

# Graphic analysis of population structure on genome-wide Rheumatoid Arthritis data

Jun Zhang<sup>1§</sup>, Chunhua Weng<sup>2</sup>, Partha Niyogi<sup>3</sup>

1. Department of Radiology, The University of Chicago, 5841 South Maryland Avenue, Chicago, Illinois 60637
2. Department of Biomedical Informatics, Columbia University, 622 West 168 Street, New York, NY 10032
3. Departments of Statistics and Computer Science, The University of Chicago, 1100 E. 58th, Chicago, Illinois 60637

§ Corresponding author

Email addresses:

Jun Zhang: junzhang@uchicago.edu

Chunhua Weng: chunhua.weng@dbmi.columbia.edu

Partha Niyogi: niyogi@cs.uchicago.edu

## Abstract

Principal component analysis (PCA) has been used for decades to summarize the human genetic variation across geographic regions and to infer population migration history. Reduction of spurious associations due to population structure is crucial for the success of disease association studies. Recently PCA has also become a popular method for detecting population structure and correction of population stratification in disease association studies. Motivated from manifold learning, we propose a novel method based on spectral graph theory. Regarding each study subject as a node with suitably defined weights for its edges to close neighbors, one can form a weighted graph. We suggest using the spectrum of the associated graph Laplacian operator, namely, Laplacian eigenfunctions, to infer population structures instead of principal components (PCs). For the whole genomewide association study for North American Rheumatoid Arthritis Consortium (NARAC) provided by GAW16, Laplacian eigenfunctions revealed more meaningful structures of the underlying population, compared with PCA. The proposed method has connection with PCA, and it naturally includes PCA as a special case. Our simple method works computationally fast and is suitable for disease studies at the genomewide scale.

## Introduction

As is well known that unidentified population structure can cause spurious associations in genomewide association studies [1,2]. Such associations typically occur when the disease frequencies vary across subpopulations thereby resulting in that affected individuals are more likely to be sampled from overrepresented subpopulations. To correctly identify population structures from population genotype data is critical for the rapidly planned and carried out genomewide association studies. Though this topic has been extensively studied, the prevailing methods such as genomic control and structured association still have limitations [3]. Recently principal components has been employed to summarize the genetic background variations [4,5]. Price et al [3] suggested another solution to include a few top PCs as covariates in a regression setting. However, there is potential concern about the interpretation of PCs. Recently, John Novembre and Matthew Stephens [6] showed that the patterns such as gradients and waves appear in the PC analysis of continuous genetic data sometimes resemble sinusoidal mathematical artifacts that arise generally when PCs are applied to spatially correlated data. Nevertheless,

PCA can provide evidence of major demographic migration events and is still widely used in many contexts of genetic data analysis.

Here we propose a novel approach for detecting population structure motivated from graph theory. It is nonlinear and can reveal the local dependence structures of population samples. One regards cases and controls as vertices of a weighted graph [7] and each vertex is connected to its close neighbors in the sense of correlation via edges. This reflects the fact that distances between vertices that are far apart are usually meaningless, and therefore need not be preserved. The weight of the edge for each pair of individuals measures the degree of being correlatedness (see Method). The eigenfunctions of the associated graph Laplacian operator on the graph are generalized geometric harmonic functions [8], which contain useful geometric structure information of the population dependence graph. The eigenvectors of the graph Laplacian are the first-order linear approximations of Laplacian eigenfunctions. Therefore they are much more meaningful than the usual PCs as they relate to the intrinsic structure of the data.

The Laplacian eigenmap formed by embedding each subject to a lower dimensional Euclidean space via the top few eigenfunctions has extremely important locality preserving property, that is, the distance between a pair of subjects in the Laplacian eigenmap reflects the degree of their being correlated. The more they are correlated, the closer they are mapped to. Immediately, Laplacian eigenmap leads to cluster-like structures for subjects who either come from the same discrete subpopulation or share more common ancestry in an admixed population. We suggest using Laplacian eigenvectors instead of PCs to study population structure and regressing on Laplacian eigenvectors in disease association testings.

We demonstrate our method, LAPSTRUCT, on the North American Rheumatoid Arthritis Consortium (NARAC) data provided by GAW16. Rheumatoid Arthritis (RA) is a complex and chronic inflammatory human disorder, with both a moderately strong genetic component and environmental factor. It has been observed that PTPN22 and TRAF1-C5 genes are associated with RA [9].

