GWAS PCA investigation (title pending)

Alex, Yiqi

October 25, 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 PCA GWAS is Equivalent to linear regression with covariate

The original model in our project has the same structure to linear regression. We assume that in this model, there are n individuals and m genetic marker. The formula can be written

as:

$$Y = \mu + X\vec{\beta} + \vec{\epsilon}$$

. Here, Y is a n*1 vector which represents trait value for each individual and X is a n*m genotype matrix. In addition, $\vec{\beta}$ and $\vec{\epsilon}$ are a n*1 vectors representing the coefficient of genetics marker and residuals separately. Here ϵ follows a normal distribution $N(0, \sigma)$. This model sometimes fails beacause the number of genetic market is much larger than the number of indivudual. Then, we introduce PCA to make approximations

In a PCA linear regression model, we can write it in the form of

$$Y = \mu + \vec{X_j}\beta_j + U_{1:i}\nu + \vec{\epsilon}$$

Here Y still repesents the numerical value of trait of different individual and μ is the intercept. Both of them are n*1 vector. Meanwhile, $\vec{X_j}$ is a n*1 vector of j_{th} genetic marker and in this csae, β is a scale regression coefficient. Then, $U_{1:i}$ is a n*i matrix which is the first i Principal Components and ν is a n*1 coefficient vector for $U_{1:i}$. In the end, ϵ represnets the residual which follows a normal distribution $N(0, \sigma)$ which is the same to previous model. Then, we need to test the significance of each gemetic marker. The null hypothesis is β_j equals to 0 and the model in this case can be written as $Y = \mu + U_{1:i}\nu + \vec{\epsilon}$ for j_{th} market. The alternative hypothesis is β_j does not equal to 0. Therefore, we will conduct the F test to investigate whether the reduced model can have the same statistical power. If the p-values is small, which indicates this market is associated with trait.

In our simulation, each individual will be test 100000 SNPs (Single Nucleotide Polymorphism) and each for each loci (in total 100000), we will conduct linear regression separately. The genotype of each loci will be combined with eigenvector of principal components matrix which is composed of eigenvectors we used in PCA.Regarding the eigenvector of principal component, it is calculated by decomposing the kinship matrix.

2.3 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.

2.4 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.
In our simulation,the function of trait can be written as

$$Y = G + \epsilon$$

, where G represents the effect of genotype and ϵ represents the noise. The noise follows a normal distribution with mean zero and variacne equals to one minus hertability rate times desired parametric variance factor of the trait which is 1 in default. To obtain the genotype effect, marginal allele frequency will be calculated first and then, SNP will be randomly selected as causal index with random SNP coefficients. Then coefficients of causal index will be scaled and centered to estimate the genotype effect, thereby obtaining the numerical value of traits of each causal loci.

2.5 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.

2.5.1 Precision-Recall Curves

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.

2.5.2 Uniform P-Value Test (RMSD)

Due to the existence of multiply hypothesis test in our simulation, the frequency of erroneous inferences will increase. Altough Bonferroni Correction can be used to deal with this problem, it can result in high false negative rate (FNR). The better strategy is to controlling the False Discovery Rate (FDR). FDR is calculated as the ratio of flase positives and the sum of true positives and false positives.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 case, a bette strategy is to use the q-value rather than p-value to control the FDR. Therefore, we should evaluate the performance of PCA by quantifying the distribution of p-value of null hypothesis. If distribution of p-vlaues of null hypothesis is significantly below the quantiles of uniform distribution, the resulting q-values are anti-conservative.Thus, we use root mean square deviation (RMSD) to measure the fittness of quantiles p-values to the expected quantiles of uniform distribution. 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.

In many other papers, genomic control is another popular approach to detect the existence of population stratification and written as λ_{GC} . The definition of λ_{GC} is median chi-square asscociation statistic which has one degree of freedom through SNPs divided by theoretical median based on the null hypothesis. Hence, if hte λ_{GC} close to 1, it indicates that there is no or little population stratification. However, if λ_{GC} is larger than 1 significantly, it means there exists population stratification, or other confounder factors. Genomic control λ_{GC} only use median to measure the existence of stratification, whereas, RMSD make full use of data which should be more powerful. Hence, in this paper, RMSD is used to measure the performance of PCA in terms of p-value (type 1 error).

2.6 Comparison among PCA and Existing Implementations

SNP is one package on Github which aims to accelerate the computation of PCA.

2.7 true or biased kinship matrices has the same performance

3 Result

We use simulation data where genotypes and trait will be simulated following procedure mentioned above. We first set the sample size to be 1000 and then, we reduce the sample size to 100 to investigate whether PCA still have similar performance under new scenario. We will conduct 10 times simulation so that the extra variance can be reduced. For each simulation, performance of PCA will be collected in terms of RMSD and AUC for PCs from 2 to 90. Also, real data set will also be introduced to test the performance of PCA and trait will be simulated in the same way to simulation data. Fianlly, We will test the performance of PCA when family structure exists with sample size equals to 1000.

3.1 RMSD Evaluation

3.1.1 No Family Structure & N=1000

RMSD is used to measure the approximation of distribution of p-values of null hypothesis to uniform distribution. If RMSD is low, it shows good control of FDR, whereas, high RMSD represents bad control of FDR. Results will be illustrated by boxplot which can demonstrate the distribution and tendency of test statistics.
According to the result of RMSD boxplot 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 or kinship matrix which is (k-1). It demonstrates that the distribution of p-values of null hypothesis deviates from the expected quantiles of uniform distribution and therefore, PCA fails to control FDR.Though the performance of PCA is relatively bad, there still exists an decreasing tendency. It indicates that when the number of PCs used in PCA is smaller than true rank of genotype matrix, PCA will benefit forom using more PCs in terms of controlling FDR. However, once the number of PCs used in PCA reaches the actual rank of genotype matrix or kinship matrix, the RSDM will junmp to a small value and in this case the RMSD will be XXX. It remains stable as the number of PCA increases.

3.1.2 No Family Structure & N=100

Here we reduce the sample size to 100 and in this scenario, there is not significant difference between boxplot of sample size equals to 1000 and boxplot of sample size equals to 100. PCA fails to control FDR when the number of PCs is smaller than rank of genotype matrix but there exists a decreasing tendency of RMSD. When the number of PCs no longer smaller

than true rank of genotype matrix, RMSD will be small and keep stable. Hence, PCA is roboust to sample size in terms of FDR or type 1 error.

3.1.3 Family Structure Exists & N=1000

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

PCA GWAS still works even though PCs of eigenvectors are excessively used. From the box b

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