Inference of Continuous and Discrete Population Structure

Gideon S. Bradburd[1,a], Peter L. Ralph[2,b], Graham M. Coop[3,c]

[1]Ecology, Evolutionary Biology, and Behavior Group, Department of Integrative Biology, Michigan State University, East Lansing, MI 48823

[2]Institute of Ecology and Evolution, Department of Mathematics, University of Oregon, Eugene, OR 97403

[3]Center for Population Biology, Department of Evolution and Ecology, University of California, Davis, CA 95616

[a]bradburd@msu.edu; [b]plr@uoregon.edu; [c]gmcoop@ucdavis.edu

1 Conceptual framework

We wish to model covariance in allele frequencies between individuals or demes as a mixture of spatial processes, with each process representing a discrete cluster or population, within which relatedness decays with distance between samples.

2 Drift, admixture, and allele frequencies

We imagine the allele frequencies at each locus in each sample to be the sum of 3 components: the ancestral allele frequency ϵ shared by all samples, the deviation from that ancestral mean in the kth population, $\Delta^{(k)}$, which is shared by all samples with 100% ancestry in that population, the deviation specific to the ith sample, $\Delta^{(i)}$, which captures drift not shared by all samples at the population level (i.e., inbreeding or subpopulation-specific drift).

If all samples drew all of their ancestry from any one of the K clusters, the allele frequency in the ith sample at the ℓth locus is therefore given by:

$$F_{i,\ell} = \epsilon_\ell + \Delta_\ell^{(k)} + \Delta_\ell^{(i)} \tag{1}$$

3 Normal approximation to drift - MVN and Wishart

We can model the allele frequencies at each locus across populations as multivariate normal. This will be a good approximation if the amount of elapsed drift in each sample is relatively small, as would be expected if time since divergence from a common ancestral population is short, or effective population sizes are large.

Beginning with the simplest case of two clusters ($K = 2$), with all samples drawing 100% of their ancestry from either one or the other cluster, we can model allele frequencies at locust ℓ as:

$$F_\ell \sim MVN\left(\mu = \epsilon_\ell + \iota^{(k)}\Delta_\ell^{(k)}, \Sigma = \Omega\right) \tag{2}$$

where $\iota^{(k)}$ is an indicator variable that equals 1 for samples with 100% membership in cluster k and 0 otherwise. In Eqn. (2), all entries of the covariance Ω have value

0 except the diagonals, which describe sample-specific drift, and have the following value:

$$\Omega_{i,i} = \text{Var}\left(\Delta^{(i)}\right) \tag{3}$$

We can describe the covariance induced by both the shared ancestral allele frequency and shared population-level drift with the mean of the multivariate normal, or, alternatively, we can describe it directly within the MVN covariance as follows:

$$F_\ell \sim MVN\left(\mu = 0, \Sigma = \Omega\right) \tag{4}$$

Here, the covariance Ω is parameterized as:

$$\Omega_{i,j} = \begin{cases} \text{Var}(\epsilon) + \text{Var}\left(\Delta^{(i)}\right) + \text{Var}\left(\Delta^{(i)}\right), & \text{if } i = j \text{ and } \iota_i^{(k)} = 1 \\ \text{Var}(\epsilon) + \text{Var}\left(\Delta^{(k)}\right), & \text{if } \iota_i^{(k)} = \iota_j^{(k)} = 1 \text{ and } i \neq j \\ \text{Var}(\epsilon) + \text{Var}\left(\Delta^{(i)}\right), & \text{if } i = j \text{ and } \iota_i^{(k)} \neq 1 \\ \text{Var}(\epsilon), & \text{otherwise} \end{cases} \tag{5}$$

This then implies that the covariance of the sampled allele frequencies should be Wishart-distributed.

$$FF^T \sim \mathcal{W}\left(\Omega, L\right) \tag{6}$$

where Ω is parameterized as given in Eqn. (5) and the degrees of freedom of the Wishart is L, the number of loci, if all loci are independent, and $L\phi$ otherwise, where ϕ is some fudge factor that is less than 1, and describes non-independence between loci due to LD.

4 Admixture model

The model above describes the simple case when samples draw 100% of their ancestry from only a single cluster each.

