Weekly logs of Master thesis

Zhang Leyi[1]

[1]MSc in BiRC, Aarhus University, Aarhus, Denmark

2023-02

Contents

1 Week 1

Date: 2023/2/8

1.1 Software Download

REGENIE: Stacked block ridge regression method for Mixed Linear Model.

Advantages:

1. Fast and Memory friendly.

2. Can process both quantitative and binary traits.

3. When Case-control unbalanced, h^2_{SNP} will get too high due to the too low MAC(minor allele count). REGENIE can fix this.

Plink

Bolt-LMM

GEMMA

SAIGE problems occur

1.2 QC

Filter out SNPs with genotype missingness $> 10\%$, samples with $> 10\%$ missingness, MAF $< 5\%$, minor allele count(MAC) < 100.

Cmd:

./ software / plink ——bfile data ——geno 0.1 ——mind 0.1 ——maf 0.05 ——mac 100 \
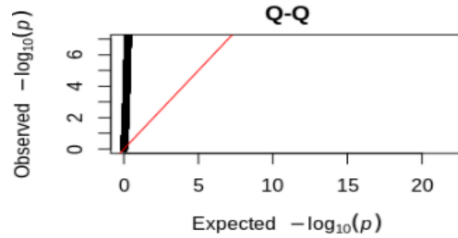——make—bed ——out data_qc

Output: data_qc

1.3 Plink

Worst result ever.

Cmd:

```
./software/plink --bfile data_qc --linear --pheno height.pheno --allow-no-sex \
--out data_plink_height
```

Q-Q plot: Figure 1

(a) Manhattan plot of REGENIE on height

Figure 1: **QQ plot of plink**

1.4 REGENIE for association study

I took height as the phenotype.

Step 1:

Input files: data_qc(3 files), covar1.covars, height1.pheno. Since REGENIE required labels in each column, I modified covar.covars and height.pheno with labels.

Cmd:

```
regenie \
--step 1 \
--bed data_qc \
--covarFile covar1.covars \
--phenoFile height1.pheno \
--bsize 100 \
--out data_regenie_out
```

Output: A predicting matrix W for h^2_{SNP}, in file: data_regenie_out_pred.list

Step 2: Association test and LRT

Cmd:

```
regenie \
--step 2 \
--bgen data_qc.bgen \
```
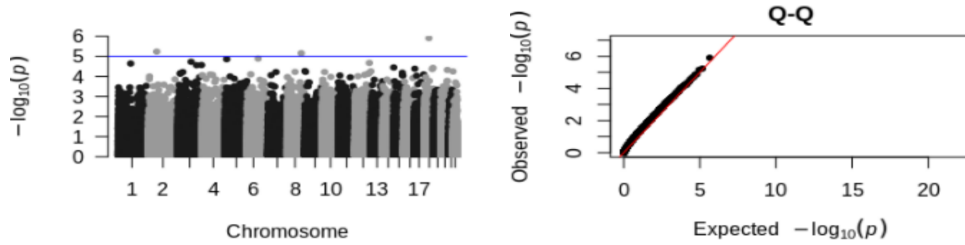
```
——covarFile covar1.covars \
——phenoFile height1.pheno \
——bsize 200 \
——qt \
——firth ——approx \
——pThresh 0.01 \
——pred data_regenie_out_pred.list \
——out data_regenie_out_firth
```

Output: data_regenie_out_firth_Phenotype.regenie

The result can be seen in Figure 2

(a) Manhattan plot of REGENIE on height   (b) QQ plot of REGENIE on height

Figure 2: **GWAS results from REGENIE on height(a)**Manhattan plot. **(b)**QQ plot.

