Migrate tutorial

From MolEvol

Contents

Tutorial Overview

How to get MIGRATE

Locally, the most recent copy of the software on the server is in the /class/shared/migrate_demo/local_distribution directory, you can either use the program that is already installed on the class server or use these files in the local_distribution directory to install MIGRATE on your own computer. if this fails download MIGRATE from the Migrate downloadwebsite (http://popgen.sc.fsu.edu/Migrate/Download.html).

Comparison of gene flow models using Bayes Factors with MIGRATE

Most are familiar wit the concept of likelihood ratio tests, or Akaike's Information criterion for model comparison. This tutorial describes how to compare models using Bayes Factors. These allow comparing nested and un-nested models, without assuming Normality, or large samples. Bayes factors are ratios of marginal likelihoods. In contrast to maximum likelihood, the marginal likelihood is the integral of the likelihood function over the complete parameter range. MIGRATE can calculate such marginal likelihoods for a particular migration model (Beerli and Palczewski 2010). This tutorial steps through all necessary program runs to calculate Bayes factors for comparing different gene flow models. We need to do the following:

- Decide on the models that are interesting for a comparison. The method does not work well for a fishing expedition where one would try to evaluate all models; this is possible only for a small model. It will be possible to enumerate all models for three populations but more will be very daunting.
- Run each model through MIGRATE. Use the same prior settings for each of them because the prior distribution has some influence on the Bayes factors. Use the heating menu to allow for at least four chains. The menu supplies a shortcut to specify the temperatures, it is #. It generates temperatures that are spaced in a particular way: they are spaced so that the inverse of the temperature are regularly spaced on the interval 0 to 1. For example, the 4 different chains have temperatures 1.0, 1.5, 3.0, 100,000.0, this results in the spacing 1.0, 0.666, 0.333, and 0.0.
- Compare the marginal likelihood of the different runs and calculate the Bayes factor and calculate the probability for each model.

The following pages detail all steps using a small example. We use a simulated dataset that was generated using parameters that force a direction of migration from the population Aadorf (A) to the population Bern (B). The Bern population is larger than the Aadorf population and no individual from Bern ever goes to Aadorf, but Bern receives about 1 migrant per generation from Aadorf. The dataset name is **twoswisstowns** (if you are not at the workshop then download here (http://popgen.sc.fsu.edu/tutorials/BF_migrate_tutorial/twoswisstowns)) We will evaluate two sets of models: first set will have 5 models, second set has 2 Models

1. a full model with two population sizes and two migration rates (from A to B and from B to A);
2. a model with two population sizes and one migration rate to Bern;
3. a model with two population sizes and one migration rate to Aadorf;
4. a model where Aadorf and Bern are part of the same panmictic population.
5. a model where Aadorf is ancient city and Bern was built anew from people who left Aadorf.

We know the truth therefore we have some prejudice about the ranking of the models, **model 2** should be *best*, **model 1**, because it allows for the same migration direction as **model 2** should be ranked *second*. Whether **model 3** is better than **model 4** is unknown a priori and may depend on the strength of the data. First we need to figure out how to run the dataset efficiently in MIGRATE. For that we pick the most complicated **model 1** and experiment with run conditions until we are satisfied that the run converges and delivers posterior distributions that look acceptable. Here are detailed instructions how to rank population genetics models for a particular dataset.

Familiarize with MIGRATE [Tutorial Start]

- Make a new directory and download or copy the datafile

```
#if you are AT THE WORKSHOP use this
rsync -avz /class/molevol-shared/migrate_lab .
cd migrate_lab
```

```
#if you are NOT at the workshop use this
mkdir migrate_lab
cd migrate_lab
wget http://popgen.sc.fsu.edu/tutorials/BF_migrate_tutorial/twoswisstowns
wget http://popgen.sc.fsu.edu/tutorials/BF_migrate_tutorial/twoswisstownsdiv
```

- Start the program: the regular distribution comes in two flavors the single cpu processor version called **migrate-n** and the parallel processing version that runs on cluster or computers with multiple cores is called **migrate-n-mpi**, but we use a a brandnew version of MIGRATE: version 4.0. To avoid confusion this is called on the cluster **migrate4**. (In this text I will call the program from now on simply MIGRATE).

