Practical 5 (12-13 Feb 2015)

The not-too-long practical: Joint estimation of effective population size and migration rate

Background

Effective population size (N_e) is defined as the size of an ideal population that has the same rate of change in allele frequencies as the observed population (Fisher, 1930; Wright, 1931). Migration rate (m) here refers to the fraction of individuals that move from one deme to another per generation.

It is known that by studying the change in allele frequency over time we may be able to estimate N_e (Nei and Tajima, 1981; Waples, 1989; Williamson and Slatkin, 1999; Wang, 2001). However all these methods assume a closed population with no migration.

The island model proposed by Sewell Wright (1931) assumes that an ancestral population was broken up into partially isolated demes with effective population size N_e. In the absence of other genetic forces, genetic differentiation among subpopulations (measured by F_{ST}) results from an equilibrium between genetic drift and migration. At equilibrium, $F_{ST} \approx \frac{1}{1+4N_em}$. The quantity N_em can be interpreted as the actual number of migrations exchanged among demes per generation. Although some statistics were developed to estimate F_{ST} (and then N_em) from genetic data, estimating N_e and m alone remains a challenge to population geneticists.

We wish to extend the study to jointly estimate N_e and migration rate m between two demes (Wang and Whitlock, 2003). The model assumes that the two demes have the same and constant effective population size and migration rate (the two-island model). Other genetic forces, such as selection and mutation, are negligible compared to genetic drift and migration.

RData file

The file `two_island.RData` contains everything you need for this exercise. Open a new R session and load it in via `load()` command. You can use `ls()` to view the objects.

Data

Two datasets named `sample.island.A` and `sample.island.B` are provided, and they refer to sampled allele counts at island A and island B respectively. Each dataset contains 2 columns: the first column are the samples at $t = 0$, the second column are the samples at $t = 8$. The dataset also has 500 rows, representing 500 loci sampled at each time

point. For example, the first row of `sample.island.A` is `(63,75)`, indicating the allele count for locus 1 on island A changes from 63 to 75 over $t = 8$ generations. Sample size at each time is 100 diploid individuals, so the maximum of allele counts is 200.

The log-likelihood model

Do not worry about the likelihood model as I've written for you. It is extremely complex and it took me several days to code the log-likelihood function and weeks to design it. Look for the function `log.likelihood.migration.model` and it is your log-likelihood function. It has several arguments: `parms, pop1, pop2, t, n`

`parms`: a vector of two parameters. The first element is N_e and the second is m.
`pop1`: the input dataset for the first population, `sample.island.A`
`pop2`: the input dataset for the second population, `sample.island.B`
`t`: time (in generations) between samples. In our case, 8
`n`: sampled diploid individuals (per deme). Here is 100

Tasks

1. To use the `optim()` function to find the maximum likelihood estimates for the two parameters, given set of data and model. It is known that N_e should be between 10 and 10000 (a very wide range indeed!), and m is always between 0 and 0.5. Return the hessian matrix as well. Save the results and call it `M`. What are the estimates?

2. Produce a contour plot of your log-likelihood function over a wide range of parameter values. In you plot, your x-axis will be N_e, running from 100 to 1000, with an increment of 20 in each step. Similarly, your y-axis will be m, ranging from 0 to 0.5, with step size 0.02. You may need `seq()` to define your parameter values, and `for` loop in calculating the log-likelihood values for every pair of parameters. Contour lines for every 10 units. Remember to subtract your log-likelihood values by the maxima; only the relative log-likelihood value matters.

3. Similar to task 2, produce a refined contour plot with a higher resolution. In task 2 we roughly know how the log-likelihood value decreases from its maxima, and we are now interested in the region surrounding the MLE (the peak). Plot another contour plot, but only contain points that are about -10 (or smaller if you prefer) from the maximum. Use a smaller step size in both N_e and m (I would suggest 1 for N_e and 0.005 for m). Contour lines for every 1 unit. Draw a cross representing the peak (maxima). After plotting the graph, compare yours with mine on the last page. They may not look exactly the same, but they should be of the same shape.

4. Find out the 95% confidence interval for N_e and m separately using likelihood method.

5. Find the joint 95% confidence region? Circle the region with a red dotted line.

6. Do you reject $\mathrm{H}_0: N_e = 500$ alone?

7. How about $\mathrm{H}_0: (N_e, m) = (500, 0.12)$? and why?

8. What can you tell about the correlation between the two parameter estimators? Can you work out the approximate variance-covariance structure of the two estimators? Can you find the approximate 95% Wald CI (that is, based on the variance-covariance structure and normality) for N_e and m separately?

9. [Extra] Multivariate testing. One of the properties of MLE is approximate normality. It means that when we repeat the experiments many many times and obtain many many mles, these mles will follow a (multivariate) normal distribution. In class we demonstrated how we can obtain the approximate variance-covariance matrix from the outputs of optim(). Now, we would like to go a step further. In univariate case, we can always perform a z-test to test whether the mean equals to a certain value. There is a multivariate version of the z-test, which is the chi-square test. The formula for this test is:

$$W = (\hat{\theta} - \theta_0)^T [\hat{V}(\underline{\theta})]^1 (\hat{\theta} - \theta_0) \dot{\sim} \chi_r^2$$

W is the test statistic and follows a chi-square distribution with degrees of freedom equals to the number of parameters considered jointly. We can test for the hypothesis $\mathrm{H}_0: (N_e, m) = (500, 0.12)$ by assuming MLE normality.

```
# MULTIVARIATE TESTING
d<-c(500-M$par[1], 0.12-M$par[2])
W<-t(d)%*%(-M$hessian)%*%d
W
```

Then you may check your observed W value against the chi-square table with 2 degrees of freedom (as we are testing two parameters at a time). If you W is larger than the critical value, then you reject the null hypothesis (that the population size is 500 AND migration rate is 0.12).

Partial answer for task 3