PCA and its applications in GWAS

Hengrui Luo

Department of Statistics

The Ohio State University

For STAT8810, 2018 and Based on

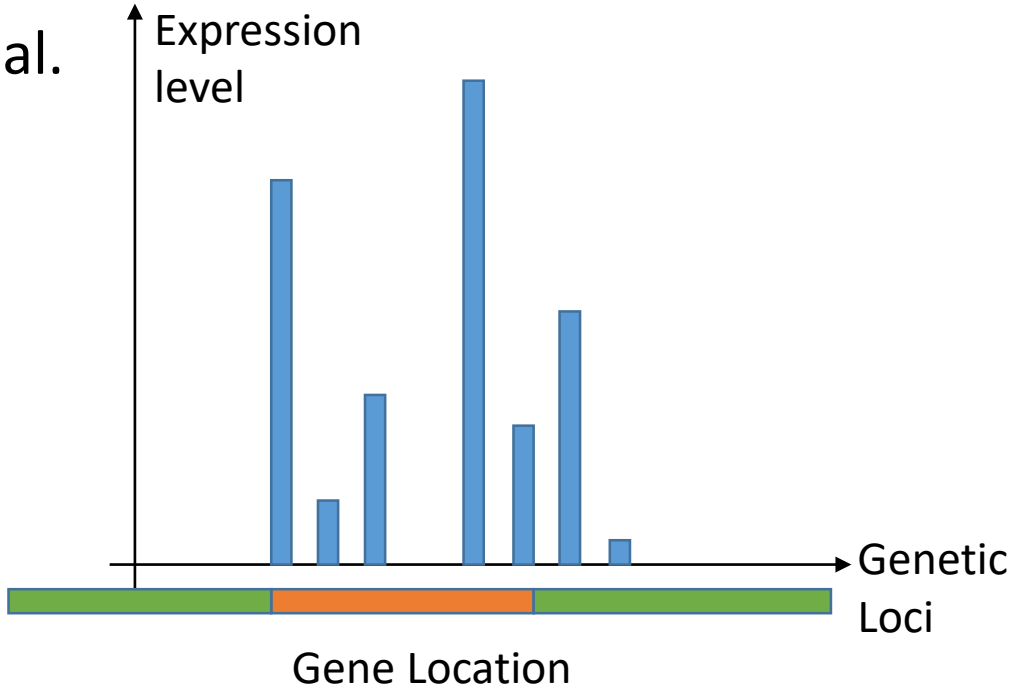
Aschard, Hugues, et al. "Maximizing the power of principal-component analysis of correlated phenotypes in genome-wide association studies." *The American Journal of Human Genetics*94.5 (2014): 662-676.

GWAS Data

- **G**enome-**W**ide **A**ssociation **S**tudies aims at associating the traits/phenotypes and genes/genotypes based on DNA sequencing data.
- The data is organized as gene expression levels X and (continuous or discrete) trait variables Y across different individuals.
  - Our goal is to determine whether the effect of the expression level of a certain gene on a given phenotype is statistically significant or not. In the current paper we use the minor allele frequency (MAF) of SNPs to represent the expression levels of locis.
- For instance, when genetic loci X usually has a very high expression level on individuals with disease A, then we can reasonably associate X with phenotype Y_A.
  - This can be formulated as a hypothesis testing problem about testing the significance of such an association when the statistical model for X and Y is assumed.

Challenges in GWAS

- Univariate analysis methods on each loci will inflat type I error when the traits are correlated. E.g. Fisher's exact test
- Multivariate analysis could increase the power of test detecting positively correlated traits.
  - Integration of correlated signals are nontrivial.
  - Some of these methods also assume relatedness of phenotypes.
  - Usually these methods are computationally intensive.



