

# Non-Parametric Bayesian Method on Semi Supervised Problems

Khadim SENE (khadim.sene@aims-senegal.org)  
African Institute for Mathematical Sciences (AIMS), Senegal

March 14, 2020

*Submitted in partial fulfillment of the requirements for the award of Master of Science in Mathematical Sciences at AIMS  
Senegal*



# AIMS


**African Institute for  
Mathematical Sciences  
SENEGAL**

# DECLARATION

This work was carried out at AIMS Senegal in partial fulfillment of the requirements for a Master of Science Degree.

I hereby declare that except where due acknowledgment is made, this work has never been presented wholly or in part for the award of a degree at AIMS Senegal or any other University.

Student: Khadim SENE

A handwritten signature in black ink, appearing to read 'Khadim SENE', with several horizontal strokes above and below it.

# ACKNOWLEDGMENTS

At the end of this work, it is my pleasure, before starting this thesis, to express some thanks to all those who have allowed me, directly or indirectly, to write this manuscript. I ask in advance all the people I will forget to apologize to, the memory has always been lacking in me.

My first thanks will obviously go to all the members of the AIMS Senegal administration for the many opportunities they have given us and also for their many efforts that they never cease to renew for the good success of the students. Allow me to mention a few names such as Prof. Youssef Travele, Dr. Franck K. Mutombo, Mrs. Layih Butake, Mr. Diallo, Mrs. Nathalie Texiera, Mr. Bouba, Dr. Charle Kimpolo, ect.

I would like to express my deep and sincere gratitude to Prof. Ibrahima FAYE who advised and supported me throughout the preparation of this brief. I would like to express my deep gratitude to him for his great availability and kindness, which he has never missed during this year. He has always been of good advice.

My thanks go to Mr. Cheikh Birahim Ndao who spared no effort for the success of this work and his comments were constructive. Through him, may all AIMS Senegal staff and especially the tutors' corps feel thanked. May providence give them back a hundredfold the fruit of their labor.

A big thank you to all my classmates who supported me and to all the thousands of anonymous people who carried me in their prayers.

I cannot conclude without a special thank you to all the members of my family: parents, brothers and sisters for their infinite patience, their encouragement, their great availability, their generosity and finally for the moral support they have always given me.

A strong and pious thought to my generous father, who has affected us greatly his death. Thank you Mummy for your love, sacrifices and guidance towards me, you who have supported, encouraged and pushed me during all my years of study. I would say that without you, the story would not have been the same.

# DEDICATION

To my family and my friends.

To Professor Ibrahima FAYE who, through his teachings, fed my mind!

To all the teachers who deprive themselves of their families for three weeks every year to become missionaries of mathematics at AIMS Senegal, thousands of miles from their home!

To all those who helped me to produce this modest thesis.

To all those who never give up.

# Abstract

This essay studies non-parametric Bayesian methods for semi-supervised learning problems. It has seen rapid and sustained growth over the past 25 years.

Bayesian nonparametric (BNP) models are prior models for infinite-dimensional parameters, such as an unknown probability measure  $F$  or an unknown regression mean function  $f$ . We review some of the most widely used BNP, including the Dirichlet process (DP), DP mixture, and Gaussian process (GP) priors. Semi-supervised learning algorithms have become a hot topic of research as an alternative to traditional classification methods, exploiting the explicit classification information of labeled data with the knowledge hidden in the unlabeled data for building powerful and effective classifiers.

In the present work, a recently-developed alternative perspective on Bayesian prediction is reviewed. This approach facilitates recursive Bayesian prediction without computing a posterior, allowing insurers to perform real time updating of risk measures to assess solvency risk, and providing them with a tool for carrying out dynamic risk management strategies in today's "big data" era.

**Keywords :** non-parametric Bayesian method, Bayesian prediction semi-supervised learning, solvency risk.

# Contents

<b>Declaration</b>	<b>i</b>
<b>Acknowledgements</b>	<b>ii</b>
<b>Dedication</b>	<b>iii</b>
<b>Abstract</b>	<b>iv</b>
<b>Introduction</b>	<b>1</b>
<b>1 Generalities on Non-Parametric Bayesian Methods</b>	<b>5</b>
1.1 Bayesian Methods . . . . .	5
1.1.1 Nonparametric Bayesian Methods . . . . .	5
1.2 Bayesian Analysis . . . . .	6
1.2.1 Prior and Posterior distributions . . . . .	6
1.2.2 Bayes' Theorem . . . . .	7
1.2.3 Data Models . . . . .	8
1.3 Bayesian inference . . . . .	8
1.4 Bayesian Models . . . . .	8
1.4.1 Bayesian Nonparametric Models . . . . .	9
1.5 Dirichlet Process . . . . .	9
1.6 Dirichlet Process Mixture Models . . . . .	10
1.7 Gaussian Processes . . . . .	11
1.8 Gaussian mixture models . . . . .	11
1.9 Finite Gaussian mixture models . . . . .	11
1.10 Markov Chain Monte Carlo (MCMC) . . . . .	12
1.10.1 Gibbs Sampler . . . . .	12
<b>2 An overview of machine learning concepts</b>	<b>13</b>

2.1	Machine Learning . . . . .	13
2.1.1	Supervised, Unsupervised and Semi-Supervised Learning . . . . .	14
2.1.2	Bayesian Methods in Machine Learning . . . . .	15
<b>3</b>	<b>Solvency Risk in Insurance Regulation</b>	<b>17</b>
3.1	Solvency Risk . . . . .	17
3.2	Insurance Regulation . . . . .	17
3.3	Implementation and Results . . . . .	18
	<b>General Conclusion and Perspectives</b>	<b>22</b>
	<b>References</b>	<b>23</b>

# List of Figures

2.1	Types of Machine Learning	14
3.1	Data histogram	20
3.2	Risk capital evolution	20



# Introduction

Nonparametric statistical models are increasingly replacing parametric models, to overcome the latter's inflexibility to address a wide variety of data. A nonparametric model involves at least one infinite-dimensional parameter and hence may also be referred to as an "infinite-dimensional model". Indeed, the nomenclature "nonparametric" is misleading in that it gives the impression that there is no parameter in the model, while in reality there are infinitely many unknown quantities. However, the term nonparametric is so popular that it makes little sense not to use it. The infinite-dimensional parameter is usually a function or measure. In a canonical example of nonparametric model the data are a random sample from a completely unknown distribution  $P$ . More generally, functions of interest include the cumulative distribution function, density function, regression function, transition density of a Markov process, spectral density of a time series, response probability of a binary variable as a function of covariates, false discovery rate as a function of nominal level in multiple testing, receiver operating characteristic function between two distributions. Non-Bayesian methods for the estimation of many of these functions have been well developed, partly due to the availability of simple, natural and well respected estimators (e.g. the empirical distribution), and partly driven by the greater attention historically given to the classical approach. Bayesian estimation methods for nonparametric problems started receiving attention in the last three decades.

