PROCESS OF RANDOM DISTRIBUTIONS CLASSIFICATION AND PREDICTION

By Richard Emilion
Université d'Orléans

We define a continuous time stochastic process such that each is a Ferguson-Dirichlet random distribution. The parameter of this process can be the distribution of any usual such as the (multifractional) Brownian motion. We also extend Kraft random distribution to the continuous time case.

We give an application in classifiying moving distributions by proving that the above random distributions are generally mutually orthogonal. The proofs hinge on a theorem of Kakutani.

1. Introduction

A random distribution (RD) is a measurable map from a probability space $(\Omega, \mathcal{F}, \mathcal{P})$ to the space $\mathbf{P}(V)$ of all probability measures defined on a fixed measurable set $(V, \mathcal{V})$. This notion is needed when dealing with the description of the distribution of random elements which are themselves probability distributions. The interest of such descriptions is emphasized for example in the nice paper of J.F.C. Kingman (1975) on random discrete distributions. Indeed there are many situations where one is faced with observations modelized by probability distributions rather than real vectors. Various references are mentionned by Pitman and Yor (1996) : models in ecology, in population genetics, in storage and search, prior in nonparametric Bayesian statistics, zero sets of stochastic processes, asymptotic distributions in number theory, representation of partition structures.

The motivation for the present work comes from a situation which is oftently encountered nowadays.When having a very large data set with a huge number of records described by an attribute, the analysis of this large set could turn out to be of great complexity but also be not very meaningfully. The records are then grouped into units according to a user specification, each unit being described by the distribution of the attribute within the unit. It is clear that these units can be considered as random elements which are themselves probability distributions.

We study here the finite mixture problem for a RD $X : (\Omega, \mathcal{P}) \longrightarrow \mathbf{P}(V)$. It consists in estimating the distribution $\mathcal{P}_X$ of X as a finite mixture, that is a convex combination $\sum_{k=1,\ldots,K} p_k P_k$, the P_k's being distributions on $\mathbf{P}(V)$ belonging to a specific family. This generalizes the well-known case where the observations are real vectors. Actually, as reported by B. Schweizer (2002), the idea of operating with distributions as data is already explicit in K. Menger's early writings. Our solution answers a question posed by E. Diday in the frame of Symbolic Data Analysis.

The paper is organized as follows.

In section 2 we give some examples of distributions of RDs that we will use as components of the mixture : Ferguson (1973, 1974), Dykstra and Laud (1981) and Lo (1982), Kraft (1964). We show that these components are generally mutually orthogonal, the proof hinging on a very nice theorem of Kakutani.

In Section 3 we will recall the mixture problem in R^p and some clustering-based algorithms which estimate the solutions.

[0] *AMS* 1991 *subject classifications.* Primary 65U05, 62F10 ; secondary 62M30, 60K35.

Key words and phrases. Bayesian, Clustering, Dirichlet distributions, Dirichlet processes, E.M. algorithm, gamma processes, Kraft processes, mixture, nonparametric estimation, random distributions, S.A.E.M. algorithm, weighted gamma processes.

Invited paper, received on February 1, 2005

Section 4 and 5 contains the method and the main result : the observations are n distributions f_i, the set V is split into a finite partition (V_k) and the above algorithms are applied to the vectors $f_i(V_k)$ in order to estimate finite dimensional mixtures. We then prove the convergence of these estimations as the partitions are refined. Clusters of the n observations are given by the disjoint supports of the components.

The proposed method differs from that of Antoniak (1974) and extends it to normalized weighted gamma processes and Kraft ones.

In Section 6 we define two continuous time processes of random distributions.

The last section is devoted to an application in the classification of Internet flows.

