MAPS: estimating recent Migration And Population size Surfaces

Hussein Al-Asadi, Desislava Petkova, John Novembre, Matthew Stephens

1 Methods

MAPS assumes a population genetic model consisting of a d x d triangular grid of demes with symmetric migration. The density of the grid is prespecified by the user with the consideration that the computational complexity is $O(d^3)$. We use Bayesian inference to estimate the MAPS parameters: m and q. Its key components are the likelihood, which measures how well the parameters explain the observed data, and the prior, which captures the expectation that m and q have some spatial structure (in particular, the idea that nearby edges will tend to have similar migration and coalescent rates).

1.1 The Likelihood

Let $N^u_{i,j}$ be the number of segments greater than u centimorgans between individuals i and j that have not experienced any intervening recombination. These UnRecombined segments (UnR) are commonly referred to as “IBD” segments; however, we prefer not use this ambigious notation. As u increases, UnR segments have shorter coalescent times and become increasingly sensitive to recent population structure and in practice, become a function of only recent population structure. In this study, we focus on characterizing population structure in the recent past and accordingly focus on UnR segments with short coalescent times which are very long (e.g. $\geq$ 4cM). Fortuitously, methods for calling UnR segments at this range (aka IBD calling) have high power and low false positive rates (Browning & Browning, 2007; Lowe et al, 2008; Raph & Coop, 2008 and Palamara et al., 2012).

Let $\theta = (m, q)$ encode the parameters of the demographic model, which are the migration rates and population sizes of a d by d grid of demes. Like Petkova et al 2015, we only allow migration between nearby demes to enfore the isolation by distance assumption.

Computing the likelihood requires many steps. following Palamara et al. 2012, the expectation of $N^u_{i,j}$ can be computed as,

$$E[N^u_{i,j}] \approx \frac{L\,E[f|\theta]}{E[s|\theta]} \tag{1}$$

where $E[f|\theta]$ is the expected fraction of the genome that lies in UnR segments greater than ucM, L the length of the genome in units of cM, and $E[s|\theta]$ the expected size of a UnR segment conditional on it is at least ucM. To gain intuition, it is helpful to see that equation (1) equates to saying that the average size of each UnR segment ($E[s|\theta]$) multiplied by the average number of segments ($E[N^u_{i,j}]$) is approximately the total length of the genome that lies in UnR segments ($L\,E[f|\theta]$).

Furthermore, following Palamara et al. 2012,

$$E[s|\theta] = \frac{\int_u^\infty p(l|\theta)dl}{\int_u^\infty p(l|\theta)/l\,dl}$$

$$E[f|\theta] = \int_u^\infty p(l|\theta)dl$$

where $p(l|\theta)$ is the probability that a randomly chosen base pair in the genome is contained with an UnR segment of at least ucM. Combining both equations together,

$$E[N^u_{i,j}|\theta] \approx L \int_u^\infty p(l|\theta)/l\,dl$$

Expanding $L\int_u^\infty p(l|\theta)/l\,dl$,

$$E[N^u_{i,j}|\theta] \approx L \int_u^\infty p(l|\theta)/l\,dl = L \int_u^\infty \int_0^\infty p(l,t|\theta)/l\,dt\,dl = L \int_0^\infty p(t|\theta) \int_u^\infty p(l|t)/l\,dl\,dt \tag{2}$$

where $p(l,t|\theta) = P(l|t,\theta)p(t|\theta) = p(l|t)p(t|\theta)$.

$p(l|t)$ is the probability of observing a UnR segment of exactly length l given a coalescent event at time t, this quantity can be derived by considering the probability of no recombination for t generations (see Hein et al, 2005) for a pair of individuals. Namely, given a site a, the probability of no recombination in t generations for x base pairs to the right of a is simply e^{-rtx} (where r is the recombination rate for one chromosome) and e^{-2rtx} (assuming independence) for both chromosomes. Consider the length of no recombination to left of a (X_L) and to the right of a (X_R), then $l = X_L + X_r$ is a convolution of two exponential random variables. Therefore,

$$p(l|t) = 4r^2t^2le^{-2trl} \tag{3}$$

Combining equation 2 with 3, the inner integral $\int_u^\infty p(l|t)/l\,dl$ can be computed analytically, leading to

$$E[N^u_{i,j}|\theta] \approx L \int_0^\infty p(t|\theta)2rte^{-2tru}dt$$

The next quantity to consider to evaluate the integral is $p(t|\theta)$, where θ encodes the demographic model. Under the structured coalescent,

