Improved Eigenanalysis of Discrete Subpopulations and Admixture Using the Minimum Average Partial Test

Daniel Shriner

Center for Research on Genomics and Global Health, National Human Genome Research Institute, Bethesda, Md., USA

Key Words

Admixture · Population stratification · Population structure · Principal components analysis

Abstract

Principal components analysis of genetic data has benefited from advances in random matrix theory. The Tracy-Widom distribution has been identified as the limiting distribution of the lead eigenvalue, enabling formal hypothesis testing of population structure. Additionally, a phase change exists between small and large eigenvalues, such that population divergence below a threshold of F_{ST} is impossible to detect and above which it is always detectable. I show that the plug-in estimate of the effective number of markers in the EIGENSOFT software often exceeds the rank of the sample covariance matrix, leading to a systematic overestimation of the number of significant principal components. I describe an alternative plug-in estimate that eliminates the problem. This improvement is not just an asymptotic result but is directly applicable to finite samples. The minimum average partial test, based on minimizing the average squared partial correlation between individuals, can detect population structure at smaller F_{ST} values than the corrected test. The minimum average partial test is applicable to both unadmixed and admixed samples, with arbitrary numbers of dis-

crete subpopulations or parental populations, respectively. Application of the minimum average partial test to the 11 HapMap Phase III samples, comprising 8 unadmixed samples and 3 admixed samples, revealed 13 significant principal components.

Copyright © 2012 S. Karger AG, Basel

Introduction

The availability of dense, genome-wide genotype or sequence data has enabled powerful detection of population structure, which is critical to both medical genetics [1] and population genetics [2]. Principal components analysis is a model-free technique widely used for identifying population structure in genetic data. The interpretation of principal components analysis is straightforward when the data are sampled from a multivariate normal distribution, $N(\mathbf{0}, \Sigma)$.

Under the null hypothesis of no population structure, p independent and homoscedastic normal variables form a p -dimensional sphere [3]. Graphically, a scatter plot of any two principal components should reveal all data points clustering within a circle centered at the origin if there is no population structure. Statistically, the variance-covariance matrix Σ equals a constant variance σ^2

times the identity matrix under the null hypothesis, in which the identity matrix reflects independence (the off-diagonal values are all 0) and homoscedasticity (the diagonal values are all equal) [3]. A correlation matrix can also be tested for sphericity, but since a correlation matrix is by definition homoscedastic, testing sphericity of a correlation matrix tests only independence. In the case of discrete subpopulations, a scatter plot will reveal distinct clusters on either side of the origin. In the case of admixture, a scatter plot will reveal that the admixed individuals fall on a line defined by the parental populations [2, 4]. In both cases, population structure results in a sphere embedded in a p -dimensional ellipsoid [3].

EIGENSOFT is an implementation of principal components analysis for genetic data [1, 2]. In EIGENSOFT, the sample covariance matrix computed from genotype data is decomposed into mutually orthogonal eigenvectors, each with an associated eigenvalue that quantifies the proportion of variance explained [2]. In the notation of Patterson et al. [2], the centered genotype matrix $\mathbf{M}$ has dimensions $m \times n$ for m individuals and n markers. Traditional principal components analysis is based on decomposition of the $n \times n$ sample covariance matrix of the form $\mathbf{M}^T\mathbf{M}$, which reflects the pair-wise covariances between markers across all individuals. In contrast, EIGENSOFT is based on decomposition of the $m \times m$ sample covariance matrix of the form $\mathbf{M}\mathbf{M}^T$, which reflects the pair-wise covariances between individuals across all markers [2]. Thus, EIGENSOFT clusters individuals not markers. The eigenvectors are conventionally presented in order of decreasing eigenvalue. The null hypothesis of no structure can be formulated in terms of the eigenvalues: $\lambda_1 = \lambda_2 = \dots = \lambda_m$. Under the alternative hypothesis of population structure, not all eigenvalues are equal, with interest being in 'large' eigenvalues. According to random matrix theory, the Tracy-Widom distribution is the limiting distribution of the lead eigenvalue [5]. The three values required to calculate the test statistic to determine if the lead eigenvalue is 'large' are the lead eigenvalue and the dimensions of the genotype matrix [5]. After eigendecomposition, the number of individuals equals the number of eigenvalues but the nominal number of markers is no longer pertinent. Thus, a key step in EIGENSOFT is the estimation of the effective number of markers from the distribution of eigenvalues [2].

