Short Course:
"Applied Bayesian Nonparametrics"

ALEJANDRO JARA

Department of Statistics
Faculty of Mathematics
Pontificia Universidad Católica de Chile

COURSE CONTENT

Bayesian nonparametric statistics is a relative new area of statistics. The intersection between Bayesian and non-parametric statistics was almost empty until the late sixties where the first advances were made, primarily on the mathematical formulations. It was only in the early nineties with the advent of sampling based methods, in particular Markov chain Monte Carlo methods, that substantial progress has been made in the area. Posterior distributions ranging over functional spaces are highly complex and hence sampling methods play a key role.

This course is designed to provide basic coverage of Bayesian nonparametrics and includes advanced use of the R package DPpackage. The course will include theoretical input, but also practical elements and participants will be involved hands-on in the use of DPpackage. Emphasis on the course is placed on the probability models for probability distributions and their practical applications for model building. Some knowledge on Bayesian computation is assumed.

The course consists of the following sessions:

SESSION 1: BACKGROUND MATERIAL

Description:

In this session the statistical problem is formalized and the concept of statistical model is discussed. The main differences between the classical and Bayesian models are discussed. The technical definitions of parametric, semi- and non-parametric models are presented. Several illustrative examples will be discussed in order to emphasize on the main differences between the models. These considerations will be developed in the following subsections:

1.1 Statistical Models: The specification problem according to Fisher (1922) and Koopmans and Reiersøl (1950).
1.2 Bayesian vs classical statistical models.
1.3 Bayesian parametric, semi- and non-parametric statistical models.

1.4 Examples of Bayesian semi- and non-parametric statistical models.

SESSION 2: CONSTRUCTING PROBABILITY MODELS FOR PROBABILITY DISTRIBUTIONS

Description:
In this session the construction of random probability measures (RPM) is revised from an historical point of view. Different strategies for constructing RPM are presented and illustrated. Notice that a RPM construction is a technique for specifying a probability measure (prior) on the space of probability measures. An advantage of the approaches discussed here over parameter-based priors is that the former have large support on the space of probability measures. Hence the general feeling that inferences based upon nonparametric priors may avoid problems associated to the miss-specification of a particular parametric family. The material of this session will be developed in the following sections:

2.1 Constructing random probability measures.
2.2 Common features and motivations.
2.3 Dubins and Freedman random distribution functions.
2.4 Random re-scaling random probability measures.
2.5 Sequential barycenter array constructions.
2.6 Polya urn constructions.
2.7 Polya and exchangeable tree constructions.

SESSION 3: USEFUL PROPERTIES AND COMPUTATIONAL ISSUES FOR DIRICHLET PROCESS

Description:
In this session we provide an overview of the theoretical properties and the applied associated topics for Dirichlet process-based priors. The Dirichlet process was introduced as a prior probability model for random probability measures by Ferguson (1973), and has become the most widely used Bayesian nonparametric model (Mueller and Quintana, 2004). The intention here is not to be complete or exhaustive, but rather to touch on areas of interest for the most commonly used models. Specifically, we try to give a complete reference covering the theory and required algorithms. The material will be developed in the following sections:

3.1 Dirichlet processes.
3.2 Processes derived from Dirichlet processes.
3.3 Generalizations of the Dirichlet processes.
3.4 Posterior sampling for Dirichlet processes.

SESSION 4: USEFUL PROPERTIES AND COMPUTATIONAL ISSUES FOR POLYA TREES

Description:
In this session we provide an overview of the theoretical properties and the applied associated topics for Polya tree priors. They can be viewed as generalizations of the Dirichlet processes (Ferguson 1974). However, they can be specified in such a way that they admit continuous distributions directly. This material will be developed in the following sections:

4.1 Polya tree processes.
4.2 Process derived from the Polya tree processes. Uni- and Multi-variate extensions.
4.3 Posterior sampling for Polya tree processes

SESSION 5: ILLUSTRATIVE EXAMPLES, PART I

Description: While Bayesian nonparametric methods are extremely powerful and have a wide range of applicability within several prominent domains of statistics, they are not as widely used as one might hope. At least part of the reason for this is the gap between the type of software that many applied users would like to have for fitting models and the software that is currently available. The most popular programs for Bayesian analysis are generally unable to cope with nonparametric models.
In this session we introduce an R package, DPpackage, designed to help bridge the previously mentioned gap. Several illustrative examples will be analyzed using DPpackage.
This material will be developed in the following sections:

5.1 The general syntax and design philosophy of the package.
5.2 Uni- and multi-variate density estimation.
5.3 Survival analysis.
5.4 Generalized linear mixed models.

SESSION 6: DP-BASED MODELS FOR RELATED PROBABILITY DISTRIBUTIONS

Description: In recent years there has been a surge of interest in models for dependent (in predictors) random probability models. In this session we present dependent models based on extensions of the Dirichlet process prior. The dependent DP proposed by MacEachern (1999) will be introduced. Several special cases where the the point masses are stochastic processes indexed by predictors (e.g. Gelfand et al. 2005) or simple regressions functions (e.g. De Iorio et al. 2004, De Iorio et al. 2009) will be discussed. The approach proposed by Mueller et al. (1996), originally proposed for regression models, will be also presented as an alternative method. The material will be developed in the following sections:

6.1 The Dirichlet process mixture model for density regression.

6.2 The dependent Dirichlet processes.
6.3 Linear dependent Dirichlet processes.
6.4 The Hierarchical Dirichlet process.

SESSION 7: PT-BASED MODELS FOR RELATED PROBABILITY DISTRIBUTIONS

Description:
In this Session we present dependent models based on extensions of Polya tree priors. We introduce a novel class of dependent processes in which tailfree parameters follow the logistic-normal model. Density shape is regressed on one or more predictors through the tailfree conditional probabilities. Prior specification is developed in detail. The resulting process is remarkably flexible and easy to fit using standard algorithms for generalized linear models. The general method is developed and applied to growth curve analyses, evolving random effects distributions in generalized linear mixed models, and median survival modeling with censored data and covariate-dependent errors. The material will be developed in the following sections:

7.1 Dependent Tailfree models.
7.2 The linear dependent Tailfree process.
7.3 Dependent predictive updating model.

SESSION 8: ILLUSTRATIVE EXAMPLES, PART II

Description:
In this Session several R functions for modeling related probability distributions are illustrated. This material will be developed in the following sections:

8.1 Generalized linear mixed models.
8.2 Median regression models.
8.3 Survival models.
8.4 Growth curve modeling.

THE SPEAKER

Alejandro Jara is Assistant Professor of Statistics, Department of Statistics, Faculty of Mathematics, Pontificia Universidad Católica de Chile, Chile.

WHO SHOULD ATTEND The course is intended for statisticians. The prerequisites are a background in probability theory, mathematical statistics, and Bayesian statistics. In addition, the course may be of interest for epidemiologists and public health workers with a strong interest in data analysis.

WHY ATTEND Participants will gain an in-depth understanding of the basic theoretical issues, methods and techniques used in the Bayesian nonparametric model building.