Half-Day Tutorial: A tutorial on Dirichlet Process mixture modeling

Instructor: Yuelin Li, PhD. Memorial Sloan Kettering Cancer Center

Bayesian nonparametric (BNP) models are becoming increasingly important in cognitive psychology, both as theoretical models of cognition and as analytic tools. However, existing expositions tend to be at a level of abstraction largely impenetrable by non-technicians. This tutorial aims to explain BNP to the curious non-technicians using the Dirichlet Process (DP) as an illustrative example. DP is one of the most widely used BNP methods. A student researching these topics may encounter terms such as DP and the Chinese Restaurant Process (CRP, one of the construction methods of DP), but he or she may only have a vague impression as to the origin of these somewhat abstract concepts. This tutorial aims to make these concepts more concrete, explicit, and transparent.

This tutorial will: 1) show you what the DP and CRP look like; 2) explain the essential mathematical derivations often omitted in existing expositions you find online; and 3) demonstrate how to write a simple program in the statistical language R to fit a DP mixture model (DPMM). The R program will be explained line by line so that you know precisely how the computation algorithm works. The mathematics will be no more than basic conditional probability and sampling from standard probability distributions. The overall goals are to help you understand more fully the theory and application so that you may apply DP in your own work and leverage the technical details in this tutorial to develop novel methods. By working through the R program and simulated data, you will learn the key feature of DP. The number of clusters is not required to be fixed in advance. The number of clusters used by a DP cognitive theory grows as data accrue and tops when additional clusters no longer explains the data. This tutorial should enhance your appreciation of other tutorials of the DP (e.g., Gershman & Blei, 2012, J Math Psych; Austerweil, Gershman, Tenenbaum, and Thomas L. Griffiths, 2015, In Busemeyer et al., Oxford Handbook of Computational and Mathematical Psychology).

Prerequisite knowledge:

1. Basic familiarity with R (e.g., comfortable with logistic regression in R).
2. Experience with R programming also helps (unfortunately, DP is not yet supported by statistical packages frequently used by behavioral scientists, such as SPSS, Mplus, Stata or SAS). But the programming skills required are no more complicated than writing simple functions.
3. Consider bringing a laptop with R already installed so that you can run the R program right away.

Yuelin Li is an Associate Attending Behavioral Scientist at Memorial Sloan Kettering Cancer Center in New York City, where he has been since 2005. He received his Ph.D. in Cognitive Psychology from Columbia University in 1999 working with David Krantz. He wrote a textbook with Jonathan Baron on Behavioral Research Data Analysis with R (2012, Springer). He directs the Behavioral Statistics Laboratory at MSKCC, where he applies Bayesian methods in explaining patient-reported outcomes in cancer (e.g., Li & Baser, 2012; Li, Root, Atkinson, and Ahles, 2016; Li, Lord-Bessen, Shiyko, Loeb, 2018). His recent research interests are in Bayesian nonparametric methods (forthcoming tutorial in J Math Psych) and text analysis by Latent Dirichlet Allocation (Li, Rapkin, Schofield, Atkinson, Bochner, in press).