Estimating the size of the blob

Richard D. Gill

Bachelor seminar project 2015

You are a zoologist on an expedition in the Amazon jungle, searching for new species of animals. From time to time you see animals of previously unknown species. For instance, you might see altogether $N = 7$ animals of previously unknown species, of which 3 all belong to one species, 2 to another, and the other 2 each to two more different species. The $7 = 3+2+1+1$ animals observed belonged to just $L = 4$ different new species.

Think of these animal species as a random sample of size N from a set $\mathcal{A}$ of previously unknown species α with probabilities p_α. For each pair of animals in the sample we can see if they belong to the same or to different species, but that's all. Let n_i denote the total number of times the i'th most common species in the sample was observed. The observed data can be summarised by the sequence of numbers $n_1 \geq n_2 \geq \cdots \geq n_L > 0$ where L is the number of different species observed in the sample and $N = \sum_{i=1}^{L} n_i$.

You want to estimate the vector of unknown probabilities $(p_\alpha)_{\alpha\in\mathcal{A}}$. The set $\mathcal{A}$ is unknown, but you can imagine all unknown species as ordered by probability, from large to small. So you can take $\mathcal{A}$ to be the set of nonnegative integers, and $p_1 \geq p_2 \geq p_3 \geq \ldots$.

A naive estimator of $\theta = (p_\alpha)_{\alpha=1,2,\ldots}$ is the vector F of ordered observed relative frequencies $f_i = n_i/N$, $i = 1, 2, \ldots$, where for $i > L$ we set $f_i = n_i = 0$. The estimator "guesses" that the i'th most frequent species in the sample is the i'th most frequent species in the population. And it guesses that there are no unobserved species left out there in the wild. But there is no reason at all for this to be true!

In the project I want to investigate a little known alternative estimator of θ called the *hi-profile estimator* by computer-scientist Alain Orlitsky. Basically, it is just the maximum likelihood estimator of θ after adjusting the model in the following way: we pretend that in the population there are at most a finite number K of species of positive probability, together with uncountably many species each of probability 0 but making up together a "blob" of zero-probability species of total probability $p_0 = 1 - \sum_{\alpha=1}^{K} p_\alpha$. I propose to compute the estimate by something called "stochastic approximation Metropolis-Hastings EM". The idea is that behind the observed data there corresponds an unobserved mapping χ from population species to sample species. The likelihood is the sum over all possible mappings of the probability of the data under each mapping. The algorithm performs a random walk on the set of all mappings χ, building up information about the relative probabilities of different mappings, and continuously improving the estimate of θ.

The aim of the project is to study different possible ways to define this random walk, with a view to finding a reliable way to compute the hi-profile estimator in challenging real world applications from various fields.