Principal components analysis of population genetic data

Graham Coop[1]

[1] Department of Evolution and Ecology & Center for Population Biology,

We have a dataset consisting of N individuals at S sites (single nuc. polymorphisms, SNPs). At each SNP an individual's data takes 0,1, or 2 (corresponding to the number of copies of an allele an individual carrys at this SNP). We can think of this as a N x S matrix (where usually $N \ll S$).

We usually in population genetics standardize the columns of this matrix (the info. from each SNP) in the following way: Denoting the mean allele freq at SNP j by ϵ_j, at each SNP we take the minus off the mean (2ϵ) and divide through by the expected variance assuming that alleles are sampled binomial from the mean frequency ($\epsilon(1-\epsilon)$). Call this data matrix X, it is an N x S matrix (where usually $N \ll S$).

To do principal components we'd usually do the eigenvalue decomposition

$$X^T X = P \Lambda P^T \tag{1}$$

P is an $S \times S$ matrix, Λ is an $N \times N$ matrix. Then we can calculate the genome-wide projection of an individual i on the j^{th} principal component as

$$Z_{i,j} = \sum_{s=1}^{S} P_{js} X_{is} \tag{2}$$

However, our matrices are usually so big that we generally can not do the eigen-decomp (eqn. 1) as $X^T X$ is too big, so we do instead

$$X X^T = Q \Lambda Q^T \tag{3}$$

where Q is a $S \times S$ matrix, Λ is an NxN matrix, this is the eigen-decomposition of the individual by individual covariance matrix. The k column of this matrix is the position of each individual on the k eigen-vector (this is the projection given by 2). We often plot individuals against each other on the first few principal components to visualize the major axes of variation and explore population structure.

some further notes The two eigen-value decompositions are closely related to each other as

$$X = Q \Lambda^{1/2} P^T \tag{4}$$

which is the singular value decomposition of X (if X is symmetric and square then eqn. 1 and eqn. 3 are the same and are the square of the SVD (4)). Helpfully we can move between these projections. To obtain the j^{th}

$$Q_j^T X = (Q_k^T Q)\Lambda^{1/2} P^T = \sqrt{(\lambda_j)} P_j \tag{5}$$

and

$$(XP)_j = (Q\Lambda^{1/2} P^T P)_j = Q_j \sqrt{(\lambda_j)} \tag{6}$$