Testing the effectiveness of principal components in adjusting for relatedness in genetic association studies

Yiqi Yao, Alejandro Ochoa

November 14, 2019

1   Introduction

The goal of a genome-wide association study (GWAS) is to identify loci whose genotypes are correlated significantly with a certain trait. An important assumption made by basic association tests is that genotypes at non-associated loci are drawn independently from a common allele frequency, so that they are in Hardy-Weinberg Equilibrium (HWE). However, HWE does not hold for structured populations, which includes multiethnic cohorts and admixed individuals. When approaches for association not designed for structured populations are incorrectly applied to structured populations, association statistics (such as χ^2) become inflated relative to the null expectation, resulting in greater numbers of false positives than expected (Devlin and Roeder, 1999; Astle and Balding, 2009).

Modern approaches for conducting genetic association studies with structured populations involve modeling the population structure via covariates. Such covariates may be inferred ancestry proportions (Pritchard et al., 2000) or transformations of these. Principal components analysis (PCA) represents the most common of these variants, in which the top eigenvectors of the kinship matrix are used to model the population structure (Price et al., 2006). These top eigenvectors are commonly referred to as Principal Components (PCs) in the genetics literature (the convention we adopt here), but it is worth noting that in other fields the PCs would instead denote the projections of the data onto the eigenvectors. Various works have found that PCs map to ancestry, and in GWAS work as well as ancestry but without the computational expense of inferring ancestry proportions.

The other dominant approach for genetic association studies under population structure is the Linear Mixed-effect Model (LMM), in which population structure is a random effect drawn from a covariance model parametrized by the kinship matrix. LMM and PCA share

deep connections that suggest that both models ought to perform similarly (Hoffman, 2013). However, many previous studies have found that LMM outperforms the PCA approach, although many evaluations are inconclusive or are limited to unrealistic population structures often with unrealistically low differentiation (Astle and Balding, 2009; Kang et al., 2010; Wang et al., 2013). Moreover, various explanations for why LMM outperforms PCA are vague and have not been tested directly. Since LMMs tend to be considerably slower than the PCA approach, it is important to understand when the difference in performance between these two approaches is outweighed by their difference in runtime.

In this work, we study the performance of the PCA method for GWAS, comparing it to a leading LMM approach, characterizing its behavior under various numbers of PCs and varying sample sizes, under a reasonable admixture model and a model with admixture and family structure. Our evaluation is more thorough than previous ones, directly measuring the uniformity of null p-values (as required for accurate FDR control via q-values; Storey, 2003; Storey and Tibshirani, 2003) and orthogonally measuring predictive power by calculating the area under precision-recall curves. We find that the performance of the PCA approach is favorable when sample sizes are large (at least 1,000 individuals), matching the performance of LMMs as long as enough PCs are used. Remarkably, the approach is robust even when the number of PCs far exceeds the optimal number, suggesting that the degrees of freedom is not a concern in reasonably large studies. However, for smaller studies (100 individuals) there is a more pronounced loss of power when the number of PCs exceeds the optimal number. Moreover, PCA fails to perform as well as LMM in the presence of family structure, which is a scenario where the problematic structure is not low-dimensional so PCA naturally cannot model it entirely. All together, our simulation studies provide clear criteria under which use of PCA results in acceptable performance compared to LMMs.

Missing:

- Proposed simulation to infer optimal number of PCs from real data, predict gap between PCA and LMM?

- ...

2 Methods

2.1 The additive quantitative trait model and PCA approximation

Suppose there are n individuals and m loci, $\mathbf{X}$ is their $n \times m$ genotype matrix, and $\mathbf{y}$ is the length-n (column) vector which represents trait value for each individual. The PCA approach

can be derived from the following additive linear model for a quantitative (continuous) trait:

$$\mathbf{y} = \mathbf{1}\alpha + \mathbf{X}\beta + \epsilon,$$

where $\mathbf{1}$ is a length-n vector of ones, α is the scalar intercept coefficient, β is the length-m vector of locus effect sizes, and ϵ is a length-n vector of residuals. The residuals are assumed to follow a normal distribution: $\epsilon \sim \text{Normal}(0, \sigma^2)$, for some residual variance parameter σ^2.

