Bayes, Bootstrap, Moments*

by

Jean-Pierre Florens** **and Jean-Marie Rolin*****

June 1994

Abstract

This paper presents a bayesian analysis of semi parametric models in which the vector of parameters of interest is characterized by a moment equation. Powerful representations of the Dirichlet processes are used and provide an efficient numerical strategy to deal with such models. So called non informative prior specification gives a bayesian interpretation of the bootstrap method but some pathologies of this prior measure are pointed out. Several numerical applications illustrate the presentation.

Key words: Bayesian semi parametric models, Dirichlet processes, Bootstrap method, moment estimation.

AMS. MOS: primary: G2F15;
secondary: G2G05 G2G25 G5V05 G0G57

---

* This research follows from our previous works on non parametric bayesian models but this particular paper started by very stimulating discussions with Tony Lancaster during his visit in Toulouse in April 1993. His previous contributions was the seeds of numerous ideas of this paper and we thank him warmly.
We also thank Michel Mouchart for all the history of our joint research on this field and for specific suggestions on the topic of this paper and Jean-Pierre Raoult for very helpful discussions.
Computations presented in this paper have been performed in a very efficient and rigourous way by Vincent Florens and Florent Simondin.

** GREMAQ and IDEI, Université des Sciences Sociales de Toulouse, Toulouse, France.

*** Institut de Statistique and CORE, Université Catholique de Louvain, Voie du Roman Pays 34, 1348 Louvain-la-Neuve, Belgium.

1 Introduction

This paper is devoted to the analysis of the following statistical structure :

- the sample $x = (x_i)_{i=1,\cdots,n}$ is independently identically distributed and P is the probability which generates any x_i.

- the probability P is unknown and the statistician specifies a prior probability distribution on P. This prior probability measure will be a Dirichlet process in general but this family of prior may be extended, if necessary, to more general distributions, such as mixtures of Dirichlet processes. This prior distribution could be "low informative" or "non informative".

- the parameter of interest θ is a vector of real numbers associated to P by a functional relation

$$\int \psi(x_i, \theta) P(dx_i) = 0 \tag{1.1}$$

which can be solved in a unique solution for any P (moment relation) or which determines a set of restrictions on P (generalized moment relation). The parameter of interest θ is provided by a prior probability measure deduced from the prior assumptions on P but this prior knowledge may be improved by a specific prior specification on the vector θ.

The main conclusions of our treatment of the previous statistical model are the following :

i) Using suitable representations of the Dirichlet processes, the treatment of this semi parametric statistical model is extremely simple and the computation of the posterior distribution or moments of θ may be numerically achieved even for relatively large dimension of θ and relatively tricky moment equations.

ii) The implicit information on the parameters of interest carried by the so called non informative specification on the baseline probability P is pointed out and some standard assertions like "the posterior precision obtained through a proper prior is smaller then the one following from a non informative prior", familiar in most usual parametric models becomes false in our framework.

iii) Our model provides a bayesian interpretation of the bootstrap method, in a more general way than in the Rubin (1981) paper (see also Lancaster (1992)). The bootstrap distribution of θ will appear as an approximation of the bayesian posterior distribution of θ in the case of a non informative prior specification on P. As a consequence, all the pathologies implied by the non informative prior specification are also present in the bootstrap treatment.

Our presentation is devoted to the basic model in which the prior on P is a Dirichlet process, and the prior on θ is deduced from this prior. The moment equation does not induce overidentification restriction and does not require smooth estimation of some characteristics of P as density or regression function. These different miscellaneous extensions of our basic model will be considered in a subsequent paper.

The exposition of our basic model starts by a short recall of the bayesian non parametric model in a first section. A second section describes the representations of the Dirichlet processes and gives the fundamental probabilistic results of our construction. A general numerical method is given in the third section and some examples (expectation, linear regression coefficient) are treated in the last section. Some general statements and proofs are given in an appendix.

2 Bayesian non parametric and semi parametric models

The basic statistical model we consider is simply a standard i.i.d. non parametric model provided with a Dirichlet prior process and enlarged by a functional which associates a vector of real parameters to the unknown sampling probability. The history of non parametric bayesian statistics starts with the pioneering papers of Ferguson (1973, 1974) which pointed out the role of the Dirichlet process as a natural conjugate prior specification in the non parametric model. Properties of such processes are also considered by Ferguson and in some contemporary articles (see e.g. Blackwell (1973)). Let us first recall the construction of this statistical model.

2.1 The basic non parametric model

The sample x is a sequence $(x_i)_{i=1,\cdots,n}$ of vectors $x_i \in I\!R^p$ independently generated from a probability P on $(I\!R^p, \mathcal{B}^p)$ where $\mathcal{B}^p$ is the Borelian σ-field on $I\!R^p$. The probability P is unknown and is the functional parameter of the model. This probability is an element of the set $\mathcal{P}^p$, set of all the probability measures on $\mathcal{B}^p$ and $\mathcal{P}^p$ may be considered as a subset of all functions from $\mathcal{B}^p$ to $[0,1]$. This set of functions is canonically provided with a σ-field * and the trace of this σ-field defines the σ-field on $\mathcal{P}^p$. Kolmogorov's theorem on projective systems (see Ferguson (1973)) shows that a probability Λ on $\mathcal{P}^p$ is fully characterized by the distributions of the random vectors $(P(B_\ell))_{\ell=1,\cdots,L}$ where $(B_\ell)_{\ell=1,\cdots,L}$ is a mesurable partition of $I\!R^p$. A projective condition is required: if $(B_{\ell'})_{\ell'=1,\cdots,L'}$ is an other partition obtained by regrouping elements of the previous partition $(B_{\ell'} = \bigcup_{\ell \in I_{\ell'}} B_\ell$ where $(I_{\ell'})_{\ell'=1,\ldots,L'}$ is a partition of $\{1,\cdots,L\})$ the distribution of $(P(B_{\ell'}))_{\ell'=1,\ldots,L'}$ must be identical to the distribution of $(\sum_{\ell \in I_{\ell'}} P(B_\ell))_{\ell'=1,\ldots,L'}$.

A Dirichlet process is such a distribution on $\mathcal{P}^p$ and is characterized by two elements n_0 and P_0, where n_0 is a positive real number and P_0 is a probability on $(I\!R^p, \mathcal{B}^p)$. The probability on $\mathcal{P}^p$ is then defined by the property: the random vector $(P(\mathcal{B}_\ell))_{\ell=1,\cdots,L}$ of S_L, the $L-1$ dimensional simplex in $I\!R^L$, has a Dirichlet distribution whose parameters are equal to $(n_0 P_0(B_\ell)_{\ell=1,\cdots,L})$.

The density of $P(B_\ell)$ is then equal to:

$$\Gamma(n_0) \prod_{\ell=1}^{L} \frac{P(B_\ell)^{n_0 P_0(B_\ell)-1}}{\Gamma(n_0 P_0(B_\ell))} 1\!\!1((P(B_\ell))_{\ell=1,\ldots,L} \in S_L) \tag{2.1}$$

Equivalently Wilks (1962) described the Dirichlet distribution by the property:

$$\forall a_\ell > -n_0 P_0(B_\ell) \qquad \ell = 1,\cdots,L \quad a = \sum_{\ell=1}^{L} a_\ell \tag{2.2}$$

$$E\left(\prod_{\ell=1}^{L} P(B_\ell)^{a_\ell}\right) = \frac{\Gamma(n_0)}{\Gamma(n_0+a)} \prod_{\ell=1}^{L} \frac{\Gamma(n_0 P_0(B_\ell) + a_\ell)}{\Gamma(n_0 P_0(B_\ell))} 1\!\!1(P_0(B_\ell) > 0)$$

*Let E be a set, $(F, \mathcal{F})$ a measurable space and Φ the set of all functions from E to F. The canonical σ-field on Φ is generated by the family of the sets $A_{e,J} = \{f \in \Phi : f(e) \in J\}$ for $e \in E$ and $J \in \mathcal{F}$.

If P is a random probability measure generated by a Dirichlet process whose parameters are equal to n_0 and P_0 we write:

$$P \sim \mathcal{D}i(n_0, P_0) \tag{2.3}$$

From (2.1) it follows in particular that the random variable $P(B), B \in \mathcal{B}^p$, has a Beta distribution whose parameters are equal to $n_0 P_0(B)$ and $n_0(1 - P_0(B))$. Then:

$$\begin{aligned} E(P(B)) &= P_0(B) \\ V(P(B)) &= \frac{P_0(B)P_0(B^c)}{n_0 + 1} \end{aligned} \tag{2.4}$$

The first result is summarized by $E(P) = P_0$ and the second equality shows that n_0 is a prior precision parameter.

If this Dirichlet process is taken as a prior specification of P, its follows (see Ferguson (1973)) that the posterior distribution of P is also a Dirichlet process whose parameters satisfy:

$$\begin{aligned} n_* &= n_0 + n \\ P_* &= \frac{n_0}{n + n_0} P_0 + \frac{n}{n + n_0} P_n \end{aligned} \tag{2.5}$$

where P_n is the empirical probability measure:

$$P_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i} \quad \text{where} \quad \delta_{x_i}(B) = \mathbb{1}(x_i \in B) \qquad \forall B \in \mathcal{B}^p \tag{2.6}$$

An intuitive argument which illustrates this result is the following: select a measurable partition $(B_\ell)_{\ell=1,\cdots,L}$ of $\mathbb{R}^p$ and consider as parameters the vector $(P(B_\ell))_{\ell=1,\ldots,L}$. Its prior density is given by (2.1) and its likelihood is:

$$\prod_{i=1}^{n} \prod_{\ell=1}^{L} P(B_\ell)^{\delta_{x_i}(B_\ell)} = \prod_{\ell=1}^{n} P(B_\ell)^{nP_n(B_\ell)} \tag{2.7}$$

Then the posterior distribution of $(P(B_\ell))_{\ell=1,\ldots,L}$ is a Dirichlet distribution whose parameters are n_* and $(P_*(B_\ell))_{\ell=1,\ldots,L}$. As this result is valid for any partition we obtain the previous posterior process on P.

The Dirichlet process is the natural conjugate prior distribution in the non parametric case because the posterior distribution belongs to the same family and the inference mecanism reduces to the revision, described in formulae (2.5), of the parameters defining the Dirichlet process. The observation of the sample P_n increases the precision parameter and the posterior mean is a convex combination of the prior mean P_0 and of the sufficient statistics of the sample, which is in this case the empirical probability. Moreover this result suggests a non informative prior specification: if n_0 is taken equal to 0, in the posterior distribution, P becomes distributed as a Dirichlet process whose parameters are equal to n and P_n. In particular the posterior expectation of P becomes equal to P_n, the natural estimator of P in the sampling theory framework. However this non informative specification determines some particular features of the model which will be investigated in the sequel of the paper.

In a Bayesian model, we also want the description of the predictive probability, i.e. the probability on the sample after integrating out the (functional) parameter. This predictive probability on $\mathbb{R}^{pn}$ is less known than the posterior distribution and has some specific features. We will present two alternative descriptions of this predictive distribution, assuming a proper prior specification ($n_0 > 0$).