Benchmark:

See in Table 1

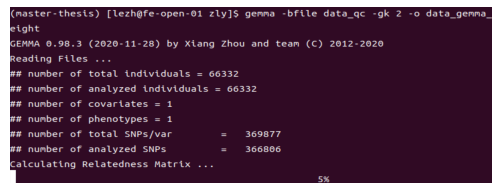| Method	Step	CPU time	Elapsed time(s)	Memory usage(GB)
REGENIE(null)	1		1492.8	10
REGENIE-Firth	2		4992.47	0.289

Table 1: **Computational performance of REGENIE-Firth**

Refer to: Mbatchou et al. (2021)

1.5   GEMMA

1. Before using GEMMA, the 6th column of .fam should be replaced by real phenotypes.

2. Calculate Kinship Matrix

```
gemma −bfile data_qc −gk 2 −o data_gemma_height
```

While it stuck here for hours, as in Figure

(a) Problem with GEMMA

Figure 3: **Problem with GEMMA**

1.6 fastGWA

1. After download, see: **gcta** in ./software folder

2. GCTA-GRM: calculating the genetic relationship matrix (GRM) from all the autosomal SNPs:

```
./software/gcta −−bfile data_qc −−chr 1 −−maf 0.01 −−make−grm \
−−out data_qc_chr1 −−thread−num 10
./software/gcta −−bfile data_qc −−chr 2 −−maf 0.01 −−make−grm \
−−out data_qc_chr2 −−thread−num 10
...
./software/gcta −−bfile data_qc −−chr 22 −−maf 0.01 −−make−grm \
−−out data_qc_chr22 −−thread−num 10
```

Output: .grm.bin, .grm.N.bin, .grm.id

3. To generate a sparse GRM from SNP data:

```
./software/gcta −−bfile data_qc −−autosome −−maf 0.01 −−make−grm \
−−out data_qc_gcta −−thread−num 10
./software/gcta −−grm data_qc_gcta −−make−bK−sparse 0.05 \
−−out sp_grm_gcta
```

4. Association study

I didn't use PCs

```
./software/gcta −−bfile data_qc −−grm−sparse sp_grm_gcta \
−−fastGWA−mlm −−pheno height.pheno −−qcovar covar.covars \
−−thread−num 10 −−out data_fastgwa_height
```

Problem comes from step 4, Figure 4:


```
Reading the sparse GRM file from [sp_grm_gcta]...
After matching all the files, 66151 individuals to be included in the analysis.
Estimating the genetic variance (Vg) by fastGWA-REML (grid search)...
```

(a) Problem with GEMMA

Figure 4: **Problem with fastGWA**

It is stuck here also for hours, is it normal? I'll open the computer for the night and see.

Still stuck there in the morning...

1.7   Bolt-lmm

./ software /BOLT–LMM_v2.4/ bolt —bfile=data_qc —phenoFile=height1 . pheno —phenoCol=I
—covarFile=covar1 . covars —qCovarCol —lmmForceNonInf —statsFile=data_bolt_height

The problem occurs at Figure 5.


```
Reading bed file #1: data_qc.bed
    Expecting 6117025826 (+3) bytes for 66151 indivs, 369877 snps
ERROR: Wrong file size or reading error for bed file: data_qc.bed
```

(a) Problem

Figure 5: **Problem with Bolt-LMM**

1.8   Question

1. What should I do with CV and LOOCV? What for?

2. How can I detect the Type I errors? By prediction? 3. When I tried SAIGE, I found it need me to use Rscript, how to do?

```
Rscript  createSparseGRM.R           \
—plinkFile=${LD pruned PLINK file } \
—nThreads=72   \
—outputPrefix=${OUTNAME}           \
—numRandomMarkerforSparseKin=5000         \
—relatednessCutoff=0.05
```

2 Week 2

2.20 - 2.26

Since there are a lot problems occuring with UKBB data, I use 1000g as a demo. If it can work, I'll run on UKBB.
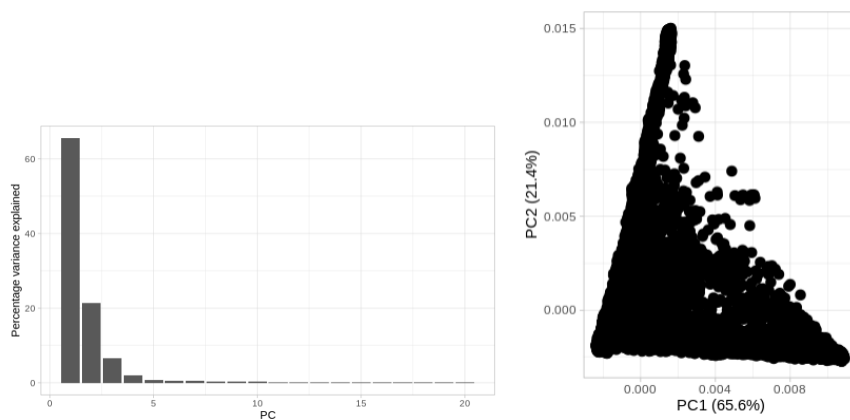
See in

