Heredity (2011) 107, 413–420

www.nature.com/hdy

ORIGINAL ARTICLE

Investigating population stratification and admixture using eigenanalysis of dense genotypes

D Shriner
Center for Research on Genomics and Global Health, National Human Genome Research Institute, Bethesda, MD, USA

Principal components analysis of genetic data is used to avoid inflation in type I error rates in association testing due to population stratification by covariate adjustment using the top eigenvectors and to estimate cluster or group membership independent of self-reported or ethnic identities. Eigendecomposition transforms correlated variables into an equal number of uncorrelated variables. Numerous stopping rules have been developed to identify which principal components should be retained. Recent developments in random matrix theory have led to a formal hypothesis test of the top eigenvalue, providing another way to achieve dimension reduction. In this study, I compare Velicer's minimum average partial test to a test on the basis of Tracy–Widom distribution as implemented in EIGENSOFT, the most widely used implementation of principal components analysis in genome-wide association analysis. By computer simulation of vicariance on the basis of coalescent theory, EIGENSOFT systematically overestimates the number of significant principal components. Furthermore, this overestimation is larger for samples of admixed individuals than for samples of unadmixed individuals. Overestimating the number of significant principal components can potentially lead to a loss of power in association testing by adjusting for unnecessary covariates and may lead to incorrect inferences about group differentiation. Velicer's minimum average partial test is shown to have both smaller bias and smaller variance, often with a mean squared error of 0, in estimating the number of principal components to retain. Velicer's minimum average partial test is implemented in R code and is suitable for genome-wide genotype data with or without population labels.
Heredity (2011) **107,** 413–420; doi:10.1038/hdy.2011.26; published online 30 March 2011

Keywords: admixture; population stratification; principal components; stopping rule; vicariance

Introduction

An active area of research for the past decade has been the exploration of evidence for population structure from genome-wide genetic data. Accounting for population structure is critical in association studies, in which population stratification or cryptic relatedness can lead to inferential errors. Inferences about population structure are also critical in understanding group differences in studies of evolutionary and demographic histories.

Principal components analysis is widely used for identifying population structure in genetic data. EIGENSOFT implements principal components analysis for the purposes of detecting and correcting for population stratification in genome-wide association studies (Price *et al*., 2006) and detecting structure in population genetic studies (Patterson *et al*., 2006). EIGENSOFT is based on principal components analysis of the normalized sample covariance matrix, in which the normalization assumes Hardy–Weinberg equilibrium (Price *et al*., 2006). Formal hypothesis testing for significance of eigenvalues using Tracy–Widom statistics permits a determination of the number of significant principal components, or, equivalently, the number of covariates necessary to control for population stratification (Patterson *et al*., 2006).

Many stopping rules have been developed to determine the number of significant principal components (for a comparative study of 20 stopping rules, see Peres-Neto *et al*. (2005)). In this study, I explore Velicer's minimum average partial test (Velicer, 1976; O'Connor, 2000) as an alternative to Tracy–Widom statistics. Rather than performing formal hypothesis testing using an external reference distribution and subjective significance thresholds, Velicer's minimum average partial test is based on an objective minimization function of partial correlations (Velicer, 1976).

The motivations of this study are twofold. One, in their original description of EIGENSOFT, Patterson *et al*. (2006) noted an overestimation of significant principal components for admixed data. Two, analysis of an empirical admixed African-American data set by EIGENSOFT yielded 16 significant principal components, whereas the expectation for a two-way admixed sample was one significant principal component. Herein, by computer simulation, Velicer's minimum average partial test estimated the number of principal components to retain with a smaller mean squared error than EIGENSOFT, with EIGENSOFT's estimates being biased upward. Computer simulation also revealed that EIGENSOFT yielded even more highly upwardly biased estimates for admixed samples than for unadmixed samples, whereas Velicer's minimum average partial test yielded a low mean squared error for both unadmixed and admixed samples. For the empirical data, Velicer's

Correspondence: Dr D Shriner, Center for Research on Genomics and Global Health, National Human Genome Research Institute, Building 12A, Room 4047, 12 South Dr., MSC 5635, Bethesda, MD 20892-5635, USA.
E-mail: shrinerda@mail.nih.gov
Received 26 November 2010; revised 3 February 2011; accepted 22 February 2011; published online 30 March 2011

minimum average partial test indicated retention of only one principal component, matching the expectation for a two-way admixed sample.

Materials and methods

Simulations

All work was performed in R (R Development Core Team, 2009). R code is available upon request.

Two populations: Under a coalescent model of vicariance (McVean, 2009), suppose A and B represent two populations that diverged at some time t in the past. To mimic an admixed African-American population, suppose A represents individuals of West African ancestry and suppose B represents individuals of Western European ancestry. A sample of 216 haplotypes (108 diploid individuals, see real data analysis below) from population A and 218 haplotypes (109 diploid individuals) from population B were simulated with divergence times $t=\{0, 0.001, 0.01, 0.1, 1.0\}$ in units of $2N_e$ generations (Figure 1a). Each data set consisted of 10 000 unlinked sites, that is, each site had an independently realized coalescent genealogy. As a result, there was neither background linkage disequilibrium nor extended linkage disequilibrium due to admixture. Mutations were placed on a genealogy proportional to branch lengths; the effective population size N_e canceled out in this step. As a consequence of this mutational scheme, sites were ascertained to be polymorphic, but were not ascertained to be ancestrally informative. Haplotypes were randomly paired within each population to generate diploid individuals. These data sets were used to test population stratification.

