Laplacian Eigenfunctions Learn Population Structure

Jun Zhang[1]*, Partha Niyogi[2], Mary Sara McPeek[3]

[1] Department of Radiology, The University of Chicago, Chicago, Illinois, United States of America, [2] Departments of Statistics and Computer Science, The University of Chicago, Chicago, Illinois, United States of America, [3] Departments of Statistics and Human Genetics, The University of Chicago, Chicago, Illinois, United States of America

Abstract

Principal components analysis has been used for decades to summarize genetic variation across geographic regions and to infer population migration history. More recently, with the advent of genome-wide association studies of complex traits, it has become a commonly-used tool for detection and correction of confounding due to population structure. However, principal components are generally sensitive to outliers. Recently there has also been concern about its interpretation. Motivated from geometric learning, we describe a 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 structure. In simulations and real data on a ring species of birds, Laplacian eigenfunctions reveal more meaningful and less noisy structure of the underlying population, compared with principal components. The proposed approach is simple and computationally fast. It is expected to become a promising and basic method for population genetics and disease association studies.

Citation: Zhang J, Niyogi P, McPeek MS (2009) Laplacian Eigenfunctions Learn Population Structure. PLoS ONE 4(12): e7928. doi:10.1371/journal.pone.0007928

Editor: Dennis O'Rourke, University of Utah, United States of America

Received June 16, 2009; **Accepted** October 23, 2009; **Published** December 1, 2009

Funding: Support from NIH grant R01 HG001645 (to M.S.M.) is gratefully acknowledged. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: junzhang@galton.uchicago.edu

Introduction

Principal Components Analysis (PCA) is a classical statistical tool to achieve dimension reduction through consideration of linear combinations of the original variables. The top few principal components (PCs) are the linear combinations that explain the greatest amount of variation in the data. The use of PCA in population genetics has a long history, including early work of Cavalli-Sforza and colleagues [1,2], who considered high dimensional genetic variants from population samples at many different continental locations and used the top PCs to summarize the genetic variation across space. While legitimate concerns have been raised about the interpretation of such PC maps [3], PCA can still provide useful information and is a commonly-used tool in various contexts of genetic data analysis [4]. For example, there is known to be a close connection between the spectral decomposition of the migration matrix and that of the genetic covariance matrix [5]. More recently, in genome-wide disease association studies, PCA has been employed to detect and correct population stratification [6–8], in which systematic ancestry differences between cases and controls can lead to false positive association between phenotype and genotype. Such spurious associations [9–11] can occur when the disease frequency varies across subpopulations, resulting in affected individuals being more likely than unaffected individuals to be sampled from certain subpopulations [12]. Though this topic has been extensively studied, PCA has advantages [6] over other methods such as genomic control [13] and structured association [14].

Motivated from geometric learning [15], we describe LAPSTRUCT, a Laplacian eigenfunction approach based on graph theory which we briefly introduced in Genetic Analysis Workshop (GAW) 16 [16]. One regards each subject as a vertex of a weighted graph [17], where the weight associated to the edge for each pair of subjects is chosen as a function of their genetic relatedness, with higher weight given when individuals are genetically closer (see Methods). Thus, in this context, one thinks of the distance between each pair of subjects as being based on their degree of genetic relatedness, not on their geographical proximity. The resulting adjacency graph approximates the underlying manifold of the dependence structure of the sample. The eigenfunctions of the Laplace-Beltrami operator [18] on the manifold are generalized geometric harmonic functions, which contain useful intrinsic geometric structure information on the population. The eigenvectors of the associated graph Laplacian matrix (see **Methods**) are first-order linear approximations of the Laplacian eigenfunctions, and they relate to the intrinsic dependence structure of the data. The Laplacian eigenmap formed by embedding each subject to a lower dimensional Euclidean space via the top few eigenfunctions has a 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 together they are mapped. As a result, the 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.

The Laplacian eigenfunction method is part of a large class of spectral methods that includes PCA as a special case. However, the approach we use improves on PCA in that each vertex is connected by edges to only its close neighbors, rather than to all other individuals (where, here, closeness refers to genetic relatedness rather than physical proximity). A justification for this results from the connection between spectral clustering and approximate solutions to graph cut problems (see previous work

[19,20] for details). The result is that the Laplacian eigenfunction method tends to emphasize substructure that affects many data points rather than just a few extreme points, so the proposed nonlinear algorithm is robust to outliers, in contrast to PCA. Therefore we suggest using Laplacian eigenvectors instead of PCs to study population structure. A similar approach based on spectral graph theory is also treated by Lee et al. [20] with a nice illustration on the POPRES data [21], but with different choices of weight and data renormalization (see Methods and Discussion).