Competing Strategies

- Strategies for detection of genetic associations in correlated phenotypes
  - Regression models with designed correlation matrices. (Mixed effect models)
    - In this approach, the covariance structure of the joint model describes the correlation between phenotypes Y_A ,Y_B ,…. The covariance structure can also handle the population genetic heterogeneity across different populations when we want to combine data from different populations.
  - P-value correction of univariate analysis. (P-value approximation/correction)
    - Combine the p-values obtained from univariate tests for each phenotype. Meta-analysis will join all the p-values together and produce a “new” p-value after taking correlation into consideration.
  - Data reduction methods. (PCA)
    - Compose a linear combination of phenotypes Y with the highest correlations, which also maximizes the variance or heritability. This approach allows us to integrate those correlated data and potentially we can choose the PCs with “high variances” so that reduce the dimension of the data.

Fundamental Problem

- Problems: When we choose those PCs with high variances,
  - How do we choose a threshold to throw away those low variances?
  - How to combine associations across PCs and give an interpretation?
- Misconception: "…there appears to have been a growth in the misconception that the principal components with small eigenvalues will very rarely be of any use in regression." [Jolliffe]
  - (Significant low variance components) Many of the coefficients with low eigenvalues can be significantly different from zero because of the size of the estimates and the small value of noise.
  - (Correlated inside components) The component with the low variance can be highly collinear/correlated with the component with the high variance.
  - (PCs with similar explained variances) There can be many components with almost the same variances when the data is equally spread in each directions.
  - (Non-Gaussian) Non-Gaussian distribution will generate data with high variance in certain skewed direction but regular PCA requiring orthogonality inbetween components.
  - (Non-linearity) Non-linearity hidden in principal components will also hinder the performance of PCA. (Kernel PCA)
- Discarding the low variance components will decrease the power of detecting genetic variants associated with traits, since these components can capture substantial proportion of genetic variance.

Regression model with two phenotypes

- Model: Phenotype variables are described by a normal model [Aschard et.al] $\epsilon_i \sim N(0,1), i = 1,2$: $y_i = \sqrt{c} \cdot u + \sqrt{v_i} \cdot g + \sqrt{1 - c - v_i} \cdot \epsilon_i$ with unknown variable u and genotype variable g.

- PCA can provide two PCs in terms of linear combinations of u and g.

$$PC_1 = \frac{1}{\sqrt{2}}\left[(2\sqrt{c}) \cdot u + \left(\sqrt{v_1} + \sqrt{v_2}\right) \cdot g + \sqrt{1 - c - v_1} \cdot \epsilon_1 + \sqrt{1 - c - v_2} \cdot \epsilon_2\right]$$

$$PC_2 = \frac{1}{\sqrt{2}}\left[(0) \cdot u + \left(\sqrt{v_1} - \sqrt{v_2}\right) \cdot g + \sqrt{1 - c - v_1} \cdot \epsilon_1 - \sqrt{1 - c - v_2} \cdot \epsilon_2\right]$$

With corresponding explained variances (singular values)

$$v_{PC_1} = \frac{v_1 + v_2 + 2 \cdot \sqrt{v_1 v_2}}{2(1 + c + \sqrt{v_1 v_2})}, v_{PC_2} = \frac{v_1 + v_2 - 2 \cdot \sqrt{v_1 v_2}}{2(1 - c - \sqrt{v_1 v_2})}$$

Testing association

- Combined Method: Combine/Sum the Wald testing statistics
- Advantage: The test can be applied to any number of genes (dimension of g) and any number of traits(dimension of u).
- We conduct a Wald association test at level α with power

$$Power = 1 - F_{\chi_1^2(\delta)}(q_{\chi^2_{1,1-\alpha}})$$

on the hypothesis whether the explained variance of each PC is zero i.e. whether the specific PC explained the phenotype Y.

