GWAS PCA investigation (title pending)

Alex, Yiqi

September 27, 2019

1 Introduction

Genome-wide association study (GWAS) has been widely used to investigate whether target Single Nucleotide Polymorphism (SNP) is associated with certain trait (association study). However, considering admixture population is involved in recent study, linkage disequilibrium (LD) exists due to the chromosomal segments of different sub-population's (k) ancestry [1]. The existence of linkage disequilibrium or failure to correct for population structure can reduce statistical power. Hence, principal component analysis (PCA) which is a dimensionality-reduction method is used to provide examination for admixture population to identify the causal locus [2,3]. Although in recent study PCA has been a standard method to investigate the GWAS, doubts are cast on its statistical power when comparing with other existing implementation such as linear mixed model (LMM) [4]. Since current studies mainly focus on simple simulations or observation on real data [4, 5], existing evaluation can be limited to fully investigate the statistical power of PCA. In this paper, both simulation and real data set are used to evaluate the performance of PCA under different situations.

According to the result of our simulation,

2 Method

2.1 Theory connecting to kinship

2.2 Admixture Simulation

The construction of admixture population is mainly based on admixture simulation of Alex (2016). The related code of admixture simulation has been uploaded to Github with a R package called "bnpsd". Some parameters are changed in order to better simulate under

different situation. According to Alex (2016), the number of independent loci is 30000, in this paper the number of independent locus is 10000. The default value of Alex (2016) is 3, whereas in this project, it can be variable. Considering the difference among number of sub-population, the sample size of i_{th} sub-population will be set as the smaller integer of the ratio of total sample divided by number of sub-population.
Regarding the family structure, the generation will be set to be 20 to simulate admixture population with the existence of family structure. Considering the large calculation cost during the generation process of admixture population with family structure, this data set will be treated as real data set. It will be used repeatedly to test performance of PCA under different situations with random traits.

2.3 Trait Simulation

The construction of trait simulation is based on a R package called "simtrait". This package constructs the complex trait simulation with user-defined causal loci and the desirable heritability of the trait. It can be used in both simulated data set and real data set if the kinship matrix is estimated correctly.

2.4 Result Examination Method

In this paper, precision-recall curves (AUC) and uniformity p-value test (RMSD) will be used to test the performance of PCA under different scenarios. Both two methods will be illustrated by boxplot.

2.4.1 Precision-Recall Curves (AUC)

Precision-recall curve (AUC) is a plot whose y-axis represents precision and x-axis represents recall. Precision is calculated as the number of true positives divided by sum of both true positives which indicates the performance of model in predicting positives. Similarly, recall measures the ratio of number of true negatives and sum of true negatives and false negatives, which measures the performance of model in predicting negatives. The higher value of AUC, the better performance of statistical model.

2.4.2 Uniform P-Value Test (RMSD)

The For multiple independent and identical hypothesis tests, if the null hypothesis is true, the distribution of p-values will approximate to a uniform distribution [8]. In this paper,error

metric (RMSD) is used to measure the performance of PCA in terms of p-values. The RMSD is calculated as the root of mean square of the difference among sorted p-values of null hypothesis and expected quantiles of sorted p-values of null hypothesis after removing causal indexes.The numerical values of RMSD is inversely performance of PCA. The formula of RMSD is:

$$RMSD = \sqrt{(p_{uniform} - p_{null})^2}$$

where p_{null} is a list of p-values of null hypothesis after removing causal index and $p_{uniform}$ is a list of expected quantiles of p_{null} in uniform distribution.

2.5 PCA GWAS is Equivalent to linear regression with covariate

In some aspects, PCA GWAS can be deemed as equivalent to linear regression, based on result of principal component analysis. The trait in PCA GWAS is regressed by principal component of covariates rather than the set of covariates. Typically, the high eigenvalues of covarites corresponding to high variance will be selected in PCA GWAS.

2.6 Comparison among PCA and Existing Implementations

2.7 true or biased kinship matrices has the same performance

3 Result

3.1 RMSD Evaluation

3.1.1 No Family Structure

We compare the performances of PCA with different quantity of eigenvectors when the number of sub-population (k) equals to 10 and 50 separately in terms of RMSD. In this case, RMSD is used measure to find the optimal p under different situation. At each p, the simulations are repeated for 10 times in order to remove the extra variance.