2. Orthogonal random distributions

Let $(\Omega, \mathcal{F}, \mathcal{P})$ be a probability space and $\mathbf{P}(V)$ the set of all probability measures defined on a measurable space $(V, \mathcal{V})$. Let $\mathcal{B}$ be the smallest $\sigma-$field on $\mathbf{P}(V)$ such that the mappings $Q \longrightarrow Q(A)$ defined on $\mathbf{P}(V)$ are measurable for any $A \in \mathcal{V}$. If V is a complete separable metric space, then $\mathcal{B}$ is also the Borel $\sigma-$field when $\mathbf{P}(V)$ is topologized by weak convergence. In most of applications V will be product space $V_1 \times ... \times V_p$ but for sake of simplicity and without loss of generality, we will suppose, throughout this paper, that $V = [0,1]$ with its standard uniform probability measure λ.

Definition 1 A random distribution (RD) is a measurable map from $(\Omega, \mathcal{F})$ to $(\mathbf{P}(V), \mathcal{B})$.

If $X : \Omega \longrightarrow \mathbf{P}(V)$ is a RD, its distribution $\mathcal{P}_X$ is then a probability measure on $\mathbf{P}(V)$. Le us mention some examples of RDs, starting with the case of a finite set.

2.1. Case of a finite set V.

If V is a finite set with $l = \#V$ then $\mathbf{P}(V)$ can be identified to the set

$$\{y = (y_1, ..., y_l), y_j \geq 0, \sum_{j=1}^{l} y_j = 1\}$$

and any RD to a random vector

$$X = (X_1, ..., X_l) : (\Omega, \mathcal{P}) \to R_+^l \text{ such that} \sum_{k=1}^{l} X_k = 1.$$

In that case, distributions on $\mathbf{P}(V)$ can be obtained in considering the distribution of a positive random vector divided by the sum of its coordinates. This the case of standard Dirichlet distributions or, more generally, normalized weighted gamma distributions which better encompass concrete situations.

2.1.1. *Dirichlet distributions* $\mathcal{D}(\alpha_1, ..., \alpha_l)$. Let $\alpha = (\alpha_1, ..., \alpha_l)$, with $\alpha_1 > 0, ..., \alpha_l > 0$, and let $Z_1, ..., Z_l$ be l independent real random variables with gamma distributions $\gamma(\alpha_1, 1), ..., \gamma(\alpha_l, 1)$ respectively, where

$$\gamma(a,b)(x) = \frac{1}{\Gamma(a)} b^a e^{-bx} x^{a-1} I_{(x>0)}.$$

The Dirichlet distribution $\mathcal{D}(\alpha_1, ..., \alpha_l)$ is defined as the distribution of the random vector $(\frac{Z_1}{Z}, ..., \frac{Z_l}{Z})$ where $Z = Z_1 + ... + Z_l$.

This distribution is singular w.r.t. Lebesgue measure since its support has Lebesgue measure 0, however if $l \geq 2$ then $(\frac{Z_1}{Z}, ..., \frac{Z_{l-1}}{Z})$ has the following density

$$d(\alpha \mid y) = \frac{\Gamma(\alpha_1 + ... + \alpha_l)}{\Gamma(\alpha_1)...\Gamma(\alpha_l)} y_1^{\alpha_1 - 1} ... y_{l-1}^{\alpha_{l-1}-1} (1 - \sum_{h=1,...,l-1} y_h)^{\alpha_l - 1} I_{S_l}(y) \tag{2.1}$$

where S_l denotes the simplex

$$S_l = \{y = (y_1, ..., y_{l-1}), y_j \geq 0, \sum_{j=1}^{l-1} y_j \leq 1\}. \tag{2.2}$$

This completely determines $\mathcal{D}(\alpha_1, ..., \alpha_l)$ since $\frac{Z_l}{Z} = 1 - \sum_{h=1,...,l-1} \frac{Z_h}{Z}$.