I extended this model of vicariance to include admixture. A sample of 1018 admixed individuals was generated instantaneously using parental populations A and B. For each admixed individual, the average genome-wide admixture proportion p was determined by drawing a random deviate from the beta-distribution β(10.18508,2.837815), yielding an expected genome-wide admixture proportion $\bar{p}=0.782$. The parameters of the beta-distribution were chosen such that the first two moments matched the first two moments of the empirical distribution of individual admixture proportions for the African-American data set described below in the real data analysis subsection. For each site independently, the individual was assigned the state of a randomly selected haplotype from population A if a random deviate from the uniform distribution U(0,1) $\leqslant p$ and assigned the state of a randomly selected haplotype from population B otherwise. For each divergence time t, 1000 independent replicate data sets were generated.

Three populations: I also extended the model of vicariance to three populations. In the same coalescent framework, suppose A, B and C represent three ancestral populations that diverged at two times in the past. Populations B and C diverged $t_1=\{0.0001, 0.001, 0.01, 0.1, 1.0\}$ in units of $2N_e$ generations ago, and population A diverged $t_2=10t_1$ in units of $2N_e$ generations ago (Figure 1b). Samples of 200 haplotypes from each of the three populations were simulated. Each data set consisted of 10 000 unlinked sites. Haplotypes were randomly paired within each population to generate diploid individuals. These data sets were used to test population stratification.

A sample of 1018 admixed individuals was generated instantaneously using parental populations A, B and C. For each admixed individual, p_1 was a random deviate from β(0.8,7.2) and p_2 was a random deviate from β(12,12). For each site independently, the individual was assigned the state of a randomly selected haplotype from population A if a random deviate from U(0,1) $\leqslant p_1$. If the random uniform deviate $>p_1$, then the individual was assigned the state of a randomly selected haplotype from population B if a random deviate from U(0,1) $\leqslant p_2$ and assigned the state of a randomly selected haplotype from population C otherwise. The expected genome-wide proportion of haplotypes from populations A, B and C were $p_A=p_1=0.10$, $p_B=(1-p_1)p_2=0.45$ and $p_C=1-p_1-(1-p_1)p_2=0.45$, respectively, intended to mimic African, European and Native American admixture proportions in Latino populations (Martinez-Marignac *et al*., 2007; Price *et al*., 2007). For each divergence time t_1, 1000 independent replicate data sets were generated.

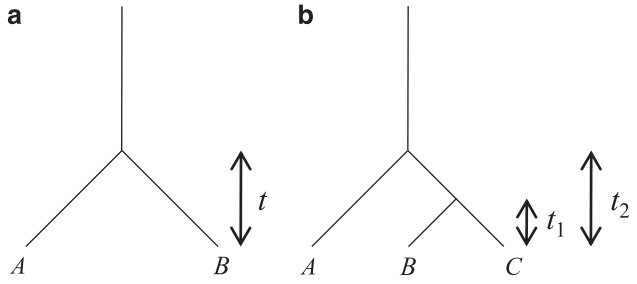

Figure 1 Genealogical representation of the coalescent simulations. (**a**) Two populations with a single divergence event $2tN_e$ generations ago. (**b**) Three populations with the first divergence event $2t_1N_e$ generations ago and the second divergence event $2t_2N_e$ generations ago.

EIGENSOFT

Let **G** be the $M\times N$ matrix of genotypes for $i=1$ to M SNPs and $j=1$ to N individuals, with genotypes coded as 0, 1 or 2 copies of the minor allele. The rows of **G** were centered by subtracting

$$\mu_i=\sum_{j=1}^{N} g_{ij}\Big/N$$

from each entry in row i (Price *et al*., 2006). Eigendecomposition was performed on the standardized $N\times N$ sample covariance matrix. Significance of the leading eigenvalue was determined by a formal hypothesis test on the basis of the Tracy–Widom distribution (Johnstone, 2001; Patterson *et al*., 2006). The rank of the sample covariance matrix was one less than the number of individuals due to centering. The number of eigenvalues/eigenvectors equaled the rank of the covariance matrix. Given orthogonality of eigenvectors, P-values were Bonferroni corrected for the number of eigenvectors. Thus, the significance level was $0.05/(N-1)$.