The proposed method, LAPSTRUCT, has arisen from the idea of studying the geometry of the intrinsic dependence structure of sample populations, which can be creatively regarded as a weighted graph, together with a metric measuring the degree of relatedness for each pair of individuals. The paradigm of the approach is that local infinitesimal structure integrates out global macroscopic structure. Another interpretation to this is to define a random walk on the weighted graph constructed above, with a suitably normalized transition probability between two nodes reflecting their connectivity. Then one can use the top spectrum of the Markov transition matrix to map the data to a lower dimenional Euclidean space. This idea has clear antecedents in earlier work in population genetics (e.g. [5]).

The results on both the Greenish warbler (a ring species) data set [3,22] and a simulated data set with a spatially correlated population give better approximations to the true population structure than does PCA. Because Laplacian eigenfunctions are generalized harmonic functions, the patterns observed from the PC map on spatially correlated genetic data [3] are also present in the Laplacian eigenmap. Therefore, any hypotheses of historic migration suggested by LAPSTRUCT would require additional evidence before a conclusion is made.

Results

Simulation Study A

In our simulations, we compare the results of LAPSTRUCT with those of the PC-based method EIGENSTRAT [6]. **Figure 1** illustrates the population structure dectected by EIGENSTRAT and by LAPSTRUCT in the discrete population consisting of two subpopulations (see Methods). In this example, the population structure is perfectly captured by the vector, v, of length N, having entry $\frac{-N_2}{\sqrt{N_1N_2N}} = -0.0316$ for each individual in population 1 and entry $\frac{N_1}{\sqrt{N_1N_2N}} = 0.0316$ for each indvidual in population 2, where N_1 and N_2 are the total numbers of individuals from subpopulations 1 and 2, respectively, and $N = N_1 + N_2$ (see **Text S1** online for details). Both the PC and the Laplacian eigenvector appear to be approximating v, but the Laplacian approach is clearly giving a much more accurate approximation. While both approaches are effective at clustering the data, the more accurate approximation of the ancestry vector, v, by the Laplacian approach suggests that ancestry should be more accurately accounted for in downstream analyses such as association mapping. In principle, this should increase power, though in our simulation the effect was slight (see **Table 1**). **Figure 2** shows the population structure identified by EIGENSTRAT and by LAPSTRUCT in the admixed population. The PC map shows the expected uniform distribution of ancestry proportion. However, the Laplacian eigenmap shows a tendency to shrink the points toward two clear clusters, indicating the two ancestral populations. For disease association studies conducted in both simulations by simply replacing the PCs by Laplacian eigenfunctions in the regression setting introduced in reference [6], LAPSTRUCT peforms as well as EIGENSTRAT (see **Table 1**).

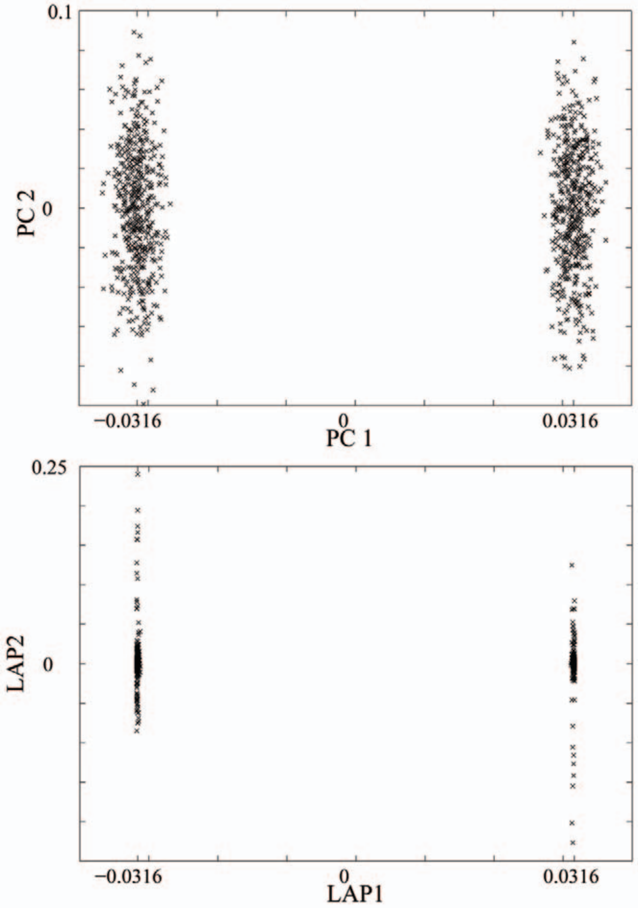

Figure 1. Structure of a simulated discrete population. Population structure detected by PCA (top) and by Laplacian with $\epsilon = 1.0$ (bottom), for the discrete population consisting of two subpopulations.
doi:10.1371/journal.pone.0007928.g001

