Bayesian Nonparametric priors for genomic variant discoveries

A seminar by Federico Camerlenghi

University of Milano-Bicocca

Friday 18 Nov 2022 | 12:30 p.m.
Room Benvenuti and live Zoom
Department of Statistical Sciences

The estimation of the number of unseen features is an important problem in biological sciences, for example to predict the number of 7 5hitherto unseen genomic variants. We investigate two classes of Bayesian nonparametric priors for the unseen features problem, and we shed light on their behaviour in terms of predictive inference.

We first focus on the popular class of completely random measures (CRMs), which include the three-parameter Beta process, and we show how, for fixed prior's parameters, CRMs all lead to Poisson posterior distribution for the number of unseen features, which depends on the sampling information only through the sample size. CRMs are thus not flexible enough to face the unseen features problem. Therefore, we introduce the Stable-Beta Scaled Process (SB-SP) prior, and we show that it allows to enrich the posterior distribution of the number of unseen features arising under CRM priors, while maintaining its analytical tractability and interpretability. That is, the SB-SP prior leads to a negative Binomial posterior distribution for the unseen features, which depends on the sampling information through the sample size and the number of distinct features.

The proposed approach turns out to be simple and computationally efficient. We apply our BNP proposal to synthetic data and to real cancer genomic data, showing that: i) it outperforms the most popular parametric and nonparametric competitors in terms of estimation accuracy; ii) it provides improved coverage for the estimation with respect to a BNP approach under CRM priors.

800 1222·2022 ANNI

Università degli Studi di Padova

Dipartimento di Scienze Statistiche