

ESTIMATING VOCAL REPERTOIRES FROM LIMITED SAMPLES

RICHARD MCELREATH

May 25, 2023

1. PURPOSE

Suppose a population of N individuals (monkeys, people) who produce among them M different vocalizations. Each individual has potentially a different repertoire of these M vocalizations. We want to characterize repertoire at the individual level, so that we might study its relationship with development for example.

One problem is the undercounting of repertoire which is expected in any finite sample. If some vocalizations are rare, and recordings are short or small in number, then on average we will undercount true repertoire size.

2. MODELING STRATEGY

We seek a vector ϕ of posterior probability distributions of length M for each individual i . In the case that i was observed to produce a m , then ϕ_m is 1. But when m was not observed, we can compute the posterior probability that this was due to the finite sample, an undercount. In the simplest model, assume that each vocalization m is produced at a rate λ_m , when an individual possesses it. This produces a Poisson distribution of the counts $Y_{i,j,m}$ of each possessed vocalization m in a focal recording j of duration d_j .

$$Y_{i,j,m} \sim \text{Poisson}(\lambda_m d_j x_{i,m})$$

where $x_{i,m}$ is an indicator variable for whether individual i possesses vocalization m . When the observed $y = Y_{i,j,m} > 0$, then the probability is:

$$\Pr(y|y > 0) = p_m \frac{(\lambda_m d_j)^y \exp(-\lambda_m d_j)}{y!}$$

where p_m is the probability an individual possesses m . When $y = 0$ is observed, we require a mixture:

$$\Pr(y|y = 0) = p_m \exp(-\lambda_m d_j) + (1 - p_m)(1)$$

DEPARTMENT OF HUMAN BEHAVIOR, ECOLOGY AND CULTURE, MAX PLANCK INSTITUTE FOR EVOLUTIONARY ANTHROPOLOGY, DEUTSCHER PLATZ 6, 04103 LEIPZIG, GERMANY

E-mail address: richard_mcelreath@eva.mpg.de.

Date: May 25, 2023.

Let $X_{i,m}$ be the total number of times i was observed to use m and D_i be the total sampling effort (duration of recordings). If $X_{i,m} > 0$, then i possesses the vocalization. But what about when $X_{i,m} = 0$? In that case, the posterior probability that individual i possesses vocalization m follows from Bayes' rule:

$$\begin{aligned} \Pr(x_{i,m} = 1 | X_{i,m} = 0, D_i) &= \frac{\Pr(0 | x_{i,m} = 1) \Pr(x_{i,m} = 1)}{\Pr(0)} \\ &= \frac{\Pr(0 | \lambda_m D_i) p_m}{\Pr(0 | \lambda_m D_i) p_m + (1 - p_m)} \end{aligned}$$

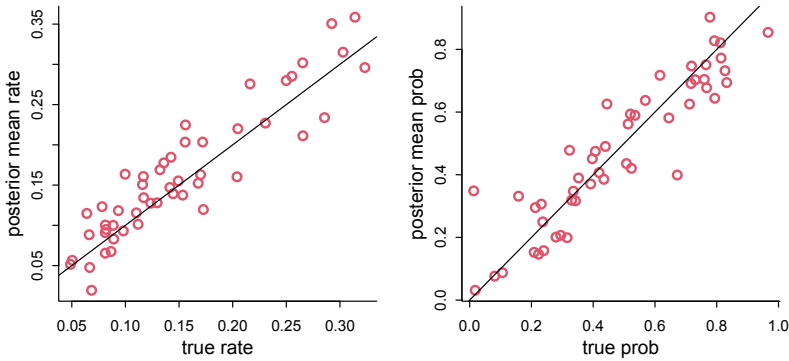
25 And this is sufficient to solve the problem.

In the Poisson case, we need only the total duration of the recording. This model can be modified to allow for Markov transitions between vocalizations, but then we require time stamps for each vocalization.

IMPLEMENTATION

30 We developed a generative model of the sampling design, in which each individual has a variable number of recordings of variable length. Then we used Stan to sample from the posterior distribution of the inferred individual repertoires.

First we can ask how well the model recovers the rate λ and probability p variables.

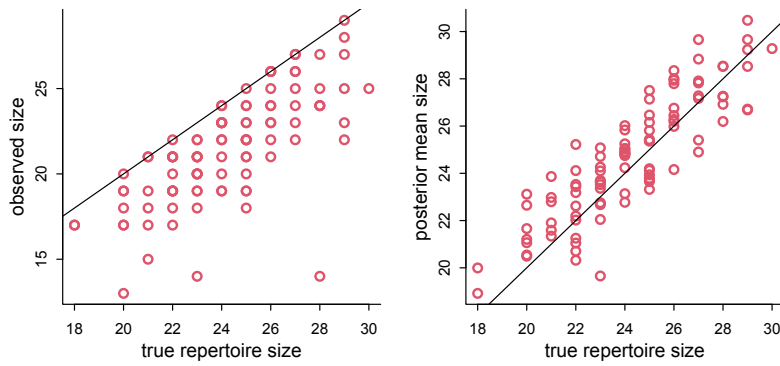


35

The left compares the true λ values (horizontal) to the posterior means. The diagonal line shows identity. The right compares the true p parameters (horizontal) to the posterior means. This is only one example, but the model unsurprisingly has no trouble estimating Poisson rates and the base rates of each vocalization, except in cases in which a vocalization is never performed in a particular sample.

40

Now we can ask how the individual repertoires are captured.



45 The left compares the true repertoire size of each individual (horizontal) to the raw observed (vertical). Note the undercounting. The right compares the same true repertoires (horizontal) to the posterior mean sizes. These are still uncertain, but less biased.