Typically the number of loci m is in the order of millions while the number of individuals n is in the thousands. Hence, the full model above cannot be fit in this typical $n \ll m$ case, as there are only n datapoints to fit (the trait vector) but there are $m+1$ parameters to fit (α and the β vector). The PCA model with p PCs corresponds to the following approximation to the full model, corresponding to a model fit at a single locus i:

$$\mathbf{y} = \mathbf{1}\alpha + \mathbf{x}_i\beta_i + \mathbf{U}_p\gamma_p + \epsilon,$$

where $\mathbf{x}_i$ is the length-n vector of genotypes at locus i only, β_i is the effect size coefficient for that locus, $\mathbf{U}_p$ is an $n \times p$ matrix of PCs, and γ_p is the length-p vector of coefficients for the PCs. This approximation is explained by first noticing that the genotype matrix has the following singular value decomposition: $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\mathsf{T}$, where $\mathbf{U}$ are its left singular vectors of $\mathbf{X}$, $\mathbf{V}$ are its right singular vectors, and $\mathbf{D}$ is a diagonal matrix of its singular values. Thus, in the full model we have $\mathbf{X}\beta = \mathbf{U}\gamma$, where $\gamma = \mathbf{D}\mathbf{V}^\mathsf{T}\beta$. The approximation consists solely of replacing $\mathbf{U}\gamma$ (the full set of n left singular vectors and their coefficients) with $\mathbf{U}_p\gamma_p$ (the top p singular vectors only, which constitutes the best approximation of rank p). Thus, the extra terms in the PCA approach approximate the polygenic effect of the whole genome, and assumes that the locus i being tested does not contribute greatly to this signal. An equivalent model can be arrived at after standardizing each locus of the genotype matrix, in which case $\mathbf{U}$ is also equal to the eigenvectors of the kinship matrix.

Statistical significance is tested as follows. The null hypothesis is $\beta_j = 0$ (no association). The null and alternative models are each fit (fitting the coefficients of the multiple regression, where β_j is excluded under the null while it is fit under the alternative). The resulting regression residuals are compared to each other using the F-test, which results in a p-value where here a two-tailed test is used. Note that many common PCA implementations trade the more exact F-test for a χ^2 test, which is simpler to implement but only asymptotically accurate. As this is a multiple hypothesis test, there are a large number of loci (m) tested for association, so it is best to control the FDR rather than setting a fixed p-value threshold. We recommend estimating q-values and setting a threshold of $q < 0.05$ so that the FDR is controlled at the 5% level.

2.2 Simulations

2.2.1 Genotype simulations from admixture and family models

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.

2.2.2 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. The rank of genotype effect equals to the number of true subpopulation minus one as we take intercept into consideration as well.

Say what $\pi_1 = m_{\text{causal}}/m$ below is.

2.3 Evaluation of performance

All of the approaches considered here are evaluated in two orthogonal dimensions. The first one—the RMSD_p statistic below—quantifies the extent to which null p-values are uniform, which is a prerequisite for accurate control of the type-I error and successful FDR control via q-values. The second measure—the area under the precision-recall curve—quantifies the predictive power of each method, which makes it possible to qualitatively compare the statistical power of each method without having to select a single threshold, and most importantly, overcoming the problem that methods may not have accurate p-values.

2.3.1 RMSD_p: a measure of p-value uniformity

From their definition, correct p-values (for continuous test statistics) have a uniform distribution when the null hypothesis holds. This fact is crucial for accurate control of the type-I error, and is a prerequisite for the most common approaches that control the FDR, such as q-values (Storey, 2003; Storey and Tibshirani, 2003). We use the Root Mean Square Deviation (RMSD) to measure the disagreement between the observed p-value quantiles and the expected uniform quantiles:

$$\text{RMSD}_p = \sqrt{\frac{1}{m_0} \sum_{i=1}^{m_0} \left(u_i - p_{(i)} \right)^2},$$

where m_0 is the number of null loci ($\beta_i = 0$ cases only), here i indexes null loci only, $p_{(i)}$ is the ith ordered null p-value, and $u_i = i/m_0$ is its expectation. Thus, $\text{RMSD}_p = 0$ corresponds to the best performance in this test, and larger RMSD_p values correspond to worse performance.