Simulation Study B

The sensitivity of PC to outliers is illustrated by the analysis of the spatially correlated population that consists of subpopulations arranged on a circle and an additional isolated subpopulation. When 10 individuals from the isolated subpopulation are included

Table 1. Simulated Association Testing.

	EIGENSTRAT	LAPSTRUCT ($\epsilon = 1.0$)	LAPSTRUCT ($\epsilon = 2.0$)
Discrete population			
Random SNPs	0.0001	0.0001	0.0001
Differentiated SNPs	0.0001	0.0001	0.0001
Causal SNPs	0.4735	0.4762	0.4739
Admixed population			
Random SNPs	0.0001	0.0001	0.0001
Differentiated SNPs	0.0001	0.0001	0.0001
Causal SNPs	0.4891	0.4919	0.4863

Proportion of association reported as significant by EIGENSTRAT and LAPSTRUCT at significance level 10^{-4}, based on 100,000 simulations.
doi:10.1371/journal.pone.0007928.t001

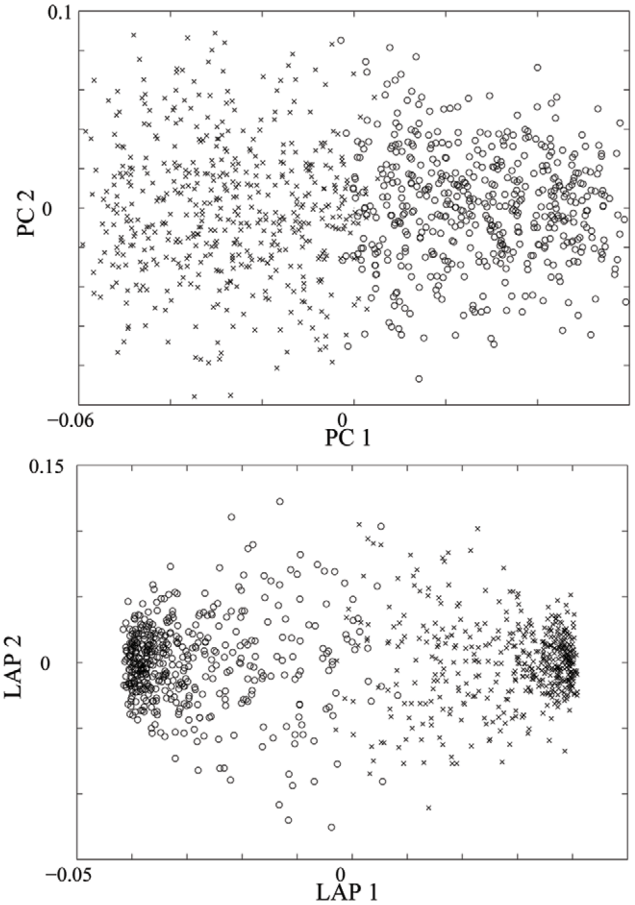

Figure 2. Structure of a simulated admixed population. The ancestral population structure detected by PCA (top) and by Laplacian with $\epsilon = 1.0$ (bottom), for the admixed population with two ancestral populations, where crosses (circles) stand for individuals whose ancestry proportion from ancestral population 1 is larger (smaller) than one-half.
doi:10.1371/journal.pone.0007928.g002

in the sample, the top PC focuses on isolating those outliers, and the PC map based on the top 2 components does not capture the full structure of the data, missing the circle configuration of the population structure (see **Figure 3**). With the outliers removed from the sample, the PC map based on the top two PCs does give the ring shape of the population structure. In contrast, the Laplacian eigenmap based on two components identifies the full population structure even in the presence of outliers, demonstrating that it is much more robust to outliers than is PC. The additional smoothness in the Laplacian eigenmap compared to the PC map might be due to the fact local correlation is weighted more highly, which gives a local smoothing effect.

Phylloscopus trochiloides

Figure 4 below illustrates the population structure detected by the PCA and Laplacian methods, respectively, where one can more clearly observe the ring-shape structure in the Laplacian eigenmap, compared to the vague structure shown in the PC map.

