Estimating edge weights connecting HJA columbine genetic networks

D. G. Gannon

February 2021

Data

We have SNP data on $n = 192$ *Aquilegia formosa* individuals from 25 meadows, which we consider to be "sub-populations" of the H.J. Andrews *A. formosa* population. We work with the allelic scatter matrix $\mathbf{S}$ computed from an allelic frequency matrix $F_{(n\times\ell)}$, where ℓ is the number of SNP loci sequenced and n is the number of plants sampled. Allelic frequencies are coded as $f_{ik} = 0$ if individual i is homozygous for the randomly selected reference allele at locus k, $f_{ik} = 0.5$ if individual i has a heterozygous genotype at site k, and $f_{ik} = 1$ if individual i is homozygous with two copies of the alternative allele at site k. The data were previously filtered to include only bi-allelic loci.

I computed $\mathbf{S}$ as

$$\mathbf{S} = \left(\mathbf{F} - \frac{1}{2}\mathbf{J}_{(n\times\ell)}\right)\left(\mathbf{F} - \frac{1}{2}\mathbf{J}_{(n\times\ell)}\right)',$$

where $\mathbf{J}_{(n\times\ell)}$ is an all-ones matrix and $'$ denotes the matrix transpose. This definition is the matrix representation of the allelic covariance matrix defined by Bradburd, Coop, and Ralph (2018) (**This we might need to adjust to set the diagonal to 0.25**).

Model

Following Bradburd, Coop, and Ralph (2018) and Peterson et al. (2019), We assume that $\mathbf{S} \sim \text{Wishart}(\ell,\ \boldsymbol{\Sigma})$, where the number of SNP loci, ℓ, is the degrees of freedom parameter and $\boldsymbol{\Sigma}$ is the scale matrix. To model spatial dependence among individual genotypes (i.e., isolation by distance (Wright 1943) or isolation by resistance (McRae 2006)), we let

$$\boldsymbol{\Sigma} = (\mathbf{M} - \rho\mathbf{W})^{-1}.$$

Above, $\mathbf{W}_{(n\times n)}$, $w_{ij} = 0$ for $i = j$, is the weights matrix. It determines the degree of connectivity among nodes (defined as plants here) in a spatial network and is the parameter of interest. The parameter $\mathbf{M}$ is a diagonal matrix with $m_{ii} = \sum_{j=1}^{n} w_{ij}$ and all off-diagonal elements equal to zero, and ρ controls the amount of spatial autocorrelation among the genotypes of nearby individuals. We model the weight between plants i and j as a log-linear combination of covariates and regression parameters such that

$$w_{ij} = \exp\{\mathbf{x}'_{ij}\boldsymbol{\beta}\},$$

where $\mathbf{x}_{ij}$ is a vector of covariates for individuals i and j and $\boldsymbol{\beta}$ is a vector of regression parameters. Our covariates of interest are with respect to *edges* in the network. Therefore, we average the covariates that are measurements on a node. For example, if d_i, d_j are the flowering plant densities around plants i and j (respectively), then the explanatory variable for the edgeweight $w_{ij} = w_{ji} = (d_i + d_j)/2$. This helps ensure a symmetric weights matrix where explanatory variables measured at node i are given equal weights as the same measures at node j. The list of explanatory variables is:

- **Geographic distance between plants i and j (km)**.
- **Average flowering plant density within 5m of plants i and j.** We expect that high flowering plant densities should attract more pollinators. Therefore, a given plant in a high density patch may export more pollen and sire more offspring in the surrounding area than a plant in a low density patch. This should result in higher allelic covariance between this plant and a randomly selected plant on the landscape.
- **Average proportion of forested cells within a 500m radius around plants i and j.** We use this as a measure of average "functional connectivity", hypothesizing that the more forest around a plant, the fewer pollinators will find it and therefore the less connected it is to other plants. We selected a cutoff of 500m based on previous work with hummingbirds which indicated a 50% reduction in the probability of movement between two locations with an increase of 500m.
- **Average canopy cover over plants i and j.** The logic for including this explanatory variable is that well-connected plants (those not surrounded by lots of forest) may still be growing under a tree. This may reduce the chances of being visited by a pollinator if pollinators flying overhead are less likely to see the plant.