In previous evaluations, test statistic inflation has been used to measure the success of corrections for population structure (Astle and Balding, 2009). The inflation factor λ is defined as the median χ^2 association statistic divided by theoretical median under the null hypothesis (Devlin and Roeder, 1999). Hence, when null test statistics have their expected distribution, we get $\lambda = 1$ (same as $\text{RMSD}_p = 0$ above). However, any other null test statistic distribution with the same median results in $\lambda = 1$ as well, which is a flaw of this test that RMSD_p overcomes ($\text{RMSD}_p = 0$ if and only if null test statistics have their expected distribution). The $\lambda > 1$ case (gives $\text{RMSD}_p > 0$) corresponds to inflated statistics, which occurs when residual population structure is present. $\lambda < 1$ is possible but not expected for genetic association studies (also gives $\text{RMSD}_p > 0$). Note that λ only use the median of the null distribution, whereas the RMSD_p makes use of the complete p-value distribution to evaluate its uniformity, which is more accurate.

2.3.2 The area under the precision-recall curve

Precision and recall are two common measures for evaluating binary classifiers. Let c_i be the the true classification of locus i, where $c_i = 1$ for truly causal loci (if the true $\beta_i \neq 0$, where the alternative hypothesis holds), and $c_i = 0$ otherwise (null cases). For a given method and some threshold t on its per-locus test statistics, the method predicts a classification $\hat{c}_i(t)$ (for example, if t_i is the test statistic, the prediction could be $\hat{c}_i(t) = 1$ if $t_i \geq t$, and $\hat{c}_i(t) = 0$ otherwise). Across all loci, the number of true positives (TP), false positives (FP) and false negatives (FN) at the given threshold t is given by

$$\text{TP}(t) = \sum_{i=1}^{m} c_i \hat{c}_i(t),$$

$$\text{FP}(t) = \sum_{i=1}^{m} (1 - c_i) \hat{c}_i(t),$$

$$\text{FN}(t) = \sum_{i=1}^{m} c_i \left(1 - \hat{c}_i(t)\right).$$

Precision and recall at this threshold are given by

$$\text{Precision}(t) = \frac{\text{TP}(t)}{\text{TP}(t) + \text{FP}(t)} = \frac{\sum_{i=1}^{m} c_i \hat{c}_i(t)}{\sum_{i=1}^{m} \hat{c}_i(t)},$$

$$\text{Recall}(t) = \frac{\text{TP}(t)}{\text{TP}(t) + \text{FN}(t)} = \frac{\sum_{i=1}^{m} c_i \hat{c}_i(t)}{\sum_{i=1}^{m} c_i}.$$

The precision-recall curve results from calculating the above two values at every threshold t, tracing a curve as recall goes from zero (everything is classified as null) to one (everything is classified as alternative), and the area under this curve is our final measure AUC_{PR}. A method obtains the maximum $\text{AUC}_{\text{PR}} = 1$ if there is some threshold that classifies all loci perfectly. In contrast, a method that classifies at random (for example, $\hat{c}_i(t) \sim \text{Bernoulli}(p)$ for any p) has an expected precision ($= \text{AUC}_{\text{PR}}$) approximately equal to the overall proportion of alternative cases: $\pi_1 = \frac{1}{m} \sum_{i=1}^{m} c_i$. The AUC_{PR} was calculated using the R package `PRROC`, which computes the area by integrating the correct non-linear piecewise function when interpolating between points (Grau et al., 2015).

2.4 Comparison among PCA and Existing Implementations

Some current research argue that Linear Mixed Model (LMM) can outperform PCA in correcting for population strcuture. In Linear Mixed Model, population strcuture is treated as random effect. Gengxin and Hongjiang (2013) demonstrate that efficient mixed-model

association eXpedited (EMMAX) which is based on linear mixed model outperforms PCA in both the population cohort study and case-control study. In this project, LMM will be tested under the same admixture simulation and trait simulation to PCA and thus, comparison between LMM and PAC with different PCs used under different situations can be made. Additionally, PCA with no PC used is equivalent to Linear Regression (LM). We will also consider the performance comparision among PCA and LM.

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. Then, we will test the performance of PCA when family structure exists with sample size equals to 1000 and 100 separately. Finally, we test the performance of PCA with the existence of complex family structure and here we set the generation to be 20.