Velicer's minimum average partial test

Let **G** be the $M\times N$ matrix of genotypes for $i=1$ to M SNPs and $j=1$ to N individuals, with genotypes coded as

0, 1 or 2 copies of the minor allele. First, center the rows of **G** by subtracting

$$\mu_i = \sum_{j=1}^{N} g_{ij} \Big/ N$$

from each entry in row i (Price *et al*., 2006). Next, compute the $N \times N$ sample correlation matrix **R**, in which the elements are Pearson product-moment correlation coefficients. Let $\mathbf{R}^*_{(m)}$ be the $N \times N$ matrix of partial correlations after the first m principal components have been partialed out. Let r^*_{kl} be the element in the kth row and lth column of $\mathbf{R}^*_{(m)}$. Velicer (1976) proposed the summary statistic $f_m = \sum_{k \neq l} \sum \frac{(r^*_{kl})^2}{N(N-1)}$. The summary statistic f_m is the average of the squared partial correlations after the first m components are partialed out, with $m = 0$ to $N-1$ (see the Appendix for computer code) (Velicer, 1976). The stopping point is the value of m for which f_m is minimum (Velicer, 1976).

Real data analysis

To illustrate application to real world data, the number of significant principal components was estimated using both EIGENSOFT and Velicer's minimum average partial test for a previously described sample of 1018 unrelated African Americans, genotyped as part of the Howard University Family Study (Adeyemo *et al*., 2009). The 808 465 autosomal SNPs that passed quality control were pruned for linkage disequilibrium at $r^2 \geqslant 0.3$ using PLINK, version 1.06 (Purcell *et al*., 2007). Of the SNPs that remained after pruning, 10 000 were randomly selected. HapMap phase III CEU and YRI samples were used to represent the presumed parental populations (The International HapMap 3 Consortium, 2010). After quality control, 108 YRI and 109 CEU individuals remained (Chen *et al*., 2010).

Results

Simulations

For the simulation of two populations separated by one divergence event (Figure 1a), the expected number of significant principal components was one if the divergence event occurred in the distant past, or zero if the divergence event occurred in the recent past (Figure 2). These expectations also hold for analysis of admixed individuals, because the expected allele frequencies for an admixed individual are linear mixes of the parental allele frequencies (Patterson *et al*., 2006; McVean, 2009). Consequently, admixed individuals will have coordinates along the axis defined by the two parental populations. This behavior holds regardless of whether the parental populations are included in the projection, which is important because sizes of proxy or reference samples are typically much smaller than sizes of admixed samples, and unequal sample sizes can severely distort the projection (Novembre *et al*., 2008; McVean, 2009).

EIGENSOFT with Tracy–Widom statistics consistently overestimated the number of significant principal components in analysis of the two populations and performed worse in analysis of admixed individuals (Table 1). Velicer's minimum average partial test showed neither bias nor variance (Table 1). Both methods were more sensitive to lower levels of divergence in analysis of stratified populations than in the analysis of admixed individuals, despite the former analysis consisting of 10^2 individuals and the latter analysis consisting of 10^3 individuals (Table 1).

For the simulation of three populations separated by two divergence events (Figure 1b), the expected number of significant principal components was 0, 1 or 2, depending on the time since the divergence events (Figure 3). EIGENSOFT with Tracy–Widom statistics showed a larger upward bias for analysis of three populations (Table 2) than for analysis of two populations (Table 1), and again performed worse in analysis of admixed individuals (Table 2). Velicer's minimum average partial test yielded no to very small biases and variances in both the analysis of three populations and the analysis of admixed individuals.

Real data analysis

In analysis of 1018 unrelated African Americans, Adeyemo *et al*. (2009) retained two principal components on the basis of visual inspection of the scree plot (that is, eigenvalues sorted in descending order as a function of the eigenvalue index) from EIGENSOFT. Using Tracy–Widom statistics, EIGENSOFT claimed 16 significant principal components (Figure 4). Using Velicer's minimum average partial test, only one principal component should be retained (Figure 5), the minimal number of principal components necessary to explain admixture between two ancestral parental populations in the absence of residual substructure.

Discussion

Detecting population structure is important in both medical genetics and population genetics. EIGENSOFT is a widely used implementation of principal components analysis supplemented with statistics for formal hypothesis testing on the basis of the expected distribution of the largest eigenvalue (Johnstone, 2001; Patterson *et al*., 2006). In this study, I compared the test based on the Tracy–Widom distribution in EIGENSOFT to Velicer's minimum average partial test (Velicer, 1976) for determining the number of principal components to retain. Computer simulation under a coalescent model of vicariance (that is, divergence with no subsequent gene flow) revealed three trends regarding the estimation of the number of principal components to retain. One, EIGENSOFT increasingly overestimated the number of significant principal components for increasingly distant divergence events in unadmixed samples. Two, EIGENSOFT overestimated the number of significant principal components even more for samples of admixed individuals than for samples of unadmixed individuals. Three, EIGENSOFT increasingly overestimated the number of significant principal components as the number of populations increased. In contrast, Velicer's minimum average partial test showed almost no tendency to overestimate the number of principal components to retain in any of these scenarios.

