All of the correlations

Bob Week

7/7/2021

Single species

We assume diploid genotypes $\Gamma_1\Gamma_2$, where Γ_i is the ith haplotype. Assuming a genome composed of L biallelic, we write $\Gamma_i = (\gamma_1^i, \dots, \gamma_L^i)$ where γ_ℓ^i is the allelic state of an individual at the ℓth locus on the ith haplotype. For each $\ell = 1, \dots, L$, suppose $p_\ell(x)$ is the probability that an allele at locus ℓ of an individual at location $x = (x_1, x_2)$ is one (and one minus probability that it is zero).

The allele frequency surface $p_\ell(x)$ as a Beta random field

At a single locus

Suppose the allele frequency surface $p_\ell(x)$ is a Beta random field. By this we mean that for each geographical location $x \in \mathbb{R}^2$ and each locus $\ell = 1, \dots, L$, $p_\ell(x) \sim \text{Beta}(\alpha_\ell(x), \beta_\ell(x))$. Furthermore, the values of p_ℓ at distinct locations x, y are correlated following a spatially homogeneous correlation function $\kappa_\ell(\|x - y\|)$. In particular, this yields

$$\mathbb{E}[p_\ell(x)] = \frac{\alpha_\ell(x)}{\alpha_\ell(x) + \beta_\ell(x)},$$

$$\mathbb{V}[p_\ell(x)] = \frac{\alpha_\ell(x)\beta_\ell(x)}{(\alpha_\ell(x) + \beta_\ell(x))^2(\alpha_\ell(x) + \beta_\ell(x) + 1)}.$$

$$\mathbb{C}[p_\ell(x), p_\ell(y)] = \kappa_\ell(\|x - y\|)\sqrt{\mathbb{V}[p_\ell(x)]\mathbb{V}[p_\ell(y)]}.$$

For simplicity we assume spatially homogeneous parameters $\alpha_\ell(x) = \alpha_\ell > 0$, $\beta_\ell(x) = \beta_\ell > 0$. Under the assumption that $p_\ell(x)$ is Beta distributed and that, conditional on $p_\ell(x)$, $\gamma_\ell^i(x) \sim \text{Bern}(p_\ell(x))$, we can use the law of total expectation, total variance and total covariance to compute

$$\mathbb{E}[\gamma_\ell^i(x)] = \mathbb{E}[\mathbb{E}[\gamma_\ell^i(x)|p_\ell(x)]] = \frac{\alpha_\ell}{\alpha_\ell + \beta_\ell},$$

$$\mathbb{V}[\gamma_\ell^i(x)] = \mathbb{E}[\mathbb{V}[\gamma_\ell^i(x)|p_\ell(x)]] + \mathbb{V}[\mathbb{E}[\gamma_\ell^i(x)|p_\ell(x)]] = \frac{\alpha_\ell\beta_\ell}{(\alpha_\ell + \beta_\ell)^2},$$

$$\mathbb{C}[\gamma_\ell^i(x), \gamma_\ell^j(y)] = \mathbb{E}[\mathbb{C}[\gamma_\ell^i(x), \gamma_\ell^j(y)|p_\ell]] + \mathbb{C}[\mathbb{E}[\gamma_\ell^i(x)|p_\ell(x)], \mathbb{E}[\gamma_\ell^j(y)|p_\ell(y)]]$$

$$= 0 + \mathbb{C}[p_\ell(x), p_\ell(y)] = \kappa_\ell(\|x - y\|)\frac{\alpha_\ell\beta_\ell}{(\alpha_\ell + \beta_\ell)^2(\alpha_\ell + \beta_\ell + 1)},$$

where $i, j = 1, 2$ and $i = j$ when $x = y$ (since we need to think about probability of homozygosity in that case). Note the covariance of $\gamma_\ell^i(x)$ and $\gamma_\ell^j(y)$ is zero when conditioned on p_ℓ because we assume all the covariance comes from spatial patterns in the allele frequencies and not from the sampling of alleles from this spatial distribution.