3.1 PCA performance under N=1000

According to Fig. 1A which is the boxplot of RSMD when the number of subpopulation is 10 and sample size is 1000, 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 value of RMSD is close to 0. It remains stable as the number of PCA increases. Apart from this, it can be seen that before the number of PCs reaches the rank of genotype matrix, the distance among minimum and maximun of RMSD is larger. It tends to be smaller after number of PCs used in PCA is larger than the rank of genotype matrix, which shows less fluctuations in terms

Boxplot of RMSD when n=1000
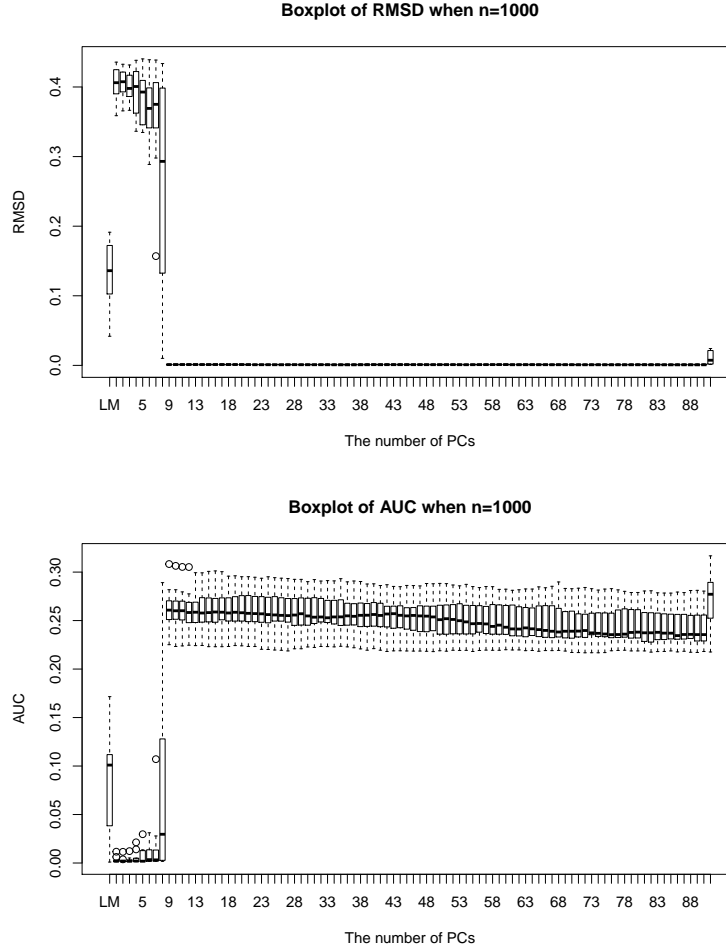
Boxplot of AUC when n=1000

Figure 1: **boxplot of rmsd and auc** The first panel is the boxplot of RMSD when sample size is 1000. Here the y axis represents value of rmsd and x axis is the number of PCs used in PCA. The second panel is the boxplot of AUC and here y axis is value of AUC. RMSD is used to measure the deviation of distribution of p values of null hypothesis from uniform distribution. In the case of multiple null hypothesis holds, the distribution of p-values of null hypothesis should approximate to unifrom distribution. The small value of RMSD implies type one error is well controlled. Regaring AUC, it's calculated by integration of predicion and recall curve. The value of AUC can be interpreted as the proportion of predictions made by this model is correct. The large AUC value implies better performance in predictive power. The result of Linear Regression is put at the position of PC equals to zero and the result of LMM is put at the position of PC equals to 91.

of type 1 error controlling. The RMSD value of LM is around 0.13 and it is smaller than PCA's RMSD value when PCs used in PCA is smaller than true rank of genotype matrix. However, once the number of PCs is larger than true rank of genotype matrix, it can be seen that PCA has better performance in type one error controlling. Compared with the performance of LMM, it can be seen that both PCA and LMM controll type one error well once the number of PCs used in PCA is larger than the true rank of genotype.