Discussion

We have developed LAPSTRUCT, a Laplacian eigenfunction approach for detection and correction of population structure in

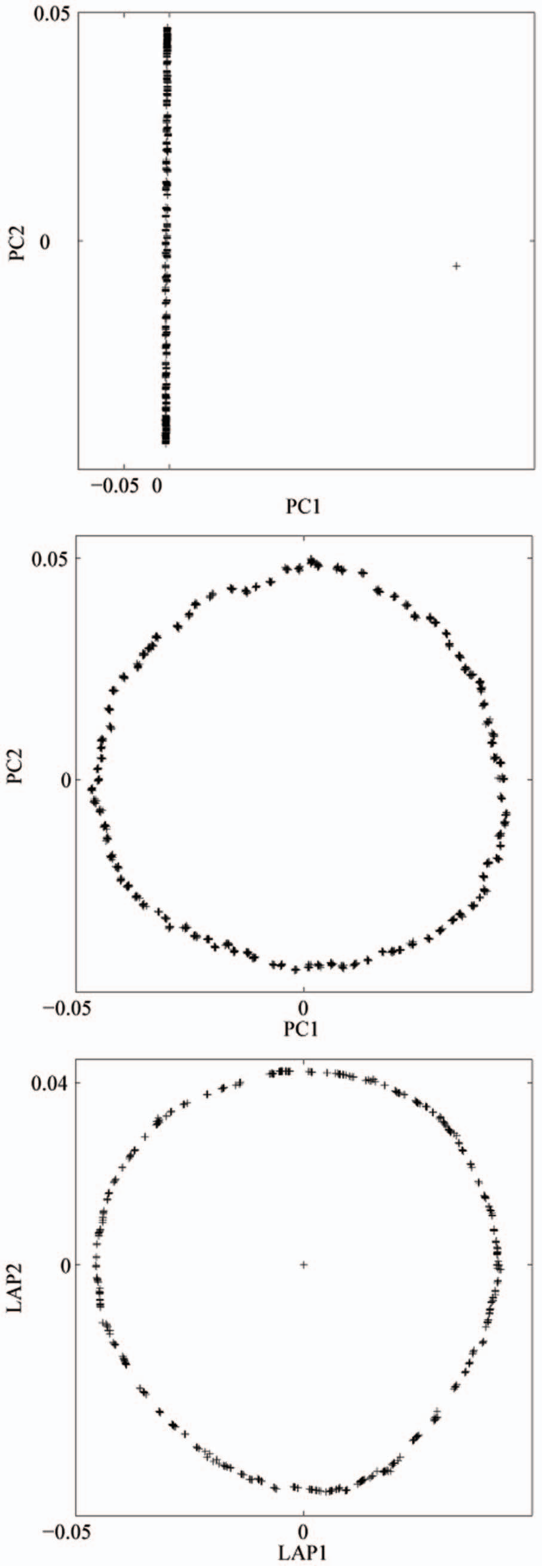

Figure 3. Structure of a simulated ring population. Population structure detected by PCA (with and without outliers present) and by Laplacian with $\epsilon = 0.4$, for the simulated ring population.
doi:10.1371/journal.pone.0007928.g003