Velicer's minimum average partial test is one of many possible stopping rules for determining the number of principal components to retain (Peres-Neto *et al*., 2005). Another possibility is sparse factor analysis (Engelhardt and Stephens, 2010), in which constraints encourage

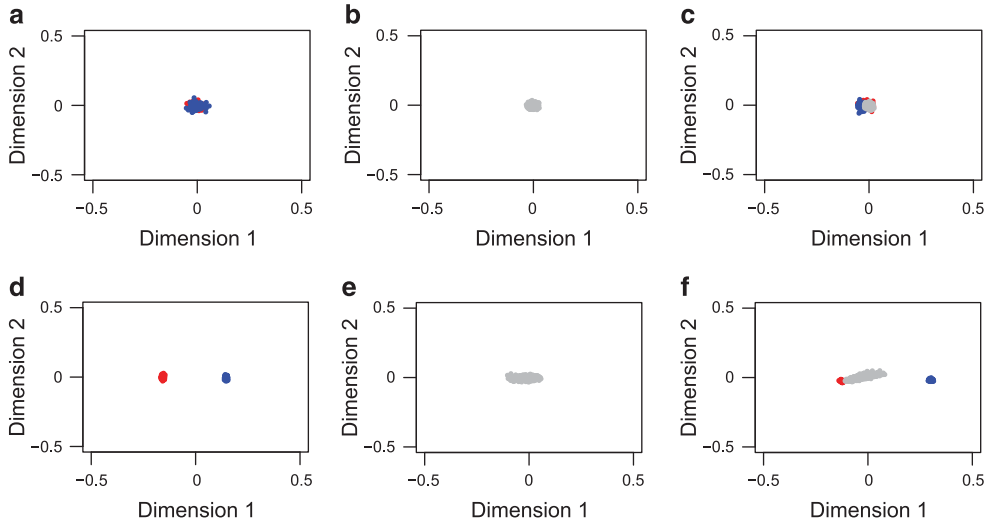

Figure 2 Representative projections of simulated data for two populations. (**a**–**c**) The divergence event between populations *A* (red circles) and *B* (blue circles) occurred 0 generations ago. (**d**–**f**) The divergence event occurred $2N_e$ generations ago. (**a**, **d**) Analysis of populations *A* and *B*. (**b**, **e**) Analysis of admixed individuals (gray circles) with average individual admixture proportions 78.2% population *A* and 21.8% population *B*. (**c**, **f**) Combined analysis of admixed individuals, population *A* and population *B*.

Table 1 Number of significant principal components in the simulation of two populations

t^a	F_{ST}	*EIGENSOFT*				*Minimum average partial test*			
		Population stratification		*Admixture*		*Population stratification*		*Admixture*	
	Mean (s.d.)	*Mean (s.d.)*	*Range*	*Mean (s.d.)*	*Range*	*Mean (s.d.)*	*Range*	*Mean (s.d.)*	*Range*
1	0.115 (0.003)	1.37 (0.62)	(1, 4)	3.23 (1.19)	(1, 7)	1.00 (0.00)	(1, 1)	1.00 (0.00)	(1, 1)
0.1	0.025 (0.001)	1.30 (0.54)	(1, 4)	2.42 (1.28)	(1, 7)	1.00 (0.00)	(1, 1)	1.00 (0.00)	(1, 1)
0.01	0.006 (0.000)	1.08 (0.28)	(1, 3)	0.12 (0.32)	(0, 1)	1.00 (0.00)	(1, 1)	0.00 (0.00)	(0, 0)
0.001	0.003 (0.000)	0.11 (0.33)	(0, 2)	0.01 (0.07)	(0, 1)	0.00 (0.00)	(0, 0)	0.00 (0.00)	(0, 0)
0	NA	0.06 (0.24)	(0, 1)	0.00 (0.03)	(0, 1)	0.00 (0.00)	(0, 0)	0.00 (0.00)	(0, 0)

Abbreviation: NA, not applicable.
[a]The time since divergence is given in units of $2N_e$ generations.

