

International Society for Bayesian Analysis, 9th World Meeting,
Hamilton Island, Australia, 2008.

A BAYESIAN NONPARAMETRIC APPROACH FOR ANALYSING AND TESTING CLUSTERING STRUCTURES

Antonio Lijoi¹, Ramsés H. Mena², Igor Prünster^{3*} and Stephen G. Walker⁴

¹ University of Pavia, Pavia and CNR–IMATI, Milan, Italy

² IIMAS–UNAM, Mexico City, Mexico

³ University of Turin and Collegio Carlo Alberto, Turin, Italy

⁴ University of Kent, Canterbury, UK

* igor@econ.unito.it

Many applications require a deep understanding of the clustering mechanism that generates the observed data. The two parameter Poisson–Dirichlet process and more general Gibbs–type priors are natural candidates for modelling data arising from discrete distributions. Here we make use of such priors and analyze their posterior behaviour in some detail. In particular, we propose methods for prediction and testing in order to assess the clustering structure featured by the data. The methodology is then applied to Expressed Sequence Tags (ESTs) data in genomics. Indeed, when studying EST data one is typically interested in evaluating the redundancy of the corresponding cDNA library and in comparing different libraries on the basis of their ability to generate new distinct genes. Our proposal has appealing properties over frequentist nonparametric methods, which become unstable when prediction is required for large future samples.