2.1.2. *Normalized weighted gamma $\mathcal{D}(\alpha_1, ..., \alpha_l; \beta)$.* Let $\beta \geq 0$ be a nonzero positive function defined on R_+ such that $\beta\gamma(\alpha, 1)$ is integrable. Then there exists a constant $c_\alpha(\beta) > 0$ such that $\gamma^\beta(\alpha, 1)$ defined by

$$\gamma^\beta(\alpha, 1)(x) = c_\alpha(\beta)\beta(x)\gamma(\alpha, 1)(x)$$

is a density function.

The normalized weighted gamma distribution $\mathcal{D}(\alpha_1, ..., \alpha_l; \beta)$ is defined as the distribution of the random vector $(\frac{Z_1}{Z}, ..., \frac{Z_l}{Z})$ with independent $Z_i \sim \gamma^\beta(\alpha_i, 1)$ and $Z = Z_1 + ... + Z_l$.

It is obviously seen that $\mathcal{D}(\alpha_1, ..., \alpha_l) = \mathcal{D}(\alpha_1, ..., \alpha_l; 1)$.

2.2. **Discrete Random Distributions.**

2.2.1. *Ferguson RDs.* Dirichlet processes are interesting RDs introduced by Ferguson (1973,1974) in fundamental papers on a Bayesian appproach to some nonparametric problems.

Let α be a probability measure on V. A random distribution $X : \Omega \longrightarrow \mathbf{P}(V)$ is a Dirichlet process $\mathcal{D}(\alpha)$ if for every $k = 2, 3, ...$ and every measurable partition $B_1, ..., B_k$ of V, the joint distribution of the random vector $(X(B_1), ..., X(B_k))$ is a Dirichlet distribution with parameters $(\alpha(B_1), ..., \alpha(B_k))$.

Ferguson proved that this definition satifies the Kolmogorov criteria which yields the existence of such random distributions. He also showed that $X(\omega)$ is a discrete probability measure and that there exists an i.i.d sequence $V_{n,\alpha}$ of random variables

$$V_{n,\alpha} \sim \alpha$$

such that the support of the distribution $X(\omega)$ is contained in the random set

$$\{V_{1,\alpha}(\omega), V_{2,\alpha}(\omega), ..., V_{n,\alpha}(\omega), ...\}.$$

Theorem 1 *Suppose that $\mathcal{P}_X = \sum_{s=1}^K p_s \mathcal{D}_s(\alpha_s; \beta_s)$ is a simple mixture of normalized weighted gamma processes where the α_s's are distinct probabilty measures on $[0, 1]$ equivalent to λ and $\beta_s > 0$. Let S_s be the support of $\mathcal{D}_s(\alpha_s; \beta_s)$ so that the support of $\mathcal{P}_X$ is $\cup_{s=1}^K S_s$. Then*

i) The finite-dimensional distributions of $\mathcal{D}_s(\alpha_s, \beta_s)$ are equivalent.

i*i) The distributions $\mathcal{D}_s(\alpha_s; \beta_s)$ are mutually singular so that the S_s are disjoint.*

Theorem 1 *Suppose that $\mathcal{P}_X = \sum_{s=1}^K p_s \mathcal{D}_s(\alpha_s; \beta_s)$ is a simple mixture of normalized weighted gamma processes where the α_s's are distinct probabilty measures*

on $[0,1]$ *equivalent to* λ *and* $\beta_s > 0$. *Let* S_s *be the support of* $\mathcal{D}_s(\alpha_s;\beta_s)$ *so that the support of* $\mathcal{P}_X$ *is* $\cup_{s=1}^{K} S_s$. *Then*

i) The finite-dimensional distributions of $\mathcal{D}_s(\alpha_s,\beta_s)$ *are equivalent.*

i*i) The distributions* $\mathcal{D}_s(\alpha_s;\beta_s)$ *are mutually singular so that the* S_s *are disjoint.*

2.2.2. *Lo RDs.* Let α be a finite additive measure on V and let $\beta > 0$ on R_+ be an $\alpha-$integrable function. A random distribution $X : \Omega \longrightarrow \mathbf{P}(V)$ is a normalized weighted gamma process $\mathcal{D}(\alpha;\beta)$ if for every $k = 2,3,...$ and every measurable partition $B_1,...,B_k$ of V, the joint distribution of the random vector $(X(B_1),...,X(B_k))$ is a normalized weighted gamma distribution with parameters $(\alpha(B_1),...,\alpha(B_k);\beta)$. This definition is equivalent to that of Lo (1982). A result similar to Thm 2 holds for these RDs.