All statistical models involve certain parametric and structural assumptions. Bayesian nonparametric inference is an increasingly widely used approach to mitigate the dependence on such assumptions. Technically, Bayesian nonparametric (BNP) models can be defined as probability models on infinite-dimensional parameter spaces, usually devised for random distributions or random mean functions. Typical examples are the Dirichlet process (DP) and the Polya tree (PT) priors for random distributions, or Gaussian process (GP) priors for random functions.

The goal of machine learning is to build robust models based on observed data that, when deployed in real-life applications, generalize well to as-yet unseen examples of the sample population. Two major paradigms heavily studied in machine learning are supervised learning and unsupervised learning.

In supervised learning each data point is coupled with a label, generally indicating a class membership or a function output, where the goal is to infer a mapping from the data into labels and employ it in predicting labels of (existing or future) unlabeled data points. In unsupervised learning samples do not have labels, where the goal is to identify patterns or substructures in the data and to describe or represent the data by those patterns/substructures.

The emergence of a new paradigm in machine learning known as semi-supervised learning (SSL) has seen benefits to many applications where labeled data is expensive to obtain. However, unlike supervised learning (SL), which enjoys a rich and deep theoretical foundation, semi-supervised learning, which uses additional unlabeled data for training, still remains a theoretical mystery lacking a sound fundamental understanding.

The goal of Bayesian analysis is to reduce the uncertainty about unobserved variables by combining prior knowledge with observations. A fundamental limitation of a parametric statistical model, including a Bayesian approach, is the inability of the model to learn new structures. The goal of the learning process is to estimate the correct values for the parameters. The accuracy of these parameters improves with more data but the model's structure remains fixed. Therefore new observations will not affect the overall complexity (e.g. number of parameters in the model). Recently, nonparametric Bayesian methods have become a popular alternative to Bayesian approaches because the model structure is learned simultaneously with the parameter distributions in a data-driven manner.

In this work, we study non-parametric Bayesian methods on semi-supervised problems.

The remainder of this thesis is organized as follow:

In **Chapter 1**, we have covered the generalities on nonparametric Bayesian methods.

In **Chapter 2**, we have introduced the different concepts of machine learning such as supervised, unsupervised and semi-supervised learning.

In **Chapter 3**, we describe the nonparametric Bayesian method on a semi-supervised problem such as solvency risk in Insurance Regulation.

In **Chapter 4**, we give a general conclusion and suggest the future work.

## Motivation and Background

Most of machine learning is concerned with learning an appropriate set of parameters within a model class from training data. The meta level problems of determining appropriate model classes are referred to as model selection or model adaptation. These constitute important concerns for machine learning practitioners, chiefly for avoidance of over-fitting and under-fitting, but also for discovery of the causes and structures underlying data. Examples of model selection and adaptation include: selecting the number of clusters in a clustering problem, the number of hidden states in a hidden Markov model, the number of latent variables in a latent variable model, or the complexity of features used in nonlinear regression.

Bayesian nonparametrics is the study of Bayesian inference methods for nonparametric and semiparametric models. In the Bayesian nonparametric paradigm a prior distribution is assigned to all unknown quantities (parameters) involved in the modeling, whether finite or infinite dimensional. Inference is made from the “posterior distribution”, the conditional distribution of all parameters given the data. A model completely specifies the conditional distribution of all observable given all unobserved quantities, or parameters, while a prior distribution specifies the distribution of all unobservables. From this point of view, random effects and latent variables also qualify as parameters, and distributions of these quantities, often considered as part of the model itself from the classical point of view, are considered part of the prior. The posterior distribution involves an inversion of the order of conditioning. Existence of a regular version of the posterior is guaranteed under mild conditions on the relevant spaces.

Bayesian nonparametric methods provide a Bayesian framework for model selection and adap-

tation using nonparametric models. A Bayesian formulation of nonparametric problems is nontrivial, since a Bayesian model defines prior and posterior distributions on a single fixed parameter space, but the dimension of the parameter space in a nonparametric approach should change with sample size. The Bayesian nonparametric solution to this problem is to use an infinite-dimensional parameter space, and to invoke only a finite subset of the available parameters on any given finite data set. This subset generally grows with the data set. In the context of Bayesian nonparametric models, “infinite-dimensional” can therefore be interpreted as “of finite but unbounded dimension”. More precisely, a Bayesian nonparametric model is a model that constitutes a Bayesian model on an infinite-dimensional parameter space and can be evaluated on a finite sample in a manner that uses only a finite subset of the available parameters to explain the sample.

There are many ways to use data, whether experimental or observational, to better understand the world and to make better decisions. The Bayesian approach distinguishes itself from other approaches with two distinct sources of sound foundational support. The first is the theory of subjective probability, developed from a set of axioms that describe rational behavior. Subjective probability provides an alternative to the relative frequency definition of probability. Under subjective probability, individuals are free to have their own assessments of probabilities. This theory leads inexorably to Bayesian methods (Savage, 1954). The second source of support is decision theory which formalizes statistical inference as a decision problem. The combination of state-of-nature (parameter) and action (say, an estimate) yield a loss, and a good inference procedure (decision rule) leads to a small expected loss.

