Genetic risk prediction: the role of SNP pre-selection in the polygenic score approach and shrinkage methods

Yiming Li[1], Timothy S.H. Mak[2], Johnny S.H. Kwan[1], Desmond D. Campbell[1,2], Pak C. Sham[1,2]

[1] Department of Psychiatry, the University of Hong Kong; [2] Centre for Genomic Sciences, the University of Hong Kong

Introduction

- Complex diseases are fundamentally determined by **genetic** and **environmental** factors.
- Estimating individual disease risk based on genotype data (**genetic risk prediction**, GRIP) has gained great interest.
- An accurate and efficient GRIP method has two steps –
  ① **select genetic markers for use**; and
    - *"The **pre-selection** process", often neglected.*

  ② **develop a GRIP model integrating the selected markers**.
    - *Plenty of those have been proposed in the past literature.*

Methods and Results

In this poster, we conduct simulation studies to compare the performances of the **"oracle predictor"** (the vector of effect sizes is known) and **different GRIP methods**, i.e. the polygenic score approach (PGS) and various shrinkage methods. The results are shown in **Table 1**.

Our SNP pre-selection procedure (step ①):

I. **Linkage disequilibrium based SNP pruning**.
   - *SNPs are selected according to a **pruning correlation**.*

II. **Per SNP chi-square tests of association**.
   - *SNPs are selected according to a **significance threshold**.*

The results (**Figure 1**) show that PGS could benefit from a more stringent pre-selection threshold, whereas the shrinkage methods, especially LASSO, perform better when no pre-selection is conducted.

Methods and Results (Cont'd)

Generally speaking, when the number of causal SNPs (*noc*) is small, LASSO outperforms the other methods with a performance comparable to the oracle predictor. Whereas when *noc* is large, ridge regression has the best performance, which is nevertheless still not very satisfactory. Better models need to be developed to tackle the large *noc* scenario.

Number of causal SNPs	Prevalence	Case	Method	AUC
10	0.1	Worst case	PGS	0.5625
10	0.1	Best case	EN	0.7646
10	0.1	Oracle	N/A	0.7684
10	0.5	Worst case	PGS	0.6172
10	0.5	Best case	LASSO	0.7430
10	0.5	Oracle	N/A	0.7578
100	0.1	Worst case	PGS	0.5763
100	0.1	Best case	LASSO	0.6887
100	0.1	Oracle	N/A	0.8095
100	0.5	Worst case	PGS	0.6075
100	0.5	Best case	LASSO	0.6768
100	0.5	Oracle	N/A	0.7473
1000	0.1	Worst case	LASSO	0.5513
1000	0.1	Best case	RR	0.6170
1000	0.1	Oracle	N/A	0.7814
1000	0.5	Worst case	PGS	0.6111
1000	0.5	Best case	RR	0.6416
1000	0.5	Oracle	N/A	0.7509

Table 1 The performances of ridge regression (**RR**), elastic net (mixing parameter = 0.5) (**EN**), **LASSO**, and the polygenic score approach (**PGS**) in GRIP compared with the **oracle** predictor. The area under the receiver operating characteristic curve (AUC) is used to evaluate their performances, and ten-fold cross validations are performed.

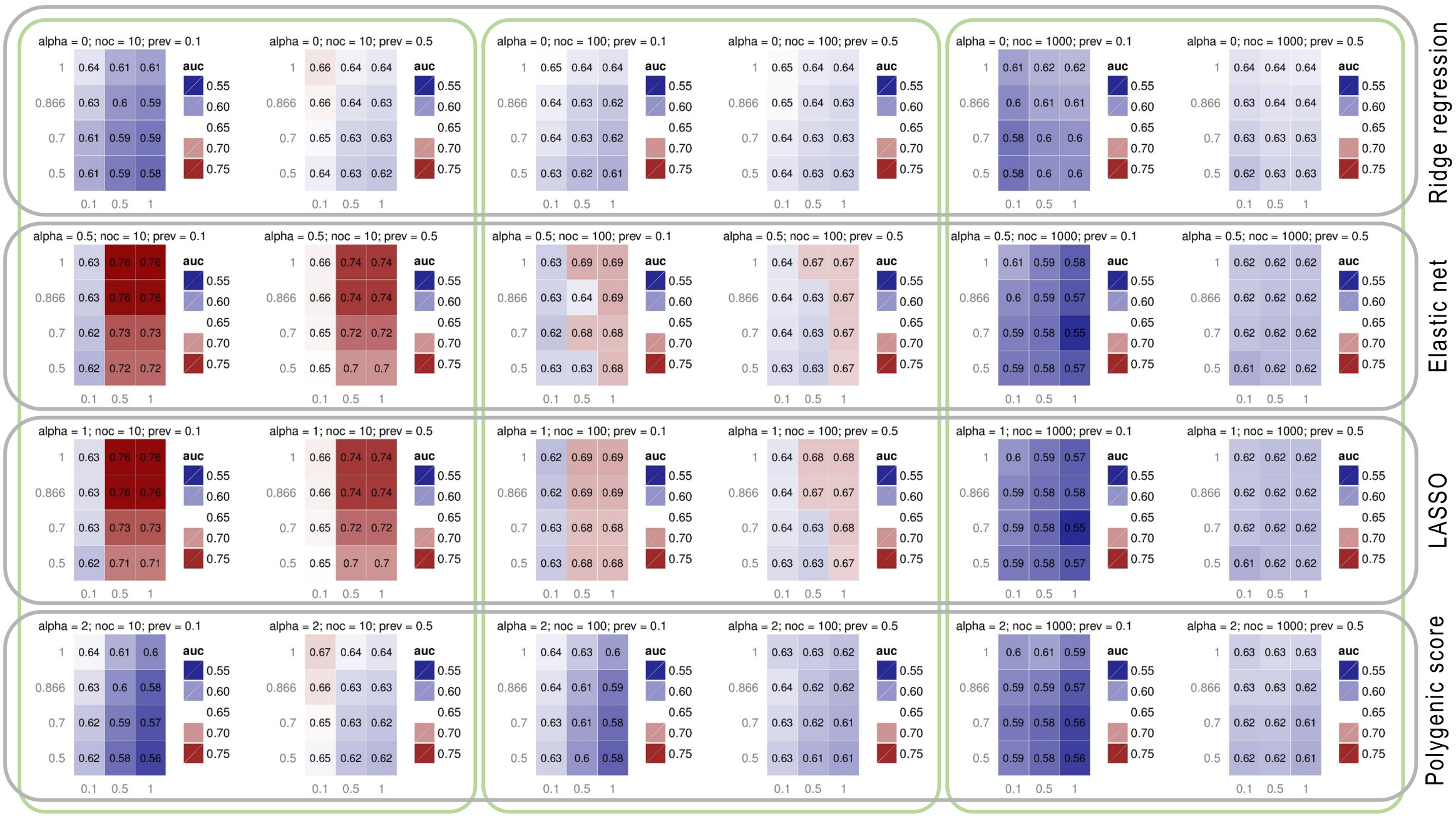

Figure 1 The performances of ridge regression, elastic net (mixing parameter = 0.5), LASSO, and the polygenic score approach in GRIP on simulated data assessed by the AUC. In each small heatmap, the x and y coordinates are corresponding to "pre-selection" parameters – the y-axis is the SNP pruning correlation we use whereas the x-axis is the significance threshold we apply along with the per SNP chi-square tests of association.