Admixture / PCA notes

Benjamin Peter

June 6, 2022

1 Introduction

Isolated populations are expected to slowly diverge (Nielsen and Wakeley, 2001). If this isolation persists, populations will eventually differ so much that they are no longer able to reproduce, resulting in the (allopatric) formation of new species. However, this process of speciation usually takes millions of years, and so it is quite common that diverged population meet again, and produce offspring populations with both ancestries. One major revelation of the genomic reovlution for evolution has been that such admixture is quite common, and many phenotypically well-differentiated species occasionally interbreed.

2 Theory

2.1 Gradients vs Admixture

2.2 Known Source Populations

This is the simplest case. Let us assume we have data from p SNPs from n populations, so that X_i is the p-dimensional vector with population allele frequencies in population i, and we could store all data in a matrix $\mathbf{X}_{[n\times p]}$ where the subscript denotes the dimension, and x_{ij} is the allele frequency of the j-th SNP in the i-th population.

2.2.1 2-way admixture

Ignoring the effects of genetic drift, let us assume that population X is the product of admixture between populations S_1 and S_2 then,

$$X = \alpha S_1 + (1-\alpha)S_2, \tag{1}$$

so that the allele frequencies of the admixed population are a linear combination of the source populations. This scenario has a straightforward geometric interpretation: We can think of S_1, S_2 as vectors in an p-dimensional space, and X will lie on the line between the two sources.

The LSQ-estimate of α is then

$$\epsilon = \sum_i^p (\alpha s_{1i} + (1-\alpha)s_{2i} - x_i)^2 \tag{2}$$

$$= \sum_i [x_i^2 + \alpha^2 s_{1i}^2 + (1 - 2\alpha + \alpha^2)s_{2i}^2 + 2\alpha(1-\alpha)s_{1i}s_{2i} - 2\alpha s_{1i}x_i - (1-\alpha)s_{2i}x_i] \tag{3}$$

$$= \alpha^2 \left(\sum_i [s_{1i}^2 + s_{2i}^2 - 2s_{1i}s_{2i}] \right) + \alpha \left(\sum_i [-2s_{2i}^2 - 2s_{1i}x_i + 2s_{1i}s_{2i} + 2s_{2i}x_i] \right) + K \tag{4}$$

$$\frac{d\epsilon}{d\alpha} = 2\alpha \left(\sum_i [s_{1i}^2 + s_{2i}^2 - 2s_{1i}s_{2i}] \right) - 2\sum_i [s_{2i}^2 + s_{1i}x_i - s_{1i}s_{2i} - s_{2i}x_i] = 0 \tag{5}$$

$$\alpha = \frac{\sum_i [s_{2i}^2 + s_{1i}s_i - s_{1i}s_{2i} - s_{2i}x_i]}{\sum_i [s_{1i}^2 + s_{2i}^2 - 2s_{1i}s_{2i}]} \tag{6}$$

$$= \frac{\sum_i (s_{1i} - s_{2i})(x_i - s_{2i})}{\sum_i (s_{1i} - s_{2i})^2} \tag{7}$$

$$= \frac{\langle S_1 - S_2; X - S_2 \rangle}{\|S_1 - S_2\|^2} \tag{8}$$

which is the length of the projection of $X - S_2$ onto $S_1 - S_2$.

Let us now assume X is not sampled immediately, but rather we observe a later population X_2 Assuming no further gene flow happens between X and X_2, then the drift should be orthogonal, i.e.

$$\langle X_2 - X; S_2 - S_1 \rangle = 0$$

and so this projection will still work.

