# Estimating edge weights connecting HJA columbine genetic networks

D. G. Gannon

February 2021

## Data

We have biallelic SNP genotype data from $n = 192$ *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 genetic distance matrix defined by $\mathbf{D} = (\mathbf{y}_i - \mathbf{y}_j)^T(\mathbf{y}_i - \mathbf{y}_j)$ where $\mathbf{y}_i,\ i = 1, 2, ..., n$ is the $K$-vector of genotypes of individual $i$ at loci $k = 1, 2, ..., K$ coded as $y_{ik} \in \{0, 1, 2\}$. $y_{ik} = 0$ if individual $i$ is homozygous for the major allele at locus $k$, $y_{ik} = 1$ if individual $i$ is heterozygous at locus $k$ and $y_{ik} = 2$ if individual $i$ is homozygous for the minor allele at locus $k$.

### Model

McCullagh (2009) showed that distance (or dissimilarity) matrices can be modeled using a *generalized Wishart* distribution, or, equivalently, that linear contrasts on distance matrices can be modeled using a Wishart distribution. Hanks and Hooten (2013) illustrated the relationship between this result and the *Isolation by Resistance* model of genetic differentiation (McRae 2006), which is a theoretical construct relating migration and gene flow to random walks along edges and among nodes of a graph, where the nodes represent populations. McRae (2006) showed that if we imagine the graph as an electrical circuit connected by a set of resistors, then the electrical resistance among nodes is proportional to $F_{ST}$, a measure of genetic differentiation among populations (Wright 1922).

Hanks and Hooten (2013) further developed a spatial regression model of genetic differentiation that utilizes the connection between graph edge weights and genetic distance and the (generalized) Wishart likelihood. We fit a similar model here, with some minor modifications.

Let $w_{ij}$ be the edge weight connecting plants $i$ and $j$, which we assume is positively related to gene flow between the spatial locations of plants $i$ and $j$. We model the edgeweights using the log-linear model

$$w_{ij} = \exp\left\{\mathbf{g}_{ij}^T\boldsymbol{\beta} + \frac{1}{2}(\mathbf{x}_i + \mathbf{x}_j)^T\boldsymbol{\gamma} + \mathbf{Z}_i\boldsymbol{\alpha} + \mathbf{Z}_j\boldsymbol{\alpha}\right\},$$

where:

- $\mathbf{g}_{ij}$ is a $p_g$ vector of explanatory variables that can be measured on the edge connecting nodes $i$ and $j$ (e.g., length of the edge, whether the edge crosses some biologically relevant landscape feature, etc.).
- $\boldsymbol{\beta}$ is a $p_g$ vector of regression coefficients.
- $\mathbf{x}_i$ is a $p_x$ vector of explanatory variables measured on the node (i.e., at the location of the plant, such as canopy cover above plant $i$).
- $\boldsymbol{\gamma}$ is $p_x$ vector of regression coefficients.
- $\underset{n\times m}{\mathbf{Z}}$ is an indicator matrix indicating to which of the $m$ groups (meadows in this case) individual $i = 1, 2, ..., n$ belongs.
- $\boldsymbol{\alpha} \sim \mathcal{N}(\mathbf{0}, \tau \underset{m\times m}{\mathbf{I}})$, $\tau > 0$, $\mathbf{I}$ is the identity matrix.

Let $\mathbf{W}$ be the matrix of edge weights with $w_{ij} = 0$ if $i = j$. A *conditional autoregressive spatial* model assumes that a given measurement, $\eta_i$, $i = 1, 2, ..., n$, in a spatial graph (a genotype in our case) is a weighted average of the values of the nodes to which it is connected, with $w_{ij}$ the weight of the influence of node $j$ on the value of node $i$. Under this model, it can be shown that the covariance matrix defined by $\text{Var}(\boldsymbol{\eta}) = \boldsymbol{\Sigma}$ is

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

where $\mathbf{M}$ is a diagonal matrix with $m_{ii} = \sum_{j=1}^{n} w_{ij}$ and all off-diagonal elements equal to zero, and $\rho$ controls the amount of spatial autocorrelation in genotypes. Using results from (McCullagh 2009), we define a model relating $\mathbf{D}$, the genetic distance matrix, to $\boldsymbol{\Sigma}$. Specifically,

$$\mathbf{L}(-\mathbf{D})\mathbf{L}^T \sim \mathcal{W}_\kappa(\mathbf{L}2(\mathbf{M} - \rho\mathbf{W})^{-1}\mathbf{L}^T),$$

where $\mathcal{W}_k(\boldsymbol{\Gamma})$ denotes a Wishart distribution with degrees of freedom parameter $\kappa$ and scale matrix $\boldsymbol{\Gamma}$.