2.1 fastGWA, 1000g
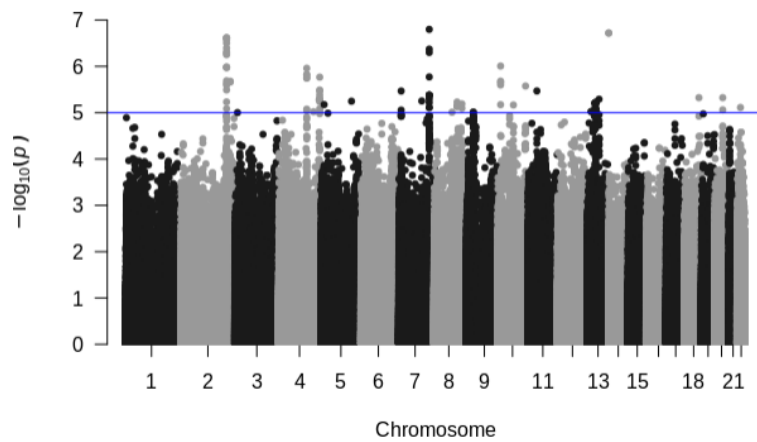
(a) Manhattan plot of fastGWA on 1000g, qt
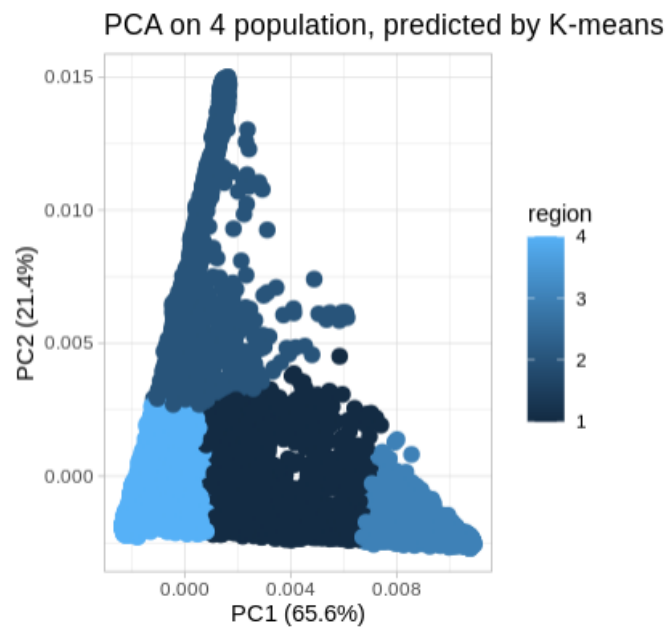
Figure 6: **fastGWA on 1000g**

2.2 PCA

(a) PCs explained

(b) PCA

Figure 7: **GWAS results from REGENIE on height**

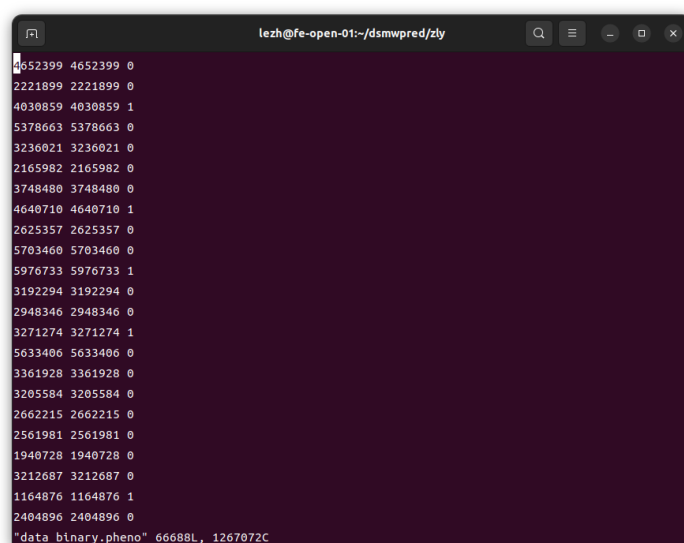## 2.3 PCA for population prediction, K-means

$y <-> PC1 + PC2$

(a) 4 populations

Figure 8: **K-means clustering on PCA**

Problem: I cannot evaluate the result.

