TeTame

Estimation of **q** *and m by maximum likelihood in the neutral model with dispersal limitation*

Jérôme Chave
Franck Jabot

*Laboratoire Evolution et Diversité Biologique UMR 5174
CNRS/UPS*

Université Paul Sabatier, Toulouse, France

February 6th, 2006
Version 1.1

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 *q* can be computed using Ewens sampling formula (Ewens, 1972), given a species abundance dataset. However, in the general case where both *q* and m can take non-trivial values, the previous methods of parameter estimation consisted in sequentially estimating *q* 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/), available online in the *Ecology Letters* e-archive.

We here provide **TeTame**, an easy-to-use and freely available software written in C++. This software estimates the parameters *q* 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.

Thanks to Rampal Etienne (U. Groningen) and David Alonso (Ann Arbor, MI) for troubleshooting the program.

Hardware and program installation

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

Running the program

1- Formatting your data
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

Order is unimportant, as species are unlabelled, however please make sure that the number are **all strictly positive integers**. Please do not add a header to the file. If you want to run more than one dataset sequentially (up to 10,000), please contact us.

2- Running the program

Double-click on the executable file (tetame.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 (bci.txt). The typical result produced on our machine is provided in the appendix below.

3- 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 BCI dataset (bci.txt), the output file name is thus "bci_out.txt".
The output file of the 'bci.txt' dataset looks like:

S	J	Theta	Std_Theta	I	Std_I	m	Std_m	loglike_min	Theta_Ewens	loglike_Ewens
225	21457	47.6743	5.18986	2211.09	158.696	0.0934248	0.00607889	308.725	34.9623	318.849

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 q, 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, m and Std_m are the immigration rate and its standard deviation (respectively), loglike_min is the minimum of the opposite of the log-likelihood, Ewens_Theta is the value of q estimated from Ewens sampling formula (assuming that $m=1$), and loglike_Ewens is the minimum of the opposite of the log-likelihood (assuming that $m=1$),.

4- Likelihood surface plo tting

It can be useful to visualize the shape of the likelihood function (see e.g. Etienne et al. 2006 for a worrisome example). Using **TeTame**, you can output a list of log-likelihoods for values of (q, *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 q 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.

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, file 'bci_outR.txt').

Troubleshooting

1. So far, **TeTame** handles sizes up to J=117,000. Up to J=47,000 and abundance of the top species equal to ca. 20,000, the exact likelihoods are computed. Beyond these values, **TeTame** computes an approximated likelihood (details of our approximation upon request).
2. If the program crashes, make sure that the executable file and your data file are saved in the same directory.
3. For problems with the R software, there is an online help on the R website http://www.r-project.org/
4. 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.
5. 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

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.

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 BCI dataset (bci.txt) to make sure that **TeTame** outputs correct results. The outputs of the run are reproduced below (computing time on a Intel Pentium 4, 3.2 GHz: 10 seconds for the maximum-likelihood estimation and 17 seconds for the computation of 2,000 points):

ESTIMATING THE PARAMETERS FOR THE DISPERSAL-LIMITED NEUTRAL THEORY
This program estimates theta and m of Hubbell's 2001 neutral theory using Etienne's 2005 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') bci
Input file: bci.txt
Reading the file stats ...
Number of samples: 1
In sample 1, number of species: 225
Number of individuals: 21457

Sample 1
Maximal abundance: 1717
Start computing Stirling numbers ...
Start computing ln(K(D,A)) ...
Compute the Ewens theta and log-likelihood ...
Ewens' -log-likelihood: 318.849
Maximizing the likelihood ...
Would you like to provide initial values for the optimization procedure? n
293 Function evaluations
164 Optimizations

RESULTS (also output in file named: bci_out.txt):
S   J   Theta Std_Theta   I   Std_I m   Std_m loglike_min   Theta_Ewens
loglike_Ewens
225   21457 47.6743 5.18986 2211.09 158.696 0.0934248   0.00607889   308.725
34.9623 318.849

Would you like to plot the likelihood surface? y
How many grid points would you like to have? 2000
Would you like to use a log scale for m? n
Value of theta_min (must be a positive number): 10
Value of theta_max (must be a positive number greater than 10): 100
*Value of m_min (must be a number between 0 and 1): 0.01*
Value of m_max (must be a number between 0.01 and 1): 0.4
Computing the likelihood of the grid points ...

Post-processing :

TeTame outputs a file named name_of_file_outR.txt (in the example, this is the file bci_outR.txt). This file contains the commands that you may use in the R software to produce graphs. You just need to copy and paste this file in the R software. Make sure that you have changed the working directory in the R software before doing this.

Plotted figure in the example :

bci
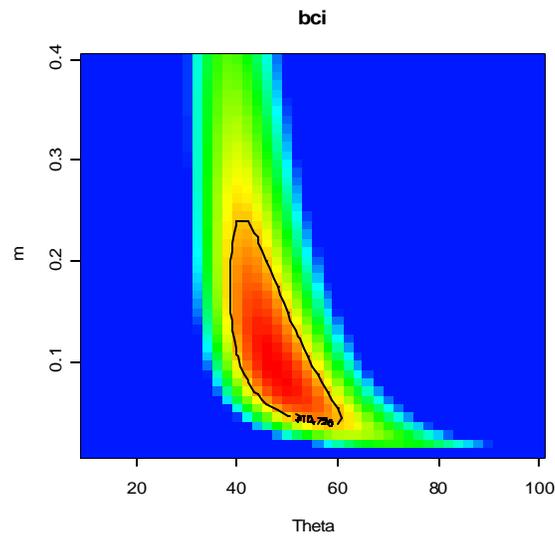
NB : The list of points is saved in two output files :
1. in name_of_file_outg.txt (in the example, this is the file bci_outg.txt), one can find the matrix of likelihoods for all the combinations of parameters Theta and m on the grid.
2. in name_of_file_out2.txt, you will find the list of points used to construct the surface (theta, m, -loglikelihood). This file has the following structure:

theta	m	llik
27.7675	0.632766	258.193
84.0761	0.389624	364.589
...	...	...

You can improve the quality of the graph by computing a larger number of points. Below, you can see the same graph for a number of points equal to 40000 (computing time on a Intel Pentium 4, 3.2 GHz: 5 minutes)

bci