2.2.3. *PY RDs.* The paper by Pitman and Yor (1996) deals with an interesting class of discrete RDs derived from random closed sets that have a property of self-similarity. We don't know however wether these RDs are generally mutually orthogonals.

2.3. **Continuous Random Distributions.**

2.3.1. *Kraft RDs.* Kraft defined a RD such that the probability measure $X(\omega)$ has a density w.r.t. the Lebesgue measure on $[0,1]$. His construction hinges on a set

$$Z = \{Z_{\frac{k}{2^r}}; r = 1,2...,k = 1,3,...,2^r - 1\}$$

of completely independent real random variables defined on $(\Omega, \mathcal{F}, \mathcal{P})$ such that

$$0 \leq Z_{\frac{k}{2^r}} \leq 1$$

and

$$E(Z_{\frac{k}{2^r}}) = \frac{1}{2}.$$

Let F_l be the sequence of random cumulative distribution functions (cdf) on $[0,1]$ defined by induction as follows :

$$F_1(0) = 0, F_1(\frac{1}{2}) = Z_{\frac{1}{2}}, F_1(1) = 1$$

$$F_1 \text{ is affine on } [0,\frac{1}{2}) \text{ and } [\frac{1}{2},1]$$

$$F_l(\frac{k}{2^l}) = F_{l-1}(\frac{k-1}{2^l})(1 - Z_{\frac{k}{2^l}}) + F_{l-1}(\frac{k+1}{2^l})Z_{\frac{k}{2^l}}$$

$$F_l \text{ is affine on the dyadic intervals.}$$

F_l has, except at $\frac{k}{2^l}$, a derivative g_l which is constant on the dyadic intervals. If x is written as $x = \sum_{r=1}^{\infty} \frac{\varepsilon_r(x)}{2^r}$ and $k_r(x)$ is defined by $\frac{k_r(x)}{2^r} < x < \frac{k_r(x)+1}{2^r}$ then we have

$$g_l(x) = 2^l \prod_{r=1}^{l} Z_{\frac{k_r(x)+1}{2^r}}{}^{1-\varepsilon_r(x)}(1 - Z_{\frac{k_r(x)}{2^r}})^{\varepsilon_r(x)}.$$

Let $\mathcal{B}_l$ denote the finite $\sigma-$ algebra generated by the dyadic intervals $[\frac{k}{2^l}, \frac{k+1}{2^l}), k = 0, ..., 2^l - 1$, then g_l is a $\mathcal{P} \otimes \lambda-$ martingale w.r.t $(\mathcal{F} \otimes \mathcal{B}_l)_l$. If there exists a constant $C > 0$ such that

$$2^l var(Z_{\frac{k}{2^l}}) \leq C \tag{2.3}$$

then there exists a cdf F, such that

$$F_l(x) \rightarrow F(x) \text{ and } g_l(x) \rightarrow g(x) = F'(x) \text{ as } l \rightarrow +\infty, \tag{2.4}$$

As g is completeley determined by Z, we will use the following notation

$$g = Kraft(Z).$$

For technical reasons which will appear in the proofs, we will assume that the distribution $d_{k,l}$ of $Z_{\frac{k}{2^l}}$ belongs to the exponential family and is equivalent to the uniform distribution on $[\frac{k}{2^l}, \frac{k+1}{2^l})$:

$$d_{k,l} \text{ is equivalent to } uniform([\frac{k}{2^l}, \frac{k+1}{2^l})) \tag{2.5}$$

Theorem 2
i) The finite-dimensional distributions of Kraft processes are equivalent

ii) $Kraft(Z^{(s)})$ *and* $Kraft(Z^{(t)})$ *are mutually singular if* $s \neq t$

iii) $\lim_{l\rightarrow\infty} t_{is}(\sigma_l) = t_{is}$ $(= 1$ *or* $0)$ *and* $\lim_{n\rightarrow\infty}\lim_{\ell\rightarrow\infty} p_s(\sigma_l) = p_s$.