To accommodate admixture between clusters, we can model allele frequencies within samples as linear combinations of the population-specific drift terms across the clusters from which they draw ancestry. Here, we introduce a term $w_i^{(k)}$ that describes the admixture proportion of sample i in cluster k, which can be interpreted as the probability an allele sampled in the ith sample came from the kth

cluster. The allele frequency in the ith sample at the ℓth locus is therefore given by:

$$F_{i,\ell} = \epsilon_\ell + \sum_K \left(w_i^{(k)} \Delta_\ell^{(k)} \right) + \Delta_\ell^{(i)} \tag{7}$$

where

$$\sum_K w_i^{(k)} = 1 \tag{8}$$

The form of the parametric covariance of samples that are admixed between K clusters is given by:

$$\Omega_{i,j} = \mathrm{Var}(\epsilon) + \sum_K \left(w_i^{(k)} w_j^{(k)} \mathrm{Var}(\Delta^{(k)}) \right) + \delta_{i=j} \mathrm{Var}(\Delta^{(i)}) \tag{9}$$

5 Modeling spatial differentiation

Equation (9) describes a model in which samples can be continuously admixed between a set of K discrete clusters. In this model, any pair of samples with 100% ancestry in a cluster have exactly the same covariance with each other (namely $\mathrm{Var}(\Delta^{(k)})$). However, as described in Section 1, we also wish to model continuous decay of covariance with spatial separation between samples. That is, we expect (and want our model to reflect) that samples within the same cluster to have higher covariance if they are sampled closer together than if they are sampled farther apart. To describe this spatial pattern, we build in a spatial component to our covariance model.

Specifically, we write that the covariance between a pair of samples i and j that draw all their ancestry from a single cluster k decays exponentially as a function of the distance between their sampling locations as follows:

$$H_{i,j}^{(k)} = \alpha_0^{(k)} \times \exp\left(\left(-(\alpha_D^{(k)} D_{i,j})^{\alpha_2^{(k)}} \right) + \mu^{(k)} \right) \tag{10}$$

where the α parameters control: the sill of the covariance (α_0), the rate at which covariance decays with distance $D_{i,j}$ (α_D), and the shape of that decay (α_2) in the kth cluster. The parameter $\mu^{(k)}$ describes the covariance shared by all samples in the kth cluster, and is included to describe the quantity $\mathrm{Var}(\Delta^{(k)})$ from Eqn. (9).

We can then build a parametric covariance that accommodates

1. continuous decay of covariance within a cluster
2. a discrete amount of covariance shared within a cluster
3. continuous admixture between clusters.

$$\Omega_{i,j} = \sum_K \left(w_i^{(k)} w_j^{(k)} H_\theta^{(k)} \right) + \delta_{i=j} \mathrm{Var}(\Delta_i) \tag{11}$$

where $H_\theta^{(k)}$ denotes the dependence of the spatial covariance within a cluster on other quantities θ: specifically, the parameters $\vec{\alpha}$ and μ, as well as on the observed quantity D, the matrix of pairwise distances between all samples.

A quick note that following Section 4, I've switched notation so that, at each locus, there is a global mean in addition to K population-specific deviates from that mean, rather than a global mean that incorporates one population-specific deviate and another $K-1$ deviates.

6 Standardization

It would be tempting to model the sample covariance of the raw allele frequency data (FF^T) as Wishart-distributed with a parametric scale matrix Ω taken from Eqn. (11). However, the raw allele frequency data are probably not well fit by a multivariate normal, so their covariance is probably not well described by the Wishart.

We therefore standardize the data by mean-centering the observations at each locus and normalizing their variance. The standardization we use proceeds as follows:

$$\begin{aligned} \bar{F}_\ell &= \frac{1}{N} \sum_N F_{\ell,n} \\ \bar{F}_\ell^* &= \frac{1}{N+1} \left(0.5 + \sum_N F_{\ell,n} \right) \\ X_{\ell,n} &= \frac{\hat{F}_{\ell,n} - \bar{F}_\ell}{\sqrt{\bar{F}_\ell^* (1 - \bar{F}_\ell^*)}} \end{aligned} \tag{12}$$

We then define the sample covariance of the standardized frequencies X, and proceed by modeling that quantity:

$$\widehat{\Omega} = XX^T \tag{13}$$

7 Likelihood

The sample covariance of the standardized allele frequencies ($\widehat{\Omega}$) has rank $N-1$, so we compute the likelihood of an $(N-1)$-dimensional projection of the data. Following SpaceMix, we create a mean-centering matrix T:

$$T_{i,j} = \delta_{i=j} - \frac{1}{N} \tag{14}$$

where $\delta_{i=j}$ is an indicator variable that equals 1 when $i = j$ and 0 otherwise. We choose a projection matrix Ψ by dropping the last column of the orthogonal matrix in the QR decomposition of T.

We then calculate the likelihood of the data as

$$P(\Psi^T\widehat{\Omega}\Psi \mid \Omega) = \mathcal{W}(\Psi^T\Omega\Psi\ ,\ L) \tag{15}$$

where L is the number of loci across which $\widehat{\Omega}$ is calculated, and Ω is the parametric covariance described above in Eqn. (11).