The first one, following Blackwell and MacQueen (1973), is sequential. From (2.4) it follows that

$$x_1 \sim P_0, \tag{2.8}$$

i.e., the distribution of the first observation is the prior mean of the Dirichlet process. Then, from (2.5), we get:

$$x_{i+1}|x_1, \cdots, x_i \sim \frac{n_0}{n_0 + i} P_0 + \frac{i}{n_0 + i} P_i \qquad i > 1 \tag{2.9}$$

where $P_i = \frac{1}{i} \sum_{j=1}^{i} \delta_{x_j}$.

This distribution is a mixture of P_0 and of an atomic measure P_i and, even if P_0 is diffuse, (i.e. $P_0(\{x\}) = 0 \quad \forall x \in \mathbb{R}^p$), P_i is atomic and generates ties. For example, if P_0 is diffuse, the probability of $\{x_2 = x_1\}$ is $\frac{1}{n_0+1}$.

An alternative description of this predictive distribution is extremely useful, in particular in semi parametric analysis. We transform (see Korwar and

Hollander (1973), Antoniak (1974) or Yamato (1984)) the random sequence $(x_1, \cdots, x_n)$ in $(c, \tilde{x}_1, \cdots, \tilde{x}_m)$. The first element c is a partition of $\{1, \cdots, n\}$ and can be viewed as a configuration of ties in the $(x_1, \cdots, x_n)$. This partition has a (random) number of elements, m, and the $\tilde{x}_1, \cdots \tilde{x}_m$ are the m distinct values of the x_i's. The distribution of $(c, \tilde{x}_1, \cdots, \tilde{x}_m)$ has the following characteristics (see Rolin (1993) for a precise description) :

i) c and $(\tilde{x}_1, \cdots \tilde{x}_m)$ are independent given the number m of distinct values.

ii) the distribution of c only depends on n_0.

iii) the sample $(\tilde{x}_1, \cdots, \tilde{x}_m)$ is an iid sample from P_0.

2.2 Semi parametric specifications

A semi parametric statistical analysis may be performed in several ways. We first present the route we follow in this paper but an alternative class of models will also be briefly discussed in this section.

Our analysis will be concentrated on parameters of interest $\theta \in I\!R^\ell$ defined through a functional transformation ϕ:

$$\theta = \phi(P) \quad \phi : \mathcal{P}^p \to I\!R^\ell \tag{2.10}$$

In our presentation we assume that ϕ is defined for all probability in $\mathcal{P}^p$. This assumption avoids two types of difficulties:

i) an overidentification problem, which would be in contradiction with the previous basic non parametric model.

For example if $\theta \in I\!R$ and if we assume that θ is both the expectation and the variance of P, a constraint is introduced on the probability which is not satisfied by a realization of a Dirichlet process. This type of constraints is a central element of the generalized moments method (GMM) (see e.g. the initial paper by Hansen (1982)).

ii) a possible problem of regularity of P.

For example let us suppose that θ is related to the density of x_i with respect to Lebesgue measure (such as the expected log-density). As we will show later, a probability generated by a Dirichlet process has no density and in this case the functional ϕ could be defined on a zero measure subset of $\mathcal{P}^p$ only. A lot of very important cases are then eliminated. Some answers to this problem will be developed in a subsequent paper. Some aspects of Bayesian estimation of the density has been analysed by Lo (1984), Florens, Mouchart and Rolin (1992) and Escobar and West (1993).

In the general context of the moments method the functional ϕ is in general introduced by a moment equation whose solution defines the parameter θ:

$$\int \psi(x_i, \theta) P(dx_i) = 0 \tag{2.11}$$

where $\psi : I\!R^p \times I\!R^\ell \to I\!R^\ell$ is a known function such that (2.11) has a unique solution in θ.

An explicit solution of (2.11) is in general not necessary and the existence of a numerical algorithm computing such a solution will be sufficient. However the simplest case reduces to a relation:

$$\begin{aligned} \theta &= \int f(x_i) P(dx_i) \\ f &: I\!R^p \to I\!R^\ell \end{aligned} \tag{2.12}$$

in which θ is explicitely the expectation of a vector valued function of a sampling result. The definition of θ by (2.11) or (2.12) eliminates the regularity problem of P but not the possible overidentification, up to integrability conditions of ψ or f.

Let us remark that if $\ell = p$ the definition of θ given in (2.12) may be replaced by:

$$\theta = \int x_i P(dx_i) \tag{2.13}$$

Indeed, as f is given, the sample may be replaced by $(f(x_i))_{i=1,\cdots,n} = (y_i)_{i=1,\cdots,n}$ whose distribution is Q a priori distributed as a Dirichlet process parametrized by n_0 and $f.P_0$ $((f.P_0)(B) = P_0(f^{-1}(B)))$. However it could be interesting to analyse jointly several moments of P and, in this case, this remark does not apply.

Our objective is mainly the analysis of the prior and of the posterior distributions of θ defined through equations (2.11) or (2.12). In particular implications of non informative or low informative prior choices will be exhibited in order to shed light on the role of the sampling process in the inference.

A theoretical study of some characterizations of the distributions of functional transformations of probability generated by Dirichlet process is very tricky but a lot of important results was obtained by Cifarelli and Reggazzini (1990), Hannum, Hollander and Langberg (1981), Rolin (1992) or Yamato (1984) for example. The characterizations provided by these papers are however difficult to implement and we present here a different strategy based on representations of Dirichlet processes which may be easily used in order to generate numerically random probabilities from this distribution.

We can draw a sample from the prior or from the posterior Dirichlet process and compute from this sample a sample of θ. If θ is defined by (2.11) P takes the successive values of the sample and θ is computed by solving (numerically) equation (2.11). The last step consists in the analysis of the sample of θ which is conducted in standard frequentist way (computation of means, variances, densities, regressions...) using if necessary techniques such as kernel smoothing.

However, the usual previous description of the Dirichlet process is not suitable for such a Monte Carlo analysis and the corner stone of our paper is in the representations of the Dirichlet random measures which will be described in the next section.

This section will be concluded by some remarks on alternative semi parametric model specifications and to their bayesian treatment. A semi parametric sampling model could be the following: $(x_i)_{i=1,\cdots,n}$ is an independent sample of a probability $P^{\lambda,Q}$ on $(I\!R^p, \mathcal{B}^p)$ indexed by a parameter $\lambda \in I\!R^k$ and a probability Q on $(I\!R^q, \mathcal{B}^q)$. The probability Q is here a nuisance functional parameter and λ is the parameter of interest. In the Bayesian approach, (λ, Q) is provided with a prior distribution: λ is a priori distributed by a probability defined by its density $\mu(\lambda)$ with respect to Lebesgue measure and Q given λ is distributed as a Dirichlet process whose parameters n_0 and Q_0 depend possibly on λ.

For example one can assume that $x_i = \lambda + u_i$ where the u_i's are generated by an independent sample of Q. As we will detail later on, the mean of u_i is a random variable and the sampling expectation of x_i is the sum of λ and the mean of u_i. This construction may be considered as more "structural" than the one we explore because λ and Q naturally appear as "parameters of comportemental equations" and as "distribution of noises". However the Bayesian analysis of this type of models is sometimes disappointing. Diaconis and Freedman (1986) have also shown that bayesian estimation of λ may even be unconsistant.

Let us consider the general class of semi parametric model in which $P^{\lambda,Q} = g_\lambda.Q$ where Q is defined on $(\mathbb{R}^p, \mathcal{B}^p)$ and $g.(.)$ is a known function from $\mathbb{R}^{p+\ell}$ to $\mathbb{R}^p$. Let us recall that $g_\lambda.Q$ is the image of Q by g_λ (i.e. $(g_\lambda.Q)(B) = Q(g_\lambda^{-1}(B)))$. Moreover we assume that g_λ is a bijective transformation of $\mathbb{R}^p$ for almost all (w.r.t. the prior measure) λ.

In this case, as it has been previously remarked by Bunke (1985) (see also Ruggerio(1994) and Florens, Mouchart, Rolin (1992)), the posterior distribution of λ in the semi parametric model is identical to the posterior distribution of λ in the parametric model obtained by replacing Q by Q_0, its prior expectation, and by retaining only the distinct values of the sample.

This result follows from the description of the predictive distribution given at the end of paragraph 2.1: given λ we integrate out Q and we compute the predictive distribution given λ. This predictive probability is factorized into the configuration of ties and the distinct values of the sample. The configuration of ties is observable because g_λ is bijective and is an ancillary statistic relatively to λ. Then the inference on λ may be based on the distribution of the distinct values of the sample (given their number) which are independently $g_\lambda.Q_0$ distributed given λ.

Some semi-parametric model gives more interesting posterior distribution of λ. For example, if $p = 1$, one can assume $P^{\lambda,Q}$ characterized by $P^{\lambda,Q}([t, +\infty)) = (Q[t, +\infty))^{h_\lambda}$ where h_λ is a positive function of λ, possibly depending on explanatory variables. This type of model is in particularly used in duration analysis (see Ruggiero (1994), Hjort(1990), Florens, Mouchart and Rolin (1993)) and the posterior distribution of λ takes into account the semi parametric specification.

3 Representations of Dirichlet processes

The usual definition of the Dirichlet processes presented in section 1 is convenient to perform the transformation of the prior distribution to the posterior but more suitable characterizations of this processes are necessary if we want to generate Monte Carlo samples of the Dirichlet random measure.

All this section is based on a general representation theorem valid for any Dirichlet process. This theorem, independently proved by Sethuraman (1994) and Rolin (1993) is recalled in appendix A. We adopt here a more didactic presentation of this topic by considering successively three differents cases.

3.1 The sampling Dirichlet process

Let us first consider the Dirichlet process whose parameters are n and $P_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}$. This distribution is obtained as a posterior distribution if the parameter n_0 of the prior is taken equal to 0. A deeper discussion about the non informative prior will be conducted later on and we just examine here the properties of the posterior. Let us denote by P_S a realisation of the $\mathcal{D}i(n, P_n)$ distribution (the subscript S is for sample). Equivalently:

$$P_S|x \sim \mathcal{D}i(n, P_n) \tag{3.1}$$

In general, the values x_i may not be distinct and ties exist in the sample. Let $\tilde{x}_j$ $(j = 1, \cdots, m)$ the distinct values of the x_i and n_j the number of occurences of $\tilde{x}_j$. This implies $\sum_{j=1}^{m} n_j = n$ and the empirical sampling probability may be written as:

$$P_n = \frac{1}{n}\sum_{j=1}^{m} n_j\delta_{\tilde{x}_j} \tag{3.2}$$

This probability is then a finite measure whose support is $\{\tilde{x}_1, \cdots, \tilde{x}_m\}$ and probability mass are $\{\frac{n_1}{n}, \cdots, \frac{n_m}{n}\}$. The sample is considered as fixed because we analyse the posterior process. Then its follows from theorem A1 that:

Proposition 3.1 *P_S may be written as:*

$$P_S = \sum_{j=1}^{m} \beta_j\delta_{\tilde{x}_j} \tag{3.3}$$

where the $\delta_{\tilde{x}_j}$ are the Dirac measures on the $\tilde{x}_j$ and the vector $(\beta_j)_{j=1,\cdots,m}$ is (generated by) a Dirichlet distribution on S_m whose parameters are $(n_j)_{j=1,\cdots,m}$. The representation is almost sure because $\beta_j = P_S(\{\tilde{x}_j\})$.
However, as far as the distribution of P_S only is concerned, P_S has the same distribution as $\sum_{i=1}^{n}\gamma_i\delta_{x_i}$ where the vector $(\gamma_i)_{i=1,\cdots,n}$ is uniformly distributed

on the simplex S_n and we will write

$$P_S = \sum_{i=1}^{n} \gamma_i \delta_{x_i} \tag{3.4}$$

This representation becomes almost sure if the values of the sample are distinct.