When considering allele frequencies sampled at locations $x^{(1)}, \dots, x^{(n)}$, the vector $(p_\ell(x^{(1)}, \dots, p_\ell(x^{(n)}))^\top$ follows a multivariate generalization of the Beta distribution. A popular choice for this multivariate distribution is the Dirichlet distribution. The Dirichlet distribution is parameterized by the vector of concentration parameters $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_n)^\top$. Setting $\alpha_0 = \sum_{i=1}^n \alpha_i$, this implies

$$\mathbb{E}[p_\ell(x^{(i)})] = \frac{\alpha_i}{\alpha_0},$$

$$\mathbb{V}[p_\ell(x^{(i)})] = \frac{\alpha_i(\alpha_0 - \alpha_i)}{\alpha_0^2(\alpha_0 + 1)},$$

$$\mathbb{C}[p_\ell(x^{(i)}), p_\ell(x^{(j)})] = -\frac{\alpha_i \alpha_j}{\alpha_0^2(\alpha_0 - 1)}$$

Setting

$$\frac{\alpha_\ell}{\alpha_\ell + \beta_\ell} = \frac{\alpha_i}{\alpha_0},$$

$$\frac{\alpha_\ell \beta_\ell}{(\alpha_\ell + \beta_\ell)^2} = \frac{\alpha_i(\alpha_0 - \alpha_i)}{\alpha_0^2(\alpha_0 + 1)},$$

$$\kappa_\ell(\|x^{(i)} - x^{(j)}\|) \frac{\alpha_\ell \beta_\ell}{(\alpha_\ell + \beta_\ell)^2(\alpha_\ell + \beta_\ell + 1)} = -\frac{\alpha_i \alpha_j}{\alpha_0^2(\alpha_0 - 1)},$$

we can solve the corresponding $\boldsymbol{\alpha}$, which depends on the sample size n. Hence, using this parameterization we can directly sample $(p_\ell(x^{(1)}, \dots, p_\ell(x^{(n)}))^\top$ from a Dirichlet distribution by specifying α_ℓ, β_ℓ and $\kappa_\ell(h)$.

At multiple loci

While $\kappa_\ell(h)$ is the spatial autocorrelation function of allele frequencies at locus ℓ, we denote by $\lambda_{\ell\ell'}(h)$ the spatial cross-correlation function of allele frequencies at loci ℓ and ℓ'. Following this model, we compute the covariance between the alleles $\gamma_\ell^i(x)$ and $\gamma_{\ell'}^j(y)$ as

$$\mathbb{C}[\gamma_\ell^i(x), \gamma_{\ell'}^j(y)] = \mathbb{E}[\mathbb{C}[\gamma_\ell^i(x), \gamma_{\ell'}^j(y)|p]] + \mathbb{C}[\mathbb{E}[\gamma_\ell^i(x)|p_\ell(x)], \mathbb{E}[\gamma_\ell^j(y)|p_{\ell'}(y)]]$$

$$= 0 + \mathbb{C}[p_\ell(x), p_{\ell'}(y)] = \lambda_{\ell\ell'}(\|x - y\|) \sqrt{\frac{\alpha_\ell \beta_\ell}{(\alpha_\ell + \beta_\ell)^2(\alpha_\ell + \beta_\ell + 1)} \frac{\alpha_{\ell'} \beta_{\ell'}}{(\alpha_{\ell'} + \beta_{\ell'})^2(\alpha_{\ell'} + \beta_{\ell'} + 1)}}.$$

- Would it be useful to model $(1 + r_{\ell,\ell'}(x))/2$ as a Beta random field?