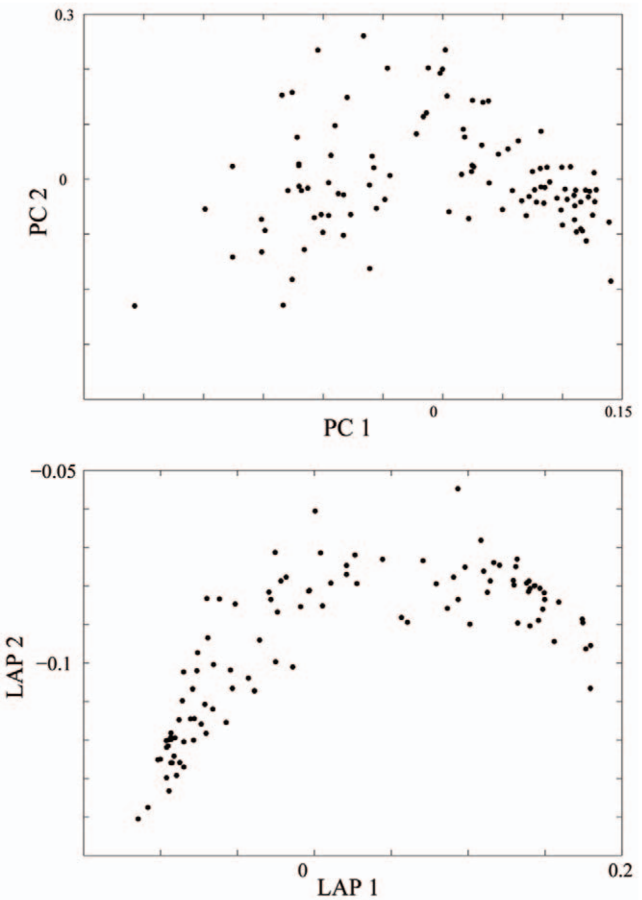

Figure 4. Ring Structure of a real dataset. Population structure detected by PC map and by Laplacian eigenmap with $\epsilon=0.90$, for Greenish Warbler dataset.
doi:10.1371/journal.pone.0007928.g004

genetic studies. LAPSTRUCT can be viewed as a robust alternative to PC-based methods such as EIGENSTRAT. Like PC, LAPSTRUCT naturally leads to population clusters according to the degree of genetic correlation among individuals. However, LAPSTRUCT is designed to be less sensitive to outliers than PC, emphasizing structure that affects many data points rather than just a few extreme points. LAPSTRUCT can reveal less noisy and richer structure at different scales by varing the parameters. It is expected to become a promising tool for population genetics.

In the simulation studies, the top Laplacian eigenfunctions identify the overall structure, while the PC approach has a tendency to highlight outliers, when they are present. For example, in the spatial simulation with outliers, PC requires three components to find the ring structure, while the Laplacian eigenfunction approach finds the ring structure with only two components. This suggests that the Laplacian eigenfunction approach could be more useful than the PC approach in contexts such as association mapping in which it is desirable to capture the population structure with as few components as possible, in order to preserve power. Additionally, only those eigenfunctions for which cases and controls have significantly different distributions need to be accounted for in the setting of association mapping, and including unnecessary eigenfunctions will lead to power loss. Further investigation in this direction is encouraged.

The Laplacian eigenmap approach we describe is part of a more general setting of spectrum-based dimension reduction techniques that includes the PC approach. The appropriate choice of the neighborhood parameter, ϵ, is what causes the Laplacian eigenmap to be less sensitive to outliers than PC. When ϵ is sufficiently large, the Laplacian eigenmap approach and the PC approach can produce very similar results. As ϵ is decreased, the Laplacian eigenmap can capture the local dependence structure at different scales. In practice, ϵ should be chosen reasonably large to make the graph connected and maintain valid type one error for association studies. For example, ϵ could be the α-th quantile for some suitable α. An alternative on the scale of neighborhood is to select each subject's K closest neighbors in terms of correlation for some reasonably large integer K. To avoid the issue of tuning parameter selection Lee et al. [20] simply take $w_{jk}=\sqrt{c_{jk}}$ if $c_{jk}>0$, otherwise $w_{jk}=0$. Generally there is room for different choices of weights which may give close performance, and the optimal weight is worth further investigation. The threshholding technique seems appropriate and it has been widely accepted. It reduces the noise from less correlated samples. We incorporate this idea in the renormalization of the genotype data, where each individual's SNP is normalized using the *local* SNP frequency estimated from only those closely correlated individuals. We note this is appropriate when the data are abundant, and one would certainly use all data instead if the sample size were relatively small.

Materials and Methods

Phylloscopus Trochiloides (Greenish Warblers) Data

Greenish warblers are most abundant in western and eastern Siberia, where they form a ring species complex. The complex consists of two main populations connected by gene flow via a narrow band of populations to the south that are arranged in a ring around the Tibetan plateau. There is no mating between the two main populations where they overlap geographically, so greenish warblers can be regarded as inhabiting a one-dimensional habitat. Irwin et al. [22] collected 105 individuals from 26 geographic sites and each individual was typed for presence or absence at 62 amplified fragment length polymorphism (AFLP) markers.

