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 Model

Ok so we have N individuals sampled throughout time and space, and, for each individual, we have the genotype at each of L loci. We are interested in modeling both continuous patterns of population differentiation (isolation across space and over time), as well as discrete population structure. In space, discrete structure might be due to a barrier to dispersal, or to a recent expansion event that has brought relatively differentiated groups into geographic contact. In time, discrete structure might correspond to population replacement, or the sudden influx of gene flow from somewhere not nearby.

To model these dual patterns of structure (continuous in time and space, as well as discrete), we say that there are K populations in space-time. Within each population (which are roughly analogous to the clusters in STRUCTURE), covariance decays continuously with spatial and temporal distance. The spatiotemporal covariance between individuals i and k takes the following form in population k:

$$F_{i,j}^{(k)} = \alpha_0^{(k)} \exp\left(\left(\frac{D_{i,j}}{\alpha_D^{(k)}} + \frac{\tau_{i,j}}{\alpha_\tau^{(k)}}\right)^{\alpha_2^{(k)}}\right) \tag{1}$$

(Peter, this form of covariance is a placeholder. I've been looking into different forms from Gneiting 2002, but would love some feedback on considerations when choosing which properties of which Matern functions are desirable).

Individuals are then admixed between those populations, choosing membership $w^{(k)}$ in population k, where $w^{(k)}$ is the probability that a sampled allele came from population k.

We model the covariance in allele frequencies, Ω, across loci between individuals i and j as a sum of their spatiotemporal covariances within each cluster, weighted by their membership proportions in each cluster.

$$\Omega_{i,j} = \sum_K \left(w_i^{(k)} w_j^{(k)} \left(F_{i,j}^{(k)} + \mu^{(k)}\right)\right) + \gamma + \delta_{i=j}\eta_i \tag{2}$$

Here, $\mu^{(k)}$ is the population-level effect of sharing some deviation away from an ancestral allele frequency, γ is the global effect of sharing an ancestral allele fre-

quency, δ is an indicator variable that takes the value 1 when index i is equal to index j, and η is the sample-specific nugget on the diagonal of the covariance matrix.

(Graham, I have two questions for you:

- First, do you remember the conversation we had in your office about short-term and long-term tweaks to spatialStructure's current model? I remember that we discussed two short-term tweaks: (1) one that involved incorporating a sample size effect onto the diagonal of the covariance matrix, and (2) **something else**. To my shame, I can neither remember what the other thing was nor (since I moved out of the office) can I find the notes that I took on it. So, question 1 is: do you remember what the other thing was?

- Question 2 has to do with the sample size effect. I've been playing around with it a little bit and hitting a conceptual wall. Here are my thoughts: From the binomial, the variance of f is $\frac{f(1-f)}{N}$, where N is the sample size. So, if the allele frequencies across all loci in a sample were identical, their variance would be just given by $\frac{f(1-f)}{N}$.

  Now, if we allow f to vary across loci, and if we say that f_i denotes the vector of allele frequencies in sample i across all L loci, then I think, when we had talked about it, that we arrived at the following:

$$Var(f_i) = Var(\epsilon) + \frac{1}{L}\sum_{L}(f_\ell(1-f_\ell)/N_i) \tag{3}$$

  where ϵ_ℓ gives the ancestral allele frequency at locus ℓ, and $Var(\epsilon)$ is hopefully captured by our γ parameter. This equation appears to hold true in tests, both from simulated and empirical datasets.

  So, the question is, how do we incorporate information about sample sizes into the model? Considerations are:

  - We don't want the sample size effect to be greater than the actual variance in any sample, as the model will not be able fit the data with any parameter (nuggets are all positive). Also note that we're adding $(+\gamma + \sum_{K} w_i^{(k)}\mu^{(k)} + \eta_i)$, so really that presents the hard limit we don't want to be greater than $Var(f_i)$
  - We may want to accommodate different sample sizes across loci, as those could be highly variable. Although, maybe that's just too unwieldy.

  Options seem to be:

  - We make the prior of the nugget centered around $1/N$. The upside here is that the model retains the flexibility to have a diagonal element

on the parametric covariance matrix that's equal to (or, at least, not necessarily greater than) the true sample variance. The downside is that maybe we don't want it to have that flexibility. This option would look like:

$$\Omega_{i,j} = \gamma + \mathbb{E}\left[\bar{f}(1-\bar{f})\right]\left(\sum_{K}\left(w_i^{(k)}w_j^{(k)}\left(F_{i,j}^{(k)}+\mu^{(k)}\right)\right)+\delta_{i=j}\left(\eta_i(\frac{1}{N_i})\right)\right) \tag{4}$$

– We just add a hard $1/N$ to the diagonal element. This option would look like:

$$\Omega_{i,j} = \gamma + \mathbb{E}\left[\bar{f}(1-\bar{f})\right]\left(\sum_{K}\left(w_i^{(k)}w_j^{(k)}\left(F_{i,j}^{(k)}+\mu^{(k)}\right)\right)+\delta_{i=j}\left(\eta_i+\frac{1}{N_i}\right)\right) \tag{5}$$

Do you have any thoughts or opinions about these options? And also, am I missing something here? I don't think the $\mathbb{E}\left[\bar{f}(1-\bar{f})\right]$ should be square-rooted, but maybe I'm wrong?)