2.2.2 Many pops

Likewise, for multiple sources,

$$X = \boldsymbol{\alpha}\mathbf{S}, \tag{9}$$

where now $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)$ is the vector containing contributions from each population, subject to the constraints that $\alpha_i \geq 0$ and $\sum_i^n \alpha_i = 1$.

$$\epsilon = \sum_i^p \left(\left(1 - \sum_{j>1} \alpha_j \right) s_{0i} + \sum_{j>1} \alpha_j s_{ji} - x_i \right)^2 \tag{10}$$

$$= \sum_i^p \left(\sum_j \alpha_j (s_{ji} - s_{0i}) - (x_i - s_{0i}) \right)^2 \tag{11}$$

$$= \sum_i^p \left(\sum_j \alpha_j s'_{ji} - y_i \right)^2 \tag{12}$$

$$= \sum_i \left(y_i^2 + \sum_j \alpha_j^2 s_{ji}^2 + 2 \sum_{j \neq k} \alpha_j \alpha_k s_{ji} s_{ki} - 2 \sum_j y_i \alpha_j s_{ji} \right) \tag{13}$$

$$= \sum_i y_i^2 + \sum_j \alpha_j^2 \sum_i s_{ji}^2 + 2 \sum_{j \neq k} \alpha_j \alpha_k \sum_i s_{ji} s_{ki} - 2 \sum_j \alpha_j \sum_i y_i s_{ji} \tag{14}$$

$$= \|\mathbf{Y}\|^2 + \sum_j \alpha_j^2 \|S_j\|^2 + 2 \sum_{j \neq k} \alpha_j \alpha_k S_j \cdot S_k - 2 \sum_j \alpha_j \mathbf{Y} \cdot S_j \tag{15}$$

$$\frac{d\epsilon}{d\alpha_j} = 2\alpha_j \|S_j\|^2 + 2 \sum_{k \neq j} \alpha_k S_j \cdot S_k - 2\mathbf{Y} \cdot S_j = 0 \tag{16}$$

$$\hat{\alpha_j} = \frac{(\mathbf{Y} - \sum_{k \neq j} \alpha_k S_k) \cdot S_j}{\|S_j\|^2} \tag{17}$$

$$\hat{\alpha_j} = \frac{\mathbf{Y} \cdot S_j}{\|S_j\|^2} \tag{18}$$

where the last line only holds when the S_j are orthogonal, in which case

$$\hat{\boldsymbol{\alpha}} = \mathbf{YS}$$

2.2.3 Motivating a set of reference pops

Let us now assume that the S_i are not directly known, but instead admixture happened earlier, such that

$$X = \sum \alpha_j k_j S_j \tag{19}$$

with the constraint that $\sum_j \alpha_j = 1$ and $0 \leq k_j \leq 1$ for all j. It seems clear from above considerations that the α_j and k_j cannot be identified independently. Thus, we can motivate the introduction of a set of reference populations $\mathbf{R}$ some reference populations that are in some way related to the populations, but are not directly admixture sources. In this scenario, let S'_i be the projection of S_i onto the reference matrix $\mathbf{P}$, (tbd, remember how $\mathbf{P}$ is derived)

$$S'_i = S_i \mathbf{P} \tag{20}$$

By definition, $\langle S_i - S'_i, R_i \rangle = 0$ for all basis vectors in R. Furthermore, by the assumption,

$$\langle S_i - S'_i, S_j - S'_j \rangle = 0$$

for all suitable sources. Finally, the target population X can be written as $X' + \Delta X$, where ΔX is, by definiton, orthogonal to $\mathbf{S}$ and $\mathbf{R}$, and thus

$$\langle X' - X, R_i - R'_i \rangle = 0 \tag{21}$$

(not sure if useful or trivial)

2.2.4 Defining a reference

This section is pure conjecture. Given a set of putative source populations **S** and a larger set of populations **R**, can we create a suitable reference space (similar to the rotation idea in qpgraph). I.e. we can think of the $S_i - S_i'$ spanning an orthogonal space.

2.2.5 Closely correlated putative sources

Another issue might be that we have closely related sources. For example, when modelling African Americans we might conjecture that many West African and European pops would give very similar results. How can we find a "better" source set (e.g a linear combination of Africans) instead of a single pop source?

References

Nielsen, R and J Wakeley (2001): Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics* 158 (2), 885–896.