In this study, I describe 4 main observations. (1) The original moments estimator of the effective number of markers overestimates the effective number of markers. The original moments estimator yields inflated test statistics, leading to systematic overestimation of the amount of

population structure [6]. For population genetics studies, this can lead to incorrect inferences about group differences. For genetic association studies, this can lead to a loss of power by overcorrecting association testing. I describe a new moments estimator that fixes this problem. (2) Random matrix theory predicts the existence of a phase change with respect to 'small' and 'large' eigenvalues [7, 8]. Below a threshold level of divergence as measured by the summary statistic F_{ST} , population structure is conjectured to be difficult to detect [2]. Similarly, above the threshold, population structure is conjectured to be easily detectable [2]. Using the new moments estimator, even with finite samples, the transition is so sharp that statistical power to detect discrete subpopulations is essentially either 0 or 1. (3) Power to detect admixture is substantially lower than power to detect discrete subpopulations. (4) I recently described an alternative procedure called the minimum average partial test, which is based on minimizing the average squared partial correlation [6]. Here, I show that the minimum average partial test is more sensitive to smaller amounts of F_{ST} than the corrected test based on the Tracy-Widom distribution, especially for admixed samples. As an illustration, the minimum average partial test provides new insight regarding population structure for the 11 HapMap Phase III samples [9].

Materials and Methods

EIGENSOFT

Testing for the presence of population structure using eigenvalues and the Tracy-Widom distribution has been detailed by Patterson et al. [2]. I ported their C code to R [10], which requires the add-on package RMTstat [11]. Briefly, assume that the data consist of m individuals genotyped at n autosomal SNPs, with genotypes coded as 0, 1, or 2 copies of the variant allele, and $n > m$. Center the data by subtracting the mean genotype frequency for each marker (which extracts a matrix of rank one, leaving a matrix of rank $m' = m - 1$) and normalize the data by dividing by the expected binomial variance of allele frequencies for each marker. Then, perform eigendecomposition on the $m \times m$ sample covariance matrix. To calculate the test statistic, use the eigenvalues to estimate the effective number of markers n' . Rescale the lead eigenvalue λ_1 using

$$l = \frac{(m-1)\lambda_1}{\sum_{i=1}^{m-1} \lambda_i}.$$

Normalize l using

$$\mu(m', n') = \frac{(\sqrt{n'-1} + \sqrt{m'})^2}{n'},$$

$$\sigma(m', n') = \frac{\sqrt{n'-1} + \sqrt{m'}}{n'} \left(\frac{1}{\sqrt{n'-1}} + \frac{1}{\sqrt{m'}} \right)^{1/3},$$

and

$$x = \frac{l - \mu(m', n')}{\sigma(m', n')}.$$

The test statistic x approximately follows the Tracy-Widom distribution. For n' , Patterson et al. [2] use the moments estimator

$$\hat{n} = \frac{m'(m' + 2)}{\frac{m'^2 \sum_{i=1}^{m'} \lambda_i^2}{\left(\sum_{i=1}^{m'} \lambda_i\right)^2} - m'}$$

whereas I use the moments estimator

$$N_{\text{eff}} = \frac{\left(\sum_{i=1}^{m'} \lambda_i\right)^2}{\sum_{i=1}^{m'} \lambda_i^2}$$

[12]. To test for additional structure, drop the top k eigenvalues and test the remaining $m' - k$ eigenvalues, treating the data matrix as $(m' - k) \times (m' - k)$.

The Minimum Average Partial Test

The minimum average partial test was performed as previously described [6, 13, 14]. Despite the name, this procedure is not a formal hypothesis test rather it is an objective minimization function. Briefly, this procedure commences with the centered genotype matrix as described above for EIGENSOFT. Next, compute the $m \times m$ sample correlation matrix $\mathbf{R}$. Perform eigendecomposition on $\mathbf{R}$ to obtain the matrix of eigenvectors $\mathbf{\Pi}$ and the diagonal matrix of eigenvalues $\mathbf{\Lambda}$. Calculate the loadings as $\mathbf{\Pi}\mathbf{\Lambda}^{1/2}$. Let α_z be the first z loadings. Let $\mathbf{R}_z^* = \mathbf{D}^{-1/2} \mathbf{\epsilon} \mathbf{D}^{-1/2}$ be the $m \times m$ matrix of partial correlations after the first z loadings have been extracted, in which $\mathbf{D} = \text{diag } \mathbf{\epsilon}$ and $\mathbf{\epsilon} = \mathbf{R} - \alpha_z \alpha_z^T$. Let r_{kl}^* be the element in the k -th row and l -th column of $\mathbf{R}_z^*$. Define the summary statistic as

$$f_z = \sum_{k=1}^m \sum_{l=1}^m \frac{r_{kl}^{*2}}{m(m-1)},$$

which is the average of the squared partial correlations after the first z principal components have been extracted. The number of principal components to retain is the value of z for which f_z is a global minimum. If f_z is minimum at $z = 0$, then no principal components should be retained, indicating no population structure. If f_z is minimum at any other value of z , indicating population structure, then the first z principal components should be retained.