■

Details on the uniform distribution on the simplex and its connexion with the Dirichlet distribution may be found in Devroye (1986).

3.2 The prior Dirichlet process

The second case we want to detail is the prior Dirichlet process whose parameter P_0 is assumed to be diffuse ($P_0(\{x\}) = 0 \quad \forall x \in I\!R^p$). Let us denote P_A (A for a priori) a random realisation of the prior process or equivalently:

$$P_A \sim \mathcal{D}i(n_0, P_0), \tag{3.5}$$

It was remarked by Ferguson (1974) that the Dirichlet process generates almost surely a totally discrete probability even if P_0 is diffuse. A realization P_A has a countable support dense in the support of P_0 and P_A has an infinite countable number of jumps.

More precisely the following proposition characterizes the probabilities generated by Dirichlet processes and is also a consequence of theorem A1.

Proposition 3.2 *Under (3.5) and if P_0 is diffuse, P_A may be written as:*

$$P_A = \sum_{k=1}^{\infty} \alpha_k \delta_{\xi_k} \qquad a.s. \tag{3.6}$$

where:

- *the sequences $(\alpha_k)_{k=1,\cdots}$ and $(\xi_k)_{k=1,\cdots}$ are independent.*
- *$(\xi_k)_{k=1,\ldots}$ is an iid infinite sample of P_0*
- *the distribution of α_k may be described by the following representation:*

$$\alpha_k = v_k \prod_{h=1}^{k-1} (1 - v_h) \tag{3.7}$$

where $(v_h)_{h=1,\cdots}$ is an infinite iid sample of a Beta distribution whose parameters are equal to 1 and n_0 ■

The representation (3.6) is also valid in distribution with arbitrary P_0 (see Sethuraman (1994)) but the almost sure equality requires P_0 diffuse as shown by Rolin (1993).

The representation (3.6) had been noticed by Ferguson (1974) but only the distribution of the order statistics of the α_k was given in this paper. The representation (3.7) of the random jumps sizes is particularly convenient and simple.

Some troubles follow from the fact that the random probability drawn from the prior is almost surely discrete when such a property is in general not required for the sampling probability.

However it was proved by Ferguson (1973) that the closed set generated by all the trajectories of a Dirichlet process contains all the probabilities dominated by P_0. The concept of closed set is, in this result, related to punctual convergence topology on the set of probabilities ($Q_n \to Q$ if $\quad \forall B \in \mathcal{B}^p, \quad Q_n(B) \to Q(B)$).

One of the most remarkable feature of the representation (3.6) is the fact that the locations of the jumps (the ξ_k's) only depend on P_0 and the sizes of the jumps (the α_k's) only depend on n_0.

In view of Monte Carlo experimentation it is natural to consider approximations of the representation (3.6).

Let us define:

$$P_A^K = \sum_{k=1}^{K} \alpha_k \delta_{\xi_k} \tag{3.8}$$

where the α_k's and the ξ_k's are generated in the same way as in proposition 3.2.

The random measures P_A^K are not probabilities (because $P_A^K(\mathbb{R}^p) < 1$) and are not generated by a Dirichlet process. However, theorem A2 provides the distribution of the approximation error and gives the density and the moments of $\varepsilon_K = \sup_{B \in \mathcal{B}^p} |P_A(B) - P_A^K(B)| = \sum_{k>K} \alpha_k$.

This distribution only depends on n_0 and K and, in particular, the mean and the variance of ε_K satisfy:

$$\begin{aligned} E(\varepsilon_K) &= \left(\frac{n_0}{n_0+1}\right)^K \\ V(\varepsilon_K) &= \left(\frac{n_0}{n_0+1}\right)^{2K} \left(\left(1+\frac{1}{n_0(n_0+2)}\right)^K - 1\right). \end{aligned} \tag{3.9}$$

The expectation and the variance go to 0 when $n_0 \to 0$ or when $K \to +\infty$. For a given value of n_0 it is simple to select K in order to obtain an approximation with a suitable precision.

Figure 1 illustrates the behavior of approximated Dirichlet processes. We generate realizations of Dirichlet processes with the same mean P_0 chosen equal to the $N(0,1)$ distribution and with different values of n_0 (from 0.1 to 50). Different values of K are selected (from 50 to 300) in connection with n_0. The random distribution functions and the normal distribution function are reported in this figure.

An alternative approximation to P_A would be $\bar{P}_A^K$ defined by:

$$\bar{P}_A^K = \frac{1}{\sum_{k=1}^{K} \alpha_k} \sum_{k=1}^{K} \alpha_k \delta_{\xi_k}. \tag{3.10}$$

The interest of $\bar{P}_A^K$ follows from the fact that $\bar{P}_A^K$ are random probabilities for any K. As shown by corollary A3, the approximation error is the same as before, i.e., $\sup_{B\in\mathcal{B}^p} |P_A(B) - \bar{P}_A^K(B)| = \sum_{k>K} \alpha_k = \varepsilon_K$.

This remark sheds light on the meaning of low informative or non informative prior specification in the framework of Dirichlet processes. In formula (3.10) let us select $K = 1$. Then $\bar{P}_A^K$ becomes δ_{ξ_1}, the Dirac probability measure concentrated on a random value δ_{ξ_1} drawn from the probability P_0. If $n_0 \to 0$, the limit of P_A is δ_{ξ_1} in the previous sense ($\sup_B |P_A(B) - \delta_{\xi_1}(B)| \to 0$ in L^2). If the precision parameter n_0 is taken very small the prior Dirichlet process is approximatively identical to a random Dirac measure. This result may be intuitively deduced from proposition 3.2: if $n_0 \to 0$ the distribution of α_1 concentrates on 1 and the distributions of $\alpha_k (k > 1)$ concentrates on 0. However the choice of δ_{ξ_1} as a prior specification generates paradoxes: if the sampling probability P is drawn by such a prior all the observed values of the sample must be almost surely equal to ξ_1, a property clearly incompatible with the data sets.

Figure 1: Some cumulative distribution functions of Dirichlet prior process
$(P_0 = N(0,1))$

Figure 1: Some cumulative distribution functions of Dirichlet prior process ($P_0 = N(0,1)$)

Figure 1: Some cumulative distribution functions of Dirichlet prior process ($P_0 = N(0, 1)$)

3.3 The posterior Dirichlet process

The last case examined here is the case of the posterior Dirichlet process deduced from a proper prior distribution. Let us denote by P_P (P for posterior) a realization of the posterior process:

$$P_P|x \sim \mathcal{D}i(n_*, P_*) \tag{3.11}$$

A last application of theorem A1 gives the following result:

Proposition 3.3 *Under (3.11) and if P_0 is a diffuse probability, P_P may be written as:*

$$P_P = (1-\gamma)P_A + \gamma P_S \qquad a.s. \tag{3.12}$$

where:

- *P_A and P_S are drawn independently according to propositions 3.1 and 3.2.*
- *γ is independent from P_A and P_S and is a Beta random variable whose parameters are equal to n and n_0.*

■

Let us remark that $P_P|x$ could be represented in distribution as in proposition (3.2) with n_0, P_0 replaced by n_* and P_*. However the decomposition (3.12) of proposition 3.3 exhibits an inference mecanism based on Bayes theorem in terms of representation function of random quantities instead of densities of probability measures. This result is illuminating and shows that a probability drawn from the posterior process is a random convex combination of two random probabilities generated by the prior process and by the sampling Dirichlet process defined in section 3.1 and 3.2.

Approximations presented in the section 3.2 may be used for the posterior process and P_P may approximated by $P_P^K = (1-\gamma)P_A^K + \gamma P_S$ or by $\bar{P}_P^K = (1-\gamma)\bar{P}_A^K + \gamma P_S$.

The posterior approximation error is equal to the prior approximation error multiplied by $(1-\gamma)$, which is a $\mathcal{B}(n_0, n)$ random variable, independent from the prior error: the quality of the posterior approximation is then better than the quality of the prior approximation.

4 Prior and Posterior distribution of parameters of interest: the Bayesian Bootstrap

Our main interest is the inference on the parameter of interest θ related to the sampling probability P by a relation whose general form is $\theta = \phi(P)$ and the objective is the analysis of the different distributions of θ deduced from different processes generating P : the so-called sampling process, the prior process and the posterior process. Our framework is numerical: we propose to generate Monte Carlo samples $P_S^e, P_A^e, P_P^e (e = 1, \cdots, N)$ of the three Dirichlet processes $\mathcal{D}i(n, P_n), \mathcal{D}i(n_0, P_0)$ and $\mathcal{D}_i(n_*, P_*)$ and to compute three samples of θ :

$$\begin{array}{ll} \text{i)} & \theta_S^e = \phi(P_S^e) \\ \text{ii)} & \theta_A^e = \phi(P_A^e) \\ \text{iii)} & \theta_P^e = \phi(P_P^e), \qquad e = 1, \cdots, N. \end{array} \tag{4.1}$$

We call *Bayesian bootstrap* the methods based on Monte Carlo sampling of Dirichlet processes in order to derive distributions of functionals of this processes. From this sample all the statistics of interest may be obtained: moments, confidence intervals, distribution functions, marginal or conditional densities (through a smoothing technique). It could objected that "classical" or "frequentist" procedures are used in a Bayesian analysis but this tools are used in the treatment of the Monte Carlo sample (typically as large as we want under the restriction of computer time) only.

4.1 Posterior distribution on θ under non informative prior

Let us first examine the generation of a sample through the $\mathcal{D}i(n, \mathcal{P}_n)$ process. The result given in proposition 3.1 provides an efficient way to draw the sample P_S^e: we draw a sample $(\beta_i^e)_{i=1,\cdots,n} (e = 1, \cdots, N)$ of a Dirichlet distribution on S_n whose parameters equal to 1 or equivalently from a uniform distribution on S_n.

Then P_S^e is equal to $\sum_{i=1}^n \beta_i^e \delta_{x_i}$ and, if θ is defined through relation (2.11), we get a sample θ_S^e by solving:

$$\int \psi(x_i, \theta) P_S^e(dx_i) = \sum_{i=1}^{n} \beta_i^e \psi(x_i, \theta) = 0 \tag{4.2}$$

In particular if $\theta_S = \int f dP_S$ this relation reduces to:

$$\theta_S^e = \sum_{i=1}^{n} \beta_i^e f(x_i). \tag{4.3}$$

This means that the non parametric, non informative posterior distribution of θ is a stochastic convex combination of the $f(x_i)$ whose coefficients are uniformly distributed on the the simplex S_n.