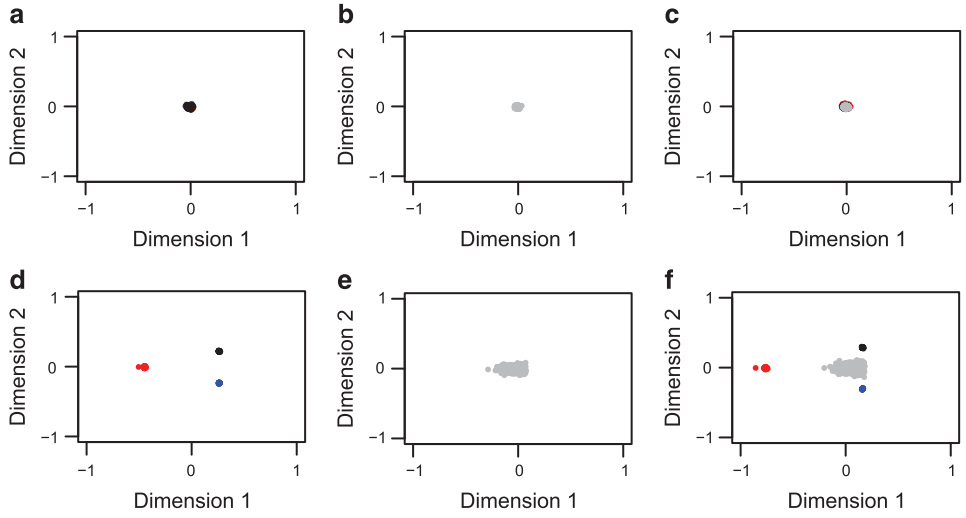

Figure 3 Representative projections of simulated data for three populations. (**a**–**c**) The divergence event between populations *B* (blue circles) and *C* (black circles) occurred $0.0002N_e$ generations ago and the divergence of population *A* (red circles) occurred $0.002N_e$ generations ago. (**d**–**f**) The divergence event between populations *B* and *C* occurred $2N_e$ generations ago and the divergence of population *A* occurred $20N_e$ generations ago. (**a**, **d**) Analysis of populations *A*, *B* and *C*. (**b**, **e**) Analysis of admixed individuals (gray circles) with average individual admixture proportions 10% population *A*, 45% population *B* and 45% population *C*. (**c**, **f**) Combined analysis of admixed individuals, population *A*, population *B* and population *C*.

Table 2 Number of significant principal components in the simulation of three populations

t_1[a]	t_2[a]	F_{ST} (AC)	F_{ST} (BC)	*EIGENSOFT*				*Minimum average partial test*			
				Population stratification		*Admixture*		*Population stratification*		*Admixture*	
		Mean (s.d.)	*Mean (s.d.)*	*Mean (s.d.)*	*Range*	*Mean (s.d.)*	*Range*	*Mean (s.d.)*	*Range*	*Mean (s.d.)*	*Range*
1	10	0.341 (0.006)	0.218 (0.005)	4.39 (1.25)	(2, 8)	15.64 (2.69)	(8, 27)	2.01 (0.07)	(2, 3)	2.00 (0.00)	(2, 2)
0.1	1	0.050 (0.001)	0.029 (0.001)	2.90 (0.84)	(2, 6)	6.23 (1.72)	(2, 12)	2.00 (0.05)	(2, 3)	2.00 (0.00)	(2, 2)
0.01	0.1	0.010 (0.000)	0.007 (0.000)	3.36 (0.96)	(2, 7)	1.98 (0.91)	(1, 5)	2.01 (0.09)	(2, 3)	0.56 (0.50)	(0, 1)
0.001	0.01	0.005 (0.000)	0.003 (0.000)	1.51 (0.66)	(1, 5)	0.27 (0.51)	(0, 3)	1.00 (0.03)	(1, 2)	0.00 (0.00)	(0, 0)
0.0001	0.001	0.003 (0.000)	0.003 (0.000)	0.28 (0.49)	(0, 2)	0.00 (0.06)	(0, 1)	0.00 (0.00)	(0, 0)	0.00 (0.00)	(0, 0)

[a]Times since divergence are given in units of $2N_e$ generations.