We will run the exercise on the server. On the server type

```
migrate4
```

On your laptop you may need to use "./migrate4", if the program is in the same directory [I hope to have binaries for macs running MacOS 10.9 available. The main menu will appear, looking like this

```
[pbeerli@class-02 migrate_lab]$ migrate4


  ============================================================
  POPULATION SIZE, MIGRATION, DIVERGENCE, ASSIGNMENT, HISTORY
  Bayesian inference using the structured coalescent
  ============================================================
  PDF output enabled [Letter-size]
  Version 4.0   [2022]
  Program started at   Sat Aug  2 19:23:59 2014

  Settings for this run:
  D       Data type currently set to: DNA sequence model
  I       Input/Output formats and Event reporting
  P       Parameters  [start, migration model]
  S       Search strategy
  W       Write a parmfile
  Q       Quit the program

  To change the settings type the letter for the menu to change
  Start the program with typing Yes or Y
===>
```

- Go to the **Input/Output formats** menu (press **I** and hit Enter), in the **INPUT** section change the Datafile name to **twoswisstowns**, Return to the main menu by typing **Y**.
- In the **Search strategy** menu: Change the **Number of recorded steps in chain** to **1000**, and also change the **Burn-in for each chain:** to **1000**. Do not worry about priors or other runtime options for the moment. Return to the main menu.
- Save the changes by using the menu item **Write a parmfile**. This will write a file named **parmfile**.
- Now run the program (**pressing Y** will start the run if you are in the main menu). For this dataset the runtime will be very short. On a modern

computer this will take under a minute. If this takes more than 3 minute, something is not set up correctly! On the server this takes about *20 seconds* (on my macOS 10.9: 5 seconds).

- The program writes considerable information during the run to the screen, that gives some information about the run. Most interesting are the acceptance ratio for the genealogy and the autocorrelations of the parameter and the genealogy. If the autocorrelation is high and the effective sample size is low (<500) then a longer run may be needed. If the priors boundaries are to tight, then you will see that the values reported are either very close or exactly at the upper prior boundary, in these cases you need to extend the prior range. See prior problems in the output, but for this dataset we will have no such problems.
- Look at the **outfile.pdf**, you will need to transfer the pdf file to your computer and use preview or acrobat or another PDF viewer. In the outputfile look at the figures labeled Bayesian analysis: Posterior distribution, you see histograms similar to the ones in Figure 1. We expect single peaks where the shading of the histogram shows one dark block in the center (50% credibility set), two light gray bars indicating the extent of the 95% credibility set, and two lighter gray bars indication the 99% credibility set.

- In your investigation of Figure 1 you recognize that the histogram does not look very smooth because our run was too short, now restart MIGRATE and set in the strategy menu the setting for change the **number of recorded steps in chain** from 1,000 to **10,000**. This will lengthen the run by a factor of 10 (my run needed 112 sec). Don't forget to write the parmfile to save the settings. Run and compare the results (Figure 2) with the histogram from before. You will recognize that the longer run has somewhat smoother histograms, and the double peaks vanish (hopefully). With your own data you may want to do another round of refinements, but eventually, by comparing the medians and modes of the parameters in the table and the shape of the histograms you should see a good agreement on similar values, if the modes of the different runs are not within the 50% credibility intervals you certainly need to run longer.

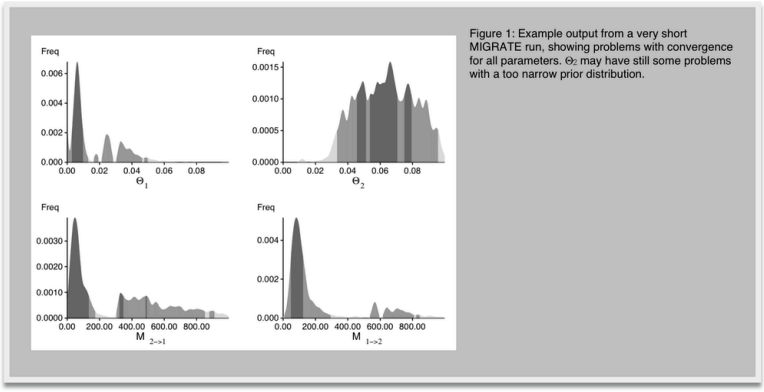


Figure 1: Example output from a very short MIGRATE run, showing problems with convergence for all parameters. Θ_2 may have still some problems with a too narrow prior distribution.