This model differs from that of Hanks and Hooten (2013) in three ways. First, we include geographic distance as an explanatory variable in the log-linear model for the weights instead of multiplying the regression equation by the inverse of distance. We felt this was more natural since it allows edge weights to go to zero at great distances if the regression parameter associated with geographic distance is negative. Second, while the theoretical connection between this model and resistance distance (McRae 2006; Hanks and Hooten 2013) is stronger in the case where $\rho = 1$ (i.e., an *intrinsic autoregressive model* - ICAR model), this defines a rank-deficient precision matrix. Because generalized inverses for rank-deficient matrices are known to be numerically unstable, we chose to put a prior on $\rho$ with most prior weight towards 1. This allows the model to tend towards the ICAR model without resulting in a rank-deficient precision matrix. Finally, Hanks and Hooten (2013) define $\kappa = K$, which is true when all loci are neutral and mutually independent. Because we do not know if this is the case for our SNP data, we assume $\kappa$ is an unknown parameter to be estimated.

**Explanatory variables of interest**

- **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 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.
- **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.
- **Interaction between distance and density of conspecifics around plants $i$ and $j$**. We hypothesized that, for two plants that are far apart, high conspecific density may have a negative effect on

genetic similarity if greater densities reduce the likelihood of any one plant being visited by a pollinator. However, at small distances,

- **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 $j$.
- **Interaction between distance and connectivity (isolation)**. Similar to the hypothesized relationship among genetic distance, geographic distance, and canopy cover, we expected that plants in isolated meadows may experience increased rates of mating among relatives if fewer pollinators travel to and from isolated locations. Thus, we would expect a positive effect of isolation on genetic similarity when comparing two plants found close in space, but a strong negative effect when comparing two plants found far apart.
- **Indicator for instances in which plants were sampled from the same meadow complex**. Due to shared histories of founding events and other demographic processes that are known to affect genetic diversity and similarity, we expected that two plants that were sampled from the same meadow complex may be more closely related than those sampled from different meadow complexes, all else equal.
- **Indicator for instances in which plants were sampled from the same meadow**. Similar reasoning to above.

**Priors**

We assume weakly informative priors for the regression parameters such that

$$\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}_P,\ 5 \cdot \mathbf{I}_{(P \times P)}).$$

For the parameter $\rho$, we place most prior weight towards 1 with the prior

$$\rho \sim \mathcal{B}(5, 1),$$

where $\mathcal{B}(\alpha, \beta)$ is a Beta distribution with shape parameters $\alpha$ and $\beta$. For the degrees of freedom parameter $\kappa$, we also put most prior weight towards the upper bound of $K$, the number of SNP loci, and away from the lower bound of $N$, the number of individuals (or nodes). Let $\kappa' \sim \mathcal{B}(5, 1)$, then

$$\kappa = (K - N)\kappa' + N$$

such that $\kappa \in (N, K),\ K > N$, and most of the density is towards the upper limit. Finally, we define the prior for the standard deviation of the meadow effects, $\tau \sim$ half-Normal$(0, 2)$.