Spatial correlation between allelic states

Following the above model, we have $\mathrm{Corr}[\gamma_\ell^i(x), \gamma_\ell^i(y)] = \kappa_\ell(\|x - y\|)$. Hence, $\kappa_\ell(h)$ captures the degree of spatial autocorrelation for alleles at locus ℓ sampled at a distance $h > 0$. For simplicity, we assume $\kappa_\ell(h)$ belongs to the Matern class of spatial correlation functions. In particular, this implies

$$\kappa_\ell(h) = M(h|\nu_\ell, \xi_\ell) = \frac{2^{1-\nu_\ell}}{\Gamma(\nu_\ell)} \left(\sqrt{2\nu_\ell}\frac{h}{\xi_\ell}\right)^{\nu_\ell} K_{\nu_\ell}\left(\sqrt{2\nu_\ell}\frac{h}{\xi_\ell}\right),$$

where here Γ is the gamma function, K_{ν_ℓ} is a modified Bessel function of the second kind, ν_ℓ captures how smooth p_ℓ is as a function of x and ξ_ℓ is the characteristic length of spatial autocorrelation in allele frequencies at locus ℓ.

Homozygosity at a locus

Denote by $\rho_\ell(x)$ the correlation between allelic states on each haplotype at locus ℓ for an individual sampled at location x. That is,

$$\rho_\ell(x) = \mathrm{Corr}[\gamma_\ell^1(x), \gamma_\ell^2(x)].$$

Then, as a function of x, $\rho_\ell(x)$ provides a spatial map of homozygosity at the locus ℓ. We might consider modeling $(1 + \rho_\ell(x))/2$ as another Beta random field.

The probability that the two alleles observed at locus ℓ sample from location x are homozygous is

$$\mathbb{P}[\gamma_\ell^1(x) = \gamma_\ell^2(x) = 1] + \mathbb{P}[\gamma_\ell^1(x) = \gamma_\ell^2(x) = 0].$$

According to this model, we have

$$\mathbb{P}[\gamma_\ell^1(x) = \gamma_\ell^2(x) = 1] = p_\ell(x)^2 + \rho_\ell(x)p_\ell(x)(1 - p_\ell(x)),$$

$$\mathbb{P}[\gamma_\ell^1(x) = \gamma_\ell^2(x) = 0] = (1 - p_\ell(x))^2 + \rho_\ell(x)p_\ell(x)(1 - p_\ell(x)).$$

Hence, the probability of homozygosity is $1 - 2p_\ell(x)(1 - p_\ell(x))(1 - \rho_\ell(x))$.

Correlation among loci sampled at different locations

We have that $\lambda_{\ell\ell'}(\|x - y\|)$ equal to the correlation between allelic states at locus ℓ sampled at location x and locus ℓ' sampled at location y. Hence, $\lambda_{\ell\ell'}(h)$ is the spatial cross-correlation between loci ℓ and ℓ' sampled at the distance $h \geq 0$. As a simple model, we can follow the literature on spatial cross-correlations and assume $\lambda_{\ell\ell'}(h) = r_{\ell\ell'}M(h|\nu_{\ell\ell'}, \xi_{\ell\ell'})$ where $r_{\ell\ell'}$ is the collocated correlation between loci ℓ and ℓ' (corresponding to the classical measure of linkage disequilibrium) and $M(h|\nu_{\ell\ell'}, \xi_{\ell\ell'})$ is a Matern correlation function with smoothness parameter $\nu_{\ell\ell'}$ and characteristic length $\xi_{\ell\ell'}$. For this model to be valid, $\nu_{\ell\ell'}$ and $\xi_{\ell\ell'}$ must satisfy specific conditions. Sufficient (and parsimonious) conditions include $\nu_{\ell\ell'} = (\nu_\ell + \nu_{\ell'})/2$ and $\xi_{\ell\ell'} = \min(\xi_\ell, \xi_{\ell'})$. If loci ℓ and ℓ' are evolving completely neutrally, then we may assume $\nu_\ell = \nu_{\ell'}$, $\xi_\ell = \xi_{\ell'}$ and $r_{\ell\ell'}$ equals some baseline amount due to equilibrium between drift, gene-flow and recombination.

- Perhaps this assumption may lead to a test for non-neutrality. What is the distribution of $r_{\ell\ell'}$ under neutrality?