Coalescent Simulations

Data were simulated under a coalescent model of vicariance [4, 6]. Briefly, let A and B represent two populations that diverged at some time t in the past. A sample of haplotypes from populations A and B were simulated with divergence times $t = \{0, 0.001, 0.01, 0.1, 1.0\}$ in units of $2N_e$ generations, with N_e being the effective population size. Each data set consisted of 10,000 unlinked sites. Mutations were placed on a genealogy proportional to branch lengths. As a consequence of this mutational scheme, the effective population size N_e canceled out and sites were ascer-

tained to be polymorphic but not ancestrally informative. Haplotypes were randomly paired within each population to generate diploid individuals. For each divergence time t , 1,000 independent replicate data sets were generated.

Using these two samples to represent parental populations, a sample of two-way admixed individuals was generated. For each admixed individual, the average genome-wide admixture proportion p was determined by drawing a random deviate from the beta distribution $\text{Beta}(10.18, 2.84)$, yielding an expected genome-wide admixture proportion $\bar{p} = 0.782$, a value representative for admixed African Americans [15]. 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 $U(0,1) \leq p$, and assigned the state of a randomly selected haplotype from population B otherwise. For each divergence time t , 1,000 independent replicate data sets were generated. Under this scheme, because of free recombination, the expected covariance of allele frequencies is zero, i.e. there is neither background linkage disequilibrium nor extended linkage disequilibrium due to admixture (either of which would violate the independence assumption and lead to a distorted distribution of eigenvalues and eigenvectors [2]).

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. Other details remained the same as above. Using the three samples, a sample of three-way admixed individuals was generated. For each admixed individual, p_1 was a random deviate from $\text{Beta}(0.8, 7.2)$ and p_2 was a random deviate from $\text{Beta}(12, 12)$. For each site independently, the individual was assigned the state of a randomly selected haplotype from population A if a random deviate from $U(0,1) \leq p_1$. 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 $U(0,1) \leq p_2$, 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 $p_C = 1 - p_1 - (1 - p_1) p_2 = (1 - p_1) (1 - p_2) = 0.45$, respectively, intended to mimic African, European, and Native American admixture proportions in Latino populations [16, 17].

In the simulations with two parental populations (with or without admixture), the true number of principal components is either 0 if the divergence time is zero, or 1 if the divergence time is non-zero. Similarly, in the simulations with three parental populations (with or without admixture), the true number of principal components is 0 if both divergence times are zero, 1 if one divergence time is zero and the other is non-zero, and 2 if both divergence times are non-zero.

Real Data Analysis

Raw genotype data for the 11 HapMap Phase III samples (1,115 individuals and 1,615,203 markers) were accessed from http://hapmap.ncbi.nlm.nih.gov/downloads/genotypes/latest_phaseIII_ncbi_b36/plink_format/. All offspring were removed, leaving 924 individuals. Markers were filtered for a minor allele frequency of 0.05 and a genotyping call rate of 95%, leaving 1,048,624 markers. Markers were then pruned for linkage disequilibrium at

an r^2 threshold of 0.3, leaving 213,062 markers. Markers were filtered to be autosomal only, leaving 208,116 markers. From this set, 10,000 random markers were retained for principal components analysis.

Results

Moments Estimators of the Effective Number of Markers

Assume the data comprise a genotype matrix for m individuals and n markers, for which $n > m$. The genotype matrix has at most a full rank of $\min(m, n) = m$ and the sample covariance matrix also has at most a full rank of m . Assuming sphericity, i.e. the null hypothesis of no structure, Patterson et al. [2] derived a moments estimator of n as

$$\hat{n} = \frac{m(m+2)}{m^2 \sum_{i=1}^m \lambda_i^2 - \left(\sum_{i=1}^m \lambda_i \right)^2 - m}$$

in which $\hat{n}$ estimates the effective number of markers [2]. As an alternative estimator of n , let

$$N_{eff} = \frac{\left(\sum_{i=1}^m \lambda_i \right)^2}{\sum_{i=1}^m \lambda_i^2},$$

derived by matching the first two moments of the distribution of eigenvalues to the first two moments of a χ^2 distribution, with $1 \leq N_{eff} \leq m$ [12]. By substitution,

$$\hat{n} = \frac{m+2}{\frac{m}{N_{eff}} - 1}.$$

Given that N_{eff} is positive, $\hat{n}$ is always greater than N_{eff} .