shared mean business

Our thoughts proceed as follows:
say X_i is the allele frequency at a given locus in population i,
and ϵ is the ancestral frequency at that locus,
and ΔX_i is the deviation in population i from ϵ at that locus:

$$\begin{aligned}
\text{Cov}(f_i, f_j) &= \mathbb{E}\left[X_i X_j\right] - \mathbb{E}\left[X_i\right]\mathbb{E}\left[X_j\right] \\
&= \mathbb{E}\left[(\Delta X_i + \epsilon)(\Delta X_j + \epsilon)\right] - \mathbb{E}\left[\Delta X_i + \epsilon\right]\mathbb{E}\left[\Delta X_j + \epsilon\right] \\
&= \mathbb{E}\left[(\Delta X_i \Delta X_j + \Delta X_i \epsilon + \Delta X_j \epsilon + \epsilon^2)\right] - (\mathbb{E}\left[\Delta X_i\right] + \mathbb{E}\left[\epsilon\right])(\mathbb{E}\left[\Delta X_j\right] + \mathbb{E}\left[\epsilon\right]) \\
&= \mathbb{E}\left[\Delta X_i \Delta X_j\right] + \mathbb{E}\left[\epsilon^2\right] - \mathbb{E}\left[\epsilon\right]^2 \\
&= \mathbb{E}\left[\Delta X_i \Delta X_j\right] + \mathbb{V}\text{ar}\left[\epsilon\right] \qquad (6)
\end{aligned}$$

So in Eqn. (2), the δ term is describing this $\mathbb{V}\text{ar}\left[\epsilon\right]$. If this math works, then it also means that we don't have to mean-center the loci; we can just model covariance in sample frequencies, and add a single parameter to describe the increased covariance between samples across loci due to the ancestral mean they share at each locus.

2 WAIC

$$WAIC = -2\widehat{elpd_{waic}} \qquad (7)$$

$$\widehat{elpd_{waic}} = \widehat{lpd} - \hat{p}_{waic} \qquad (8)$$

$$\widehat{lpd} = \sum_N log\left(\frac{1}{S}\sum_S p(y_i|\theta^s)\right) \qquad (9)$$

$$\hat{p}_{waic} = \sum_N Var_S\left(log\left(p(y_i|\theta^s)\right)\right) \qquad (10)$$

$$WAIC = -2\left(\sum_N\left[log\left(\frac{1}{S}\sum_S p(y_i|\theta^s)\right)\right] - \sum_N\left[Var_S\left(log\left(p(y_i|\theta^s)\right)\right)\right]\right) \qquad (11)$$

And now, plugging in our likelihood function,

$$p(y_i|\theta^s) = \frac{|X|^{\frac{n-p-1}{2}} e^{\frac{-tr(V_S^{-1}X)}{2}}}{2^{\frac{np}{2}}|V_S|^{\frac{n}{2}}\Gamma_p\left(\frac{n}{2}\right)} \tag{12}$$

$$WAIC = -2\left(\sum_N\left[log\left(\frac{1}{S}\sum_S \frac{|X|^{\frac{n-p-1}{2}} e^{\frac{-tr(V_S^{-1}X)}{2}}}{2^{\frac{np}{2}}|V_S|^{\frac{n}{2}}\Gamma_p\left(\frac{n}{2}\right)}\right)\right] - \sum_N\left[Var_S\left(log\left(\frac{|X|^{\frac{n-p-1}{2}} e^{\frac{-tr(V_S^{-1}X)}{2}}}{2^{\frac{np}{2}}|V_S|^{\frac{n}{2}}\Gamma_p\left(\frac{n}{2}\right)}\right)\right)\right]\right) \tag{13}$$

$$WAIC = -2\left(\sum_N\left[log\left(\frac{|X|^{\frac{n-p-1}{2}}}{S2^{\frac{np}{2}}\Gamma_p\left(\frac{n}{2}\right)}\sum_S \frac{e^{\frac{-tr(V_S^{-1}X)}{2}}}{|V_S|^{\frac{n}{2}}}\right)\right] - \sum_N\left[Var_S\left(log\left(|X|^{\frac{n-p-1}{2}}\right) + log\left(e^{\frac{-tr(V_S^{-1}X)}{2}}\right) - log\left(2^{\frac{np}{2}}\right) - log\left(|V_S|^{\frac{n}{2}}\right) - log\left(\Gamma_p\left(\frac{n}{2}\right)\right)\right)\right]\right) \tag{14}$$

$$WAIC = -2\left(\sum_N\left[log\left(\frac{|X|^{\frac{n-p-1}{2}}}{S2^{\frac{np}{2}}\Gamma_p\left(\frac{n}{2}\right)}\right) + log\left(\sum_S \frac{e^{\frac{-tr(V_S^{-1}X)}{2}}}{|V_S|^{\frac{n}{2}}}\right)\right] - \sum_N\left[Var_S\left(log\left(e^{\frac{-tr(V_S^{-1}X)}{2}}\right) - log\left(|V_S|^{\frac{n}{2}}\right)\right)\right]\right) \tag{15}$$

$$WAIC = -2\left(N \times log\left(\frac{|X|^{\frac{n-p-1}{2}}}{S2^{\frac{np}{2}}\Gamma_p\left(\frac{n}{2}\right)}\right) + \sum_N\left[log\left(\sum_S \frac{e^{\frac{-tr(V_S^{-1}X)}{2}}}{|V_S|^{\frac{n}{2}}}\right)\right] - \sum_N\left[Var_S\left(log\left(e^{\frac{-tr(V_S^{-1}X)}{2}}\right) - log\left(|V_S|^{\frac{n}{2}}\right)\right)\right]\right) \tag{16}$$

(17)