Report marginal likelihoods of Model 1 (4 Parameter)

[remember work with your neighbor, the runs will take a while and the best strategy to work together entering the options on one computer, starting the analysis, then update the options on the other computer and work on the next step in the tutorial]

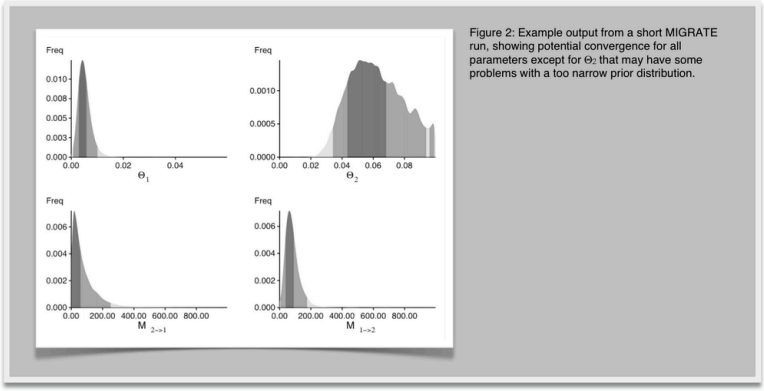
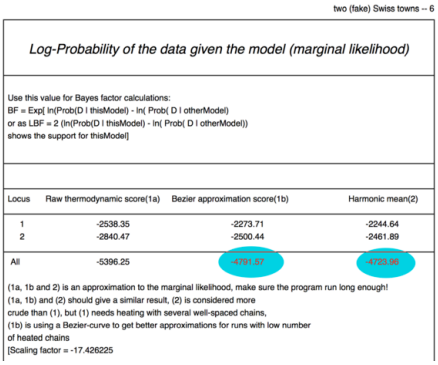


Figure 2: Example output from a short MIGRATE run, showing potential convergence for all parameters except for Θ_2 that may have some problems with a too narrow prior distribution.

- Let's assume that our runs are all satisfactory. We turn now to the best estimation of the marginal likelihood to compare models. Because we want to use the thermodynamic integration method, we need to turn on heating. Start MIGRATE, use the strategy menu and turn on heating, use static heating. MIGRATE will tell what to do next, you will need to enter 4 chains sampling at every tenth (10) interval using the temperature scheme that is suggested with the character #. Save the parmfile, and run. This will take about 4x longer than before. It should give a better posterior distribution histogram and will add a full table of (natural) log marginal likelihoods is shown towards the end of the outfile.pdf. On the server this takes about 16 minutes.

- Come to the front and write down the log marginal likelihood into the spreadsheet (look at the example figure labeled **Log-Probability of the data given model (marginal likelihood)**) You will need the numbers from the row labeled All, in the table there are three columns, report the values for the Bezier approximation. (This was our first model, we will compare the different models at the end of this exercise: my log marginal likelihood values for the Bezier approximated score are -4803.07, respectively (**see example figure on the right**)



- To save what we have done so far, copy the parmfile to parmfile.4param, copy outfile to outfile.4param, and copy outfile.pdf to outfile.4param.pdf. MIGRATE allows to specify filenames in the menu but copying in the terminal is fast and also allows us to use the **parmfile** as a template for the other models, so that we do not need to change the run-length and heating parameters again.

```
cp parmfile parmfile.4param
cp outfile.pdf outfile.4param.pdf
cp outfile outfile.4param
```

Report marginal likelihoods of Model 4!! (1 Parameter)

- We start now to work on those other models. We pick the easiest first: **model 4**.