## Stan Model Code

```
functions{

// Function to create contrasts matrix L
  matrix contrasts(int N){
    matrix[N-1,N] L;
    for(i in 1:(N-1)){
      for(j in 1:N){
        if(i==j){L[i,j] = 1;}
        else if(j == (i+1)){L[i,j] = -1;}
        else{L[i,j] = 0;}
      }
    }
    return(L);
  }

// function to make weights matrix from regression array X and
// paramater vector v
  matrix make_W(int N, vector v, real[,,] X, vector a, matrix Z, real s){
    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 +
                          (Z[i,]*a)*s +
                          (Z[j,]*a)*s);
          Wmat[j,i] = Wmat[i,j];
        }
      }
    }
    return(Wmat);
  }

}


data{

  int<lower=1> N;          //number of individuals sampled
  int<lower=1> K;          //number of SNP loci
  int<lower=1> P;          //number of landscape and node variables
  int<lower=1> G;          // number of groups (meadows)

  real X[N,N,P];           //design array
  matrix[N,G] Z;           // random effect design matrix
  matrix[N,K] al_LD;       //allelic load matrix

}

// Transform allelic load matrix into contrasts on distances
transformed data{

  matrix[N-1,N] L = contrasts(N);
```

```
  matrix[N-1,N-1] S = 2*(L*al_LD)*((L*al_LD)');

}


parameters{

  vector[P] beta;              //regression parameters
  real<lower=0,upper=1> rho;   //spatial dependence
  vector[G] alpha_raw;         // random meadow effects
  real<lower=0> tau;           // scale parameter for meadow effects
  real<lower=0,upper=1> kappa_std; // degrees of freedom parameter

}

transformed parameters{

  matrix[N,N] W;                // spatial weights or conductance matrix
  vector[N] W_sum;              // vector of row sums of W
  matrix[N,N] M;                // diagonal matrix with W_sum along diagonal
  cov_matrix[N] Sigma;          // inverse of (M-rho*W)
  real<lower=N,upper=K> kappa;  // scaled and shifted degrees of freedom parameter


  W = make_W(N, beta, X, alpha_raw, Z, tau);  // define W based on loglinear model

  for(i in 1:N){
    W_sum[i] = sum(W[i,]);                  // sum of each row
  }

  M = diag_matrix(W_sum);                   // diagonal matrix M
```

```
  Sigma = inverse((M - (rho*W)));            // definition of Sigma

  kappa = (K-N)*kappa_std + N;               // shift and scale degrees of freedom parameter

}

model{

//priors
  beta ~ normal(0,5);
  rho ~ beta(5,1);
  kappa_std ~ beta(5, 1);
  alpha_raw ~ normal(0,1);
  tau ~ normal(0,2);



//likelihood
  S ~ wishart(kappa, L*(2*Sigma)*(L'));

}

generated quantities{

  matrix[N-1,N-1] pred;    // Posterior predictive draws
  real loglik;             // loglikelihood of posterior draws given data

  loglik = wishart_lpdf(S | kappa, L*(2*Sigma)*(L'));
  pred = wishart_rng(kappa, L*(2*Sigma)*(L'));

}
```

## Loading and formatting data

```
  load(here("Data", "Indiv_lndscp_gen.RData"))
```

**Format the design matrix into an $n \times n \times p$ array**

```
  X_comb <- array(dim = c(dim(X)[c(1,2)],10))
  X_comb[,,c(1,2)] <- X[,,c(1,2)]

# now average and standardize them
  X_comb[,,3] <- (X[,,"plant_density_i"] +
    X[,,"plant_density_j"])/2

  X_comb[,,4] <- ((X[,,"cover_i"]+
                   X[,,"cover_j"])/2)

# isolation
  X_comb[,,5] <- ((X[,,"forest_500m_i"]+
            X[,,"forest_500m_j"])/2)

# Each variable may interact with distance
  X_comb[,,6] <- X_comb[,,2]*X_comb[,,3]
  X_comb[,,7] <- X_comb[,,2]*X_comb[,,4]
  X_comb[,,8] <- X_comb[,,2]*X_comb[,,5]

# create a within-complex effect
  complexes <- group_by(col_popgen_data, COMPLEX) %>%
    summarise(n=n())

  complex_submats <- map(1:nrow(complexes),
                         ~matrix(data = 1,
                                 nrow = complexes$n[.x],
```

```
                                  ncol = complexes$n[.x]))
  X_comb[,,9] <- as.matrix(bdiag(complex_submats))

# create within-meadow effect
  meadows <- group_by(col_popgen_data, MEADOW_ID) %>%
    summarise(n=n())

  meadow_submats <- map(1:nrow(meadows),
                        ~matrix(data = 1,
                                nrow = meadows$n[.x],
                                ncol = meadows$n[.x]))
  X_comb[,,10] <- as.matrix(bdiag(meadow_submats))




  dimnames(X_comb)[[3]] <- c("intercept",
                             "dist",
                             "avg_pl_density_ij",
                             "avg_cover_ij",
                             "avg_forest500_ij",
                             "dens_by_dist",
                             "cover_by_dist",
                             "forest_by_dist",
                             "Isame_complex",
                             "Isame_meadow")
```