Fig. 1B illustrates the pattern of AUC for the same scenario to the first panel. In this case, we can see a increasing tendency for AUC when the number of PCs is samller than the true rank of genotype or kinship matrix. Such a tendency indicates that the increase of number of PCs can increase predictive power of PCA when it is smaller than the rank of genotype matrix. Once the number of PCs used in PCA exceeds the rank of PCA, the value of AUC will be increased to around 0.8, which indicates that 80% predictions of PCA are correct. After that, PCA does not benefit from adding more PCs as it can not increase the value of AUC obviously. Similarly, the range of AUC tends to be smaller than it when the number of PCs used in PCA is greater than rank of genotype matrix. It is clear that LM has better performance when Similar to the previous pandel, PCA outperforms LM when the number of PCs is no smaller than the true rank of genotype matrix. In terms of LMM, the average AUC value is slightly higher than AUC of PCA even the number of PCs is larger than true rank of genotype matrix.

3.2 PCA performance under N=100

Then, we want to investigate the performance of PCA when samlpe size is small. Here we set the sample size to be 100. From Fig. 2A it can be seen that the pattern is similar to the boxplot of RMSD in the case of sample size equals to 1000. It illustrates an decreasing when the number of PCs is smaller than true rank of genotype matrix. It indicates that PCA can better control type 1 error in this situation. If the number of PCs is greater than true rank of genotype matrix, the RSMD will decrease immediately and approximate to zero, which shows type one error is excellently controlled. Hence, PCA is insensitive to the smaple size in terms of RMSD or type 1 error controlling. Compared with the RMSD value of LM, RMSD value of PCA is smaller when the number of PCs is greater than 3. It implies that PCA control type one error better than LM. Moreover,the performance of LMM in this case is similar to Fig. 2 which is close to previous figure. The RMSD value of LMM approximates to 0 which indicates that there is no significant difference among PCA and LMM in this case.

Fig. 2B concerned with the boxplot of AUC, the pattern is obviously different from the