## Materials and Methods

The study sample of NARAC includes 868 cases and 1194 controls across North America, who has been treated at rheumatoid clinics. The individuals from NARAC were genotyped by SNP array Illumina 550K chip in the whole

genome, with total 545,080 SNP scans. 507,246 SNPs passed quality control with the criteria: Hardy-Weinberg equilibrium (HWE) P-value  $< 10^{-5}$  (using  $\chi^2$  statistic) in controls, SNP genotype call rates  $>90\%$  completeness and minor allele frequency (MAF)  $<0.01$ . Each individual's affection status (unaffected as 0, affected as 1) was regarded as the phenotype. All the 2026 individuals in NARAC were included in the analysis.

Regard each individual  $j$  as a vertex  $V_j$  in a weighted graph  $G = (V, E)$ , where  $j = 1$  to  $N$ . Set the weight between individuals  $j$  and  $k$  to be a Gaussian kernel  $W_{jk} = e^{-\frac{\|V_j - V_k\|^2}{t}}$  for  $j \neq k$  and  $\|V_j - V_k\| < \epsilon$ , and  $W_{jj} = 1.0$  for all  $j$ . Here  $t$  and  $\epsilon$  are some selected positive real numbers. The constant  $t$  stands for the global diffusion scale on the graph and we set  $t=1.0$  in the computations. The  $\epsilon$  measures the size of each subject's neighborhood in terms of correlations, that is, all individuals to whom the distance is within  $\epsilon$  are regarded as one's close neighbors. The  $\|V_j - V_k\|$  measures the distance between vertex  $V_j$  and  $V_k$ . Here we take  $\|V_j - V_k\| = 1 - C_{jk}$ , where  $C_{jk}$  is the standard sample correlation coefficient between individuals  $j$  and  $k$  after suitable normalization of genotypes. It is obtained as follows. Let  $g_{ij}$  denote the matrix of genotype (0, 0.5, 1) of individual  $j$  at SNP  $i$ . We normalize each SNP  $i$  by subtracting off the row mean  $\mu = \frac{1}{N} \sum_j g_{ij}$ , and then divide each entry by  $\sqrt{\frac{1}{2}p_i(1-p_i)}$ , where  $p_i$  is a posterior estimate of the allele frequency at SNP  $i$  given by  $p_i = \frac{\frac{1}{2} + \sum_j g_{ij}}{1+N}$ , all missing entries are excluded from the computation. Let's still use  $g_{ij}$  denote the normalized genotype matrix, then  $C_{jk} = \frac{1}{M} \sum_i g_{ij}g_{ik}$ .

Let  $D$  be a diagonal matrix of size  $N \times N$  with entries  $D_{jj} = \sum_k W_{jk}$ , which is a natural measure on the vertices. The Laplacian matrix on graph  $G$  is defined as  $L = W - D$ . Note that  $L$  is a symmetric and positive semidefinite matrix, and we restrict to the normalized version  $D^{-1}L$  which is not symmetric anymore. The eigenfunctions of the normalized equation  $Le = \lambda De$  are denoted by  $e_j = (e_{j1}, \dots, e_{jN})^T$  for each  $j$ , ranked according to the increasing of their corresponding eigenvalues, i.e.,  $\lambda_0 \leq \lambda_1 \leq \lambda_2 \leq \dots$ . It is easy to see that 0 is always an eigenvalue with constant eigenvector consisting of all 1's. These eigenfunctions generalize the low frequency Fourier harmonics on a manifold approximated by the graph  $G$ . The Laplacian eigenmap with first  $n$  (usually small, 2 or 3) eigenvectors is defined as  $f : k \rightarrow (e_{1k}, e_{2k}, \dots, e_{nk})$  for individual  $k$  to achieve dimension reduction. Note the situation here is different from PCA, where one takes the PCs corresponding to the largest

eigenvalues which account for the largest amount of variation in the data.

Next we follow Price et al [3] to regress genotypes and phenotypes on the top 10 Laplacian eigenvectors for each individual and compute the adjusted  $\chi^2$  statistic of the residues.

[Figure 1 about here.]