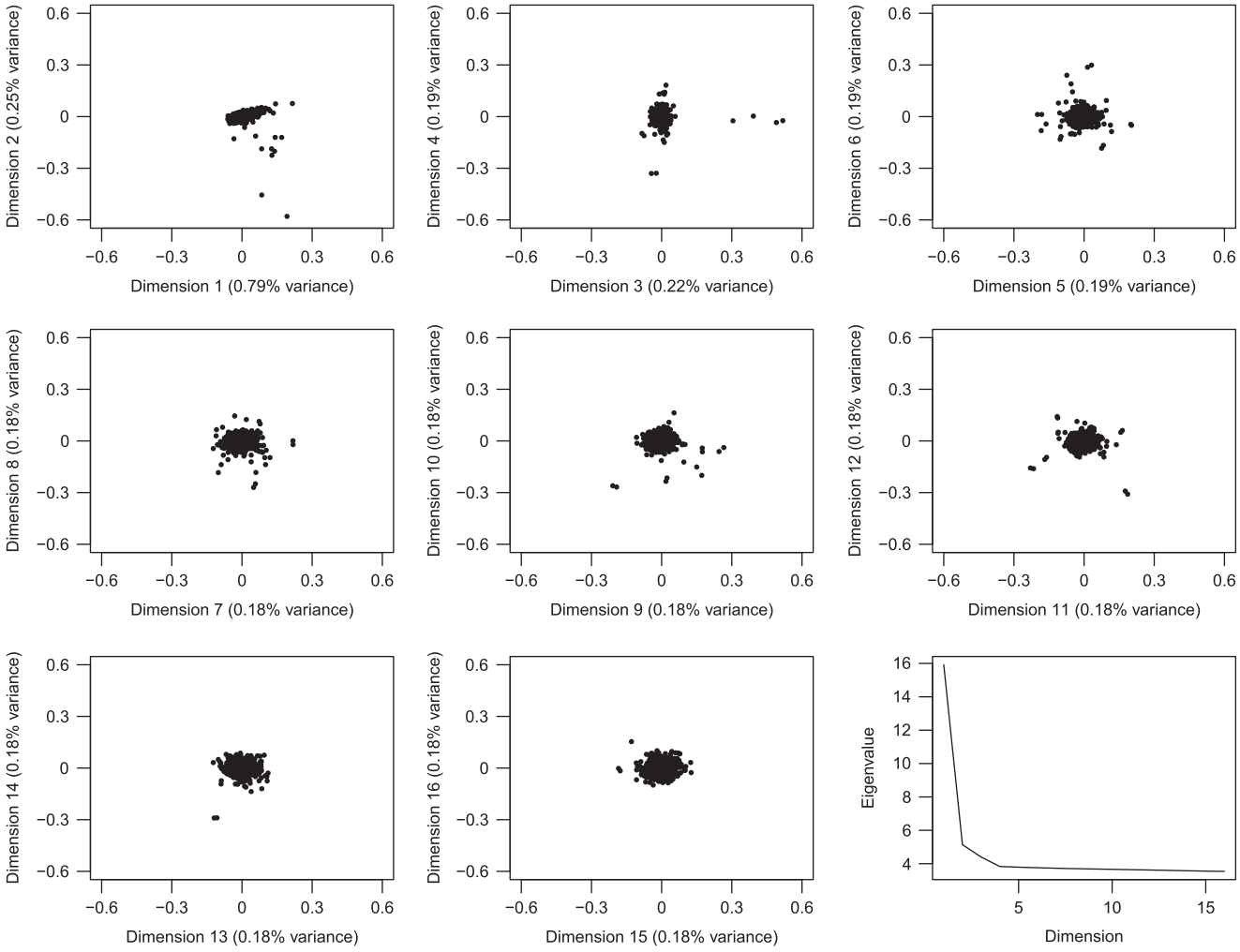

Figure 4 Top 16 principal components for the Howard University Family Study data using EIGENSOFT. All 16 principal components are statistically significant according to Tracy–Widom statistics. The bottom right panel shows the scree plot.

loadings to shrink toward zero. Whereas sparse factor analysis involves numerical optimization, which is stochastic, and can therefore yield different results from run to run, Velicer's minimum average partial test is deterministic and need only be run once. Also, sparse factor analysis does not provide a rule for determining the number of principal components to retain, whereas both Velicer's minimum average partial test and EIGENSOFT implement stopping rules.

EIGENSOFT was designed to work with genome-wide genotype data while accounting for linkage disequilibrium. The model of admixture used in the simulations assumes free recombination. Furthermore, the model assumes that haplotypes in the admixed individuals are inherited identical by descent from the parental populations, regardless of the number of generations since admixture. Taken together, these modeling assumptions eliminate the possibility that the observed overestima-