Boxplot of RMSD when n=100

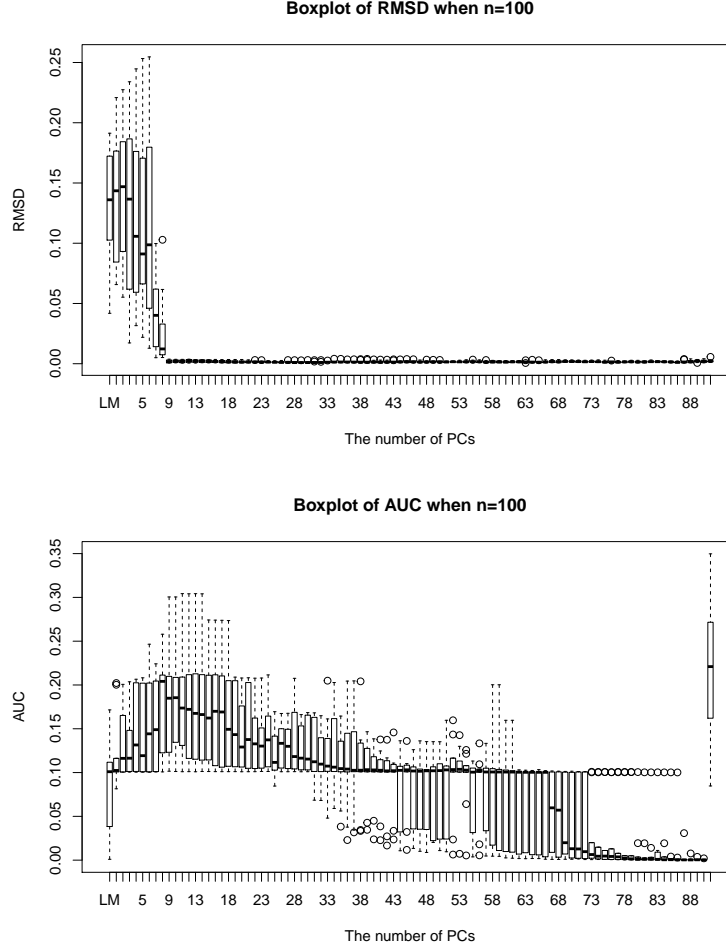**Boxplot of AUC when n=100**

Figure 2: **boxplot of rmsd and auc** The first panel is the boxplot of RMSD when sample size is 100. The pattern of panel A in here is similar to Fig. 1A. The converging value of RMSD here is also close to zero indicating that even in a small sample size, PCA is still robust to type one error controlling. Moreover, the LMM and PCA has similar RSMD values which demonstrates that these two methods have almost the same performance in type one error controlling. The pattern of Fig. 1B is bell-shapled. The AUC value reaches the peak when the number of PCs used in PCA equals to 9 and begins to decrease with the increase of PCs used. It illustrates that PCA fails in terms of power when the sample size is small. The decrease tendency is more obvious when the number of PCs is around 68 and it can be seen that LMM which is located at the last point of y-axis has better performance in power, compared with PCA and LM.

10

boxplot of AUC when smaple size is 1000. Although we still can find an increasing pattern of AUC before the number of PCs reaches the k. However, it can be seen that there is an decreasing pattern of AUC value once the number of PCs exceeds the true rank of PCA. It demonstrates that, even though fluctuations exist, performance of PCA in terms of power can be wrose with the increase of PCs used in PCA. The position of peak of AUC in this case is located at PCs equal to 9 which is exactly the true rank of genotype. Since we take the intercept into consideration, the true rank of genotype matrix equals to the number of subpopulation minus one. The bell-shaled distribution of AUC boxpolot implies that PCA receives punishment in power once the number of PCs used in PCA is larger than true rank of genotype matrix. Moreover, in terms of LM, it can be seen that even only one PC is used in PCA, the AUC value of PCA is still no worse than LM. However, when excessive PCs are used, performance of LMM can be better with a larger AUC value. Considering LMM, the average' AUC value is 0.25, whereas, the maximum AUC of PCA is 0.2 and it decreases immediately due to the excessive use of PCs. In this case, LMM performs better than PCA in terms of power.

3.2.1 PCA performance when familty structure exists

The introduction of family structure will make the original admixture populatio more complicated. In this case, we assume that there are 20 generations in total while other factors fixed. Based on the result of Fig. 3A indicates that there is a decreasing pattern of RMSD for the first three PCs. It addition, the value of RMSD in this situation is large, which indicates that the distribution of p-value of null hypothesis does not approximate to uniform distribution. It illustrates that PCA fails to control type one error when the number of PCs is not enough. When the number of PCs is larger than 4, values of RMSD beome stable around 0.05 which demonstrates that there exists evidence of the distribution of p-value of null hypothesis does not deviate from uniform distribution but the evidence is weaker than previous simulation where RSDM approximates to 0. Considering the range of RMSD, it decreases first and then begins to increase. It shows that excessive use of PCs will lead to extra variance. For AUC, it illustrates an increasing tendency for the first three PCs (Fig. 3B). When the number of PCs is larger than 3, AUC fluctuates around 0.2. This shows the power of PCA is small and hence, PCA's performance is not pretty well in this csae in terms of power even though enough PCs are used.

Boxplot of RMSD when n=1000 & family structure exists

RMSD

The number of PCs

Boxplot of AUC when n=1000 & family structure exists
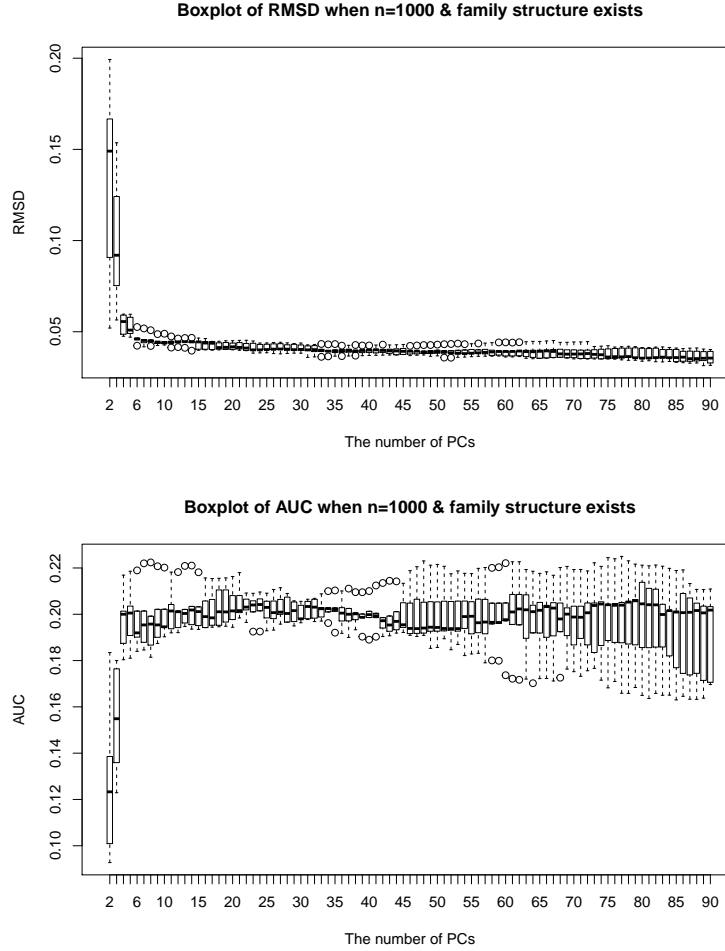