Doing descriptive stats

At a single locus

To describe spatial patterns of genetic variation, we can use the model described above to fit, for each locus ℓ, the four parameters: $\alpha_\ell, \beta_\ell, \nu_\ell, \xi_\ell$. The parameters α_ℓ, β_ℓ determine the expected allele frequency averaged across space along with the magnitude of spatial fluctuations in allele frequency. On the other hand, ν_ℓ and ξ_ℓ respectively determine the roughness of $p_\ell(x)$ and characteristic length of spatial autocorrelation in $p_\ell(x)$. If we consider datum to take the form $(\gamma_\ell^i(x), \gamma_\ell^i(y), \|x-y\|)$, then with a sample of size n, this makes for $2n(2n-2)$ points in our data set to infer the four parameters $\alpha_\ell, \beta_\ell, \nu_\ell$, and ξ_ℓ.

At multiple loci

Following our model of spatially cross-correlated loci, we need to identify the collocated correlations r_{ij}, $i, j = 1, \ldots, L$. That makes for $L(L-1)$ parameters to estimate. If we consider datum to take the form $(\gamma_\ell^i(x), \gamma_{\ell'}^j(y), \|x-y\|)$, then with a sample of size n, this makes X points in our data set.

Two species

Transpecific linkage

Since we are interested in studying spatial patterns of transpecific linkage disequilibrium, we extend our notation to account for two species. Set L_1, L_2 the number of loci in each species. Since sampled individuals from each species will likely be sampled from distinct locations, we need to consider spatial cross-covariance between loci in the different species. In particular, say we sample an individual from species one at location x and an individual from species two at location y, then we can write the correlation of alleles at locus i in our sample from species one and locus j in our sample from species two as $\tau_{ij}(\|x-y\|)$, where we assume this covariance depends only on the distance between the sampled individuals and so is spatially homogeneous.

$$\tau_{ij}(h) = \theta_{ij} M(h|\nu_{ij}, \xi_{ij})$$

where $\nu_{ij} = (\nu_i + \nu_j)/2$ and $\xi_{ij} = \min(\xi_i, \xi_j)$.

A multivariate Beta distribution

Suppose we want n Beta distributed variables $B_1, \dots, B_n$ that may be correlated with each other. These can be constructed from a 2^n dimensional Dirichlet distribution. Set $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_{2^n})$ such that $\alpha_i \geq 0$ for each $i = 1, \dots, 2^n$ and $\boldsymbol{X} = (X_1, \dots, X_{2^n})^\top \sim \mathrm{Dir}(\boldsymbol{\alpha})$.

For $n = 4$:

We have $\boldsymbol{X} = (X_1, \dots, X_{16})^\top$

$$\mathcal{X}_0 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

$$\mathcal{X}_1 = \begin{pmatrix} X_5 & 0 & 0 & 0 \\ 0 & X_4 & 0 & 0 \\ 0 & 0 & X_3 & 0 \\ 0 & 0 & 0 & X_2 \end{pmatrix}$$

$$\mathcal{X}_2 = \begin{pmatrix} 0 & 0 & 0 & X_8 & X_7 & X_6 \\ 0 & X_{10} & X_9 & 0 & 0 & X_6 \\ X_{11} & 0 & X_9 & 0 & X_7 & 0 \\ X_{11} & X_{11} & 0 & X_8 & 0 & 0 \end{pmatrix}$$

$$\mathcal{X}_3 = \begin{pmatrix} 0 & X_{14} & X_{13} & X_{12} \\ X_{15} & 0 & X_{13} & X_{12} \\ X_{15} & X_{14} & 0 & X_{12} \\ X_{15} & X_{14} & X_{13} & 0 \end{pmatrix}$$

$$\mathcal{X}_4 = \begin{pmatrix} X_{16} \\ X_{16} \\ X_{16} \\ X_{16} \end{pmatrix}$$

$$2^4 = 16 = 1 + 4 + 6 + 4 + 1 = \mathcal{C}_0^4 + \mathcal{C}_1^4 + \mathcal{C}_2^n + \mathcal{C}_3^4 + \mathcal{C}_4^4$$

where $\mathcal{C}_k^n = n!/k!(n-k)!$.