Nearly all agree that inadmissible inference procedures are to be avoided. The complete class theorems show that the entire set of admissible inference procedures is comprised of Bayesian procedures and of procedures that are close to Bayesian in a technical sense (Berger, 1985). Procedures that are far from Bayesian can be useful, but they must be justified on special grounds for example, our inability to discover a dominating procedure, our inability to implement the dominating procedure due to computational limitations, or to address robustness issues, perhaps due to the shortcomings of our formal mathematical model. The twin perspectives of subjective probability and decision theory have convinced many that statistical inference should be driven by Bayesian methods. However, neither of these perspectives describes how to conduct a sound Bayesian analysis. Surely, we have all seen examples of poor Bayesian analyses, and so we seek guiding principles to help in the use of Bayesian methods. [(MacEachern, 2016)]

## Main Objective

The main objective work is to this study nonparametric Bayesian methods in Machine Learning.

**Specific Objectives**

In order to achieve the main objective, we have subdivided it into several specific objectives:

1. To review the concepts around the non-parametric Bayesian method and Machine Learning.
2. To explore nonparametric Bayesian method in a semi supervised problem.

# 1. Generalities on Non-Parametric Bayesian Methods

In this section we are going to give some basic concepts of the Non-Parametric Bayesian Methods.

## 1.1 Bayesian Methods

Bayesian methods is a term which may be used to refer to any mathematical tools that are useful and relevant in some way to Bayesian inference, an approach to statistics based on the work of Thomas Bayes (1701–1761). Bayes was an English mathematician and Presbyterian minister who is best known for having formulated a basic version of the well-known Bayes' Theorem.

Bayesian inference is different to classical inference (or frequentist inference) mainly in that it treats model parameters as random variables rather than as constants. The Bayesian framework (or paradigm) allows for prior information to be formally taken into account. It can also be useful for formulating a complicated statistical model that presents a challenge to classical methods. [(Puza, 2015)]

### 1.1.1 Nonparametric Bayesian Methods

**Definition 1.1.1.** For parametric Bayesian models, the parameter space is pre-specified. No matter how the data changes, the number of parameters is fixed. This restriction may cause limitations on model capacity, especially for big data applications, where it may be difficult or even counter-productive to fix the number of parameters a priori. For example, a Gaussian mixture model with a fixed number of clusters may fit the given dataset well; however, it may be sub-optimal to use the same number of clusters if more data comes under a slightly changed distribution.

Nonparametric Bayesian (NPB) methods provide an elegant solution to such needs on automatic adaptation of model capacity when learning a single model. Such adaptivity is obtained by defining stochastic processes on rich measure spaces. Classical examples include Dirichlet process (DP) and Gaussian process (GP). Below, we briefly review DP and GP. [(Kurihara et al., 2007)]

**Definition 1.1.2.** In parametric Bayesian inference we have a model  $\mathcal{M} = \{f(y|\theta) : \theta \in \Theta\}$  and data  $Y_1, \dots, Y_n \sim f(y|\theta)$ . We put a prior distribution  $\Pi(\theta)$  on the parameter  $\theta$  and compute the posterior distribution using Bayes' rule :

$$\Pi(\theta|Y) = \frac{\mathcal{L}_n(\theta)\Pi(\theta)}{m(Y)} \quad (1.1.1)$$

where  $Y = (Y_1, \dots, Y_n)$ ,  $\mathcal{L}_n(\theta) = \prod_i f(Y_i|\theta)$  is the likelihood function and

$$m(y) = m(y_1, \dots, y_n) = \int f(y_1, \dots, y_n|\theta)\Pi(\theta) d\theta = \int \prod_{i=1}^n f(y_i|\theta)\Pi(\theta) d\theta$$

is the marginal distribution for the data induced by the prior and the model. We call  $m$  the induced marginal. The model may be summarized as :

$$\begin{aligned}\theta &\sim \Pi \\ Y_1, \dots, Y_n|\theta &\sim f(y|\theta).\end{aligned}$$

We use the posterior to compute a point estimator such as the posterior mean of  $\theta$ . We can also summarize the posterior by drawing a large sample  $\theta_1, \dots, \theta_N$  from the posterior  $\Pi(\theta|Y)$  and the plotting the samples.

In nonparametric Bayesian inference, we replace the finite dimensional model  $\{f(y|\theta) : \theta \in \Theta\}$  with an infinite dimensional model such as

$$\mathcal{F} = \{f : \int (f''(y))^2 dy < \infty\} \quad (1.1.2)$$

Typically, neither the prior nor the posterior have a density function with respect to a dominating measure. But the posterior is still well defined. On the other hand, if there is a dominating measure for a set of densities  $\mathcal{F}$  then the posterior can be found by Bayes theorem :

$$\Pi_n(A) = \mathcal{P}(f \in A|Y) = \frac{\int_A \mathcal{L}_n(f) d\Pi(f)}{\int_{\mathcal{F}} \mathcal{L}_n(f) d\Pi(f)} \quad (1.1.3)$$

where  $A \subset \mathcal{F}$ ,  $\mathcal{L}_n(f) = \prod_i f(Y_i)$  is the likelihood function and  $\Pi$  is a prior on  $\mathcal{F}$ . If there is no dominating measure for  $\mathcal{F}$  then the posterior still exists but cannot be obtained by simply applying Bayes' theorem. An estimate of  $f$  is the posterior mean

$$\hat{f}(y) = \int f(y) d\Pi_n(f). \quad (1.1.4)$$

A posterior  $1 - \alpha$  region is any set  $A$  such that  $\Pi_n(A) = 1 - \alpha$ .

## 1.2 Bayesian Analysis

### 1.2.1 Prior and Posterior distributions

#### Prior distribution

- In Bayesian statistical inference, a prior probability distribution, often simply called the prior, of an uncertain quantity is the probability distribution that would express one's beliefs

about this quantity before some evidence is taken into account. For example, the prior could be the probability distribution representing the relative proportions of voters who will vote for a particular politician in a future election. The unknown quantity may be a parameter of the model or a latent variable rather than an observable variable.

- $Y_1, Y_2, \dots, Y_n$  Random Variables associated with a sample of size  $n$ ,  $L(y_1, y_2, \dots, y_n | \theta)$  the likelihood of the sample.

In the Bayesian approach, the unknown parameter  $\theta$  is viewed to be a random variable with a probability distribution, called the prior distribution of  $\theta$ . This prior distribution is specified before any data are collected and provides a theoretical description of information about  $\theta$  that was available before any data were obtained.