- With **orthogonality(i.e. independence) between PCs**, we can conduct a combined test, i.e. with higher power

$$Power = 1 - F_{\chi_2^2(\delta')}\left(q_{\chi^2_{2,1-\alpha}}\right)$$

Where δ' is the estimated explained variance explained by PC_1, PC_2 multiplied by number of individuals (dimension of Y)

Summary for bivariate case

- PC decomposition produces orthogonality, hence we can have independent PCs to conduct hypothesis testing about the association between PCs (linear combination of model variable like genes/unknown factors/noise). In the analysis of more than two phenotypes, combined test using PCs also prove to be powerful in most scenarios.
- When the PCs are independent, their testing statistics can be summed up and yields increased power. Instead of selecting "high-variance" components, we simply use all the information contained. In addition, in Table 1 of [Aschard et.al]:

Table 1. Rationale for Testing Genetic Association with PCs in a Bivariate Analysis

Genetic Model	PC_1	PC_2	Combined PCs
$\beta_{Y1} \neq 0$ and $\beta_{Y2} = 0$	almost no power, converging to 0 with increase correlation	most powerful for small sample size and lowβ_{Y1}; power increases with correlation	most powerful for large sample size and large β_{Y1}; power increases with correlation
β_{Y1} in the same direction as β_{Y2}	most powerful when correlation and $\beta_{Y.}$ are moderate	very low power; power increase slightly with correlation	most powerful when correlation and $\beta_{Y.}$ are high
β_{Y2} opposite to β_{Y1}	almost no power; minor variation with increase correlation	very powerful; power increase with correlation	very powerful; power increase with correlation

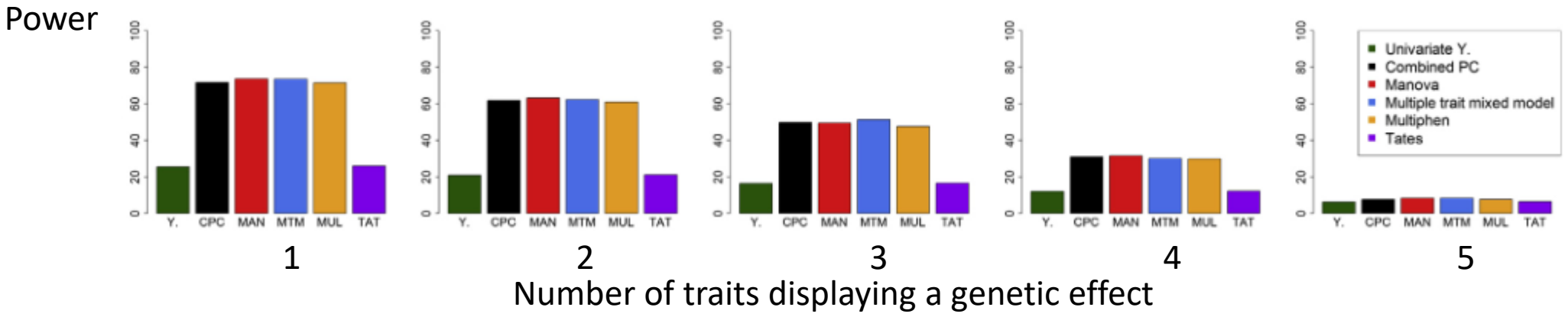
The two positively traits are denoted Y_1 and Y_2 and and genetic effect of G on Y_1 (β_{Y1}) and Y_2 (β_{Y2}).

More than two phenotypes

- The factor that affects the powers of the combined test in bivariate case:
  - Correlation between phenotypes Y_1, Y_2.
  - Ratio of $\frac{v_1}{v_2}$ in the assumed model.
  - Sample size of individuals available in data.
- Simulation shows that when a relatively small number of phenotypes are analyzed jointly the genetic association signals is usually spread across most or all of the PCs. **Therefore discarding low variance components will cause loss of information.**
- This does not have to be the case when there are a very large number of phenotypes jointly analyzed. Leading PCs may have good detection power. The large increase in degree of freedom may outweigh the benefit in combining small associations.
  - The optimal strategy depends on the model.
  - Analyzing the PCs based on grouping by eigenvalues magnitudes can improve power.