Next consider the bounds of $\hat{n}$ and N_{eff} . If $\lambda_1 \neq 0$ and $\lambda_i = 0$ for $2 \leq i \leq m$, then

$$N_{eff} = \frac{\lambda_1^2}{\lambda_1^2} = 1$$

whereas

$$\hat{n} = \frac{m+2}{m-1} > 1.$$

More generally, if the first k eigenvalues are equal and non-zero and the remaining $m - k$ eigenvalues are zero,

then the expected value for the effective number of markers is k . For all such configurations, $N_{eff} = k$, indicating that N_{eff} is unbiased and $\hat{n}$ is biased upwards. If all m eigenvalues are equal, then $N_{eff} = m$, indicating no population structure, whereas $\hat{n} = \infty$, which is invalid because it exceeds m . Furthermore, $\hat{n}$ exceeds the full rank m when

$$N_{eff} > \frac{m^2}{2m+2} \approx \frac{m}{2}.$$

If the first k eigenvalues are unequal and non-zero and the remaining $m - k$ eigenvalues are zero, then $1 \leq N_{eff} < k$, but the expected value is unknown, so it is unclear if N_{eff} is also unbiased for these configurations.

To establish the consequences of overestimating the effective number of markers, consider the mean and the standard deviation used in the test statistic based on the Tracy-Widom distribution. The mean is given by

$$\mu = \frac{(\sqrt{n-1} + \sqrt{m})^2}{n}$$

[5], according to which μ is monotonically decreasing as a function of n . Therefore, if n is overestimated, μ is underestimated and the test statistic is overestimated. The standard deviation is given by

$$\sigma = \frac{\sqrt{n-1} + \sqrt{m}}{n} \left(\frac{1}{\sqrt{n-1}} + \frac{1}{\sqrt{m}} \right)^{1/3}$$

[5], according to which σ is also monotonically decreasing as a function of n . Therefore, if n is overestimated, σ is also underestimated and the test statistic is again overestimated.

The numbers of significant principal components were evaluated via simulation experiments. In the analysis of two discrete subpopulations (table 1) or two-way admixed individuals (table 2), the use of $\hat{n}$ led to overestimation, whereas the use of N_{eff} led to nearly no bias and no variance. For a fixed value of F_{ST} , detecting discrete subpopulations was more powerful than detecting admixture (tables 1 and 2). Similar behaviors were observed in the analysis of three discrete subpopulations (table 3) and three-way admixed individuals (table 4). Overestimation resulting from the use of $\hat{n}$ worsened as population divergence increased.

The Phase Change

Based on random matrix theory [7, 8], Patterson et al. [2] conjectured that a phase change existed such that the test statistic for the lead eigenvalue tends to the null distribution below some threshold and tends to positive in-

Table 1. Simulation results for two discrete subpopulations

t	Rank ^a	F_{ST}	Old moments estimator					New moments estimator				
			1st $\hat{n}$	2nd $\hat{n}$	max. $\hat{n}$	mean number of PCs	range	1st N_{eff}	2nd N_{eff}	max. N_{eff}	mean number of PCs	range
1	200	0.117	17.6	4,637.3	∞	1.32	[1, 4]	16.0	189.8	190.3	1.00	[1, 1]
0.1	200	0.025	334.3	4,744.4	∞	1.18	[1, 4]	124.2	190.0	190.5	1.00	[1, 1]
0.01	200	0.006	3,598.9	4,705.5	∞	1.09	[1, 3]	188.5	189.9	190.4	0.00	[0, 0]
0.001	200	0.003	4,711.9	4,765.5	∞	0.09	[0, 2]	190.9	190.0	191.3	0.00	[0, 0]
0	200	0.000	4,730.8	4,783.3	∞	0.06	[0, 2]	190.9	190.1	191.3	0.00	[0, 0]

^a A sample consisted of 100 individuals from each subpopulation and 10,000 unlinked markers.

t = Time since the split in units of $2N_e$ generations; max. = maximum; PC = principal component.

Table 2. Simulation results for two-way admixed individuals

t	Rank ^a	F_{ST}	Old moments estimator					New moments estimator				
			1st $\hat{n}$	2nd $\hat{n}$	max. $\hat{n}$	mean number of PCs	range	1st N_{eff}	2nd N_{eff}	max. N_{eff}	mean number of PCs	range
1	1,000	0.117	5,394.4	8,601.9	∞	3.11	[1, 8]	842.4	894.1	897.0	1.00	[1, 1]
0.1	1,000	0.025	8,953.0	9,173.2	∞	1.60	[1, 5]	898.5	899.9	901.8	0.00	[0, 0]
0.01	1,000	0.006	9,602.7	9,622.4	∞	0.15	[0, 2]	904.7	904.0	905.9	0.00	[0, 0]
0.001	1,000	0.003	9,762.0	9,780.3	∞	0.00	[0, 0]	906.1	905.4	907.1	0.00	[0, 0]

^a A sample consisted of 1,000 admixed individuals and 10,000 unlinked markers.

t = Time since the split in units of $2N_e$ generations; max. = maximum; PC = principal component.