### Posterior distribution

In Bayesian statistics, the posterior probability distribution is the probability distribution of an unknown quantity, treated as a random variable, conditional on the evidence obtained from an experiment or survey.

The posterior probability is the probability of the parameters  $\theta$  given the evidence  $X$ :  $\mathbb{P}(\theta|X)$ . It contrasts with the likelihood function, which is the probability of the evidence given the parameters:  $\mathbb{P}(X|\theta)$ .

The two are related as follows:

Let us have a prior belief that the probability distribution function is  $\mathbb{P}(\theta)$  and observations  $x$  with the likelihood  $\mathbb{P}(x|\theta)$ , then the posterior probability is defined as

$$\mathbb{P}(\theta|x) = \frac{\mathbb{P}(x|\theta)}{\mathbb{P}(x)} \mathbb{P}(\theta)$$

The posterior probability can be written in the memorable form as :

Posterior probability  $\propto$  Likelihood  $\times$  Prior probability.

### 1.2.2 Bayes' Theorem

The foundation of Bayesian statistics is Bayes' theorem. Suppose we observe a random variable  $y$  and wish to make inferences about another random variable  $\theta$ , where  $\theta$  is drawn from some distribution  $\mathbb{P}(\theta)$ . From the definition of conditional probability,

$$\mathbb{P}(\theta|y) = \frac{\mathbb{P}(y, \theta)}{\mathbb{P}(y)} \quad (1.2.1)$$

Again from the definition of conditional probability, we can express the joint probability by conditioning on  $\theta$  to give

$$\mathbb{P}(y, \theta) = \mathbb{P}(y|\theta) \mathbb{P}(\theta) \quad (1.2.2)$$

Putting these together gives Bayes' theorem:

$$\mathbb{P}(\theta|y) = \frac{\mathbb{P}(y|\theta) \mathbb{P}(\theta)}{\mathbb{P}(y)} \quad (1.2.3)$$

With  $n$  possible outcomes  $(\theta_1, \dots, \theta_n)$ ,

$$\mathbb{P}(\theta_j|y) = \frac{\mathbb{P}(y|\theta_j) \mathbb{P}(\theta_j)}{\mathbb{P}(y)} = \frac{\mathbb{P}(y|\theta_j)}{\sum_{i=1}^n \mathbb{P}(\theta_i) \mathbb{P}(y|\theta_i)} \quad (1.2.4)$$

### 1.2.3 Data Models

The first step in a Bayesian analysis is to choose a probability model for the data. This process, which is analogous to the classic approach of choosing a data model, involves deciding on a probability distribution for the data if the parameters were known. If the  $n$  data values to be observed are  $y_1, \dots, y_n$ , and the vector of unknown parameters is denoted  $\theta$ , then, assuming that the observations are made independently, we are interested in choosing a probability function  $\mathbb{P}(y_i|\theta)$  for the data (the vertical bar means “conditional on” the quantities to the right). In situations where we have extra covariate information,  $x_i$ , for the  $i$ th case, as in regression models, we would choose a probability function of the form  $\mathbb{P}(y_i|x_i, \theta)$ .

When the data are not conditionally independent given the parameters and covariates, we must specify the joint probability function,  $\mathbb{P}(y_1, \dots, y_n|x_1, \dots, x_n, \theta)$ .

## 1.3 Bayesian inference

Bayesian inference is a method of statistical inference in which Bayes' theorem is used to update the probability for a hypothesis as more evidence or information becomes available. Bayesian inference is an important technique in statistics, and especially in mathematical statistics.

### 1.4 Bayesian Models

Bayes' formula extends naturally to statistical models. A Bayesian model is a parametric model in the classical (or frequentist) sense, but with the addition of a prior probability



distribution for the model parameter, which is treated as a random variable rather than an unknown constant. The basic components of a Bayesian model may be listed as:

- the data, denoted by  $y$
- the parameter, denoted by  $\theta$
- the model distribution, given by a specification of  $f(y|\theta)$  or  $F(y|\theta)$  or the distribution of  $(y|\theta)$
- the prior distribution, given by a specification of  $f(\theta)$  or  $F(\theta)$  or the distribution of  $\theta$ .

Here,  $F$  is a generic symbol which denotes cumulative distribution function (cdf), and  $f$  is a generic symbol which denotes probability density function (pdf) (when applied to a continuous random variable) or probability mass function (pmf) (when applied to a discrete random variable). For simplicity, we will avoid the term pmf and use the term pdf or density for all types of random variable, including the mixed type.

### 1.4.1 Bayesian Nonparametric Models

#### Definition

A Bayesian nonparametric model is a Bayesian model on an infinite-dimensional parameter space. The parameter space is typically chosen as the set of all possible solutions for a given learning problem. For example, in a regression problem the parameter space can be the set of continuous functions, and in a density estimation problem the space can consist of all densities.

Bayesian nonparametric models have recently been applied to a variety of machine learning problems, including regression, classification, clustering, latent variable modeling, sequential modeling, image segmentation, source separation and grammar induction. (Orbanz and Teh, 2010)

## 1.5 Dirichlet Process

The Dirichlet process is a stochastic process used in Bayesian nonparametric models of data, particularly in Dirichlet process mixture models (also known as infinite mixture models). It is called a Dirichlet process because it has Dirichlet distributed finite dimensional marginal distributions, just as the Gaussian process, another popular stochastic process used for Bayesian nonparametric regression, has Gaussian distributed finite dimensional marginal distributions.

**Definition 1.5.1.** A Dirichlet process  $DP(\alpha, H)$  parametrized by a concentration parameter  $\alpha > 0$  and a base distribution  $H$  is a prior over distributions (probability measures)  $G$  such that, for any finite partition  $A_1, \dots, A_m$  of the parameter space, the induced random vector  $(G(A_1), \dots, G(A_m))$  is Dirichlet distributed with parameters  $(\alpha H(A_1), \dots, \alpha H(A_m))$ .