AUC

The number of PCs

Figure 3: **boxplot of rmsd and auc** The first panel is the boxplot of RMSD when sample size is 100. The pattern of panel A here is similar to Fig. 1A. The converging value of RMSD here is also close to one indicating that even in a small sample size, PCA is still robust to type one error controlling. The pattern of panel B is bell-shapled. The AUC value reaches the peak when the number of PCs used in PCA equals to 9 and begins to decrease with the increase of PCs used. It illustrates that PCA fails in terms of power when the sample size is small.

12

4 Discussion

4.1 PCA GWAS fails without enough PCs

Right now, both LMM and PCA have become standard approaches to correct for admixture population. In current PCA GWAS research, the number of PCs used in PCA is ususaly assmued to be 10. For instance, based on the simulation of Hoffman (2013), 10 PCs are randomly selected from the first 30 PCs. Furthermore, Wojcik et al. (2019) also performed PCA GWAS for 26 traits with first 10 PCs. According to the result of Wojcik et al. (2019), the correlation plot between SNP genotype and PC1–PC10 illustrates different populations has different correlations over some PCs. It seems that there exists a tradition that when PCA GWAS is performed, the top 10 PCs will be used. Since the result of PCA GWAS in this convention is good, right now most paprs will assume the number of PCs used in PCA GWAS to be 10. However, in this paper, we further investigate about the optimal choice of number of PCA under different population structures. In most current research, the number of subpopulation is smaller 10 and thus, based on results of previous three scenarios, the number of PCs used in PCA is enough which certifies the good performance of PCA in current research. For instance, based on the scatter plot of first two principal components with HapMap3 dataset which contains 11 populations, it can be divided into three subpopulations (Gad and Machiel, 2014). On the contraty, the lack of PCs will lead to the failure of PCA regardless of the sample size or family structure in both type one error and power. Seokho et al.(2012) state that considering the large number of single-nucleotide polymorphisms (SNP) used in GWAS to infer structure, it is necessary to remove SNPs that have negligible loadings in PCA. Whereas, our simulation indicates that if there is no enough PCs used in PCA GWAS, its performance can be bad as SNPs that have significant loadings are vanished from the analysis.

4.2 PCA GWAS still works even too many PCs are used

PCA GWAS is still robust even though PCs are excessively used for type one error controlling. However, for power, it will receive punishment if the number of PCs are excessive and the sample size is small. Jason and Anna (2009) point out that PCA with large sample size has better sample size than it with small sample size. This is because that PCA with large sample size can have smaller probability error and larger accuracy of population estimation (Jason and Anna ,2009). The small sample size in our research project only has 100 individuals in total. In a GWAS study, this is an extreme situation which it not realistic in research. The AUC boxplot of small sample size indicates that the peak of AUC is close to AUC

value for large sample size. This may result from that the number of causal loci in small sample size is reduced from 100 to 10. However, the boxplot of AUC has a bell shap with downward-sloping line on each side of the peak. It demonstrates that in small sample size, punishment of excessive use of PCs will come occur immdeiately. Considering in the study of GWAS, we will expect to use thousnds of SNPs and number of individuals are also much larger than 100, the situation of small sample size may not be common (Seokho et al., 2012). Hence, although PCA will fail when the number of PCs is much larger than the true rank of genotype matrix, it is still encouraged to use more PCs in PCA GWAS .

