Inference of Continuous and Discrete Population Structure Through Space and Time

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

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

[2]Department of Molecular and Computational Biology, University of Southern California, Los Angeles, CA 90089

[a]gbradburd@ucdavis.edu; [b]pralph@usc.edu; [c]gmcoop@ucdavis.edu

1 Data

At each variable, bi-allelic locus, we have picked an allele to count in each sample, and we also know the number of chromosomes genotyped in each sample. We can then define the sample allele frequency at locus ℓ in sample n as the quotient of the allele counts (C) and the sample size (S):

$$\hat{F}_{\ell,n} = \frac{C_{\ell,n}}{S_{\ell,n}} \tag{1}$$

Our inference framework proceeds by modeling the covariance in allele frequencies as a Wishart-distributed random variable, but the sample allele frequencies are not multivariate normally distributed. so we standardize the allele frequencies. The standardization we use proceeds as follows:

$$\begin{aligned} \bar{F}_\ell &= \frac{1}{N}\sum_N F_{\ell,n} \\ \bar{F}^*_\ell &= \frac{1}{N+1}\left(0.5 + \sum_N F_{\ell,n}\right) \\ X_{\ell,n} &= \frac{\hat{F}_{\ell,n} - \bar{F}_\ell}{\sqrt{\bar{F}^*_\ell(1-\bar{F}^*_\ell)}} \end{aligned} \tag{2}$$

We then can define the covariance of the standardized allele frequencies across all L variable, and we model that covariance.

$$\widehat{\Omega} = \frac{1}{L}X^T X \tag{3}$$

2 Model

We define a parametric covariance, Ω, which will serve as the mean of the Wishart distribution from which we are treating the sample covariance of the standardized sample allele frequencies as a draw. This parametric covariance is a function both discrete population structure (i.e., samples can draw ancestry from one or more discrete clusters) as well as a continuous decay of covariance with distance within each cluster (i.e., a pair of samples drawing all of their ancestry from a single cluster are expected to have lower covariance with each other the farther apart they occur).

We begin by defining the spatial covariance function in the k^{th} cluster, $F^{(k)}$, as a function of the geographic distance $D_{i,j}$ between a pair of samples, i and j:

$$F_{i,j}^{(k)} = \alpha_0^{(k)} \times \exp\left(-\left(\alpha_D^{(k)} * D_{i,j}\right)^{\alpha_2^{(k)}}\right) + \mu^{(k)} \tag{4}$$

We can then define the admixed covariance function between the same pair of samples, which incorporates both the discrete, between-cluster structure, as well as the continuous differentiation within each cluster.

$$\Omega_{i,j} = \sum_K \left(w_i^{(k)} w_j^{(k)} F_{i,j}^{(k)}\right) + \delta_{i=j}\left[\left(1 - \sum_K \left(w_i^{(k)} w_j^{(k)} F_{i,i}^{(k)}\right)\right) \times \left(\eta_i + \frac{1}{S_i}\right)\right] \tag{5}$$

Above, $\delta_{i=j}$ is an indicator variable that has value 1 when $i = j$ and 0 otherwise. The variable η_i is the nugget in the i^{th} sample, and, as in Eqn. 1, S_i gives the number of genotyped chromosomes in the i^{th} sample.

There are 4 models in this conceptual framework (3 others in addition to the one listed above):

1. spatial model with multiple clusters

$$\Omega_{i,j} = \sum_K \left(w_i^{(k)} w_j^{(k)} F_{i,j}^{(k)}\right) + \delta_{i=j}\left[\left(1 - \sum_K \left(w_i^{(k)} w_j^{(k)} F_{i,i}^{(k)}\right)\right) \times \left(\eta_i + \frac{1}{S_i}\right)\right]$$

2. nonspatial model with multiple clusters

$$\Omega_{i,j} = \sum_K \left(w_i^{(k)} w_j^{(k)} \mu^{(k)}\right) + \delta_{i=j}\left[\left(1 - \sum_K \left(w_i^{(k)} w_j^{(k)} \mu^{(k)}\right)\right) \times \left(\eta_i + \frac{1}{S_i}\right)\right]$$

3. spatial model with a single cluster

$$\Omega_{i,j} = F_{i,j} + \delta_{i=j}\left[(1 - F_{i,i}) \times \left(\eta_i + \frac{1}{S_i}\right)\right]$$

4. nonspatial model with a single cluster

$$\Omega_{i,j} = \mu + \delta_{i=j}\left[(1 - \mu) \times \left(\eta_i + \frac{1}{S_i}\right)\right]$$

3 Likelihood

The sample covariance of the standardized allele frequencies 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, where

$$T = \delta_{i,j} - \frac{1}{N} \tag{6}$$

and 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

$$P(\widehat{\Omega} \mid \Omega) = \mathcal{W}(L\Psi^T \Omega \Psi \ , \ L) \tag{7}$$