Table 3. Simulation results for three discrete subpopulations

t_1	t_2	Rank ^a	F_{ST}		Old moments estimator						New moments estimator					
			(AC)	(BC)	1st $\hat{n}$	2nd $\hat{n}$	3rd $\hat{n}$	max. $\hat{n}$	mean num- ber of PCs	range	1st N_{eff}	2nd N_{eff}	3rd N_{eff}	max. N_{eff}	mean num- ber of PCs	range
10	1	300	0.341	0.218	5.9	14.6	3,687.9	∞	4.43	[2, 10]	5.8	13.8	274.7	276.5	2.00	[2, 2]
1	0.1	300	0.049	0.029	149.5	639.6	4,747.0	∞	2.85	[2, 7]	99.2	202.8	279.4	280.5	2.00	[2, 2]
0.1	0.01	300	0.016	0.007	1,324.5	4,004.5	4,475.7	∞	3.37	[2, 7]	243.6	277.2	278.4	279.5	1.00	[1, 1]
0.01	0.001	300	0.006	0.003	3,880.1	4,569.0	4,607.4	∞	1.52	[1, 4]	277.5	279.6	278.9	280.4	0.00	[0, 0]
0.001	0.0001	300	0.003	0.003	4,707.4	4,744.5	4,778.4	∞	0.27	[0, 4]	281.0	280.3	279.5	281.6	0.00	[0, 0]

^a A sample consisted of 100 individuals from each subpopulation and 10,000 unlinked markers.

t_1 = Time since the split of population A from populations B and C in units of $2N_e$ generations; t_2 = time since the split of populations B and C in units of $2N_e$ generations; max. = maximum; PC = principal component.

Table 4. Simulation results for three-way admixed individuals

t_1	t_2	Rank ^a	F_{ST}		Old moments estimator						New moments estimator					
			(AC)	(BC)	1st $\hat{n}$	2nd $\hat{n}$	3rd $\hat{n}$	max. $\hat{n}$	mean num- ber of PCs	range	1st N_{eff}	2nd N_{eff}	3rd N_{eff}	max. N_{eff}	mean num- ber of PCs	range
10	1	1,000	0.341	0.218	1,367.5	6,107.3	7,227.5	∞	15.63	[7, 24]	575.6	857.5	875.9	881.2	2.00	[2, 2]
1	0.1	1,000	0.049	0.029	8,043.1	8,549.5	8,597.4	∞	6.25	[2, 13]	888.4	893.5	893.2	896.6	0.00	[0, 1]
0.1	0.01	1,000	0.016	0.007	8,958.3	9,005.5	9,027.1	∞	1.98	[1, 6]	898.6	898.3	897.7	901.3	0.00	[0, 0]
0.01	0.001	1,000	0.006	0.003	9,245.9	9,265.1	9,282.7	∞	0.21	[0, 2]	901.4	900.8	900.1	903.3	0.00	[0, 0]
0.001	0.0001	1,000	0.003	0.003	9,602.1	9,620.2	9,637.4	∞	0.01	[0, 1]	904.7	904.0	903.4	905.8	0.00	[0, 0]

^a A sample consisted of 1,000 admixed individuals and 10,000 unlinked markers.

t_1 = Time since the split of population A from populations B and C in units of $2N_e$ generations; t_2 = time since the split of populations B and C in units of $2N_e$ generations; max. = maximum; PC = principal component.

finitly above that threshold. In the context of population genetics, they conjectured that for two discrete subpopulations of equal size, the threshold in terms of F_{ST} is equal to the reciprocal of the geometric mean of the number of individuals m and the number of markers n ,

$$F_{ST} \approx \frac{1}{\sqrt{mn}}$$

[2]. However, the upper bound of the rank of the sample covariance matrix is $\min(m, n)$. Assuming $n > m$, the estimate of n from the sample covariance matrix should not exceed m , so that the threshold occurs at

$$F_{ST} \approx \frac{1}{m}.$$

Based on simulations of $m = 200$ individuals comprising two equal-sized discrete subpopulations such that $m_s = 100$, the phase change is observable at

$$F_{ST} = \frac{1}{100} = \frac{1}{m_s}$$

(fig. 1a). The threshold is higher for admixture than for discrete subpopulations (fig. 1b), indicating that the test is more powerful for detecting discrete subpopulations than admixture.