4.3 PCA GWAS fails with the existence are close relatives

(Price et al., 2010) point out that GWAS may fail in the case of dataset contains family structure. The result of our simulation also supports this argument. Fig. 3 shows that ...

4.4 LMM outperforms PCA

Wang et al. (2013) argues that mixed effects model is preferred in the case of existence cryptic relatedness but not population stratification. In their paper, only first four PCs are used in PCA and performance of PCA may be underestimated. Fruthermore, EMMAX which is a kind of linear mixed model is claimed to be better than PCA (Gengxin and Hongjiang, 2013). Based on the result of our simulations, it can be seen that in both large sample size and small sample size, LMM are slightly better than PCA in terms of power. When sample size is large such as the scenario in Fig. 1B, it can be seen that although both two methods's are not good. However, the maximum AUC value of PCA is around 0.25 and LMM's AUC value is slightly larger than PCA. In addition to this, in the case of small smaple size, advantage of LMM is more obvious. As mentioned before, PCA will receive punishment when the number of PCs is much larger than the true rank of genotype matrix. The maximum of AUC value of PCA in this scenario decreases to 0.2 and the average AUC value of LMM is around 0.24. This two simulations illustrates that

(Sul and Eskin, 2013) (Price et al., 2013)

(Tucker et al., 2014)

References

Astle, William and David J. Balding (2009). "Population Structure and Cryptic Relatedness in Genetic Association Studies". *Statist. Sci.* 24(4). Mathematical Reviews number (MathSciNet): MR2779337, pp. 451–471.

Devlin, B. and Kathryn Roeder (1, 1999). "Genomic Control for Association Studies". *Biometrics* 55(4), pp. 997–1004.

Grau, Jan, Ivo Grosse, and Jens Keilwagen (1, 2015). "PRROC: computing and visualizing precision-recall and receiver operating characteristic curves in R". *Bioinformatics* 31(15), pp. 2595–2597.

Hoffman, Gabriel E. (2013). "Correcting for population structure and kinship using the linear mixed model: theory and extensions". *PLoS ONE* 8(10), e75707.

Kang, Hyun Min et al. (2010). "Variance component model to account for sample structure in genome-wide association studies". *Nat. Genet.* 42(4), pp. 348–354.

Price, Alkes L. et al. (2006). "Principal components analysis corrects for stratification in genome-wide association studies". *Nat. Genet.* 38(8), pp. 904–909.

Price, Alkes L. et al. (2010). "New approaches to population stratification in genome-wide association studies". *Nature Reviews Genetics* 11(7), pp. 459–463.

— (2013). "Response to Sul and Eskin". *Nature Reviews Genetics* 14(4), p. 300.

Pritchard, Jonathan K. et al. (2000). "Association Mapping in Structured Populations". *The American Journal of Human Genetics* 67(1), pp. 170–181.

Storey, John D. (2003). "The positive false discovery rate: a Bayesian interpretation and the q-value". *Ann. Statist.* 31(6). Mathematical Reviews number (MathSciNet): MR2036398; Zentralblatt MATH identifier: 02067675, pp. 2013–2035.

Storey, John D. and Robert Tibshirani (2003). "Statistical significance for genomewide studies". *Proceedings of the National Academy of Sciences of the United States of America* 100(16), pp. 9440–9445.

Sul, Jae Hoon and Eleazar Eskin (2013). "Mixed models can correct for population structure for genomic regions under selection". *Nature Reviews Genetics* 14(4), p. 300.

Tucker, George, Alkes L. Price, and Bonnie Berger (1, 2014). "Improving the Power of GWAS and Avoiding Confounding from Population Stratification with PC-Select". *Genetics* 197(3), pp. 1045–1049.

Wang, Kai, Xijian Hu, and Yingwei Peng (2013). "An Analytical Comparison of the Principal Component Method and the Mixed Effects Model for Association Studies in the Presence of Cryptic Relatedness and Population Stratification". *HHE* 76(1), pp. 1–9.

Wojcik, Genevieve L. et al. (2019). "Genetic analyses of diverse populations improves discovery for complex traits". *Nature* 570(7762), pp. 514–518.