According to the result of RMSD box-plot of PCA when k equals to 10, it can be seen that RMSD values remain relatively high when p is smaller than 9, which satisfies the actual rank of genotype matrix (k-1). It illustrates that the performances of PCA are bad of piece of eigenvalues is smaller than rank of genotype. Additionally, there exists an obvious tendency that RMSD values decreases as p increases before p increases to 9. Once p reaches 9, RMSD value jump downwards immediately. After that, RMSD values are relatively small and stable. It shows that the piece of eigenvalues can not impose extra statistical power of

PCA after p reaches the rank of genotype matrix

Similarly to previous result, RMSD box-plot indicates that pieces of eigenvalues should be at least the rank of genotype matrix in order to obtain a good performance of PCA. Hence, these two examples show that RMSD requires p to be no smaller than the true rank of genotype matrix or number of sub-population minus 1.

3.1.2 Family Structure Exists

3.2 AUC Evaluation

3.2.1 No Family Structure

The situation of AUC evaluation is slightly different from RMSD evaluations. Although AUC still requires the piece of eigenvalues to be no smaller than the rank of genotype matrix, our simulation indicates that the AUC value will decrease when number of eigenvectors used in PCA is extremely large while the true rank of kinship matrix in small or sample size is small. It means the excessively adding number of eigenvectors in PCA will not strengthen the performance of PCA in GWAS. When sample size is 1000, when PCs of eigenvectors used in PCA is among the interval from 1p to 100, there exists fluctuation in terms of AUC, but no obvious decreasing tendency. Then we test the AUC value with the same environment but increase PCs of eigenvector to 200 with 5 repeats. Compared with AUC values of PCs from 10 to 100, AUC values is much smaller when PCs is 200.
Considering the large calculation cost when PCs of eigenvectors used in PCA are large, we reduce the individual number from 1000 to 100. The reason for us to set the sample size to 100 is that it can better illustrate the tendency of AUC and save time for simulation. Based on the boxplot of AUC, it can be seen that the AUC will keep increasing when number of eigenvectors is smaller than 10. Once the PCs of eigenvectors reach 10, there exists a decreasing pattern of AUC though fluctuations exist. To better illustrate the pattern, we take absolute logarithm of AUC and then generate the boxplot again. In this case, the plot absolute logarithm of AUC has a bell-shaped and hence, it can be seen that the excessive use of eigenvector can impose negative influences on prediction accuracy.

3.2.2 Family Structure Exists

4 Discussion

4.1 PCA GWAS fails without enough PCs

According to the result of our simulation, it can be seen that PCA perform poorly when the number of PCs is smaller than the ranks of genotype matrix or kinship matrix. When the number of PCs is smaller than rank of kinship matrix and genotype matrix, RMSD values will be quite high which indicates that there is no significant association among alleles and trait. In addition, AUC values will very low indicating bad performance of PCA in predicting positives.

4.2 PCA GWAS still works even too many PCs are used

4.3 PCA GWAS fails with the existence are close relatives

5 Reference List

1$Pasaniuc, B., Zaitlen, N., Lettre, G., Chen, G.K., Tandon, A., Kao, W.L., ...\&Larkin, E.$(2011).$Enhance$
$assessmentusingAfricanAmericansfromCAReandaBreastCancerConsortium.PLoSgenetics$, 7(4), $e10$

2$Bryc, K., Auton, A., Nelson, M.R., Oksenberg, J.R., Hauser, S.L., Williams, S., ...\&Bustamante, C.1$
$widepatternsofpopulationstructureandadmixtureinWestAfricansandAfricanAmericans.Proceedingso$
791.

3$Price, A.L., Weale, M.E., Patterson, N., Myers, S.R., Need, A.C., Shianna, K.V., ...\&Goldstein, D.E$
$rangeLDcanconfoundgenomescansinadmixedpopulations.TheAmericanJournalofHumanGenetics$, 83(
135.

4$Wang, K., Hu, X., \&Peng, Y.$(2013).$Ananalyticalcomparisonoftheprincipalcomponentmethodandthe$
9.

5$Ochoa, A., \&Storey, J.D.$(2016).$FSTandkinshipforarbitrarypopulationstructuresI : Generalizedde$
6$Martin, A.R., Gignoux, C.R., Walters, R.K., Wojcik, G.L., Neale, B.M., Gravel, S., ...\&Kenny, E.E$
649.

8$SimonsohnU, NelsonLD, SimmonsJP.P-curve: akeytothefile-drawer[J].Journalofexperimenta$
$General, 2014, 143(2): 534.$