This presentation extends Rubin's (1981) definition of Bayesian Bootstrap and includes this methodology in the framework of bayesian non parametric inference.

The posterior distribution of θ determined by (4.2) or (4.3) is almost identical to the bootstrap distribution in which the β_i^e are replaced by $\tilde{\beta}_i^e$ equal to the frequency of the index i when n draws among $\{1, \cdots, n\}$ with equal probabilities are performed. Equivalently the vector $(n\tilde{\beta}_i^e)$ is generated by a multinomial distribution whose parameters are equal to $(n, \frac{1}{n}, \cdots, \frac{1}{n})$. When n is large the multinomial distribution divided by n is approximatively Dirichlet as discussed for example in Johnson and Kotz (1989 vol I chap II section 2.4). However it is useful to note that the Dirichlet distribution which approximates the distribution of the Bootstrap coefficient with the same two moments is the distribution whose parameters are equal to $\frac{n-1}{n}$ and not to 1. Thus, excepted for the mean, moments of $\tilde{\beta}_i^e$ and of β_i^e are different, at least for small n.

The distribution of θ deduced from (4.1) i) (or in particular cases from (4.2) or (4.3)) will be called the *Bayesian sampling bootstrap distribution of* θ. (BSB distribution).

As we said above, this distribution is in general analyzed numerically through a Monte Carlo sample of θ.

In some particular cases, some statistics may be computed analytically. If for example $\theta_S = \int f dP_S$ we get :

$$E(\theta_S|x) = \frac{1}{n} \sum_{i=1}^{n} f(x_i) = E_S(f) = \bar{f} \tag{4.4}$$

$$V(\theta_S|x) = \frac{1}{n+1} V_S(f) \tag{4.5}$$

where $V_S(f)$ is the empirical variance matrix of the $(f(x_i))_{i=1,\cdots,n}$. This last result follows from the computation of the variance matrix of the uniform distribution on the simplex which is equal to

$$V(\beta) = \frac{1}{n(n+1)} (I - \frac{1}{n} \mathbb{1}\mathbb{1}') \tag{4.6}$$

where I is the identity matrix and $\mathbb{1}$ the vector composed by n elements equal to 1.

A comparison with the classical bootstrap moments should be illuminating: if $(n\tilde{\beta}_i)_{i=1,\ldots,n}$ is generated from a multinomial distribution whose parameters are equal to $(n, \frac{1}{n}, \cdots, \frac{1}{n})$ the bootstrap parameter

$$\tilde{\theta} = \sum_{i=1}^{n} \tilde{\beta}_i f(x_i)$$

has expectation and variance equal to

$$E(\tilde{\theta}|x) = E_S(f) = \bar{f}$$

$$V(\tilde{\theta}|x) = \tfrac{1}{n} V_S(f).$$

The representation (4.3) suggests an updating formula: dropping out for simplicity the index e we get the following decomposition.

$$\theta_S^{(n)} = \beta_n f(x_n) + (1-\beta_n)\theta_S^{(n-1)} \tag{4.7}$$

where $\theta_S^{(n)}$ is a draw of θ given the sample of size n and $\theta_S^{(n-1)} = \sum_{1 \leq i \leq n-1} \rho_i f(x_i)$ where $\rho_i = \beta_i/(1-\beta_n)$.

It follows from the properties of the Dirichlet distribution that β_n is $B(1, n-1)$ distributed and is independent of $\theta_S^{(n-1)}$.

A recursive relation is deduced from (4.7) for the densities of θ_S. If the number of distinct values of the sample is greater than 1 (in the scalar case and to the dimension of θ in the vector case), the distribution of θ_S is dominated by Lebesgue measure (for example if the first two values are different, θ_S is uniform on the interval they define) and if $\mu^{(n)}$ is the density of $\theta_S^{(n)}$ we have :

$$\mu^{(n)}(\theta) = (n-1)\int_0^1 (1-\beta)^{n-2-\ell}\mu^{(n-1)}\Big(\frac{\theta - \beta f(x_n)}{1-\beta}\Big)d\beta \tag{4.8}$$

Finally equation (4.7) gives a recursive relation for the moments : In case $\ell = 1$,

$$E[(\theta_S^{(n)})^r] = \frac{(n-1)r!}{(n+r-1)!}\sum_{j=0}^{r}\frac{(n+j-2)!}{j!}f(x_n)^{r-j}E[(\theta_S^{(n-1)})^j] \tag{4.9}$$

If θ_S is the solution of (4.2), its distribution is dominated by Lebesgue measure if the number of distinct values of the sample is greater than the dimension of θ. This property requires smoothness conditions on ψ such that θ_S is a differentiable function of the vector β.

4.2 Prior distribution of θ

Our analysis of the prior distribution of θ will be essentially numerical but some general properties fo this distribution are easily deduced from the representation theorem given in proposition 3.2. We assume P_A generated by a Dirichlet process where P_0 is diffuse and θ_A is related to P_A by the relation (1.1). Using proposition 3.2, P_A is equal to $\sum_k \alpha_k \delta_{\xi_k}$ and θ is defined as the solution of

$$\sum_{k=1}^{\infty} \alpha_k \psi(\xi_k, \theta) = 0 \tag{4.10}$$

It is then natural to analyse the convergence property of the left hand side of this equation. This convergence is obtained by the following proposition:

Proposition 4.1 *Under the assumptions of proposition 3.2 and if*

$$\int |\psi(\xi_k,\theta)| P_0(d\xi_k)$$

is finite for any θ, the series $\sum_{k=1}^{\infty} \alpha_k \psi(\xi_k, \theta)$ converges almost surely and in L^1, and then $\int \psi(\xi_k,\theta)P(d\xi_k)$ is almost surely finite for any θ. As will be seen in the appendix (theorem A.6), under some regularity conditions, the distribution of θ is dominated by the Lebesgue measure. ■

In particular if $\theta_A = \int f dP_A$ the integrability of f with respect to P_0 implies the integrability of f with respect to almost all probability generated by the Dirichlet process. In this specific case, more can be said on the distribution of θ. This distribution was previously analysed in the litterature (see e.g. Cifarelli and Reggazzini (1990) or Yamato (1984)) but the representation given in proposition 3.2 implies some useful properties.

Proposition 4.2 *Under the assumptions of proposition 3.2 let f be a vector valued P_0 integrable function and let $\eta_k = f(\xi_k), k = 1 \cdots,$ and $\theta_A = \int f dP_A$. Then:*
i) θ_A is almost surely finite and its distribution is dominated by Lebesgue measure.
ii) θ_A satisfies the following equation :

$$\theta_A = \alpha_1 \eta_1 + (1 - \alpha_1)\lambda_A \tag{4.11}$$

where

- *α_1, λ_A and η_1 are independent*
- *α_1 is Beta distributed with parameters $(1, n_0)$*
- *η_1 is $Q_0 = f P_0$ distributed*
- *θ_A and λ_A have the same distribution.*

iii) The density μ_A of θ_A w.r.t. Lebesgue measure satisfies the following integral equation:

$$\begin{aligned}
\mu_A(\theta) &= \int_{\mathbb{R}^\ell} Q_0(d\eta) \int_0^1 \mu_A \left[\frac{1}{1-\alpha}\theta - \frac{\alpha}{1-\alpha}\eta \right] n_0 (1-\alpha)^{n_0 - \ell - 1} d\alpha \\
&= E \int_0^1 \mu_A \left[\frac{1}{1-\alpha}\theta - \frac{\alpha}{1-\alpha}\eta_1 \right] n_0 (1-\alpha)^{n_0 - \ell - 1} d\alpha \\
&= E \int_0^\infty \mu_A \left[\theta + u(\theta - \eta_1) \right] n_0 (1+u)^{-(n_0 - \ell + 1)} du
\end{aligned}$$

(iii)' If the distribution of η_1 is dominated by the Lebesgue measure with density $q_0(\eta)$, then the density μ_A of θ_A w.r.t. Lebesgue measure satisfies the following integral equation :

$$\mu_A(\theta) = \int_{\mathbb{R}^\ell} K(\theta,t)\mu_A(t)dt$$

where

$$\begin{aligned} K(\theta,t) &= \int_0^1 q_0\left[t + \frac{1}{\alpha}(\theta - t)\right] n_0 \frac{(1-\alpha)^{n_0-1}}{\alpha^\ell} d\alpha \\ &= \int_1^\infty q_0\left[t + \lambda(\theta - t)\right] n_0 \frac{(\lambda-1)^{n_0-1}}{\lambda^{n_0-\ell+1}} d\lambda \end{aligned}$$

(iii)" In the scalar case, i.e. $\ell = 1$, *then*

$$\begin{aligned} K(\theta,t) &= E\left[n_0 \frac{(\eta_1-\theta)^{n_0-1}}{(\eta_1-t)^{n_0}} 1\{\eta_1 > \theta\}\right] \quad \textit{if} \quad \theta > t \\ &= E\left[n_0 \frac{(\theta-\eta_1)^{n_0-1}}{(t-\eta_1)^{n_0}} 1\{\eta_1 < \theta\}\right] \quad \textit{if} \quad \theta < t \end{aligned}$$

(iv) In the scalar case, i.e. $\ell = 1$, *if* f *is* m*-integrable w.r.t.* P_0*, the prior moments of* θ_A *satisfies for all* $r \leq m$*, the recurrence relation*

$$E(\theta_A^r) = \frac{n_0\Gamma(r)}{\Gamma(n_0+r)} \sum_{j=0}^{r-1} \frac{\Gamma(n_0+j)}{\Gamma(j+1)} E\left[\eta_1^{r-j}\right] E\left[\theta_A^j\right]$$

■

For example we get :

$$E(\theta_A) = E(\eta_1) = \int f dP_0 = E_0(f)$$

$$V(\theta_A) = \frac{1}{n_0+1} V(\eta_1) = \frac{1}{n_0+1} V_0(f)$$

The expectation of θ_A does not depend on n_0 and $V_0(f)$ is an upper bound for the prior variance of θ_A.

The relation given in (ii) and (iii)' can be interpreted in the following way : the expression $\alpha_1\eta_1 + (1-\alpha_1)\lambda_A$ defines a conditional distribution of θ_A given λ_A which can be viewed as a Markov kernel whose density is equal to $K(\theta,t)$ and the density of θ_A is the density of the invariant distribution of this Markov kernel :

$$\mu_A(\theta) = \int_{\mathbb{R}^\ell} K(\theta, t)\mu_A(t)dt$$

We omit the proof of proposition 4.2 since it is a simple matter of computation.

Let us now consider numerical treatment of the prior distribution on θ.

As it was remarked in section 3 only aproximations P_A^K of the prior process may be drawn. Let $(P_A^{K,e})_{e=1,\cdots,N}$ an iid Monte Carlo sample of the approximated prior, i.e.

$$P_A^{K,e} = \sum_{k=1}^{K} \alpha_k^e \delta_{\xi_k^e} \tag{4.12}$$

where the ξ_k^e's are independently generated from P_0 and the α_k^e's satisfy:

$$\alpha_k^e = v_k^e \prod_{h=1}^{k-1}(1 - v_h^e) \tag{4.13}$$

with v_h^e is an iid sample of the $\mathcal{B}(1, n_0)$ distribution. Then, if θ is defined by (1.1), a sample $(\theta_A^{K,e})_{e=1,\ldots,N}$ is obtained by solving for any e the equation:

$$\sum_{k=1}^{K} \alpha_k^e \psi(\xi_k^e, \theta) = 0 \tag{4.14}$$

The distribution on θ obtained in this way will be called the *Bayesian prior bootstrap(BPB) distribution of* θ.

