BISP8

Eighth Workshop on
BAYESIAN INFERENCE IN STOCHASTIC PROCESSES

Bayesian nonparametric estimation of discovery probabilities

Stefano Favaro[1] Antonio Lijoiy[1] and Igor Prünster[1]

[1]University of Torino, Italy

Species sampling problems have a long history in ecological and biological studies and a number of statistical issues, including the evaluation of species richness, are to be addressed.
Such inferential problems have recently emerged also in genomic applications where one has to deal with very large genomic libraries containing a huge number of distinct genes and only a small portion of the library has been sequenced. These aspects motivate the Bayesian nonparametric approach we undertake, since it allows to achieve the degree of flexibility typically needed in this framework. Basing on an initial observed sample of size n, focus will be on prediction of a key aspect of the outcome from an additional sample of size m, namely the so-called discovery probability. In particular, conditionally on the observed initial sample, we derive a novel estimator of the probability of detecting, at the (n+m+1)-th observation, species that have been observed with any given frequency in the enlarged sample of size n+m. The result we obtain allows us to quantify both the rate at which rare species are detected and the achieved sample coverage of abundant species, as m increases. Natural applications are represented by the estimation of the probability of discovering rare genes within genomic libraries and the results are illustrated by means of two Expressed Sequence Tags (EST) datasets.

Keywords:
Bayesian Nonparametrics; Gibbs-type priors; Ewens-Pitman sampling formula;
Rare species discovery; Species sampling models;
Two parameter Poisson-Dirichlet process

ABSTRACT BISP8.30
TYPE Contributed poster