Performance Comparison

I previously described a procedure to detect population structure based on the minimum average partial test [6]. The minimum average partial test determines the number of principal components to retain using an objective minimization function of the squared partial correlations [13]. Briefly, the minimum average partial test cumulatively ex-

tracts principal components until the average squared partial correlation is globally minimized [13]. Under the null hypothesis of no population structure (i.e. $F_{ST} = 0$), the minimum average partial test detected population structure in 0 of 1,000 independent replicates (see table 1 in [6] and fig. 1). The minimum average partial test shows a similar phase change behavior but at a lower F_{ST} threshold, i.e. the minimum average partial test is more sensitive than the test based on the Tracy-Widom distribution (fig. 1a). For example, given a sample comprised of two subpopulations both of size 100, the minimum average partial test can detect population structure for $F_{ST} \geq 0.0044$, whereas the test based on the Tracy-Widom distribution can detect population structure for $F_{ST} \geq 0.01$ (fig. 1a). Given a sample comprised of 100 admixed individuals with an average admixture proportion of 0.8, the minimum average partial test can detect population structure for $F_{ST} \geq 0.08$ between the parental populations, whereas the test based on the Tracy-Widom distribution can detect population structure for $F_{ST} \geq 0.6$ between the parental populations (fig. 1b). If the sample size is increased to 1,000 admixed individuals, the minimum average partial test can detect population structure for $F_{ST} \geq 0.013$ between the parental populations, whereas the test based on the Tracy-Widom distribution can detect population structure for $F_{ST} \geq 0.05$ between the parental populations. Thus, the minimum average partial test is relatively more sensitive than the test based on the Tracy-Widom distribution for admixed samples compared to samples of discrete subpopulations.

The test based on the Tracy-Widom distribution depends on eigenvalues, whereas the minimum average

Fig. 1. Sensitivity to F_{ST} . **a** Power to detect population structure in the form of discrete subpopulations. Samples consisted of 100 individuals from two discrete subpopulations. The dashed line represents the Tracy-Widom test and the solid line represents the minimum average partial test. **b** Power to detect population structure in the form of admixture. Samples consisted of 100 admixed individuals, with an average admixture proportion of 0.8. The dashed line represents the Tracy-Widom test and the solid line represents the minimum average partial test.

partial test depends on both eigenvalues and eigenvectors. To evaluate how the information provided in the eigenvalues and eigenvectors contributes to the sensitivity of the minimum average partial test, I performed two additional simulations. In the first simulation, all of the eigenvalues were set to 1 and the eigenvectors were kept intact. In the second simulation, the eigenvalues were kept intact and the coefficients within each eigenvector were randomly permuted. For samples consisting of two equal-sized subpopulations of 100 individuals each and with average $F_{ST} = 0.117$ between subpopulations, sensitivity in the first simulation was 1 and sensitivity in the second simulation was 0. Therefore, eigenvectors contribute substantially more information to detecting population structure than do eigenvalues in the minimum average partial test.

Real Data Analysis

To illustrate with real data, I analyzed the 11 HapMap Phase III samples [9] using the minimum average partial test and both versions of the test based on the Tracy-Widom distribution. The minimum average partial test yielded 13 significant principal components (fig. 2). For comparison, the original test based on the Tracy-Widom distribution yielded 36 significant principal components and the corrected test yielded 5. The African samples were separated from the Asian and European samples on the first dimension (fig. 2a). The Asian samples were separated from the European samples on the second dimension (fig. 2a). Across the first two dimensions, the African American individuals fell on a line defined by the African and European samples, whereas the GIH and MXL individuals fell on a line defined by the Asian and European samples (fig. 2a). The GIH sample was separated from all other samples on the third

dimension (fig. 2b). The GIH individuals all fell on a line, anchored at one end by all of the other samples (fig. 2b). The other anchor was not represented by any of the HapMap samples (fig. 2b). The three African samples were separated along the fourth dimension (fig. 2b). The MXL sample was separated from all other samples on the fifth dimension (fig. 2c). As with the GIH sample, the MXL individuals all fell on a line, anchored at one end by all of the other samples (fig. 2c). The other anchor was not represented by any of the HapMap samples (fig. 2c). Dimensions 6 and 8–13 reflected structure within the MKK sample (fig. 2c–g). The CHB and CHD samples were separated from the JPT sample on the seventh dimension (fig. 2d).

Additional insight into the ancestry of the admixed GIH and MXL samples can be gained (fig. 3). The GIH individuals fell on a line anchored by the European samples but not by the East Asian samples (compare fig. 2b to fig. 3a and b). Similarly, the MXL individuals fell on a line anchored by the European samples but not by the East Asian samples (compare fig. 2c to fig. 3c and d). There was also one African American individual (NA19625) who was pulled in the direction of the MXL sample (fig. 2a, c, and fig. 3d).