## Results

The PC map (see **Figure 1**) depicts the European population structure that has previously been reported in a different study [3]. In the Laplacian eigenmap one also observes the compact trend from center to east-south and a long tail-like trend to east. Surprisingly, these presumably assumed two trends are remarkably separated in the unnormalized version of Laplacian eigenmap. We compared the results for two SNPs that have been reported to be associated with RA (see **Table 1**). The results are consistent with the prevailing principal components based approach, EIGENSTRAT.

[Table 1 about here.]

## Discussion

By setting a constant weight for each pair of individuals and sufficiently large  $\epsilon$  to include all individuals into everyone's neighborhood, the proposed approach naturally includes PCA as a special case. This fact follows from the observation below. If all weights  $W_{ij}$  are equal, say,  $\frac{1}{N^2}$ , where  $N$  is the total number of individuals, then  $D_{jj} = \frac{1}{N}$  and  $L = \frac{1}{N}I - \frac{1}{N^2}ee^t$ , where  $e = (1, \dots, 1)^t$ . Let  $g = (g_1, \dots, g_N)^t$  denote the genotype data of all individuals, where each  $g_i$  stands for the genotype vector for the  $i^{th}$  individual and  $\mu$  denote the sample mean vector of genotypes. Then one has  $gLg^T = \hat{E}[(g - \mu)(g - \mu)^t]$ . Since  $\hat{E}(g - \mu)(g - \mu)^t$  is the sample covariance matrix of the individuals, the Laplacian eigenfunctions equal the principal components.

In general, for sufficiently large  $\epsilon$ , the top Laplacian eigenfunctions describe global variations instead of local dependence structures, and they numerically approximate to the top PCs. As  $\epsilon$  decreases, Laplacian eigenmap describes the local dependence structures at different scales. When  $\epsilon$  becomes so small that each subject's neighborhood shrinks to itself, Laplacian eigenmap cannot detect any structure. In practice, the  $\epsilon$  should be chosen

reasonably large to make the graph connected and maintain valid type one error for association studies.

We have introduced a novel method for population structure detection which preserves local dependence structures. The Laplacian eigenmap naturally leads to population clusters according to the degree of correlatedness among individuals. It is less noisy, compared with PCA method. For disease association testing, LAPSTRUCT has comparable performance with EIGENSTRAT.

## ACKNOWLEDGEMENTS

J.Z is grateful to Matthew Stephens for his interest and great advices to improve the presentation. The Genetic Analysis Workshops are supported by NIH grant R01 GM031575 from the National Institute of General Medical Sciences.

## References

1. Marchini J: **The effects of human population structure on large genetic association studies.** *Nat. Genet.* 2004, **36**:512–517.
2. Freedman Mea: **Assessing the impact of population stratification on genetic association studies.** *Nat. Genet.* 2004, **36**:388–393.
3. Price AL, Patterson N, Plenge RM, Weinblatt ME, Shadick NA, Reich D: **Principal components analysis corrects for stratification in genome-wide association studies.** *Nature Genetics* 2006, **38**:904–909.
4. Zhu X, Zhang S, Zhao H, Cooper R: **Association mapping, using a mixture model for complex traits.** *Genet. Epidemiol.* 2002, **23**:181–196.
5. Chen H, Zhu X, Zhao H, Zhang S: **Qualitative semi-parametric test for genetic associations in case-control designs under structured populations.** *Ann. Hum. Genet.* 2003, **67**:250–264.
6. Novembre J, Stephens M: **Interpreting principal component analyses of spatial population genetic variation.** *Nature Genetics* 2008, **40**:646–649.

7. Chung FRK: *Spectral Graph Theory*. American Mathematical Society 1997.
8. Rosenberg S: *The Laplacian on a Riemannian Manifold*. Cambridge University Press 1997.
9. Plenge R: **TRACF1-C5 as a risk locus for rheumatoid arthritis - a genomwide study**. *NEJM* 2007, **357**:1199–209.