## 1.6 Dirichlet Process Mixture Models

The Dirichlet process (DP) mixture model, studied in Bayesian nonparametric (Ferguson 1973, Blackwell and MacQueen 1973, Berry and Christensen 1979, Escobar and West 1995, Liu 1996, MacEachern and Muller 1998, Neal 2000, Rasmussen 2000, Ishwaran and James 2002), allows for exactly this possibility. When a new data point arrives, it either shares the component of some previously drawn value, or it uses a newly generated component realized from a distribution  $G_0$ . The frequency with which new components are generated is controlled by a parameter  $\alpha > 0$ . The DP mixture model has found widespread application in recent statistical research (Rasmussen and Ghahramani 2002, Carota and Parmigiani 2002, Gelfand and Kottas 2002, Ishwaran and James 2002, Brown and Ibrahim 2003, Iorio et al. 2004, Muller et al. 2004).

**Definition 1.6.1.** Let  $\eta$  be a continuous random variable, let  $G_0$  be a non-atomic probability distribution for  $\eta$ , and let  $\alpha$  be a positive, real-valued scalar. A random measure  $G$  is distributed according to a Dirichlet process (DP) (Ferguson, 1973), with scaling parameter  $\alpha$  and base distribution  $G_0$ , if for all natural numbers  $k$  and  $k$ -partitions  $\{B_1, \dots, B_k\}$ ,

$$(G(B_1), G(B_2), \dots, G(B_k)) \sim \text{Dir}(\alpha G_0(B_1), \alpha G_0(B_2), \dots, \alpha G_0(B_k)). \quad (1.6.1)$$

Integrating out  $G$ , the joint distribution of the collection of variables  $\{\eta_1, \dots, \eta_n\}$  exhibits a clustering effect; conditioning on  $n - 1$  draws, the  $n$ th value is, with positive probability, exactly equal to one of those draws:

$$p(\cdot | \eta_1, \dots, \eta_{n-1}) \propto \alpha G_0(\cdot) + \sum_{i=1}^{n-1} \delta_{\eta_i}(\cdot). \quad (1.6.2)$$

Thus, the variables  $\{\eta_1, \dots, \eta_{n-1}\}$  are randomly partitioned according to which variables are equal to the same value, with the distribution of the partition obtained from a Pólya urn scheme (Blackwell et al., 1973). Let  $\{\eta_1^*, \dots, \eta_{|c|}^*\}$  denote the distinct values of  $\{\eta_1, \dots, \eta_{n-1}\}$ , let  $c = \{c_1, \dots, c_{n-1}\}$  be assignment variables such that  $\eta_i = \eta_{c_i}^*$ , and let  $|c|$  denote the number of cells in the partition. The distribution of  $\eta_n$  follows the urn distribution:

$$\eta_n = \begin{cases} \eta_i^* & \text{with prob. } \frac{|j : c_j = i|}{n - 1 + \alpha} \\ \eta, \eta \sim G_0 & \text{with prob. } \frac{\alpha}{n - 1 + \alpha}, \end{cases} \quad (1.6.3)$$

Where  $|j : c_j = i|$  is the number of times the value  $\eta_i^*$  occurs in  $\{\eta_1, \dots, \eta_{n-1}\}$ .

## 1.7 Gaussian Processes

### Definition

In probability theory and statistics, a Gaussian process is a stochastic process (a collection of random variables indexed by time or space), such that every finite collection of those random variables has a multivariate normal distribution, i.e. every finite linear combination of them is normally distributed. The distribution of a Gaussian process is the joint distribution of all those (infinitely many) random variables, and as such, it is a distribution over functions with a continuous domain, e.g. time or space.

A machine-learning algorithm that involves a Gaussian process uses lazy learning and a measure of the similarity between points (the kernel function) to predict the value for an unseen point from training data. The prediction is not just an estimate for that point, but also has uncertainty information, it is a one-dimensional Gaussian distribution (which is the marginal distribution at that point).

## 1.8 Gaussian mixture models

### Definition

Mixture modeling remains one of the most useful tools in statistics, machine learning and data mining for applications involving density estimation or clustering. One of the most prominent recent developments in this field is the application of nonparametric Bayesian techniques to mixture modeling, which allow for the automatic determination of an appropriate number of mixture components. Current inference algorithms for such models are mostly based on Gibbs sampling, which suffer from a number of drawbacks.

## 1.9 Finite Gaussian mixture models

### Definition

Finite Gaussian mixture models are widely used in various fields of computer vision and image processing. The adoption of this model-based approach to data clustering and modeling brings important advantages: for instance, the selection of the number of classes or the assessment of the validity for a given model can be addressed in a formal way. However, it is well-known that the Gaussian density has some drawbacks such as the rigidity of its shape which prevents it from having a good approximation to data with outliers. [(Cheng et al., 2006)], [(McLachlan and Peel, 2000)]

## 1.10 Markov Chain Monte Carlo (MCMC)

### Definition

Markov chain Monte Carlo (MCMC) methods are stochastic simulation methods that allow to approximate a given target distribution such as the posterior law, by relying on Markov chain theory and Monte Carlo integration.

They proceed in two main steps. First, a Markov chain is built with a given transition rule such that its stationary states follow the posterior law [Hastings, 1970; Gamerman and Lopes, 2006]. Once the Markov chain has reached its stationary distribution, Monte Carlo approximation is used to infer the posterior characteristics. Because of the lack of knowledge about the posterior distribution, the Markov chain often starts at a random point far from the target high density regions. Then, if the MCMC algorithm is not run a sufficiently long time, the resulting estimators are likely to be, highly biased leading to unreliable inference and poor forecasts.

### 1.10.1 Gibbs Sampler

**Definition 1.10.1.** In statistics, Gibbs sampling or a Gibbs sampler is a Markov chain Monte Carlo (MCMC) algorithm for obtaining a sequence of observations which are approximated from a specified multivariate probability distribution, when direct sampling is difficult. This sequence can be used to approximate the joint distribution (e.g., to generate a histogram of the distribution); to approximate the marginal distribution of one of the variables, or some subset of the variables (for example, the unknown parameters or latent variables); or to compute an integral (such as the expected value of one of the variables). Typically, some of the variables correspond to observations whose values are known, and hence do not need to be sampled.

### Conclusion

In this chapter we have presented some basic concepts about Bayesian Methods, namely the non-parametric Bayesian method, the Bayes' theorem, Bayesian analysis, Gaussian process, Gaussian mixture model and so on. In the next chapter we will give in detail some Machine Learning concepts such as Supervised, Unsupervised and Semi-Supervised Learning.