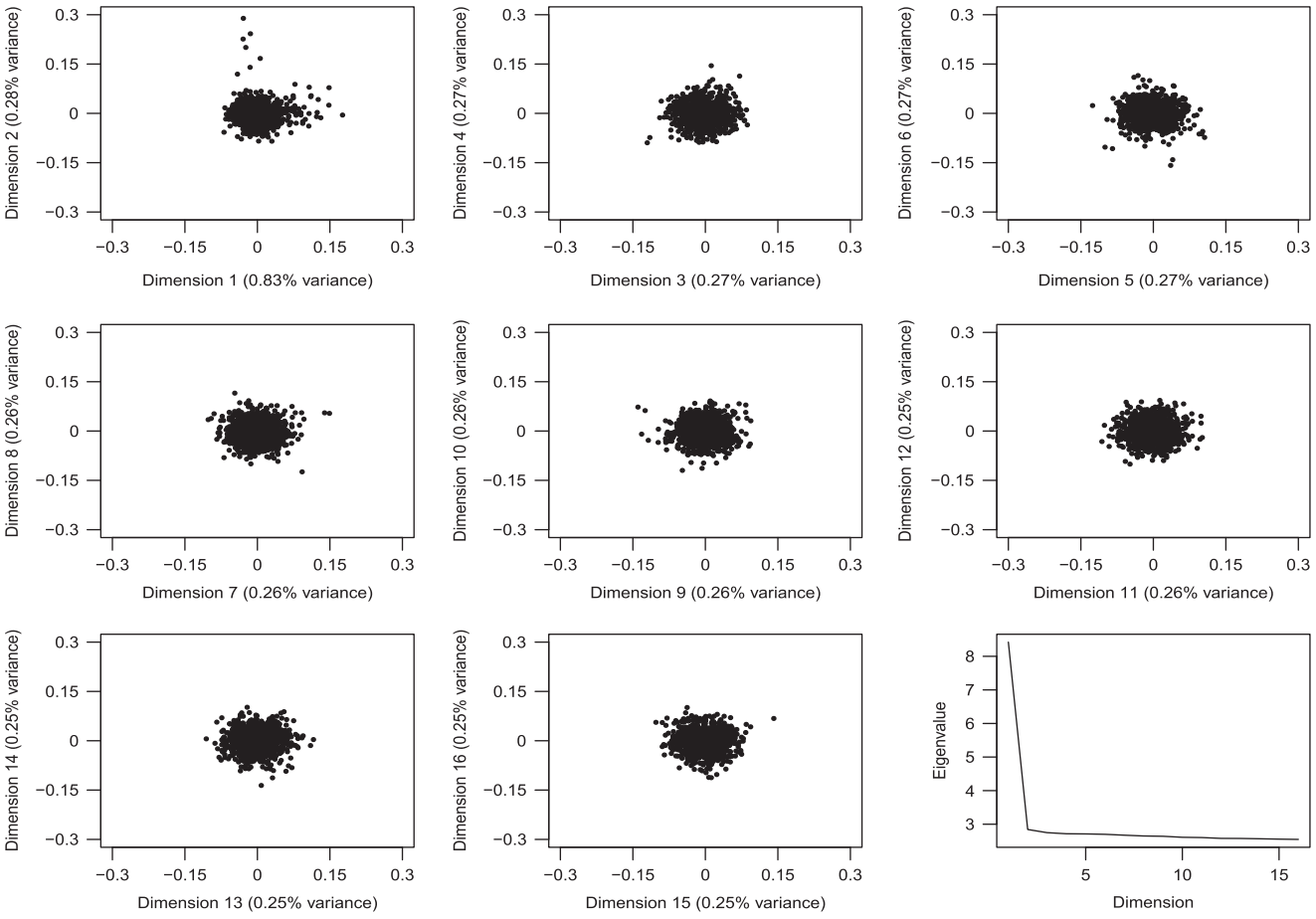

Figure 5 Top 16 principal components for the Howard University Family Study data using Velicer's minimum average partial test. Only the top principal component is statistically significant. The bottom right panel shows the scree plot.

tion of significant principal components by EIGENSOFT resulted from a failure to properly account for linkage disequilibrium. My procedure works well with 10 000 unlinked random markers (Gao and Martin, 2009; McVean, 2009), so that researchers can afford to aggressively prune genome-wide data sets to obtain a set of markers in linkage equilibrium and not have to account further for linkage disequilibrium (either background linkage disequilibrium or extended linkage disequilibrium due to admixture, both of which can induce spurious clustering (Falush *et al*., 2003; Kaeuffer *et al*., 2007)). At the same time, this number of markers is much greater than the number of individuals in the vast majority of genome-wide studies, such that sensitivity to detect population structure is limited by sample size (Patterson *et al*., 2006).

Unequal sample sizes can severely distort projections in principal component analysis (Novembre *et al*., 2008; McVean, 2009). This is a problem for analysis of admixed individuals if smaller reference samples of parental populations are included in the data set. On the one hand, this problem can be circumvented because analysis of admixed individuals does not require reference samples of parental populations (see Figures 2e and 3e). On the other hand, my procedure is less sensitive for detecting admixture without reference samples of parental populations than it is for detecting population stratification in samples of unadmixed individuals.

EIGENSOFT is known to overestimate the number of significant principal components for samples of admixed individuals (Patterson *et al*., 2006). This overestimation is not solely due to the use of ancestrally informative markers nor to extended linkage disequilibrium induced by admixture as previously suggested (Patterson *et al*., 2006), since a similar overestimation was observed in this study using random, unlinked markers. There are three major differences between EIGENSOFT and my procedure that might explain this problem. One, EIGENSOFT standardizes the sample covariance matrix, estimated from the genotype matrix, using the binomial variance of estimated allele frequencies, whereas Velicer's minimum average partial test uses the correlation matrix estimated from the genotype matrix. Thus, EIGENSOFT assumes Hardy–Weinberg equilibrium whereas my procedure does not. The assumption of Hardy–Weinberg equilibrium does not hold under either population stratification or admixture. Violation of Hardy–Weinberg equilibrium may partially explain why EIGENSOFT systematically overestimates the number of significant principal components.

Two, EIGENSOFT and my procedure use the results of eigendecomposition differently. EIGENSOFT performs formal hypothesis testing to determine the number of significant principal components by testing each eigenvalue using approximations to an external reference distribution, *viz*. the Tracy–Widom distribution

(Patterson *et al*., 2006). When assessing a given eigenvalue for significance, EIGENSOFT discards all leading eigenvalues; consequently, EIGENSOFT systematically overestimates the proportion of variance explained for every eigenvalue after the first one. In contrast, Velicer's minimum average partial test does not rely on any external reference distribution. Rather, it determines the number of principal components to retain using an objective minimization function, in which the stopping point is on the basis of data-dependent estimation of systematic noise (Velicer, 1976). Also, Velicer's minimum average partial test is based on partialing out the cumulative effect of all leading eigenvalues and eigenvectors from the original correlation matrix when it assesses a given eigenvalue and eigenvector for retention.

Three, EIGENSOFT estimates the effective number of markers on the basis of the empirical distribution of eigenvalues (Patterson *et al*., 2006). This value is used as a plug-in when estimating the mean and s.d. for the Tracy–Widom test statistic (Patterson *et al*., 2006). This moment estimator was derived under the null hypothesis of no structure (Patterson *et al*., 2006); its validity under the alternative hypothesis is unknown. These three differences may contribute to EIGENSOFT's overestimation of significance. Regardless, the principal components that are retained by Velicer's minimum average test can be used in the same way as those deemed significant by EIGENSOFT, for example, as covariates in association testing (Price *et al*., 2006).