Laplacian Eigenfunctions

Regard each individual j as a vertex V_j in a weighted graph $G=(V,E)$, where $j=1$ to N. Let the weight between individuals j and k be a Gaussian kernel $W_{jk}=e^{-\frac{\|V_j-V_k\|^2}{t}}$ if $j\neq k$ and $\|V_j-V_k\|<\epsilon$, and $W_{jk}=0$ otherwise. Here t and ϵ are some selected positive real numbers. The ϵ measures the size of each subject's neighborhood. The constant t stands for the global diffusion scale on the graph and we set $t=1.0$ in all the computations. (For information on the effects of ϵ and t on detection of population structure, see **Figure S1** online.) The $\|V_j-V_k\|$ measures the *distance* between vertex V_j and V_k. We set the distance $\|V_j-V_k\|=1-C_{jk}$, where C_{jk} is the estimator of genetic correlation [6] between individuals j and k. Specifically, let g_{ij} denote the genotype $\left(0,\frac{1}{2},1\right)$ of individual j at SNP i. We normalize the vector of genotypes for SNP i by subtracting off its average, $\mu_i=\frac{1}{N}\sum_j g_{ij}$, and then dividing each entry by $\sqrt{\frac{1}{2}p_i(1-p_i)}$, where p_i is an 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 X_{ij} be the resulting normalized genotype for SNP i in individual j. Then we set $C_{jk}=\frac{1}{N}\sum_i X_{ij}X_{ik}$.

To avoid the effects of population structure in the allele frequency estimation, the same idea above leads to an alternative local SNP frequency estimation and genotype updating approach. Instead of estimating a single allele frequency per marker, we compute a local SNP frequency f_{ij} for each individual j at SNP i simply by including only those individuals whose correlation with individual j is larger than $1-\epsilon$. That is, $f_{ij}=\frac{1}{\#\{k:C_{kj}\geq 1-\varepsilon\}}\sum_{\{k:C_{kj}\geq 1-\varepsilon\}} g_{jk}$. Next we denote the updated genotype matrix G from the original genotype matrix g by $G_{ij}=\frac{g_{ij}-f_{ij}}{\sqrt{\frac{1}{2}f_{ij}(1-f_{ij})}}$.

Let D be a diagonal matrix of size $N\times N$ with entries $D_{jj}=\sum_k W_{jk}$, a natural measure on the vertices. The Laplacian matrix on graph G is defined to be $L=D-W$. 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 $e_j=(e_{j1},\ldots,e_{jN})^T$ for each j, ranked according to the reverse order of their corresponding eigenvalues, i.e., $\lambda_0\leq\lambda_1\leq\lambda_2\leq\ldots$. 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. To achieve dimension reduction, the Laplacian eigenmap with first n (usually small, 2 or 3) eigenvectors is defined by $f:k\rightarrow(e_{1k},e_{2k},\ldots,e_{nk})\in\mathbb{R}^n$ for individual k. Note that 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. The justification is given below. We remark that a symmetrically normalized version of L is given by $D^{-\frac{1}{2}}LD^{-\frac{1}{2}}$. The Laplacian eigenmap using the corresponding spectrum gives comparable performance. For the relationship between these two versions, see [19].

The Laplacian eigenmap approach we describe is part of a more general setting of spectrum-based dimension reduction techniques that includes the PC approach. The appropriate choice of the neighborhood parameter, ϵ, is what causes the Laplacian eigenmap approach to be less sensitive to outliers than PC. When ϵ is sufficiently large, the Laplacian eigenmap approach and the PC approach can produce very similar results. This is shown in **Figure 5** for the simulated discrete population model. As ϵ is decreased, the Laplacian eigenmap can capture the local dependence structure at different scales. See **Figure S2** online for an illustration.

To apply the Laplacian eigenmap method to disease association studies, one can follow a multiple regression approach as in [6]. For example, one could regress genotypes and phenotypes on the top K Laplacian eigenvectors for each individual, and then compute the adjusted χ^2 statistic of the residuals. In the simulations, we set K equal to 10, in order to make a comparison with EIGENSTRAT.

Justification of Weight Kernel and Laplacian Eigenmap

The selected Gaussian weight is optimal in a certain sense, and it has a deep connection to the heat kernel on a manifold that gives the general solution to the heat equation. In the discrete case, the Laplacian of a function can be expressed as combinations of heat kernels which locally approximate the Gaussian kernel. For the mathematical details, see references [15,17]. The locality preserving property of the Laplacian eigenmap follows from the fact that the cost function of a weighted graph equals the Laplacian of the map function, that is, $.5\sum_{i,j}W_{ij}\left(f(x_i)-f(x_j)\right)^2=f(\vec{x})^t Lf(\vec{x})$, where $\{x_i\}$ are the collection of nodes and $f(\vec{x})=(f(x_1),\ldots,f(x_N))^t$. So the minimization problem reduces to finding $f(x)$ that minimizes $f(\vec{x})^t Lf(\vec{x})$, subject to the constraint $f(x)^t Df(x)=1$, and this is equivalent to the generalized eigenvalue problem stated above. This also explains why the Laplacian eigenmap ranks the eigenvalues in *increasing* order.