Corollary A4 shows that, under usual regularity conditions on $\psi, \theta_A^{K,e}$ converges in distribution to θ_A^e defined in (4.1) ii) for any $n_0 > 0$ when $K \to \infty$ or for any K when $n_0 \to 0$. This result proves in particular (take $K = 1$) that the prior distribution of θ converges to the distribution of θ^0, solution of

$$\psi(\xi_1, \theta) = 0 \tag{4.15}$$

when $n_0 \to 0$. In (4.15) ξ_1 is a one dimensional sample of P_0. In particular if $\theta = \int f dP$, the prior distribution of θ is almost identical, for small n_0 to the prior distribution of $f(\xi_1)$ where ξ_1 is P_0 distributed. Then a very low informative prior distribution on P, defined by a very small value of n_0, induces a prior distribution on θ which can be very concentrated. A statistician who has no prior information on P and on θ must specified his prior Dirichlet process by selecting a small value of n_0 and a probability P_0 generating a large variance distribution on θ^0.

4.3 Posterior distribution of θ

The last Dirichlet process which may be sampled is the posterior process obtained through a proper prior on the unknow probability.

If prior and sampling processes were previously used to draw $P_A^{e,K}$ and $P_S^e(e = 1, \cdots, N)$ it is sufficient, from proposition 3.3, to generate independently N independent random numbers γ^e from a Beta (n, n_0) distribution and to compute:

$$\overline{P}_P^{K,e} = (1-\gamma^e)\overline{P}_A^{K,e} + \gamma^e P_S^e \tag{4.16}$$

Only approximated sample of the posterior process are computed. The approximation error is $(1-\gamma^e)\varepsilon_K^e$, i.e., the prior approximation error multiplied by $(1-\gamma^e)$. If θ is characterized by (1.1), an (approximated) posterior sample of θ is given by solving the equation:

$$(1-\gamma^e)(\sum_{k=1}^K \alpha_k^e)^{-1} \sum_{k=1}^K \alpha_k^e \psi(\xi_k^e, \theta) + \gamma^e \sum_{i=1}^n \beta_i^e \psi(x_i, \theta) = 0 \tag{4.17}$$

As in the sampling and in the prior cases, the distribution of θ obtained through the generation of the $\theta_P^{K,e}$ solutions of (4.17) is the *Bayesian posterior bootstrap distribution of* θ.

In the case where θ is defined by $\theta = \int f dP$, the inference process on the parameter may be represented by the relation:

$$\theta_P = (1-\gamma)\theta_A + \gamma\theta_S \qquad \text{a.s.} \tag{4.18}$$

For simplicity we drop out the index e and we neglect the approximation problem. Let us recall that this relation means that a draw θ_P of the posterior distribution of θ is almost surely equal to a stochastic convex combinaison of a draw θ_A of the prior probability measure and a draw θ_S of a sampling measure on θ detailled in section 4.1. The coefficient of this combination is generated by a Beta distribution whose parameters reflect the weights of the prior knowledge and of the sampling mecanism in the inference learning. The relation (4.18) is derived from the application of Bayes theorem on the non parametric model but in practice it can be used by itself and is the foundation of a Bayes revision of knowledge without specification of underlying distributions.

In particular (4.18) determines a relation between moments. In the scalar case we obtain:

$$E(\theta_P^r|x) = \sum_{q=0}^{r} a(q,r)E(\theta_A^{r-q})E(\theta_S^q|x) \tag{4.19}$$

where

$$a(q,r) = \frac{\Gamma(n+n_0)}{\Gamma(n+n_0+r)}\frac{\Gamma(n+q)}{\Gamma(n)}\frac{\Gamma(n_0+r-q)}{\Gamma(n_0)}\frac{\Gamma(r+1)}{\Gamma(q+1)\Gamma(r-q+1)} \tag{4.20}$$

The posterior moment of order r of θ is a linear combination of products of the prior moments of order $r-q$ and the sampling moment of order q. Computation of coefficients $a(q,r)$ follows from Newton's formula and from the expression of $E(\gamma^q(1-\gamma)^{r-q})$. Application of (4.19) to $r=1$ gives first the natural relation between the different expectations:

$$\begin{aligned} E(\theta_P|x) &= \frac{n_0}{n+n_0}E(\theta_A) + \frac{n}{n+n_0}E(\theta_S|x) \\ &= \frac{n_0}{n+n_0}E_0(f) + \frac{n}{n+n_0}\bar{f} \end{aligned} \tag{4.21}$$

if $r=2$ we obtain, in the vector case, the following relation between the variances:

$$\begin{aligned} V(\theta_P|x) = \frac{1}{(n+n_0)(n+n_0+1)}\Big[&n_0(n_0+1)V(\theta_A) + n(n+1)V(\theta_S|x) \\ &+\frac{nn_0}{n+n_0}(E(\theta_S|x)-E(\theta_A))(E(\theta_S|x)-E(\theta_A))'\Big] \end{aligned}$$

$$= \frac{1}{(n+n_0)(n+n_0+1)}\left[n_0V_0(f) + nV_S(f) + \frac{nn_0}{n+n_0}(\bar{f}-E_0(f))(\bar{f}-E_0(f))'\right] \tag{4.22}$$

It could happen that the posterior variance $V(\theta_P|x)$ may be greater than the sampling variance $V(\theta_S|x)$ in particular if the sampling mean $\bar{f}$ is very different from the prior mean $E_0(f)$.

Let us conclude this section by an other look to the non informative case. If n_0 is very small the inference mecanism reveals a paradox: the prior distribution on θ is a well defined proper prior approximatively equal to the image of P_0 by f but the posterior distribution on this parameter is approximatively identical to the distribution of θ_S and then does not depend on the prior specification.

5 Numerical examples

In this section previous analysis will be illustrated by two numerical examples. The first one treats the case of the estimation of the mean of a univariate distribution. This problem is a generic example for all cases in which $\theta = \int f dP$. The second example analyses the linear regression coefficient in a bivariate model: the moment equation characterizing θ has an explicit solution but the solution is a non linear function of the moments. All computations have been performed with GAUSS I on a Compaq 4/25CX in less than one minute per item.

5.1 Non parametric estimation of the mean

The sample $(x_i)_{i=1,\cdots,n}$ is generated by a $N(0,1)$ distribution and three samples are selected: $n = 30, 50$ or 100. Empirical mean $(\bar{x})$ and variance (s^2) are computed.

We first consider the sampling Dirichlet process: a Monte Carlo sample β^e of size $N = 1000$ of a uniform distribution on the simplex S_n is drawn and this sample generates a sample θ_S^e of the distribution on $\theta = \int x_i P(dx_i)$ deduced from the sampling process $\mathcal{D}i(n, P_n)$. This sample is computed by

$$\theta_S^e = \sum_{i=1}^{n} \beta_i^e x_i \tag{5.1}$$

From $(\theta_S^e)_{e=1,\ldots,N}$ are computed mean, variance and density (using a $N(0,1)$ Kernel and a bandwith equal to $N^{-\frac{1}{5}}$ multiplied by the standard deviation of the sample).

The results are given in Table 1 and some density estimation are given in Figure 2.

Table 1 also gives the exact mean ($\bar{x}$) and variance of the distribution of θ_S given x (see section 4.1 and equal to $\frac{1}{n+1}s^2$) and the mean and variance of the posterior distribution of θ in a correctly specified parametric model:

$$x_i|\theta \sim iN(\theta, 1) \tag{5.2}$$

provided with a uniform prior measure on $I\!R$, wich entails.

$$\theta|x \sim N[\overline{x}, \frac{1}{n}] \tag{5.3}$$

Table 1. Non Informative Means and Variances of the Mean

	Parametric Model	Theoretical Non Parametric Model	Simulated Non Parametric Model
$n = 30$	$\overline{x} = -0,21572$	$s = 1,06246$	
Means Variances	-0,21572 0,03333	-0,21572 0,03520	-0,21577 0,03615
$n = 50$	$\overline{x} = 0,16049$	$s = 1,14469$	
Means Variances	0,16049 0,02000	0,16049 0,02518	0,16859 0,02730
$n = 100$	$\overline{x} = 0,03607$	$s = 1,03413$	
Means Variances	0,03607 0,01000	0,03607 0,01048	0,03961 0,01033

A second step of the simulation exercice is devoted to the prior distribution of θ deduced from a Dirichlet process $\mathcal{D}i(n_0, P_0)$. We have chosen $n_0 = 20$ and P_0 equal to $N(2,2)$ and the approximation level K has been chosen equal to 100. Samplings of $(\xi_k^e)_{\substack{e=1,\dots,N,\\ k=1,\dots,K}}$ and of $(v_k^e)_{\substack{e=1,\dots,N,\\ k=1,\dots,K}}$ have been generated according to proposition 3.2. Finally, a sampling $(\gamma^e)_{e=1,\dots,N}$ from a $Be(n, n_0)$ distribution has been generated and we have computed

$$\theta_P^{K,e} = (1-\gamma^e)\theta_A^{K,e} + \gamma^e \theta_S^e \tag{5.4}$$

Theoretical and simulated means and variances are computed in the prior, sampling and posterior analyses. These results are given in Table 2 and kernel smoothing of the densities of θ are plotted in Figure 2.

Table 2. Bayesian Theoretical and Simulated Means and Variances of the Mean

	Prior	Sampling	Posterior
$n = 30$	$\overline{x} = -0,21572$	$v^2 = 1,09120$	
Theoretical Mean	2,000002	-0,21572	0,67057
Simulated Mean	1,99321	-0,21577	0,66963
Theoretical Variance	0,09521	0,03520	0,05162
Simulated Variance	0,10026	0,03615	0,05195
$n = 50$	$\overline{x} = 0,16049$	$v^2 = 1,28411$	
Theoretical Mean	2,00000	0,16049	0,68607
Simulated mean	1,98277	0,16859	0,69138
Theoretical Variance	0,09521	0,02518	0,03069
Simulated Variance	0,09649	0,02730	0,03324
$n = 100$	$\overline{x} = 0,03607$	$v^2 = 1,05872$	
Theoretical Mean	2,00000	0,03607	0,36339
Simulated Mean	1,98744	0,03961	0,36362
Theoretical Variance	0,09521	0,01048	0,01447
Simulated Variance	0,10122	0,01033	0,01384

Figure 2

Figure 2

Figure 2

5.2 Non parametric estimation of the regression coefficient

The sample $x_i = (y_i, z_i)$, $i = 1, \ldots, n$ is generated by a $N_2 \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix} ; \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix} \right]$ distribution with $n = 50$. Now

$$\psi(x_i, \theta) = y_i z_i - \theta z_i^2 \tag{5.5}$$

so that

$$\theta_P^{K,e} = \int_{I\!R^2} y_i z_i F_P^{K,e}(dy_i, dz_i) / \int_{I\!R^2} z_i^2 F_P^{K,e}(dy_i, dz_i) \tag{5.6}$$

where

$$F_P^{K,e} = (1 - \gamma^e)(\sum_{k=1}^{K} \alpha_k^e)^{-1} \sum_{k=1}^{K} \alpha_k^e \delta_{(\eta_k^e, \zeta_k^e)} + \gamma^e \sum_{i=1}^{n} \beta_i^e \delta_{(y_i, z_i)} \tag{5.7}$$

and the true value of θ is equal to one. In the sampling Dirichlet process, ($\gamma^e = 1$ in (5.7)), we have drawn as before a sample β_i^e of size $N = 1000$ and computed θ_S^e. From this last sample of $\mathcal{D}i(nF_n)$ are computed mean, variance and density using the same kernel as in section 5.1.. This may be compared with the posterior distribution of θ in a correctly specified parametric model:

$$x_i | \theta \sim N_2 \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix} ; \begin{pmatrix} 1 & \theta \\ \theta & 1 + \theta^2 \end{pmatrix} \right] \tag{5.8}$$

provided with a uniform prior measure on $I\!R$, i.e.

$$\theta | x \sim N \left[\frac{\overline{yz}}{\overline{z^2}}, \frac{1}{n\overline{z^2}} \right] \tag{5.9}$$

where the empirical statistics

$$\begin{aligned} \overline{yz} &= \frac{1}{n} \sum_{i=1}^{n} y_i z_i \\ \overline{z^2} &= \frac{1}{n} \sum_{i=1}^{n} z_i^2 \end{aligned} \tag{5.10}$$

have been computed. These results are reported in Table 3.