Discussion

In the original description of EIGENSOFT, Patterson et al. [2] predicted that ‘most large genetic datasets with human data will show some detectable population structure.’ Indeed, the threshold for the phase change is so low with both the corrected test based on the Tracy-Widom distribution and the minimum average partial test that population structure is observable even in small

Fig. 2. The 13 significant principal components from the HapMap Phase III samples. Population descriptors: ASW = African ancestry in Southwest USA; CEU = Utah residents with Northern and Western European ancestry from the CEPH collection; CHB = Han Chinese in Beijing, China; CHD = Chinese in metropolitan

Denver, Colorado; GIH = Gujarati Indians in Houston, Texas; JPT = Japanese in Tokyo, Japan; LWK = Luhya in Webuye, Kenya; MKK = Maasai in Kinyawa, Kenya; MXL = Mexican ancestry in Los Angeles, California; TSI = Tuscan in Italy; YRI = Yoruba in Ibadan, Nigeria.

to moderately sized samples. The new plug-in moments estimator N_{eff} solves the problem of the excess of dimensions originally reported by Patterson et al. [2], particularly noticeable for admixture. Admixture can be detected without reference to samples from the parental populations or even proxies thereof, but it is more difficult to detect than discrete subpopulations. Thus, statistical evidence for admixture requires considerably larger sample sizes if the parental populations have not been sampled.

Linkage disequilibrium distorts the distribution of eigenvalues and eigenvectors [2]. For both the test based on the Tracy-Widom distribution and the minimum average partial test, a set of p independent and homoscedastic multivariate normal random variables forms a p -dimensional sphere under the null hypothesis of no population structure, with the variance-covariance matrix equaling a constant times the identity matrix. Both background linkage disequilibrium and extended linkage disequilibrium due to admixture result in non-zero off-diagonal elements in the variance-covariance and

Fig. 3. Principal components analysis of the admixed GIH individuals (**a, b**) and the admixed MXL individuals (**c, d**). The population descriptors are provided in the legend to figure 2.

correlation matrices. Consequently, large eigenvalues may reflect linkage disequilibrium rather than population structure. Thus, it is critical to ensure that markers are independent by pruning the data before eigendecomposition.

The minimum average partial test is iterative in that it detects all significant principal components, but it does so without discarding samples. This is important for the interpretation of the three admixed HapMap Phase III samples. The admixed African Americans are well defined by the YRI and CEU samples over the first two dimensions. One outlier, ASW individual NA19625, was previously suggested to have a small amount of East Asian ancestry [9]. Rather, NA19625 is closer to the MXL sample. Examination of just the first two dimensions would seem to indicate that the GIH and MXL samples are well defined by the East Asian and European samples. Instead, dimensions 3 and 5 indicate that the admixed GIH and MXL individuals do not share ancestry with the East Asian samples but rather both have substantial ancestry with the European samples and populations not sampled in HapMap Phase III. For the MXL sample, the presumed other ancestral population is American Indian and the results indicate that the East Asian samples are

not good proxies. There is no significant evidence that the GIH or LWK samples are residually structured or include admixed individuals [9], but the MKK sample is highly residually structured.

The increased power of the minimum average test also revealed separation of the JPT sample from the CHB and CHD samples, despite the low values of F_{ST} of 0.0070 between the JPT and CHB samples and of 0.0080 between the JPT and CHD samples [9]. Thus, grouped East Asian samples (e.g. CHB + JPT) should be used with caution [18]. Although the sizes of the CEU and TSI samples (109 and 77, respectively) are currently not large enough to lead to separation, given $F_{ST} = 0.0040$ [9], only slight increases in sample size are necessary to achieve separation of these two samples as well. Thus, grouped European samples (i.e. CEU + TSI) [9] also should be used with caution.

In summary, I identify the source of systematic overestimation of the number of significant principal components in the test for population structure implemented in EIGENSOFT. I also show that the minimum average partial test is more powerful than the test based on the Tracy-Widom distribution, even more so for admixture than for discrete subpopulations. The minimum average partial

test is modular in that it can replace the test based on the Tracy-Widom distribution within EIGENSOFT. These tests are equally applicable to chip-based genotype data and whole-exome or whole-genome sequence data, subject to the issue presented by linkage disequilibrium. While data on self-reported ethnicity continues to be collected, genetic methods that function independently of self-identified labels and that powerfully detect both discrete subpopulations and admixture are needed, particularly as consortia for association studies of complex diseases expand to include multiple ancestries.

Acknowledgements

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. 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). I thank Ao Yuan for mathematical assistance, Nick Patterson, Adeyemo Adebowale, and Guanjie Chen for helpful discussions, Charles Rotimi for critically reviewing the manuscript, and the anonymous reviewers for their comments.