- Start MIGRATE and choose the menu **Parameter settings**. Choose the entry about **sampling locations**. We want to use the data as if we would have sampled a single population, therefore we need to claim that the two locations Aadorf and Bern belong to the same panmictic population. MIGRATE's default is to assume that every location is a individual population. The dialog (**figure on the right gives an example of this interaction with the menu**) will ask first how many locations are in the dataset (for our example we have **2**). After that you will need to assign the locations to a population. For this model we need to assign each location to the same population. You need to enter **1 1** (one space one). With multiple populations more complicated settings are possible. Run MIGRATE, check the histogram, if it looks OK, come to the front and write down the log marginal likelihoods (again the row labeled All, Bezier and Harmonic score) into the spreadsheet under model 4. My run took 300 seconds and delivered the log marginal likelihood -4819.48.

```
Associate sampling locations with populations
---------------------------------------------
This menu allows to combine sample locations into populations
For example there are 4 locations: 1, 2, 3, 4
They can be combined into 2 populations
by mapping the 4 positions 1, 2, 3, 4 to 1, 1, 1, 2
Migrate will now combine the first three locations

Give (1) the number of populations <return> then (2) the mappi

How many localities are in the data set?
[Default: every sampling location is a population]
> 2
Enter now the remappings (little checking is done with this, e
 1 2
>1 1
```

- Copy the parmfile to parmfile.1param, copy outfile to outfile.1param, and copy outfile.pdf to outfile.1param.pdf

```
cp parmfile parmfile.1param
cp outfile.pdf outfile.1param.pdf
cp outfile outfile.1param
```

Report marginal likelihoods of Model 2 (3 Parameter)

- Now you need to consider the models with unidirectional migration.

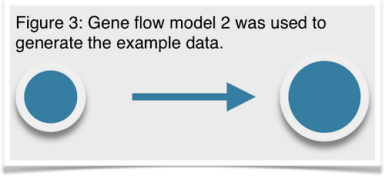


Figure 3: Gene flow model 2 was used to generate the example data.

```
cp parmfile.4param parmfile
```

- Start MIGRATE, choose the parameter menu. Choose the entry labeled **Model is set to**. MIGRATE will now show a dizzying list of options, don't panic, we will only use few of them. MIGRATE will ask you how many populations are used: enter 2. For a 2-population model we can have 4 parameters. Two population sizes and two migration rates. Before you enter values, please read this whole paragraph. A * or x means that that particular parameter will be unrestrictedly estimated, a zero (0) means that that particular parameter will not be estimated (is not used). Our goal is to set one of the migration/divergence parameters to zero. We start with model 2 (Figure 3). MIGRATE needs to know how to treat all connections between the populations. How migration rates or divergences are specified. We also need to give instructions how the program will treat the population sizes. Because we want to estimate both population sizes and one migration rate, we will use the * and a zero for the unused migration rate. The connection matrix is square so we can label it like it is shown in the first table below.

Table 1: Layout of connection matrix for our example for model 2.

	Aadorf	**Bern**
Aadorf	Population size	Migration to Aadorf
Bern	Migration to Bern	Population size

- MIGRATE asks now that you input each row, this can be done by either specifying *** 0** (see second table) and then return and then entering the next line * * return (second row in second table), or you can enter the whole matrix as *** 0 * ***.

Table 2: Syntax for model 2.

	Aadorf	**Bern**
Aadorf	*	0
Bern	*	*

Exit the parameter menu, write the parmfile, run MIGRATE, check the histogram, report the log marginal likelihoods. My run took 420 seconds, and delivered this log marginal likelihood: -4798.23. Cautionary note: if you use this tutorial for your own work, please recognize that a standard run in migrate can only use migration model that allow to draw a complete genealogy, so for example a model * 0 0 *, that has no migration among the population, does not work out of the box. As a rule of thumb each population must be connected to at least one other population.