## 2. An overview of machine learning concepts

In this section we are going to give some concepts of Machine learning. In order to understand the nature of semi-supervised learning, it will be useful first to take a look at supervised and unsupervised learning.

### 2.1 Machine Learning

Machine learning (ML) is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns and inference instead. It is seen as a subset of artificial intelligence. Machine learning algorithms build a mathematical model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to perform the task. Machine learning algorithms are used in a wide variety of applications, such as email filtering and computer vision, where it is difficult or infeasible to develop a conventional algorithm for effectively performing the task.

Bayesian machine learning allows us to encode our prior beliefs about what those models should look like, independent of what the data tells us. This is especially useful when we don't have a ton of data to confidently learn our model.

We have three types of Machine Learning : supervised learning, unsupervised learning, and reinforcement learning.

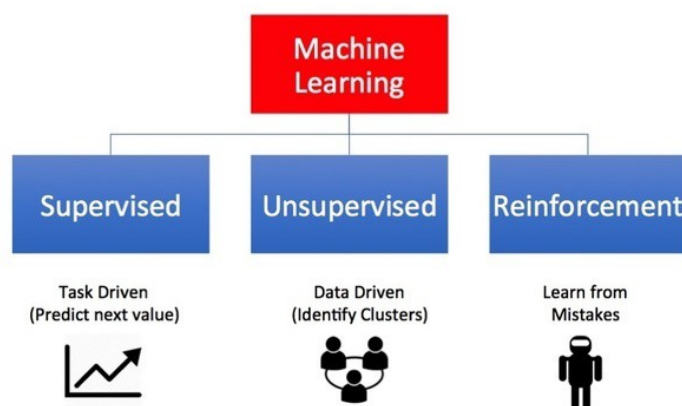


Figure 2.1: Types of Machine Learning

## 2.1.1 Supervised, Unsupervised and Semi-Supervised Learning

### 2.1.1.1 Supervised and Unsupervised Learning

Traditionally, there have been two fundamentally different types of tasks in machine learning.

- The first one is unsupervised learning. Let  $X = (x_1, \dots, x_n)$  be a set of  $n$  examples (or points),  $x_i \in \mathcal{X}$  for all  $i \in [n] = 1, \dots, n$ . Typically it is assumed that the points are drawn i.i.d (independently and identically distributed) from a common distribution on  $\mathcal{X}$ . It is often convenient to define the  $(n \times d)$  – *matrix*  $\mathbf{X} = (x_i^T)_{i \in [n]}^T$  that contains the data points as its rows.

The goal of unsupervised learning is to find interesting structure in the data  $X$ . It has been argued that the problem of unsupervised learning is fundamentally that of estimating a density which is likely to have generated  $X$ . However, there are also weaker forms of unsupervised learning, such as quantile estimation, clustering, outliers detection, and dimensionality reduction.

- The second task is supervised learning. The goal is to learn a mapping from learning  $x$  to  $y$ , given a training set made of pairs  $(x_i, y_i)$ . Here, the  $y_i \in \mathcal{Y}$  are called the labels or targets of the examples  $x_i$ . If the labels are numbers,  $\mathbf{y} = (y_i^T)_{i \in [n]}^T$  denotes the column vector of labels. Again, a standard requirement is that the pairs  $(x_i, y_i)$  are sampled i.i.d from some distribution which here ranges over  $\mathcal{X} \times \mathcal{Y}$ .

The task is well defined, since a mapping can be evaluated through its predictive performance on test examples. There are two families of algorithms generative for supervised learning.

When  $\mathcal{Y} = \mathbb{R}$  or  $\mathcal{Y} = \mathbb{R}^d$  (or more generally, when the labels are continuous), the task is called regression.

The case where  $y$  takes values in a finite set (discrete labels) is called classification.

### 2.1.1.2 Semi-Supervised Learning

Semi-supervised learning is a class of machine learning tasks and techniques that also make use of unlabeled data for training typically a small amount of labeled data with a large amount of unlabeled data. Semi-supervised learning falls between unsupervised learning (without any labeled training data) and supervised learning (with completely labeled training data).

Many machine-learning researchers have found that unlabeled data, when used in conjunction with a small amount of labeled data, can produce considerable improvement in learning accuracy. The acquisition of labeled data for a learning problem often requires a skilled human agent (e.g. to transcribe an audio segment) or a physical experiment (e.g. determining the 3D structure of a protein or determining whether there is oil at a particular location).

## 2.1.2 Bayesian Methods in Machine Learning

Bayesian statistics has been applied to almost every Machine Learning (ML) task, ranging from the single-variate regression classification to the structured output predictions and to the unsupervised/semi-supervised learning scenarios (Bishop, 2006). In essence, however, there are several basic tasks that we briefly review below.

**Prediction :** after training, Bayesian models make predictions using the distribution:

$$\mathbb{P}(x|\mathcal{D}) = \int \mathbb{P}(x, \Theta|\mathcal{D}) d\Theta = \int \mathbb{P}(x|\Theta, \mathcal{D})\mathbb{P}(\Theta|\mathcal{D}) d\Theta \quad (2.1.1)$$

where  $\mathbb{P}(x, \Theta|\mathcal{D})$  is often simplified as  $\mathbb{P}(x|\Theta)$  due to the i.i.d assumption of the data when the model is given. Since the integral is taken over the posterior distribution, the training data is considered.

**Model Selection :** Model selection is a fundamental problem in statistics and machine learning (Kadane and Lazar, 2004). Let  $M$  be a family of models, where each model is indexed by a set of parameters  $\Theta$ . Then, the marginal likelihood of the model family (or model evidence) is

$$\mathbb{P}(\mathcal{D}|M) = \int \mathbb{P}(\mathcal{D}|\Theta) \mathbb{P}(\Theta|M) d\Theta, \quad (2.1.2)$$

where  $\mathbb{P}(\mathcal{D}|M)$  is often assumed to be uniform if no strong prior exists.

For two different model families  $M_1$  and  $M_2$ , the ratio of model evidences  $k = \frac{\mathbb{P}(\mathcal{D}|M_1)}{\mathbb{P}(\mathcal{D}|M_2)}$  is called Bayes factor (Kass and Raftery, 1995).