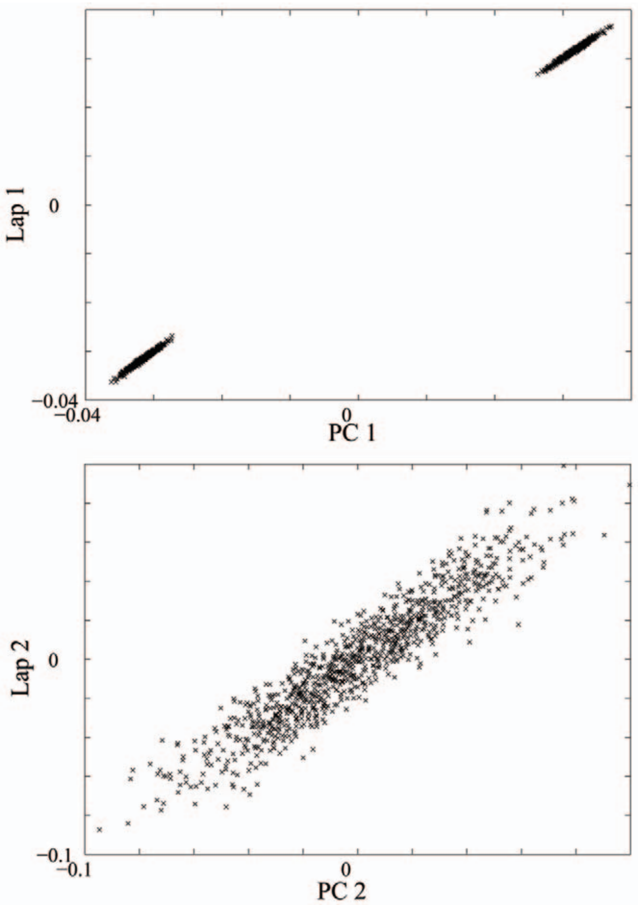

Figure 5. QQ-plot of PCA and Laplacian. QQ-plot of the top two PCs and Laplacian eigenfunctions with $\epsilon=2.0$ for the simulated discrete population.
doi:10.1371/journal.pone.0007928.g005

Simulation Study A. Discrete and Admixed Populations

To simulate a discrete population consisting of two subpopulations, we follow a model of population structure used in reference [10] (see also [6]). Each subpopulation is generated by the Balding-Nichols model, but with each subpopulation having its own generalized F_{st} value (0.01 and 0.05, respectively, for subpopulations 1 and 2), instead of the same value for both subpopulations (see [10] for details). The population allele frequency of each random SNP is sampled uniformly from [0.1,0.9]. The allele frequency within each subpopulation is drawn from a beta distribution, $Beta\left(\frac{p(1-F_{st})}{F_{st}},\frac{(1-p)(1-F_{st})}{F_{st}}\right)$. For each individual, 10,000 SNPs were generated. The sample consists of 500 cases and 500 controls, where 60% of cases and 40% of controls were from subpopulation 1 and the rest were sampled from subpopulation 2. For the admixed population with two ancestral populations, the ancestral populations' generalized F_{st} values were set equal to 0.01 and 0.09 respectively. For the admixed population, 1,000 individuals were sampled, half cases and half controls. The sample's ancestral proportions are assumed uniformly distributed from 0 to 1. For the causal allele, a risk model [6] with relative risk $r=1.5$ was used for both the discrete population and the admixed population.

The allele frequencies for highly differentiated SNPs are respectively set to 0.2 and 0.8 in the two subpopulations.

Simulation Study B. Spatially Correlated Population

Following reference [3], an equilibrium population is simulated using the software MS for population genetics developed by Hudson [23]. The population consists of 100 subpopulations equally spaced on a circle, with members of an additional isolated subpopulation as outliers. Each subpopulation is assumed to consist of an equal number of diploids. During each generation backward in time, a fraction $m = 0.1$ of each subpopulation along the circle is made up of migrants from each adjacent subpopulation, and there are no gamete swaps between non-adjacent subpopulations. 1,000 SNP loci were independently simulated with one segregating site per locus, and 10 individuals were sampled from each subpopulation.

URL. Software for running LAPSTRUCT on a Linux platform is available at http://galton.uchicago.edu/˜junzhang/LAPSTRUCT.html.