For the prior Dirichlet process $\mathcal{D}i(n_0, P_0)$ ($\gamma^e = 0$ in (5.7)) we have choosen $n_0 = 20$ and P_0 equal to $N_2 \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix} ; \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right]$ and $K = 100$ as before. Monte-Carlo samples $\xi_k^e = (\eta_k^e, \zeta_k^e)$ have be drawn from P_0 and samples of α_k^e and γ^e have been drawn as in Section 5.1. Using (5.6) and (5.7), samples of $\theta_A^{K,e}$ and $\theta_P^{K,e}$ have been computed and from these samples, means variances and densities. These results are summarized in Table 3 and kernel smoothing of the densities of θ are plotted in Figure 3.

Table 3. Posterior Sampling Non parametric and Parametric Means and Variances of the regression coefficient

$\overline{yz} = 1,02397 \qquad \overline{z^2} = 1,11386 \qquad n = 50$

	Means	Variances
Parametric	0,91930	0,01796
Sampling	0,91376	0,01352
Prior	-0,01035	0,04450
Posterior	0,67098	0,01551

APPENDIX I : Some properties of Dirichlet processes

1. Trajectories of Dirichlet processes

Let P be a Dirichlet process on $(\mathbb{R}^p, \mathcal{B}^p)$ i.e., $P \sim \mathcal{D}i(\nu\Pi)$ where $\nu \in \mathbb{R}_0^+$ is a strictly positive number and Π is a probability measure on $(\mathbb{R}^p, \mathcal{B}^p)$.

Let I be the set of atoms of Π, i.e., $I = \{x \in \mathbb{R}^p : \Pi(\{x\}) > 0\}$.
Let us define :

$\Pi_d(.) = \Pi(.|I)$, $\Pi_c(.) = \Pi(.|I^c)$

$\nu_d = \nu\Pi(I)$, $\nu_c = \nu\Pi(I^c)$

$P_d(.) = P(.|I)$, $P_c(.) = P(.|I^c)$ and $\gamma = P(I)$
Clearly

$\nu\Pi = \nu_c\Pi_c + \nu_d\Pi_d$ and
$P = (1-\gamma)P_c + \gamma Pd$

Theorem A.1.

The two following properties are equivalent :

(i) $P \sim \mathcal{D}i(\nu\Pi)$

(ii)

1. $P = (1-\gamma)P_c + \gamma P_d$ *where*
    - γ, P_d *and* P_d *are independent*
    - $\gamma \sim B(\nu_d, \nu_c)$
    - $P_c \sim \mathcal{D}i(\nu_c\Pi_c)$
    - $P_d \sim \mathcal{D}i(\nu_d\Pi_d)$
2. $P_c = \sum_{1\le k<\infty} \alpha_k \delta_{\xi_k}$ *where*
    - (ξ_k) *is an infinite i.i.d. sample of* Π_c

- (v_k) *is an infinite i.i.d sample of* $B(1, \nu_c)$ *and*

  $$\alpha_k = v_k \prod_{1 \leq \ell \leq k-1} (1 - v_\ell)$$

- (ξ_k) *and* (v_k) *are independent*

3. $P_d = \sum_{x \in I} \beta_x \delta_x$ *where*

- (β_x) *has a Dirichlet distribution on the infinite dimensional simplex characterized by :*

$\forall \mathcal{J}$ *finite partition of* I*, if for* $j \in \mathcal{J}$
$\beta_j = \sum_{x \in j} \beta_x$, $\Pi(j) = \sum_{x \in j} \Pi(\{x\})$*, then* $\{\beta_j; j \in \mathcal{J}\} \sim D(\{\nu_d \Pi(j)\}; j \in \mathcal{J})$

■

The proof that (i) implies (ii) may be find in Rolin (1992 and 1993) but it is rather long especially because it proves that the representation ii) 2. is almost sure. If we only require that ii)2. is an equality in distribution, the proof may be considerably simplified. The proof of theorem A.1., will proceed through several lemmas that have an intrinsinc interest. First of all a basic tool is the relationship between Dirichlet processes and Gamma processes.

Definition A.1.1.

A random finite measure A *on* $(\mathbb{R}^p, \mathcal{B}^p)$ *is a Gamma process of parameters* $\lambda \in \mathbb{R}_0^+$ *and* ρ *finite measure on* $(\mathbb{R}^p, \mathcal{B}^p)$ *to be written as* $A \sim \mathcal{G}a(\lambda, \rho)$*, if for any measurable non trivial partition of* $\mathbb{R}^p, \{B_\ell : 1 \leq \ell \leq k\}$*,*

(i) $\{A(B_\ell), 1 \leq \ell \leq k\}$ *are independent random variables*

(ii) $A(B_\ell) \sim \Gamma(\lambda, \rho(B_\ell))$ ■

The proof of the following lemma will be omitted because it is rather well known (see , e.g., Rolin 1992).

Lemma A.1.2.

(i) $A \sim \mathcal{G}a(\lambda, \nu\Pi)$ *and* $P = \dfrac{1}{A(\mathbb{R}^p)} A$ *then* P *and* $A(\mathbb{R}^p)$ *are independent and* $P \sim \mathcal{D}i(\nu\Pi)$

(ii) If $P \sim \mathcal{D}i(\nu\Pi)$, $U \sim \Gamma(\lambda, \nu)$ and U is independent of P, then $A = UP \sim \mathcal{G}a(\lambda, \nu\Pi)$ ■

Lemma A.1.3.

Let $P \sim \mathcal{D}i(\nu\Pi)$ and $B \in \mathcal{B}^p$ such that $0 < \Pi(B) < 1$. If $P^B(.) = P(.|B)$ and $\Pi^B(.) = \Pi(.|B)$. Then $P = P(B)P^B + P(B^C)P^{B^C}$ where

- *$P(B), P^B, P^{B^C}$ are independent.*
- *$P(B) \sim B(\nu\Pi(B), \nu\Pi(B^C))$*
- *$P^B \sim \mathcal{D}i(\nu\Pi(B)\Pi^B)$*
- *$P^{B^C} \sim \mathcal{D}i(\nu\Pi(B^C)\Pi^{B^C})$*

■

By taking $B = I$, this lemma entails that (i) implies (ii) 1 in Theorem A.1. .

Proof :

Let $U \sim \Gamma(1, \nu)$ and U independent of P. Then by lemma A.1.2. (ii) $A = UP = UP(B)P^B + UP(B^C)P^{B^C} \sim \mathcal{G}a(1, \nu\Pi)$. If $A_1(.) = A(. \cap B)$ and $A_2(.) = A(. \cap B^C)$, clearly, A_1 and A_2 are independent and $A_1 \sim \mathcal{G}a(1, \nu\Pi(B)\Pi^B)$, $A_2 \sim \mathcal{G}a(1, \nu\Pi(B^C)\Pi^{B^C})$. However $A_1 = UP(B)P^B$, $A_1(I\!R^p) = UP(B) = A(B)$, $A_2 = UP(B^C)P^{B^C}$, $A_2(I\!R^p) = UP(B^C) = A(B^C)$.
Therefore by lemma A.1.2. (i), $P^B \sim \mathcal{D}i(\nu\Pi(B)\Pi^B)$, $P^{B^C} \sim \mathcal{D}i(\nu\pi(B^C)\Pi^{B^C})$ and $P^B, A(B), A(B^C), P^{B^C}$ are independent.
Hence, $P(B) = \frac{A(B)}{A(B)+A(B^C)}$ is independent of P^B and P^{B^C}, and since $A(B)$ and $A(B^C)$ are independent, $A(B) \sim \Gamma(1, \nu\Pi(B)$ and $A(B^C) \sim \Gamma(1, \nu\Pi(B^c))$, a simple change of variables shows that $P(B) \sim B[\nu\Pi(B), \nu\Pi(B^C)]$. ■

Lemma A.1.4.

If $P_1 \sim \mathcal{D}i(\nu_1\Pi_1), P_2 \sim \mathcal{D}i(\nu_2\Pi_2), \gamma \sim B(\nu_1, \nu_2)$ and γ, P_1 and P_2 are independent, then

$$P = \gamma P_1 + (1-\gamma)P_2 \sim \mathcal{D}i(\nu_1\Pi_1 + \nu_2\Pi_2)$$

■

This general lemma shows that (ii) 1 implies (i) in Theorem A.1.

Proof :

Let $U \sim \Gamma(1, \nu_1 + \nu_2)$ and U independent of γ, P_1 and P_2. Let $U_1 = \gamma U$ and $U_2 = (1-\gamma)U$. A simple change of variables, shows that U_1 and U_2 are independent (and therefore independent of P_1 and P_2), $U_1 \sim \Gamma(1, \nu_1)$ and $U_2 \sim \Gamma(1, \nu_2)$. Therefore by lemma A.1.2. (ii), $U_1P_1 \sim \mathcal{G}a(1, \nu_1\Pi_1)$ $U_2P_2 \sim \mathcal{G}a(1, \nu_2\Pi_2)$ and U_1P_1 is independent of U_2P_2. Hence $A = UP = U_1P_1 + U_2P_2 \sim \mathcal{G}a(1, \nu_1\Pi_1 + \nu_2\Pi_2)$. This follows directly from the definition of the Gamma process. Finally, by lemma A.1.2. (i), $P = \frac{1}{U}A \sim \mathcal{D}i(\nu_1\Pi_1 + \nu_2\Pi_2)$.