## Conclusion

In this chapter we have introduced the different concepts of machine learning such as supervised, unsupervised and semi-supervised learning.

## 3. Solvency Risk in Insurance Regulation

After the study of the non-parametric Bayesian method, our objective is now apply it to a semi-supervised problem more preciously on Solvency Risk in Insurance regulation. With the lack of availability of data, we going to use directly the results from the implementation of the Danish fire insurance loss data and then try to describe them in a more simple manner.

### 3.1 Solvency Risk

Solvency risk is the risk that an institution cannot meet maturing obligations as they come due for full value (even if it may be able to settle at some unspecified time in the future) even after disposal of its assets. Liquidity risk refers to the risk that involves the disposal of assets or selling of assets.

It is a critical concern for both insurers and insurance regulators. Indeed, insurance regulations often require that insurers monitor their solvency risk continuously and take into consideration of the changes of their risk profiles. For example, the European insurance regulation Solvency II requires that all insurers must calculate their risk capital at least once a year and monitor it on a continuous basis. Similar regulations can be found in the US regulation Own Risk and Solvency Assessment (ORSA, 2017). From the insurers' point of view, they should go beyond the minimum requirement set by the regulators. That is, they should have clear picture of their financial health at all times, not just at the end of each year. Therefore, it is desirable that they know their solvency risk in real time so that they can respond in case their solvency risk goes above their tolerance level. Since these questions about risk naturally involve future losses, this boils down to a prediction problem, and any sort of risk summary would be best described as a suitable feature of the predictive distribution for these future losses.

### 3.2 Insurance Regulation

Insurance regulation often dictates that insurers monitor their solvency risk in real time and take appropriate actions whenever the risk exceeds their tolerance level. Bayesian methods are appealing for prediction problems thanks to their ability to naturally incorporate both sample variability and parameter uncertainty into a predictive distribution. However, handling data arriving in real time requires a flexible nonparametric model, and the Monte Carlo methods necessary to evaluate the predictive distribution in such cases are not recursive and can be too expensive to rerun each time new data arrives.



## Value at Risk in Insurance

Value at risk (VaR) is a popular method for risk measurement. VaR calculates the probability of an investment generating a loss, during a given time period and against a given level of confidence. It gives investors an indication of the level of risk they take with a certain investment. This can help them decide whether the possible gain is worth the potential maximum loss. VaR can be calculated for either one asset, a portfolio of multiple assets of an entire firm.

### 3.3 Implementation and Results

#### Algorithm

For log-losses  $X_1, X_2, \dots$ , the following summarizes the recursive algorithm.

1. Make an initial guess of  $\hat{F}_0$  with density  $\hat{f}_0$  whose support is the whole real line.
2. Fix a grid of points,  $\{\bar{x}_m : m = 1, \dots, M\}$ , in  $\mathbb{R}$ , covering roughly the entire support of  $\hat{f}_0$ , where  $M$  is a positive integer set by the actuary.
3. For each  $m$ , compute the sequence  $(\hat{F}_n(\bar{x}_m))$  using

$$\hat{F}_n(x) = (1 - \alpha_n) \hat{F}_{n-1}(x) + \alpha_n C_p(\hat{F}_{n-1}(x), \hat{F}_{n-1}(X_n))$$

Where where  $C_p$  is given by

$$C_p(u, v) = \Phi \left( \frac{\Phi^{-1}(u) - \rho \Phi^{-1}(v)}{(1 - \rho)^2} \right)$$

which is a distribution function in  $u$  for fixed  $v$ .

Since data  $X_i$  is surely not to fall exactly on the specified grid, an interpolation procedure, like `approxfun` in R, can be used to evaluate  $\hat{F}_{i-1}(X_i)$

4. Given the distribution function  $\hat{F}_n(x)$ , the corresponding density function  $\hat{f}_n(x)$  can be readily evaluated using difference ratios, i.e.,

$$\hat{f}_n(\bar{x}_m) = \frac{\hat{F}_n(\bar{x}_m) - \hat{F}_n(\bar{x}_{m-1})}{\bar{x}_m - \bar{x}_{m-1}}$$

and, again, interpolation can be used to evaluate  $\hat{f}_n(x)$  for points  $x$  off the grid. Some additional smoothing, e.g., `smooth.spline` in R, can also be used to improve the relatively crude difference-ratio estimate above.

**Example : Danish fire insurance loss data**

The complete Danish data on fire insurance losses, hereby abbreviated as the "Danish data", has been studied by several authors; see, for example, Scollnik and Sun (2012), Cooray and Cheng (2015), Calderin-Ojeda and Kwok (2016) and the references therein. The data is comprised of  $n = 2492$  fire insurance loss entries from 1980 to 1990. To account for inflation, the data has been adjusted to reflect 1985 values. All losses are expressed in Danish Krone, and about 94% are between 1 and 7 million Kroner. Note that the data set is traditionally stored in ascending order, which does not resemble an iid sequence, so we work here with a random permutation  $X_{i_1}, \dots, X_{i_n}$  of the sorted data  $X_1, \dots, X_n$ .

A histogram of the log-losses is shown in figure 3.1, along with two final estimators  $\hat{f}_n$  from the recursive algorithm based on Student and normal initial guesses and other default settings. Since this data set is relatively large, as the convergence theory suggests, the two recursive estimators both are able to adapt to the unusual shape of the data histogram, and provide a satisfactory fit.

The recursive algorithm also provides insurers with online updating of the predictive distribution so that real-time updating of the risk capital is possible. To illustrate this, we treat the permuted data as arriving in real time. We also assume that the insurer has no past data on this line of business. Recall that Solvency II requires the insurer to set its risk capital to Value at Risk (VaR) at the level 99.5%. Figure 3.2 gives a plot of the evolution of VaR along the data sequence for both  $t_2$  and  $N(0; 4^2)$  initial guesses with the unit for the vertical axis being log-millions; the choice of  $\sigma = 4$  in the normal is to roughly match the VaR of  $t_2$ , so that the comparison essentially only involves their tail thicknesses, not an overall scale difference.