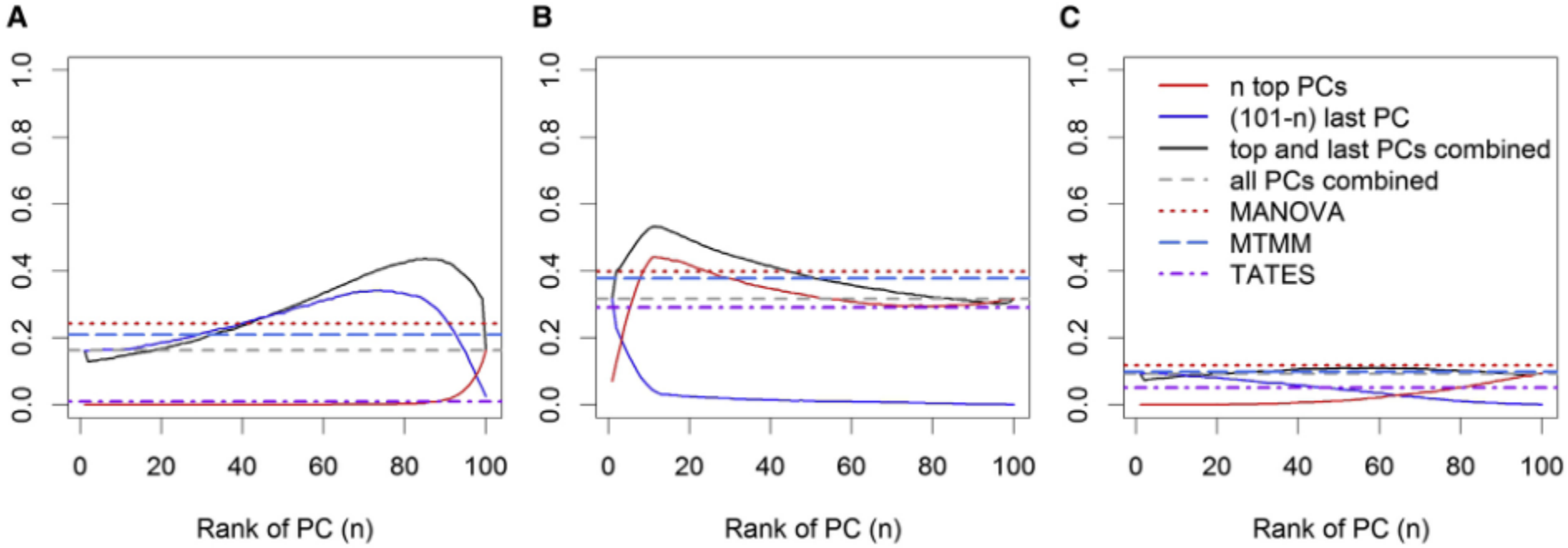
Comparison to other methods

- For more than two phenotypes, the multivariate analysis of 100 traits, the paper provides comparison study for different methods including univariate test, MANOVA, multi-trait mixed model, Multiphen model and p-value correction (TATES). The result shows that combined test achieve near optimal power under all these simulated realistic circumstances.



PCA with only high variant PCs are not optimal

- PCA based on the few components that explain the most variance in the data is not optimal.
- To this point, the paper showed with simulation evidence that (also in accordance with [Jolliffe]), in coagulation-related phenotype data all the phenotypes reflecting global coagulation activity has moderate/strong correlation. In this case the combined test of selected PCs (say the 1st and the last) even outperforms the joint test of top PCs. Also analysis shows that the combined test also outperforms other methods under different simulation schemes.