■

Clearly, by the definition of a Dirichlet process, in Theorem A.1., 3 is equivalent to $P_d \sim \mathcal{D}i(\nu_d\Pi_d)$ since $\Pi_d(I^C) = 0$ implies that $P_d(I^C) = 0$ a.s. . To finish the proof of Theorem A.1., it suffices to show that 2 is equivalent to $P_c \sim \mathcal{D}i(\nu_c\Pi_c)$. Now recall that Ferguson (1973) shows that $P \sim \mathcal{D}i(\nu\Pi)$ and $\xi|P \sim P$ if and only if $\xi \sim \Pi$ and $P|\xi \sim \mathcal{D}i(\nu\Pi + \delta_\xi)$. This is the non parametric Bayes theorem. This entails the following lemma.

Lemma A.1.5.

If $P \sim \mathcal{D}i(\nu\Pi)$ with Π diffuse, i.e., $\Pi(\{x\}) = 0$ $\forall x \in \mathbb{R}^p$, then
$P = \alpha_1\delta_{\xi_1} + (1-\alpha_1)P_1$ where

- *α_1, ξ_1 and P_1 are independent*
- *$\alpha_1 \sim B(1, \nu)$*
- *$\xi_1 \sim \Pi$*
- *$P_1 \sim \mathcal{D}i(\nu\Pi)$*

■

Proof:

Let ξ_1 be a random variable such that $\xi_1|P \sim P$. Hence, $\xi_1 \sim \Pi$ and $P|\xi_1 \sim \mathcal{D}i(\nu\Pi + \delta_{\xi_1})$. By lemma A.1.3., taking $B = \{\xi_1\}$, conditionally on ξ_1,

$$P = \alpha_1 P_2 + (1 - \alpha_1) P_1$$

and conditionally on ξ_1, α_1, P_1 and P_2 are independent. Moreover, $\alpha_1|\xi_1 \sim B(1,\nu)$ and this implies the independence between α_1 and ξ_1. Now $P_1|\xi_1 \sim \mathcal{D}i(\nu\Pi)$, because $(\nu\Pi + \delta_{\xi_1})(. \cap \{\xi_1\}^C) = \nu\Pi$ and therefore ξ_1 and P_1 are also independent. The independences between α_1 and ξ_1, ξ_1 and P_1 and α_1 and P_1 conditionally on ξ_1 are equivalent to the independence between α_1, ξ_1 and P_1. Finally, $P_2|\xi_1 \sim \mathcal{D}i(\delta_{\xi_1})$ because $(\nu\Pi + \delta_{\xi_1})(. \cap \{\xi_1\}) = \delta_{\xi_1}(.)$ and this amounts to say that $P_2 = \delta_{\xi_1}$ a.s.

For the converse we have the following lemma. ■

Lemma A.1.6.

If $P = \alpha_1 \delta_{\xi_1} + (1 - \alpha_1) P_1$ *where*

- α_1, ξ_1 *and* P_1 *are independent*
- $\alpha_1 \sim B(1,\nu)$
- $\xi_1 \sim \Pi$
- $P_1 \sim \mathcal{D}i(\nu\Pi)$

then $P \sim \mathcal{D}i(\nu\Pi)$ ■

Proof :

Conditionally on $\xi_1, \delta_{\xi_1} \sim \mathcal{D}i(\delta_{\xi_1})$ and so, by lemma A.1.4., $P|\xi_1 \sim \mathcal{D}i(\nu\Pi + \delta_{\xi_1})$. But $\xi_1 \sim \Pi$ implies that $P \sim \mathcal{D}i(\nu\Pi)$. ■

By repeated applications of lemma A.1.5. and lemma A.1.6., we obtain the following corollary.

Corollary A.1.7.

$P_c \sim \mathcal{D}i(\nu_c\Pi_c)$ *if and only if* $\forall K, P_c = \sum_{1 \le k \le K} \alpha_k \delta_{\xi_k} + \varepsilon_K P_K$, *where*

- $(\xi_k)_{k=1,\cdots,K}$ *is an i.i.d. sample of* Π_c
- $(v_k)_{k=1,\cdots,K}$ *is an i.i.d. sample of* $B(1,\nu_c)$ *and* $\alpha_k = v_k \prod_{1\le \ell\le k-1} (1-v_\ell)$ *for* $k=1,\cdots,K$ *and* $\varepsilon_K = \prod_{1\le \ell\le K}(1-v_\ell)$
- $P_K \sim \mathcal{D}i(\nu_c \Pi_c)$
- $(\xi_k), (v_k), P_K$ *are independent.*

■

The proof of Theorem A.1. will be completed if we prove that $\varepsilon_K \to 0$ as as $K\to\infty$. Note that if $P_c^K = \sum_{k=1}^{K} \alpha_k \delta_{\xi_k}$, it is clear that $\sup_{B\in\mathcal{B}^p} |P_c(B) - P_c^K(B)| = \varepsilon_K$.

Theorem A.2.

The random variable $\varepsilon_K = \sum_{k>K} \alpha_k$ *has a distribution dominated by Lebesgue measure whose density is equal to*

$$\frac{\nu_c^K}{\Gamma(K)}\left(ln\frac{1}{\varepsilon}\right)^{K-1} \varepsilon^{\nu_c - 1} 1_{(0,1)}(\varepsilon)$$

The moments of ε_K *are given by*

$$E(\varepsilon_K^p) = \left(\frac{\nu_c}{\nu_c + p}\right)^K$$

Moreover,

$$\frac{1}{K} ln\varepsilon_K \to -\frac{1}{\nu_c} \text{ a.s. as } K\to\infty.$$

Therefore, as $K\to\infty$

$$\varepsilon_K \to 0 \quad a.s. \text{ and in } L^p \ \forall p \in (0,\infty)$$

■

Proof :

By corollary A.1.7., $ln\frac{1}{\varepsilon_K} = \sum_{1\le k\le K} ln\left(\frac{1}{1-v_\ell}\right)$.
Now it is easy to see that $ln\frac{1}{1-v_\ell}$ has an exponential distribution of parameter ν_c. Therefore $\ln\frac{1}{\varepsilon_K} \sim \Gamma(\nu_c, K)$ and $\frac{1}{K}ln\frac{1}{\varepsilon_K} \to E(ln\frac{1}{1-v_1}) = \frac{1}{\nu_c}$ by the strong law of large numbers. A simple change of variables gives the density of ε_K and $E[\varepsilon_K^a] = E[(1-v_1)^a]^K = \left(\frac{\nu_c}{\nu_c+a}\right)^K$ since $(1-v_\ell)_{\ell=1,\cdots,K}$ is an i.i.d. sample of $B(\nu_c, 1)$. ■

Corollary A.3.

If $\bar{P}_c^K = \left(\sum_{k=1}^K \alpha_k\right)^{-1} P_c^K, \quad \sup_{B\in\mathcal{B}^p} |P_c(B) - \bar{P}_c^K(B)| = \varepsilon_K$
and therefore converges to 0 in L^p $\forall p \in (0,\infty)$ *if* $K\to\infty$ *or if* $\nu_c \to 0$.

■

Proof :

Note that $\bar{P}_c^K$ is a probability measure and that

$$P_c(B) - \bar{P}_c^K(B) = \sum_{k>K} \alpha_k \delta_{\xi_k}(B) - \varepsilon_K \bar{P}_c^K(B)$$

Therefore $P_c(B) - \bar{P}_c^K(B) \le \varepsilon_K$ and $P_c(B) - \bar{P}_c^K(B) \ge -\varepsilon_K$ and that these bounds are achieved. Now, $\forall p \in (0,\infty)$

$$E(\varepsilon_K^p) = \left(\frac{\nu_c}{\nu_c + p}\right)^K \to 0 \text{ if } \nu_c \to 0 \text{ or if } K \to \infty.$$

■

2. Moments of Dirichlet process

We first give a general result about the dominance of a vector valued functional of a Dirichlet process.

Theorem A.4.

Let P be a random probability measure generated by a Dirichlet process whose parameters are ν and Π and $\theta = \varphi(P), i.e., \theta \in I\!R^\ell$ is a vector of parameters associated to φ. The functional φ is assumed to be defined on the set of probability measures on $(I\!R^p, \mathcal{B}^p)$ (the supports of which are included in the support of Π) and φ is continuously Frechet differentiable in any P generated by the Dirichlet process. We denote by $\partial_P \varphi$ the derivative of φ at P, which is a linear functional on the set of measures on $I\!R^p$. Then, if one of the two following conditions is satisfied :

i) $\nu_c > 0$ and $(\partial_P \varphi(\delta_{\xi_k}))_{k=1,\cdots,\ell}$ is a regular $\ell \times \ell$ matrix.

ii) $\nu_d > 0$ and there exists $J \subset I, J \neq I$, card $(J) = \ell$ such that $(\partial_P \varphi(\delta_x))_{x \in J}$ is a regular $\ell \times \ell$ matrix,
the distribution of θ is dominated by the Lebesgue measure in $I\!R^\ell$. ■

Remarks:

- If $\nu_c = 0$ the dominance property of θ requires in particular that

$$\text{card}(I) > \ell.$$

- If ϕ is linear (e.g. if $\phi(P) = \int f dP$) the derivative of ϕ, $\partial_P \phi$, is identical to ϕ for any P. If $\phi(P) = \int f dP$ the matrix of derivatives becomes the matrix $(f(\xi_j))_{j=1,\ldots,\ell}$.

Proof:

This result is based on this following general argument. Let x, y, z be three random elements such that $dim\ (x) = dim\ (y) = \ell$. The dimension of z could be infinite. These elements are related by a measurable relation $x = f(y, z)$ and the distribution of y given z is dominated by the Lebesgue. Finally, the matrix of partial derivatives $\frac{\partial f}{\partial y'}$ is assumed to be regular and continuous in y. Then, under this set of hypothesis, the distribution of x is dominated by the Lebesgue measure. The proof of this result is obtained by noting first that the conditional distribution of x given z is dominated. Using Fubini theorem, the second step of the proof is the integration of z which preserves the dominance property.

Let us now analyses the first case. If $\nu_c > 0$ the first term of the decomposition given in theorem A.1 ii) is not degenerated. Let us consider the bijective infinitely differentiable transformation given by Theorem A.1. (ii) 2, i.e.,

$$\alpha_k = a_k(v) = v_k \prod_{1 \leq \ell \leq k-1} (1 - v_\ell)$$

and set $x = (v_1, v_2, \cdots, v_\ell)$ and $z = ((v_k)_{k>\ell}, (\xi_k)_k, (\beta_x)_x, \gamma)$

Clearly the distribution of x conditionally on z is dominated since x and z are independent by the independence properties of Theorem A.1. and since the v_k are independent and $B(1, \nu_c)$.