- Copy the parmfile to parmfile.3aparam, copy outfile to outfile.3aparam, and copy outfile.pdf to outfile.3aparam.pdf

```
cp parmfile parmfile.3aparam
cp outfile.pdf outfile.3aparam.pdf
cp outfile outfile.3aparam
```

Report marginal likelihoods of Model 3 (3 Parameter)

- Run model 3 using the same procedure as for model 2. The string for the migration connection matrix is *** * 0 ***. Write parmfile, run, report. My run took 151 seconds and the log marginal likelihood was -4802.26 .
- Copy the parmfile to parmfile.3bparam, copy out file to outfile.3bparam, and copy outfile.pdf to outfile.3bparam.pdf
- Once about more than half of the class has reached this point we will talk about the marginal likelihoods found.

```
cp parmfile parmfile.3bparam
cp outfile.pdf outfile.3bparam.pdf
cp outfile outfile.3bparam
```

Report marginal likelihoods of Model 5 (3 Parameter)

- Run model 5 using the same procedure as for model 3. The string for the connection matrix is *** 0 d ***.

In contrast to the models before this model will introduce a divergence event, the **d** marks the populaiton that split of from the ancestor, in our example this means Bern split off from Aadorf. Write parmfile, run, report. Report the log marginal likelihood; I got -4795.23.

- Copy the parmfile to parmfile.3cparam, copy out file to outfile.3cparam, and copy outfile.pdf to outfile.3cparam.pdf
- Once about more than half of the class has reached this point we will talk about the marginal likelihoods found.

```
cp parmfile parmfile.3cparam
cp outfile.pdf outfile.3cparam.pdf
cp outfile outfile.3cparam
```

Compare models

- How to calculate Bayes factors? In the Table 3 I summarized all log marginal likelihoods, ln(mL), the Bayes factors are often calculated in very different ways. Here, I report the natural log Bayes factors where

$$LBF = 2\left(\ln mL(\text{model}_1) - \ln L(\text{model}_2)\right)$$

- Using the guidelines of Kass and Raftery (1995), values smaller than -2 suggest preference for 'model 2', values larger than 2 suggest preference for 'model 1'. We can use the log marginal likelihoods or the BF to order the models (see column Choice in the Table 3).

- We also can calculate the model probability. It is calculated by dividing each marginal likelihood by the sum of the marginal likelihoods of all used models:

$$\text{Prob}(\text{model}_i) = \frac{mL_{\text{model}_i}}{\sum_j^n mL_{\text{model}_j}}.$$

- Note that for the above formula uses the marginal likelihoods, *not* the *log* marginal likelihoods (which is what the program reports). The calculation of model probabilities from the reported log likelihoods is easy with computer programs that have variable precision (for example Maple or Mathematica). Calculations on a desk calculator often fail, for example the likelihood of model 1 is a remarkable small number because the likelihood is exp(-4803.07)= 1.130323625060 x 10^{-2086}, my emulated HP sci 15C calculator delivers 0.0000. But you can calculate the above quantities using this recipe: (1) find the largest log likelihood (-4795.23), (2) subtract that number form each log likelihood in the list (result: -2.27, 0.0, 2.5, -26.67), (3) exponentiate each element in the new list (result: 0.1033, 1.0, 0.0821, 2.6144 x 10^{-12}) , (4) sum all elements in the list up (0.1033+1.0+0.0821+2.6144 x 10^{-12}), this is the denominator (1.1854). (5) now divide each element in the list by that sum and the numbers will look

like the one in table 3 last column.

Table 3: Showing log marginal likelihoods for all models tested and model prorabilities

Model	Bezier lmL	Choice	Model probability
1: full (****)	-4803.07	3	0.0077
2: true (*0)**	-4798.23	1	0.9750
3: wrong (0*)**	-4802.26	2	0.0173
4: panmictic (*)	-4819.48	4	0.0000
5: divergence (*0d*)	-4909.51	5	0.0000

Looking at the model probabilities we can see that the “true” model has considerably higher support than the full model or the model that suggests a wrong direction of gene flow.

Summary of results

The best model (the one with the highest marginal likelihood) is model 2 (custom-migration={*0**}) if we use the the thermodynamic approximation of marginal likelihood. MIGRATE also reports the harmonic mean, but I suggest to ingore it and use thermodynamic integration (as we did) although it will be more costly to run. I was shown several times (Beerli and Palczewski 2010, Xie et al. 2011) now that the harmonic mean estimator is not a good estimator and may be misleading and prefer the more complex model (like in our example). The picture below is a sample from the class tutorial done on August 1st 2011.

