Population Structure and Eigenanalysis
by Nick Patterson, Alkes L. Price and David Reich

One question that arises when analyzing a genetic dataset is whether or not the samples are drawn from a homogeneous population. If not, then identifying the structure inherent in the data is important. Principal component analysis (PCA or eigenanalysis) is a standard tool in genetics, traditionally applied to data at a population level. This paper looks at applying eigenanalysis to individuals to investigate possible structure in datasets; the main technique developed is a test for whether eigenvectors from the analysis reflect real structure in the data or are merely due to noise. In addition, the authors find a calculable threshold for population size above which detection of structure is easy, enabling decisions on how much data is required to find population structure for a given level of genetic divergence. We will only discuss the eigenanalysis statistics, and not the construction of the threshold.

We assume that we have data on n biallelic markers from m individuals, and construct matrix C such that C_{ij} is the number of variant alleles for marker j, individual i. For example, if each individual has two chromosomes, then $C_{ij} \in \{0, 1, 2\}$. Let M be the normalized version of C that corrects for genetic drift and ensures that each data column has the same variance. Then we compute an eigenvector decomposition of the $m \times m$ matrix $X = \frac{1}{n} M M^T$, which is the sample covariance of the columns of M. Eigenvectors corresponding to "large" eigenvalues indicate nonrandom population structure; the question, of course, is determining what "large" means.

If M has $m < n$ with each entry an independent standard normal random variable, and we order the eigenvalues of X so that $\lambda_1 > \lambda_2 > \ldots > \lambda_m$, then Johnstone [1] shows that with suitable normalization and for large m and n, the largest eigenvalue λ_1 is well approximated by the Tracy-Widom distribution.

For genetic applications it cannot be assumed that the markers are unlinked and independent. Nonindependence of the columns will reduce the effective sample size; the authors give a recommended moments estimator n' to use instead of n to mitigate the effect of nonindependence.

We give an overview of part of the main theorem of the paper, which investigates the eigenvalues of the theoretical covariance matrix of the counts of the variant allele, in the case where we are sampling a marker from samples belonging to K populations. We assume the frequency of the allele in population i is P_i, that sample j belongs to population $i = i(j)$, and that the sample size for population i is $M(i)$. We let the covariance of the population frequency vector be $\mathbf{p} = (\mathbf{p_1}, \mathbf{p_2}, \ldots, \mathbf{p_k})$ and assume that there is a hidden allele frequency P with diffuse distribution across the unit interval. Conditional on P we assume that $\mathbf{p}$ has mean $P(1, 1 \ldots, 1)$ and covariance matrix $P(1-P)B$ where B is independent of P. For small population divergence, we can take the diagonal entry of B_{ii} as the divergence between P and p_i. Let $\tau_i = B_{ii}$ and assume that all the τ_i are of the order of τ, which is small. Conditional on $\mathbf{p}$, the C_j are independent, and C_j has mean p and variance $2p(1-p)$ where $p = p_{i(j)}$.

Theorem: With the above assumptions, let V^* be the covariance matrix of C^*, which is C normalized to have 0 mean. Let $\tilde{V} = \frac{V^*}{2P(1-P)}$. Then conditional on the root frequency P, $\tilde{V}$ has for each k $(1 \leq k \leq K)$, $M(k)-1$ eigenvalues equal to $1 - \tau_k$.

Sketch of proof:
Let V be the covariance of the matrix of counts C, which can be viewed as a linear operator. The covariance structure depends only on the population labels of the samples, so it follows that the vector space of column vectors of length M has an orthogonal decomposition into subspaces invariant under V consisting of:

1. A K-dimensional subspace F of vectors whose coordinates are constant within a population.
2. Subspaces S_k $(1 \leq k \leq K)$ whose vectors are zero on samples not belonging to population i, and have coordinate sum 0, so are orthogonal to F.

It follows that V has K eigenvectors in F, and for each k, $M(k) - 1$ eigenvectors in S_k, each of which have the same eigenvalue λ_k. Conditional on $\mathbf{p}$, V acts on S_k as $2p_k(1-p_k)I$, where I is the identity matrix. Then $\lambda_k = E(2p_k(1-p_k)|P)$. Since $E(p_k^2|P) = P^2 + P(1-P)\tau_k$, the eigenvalues corresponding to the eigenvectors of S_k are $\lambda_k = 2P(1-P)(1-\tau_k)$. V^* and V act identically on S_k, so we're done.

[1] Johnstone, I (2001) On the distribution of the largest eigenvalue in principal component analysis. *Ann Stat*, 29: 295-327.