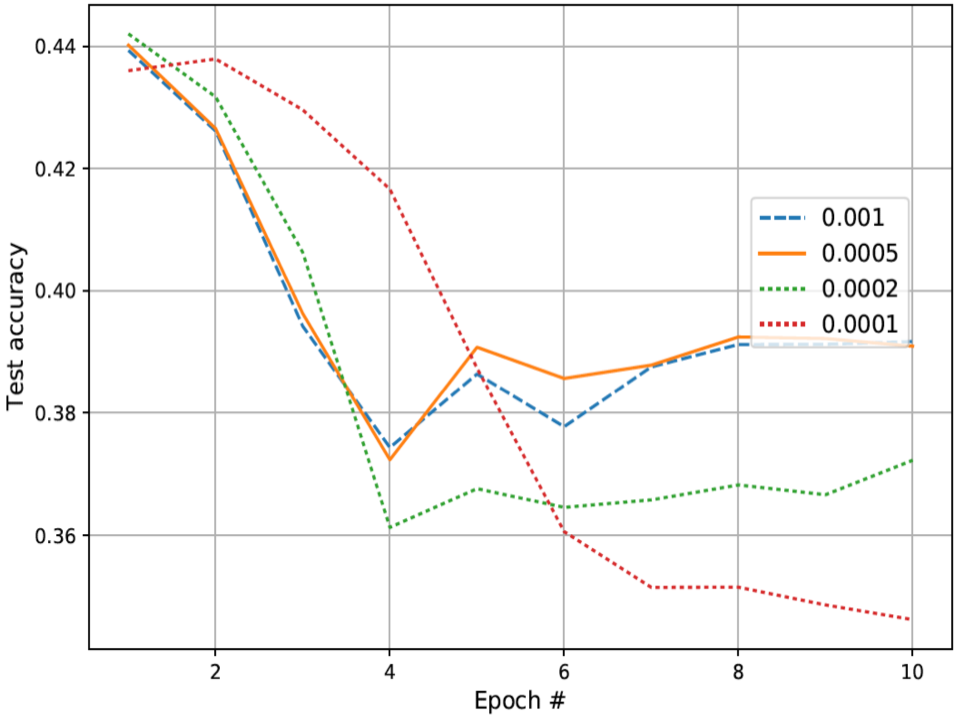
Prediction

- **BGLM**
  - Bayes B, nIter = 5000, BurnIn = 1000
- **NN**
  - Activation Function
  - Learning Rate
  - Regularization Parameters
- **PA**
  - Correlation (y_test_true, y_test_pred)



