TeTame 2.0

Estimation of neutral parameters by maximum likelihood

http://www.edb.ups-tlse.fr/equipe1/tetame.htm

Jérôme Chave
Franck Jabot

Laboratoire Evolution et Diversité Biologique UMR 5174
CNRS/UPS

Université Paul Sabatier, Toulouse, France

May 20th, 2008
Version 2.0

Introduction

Hubbell's neutral theory of biodiversity (Hubbell, 2001) proposes a simple explanation of the maintenance of biodiversity as a result of stochastic processes of birth, death, immigration and speciation. In this model, the species relative abundances in a guild is determined by two parameters: θ, that governs the appearance of new species in the regional species pool, and m that governs immigration into local communities of individuals from the regional species pool. This model is formally analogous to a continent-island model (Wright 1931).

This model is seen as a potentially useful null model in ecology. But for this assertion to hold, an efficient estimation of these two parameters is needed. In the case where m=1 (no dispersal limitation, that is, all newborn individuals are immigrants), the likelihood of the sole parameter θ can be computed using Ewens sampling formula (Ewens, 1972), given a species abundance dataset. However, in the general case where both θ and m can take non-trivial values, the previous methods of parameter estimation consisted in sequentially estimating θ and m, leading to approximated solutions (see e.g. Hubbell 2001, McGill 2003, Volkov et al. 2003, Latimer et al. 2005).

Parameter estimation can now be rigorously performed by maximum-likelihood using the sampling formula developed by Etienne (Etienne and Olff 2005, Etienne, 2005; Etienne and Alonso 2005; Etienne et al. 2006). Etienne (2005) provided a program for this estimation in the precompiled programming language "PARI" (http://pari.math.u-bordeaux.fr/).

We here provide **TeTame2**, an easy-to-use and freely available software written in C++. This software estimates the parameters θ and m using Etienne's method and produces a list of likelihood values in a parameter space that can be subsequently plotted, e.g. using the R software. This software has several advantages over the version previously published by Etienne (2005), being portable, computationally fast, and simple to use.

This new version additionally computes the immigration rates in several samples belonging to the same regional pool. The underlying assumption is that the species abundances pooled among these samples reflect regional species abundances. The theory underlying this computation can be found in Jabot et al. (2008).

Thanks to Andrea Manica (U. Cambridge), Rampal Etienne (U. Groningen) and David Alonso (U Groningen) for troubleshooting the program.

Hardware and program installation

TeTame2 for Windows has been compiled using the Linux-emulated environment cygwin on Windows and its C++ GNU compiler g++. So far, it has been tested on Windows XP Professional but is expected to run on most environments.

To install **TeTame2**, download it into the same directory as your file containing the species abundance data.

Running the program

1- Formatting your data

Case 1: you want to analyze the species abundance distribution of a **single** sample.
Save your data in a file with the extension ".txt". Your file should have the following structure:

abundance_species_1
abundance_species_2
...
abundance_species_n

Example: If your sample consists of (3 willows, 5 oaks, 1 eucalyptus), then your data entry file should read:
3
5
1

Case 2: you want to analyze the species abundance distribution of **multiple** samples.
Save your data in a file with the extension ".txt". Your file should have the following structure:

abundance_species_1_sample1
abundance_species_2_sample1
...
abundance_species_n_sample1
&
abundance_species_1_sample2
abundance_species_2_sample2
...
abundance_species_n_sample2
…
&
abundance_species_1_sampleD
abundance_species_2_sampleD
...
abundance_species_n_sampleD

Example: If your dataset consists of one sample of (3 willows, 5 oaks, 1 eucalyptus, 0 poplar), one sample of (1 willows, 15 oaks, 0 eucalyptus, 0 poplar), and one sample of (2 willows, 2 oaks, 0 eucalyptus, 3 poplars), then your data entry file should read:
3
5
1
0
&
1
15

0
0
&
2
2
0
3

NB: make sure you pressed enter at the end of the data file.

2- Running the program

Double-click on the executable file (tetame2.exe) and answer the questions in the console. Yes/no questions can be answered by using '0', 'n', or 'no' for no, and '1', 'y', or 'yes' for yes. You may want to test the program first with the dataset provided in the download webpage (test.txt). The typical result produced on our machine is provided in the appendix below.