2.3.2. *Brownian-based RDs.* Let $B_t = B_t(0, \sigma)$ be a standard centered Brownian motion in dimesion one. Then, if f is a positive continuous function, $\frac{f(B_t)}{\int_0^1 f(B_s)ds}$ $, 0 \leq t \leq 1$ clearly defines a random probability density and thus a continuous RD on $[0, 1]$. Taking $f(x) = |x|$ or $f(x) = \exp(x)$ leads to special RDs.

Finally note that RDs can be easily obtained by randomizing the parameters of any parametric distribution (e.g. $\frac{1}{\sqrt{2\pi}\sigma(\omega)} e^{-\frac{(x-m(\omega))^2}{2\sigma(\omega)^2}} dx$ defines a Gaussian RD). However even the finite dimensional distributions of such RDs need not be simple.

3. Mixtures in the real vector case

First recall the mixture problem when the observations are real vectors.

Let $x_1, x_2, ..., x_n \in R^p$ be n observations of a sample of size n from a random vector $X : (\Omega, \mathcal{P}) \longrightarrow R^p$. The problem consists in estimating the distribution $\mathcal{P}_X$ of X when $\mathcal{P}_X$ is supposed to be a simple mixture, that is a convex combination $\sum_{k=1,...,K} p_k P_k$, the distributions P_k belonging to a specific parametric family, say the exponential family. Several methods have been proposed to estimate the mixing weights p_k and the parameters of the components P_k [see e.g., Pearson (1894), Cooper (1964), Agrawala (1970), Quandt and Ramsey (1978), Makov and Smith (1976)] but we will use here one of the most efficient method : S.A.E.M. algorithm [Celeux and Diebolt (1992)], a stochastic approximation of the popular E.M. algorithm [Dempster, Laird and Rubin (1977)].

There also exists a second approach of this problem : determine K clusters from the n observations so that the estimated distribution P_k of cluster k belongs to the specified family, the number p_k representing the probability of an individual to belong to cluster k [Scott and Symons (1971)]. We will use here an algorithm based on dynamic clusters and proposed by Diday and Schroeder (1976).

The algorithms are summarized below.

3.1. ***S.A.E.M. Algorithm.*** This algorithm can estimate a mixture of density functions belonging to the exponential family, meaning that they have the form :

$$h(x,a) = d(a)e(x)\exp\{a^T b(x)\}$$

where the parameter a is a vector with transpose a^T, $d(a)$ is a normalizing factor, e and b are fixed but arbitrary functions.

The inputs are n vectors $f_i, i = 1, ..., n$. The number of components is a given integer K.

Let $\{\gamma_q\}$ be a sequence of positive real numbers decreasing to zero at a sufficiently slow rate, with $\gamma_0 = 1$. Let $c(n)$ be a treshold such that $0 < c(n) < 1$ and $\lim_{n\to\infty} c(n) = 0$.

Simulation step : generate n random numbers $t_{ik}^{(o)}$ $(i = 1, ..., n)$ which represent the initial posterior probability of cluster $k = 1, ..., K$ having observed f_i.

Stochastic step: generate multinomial numbers

$$e_{qi} = (e_{qi}^k, k = 1, ..., K)$$

of one draw of K categories with probabilities t_{i1k}^q,..., t_{i1k}^q so that all the e_{qi}^k are 0 except one of them equal to 1.