## List of Figures

- 1 Population structures detected by PCA, Laplacian and its un-normalized version both with  $\epsilon = 1.0$ . . . . . 9

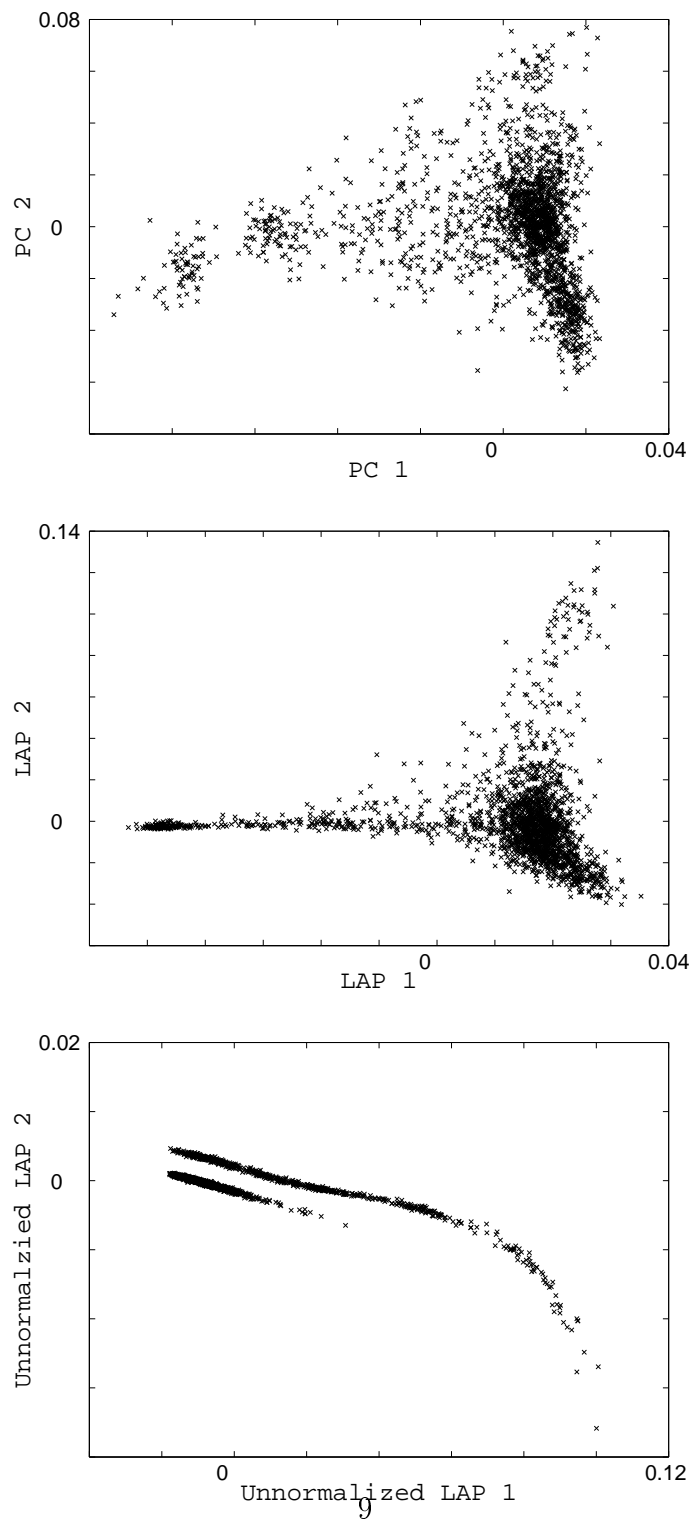


Figure 1: Population structures detected by PCA, Laplacian and its unnormalized version both with  $\epsilon = 1.0$ .

## List of Tables

|   |  |    |
|---|--|----|
| I | Association testing results for genes PTPN22 and TRAF1-C5<br>by EIGENSTRAT and LAPSTRUCT . . . . . | 11 |
|---|--|----|

Table I: Association testing results for genes PTPN22 and TRAF1-C5 by EIGENSTRAT and LAPSTRUCT

| SNP       | Chrom | EIGENSTRAT                      | LAPSTRUCT                       |
|-----------|-------|---------------------------------|---------------------------------|
| rs2476601 | 1     | 26.74 ( $2.33 \times 10^{-7}$ ) | 33.72 ( $6.36 \times 10^{-9}$ ) |
| rs3761847 | 9     | 27.57 ( $1.52 \times 10^{-7}$ ) | 25.39 ( $4.68 \times 10^{-7}$ ) |