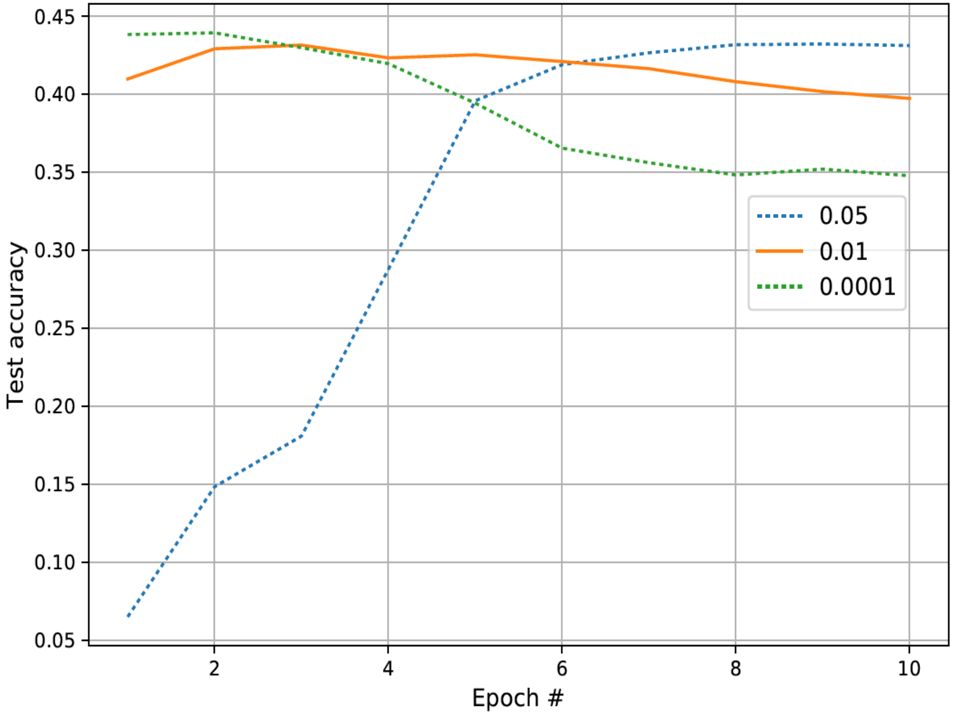
Activation Function

- **Overfit quickly**
  - Large sample size
  - Small batch
  - Simple mechanism of linear association
- **Linear vs nonlinear**
  - Linear outperform
  - Nonlinear activations fails to capture the assumed nonlinear association



Learning rate

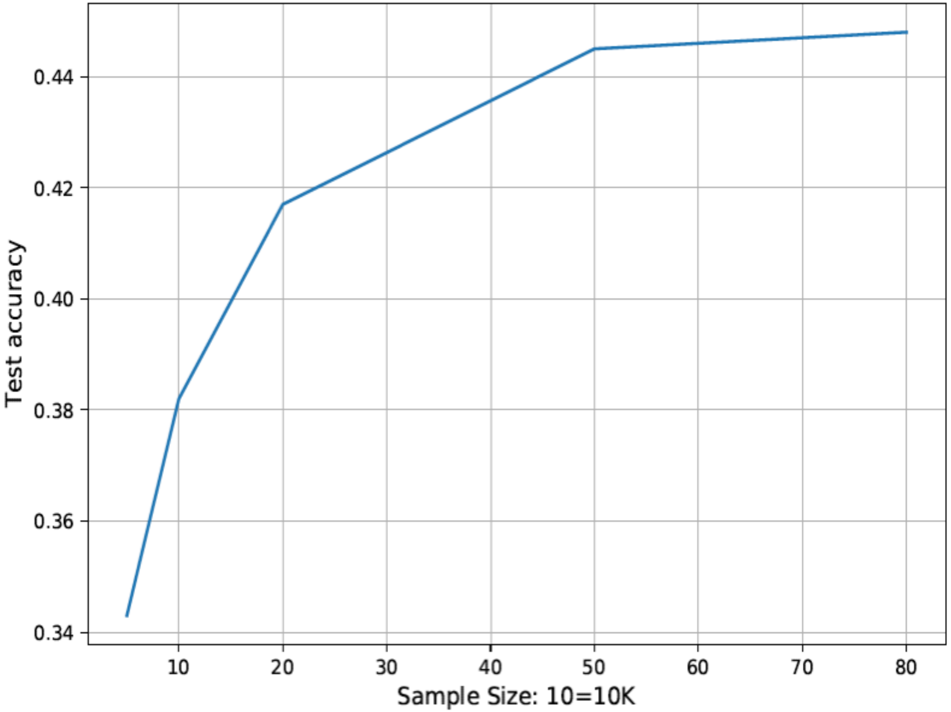
- **Slow down the speed of overfitting**



Regularization

- **Large lambda**
  - More penalize
  - offer protection from overfitting

- **Large sample**
  - Prediction accuracy is maintained after overfitting



Sample Size

- **Large Sample Size**
  - Obviously increase performance

Prediction Accuracy

- Selected SNPs >> Random SNPs
- BGLM >> NN
- NN: Linear >> Other Activations

	NN (linear)	NN (relu)	BGLM
Random, 6K	0.137	0.140	0.161
Top P value, 6K	0.452	0.442	0.459
Block BIC, 6K	0.449	0.440	0.457

Experience

Experience

- Variable Selection
  - Adjusted for age and sex; not adjusted for PCs
  - Unexpected: stepwise selection didn’t outperform top p-value selection for our case
- Prediction
  - For height, or Gaussian trait (highly additive), linear activation works well.
  - Large sample size + appropriate regularization is robust against overfitting
  - So far, generative linear model works better than neural networks

Thank you!