Case 1: single sample analysis

Maximum-likelihood estimation

The values of the different parameters are output both in the console and in the file called [name_of_file]_out.txt created during the execution and sorted in your working directory. For the test dataset (test.txt), the output file name is thus "test_out.txt".
The output file of the 'test.txt' dataset looks like:

S	J	Theta	Std_Theta	I	Std_I	m	Std_m	loglike_min	Theta_Ewens
6	355	1.7179	0.891438	11.4569	15.2071	0.0313496	0.0403067	21.1446	0.907075
loglike_Ewens		Theta2	Std_Theta2		I2	Std_I2	m2	Std_m2	loglike_min2
22.6123		1.7179	0.891438		11.4569	15.2071	0.0313496	0.0403067	21.1446

S is the number of species in the sample, *J* is the total number of individuals in the sample, Theta is the maximum-likelihood estimate of θ, Std_Theta is the standard deviation of θ (under the assumption that the posterior is a Normal distribution) I is equal to m*(J-1)/(1-m) and it is a rescaled immigration rate, Std_I is the standard deviation of I, Ewens_Theta is the value of θ estimated from Ewens sampling formula (assuming that *m*=1), loglike_Ewens is the minimum of the opposite of the log-likelihood (assuming that *m*=1), loglike_min is the minimum of the opposite of the log-likelihood, and m is the immigration rate.

NEW FEATURE of **TeTame2**: if the programs detect a second likelihood maxima, it reports this second value too, and not just the global maximum as in the previous version of **TeTame**.

Likelihood surface plotting

It can be useful to visualize the shape of the likelihood function (see e.g. Etienne et al. 2006 for a worrisome example). Using **TeTame2**, you can output a list of log-likelihoods for values of (θ, *m*) on a grid. The likelihood surface can subsequently be plotted by using the freely available R software for example (and thus to have an idea of the uncertainty of the estimates). These options are provided to the user after the maximum-likelihood estimates of θ and m have been computed.

A user-supplied number of points on a grid are generated in the rectangular domain [thetamin, thetamax]*[mmin,mmax], where the four values are also user-supplied.

NB: if you have multiple samples, launch **TeTame2** independently for each sample, otherwise you will only be able to draw the likelihood surface of the last sample.

Case 2: multiple sample analysis with Jabot et al (2008)'s method

Maximum-likelihood estimation

The values of the different parameters are output both in the console and in the file called [name_of_file]_out.txt created during the execution and sorted in your working directory. For the test dataset (test.txt), the output file name is thus "test_outm.txt".
The output file of the 'test.txt' dataset looks like:

S	J	I	Std_I	m	Std_m	loglike_min
6	112	7.4197	15.9244	0.0626559	0.126048	15.9501
6	63	5.738	13.5052	0.0847088	0.182486	20.1273
6	181	14.6712	24.4701	0.0753638	0.116226	15.9371

S is the number of species in the sample, *J* is the total number of individuals in the sample, I is equal to m*(J-1)/(1-m) and it is a rescaled immigration rate, Std_I is the standard deviation of I, loglike_min is the minimum of the opposite of the log-likelihood, and m is the immigration rate.

Post-processing with the R software

We suggest you to use the freely available R statistical software to plot the likelihood surface (http://www.r-project.org/). In R, you should make sure that your working directory is the one where you saved your data file (go to menu 'File', then 'Change the directory'). The R commands for plotting the likelihood surface are provided in the output file "name_of_file_outR.txt". (in the example, this is the file test_outR.txt .

Troubleshooting

1. For problems with the R software, there is an online help on the R website http://www.r-project.org/
2. The software's optimization can be stuck in a local maximum. You can try to avoid this issue by supplying initial parameter values. You may also check that you reached a global maximum by using the plotting device.
3. Any feedback on the software is most welcome (chave@cict.fr).

History

Version 1.0: J Chave 17-03-2005
Version 1.01: J Chave and R Etienne 17-05-2005
Version 1.02: J Chave 15-11-2005
Version 1.1: J Chave and F Jabot 02-02-2006

Version 2.0: J Chave and F Jabot 20-05-2008

References

Etienne R.S., 2005. A new sampling formula for neutral biodiversity. Ecology Letters, **8**: 253-260.

Etienne and Alonso, 2005. A dispersal-limited sampling theory for species and alleles. Ecology Letters, **8** : 1147-1156

Etienne R.S. and Olff H., 2005. Confronting different models of community structure to species-abundance data : a Bayesian model comparison. Ecology Letters, **8**:493-504

Etienne R.S. et al, 2006. Comment on "Neutral Ecological Theory Reveals Isolation and Rapid Speciation in a Biodiversity Hot Spot". Science, **311**, 610b.

Ewens W.J.,1972. The sampling theory of selectively neutral alleles. Theor. Popul. Biol., **3**, 87–112.

Hubbell S.P., 2001. The Unified Neutral Theory of Biodiversity and Biogeography. Princeton University Press, Princeton, NJ.