We then get a partition $C_q = (C_{q1}, ..., C_{qK})$ of the sample by letting

$$C_{qk} = \{f_i : e_{qi}^k = 1\}$$

If $\frac{\sum_{i=1,...,N} e_{qi}^k}{N} < c(n)$ for some k, then go to the simulation step (because C_{qk} is too small)

Maximization step : estimate the mixing weights

$$p_{(q+1)k} = \frac{1}{n}[(1-\gamma_q)\sum_{i=1,...,n} t_{ik}^q + \gamma_q \sum_{i=1,...,n} e_{qi}^k].$$

Estimate the parameters of the distribution density h_{qk} of class C_{qk}

Estimation step : the density function h_{qk} depending on a parameter a_{kq}, update a_{kq} and t_{ik}^q according to :

$$a_k^{q+1} = (1-\gamma_q)\frac{\sum_{i=1,...,n} t_{ik}^q b(f_i)}{\sum_{i=1,...,n} t_{ik}^q} + \gamma_q \frac{\sum_{i=1,...,n} e_{qi}^k b(f_i)}{\sum_{i=1,...,n} e_{qi}^k}$$

$$t_{ik}^{q+1} = \frac{p_{(q+1)k} h_{(q+1)k}(f_i)}{\sum_{r=1...K} p_{(q+1)r} h_{(q+1)r}(f_i)}.$$

The mixture is then estimated since p_{qk}, t_{ik}^q and the density parameters converge a.s. as $q \to \infty$ [Celeux, Diebolt (1992)].

3.2. **DS *algorithm.*** Similarly a clustering algorithm [Diday and Schroeder (1976)] can be applied to the vectors $f_1, ..., f_n$ in order to get a partition into K disjoint clusters :

Step 1 - Start with K arbitrary disjoint clusters
Step 2 - Estimate as above the distribution h_k of each cluster $k = 1, ..., K$
Step 3 - Redefine clusters : affect each vector f_i to a cluster j such that

$$h_j(f_i) = \max_k h_k(f_i)$$

Step 4 - Go to step 2 until the goodness of the clustering reaches a desired level.

Step 5 - Having obtained a good partition, the mixture is then estimated by $\sum_{k=1,...,K} p_k h_k$ where

$$p_k = \frac{\#cluster\ k}{n}.$$

4. Mixtures of discrete RDs

We now consider the mixture problem for a RD $X : \Omega \longrightarrow \mathbf{P}(V)$ where V need not be finite. Let $f_i \in \mathbf{P}(V), i = 1, ...n$, be the given observations from a sample. Our method consists in first splitting $V = [0, 1)$ into a dyadic partition

$$\sigma_l = (V_{kl} = [\frac{k-1}{2^l}, \frac{k}{2^l}), k = 1, ..., 2^l)$$

where $l \geq 2$ is an integer, and then applying the preceding algorithms to the probability vectors $f_i(V_{1l}), ..., f_i(V_{kl}), ..., f_i(V_{2^l l}), i = 1, ...n$, more precisely to the vectors $f_i(V_{1l}), ..., f_i(V_{kl}), ..., f_i(V_{2^l - 1l})$ as emphasized below.

Two problems then appear : the choice of the components and the stability of the mixture when we refine the partitions.

We are going to prove that Ferguson and Lo RDs yield a solution to both problems in the discrete case.

Define the random vector $X(\sigma_l)$ by

$$X(\sigma_l)(\omega) = (X(\omega)(V_{1l}), ..., X(\omega)(V_{2^l l})).$$

It follows from the definition that if $X \sim \mathcal{D}(\alpha)$ is a Ferguson RD then $X(\sigma_l)$ $\sim$ a finite dimensional Dirichlet $\mathcal{D}(\alpha(V_{1l}), ..., \alpha(V_{2^l l})) = \mathcal{D}^{\sigma_l}(\alpha)$. Similarly, if $\mathcal{P}_X$ is a simple mixture $\sum_{s=1}^{K} p_s \mathcal{D}_s(\alpha_s; \beta_s)$ of Lo RDs, then the distribution $\mathcal{P}_X^{\sigma_l}$ of the random vector $X(\sigma_l)$ is $\sum_{s=1}^{K} p_s \mathcal{D}_s^{\sigma_l}(\alpha_s; \beta_s)$.