Supporting Information

Text S1 Supporting Text
Found at: doi:10.1371/journal.pone.0007928.s001 (0.09 MB PDF)

Figure S1 Here we consider the simulated discrete population consisting of two subpopulations, analyzed with $\varepsilon = 1.0$ in all cases. When the scale parameter t is sufficiently small, the Laplacian matrix L degenerates to the identical matrix I and no structure can be detected. When $t = 0.1$, the second Laplacian eigenfunction degenerates approximately to zero for one of the subpopulations. For larger t values, there are little difference in the detected structures.
Found at: doi:10.1371/journal.pone.0007928.s002 (0.08 MB PDF)

Figure S2 Here we consider the simulated discrete popualtion consisting of two subpopulations, and $t = 1.0$ in all cases. When $\varepsilon = 0.96$, the graph has two connected components representing two subpopulations and the top two Laplacian eigenfunctions degenerate to 0 and $-1/\sqrt{500} = -0.0447$. When $\varepsilon \geq 1.0$, the graph is connected. As ε increases, the local correlation structures revealed by the Laplacian eigenmap evolve to global structures which approximate to PCs.
Found at: doi:10.1371/journal.pone.0007928.s003 (0.11 MB PDF)

Acknowledgments

The authors are grateful to Matthew Stephens for numerous discussions and advice to improve the presentation and John Novembre for help on spatial simulations. Thanks also goes to Trevor Price and Darren Irwin for generously providing the greenish warblers data set.

Author Contributions

Conceived and designed the experiments: JZ PN MSM. Performed the experiments: JZ. Analyzed the data: JZ. Contributed reagents/materials/analysis tools: JZ. Wrote the paper: JZ PN MSM.

References

1. Cavalli-Sforza L, Edwards AWF Analysis of human evolution. Genetics Today 3.
2. Menozzi P, Piazza A, Cavalli-Sforza L (1978) Synthetic maps of human gene frequencies in europeans. Science 201: 786–792.
3. Novembre J, Stephens M (2008) Interpreting principal component analyses of spatial population genetic variation. Nature Genetics 40: 646–649.
4. Reich D, Price A, Patterson N (2008) Prinicpal component analysis of genetic data. Nature Genetics 40: 491–2.
5. Felsenstein J (2002) Contrasts for a within-species comparative method. Modern Developments in Theoretical Population Genetics: the legacy of Gustave Malecot Slatkin M, Veuille M, eds. New York: Oxford University Press. pp 118–129.
6. Price AL, Patterson N, Plenge RM, Weinblatt ME, Shadick NA, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics 38: 904–909.
7. Chen H, Zhu X, Zhao H, Zhang S (2003) Qualitative semi-parametric test for genetic associations in case-control designs under structured populations. Ann Hum Genet 67: 250–264.
8. Zhu X, Zhang S, Zhao H, Cooper R (2002) Association mapping, using a mixture model for complex traits. Genet Epidemiol 23: 181–196.
9. Lander E, Schork N (1994) Genetic dissection of complex traits. Science 265: 2037–2048.
10. Marchini J, Cardon LR, Phillips MS, Donnelly P (2004) The effects of human population structure on large genetic association studies. Nat Genet 36: 512–517.
11. Freedman Mea (2004) Assessing the impact of population stratification on genetic association studies. Nat Genet 36: 388–393.
12. Pritchard J, Rosenberg N (1999) Use of unlinked genetic markers to detect population stratification in association studies. Am J Hum Genet 65: 220–228.
13. Devlin B, Roeder K (1999) Genomic control for association studies. Biometrics 55: 997–1004.
14. Pritchard J, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155: 945–959.
15. Belkin M, Niyogi P (2003) Laplacian eigenmaps for dimensionality reduction and data representation. Neural Computing 13: 1373–1397.
16. Zhang J, Weng C, Niyogi P (2009) Graphical analysis of population structure on rheumatoid arthritis data. BMC Proceedings, in press.
17. Chung FRK (1997) Spectral Graph Theory. American Mathematical Society.
18. Rosenberg S (1997) The Laplacian on a Riemannian Manifold. Cambridge University Press.
19. von Luxburg U (2007) A tutorial on spectral clustering. Stat Comput 17: 395–416.
20. Lee A, Luca D, Klei L, Devlin B, Roeder K (2009) Discovering genetic ancestry using spectral graph theory. Genetic Epidemiology 33(5).
21. Nelson MR, Bryc K, King KS, Indap A, Boyko AR, et al. (2008) The population reference sample, popres: a resource for population, disease, and pharmacological genetics research. Am J Hum Genet 83: 347–358.
22. Irwin DE, Bensch S, Irwin JH, Price TD (2005) Speciation by distance in a ring species. Science 307: 414–6.
23. Hudson RR (2002) Generating samples under a wright-fisher neutral model. Bioinformatics 18: 337–8.