Appendix

Let our data be the $n \times m$ genotype matrix for n single nucleotide polymorphisms and m individuals, with genotypes coded as 0, 1, or 2 copies of the variant allele. The following R function writes the results of the Tracy-Widom test of lead eigenvalues to a file. Note that the add-on package RMTstat is required.

```
TWtest <- function(dat) {
  library(RMTstat)

  #estimate posterior allele frequencies
  p <- vector("numeric",nrow(dat))
  for (i in 1:nrow(dat)) p[i] <- (1+sum(dat[i,],na.rm=TRUE))/(2+2*sum(!is.na(dat[i,])))

  #center
  mu <- apply(dat,1,mean,na.rm=TRUE)
  dat <- dat - mu
  dat[is.na(dat)] <- 0

  #normalize
  for (i in 1:nrow(dat)) dat[i,] <- dat[i,]/sqrt(p[i]*(1-p[i]))

  #perform eigendecomposition of the covariance matrix
  a <- eigen(cov(dat))
  m <- length(a$values)

  #test each dimension for
  (j in 1:(m-1)) {
    L1 <- sum(a$values[j:m])
    L2 <- sum(a$values[j:m]^2)
    lambda <- a$values[j]*(m-j)/L1
    nhathat <- L1^2/L2
    mu <- (sqrt(nhathat-1)+sqrt(m-1))^2/nhathat
    sigma <- (sqrt(nhathat-1)+sqrt(m-1))/nhathat*(1/sqrt(nhathat-1)+1/sqrt(m-1))^(1/3)
    twstat <- (lambda-mu)/sigma
    twpvalue <- ptw(twstat,lower.tail=FALSE)
    write.table(c(j,lambda,nhathat,twstat,twpvalue),
               "results.txt",row.names=FALSE,col.names=c("Dimension","Eigenvalue",
               "nhathat","TWstat","P-value"),quote=FALSE,sep="\t")
  }
}
```

References

- 1 Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D: Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 2006;38:904–909.
- 2 Patterson N, Price AL, Reich D: Population structure and eigenanalysis. *PLoS Genet* 2006;2:e190.
- 3 Basilevsky A: *Statistical Factor Analysis and Related Methods: Theory and Applications*. New York, John Wiley & Sons, Inc., 1994.
- 4 McVean G: A genealogical interpretation of principal components analysis. *PLoS Genet* 2009;5:e1000686.
- 5 Johnstone I: On the distribution of the largest eigenvalue in principal components analysis. *Ann Stat* 2001;29:295–327.
- 6 Shriner D: Investigating population stratification and admixture using eigenanalysis of dense genotypes. *Heredity* 2011;107:413–420.
- 7 Baik J, Ben Arous G, Pécché S: Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Ann Probab* 2005;33:1643–1697.
- 8 Baik J, Silverstein JW: Eigenvalues of large sample covariance matrices of spiked population models. *J Multivariate Anal* 2006;97:1382–1408.
- 9 The International HapMap 3 Consortium: Integrating common and rare genetic variation in diverse human populations. *Nature* 2010;467:52–58.
- 10 R Development Core Team: *R: a language and environment for statistical computing*. Vienna, The R Foundation for Statistical Computing, 2009.
- 11 Johnstone IM, Perry PO, Ma Z, Shahram M: RMTstat: distributions, statistics, and tests derived from random matrix theory, version 0.2, 2009.
- 12 Bretherton CS, Widmann M, Dymnikov VP, Wallace JM, Bladé I: The effective number of spatial degrees of freedom of a time-varying field. *J Climate* 1999;12:1990–2009.
- 13 Velicer WF: Determining the number of components from the matrix of partial correlations. *Psychometrika* 1976;41:321–327.
- 14 O'Connor BP: SSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behav Res Methods Instrum Comput* 2000;32:396–402.
- 15 Chen G, Shriner D, Zhou J, Doumatey A, Huang H, Gerry NP, Herbert A, Christman MF, Chen Y, Dunston GM, Faruque MU, Rotimi CN, Adeyemo A: Development of admixture mapping panels for African Americans from commercial high-density SNP arrays. *BMC Genomics* 2010;11:417.
- 16 Martinez-Marignac VL, Valladares A, Cameron E, Chan A, Perera A, Globus-Goldberg R, Wachter N, Kumate J, McKeigue P, O'Donnell D, Shriver MD, Cruz M, Parra EJ: Admixture in Mexico City: implications for admixture mapping of type 2 diabetes genetic risk factors. *Hum Genet* 2007;120:807–819.
- 17 Price AL, Patterson N, Yu F, Cox DR, Waliszewska A, McDonald GJ, Tandon A, Schirmer C, Neubauer J, Bedoya G, Duque C, Villegas A, Bortolini MC, Salzano FM, Gallo C, Mazzotti G, Tello-Ruiz M, Riba L, Aguilar-Salinas CA, Canizales-Quinteros S, Menjivar M, Klitz W, Henderson B, Haiman CA, Winkler C, Tusie-Luna T, Ruiz-Linares A, Reich D: A genomewide admixture map for Latino populations. *Am J Hum Genet* 2007;80:1024–1036.
- 18 The International HapMap Consortium: A second generation human haplotype map of over 3.1 million SNPs. *Nature* 2007;449:851–861.