**Create meadow indexing matrix Z**

```
# Sanity check that things are in the right order
  all.equal(rownames(col_alLD), col_popgen_data$Sample)
```

```
## [1] TRUE
```

```
# define empty matrix
  Z <- matrix(data = 0,
              nrow = nrow(col_alLD),
              ncol = length(unique(col_popgen_data$MEADOW_ID)))

# get list of meadows
  mdws <- unique(col_popgen_data$MEADOW_ID)

# loop through to put 1's in blocks of Z
# that correspond to each meadow
for (i in 1:nrow(Z)) {
  col_id <- which(mdws == col_popgen_data$MEADOW_ID[i])
  Z[i, col_id] <- 1
}
```

**Remaining data inputs**

```
# list remainder of data inputs

  N <- dim(col_alLD)[1]
  num_covs <- dim(X_comb)[3]
  num_snps <- dim(col_alLD)[2]
  num_meadows <- length(mdws)

  mod_data <- list(N=N,
                   P=num_covs,
                   K=num_snps,
                   G=num_meadows,
                   X=X_comb,
                   Z=Z,
                   al_LD=col_alLD)
```

### Fitting the model

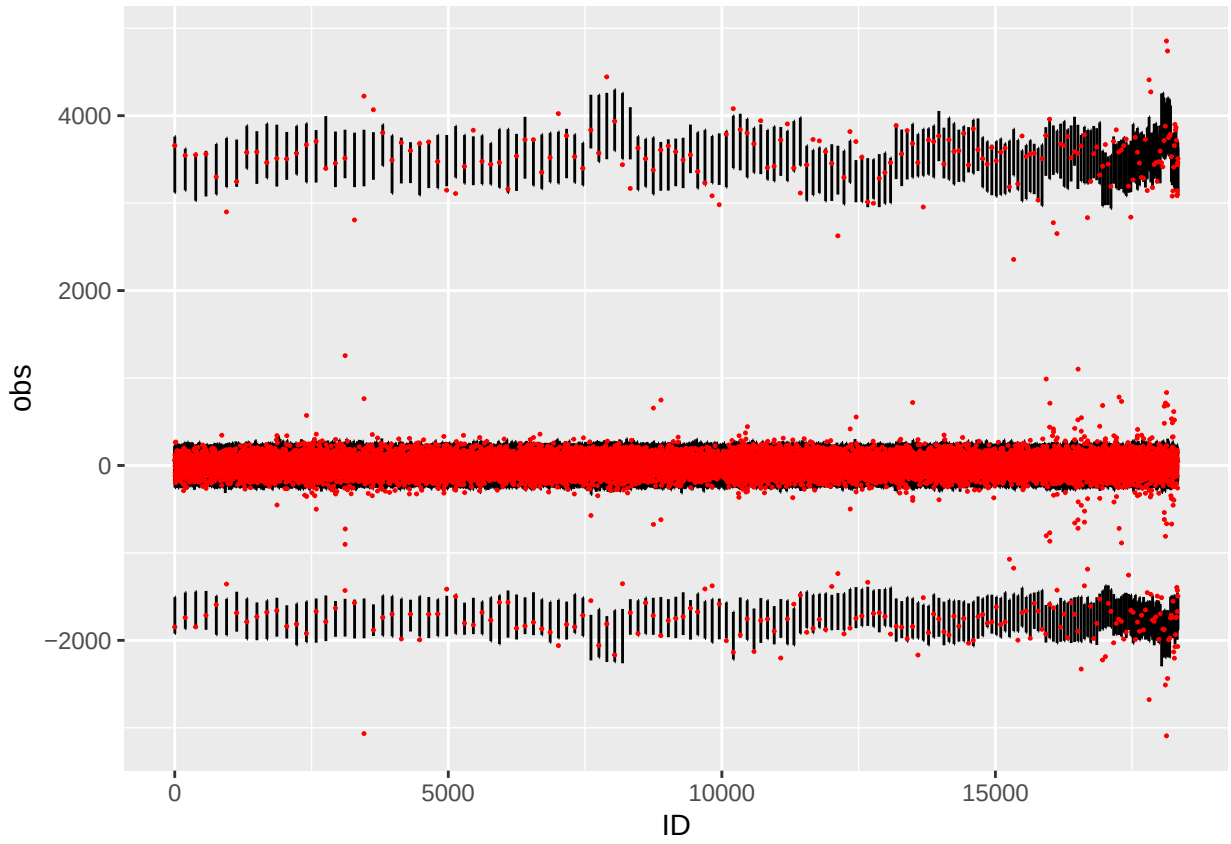
```
# fit the model
  mfit <- sampling(wishart_edgeweight_model_re,
                   data=mod_data,
                   iter=2000,
                   chains=3,
                   control=list(max_treedepth=12))

  # saveRDS(mfit, file =
  #           here(
  #             "Data",
  #             "mfit_full_dist_interactions_wm_wc_re.rds"
  #           ))
```

