

Discrete Markov Chain Monte Carlo

Prerequisite: Linear algebra

TuTh

Fall 2007

George W. Cobb

Although discrete Markov Chain Monte Carlo is an area within mathematics, it has a number of applications, and so, to be concrete, I begin this description with an example of an application.

Applied problem: Galapagos Finches. Community ecologists study data like the matrix two pages ahead, which shows the distribution of 13 species of finches among the 17 islands of the Galapagos chain. For simplicity's sake, consider here just a tiny version of the system, with only 3 species and 3 islands, distributed as follows:

Species	Island			Total
	1	2	3	
A	1	0	1	2
B	0	1	0	1
C	1	0	1	2
Total	2	1	2	5

Although I've used an ecology context, with species as rows and islands as columns, there are many other applications outside of ecology. In educational testing, for example, rows could be test-takers, and columns could be multiple choice items, with a 1 to indicate a correct answer. In computer vision, the array of 1s and 0s might correspond to pixels.

Returning to the islands and species, notice that Species A and B never appear together on the same island: they form what island bio-geographers call a "checkerboard," a species pair with no islands in common. Counting the number of checkerboards in a species-by-island matrix gives one way to measure the degree of competition between species. In the example, there are three possible species pairs, forming a total of 2 checkerboards (pairs A,B and B,C). In the finch data of the appendix, there are 78 species pairs, and 10 checkerboards.

How surprised should an ecologist be to find 10 checkerboards among 78 species pairs? Is 10 large enough to suggest that some biological cause is at work, or is 10 the sort of value you could easily get if the finch species has distributed themselves at random?

A mathematical question. Various versions of this question have been the subject of research by ecologists, statisticians, and mathematicians for more than 20 years. For the simple version with 3 species and 3 islands, there is a simple solution. If you write out all the tables of 0s and 1s that have the same row and column sums as the table above, you find there are only 5 of them. Of these 5, only the one shown above has 2 checkerboards; the other four tables have only 1 each. So if the species had distributed themselves at

random, all five tables would be equally likely, and the chance of 2 or more checkerboards would be $1/5$.

For any but the smallest tables, the questions turn out to be surprisingly difficult. For example, no one knows how many different versions of the finch data table there are with the same row and column sums, although researchers at Harvard and Stanford have developed ingenious methods to estimate the number, and obtained an estimate of 6.715×10^{17} tables. For other applications (educational testing or computer vision, for example), the data tables would have many more than 13 rows, many more than 17 columns, and the total number of tables would be astronomical.

With such a gigantic number of tables, trying to list them all is hopeless. Instead, statisticians generate random matrices through a long series of small steps. For example, in the table above, there are four different sub-tables of the form

$$\begin{array}{cc} 1 & 0 \\ 0 & 1 \end{array} \quad \text{or} \quad \begin{array}{cc} 0 & 1 \\ 1 & 0 \end{array}$$

If you pick one of these four sub-tables at random, and swap the 0s and 1s, you create a new sub-table, and thus a new 3×3 data table, with the same row and column sums as the original. Similarly, with the finch data, you can create new tables by picking a pair of rows and a pair of columns at random to get a 2×2 sub-table; if your sub-table has one of the two diagonal forms above, you swap the 1s and 0s.

This method is an instance of Markov Chain Monte Carlo, and extremely general and powerful method of computer simulation that has become one of the hottest research topics in statistics in the past decade or so. The sequence of random swaps defines a random path through the collection of all tables with the given row and column sums. Random paths of this sort have analogues in physics (Brownian motion) and molecular biology (3-dimensional structure of proteins). Trying to understand the mathematics of this method for generating random paths raises lots of questions: Does the method of random swaps give the right probability distribution? (No.) Can it be altered so that it does? (Yes.) How efficient is the method, i.e., how many random swaps do you need? (An area of ongoing research.) Can it be made more efficient, i.e., is there a way to get equally good results with far fewer swaps? (An area of ongoing research.)

This area of study offers many advantages as a topic for undergraduate students of mathematics. It is a current hot topic under active study by statisticians and mathematicians, with many open questions. Some problems are simple enough that you can understand them and stand a good chance of making progress toward an answer without having taken a lot of advanced mathematics courses. At the same time, the study of these questions brings together very different areas of mathematics, making connections that can be quite surprising and delightful. Among the areas of mathematics are probability, graph theory, linear algebra, and abstract algebra.

Distribution of 13 Finch Species among 17 islands of the Galapagos

Species	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	Total
A	0	0	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	14
B	1	1	1	1	1	1	1	1	1	1	0	1	0	1	1	0	0	13
C	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	0	14
D	0	0	1	1	1	0	0	1	0	1	0	1	1	0	1	1	1	10
E	1	1	1	0	1	1	1	1	1	1	0	1	0	1	1	0	0	12
F	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	2
G	0	0	1	1	1	1	1	1	1	0	0	1	0	1	1	0	0	10
H	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1
I	0	0	1	1	1	1	1	1	1	1	0	1	0	0	1	0	0	10
J	0	0	1	1	1	1	1	1	1	1	0	1	0	1	1	0	0	11
K	0	0	1	1	1	0	1	1	0	1	0	0	0	0	0	0	0	6
L	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	2
M	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	17
Total	4	4	11	10	10	8	9	10	8	9	3	10	4	7	9	3	3	122

The same data, with rows and columns in descending order of row and column totals

Species	Sites																Total	
	3	4	5	8	12	7	10	15	6	9	14	1	2	13	11	16		17
M	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	17
A	1	1	1	1	1	1	1	1	1	1	1	0	0	1	0	1	1	14
C	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	0	14
B	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	13
E	1	0	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	12
J	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	11
D	1	1	1	1	1	0	1	1	0	0	0	0	0	1	0	1	1	10
G	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	10
I	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	10
K	1	1	1	1	0	1	1	0	0	0	0	0	0	0	0	0	0	6
F	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	2
L	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
H	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1
Total	11	10	10	10	10	9	9	9	8	8	7	4	4	4	3	3	3	122

Density of 1s = 0.552

Checkerboards = 10 # species pairs = 78 Fraction = 0.128

A "checkerboard" is a species pair with no island inhabited by both.

2x2 sub-matrices = 10608 # CUs = 333 Fraction = 0.031

A "checkerboard unit" (CU) is a combination of species pair and island pair having the first species but not the second on one island, and the second species but not the first on the other island.

Data from <http://www.amsci.org/articles/00articles/sandersoncap3.html>