2.4 Simulate Binary Phenotype with LDAK

(a) ukbb, simulation

Figure 9: **Binary Phenotypes UKBB simulation**

2.5 Regenie for UKBB, binary
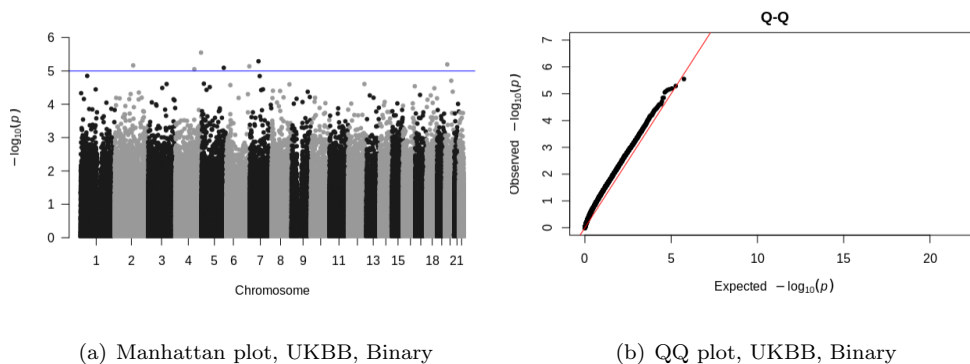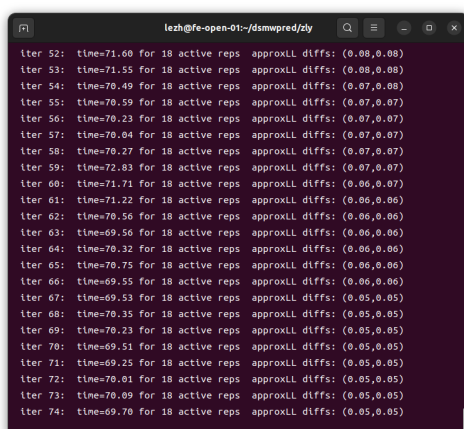
(a) Manhattan plot, UKBB, Binary

(b) QQ plot, UKBB, Binary

Figure 10: **Regenie results of UKBB on binary traits**

2.6 Bolt-LMM, 1000g, awaiting

(a) CV processing

Figure 11: **Bolt result awaiting**

The code has run for 24 hours, not consistent with the brilliant result from the paper, Loh et al. (2015).

2.7 Ancestry Inference
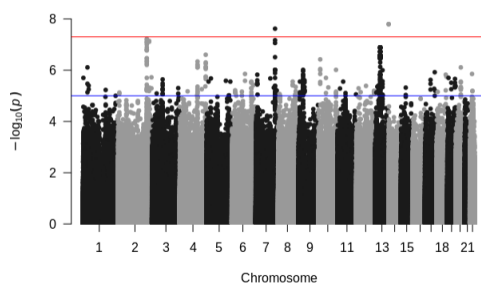
1. as shown in the section 2.3 [PCA for population prediction, K-means].

2. Using software: ADMIXTURE and fastSTRUCTURE for clustering.

Update: 2023/2/23

SNP Number: 6M

Time elapsed: 97449.4 sec

2.8   Bolt-LMM, 1000g, result

(a) Bolt result manhattan

(b) Bolt time elapsed

Figure 12:   **Bolt result and performance**

References

Loh, P.-R., Tucker, G., Bulik-Sullivan, B. K., Vilhjálmsson, B. J., Finucane, H. K., Salem, R. M., . . . Price, A. L. (2015, 3). Efficient bayesian mixed-model analysis increases association power in large cohorts. *Nature Genetics*, *47*, 284-290. DOI: 10.1038/ng.3190

Mbatchou, J., Barnard, L., Backman, J., Marcketta, A., Kosmicki, J. A., Ziyatdinov, A., . . . Marchini, J. (2021, 7). Computationally efficient whole-genome regression for quantitative and binary traits. *Nature Genetics*, *53*, 1097-1103. DOI: 10.1038/s41588-021-00870-7