Conflict of interest

The author declares no conflict of interest.

Acknowledgements

I thank Gil McVean for sharing R code to perform simulations of coalescent vicariance for two populations. The contents of this publication are solely the responsibility of the author and do not necessarily represent the official view of the National Institutes of Health. The Howard University Family Study was supported by National Institutes of Health grants S06GM008016-320107 to Charles Rotimi and S06GM008016-380111 to Adebowale Adeyemo. Participant enrollment was carried out at the Howard University General Clinical Research Center, supported by National Institutes of Health grant 2M01RR010284. Genotyping support was provided by the Coriell Institute for Medical Research. This research was supported by the Intramural Research Program of the Center for Research on Genomics and Global Health (CRGGH). The CRGGH is supported by the National Human Genome Research Institute, the National Institute of Diabetes and Digestive and Kidney Diseases, the Center for Information Technology, and the Office of the Director at the National Institutes of Health (Z01HG200362).

References

Adeyemo A, Gerry N, Chen G, Herbert A, Doumatey A, Huang H *et al.* (2009). A genome-wide association study of hypertension and blood pressure in African Americans. *PLoS Genet* **5**: e1000564.

Chen G, Shriner D, Zhou J, Doumatey A, Huang H, Gerry NP *et al.* (2010). Development of admixture mapping panels for African Americans from commercial high-density SNP arrays. *BMC Genomics* **11**: 417.

The International HapMap 3 Consortium. (2010). Integrating common and rare genetic variation in diverse human populations. *Nature* **467**: 52–58.

Engelhardt BE, Stephens M (2010). Analysis of population structure: a unifying framework and novel methods based on sparse factor analysis. *PLoS Genet* **6**: e1001117.

Falush D, Stephens M, Pritchard JK (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**: 1567–1587.

Gao X, Martin ER (2009). Using allele sharing distance for detecting human population stratification. *Hum Hered* **68**: 182–191.

Johnstone I (2001). On the distribution of the largest eigenvalue in principal components analysis. *Ann Stat* **29**: 295–327.

Kaeuffer R, Réale D, Coltman DW, Pontier D (2007). Detecting population structure using STRUCTURE software: effect of background linkage disequilibrium. *Heredity* **99**: 374–380.

Martinez-Marignac VL, Valladares A, Cameron E, Chan A, Perera A, Globus-Goldberg R *et al.* (2007). Admixture in Mexico City: implications for admixture mapping of type 2 diabetes genetic risk factors. *Hum Genet* **120**: 807–819.

McVean G (2009). A genealogical interpretation of principal components analysis. *PLoS Genet* **5**: e1000686.

Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A *et al.* (2008). Genes mirror geography within Europe. *Nature* **456**: 98–101.

O'Connor BP (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behav Res Methods Instrum Comput* **32**: 396–402.

Patterson N, Price AL, Reich D (2006). Population structure and eigenanalysis. *PLoS Genet* **2**: e190.

Peres-Neto PR, Jackson DA, Somers KM (2005). How many principal components? Stopping rules for determining the number of non-trivial axes revisited. *Comput Stat Data Anal* **49**: 974–997.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**: 904–909.

Price AL, Patterson N, Yu F, Cox DR, Waliszewska A, McDonald GJ *et al.* (2007). A genomewide admixture map for Latino populations. *Am J Hum Genet* **80**: 1024–1036.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D *et al.* (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**: 559–575.

R Development Core Team (2009). *R: A language and environment for statistical computing*. The R Foundation for Statistical Computing: Vienna, Austria.

Velicer WF (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika* **41**: 321–327.

Appendix

Let *dat* be the $M \times N$ matrix of genotypes for $i = 1$ to M SNPs and $j = 1$ to N individuals, with genotypes coded as 0, 1 or 2 copies of the minor allele. The following R function returns the number of principal components that minimizes the average squared partial correlation (O'Connor, 2000; R Development Core Team, 2009).

```
maptest <- function(dat) {
        mu <- apply(dat, 1, mean, na.rm = TRUE)
        dat <- dat - mu
        dat2 <- cor(dat, use = "complete.obs")
        a <- eigen(dat2)
        a$values[a$values < 0] <- 0
        b <- diag(a$values, nrow = length(a$values))
        loadings <- a$vectors %*% sqrt(b)
        partial <- function(x) {
                c <- loadings[,1:x]
                partcov <- dat2 - (c %*% t(c))
                d <- diag(partcov)
                if (any(is.element(NaN,d), is.element(0,d), length(d[d < 0])! = 0)) {
                        map <- 1
                } else {
                        d <- 1/(sqrt(d))
                        e <- diag(d, nrow = length(d))
                        pr <- e %*% partcov %*% e
                        map <- (sum(pr^2) - ncol(dat2))/(ncol(dat2) * (ncol(dat2) - 1))
                }
                return(map)
        }
        fm <- sapply(1:(ncol(dat2) - 1), partial)
        fm <- c((sum(dat2^2) - ncol(dat2))/(ncol(dat2) *    (ncol(dat2) - 1)), fm)
        return(max(1, which.min(fm) - 1))
}
```