Therefore if f_i , $i = 1, ..., n$ are n observations from X, we can apply S.A.E.M. algorithm to the vectors $f_i(V_{1l}), ..., f_i(V_{2^l - 1l})$ in order to estimate the mixture $\sum_{s=1}^{K} p_s \mathcal{D}_s^{\sigma_l}(\alpha_s; \beta_s)$. Indeed, we drop down the last coordinate of these vectors in order to use the density on S_{2^l} given in formulae (1) in subsection. Note that this density obviously belongs to the exponential family.

Now, as the number of iterations increases, the algorithm yields parameters t_{is}, coefficients p_s and normalized weighted gamma distributions G_s denoted by $t_{is}(\sigma_l)$, $p_s(\sigma_l)$ and $G_s(\sigma_l)$, respectively, since they depend on σ_l.

The question we need to address is then : what is the behaviour of these finite-dimensional mixtures when the partitions σ_l get refined ($l \to \infty$) and do they approximate the distribution of X when this last one is a simple mixture of normalized weighted gamma processes ? Our main result answers positively to this question and can be stated as follows :

4.1. **Main result.** Theorem 3 *Suppose that $\mathcal{P}_X = \sum_{s=1}^{K} p_s \mathcal{D}_s(\alpha_s;\beta_s)$ is a simple mixture of normalized weighted gamma processes where the α_s's are distinct probabilty measures equivalent to the Lebesgue measure on $[0,1]$ and $\beta_s > 0$. Let S_s be the support of $\mathcal{D}_s(\alpha_s;\beta_s)$ so that the support of $\mathcal{P}_X$ is $\cup_{s=1}^{K} S_s$. Then there exists measurable sets $S_s^{'} \subset S_s$ with $\mathcal{P}_X(S_s \backslash S_s^{'}) = 0$ such that if $f_i \in \cup_{s=1}^{K} S_s'$, $i = 1, ..., n$, then S.A.E.M. algorithm applied to the f_i 's and the σ_l 's yields numbers $t_{is}(\sigma_l)$ and $p_s(\sigma_l)$ such that*

$$\lim_{l\to\infty} t_{is}(\sigma_l) = 1 \ \ \textit{if} \ \ f_i \in S_s'$$

$$\lim_{l\to\infty} t_{is}(\sigma_l) = 0 \ \ \textit{if} \ \ f_i \notin S_s'$$

$$\lim_{n\to\infty} \lim_{l\to\infty} p_s(\sigma_l) = p_s.$$

A similar result holds for DS algorithm (see Section 5).

5. Mixtures of continuous RDs

We now consider the case of a RD $X : \Omega \longrightarrow \mathbf{P}(V)$ where the probability measure $X(\omega) = f(\omega).\lambda$ is absolutely continuous w.r.t. λ, with density $f(\omega)$. We will assume that these density functions f lie in some standard function spaces B where approximation by simple functions is possible (by simple function we mean a function such that there exists a finite partition of $[0,1]$ into intervals with f constant on each interval). Hence we can take $\mathrm{B} = \mathcal{C}[0,1]$ the usual Banach space of continuous functions on $[0,1]$, $\mathrm{B} = L_q[0,1]$ $(1 \le q < \infty)$, $\mathrm{B} = D[0,1]$ the usual Skorohod space, and so on.

Let $f_i \in$B , $i = 1, ..., n$ be the corresponding densities for the given observations from a sample $X^{(i)}$, $i = 1, ..., n$. The appproximation assumption on B reduces the mixture problem to the case of simple densities for which we propose the following solution.