$$p(T_{i,j} = t|\theta) = \sum_k q_k(e^{-Rt})_{(I(i),I(j))\to(k,k)} \tag{4}$$

which can be derived by considering all possible demes in which lineages i and j can coalesce (See Appendix). q_k is the coalescent rate of deme k, $I(i)$ indicates the deme which individual i belongs to, and R is the instantaneous rate matrix of the structured coalescent process for a sample size of two, see (Wakeley, 2008) for a detailed explanation of R. R is a matrix of size d^2 by d^2 and as a result becomes practically impossible to exponentiate for large d. Instead, we approximate the coalescent process as a random walk with absorption where,

$$p(T_{i,j} = t|\theta) \approx \sum_k q_k(e^{-Mt})_{I(i)\to k}(e^{-Mt})_{I(j)\to k}$$

M is a d by d matrix encoding a random walk on d vertices such that, $M_{\alpha,\beta}$ is the migration rate between individual deme α and deme β and $M = 0$ if demes α and β are unconnected. Or equivalently in matrix form,

$$P(t) \approx e^{-Mt}Qe^{-Mt}$$

where $Q = diag(q_1, ..., q_d)$ and $P(t)_{I(i),I(j)} = p(T_{i,j} = t|\theta)$. In most data-sets, not all demes have samples so we can save time by only computing $P(t)$ for sampled demes. To take advantage of this fact, we order the demes so the first o demes are observed and perform the matrix multiplication featured above so that it only takes $O(od)$ time instead of $O(d^2)$ time.

This random walk with absorption approximation effectively assumes that coalescent events before time t are neglible. Having all the components of equation 1, we are left to evaluate a one-dimensional integral which we do by gaussian quadrature. The quadrature routine only evaluates $p(T_{i,j} = t|\theta)$ at small t because the probability of observing long UnR segments with deep coalescent times is effectively zero, and therefore coalescent events in the deep past do not contribute any mass to the integral. Because the intergral is only evaluated at small t, the random walk approximation works very well because small t implies $qt \approx 0$ and as a result makes coalescent events before time t rare. We recently discovered that Baharian et al 2015 perform the same approximation.

Now that we can efficiently compute $E[N^u_{i,j}]$, we compute the distribution $N^u_{i,j}$ as

$$N^u_{i,j}|\theta \sim Pois(E[N^u_{i,j}|\theta]) \tag{5}$$

as done in Ralph & Coop 2008 and Palamara et al 2012. However, we assume a negative binomial distribution with over-dispersion parameter ϕ. By allowing overdispersion, we are fitting the extra-complexity often found in real data, and in addition, over-dispersion acts to heat the MCMC chain which allows for faster convergence.

$$N^u_{i,j}|\theta,\phi \sim NegBi()$$

such that,

$$E[N^u_{i,j}|\theta,\phi] = E[N^u_{i,j}|\theta]$$

$$Var[N^u_{i,j}|\theta, \phi] = E[N^u_{i,j}|\theta] + \phi E[N^u_{i,j}|\theta]^2$$

To find the probability of the entire data, we assume independence between pairs of individuals. Empirically, we found this to be a descent approximation. In addition, composite likelihoods have been widely used in population genetics; for example, FineStructure - a Bayesian method - computes a composite likelihood over all pairs to infer recent population structure (Lawson et al 2012).

1.2 Computing $E[N^R_{i,j}]$

Previously, we computed the expected number of UnR segments greater than u cM. It is also helpful to visualize UnR segments in a range $R = (u, v)$, this is simply,

$$E[N^R_{i,j}] = E[N^u_{i,j}] - E[N^v_{i,j}]$$

See Appendix.

1.3 The Prior

Essentially, MAPs uses the same prior as in EEMS with a few exceptions. Firstly, we estimate an additional overall mean rate for the coalescent rates wheras Petkova et al fixed the analogous parameter to one. Secondly, we estimate an overdispersion parameter in MAPs which is analogous to the degrees of freedom parameter in EEMS in that it acts as a fudge factor – i.e. it does not change the mean but only the level of uncertainity around the mean. Like the degrees of freedom parameter in EEMS, we place a log-uniform prior on ϕ.

1.4 Markov chain Monte Carlo estimation

MAPS uses the same MCMC routine outlined in Petkova et al to sample from the posterior distribution given the data. Please see Petkova et al 2015 for more details on the MCMC procedure.

2 Migration rates and Population sizes are identifiable in MAPs but not in EEMS

3 The affect of the grid on demography estimates