### Posterior predictive checks

```
# First build the contrasts matrix L
  L <- contrast_mat(
    mod_data$N
  )
# project the distance matrix down one dimension
  S_obs <- 2*(L%*%col_alLD)%*%t(L%*%col_alLD)

# Extract posterior predictions
  D_proj_pp <- rstan::extract(mfit, pars = "pred")

# convert random subset of predictions to a list of matrices
  rsamps <- sample(
    1:dim(D_proj_pp[[1]])[1],
    replace = F,
    size = 100
```

```
  )

  D_proj_pp_sub <- map(
    1:length(rsamps),
    ~as.matrix(D_proj_pp[[1]][.x,,])
  )

# convert to vectors of lower triangles
  preds <- map(
    D_proj_pp_sub,
    ~.x[lower.tri(.x, diag = T)]
  )

# convert to a matrix with columns as diagonal elements
#  of the wishart r.v.'s and
  preds_mat <- matrix(
    data = unlist(preds),
    nrow = length(preds),
    ncol = length(S_obs[lower.tri(S_obs, diag = T)]),
    byrow = T
  )

  obs_v_pred <- data.frame(
    obs = S_obs[lower.tri(S_obs, diag = T)],
    pred_mean = apply(
      preds_mat,
      MARGIN = 2,
      FUN = mean
    ),
    low = apply(
      preds_mat,
      MARGIN = 2,
```

```
      FUN = quantile,
      probs=0.025
    ),
    high = apply(
      preds_mat,
      MARGIN = 2,
      FUN = quantile,
      probs=0.975
    )
  )

  obs_v_pred$ID <- 1:nrow(obs_v_pred)

# plot the results
  ggplot(data = obs_v_pred)+
    geom_errorbar(aes(x=ID, ymin=low, ymax=high),
                  width=0.2)+
    geom_point(aes(x=ID, y=obs), colour="red",
               size=0.2)
```




```
# calculate proportion of observations outside the 95% prediction intervals


  mean(
    (obs_v_pred$obs < obs_v_pred$low) |
    (obs_v_pred$obs > obs_v_pred$high)
  )
```

```
## [1] 0.05890052
```

Hanks, Ephraim M., and Mevin B. Hooten. 2013. "Circuit Theory and Model-Based Inference for Landscape Connectivity." *Journal of the American Statistical Association* 108 (501): 22–33. https://doi.org/10.1080/01621459.2012.724647.

McCullagh, Peter. 2009. "MARGINAL LIKELIHOOD FOR DISTANCE MATRICES." *Statistica Sinica* 19 (2): 631–49.

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

Wright, Sewall. 1922. "Coefficients of Inbreeding and Relationship." *The American Naturalist* 56 (645): 330–38.