Extensions to other tests

- PC decomposition produces orthogonality, the independence between PCs can also be used to combine other univariate tests other than Wald test, as their rejection regions are shown Figure 1 of [Liu&Lin]

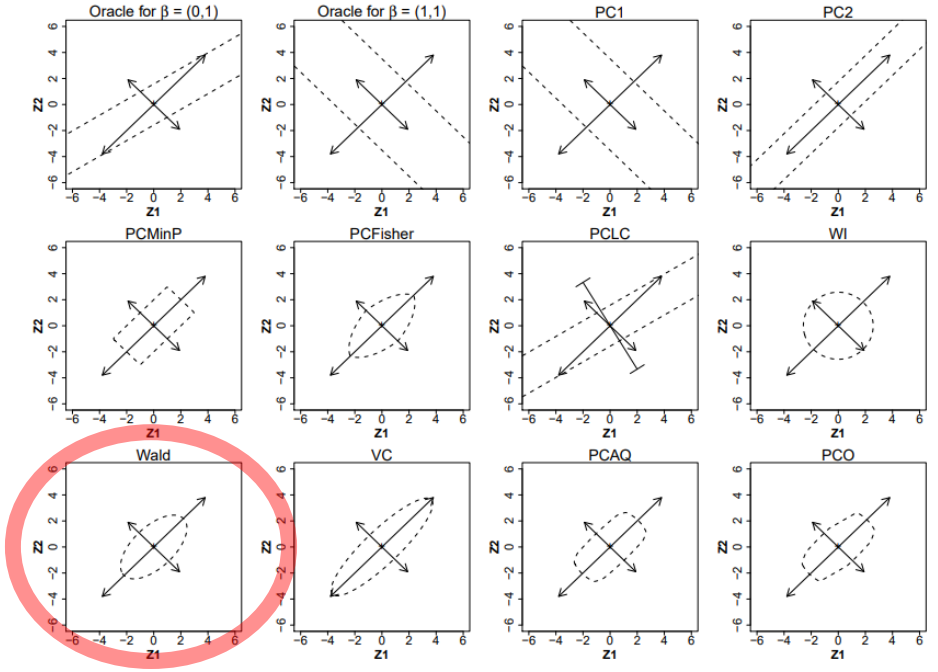


Figure 1: The rejection boundaries of the proposed PC-based tests for bivariate normal test statistics (Z_1, Z_2) with correlation equal to 0.6. The dashed lines or curves represent the boundaries that separate the acceptance and rejection regions at the significance level 0.05. The longer solid lines with arrows represent the PC_1 direction, the shorter solid lines with arrows represent the PC_2 direction, and the lengths of the longer and shorter solid lines with arrows are equal to $6\sqrt{\lambda_1}$ and $6\sqrt{\lambda_2}$ respectively. For PCLC, the added solid line with "T"-type arrows illustrates the direction for alternative β which is orthogonal to its rejection boundaries, where $\theta_1 = 76^o$ and $\theta_2 = 14^o$.

Summary and Result

- PCA is one of main approaches in the dimension reduction strategy in GWAS studies.
- PCA based on the few components that explain the most variance in the data is not optimal.
- PCA itself has intrinsic pitfalls that may not allow it to cover all possible situations occurred in sequencing data. Only include those PCs with high variances may simply cause information loss (as shown by the coagulation-related data)
- Based on the fact that PCs are orthogonal/independent, we can propose the combined test when there are more than one traits in our multivariate analysis for the linear model.
- In most realistic situations with few traits, the combined PCA testing will lead to increased power and include the least variable PC will help us detect more associations.
- When the dataset is large with many traits, the combined PCA may not out-compete naïve selection of few PCA, but multi-step approach (grouping) can be super-imposed to remedy combined PCA.

Reference

- Jolliffe, Ian T. "A note on the use of principal components in regression." *Applied Statistics* (1982): 300-303.
- Aschardet al. (2014) Maximizing the Power of Principal-Component Analysisof Correlated Phenotypesin Genome-wide Association Studies. *American Journal of Human Genetics* 94, 662–676,
- Liu, Zhonghua, and Xihong Lin. "A Geometric Perspective on the Power of Principal Component Association Tests in Multiple Phenotype Studies." *Journal of the American Statistical Association*just-accepted (2018): 1-36.