- **Squared proportion of forested cells within a 500m radius**. Based on some data visualizations, it looks like there could be a peak of allelic covariance at intermediate levels surrounding forest. This does not seem unreasonable if plants and/or pollinators prefer some shade in the heat of the day or if hummingbirds prefer partially forested areas for the sake of nesting and perching.
- **Interaction between distance and canopy cover immediately over plants i and j**. Because we think plants may be more difficult for a pollinator to find if growing beneath a tree, we might expect higher rates of mating between plants under the same tree (or trees), than two plants at the same distance but in the open. This should result in greater allelic covariance for two nearby plants under high canopy cover. However, two plants under different trees that are far apart may be very unlikely to mate. We therefore expect the effect of canopy cover to vary depending on geographic distance between plants i and j.

Priors

We assume the following weakly informative priors for the regression parameters and ρ:

$$\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}_P,\ \mathbf{I}_{(P\times P)}),$$

and

$$\rho \sim \mathcal{B}(2,2),$$

where $\mathcal{B}(\alpha,\beta)$ is a Beta distribution with shape parameters α and β.

Stan Model Code

```
functions{
// function to make weights matrix from regression array X and
// paramater vector v
  matrix make_W(int N, vector v, real[,,] X){
    matrix[N,N] Wmat;
    for(i in 1:N){
      if(i == 1){Wmat[i,i] = 0;}
      else{
        for(j in 1:(i-1)){
          Wmat[i,i] = 0;
          Wmat[i,j] = exp(to_row_vector(X[i,j,])*v);
          Wmat[j,i] = Wmat[i,j];
        }
      }
    }
    return(Wmat);
  }

}

data{

  int<lower=1> N;          //number of individuals sampled
  int<lower=1> L;          //number of SNP loci
  int<lower=1> P;          //number of landscape and node variables

  real X[N,N,P];           //design array
  matrix[N,N] S;           //scatter matrix

}

parameters{

  vector[P] beta;              //regression parameters
  real<lower=0,upper=1> rho;   //spatial dependence

}

transformed parameters{
  matrix[N,N] W;

  vector[N] W_sum;

  matrix[N,N] M;

  cov_matrix[N] Sigma;

  //matrix[N,N] Psi;


  W = make_W(N, beta, X);


  for(i in 1:N){

    W_sum[i] = sum(W[i,]);

  }


  M = diag_matrix(W_sum);


  Sigma = inverse((M - (rho*W)));


}


model{


//priors

  beta ~ normal(0,10);

  rho ~ beta(2,2);




//likelihood

  S ~ wishart(L, Sigma);


}


generated quantities{


  real loglik;
  loglik = wishart_lpdf(S | L, Sigma);


}
```

Bradburd, Gideon S., Graham M. Coop, and Peter L. Ralph. 2018. "Inferring Continuous and Discrete Population Genetic Structure Across Space." *Genetics* 210 (1): 33–52. https://doi.org/10.1534/genetics.118.301333.

McRae, Brad H. 2006. "Isolation by Resistance." *Evolution* 60 (8): 1551–61. https://doi.org/10.1111/j.0014-3820.2006.tb00500.x.

Peterson, Erin E., Ephraim M. Hanks, Mevin B. Hooten, Jay M. Ver Hoef, and Marie-Josée Fortin. 2019. "Spatially Structured Statistical Network Models for Landscape Genetics." *Ecological Monographs* 89 (2): e01355. https://doi.org/10.1002/ecm.1355.

Wright, Sewall. 1943. "Isolation by Distance." *Genetics* 28 (2): 114–38.