		Migrate Bayes Factors		BEZIER THERMODYNAMIC		
		write down the marginal likelihood				
		Model 1	Model 2	Model 3	Model 4	Model 5
Model prob (max		3.8981E-35	1	8.9858E-37	1.13976E-39	3.5754E-48
Model probability		0.91258895	1	1.4343E-06	3.94584E-12	1.384E-22
#Parameters		4	3	3	1	1
Specification		****	*0**	**0*	*	*0d*
N		25	16	18	23	16
Maximum		-4795.57	-4716.34	-4799.34	-4806.01	-4825.59
Mean		-4804.686	-4793.57688	-4807.0317	-4819.83522	-4843.9088
Standard dev.		6.57213309	23.06859589	7.56231465	3.407326157	14.5768661
		-4801.06	-4799.67	-4802.43	-4819.06	-4853.55
		-4807.35	-4800.68	-4799.34	-4820.37	-4862.81
		-4804.26	-4805.54	-4822	-4822.43	-4829.72
		-4800.56		-4803.56	-4820.66	-4827.26
		-4803.59	-4799.84	-4803.63	-4820.34	-4832.23
		-4800.96	-4800.24	-4801.64		-4851.66
		-4806.61	-4799.89	-4803.44	-4822.59	-4854.82
		-4801.55	-4819.92	-4821.74	-4821.25	-4828.59
		-4804.81	-4798.57		-4818.57	-4854.83
		-4798.17	-4798.74	-4821.47	-4820.16	-4852.88
		-4802.7	-4716.34	-4804.75	-4815.38	-4825.85
		-4800.47		-4804.39	-4820.11	-4825.59
		-4806.94	-4799.78	-4814.31	-4819.76	-4861.92
		-4795.57		-4810.19	-4820.34	
		-4800.25		-4800.81	-4822.54	-4828.95
		-4818.57	-4793.22	-4805.04	-4819.14	
		-4803.54			-4822.24	
		-4801.78	-4798.78		-4820.07	
		-4806.29	-4799.96		-4820.38	
		-4803.45				
		-4801.24			-4821.14	
		-4821.53	-4800.18	-4803.03	-4806.01	-4858.09
		-4821.5		-4801.81	-4821.65	
		-4801.42			-4822.14	
		-4802.98	-4765.88	-4802.99	-4819.88	-4853.79

a second look at model comparison

there is a second dataset in the migrate tutorial, **twoswisstownsdiv**, here the sequence data was simulated using the divergence model, it is model 5 from the earlier excercise. Try tun compare the true divergence model (model 5, but change the infile option to the **twoswisstownsdiv**) and compare it to a run that uses model2.

```
cp parmfile.3cparam parmfile
# change the datafile to twoswisstownsdiv
#run migrate
cp outfile.pdf outfileB1.pdf
cp outfile outfileB1
cp parmfile.3aparam parmfile
```

```
# change the datafile to twoswisstownsdiv
#run migrate
cp outfile.pdf outfileB2.pdf
cp outfile outfileB2
#compare the two models
```

I added a little python script that calculates the model probabilities from te text outfiles, you can run it using:

```
module load bioware
grep " All   " outfileB* | sort -n -k 4,4 | migbf.py
```

If you want to see the results of the first model exercise do:

```
grep " All   " outfile.* | sort -n -k 4,4 | migbf.py
```

References

- Beerli, P. and M. Palczewski. 2010. Unified framework to evaluate panmixia and migration direction among multiple sampling locations. Genetics 185: 313–326.
- Kass, R. E. and A. E. Raftery. Bayes factors. 1995. Journal of the American Statistical Association 90(430): 773– 795.
- Xie, W., P. O. Lewis, Y. Fan, L. Kuo, and M.-H. Chen. 2011 Improving marginal likelihood estimation for Bayesian phylogenetic model selection, Systematic Biology, 60: 150–160.