Jabot F., Etienne R.S. and Chave J., 2008. Reconciling neutral community models and environmental filtering: theory and an empirical test. Oikos.

Latimer A.M., Silander J.A. Jr. and Cowling R.M., 2005. Neutral ecological theory reveals isolation and rapid speciation in a biodiversity hot spot. Science, **309**:1722-1725.

McGill B.J., 2003. A test of the unified neutral theory of biodiversity. Nature, **422**, 881-885.

Volkov I., Banavar J.R., Hubbell S.P. and Maritan A., 2003. Neutral theory and relative species abundance in ecology. Nature, **424**: 1035-1037.

Wright S., 1931. Evolution in Mendelian populations. Genetics, **16**: 97–159.

APPENDIX

Try the test dataset (test.txt) to make sure that **TeTame2** outputs correct results:

ESTIMATING NEUTRAL PARAMETERS BY MAXIMUM LIKELIHOOD
This program can be used for:
1-estimating theta and m of Hubbell's 2001 neutral theory using Etienne's 2005 m ethod.
2-estimating m in several samples belonging to the same regional pool using Jabo t et al. 2008 method.
For more details, see the manual.
Options for answering yes/no questions: 1, y, or yes for 'yes'; 0, n, or no for 'no'.

Please enter the data file name (without'.txt') test
Input file: test.txt
Reading the file stats ...
Number of samples: 3

TeTame detected several samples in your data file.
If you want to do the simultaneous estimation of Theta and m, using Etienne (200 5)'s likelihood formula, press e and enter
If you want to estimate m in each sample, assuming that they belong to the same metacommunity, using Jabot et al.(2008)'s method, press j and enter
In the case you want to do the multi-samples m inference, make sure you entered the pooled-over-the-samples species abundances in the beginning of the data file.

Case 1: you enter 'e'
e
In sample 1, number of species: 6

In sample 2, number of species: 6
In sample 3, number of species: 6
Start computing Stirling numbers ...
Start computing ln(K(D,A)) ...
Number of individuals: 112

Sample 1
Maximal abundance: 80
Compute the Ewens theta and log-likelihood ...
Ewens' -log-likelihood: 10.7797
Maximizing the likelihood ...
Would you like to provide initial values for the optimization procedure? 0

RESULTS (also output in file named: test_out.txt):
S J Theta Std_Theta I Std_I m Std_m loglike_min Theta_Ewens loglike_Ewens Theta2 Std_Theta2 I2 Std_I2 m2 Std_m2 loglike_min2
6 112 1.19516 0.583713 3.63849e+010 7.25904e+011 1 6.08639e-008 -68.1293 1.19529 10.7797 1.19536 0.583824 2.64204e+010 4.73882e+0111 7.53553e-008 -68.1293

Would you like to plot the likelihood surface? 0
Number of individuals: 63

Sample 2
…it goes on for the different samples.

Case 2: you enter ‘j’ (after the computation, the program will close automatically so you may not see the following, but you will find the results in the file “[name_of_datafile]_outm.txt”)
j
In the total dataset, there are 6 species
In sample 1, number of species: 6
In sample 2, number of species: 6
In sample 3, number of species: 6
Number of individuals in the pooled samples: 355
In sample 1:
Number of individuals: 112
Maximal abundance: 80
Number of species present in the sample: 6
Maximizing the likelihood ...

RESULTS (also output in file named: test_outm.txt):
S J I Std_I m Std_m loglike_min
6 112 7.49668 16.0164 0.0632649 0.126612 15.9234

In sample 2:
Number of individuals: 63
Maximal abundance: 20
Number of species present in the sample: 6
Maximizing the likelihood ...

RESULTS (also output in file named: test_outm.txt):

S	*J*	*I*	*Std_I*	*m*	*Std_m*	*loglike_min*
6	*63*	*5.70176*	*13.4441*	*0.0842187*	*0.181855*	*20.1619*

In sample 3:
Number of individuals: 181
Maximal abundance: 81
Number of species present in the sample: 6
Maximizing the likelihood ...

RESULTS (also output in file named: test_outm.txt):

S	*J*	*I*	*Std_I*	*m*	*Std_m*	*loglike_min*
6	*181*	*14.6548*	*24.4522*	*0.0752862*	*0.116161*	*15.9517*

Post-processing :

If you do the simultaneous estimation of θ and m (case 1), you can output the likelihood surface in R. TeTame2 is outputting a file named name_of_file_outR.txt. (in the example, this is the file test_outR.txt). This file contains the commands that you may use in the R software to produce graphs. You just need to paste and copy this file in the R software. (Make sure that you have changed the working directory in the R software before doing this)
More details can be found in the manual of the first version of **TeTame**.