For n:

We use a known result from combinatorics:

$$2^n = \sum_{k=0}^{n} \frac{n!}{k!(n-k)!} = \sum_{k=0}^{n} \mathcal{C}_k^n.$$

Each component of this sum corresponds to a $n \times \mathcal{C}_k^n$ matrix, where $\mathcal{C}_k^n = n!/k!(n-k)!$. Hence, there will be $n+1$ matrices. Denoting the first matrix $\mathcal{X}_0$, we always have $\mathcal{X}_0 = \mathbf{0}_n$, where $\mathbf{0}_n$ is the n-dimensional zero vector. Denoting the final matrix $\mathcal{X}_n$, we always have $\mathcal{X}_n = \mathbf{1}_n$, where $\mathbf{1}_n$ is the n-dimensional vector with unit entries. Denoting the kth matrix $\mathcal{X}_k$, we have that $\mathcal{X}_k$ is a $n \times \mathcal{C}_k^n$ matrix with linearly independent columns each consisting of k unit entries and $n-k$ zero entries. Then, the transformation matrix from the 2^n-dimensional Dirichlet random variable $\boldsymbol{X}$ to the n-dimensional Beta random variable $\boldsymbol{B}$ is given by

$$\boldsymbol{\mathcal{X}} = (\mathcal{X}_0|\cdots|\mathcal{X}_n).$$

In particular, $\boldsymbol{B} = \boldsymbol{\mathcal{X}}X$.