Clearly

$$\theta = \varphi\left((1-\gamma)\sum_{k\geq 1} a_k(v)\delta_{\xi_k} + \gamma \sum_x \beta_x \delta_x\right) = f(x, z)$$

It remains to check that the matrix of derivatives of f w.r.t. $(v_1, v_2, \cdots, v_\ell)$ is regular and continuous in $(v_1, v_2, \cdots, v_\ell)$. Now, the transformation $(v_1, v_2, \cdots, v_\ell) \to (\alpha_1, \alpha_2, \cdots, \alpha_\ell)$ is clearly an infinitely differentiable bijective transformation and therefore, it suffices to check that for $\theta = g(\alpha_1, \alpha_2, \cdots, \alpha_\ell, z)$, the matrix of derivatives of g w.r.t. $(\alpha_1, \alpha_2, \cdots, \alpha_\ell)$ is regular and continuous in $(\alpha_1, \alpha_2, \cdots, \alpha_\ell)$. The probability P is differentiable w.r.t. α_j and ϕ is assumed to be continuously differentiable relatively to P. Using the theorem of the derivative of composition of functions, we get:

$$\frac{\partial}{\partial \alpha_j}\phi((1-\gamma)\sum_{k\geq 1}\alpha_k\delta_{\xi_k} + \gamma\sum_x \beta_x\delta_x) = (1-\gamma)\partial\phi_P(\partial_{\xi_j}).$$

The multiplicative term $(1-\gamma)$ is almost surely strictly positive and the regularity of the matrix of derivatives is equivalent to the assumption given in i).

If ν_d is positive the dominance of the distribution of θ could be obtained through the dominance property of the distribution of $(\beta_x)_{x\in I}$. If $J \subset I$ is a proper subset of I the vector $(\beta_x)_{x\in J}$ is dominated given all the other random components of P. The regularity of the matrix of derivatives is obtained using the assumption and an analogous computation as above.

■

Theorem A.5.

Let f be a measurable function on $I\!\!R^p$. If $\int |f|^r d\Pi < \infty$ for $r \geq 1$, $\int f dP$ is an L^r random variable. In particular, if f is Π-integrable, $\int f dP < \infty$ a.s.

■

Proof :

Indeed,

$$\left|\int f dP\right|^r \leq \left(\int |f| dP\right)^r \leq \int |f|^r dP$$

by Hölder's inequality. Therefore

$$E\left|\int f dP\right|^r \leq E\int |f|^r dP = \int |f|^r d\Pi$$

■

In fact, it may be shown that $\int |f| dP < \infty$ a.s. if and only if $\int ln(1 + |f|) d\Pi < \infty$ (see Yamato (1984)).

3. Properties of the solution of a moment equation

Let $\theta \in \Theta$, (an open set of $I\!\!R^\ell$) be characterized by the equation

$$\int \psi(x, \theta) P(dx) = 0$$

where $P \sim \mathcal{D}i(\nu, \Pi)$. This definition is based on two assumptions. First the integral $\int \psi(x, \theta) P(dx)$ is well defined and this condition may be warrented using theorem A.5 by assuming that $\int |\psi(x, \theta)| \Pi(dx)$ is finite for any θ. Then, writing $\int \psi(x, \theta) P(dx) = \Psi(P, \theta)$, we have secondly to assume that the equation $\Psi(P, \theta) = 0$ has unique solution for all P. This problem is a standard application of the implicit function theorem with two specific elements: P is a probability measure and not a finite dimensional vector and Ψ is linear in P. This linear property simplifies some computations but is not essential and the results of this section could be extended to more general Ψ.

Theorem A.6.

We assume that ψ is continuously differentiable as a function of θ. Then the distribution of θ is dominated by the Lebesgue measure if one of the following condition is satisfied :

i) $\nu_c > 0$ and $(\psi(\xi_j, \theta))_{j=1,\cdots,\ell}$ is a full rank matrix
ii) $\nu_d > 0$ and $\exists J \subset I,\ J \neq I,$ card $J = \ell$, such that $(\psi(x, \theta))_{x \in J}$ is a full rank matrix. ■

Proof :

The proof is identical to the proof of theorem A.4 where $\varphi(P)$ becomes the solution of the equation $\Psi(P, \theta) = 0$. We just have to note that, using the theorems of derivation of a composition and of an implicit function, the derivative of $\varphi(P)$ with respect to $(\alpha_j)_{j=1,\ldots,\ell}$ becomes :

$$-\left[\frac{\partial}{\partial \theta'} \int \psi(x, \theta) dP\right]^{-1} [\psi(\xi_1, \theta), \cdots, \psi(\xi_\ell, \theta)]$$

evaluated at $\theta = \varphi(P)$.

Note that the existence of a unique solution of the equation $\psi(P, \theta) = 0$ warranties the regularity of the matrices $\frac{\partial}{\partial \theta'} \int \psi(x, \theta) P(dx)$ and $\frac{\partial}{\partial \theta'} \psi(x, \theta)\ \forall P$ and x. The rank of the matrix given above is then identical to the rank of $(\psi(\xi_j, \theta))_{j=1,\cdots,\ell}$ which is assumed to be equal to ℓ. The case of the derivation with respect to β_x is analogous. ■

The last theorem is related to the convergence of the approximated solution.

Theorem A.7.

Let us assume that :

(i) $\psi(x, \theta)$ is continuous in θ for any $x \in I\!R^p$
(ii) For any θ there exists a compact neighbourhood V_θ of θ such that

$$\int \sup_{\theta' \in V_\theta} \| \psi(x, \theta') \|_1 \ \Pi_c(dx) < \infty$$

where $\| \theta \|_1 = \sum_{i=1}^{\ell} |\theta_i|$

(iii) I is finite

(iv) There exists a unique solution θ_0 to the equation $\psi(P,\theta) = 0$

(v) There exists a compact set C in $\mathbb{R}^\ell$, such that for K sufficiently large, there exists a solution $\theta_K \in C$ to the equation :

$$\psi(P^K,\theta) = \int \psi(x,\theta) P^K(dx) = (1-\gamma)\sum_{k=1}^{K} \alpha_k \psi(\xi_k,\theta) + \gamma \sum_{x\in I} \beta_x \psi(x,\theta) = 0$$

Then $\theta_K \to \theta_0$ a.s. as $K \to \infty$. ■

Proof :

By the finite covering theorem of compact sets assumption (ii) implies that, for any compact set C,

$$\int \sup_{\theta\in C} \| \psi(x,\theta) \|_1 \, \Pi_c(dx) < \infty$$

Therefore, by Theorem A.5., and assumption (iii) almost surely,

$$\int \sup_{\theta\in C} \| \psi(x,\theta) \|_1 \, P(dx)$$
$$= (1-\gamma)\sum_{k=1}^{\infty} \alpha_k \sup_{\theta\in C} \| \psi(\xi_k,\theta) \|_1 + \gamma \sum_{x\in I} \beta_x \sup_{\theta\in C} \| \psi(x,\theta) \|_1 < \infty$$

Hence, almost surely, $\psi(P,\theta)$ is continuous in θ and

$$\sup_{\theta\in C} \| \psi(P,\theta) - \psi(P^K,\theta) \|_1 \le (1-\gamma)\sum_{k>K} \alpha_k \sup_{\theta\in C} \| \psi(\xi_k,\theta) \|_1$$

and this last term tends to zero almost surely as $K \to \infty$.
Finally, if θ_{K_k} is a subsequence of θ_K in C converging to θ_∞ let us say, by uniform convergence in C,

$$0 = \lim_k \psi(P^{K_k}, \theta_{K_k}) = \psi(P,\theta_\infty)$$

By assumption (iv), $\theta_\infty = \theta_0$ and therefore, $\theta_K \to \theta_0$ a.s. ■

References

Antoniak C.E. (1974) : "Mixture of Dirichlet Processes with Applications to Bayesian Non Parametric Problems" *Annals of Statistics*, 2, 1152-1174.

Blackwell, D. (1973) : "Discreteness of Ferguson Selections", *Annals of Statistics*, 1, 356-358.

Blackwell D. and J.B. MacQueen (1973) : "Ferguson Distributions via Polya Urn Schemes", *Annals of Statistics*, 1, 353-355.

Bunke, O. (1985) : "Bayesian Estimators in Semiparametric Problems", *Tech. Rep.*, 102, Humboldt-Universitat, Berlin.

Cifarelli, O.M. and E. Regazzini (1990) : "Distribution Functions of Means of a Dirichlet Process", *Annals of Statistics*, 18, 429-442.

Devroye, L. (1988) : "A Course in Density Estimation", Birkhauser.

Diaconis P. and D. Freedman (1986) : "On Unconsistant Bayes Estimates of Location", *Annals of Statistics*, 14, 68-87.

Ferguson, T.S. (1973) : "A Bayesian Analysis of Some Non Parametric Problems", *Annals of Statistics*, 1, 209-230.

Ferguson, T.S. (1974) : "Prior Distributions on Spaces of Probability Measures", *Annals of Statistics*, 2, 615-629.

Florens, J.P., Mouchart, M. and J.M. Rolin (1992) : "Bayesian Analysis of Mixtures: Some Results on Exact Estimability and Identification", in *Bayesian Statistics*, 4, edited by J.M. Bernardo and alii, Oxford Sciences Publications, 127-145.

Florens, J.P., Mouchart, M. and J.M. Rolin (1993) : "Semi-and Non Parametric Bayesian Analysis of Duration Models: A Survey", CORE DP.

Hannum, R.C., Hollander, M. and N.A. Langberg (1981) : "Distributional Results for Random Functionals of a Dirichlet Process", *Annals of Probability*, 9, 665-670.

Hansen, L.P. (1982) : "Large Sample Properties of Generalized Method of Moments Estimators", *Econometrica*, vol 50, 1029-1054.

Hjort, N.L. (1990) : "Non Parametric Bayes Estimators Based on Beta Process in Models for Life History Data", *Annals of Statistics*, vol 18, no 3, 1259-1294.

Johnson and Kotz (1989) : "Distributions in Statistics: Discrete Distributions", John Wiley, New York.

Korwar, R. and M. Hollander (1973) : "Contributions to the Theory of Dirichlet Processes" *Annals of Probability*, 1, 705-711.

Lancaster, T. (1992) : "Bayes WESML", Discussion Paper Providence University.

Lo, A.Y. (1984) : "On a Class of Bayesian Non Parametric Estimates I : Density Estimates", *Annals of Statistics*, 12, 351-357.

Rolin, J.M. (1992) : "Some Useful Properties of the Dirichlet Process", Institut de Statistique DP 9202, Université Catholique de Louvain, Louvain-la-Neuve, Belgium.

Rolin, J.M. (1993) : "On the Distribution of Jumps of the Dirichlet Process", Institut de Statistique, DP 9302, Université Catholique de Louvain, Louvain-la-Neuve, Belgium.

Rubin (1981) : "The Bayesian Bootstrap", *Annals of Statistics*, 9, 130-134.

Ruggiero (1994) : "Bayesian Semiparametric Estimation of Proportional Hazards Models", *Journal of Econometrics*, (to appear)

Sethuraman J. (1994) : "A Constructive Definition of Dirichlet Priors", *Statistica Sinica*, 4(2), 639-650.

Sethuraman J. and R.C. Tiwari (1982) : "Convergence of Dirichlet Measures and the Interpretation of their Parameter", *Statistical Decision Theory and Related Topics*, III2, 305-315.

Wilks, S.S. (1962) : *Mathematical Statistics*, New York, Wiley.

Yamato, H. (1984) : "Characteristic Functions of Means of Distributions Chosen from a Dirichlet Process", *Annals of Probability*, 12, 262-267.

Yamato, H. (1984) : "Properties of Sample from Distribution Choosen from a Dirichlet Process", Bulletin of Informatics and Cybernetics, 21, 77-83.