Tutorial: Comparison of gene flow models using Bayes Factors

for MIGRATE 3.1.7 (http://popgen.sc.fsu.edu)

Peter Beerli, Department of Scientific Computing,
Florida State University, Tallahassee, FL
Version 1.0, June 2010.

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 suchmarginal 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.

- Decide on the models that are interesting for a comparison. The method does not work well for a fishing expedition here one would try to evaluate all models except for a small population 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 heated chains, use the # menu suggestion for best results, so that the temperatures are spaced so that the inverse of the temperature are regularly spaced on the interval 0 to 1.
- Compare the marginal likelihood of the different runs and calculate the Bayes factor and calculate the probability for each model.

The following tutorial details all steps using an 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 10x 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**. We will evaluate 4 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. 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 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.

1. Make sure that there is no file called **parmfile** in the directory you want to run our experiment.

2. Start the program MIGRATE-N (I will call it from now on simply MIGRATE). In the **Input/Output formats** menu change the **Datafile name** to **twoswisstowns**, Return to the main menu.

3. In the **Search strategy** menu change the **strategy** from the default (**likelihood**) to **Bayesian inference**. Change the **Number of recorded steps in chain** to **1000**. Do not worry about priors or other runtime options for the moment. Return to the main menu.

4. Save the changes by using the menu item **write a parmfile***.*

5. 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, if this takes more than 1 minute something is not set up correctly. On my computer this takes 5.2 seconds.

6. 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. The default acceptance method for parameters is Slice sampling (see **background information**) and their acceptance ratio is always 1.0. 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 and are either very close or exactly at the upper prior boundary, in these cases you need to extend the prior range. **See prior problems,** but for this dataset we will have no such problems.

7. Look at the outfile.pdf, you may need a PDF viewer like acroread or preview.app (on Macintosh computers use *open outfile.pdf*. In and the figures that depict the posterior distributions, you can see something like the histograms in Figure 1. Your own run may look differently.

8. 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. Don't forget to **write the parmfile** to save the settings. Run and compare the results (Figure 2) with the histogram from before.

9. With your own data you may want to do another round of refinements, but comparing the medians and means of the parameters in the table and 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. For this tutorial we now turn to the best estimation of the marginal likelihood. 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 my computer this takes about 5 minutes.

10. Come to the front and write down the log marginal likelihood into the spreadsheet. You will need the numbers from the row labeled **All**, in the table there are three columns, report the values for the Bezier approximation and the harmonic mean method. (This was our first model, we will

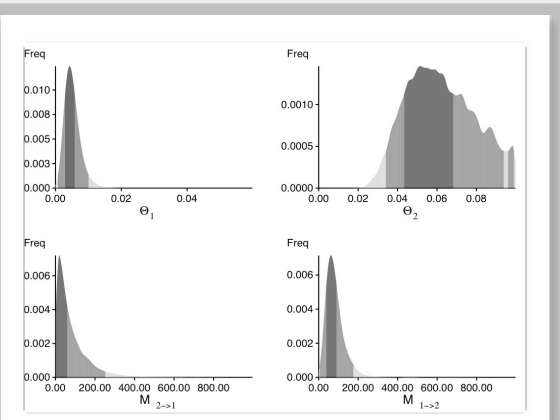


Figure 2: Example output from a short MIGRATE run, showing potential convergence for all parameters except for Θ_2 that may have some problems with a too narrow prior distribution.

compare the different models at the end of this exercise: my log marginal likelihood values for the Bezier approximated score and the Harmonic mean are -4862.85 and -4791.29, respectively.

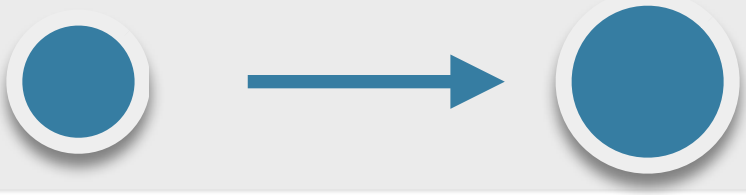
11. Copy the parmfile to parmfile.4param, copy outfile to outfile.4param, and copy outfile.pdf to outfile.4param.pdf

12. 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, this is done by telling MIGRATE that the dataset has two locations (2), and that they are in the same population by using " **1 1**". 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 183 seconds and delivered these log marginal likelihoods -4887.25 and -4803.22.

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

14. Copy the parmfile.4param to parmfile. We want to work now on the remaining two models. 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 estimated, a zero means that that particular parameter will not be estimated (is not used). Our goal is to set one of the migration parameters to zero. We start with model 2 (Figure 3). MIGRATE needs to know how to treat all connections between the populations are specified and that we also give instructions

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



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 show in the first table below. 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 * *. Exit the parameter menu, write the parmfile, run MIGRATE, check the histogram, report the log marginal likelihoods. My run took 156 seconds, and delivered these log marginal likelihoods: -4860.58, -4795.88.

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

Table 2: Syntax for model 2.

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

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

16. 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 likelihoods were -4863.08, and -4794.53.

17. Copy the parmfile to parmfile.3bparam, copy outfile to outfile.3bparam, and copy outfile.pdf to outfile.3bparam.pdf

18. Once about more than half of the class has reached this point we will talk about the marginal likelihoods found.

19. 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

$$LBF = \ln(mL(\text{model}_1)) - \ln(mL(\text{model}_2)).$$

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). The model probability is calculated by dividing each marginal likelihood by the sum of the marginal likelihoods of all used models (beware! Likelihood not log likelihood):

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

The harmonic mean fails to recover the true model for my example run and orders the models differently. The variance of the harmonic man estimator is large and therefore is not reliable, but this may depend on the dataset. 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.

Table 3: Bayes factors and log marginal likelihoods of the example.

Model	Bezier lmL	Harmonic lmL	LBF	Choice (Bezier)	Model probability
1: full	-4862.85	-4791.29	-2.27	2	0.087
2: true	-4860.58	-4795.88	0.0	1 (best)	0.844
3: wrong way	-4863.08	-4794.53	-2.5	3	0.069
4: panmictic	-4887.25	-4803.22	-26.77	4	0.000

References

Beerli, P. and M. Palczewski. Unified framework to evaluate panmixia and migration direction among multiple sampling locations. 2010. Genetics 185: 313–326.

Kass, R. E. and A. E. Raftery. Bayes factors. 1995. Journal of the American Statistical Association 90 (430): 773– 795.