Trying it out in R

```
require(stats)
require(rSPDE) # for matern cov fct
require(matlab) # ones and zeros
require(gtools) # Dirichlet
require(matrixStats)
require(ggplot2)

#
# starting by focusing on a single locus
#

# using a R-by-R region
R = 100

# sample locations uniformly
n = 3
x1 = runif(n,0,R)
x2 = runif(n,0,R)
# xi is the vector of ith coordinates of sampled individuals

# desired parameters for multivar Beta distr
a = 5
b = 5
nu = 0.5
xi = 25
m = a/(a+b)
V = a*b/((a+b)^2*(a+b+1))

# n-by-n distance matrix
geoDist = zeros(n)
for(i in 1:n){
  for(j in 1:n){
    if(i!=j) geoDist[i,j] = sqrt((x1[i]-x1[j])^2+(x2[i]-x2[j])^2)
  }
}

# n-by-n covariance matrix
covMat = ones(n)
for(i in 1:n){
  for(j in 1:n){
    covMat[i,j] = matern.covariance(geoDist[i,j],1/xi,nu,sqrt(V))
  }
}
```

```
cov2cor(covMat)

# build transformation matrix to go from 2^n-dim Dirichlet to n-dim Beta
trfM = zeros(n,1)
for(k in 1:n){
  K = choose(n,k)
  combos = combn(1:n,k)
  XX = zeros(n,K)
  for(i in 1:K){
    XX[combos[,i],i] = 1
  }
  trfM = cbind(trfM,XX)
}

# find which Dirichlet components correspond to which Beta's
DirBet = c()
for(i in 1:n){
  DirBet = rbind(DirBet,which(trfM[i,]==1))
}

# for each pair of Beta's, find which Dir components are common
DirDir = c() # common Dir comps for each pairing, makes choose(n,2) rows
BetBet = c() # indices of each pairing, also choose(n,2) rows
for(i in 2:n){
  for(j in 1:(i-1)){
    BetBet = rbind(BetBet, c(j,i))
    DirDir = rbind(DirDir,t(intersect(DirBet[i,],DirBet[j,])))
  }
}

# find the Dir comps unique to each Beta rv in each pairing
DirUnq = c() # DirUng[i,1:n] & DirUnq[i,(n+1):(2*n)] are the unique Dir comps for each member of the it
for(i in 1:choose(n,2)){
  j = BetBet[i,1]
  k = BetBet[i,2]
  uj = t(setdiff(DirBet[j,],DirDir[i,]))
  uk = t(setdiff(DirBet[k,],DirDir[i,]))
  DirUnq = rbind(DirUnq, c(uj,uk))
}

# find the Dir comps not associated with each Beta rv pairing
DirCmp = c()
for(i in 1:choose(n,2)){
  j = BetBet[i,1]
  k = BetBet[i,2]
  un = union(DirBet[j,],DirBet[k,])
  cp = t(setdiff(1:(2^n),un))
  DirCmp = rbind(DirCmp, cp)
}

#
# numerically fit the 2^n-dim param aa of the Dirichlet to desired conditions
#
```

```r
# covariance in terms of aa components
Cov <- function(a1,a2,a12,a0){
  a11 = sum(a12)
  a10 = sum(a2)
  a01 = sum(a1)
  a00 = sum(a0)
  num = a11*a00 - a10*a01
  den = (a11+a10+a01+a00)*(a11+a10+a01+a00+1)
  return(num/den)
}

Var <- function(a,aa){
  alpha = sum(a)
  beta = sum(aa)-sum(a)
  num = alpha*beta
  den = (sum(aa)+1)*sum(aa)^2
  return(num/den)
}

ftaa <- function(a){

  a = exp(a)

  # build proposal cov matrix
  K = dim(DirUnq)[2]
  prpMat = zeros(n)
  prpmn  = zeros(n,1)
  for(i in 1:choose(n,2)){
    k = BetBet[i,1]
    l = BetBet[i,2]
    a1  = a[DirUnq[i,1:(K/2)]]
    a2  = a[DirUnq[i,(1+K/2):K]]
    a12 = a[DirDir[i,]]
    a0 = a[DirCmp[i,]]
    prpMat[k,l] = Cov(a1,a2,a12,a0)
    prpMat[l,k] = Cov(a1,a2,a12,a0)
  }
  for(i in 1:n){
    prpMat[i,i] = Var(a[DirBet[i,]],a)
  }

  # l2 dist between desired and proposed cov's
  pMv = prpMat[!lower.tri(prpMat)]
  cMv = covMat[!lower.tri(covMat)]

  fit = sqrt(sum((cMv-pMv)^2)+sum((m-prpmn)^2))

  return(fit)

}

ftaa(log(rexp(2^n)))
```

```
# fit a
afit = optim(log(rexp(2^n)),ftaa,control=list(maxit=20000))

afit$convergence

# get fitted a
af = exp(afit$par)

prpMat = zeros(n)
for(i in 1:choose(n,2)){
  k = BetBet[i,1]
  l = BetBet[i,2]
  a1  = af[DirUnq[i,1:(n-1)]]
  a2  = af[DirUnq[i,n:(2*(n-1))]]
  a12 = af[DirDir[i,]]
  a0 = af[DirCmp[i,]]
  prpMat[k,l] = Cov(a1,a2,a12,a0)
  prpMat[l,k] = Cov(a1,a2,a12,a0)
}
for(i in 1:n){
  prpMat[i,i] = Var(af[DirBet[i,]],af)
}

prpMat
covMat

# draw from 2^n-dim Dirichlet with 2^n-dim param aa
X = t(rdirichlet(100,af))

# generate the Beta distr allele frequencies
p = trfM%*%X

rowMeans(p)

m

rowVars(p)

V



# # sample allele frequencies
# chr1 = c()
# chr2 = c()
# for(i in 1:n){
#   P = p[x1[i],x2[i]]
#   chr1 = c(chr1,rbinom(1,1,P))
#   chr2 = c(chr2,rbinom(1,1,P))
# }
# # chri is the vector of alleles across sampled individuals on ith chromosome
#
#
```

```
# # put it all together in a dataframe
# xs = c()
# for(X1 in 1:X){
#   for(X2 in 1:X){
#     xs = rbind(xs,c(X2,X1))
#   }
# }
# p.df = data.frame(X1=xs[,1],X2=xs[,2],p=as.vector(p))
#
# df = data.frame(id=1:n,chr1=chr1,chr2=chr2,x1=x1,x2=x2)
#
# ggplot() + geom_raster(data=p.df, aes(x=X1,y=X2,fill=p)) +
#   geom_point(data=df, aes(x=x1,y=x2,colour=factor(chr1)))
```