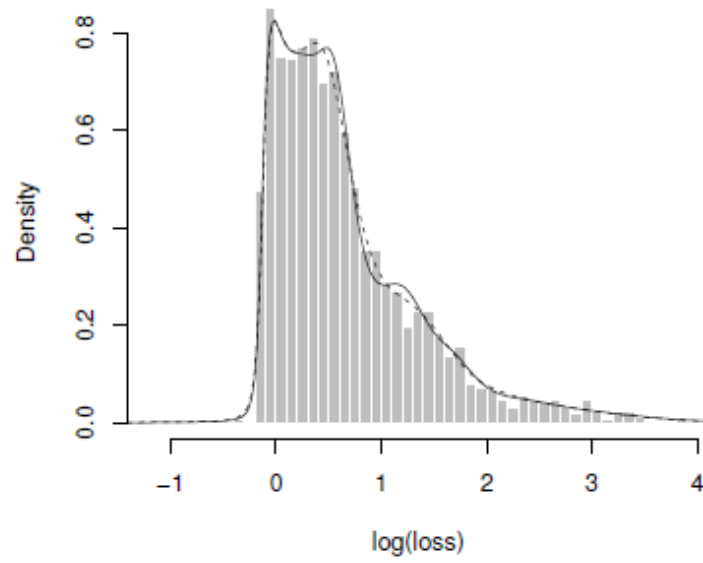


Figure 3.1: Data histogram

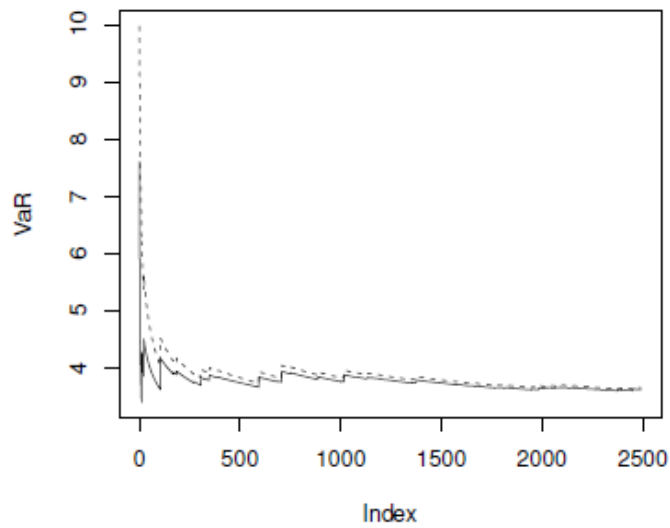


Figure 3.2: Risk capital evolution

These two figures are results for the Danish fire insurance data described in the above example.

Figure 3.1 : Data histogram with the recursive estimators with  $t_2$  initial guess (solid) and

$N(0; 4^2)$  initial guess (dashed);

Figure 3.2 : Evolution of risk capital for the same two initial guesses.

### Discussion :

Sound management of solvency risk requires the insurer to monitor their solvency risk continuously and preferably in real time. While Bayesian analysis has been successful in estimating the risk measures to gauge solvency risk, no existing Bayesian insurance models are able to allow insurers to perform real-time updating of their solvency risk. This is due to the fact implementation of Bayesian models often requires MCMC, which makes real-time updating of predictive distribution infeasible. Motivated by this, our study introduces to actuarial science a new perspective of Bayesian recursive prediction that allows insurers to recursively update predictive distribution without computing the posterior. This new approach enables insurers to update the predictive distribution recursively in real time. Though we chosen risk capital set by Solvency II as our vehicle for illustration, the same can be done for any other risk measures such as ruin probabilities and conditional tail expectation. This real data example show how insurers may use this new method to monitor its risk capital and manage its solvency risk.

# General Conclusion and Perspectives

In this work, we were challenged to give first an overview of the Non-Parametric Bayesian method then to apply it to a semi-supervised problem.

Our approach was structured in three main phases :

In a first phase, we conducted a study on the different concepts around the Bayesian statistic namely the non-parametric Bayesian method such as Bayes' theorem, Bayesian model, Gaussian Mixture Model, Markov Chain Monte Carlo and so one.

The second and final phase was dedicated to the application of the Non-Parametric Bayesian Method to our Semi Supervised Problem.

In perspective, we found it useful to see how to get a good dataset for the application part and also to apply the nonparametric method to some other semi-supervised problems such as Strucrural Health Monitoring.

# References

- Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.
- David Blackwell, James B MacQueen, et al. Ferguson distributions via pólya urn schemes. *The annals of statistics*, 1(2):353–355, 1973.
- Jian Cheng, Jie Yang, Yue Zhou, and Yingying Cui. Flexible background mixture models for foreground segmentation. *Image and Vision Computing*, 24(5):473–482, 2006.
- Thomas S Ferguson. A bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230, 1973.
- Liang Hong and Ryan Martin. Real-time bayesian non-parametric prediction of solvency risk. *Annals of Actuarial Science*, 13(1):67–79, 2019.
- Joseph B Kadane and Nicole A Lazar. Methods and criteria for model selection. *Journal of the American statistical Association*, 99(465):279–290, 2004.
- Robert E Kass and Adrian E Raftery. Bayes factors. *Journal of the american statistical association*, 90(430):773–795, 1995.
- Kenichi Kurihara, Max Welling, and Yee Whye Teh. Collapsed variational dirichlet process mixture models. In *IJCAI*, volume 7, pages 2796–2801, 2007.
- Steven N MacEachern. Nonparametric bayesian methods: a gentle introduction and overview. *Communications for Statistical Applications and Methods*, 23(6):445–466, 2016.
- Yosra Marnissi. *Bayesian methods for inverse problems in signal and image processing*. PhD thesis, 2017.
- Geoffrey McLachlan and David Peel. Wiley series in probability and statistics. *Finite Mixture Models*, pages 420–427, 2000.
- Peter Orbanz and Yee Whye Teh. Bayesian nonparametric models. *Encyclopedia of Machine Learning*, pages 81–89, 2010.
- Borek Puza. *Bayesian Methods for Statistical Analysis*. ANU Press, 2015.
- TJ Rogers, K Worden, R Fuentes, N Dervilis, UT Tygesen, and EJ Cross. A bayesian non-parametric clustering approach for semi-supervised structural health monitoring. *Mechanical Systems and Signal Processing*, 119:100–119, 2019.