5.1. **Simple densities.** Assume that f_i is constant on each interval of a partition $\sigma_l = (0 = x_1 < x_2 < ... < x_l = 1)$ of $[0,1]$ (note that this is the case when we deal with histograms). Refining all the partitions corresponding to the f_i's we may suppose that theses functions are constant on a common partition $\sigma_l = ([\frac{k-1}{2^l}, \frac{k}{2^l})$, $k = 1, ..., 2^l)$ for some integer l.

Therefore we may and do assume that the random vector X is a mixture of $(g_l^{(s)}, ..., g_l^{(s)})$, $s = 1, ..., K$ where $g_l^{(s)}$ is defined w.r.t a finite set $Z^{(s)} = \{Z^{(s)}_{\frac{k}{2^r}}$; $r = 1, 2...l$, $k = 1, 3, ..., 2^r - 1\}$. This means that

$$X = \sum_{s=1}^{K} 1_{(U=u_s)} g_l^{(s)}$$

where U is a discrete r.v, independent of the $Z^{(s)}$ and taking K distinct values $u_1, ..., u_K$.

The following algorithm yields an estimation of the finite mixture.

5.1.1. *Algorithm.* Compute $g_{i,r} = \mathcal{E}^{B_{\sigma_r}} f_i \in R_+^{2^r}$ $i = 1, 2, ..., n,$ $r = 1, 2, ..., l,$ where $\mathcal{E}^{\mathcal{B}_r}$ denotes the conditional expectation w.r.t. $\mathcal{B}_r$.

Ccompute $Z_i = (Z_{i,\ \frac{k}{2^r}}\ , k = 1, 3, .., 2^r - 1)$ for $r = 1, 2, ..., l$

according to formulas (**??**), (**??**), (**??**) given in the proof of theorem 2 (Section 7).

Apply S.A.E.M. or D.S. algorithm to (Z_i) in $R_+^{2^l}$ to get K clusters of these n vectors, the components of the mixture having for distribution the product $\otimes_k d_{k,l}$.

The algorithm estimates the parameters of $d_{k,l}$ for each component $s = 1, ..., K$ and also the mixing weights p_s. Then any discrete variable U such that $\mathcal{P}(U = u_s) = p_s$ and independent of the components yields the desired mixture.

5.2. **Mixtures of Kraft processes.** We will say that X is a mixture of Kraft processes if there exists sequences $Z^{(s)}$ as the above Z and a discrete r.v. U, independent of the $Z^{(s)}$, taking K distinct values $u_1, ..., u_K$, such that

$$X = \sum_{s=1}^{K} 1_{(U=u_s)} Kraft(Z^{(s)})$$

$$\mathcal{P}(U = u_s) = p_s.$$

With the same notation as in theorem 1, the following theorem shows that algorithm 4.3 estimates a mixture of Kraft processes :

Theorem 2
i) The finite-dimensional distributions of Kraft processes are equivalent

ii) $Kraft(Z_j^{(s)})$ *and* $Kraft(Z_j^{(t)})$ *are mutually singular if* $s \neq t$

iii) $\lim_{l\to\infty} t_{is}(\sigma_l) = t_{is}$ $(= 1$ *or* $0)$ *and* $\lim_{n\to\infty} \lim_{\ell\to\infty} p_s(\sigma_l) = p_s$.

5.3. **Weak convergence.** We finally observe that the finite-dimensional mixtures also approximate weakly the distribution of X because the finite-dimensional distributions of X converge weakly as seen in the proposition below.

For any integer $l \geq 2$ let

$$U_l = \{(u_1, ..., u_r ..., u_l) : u_r \geq 0 \text{ and } \sum_{r=1}^{l} u_r = 1\}.$$

Let

$$\sigma_l = (0 = x_1 < x_2 < ... < x_l = 1)$$

be a partition of $[0, 1]$ such that

$$|\sigma_l| = \max_{i=1,...,l-1} |x_{i+1} - x_i| \to 0 \text{ as } l \to +\infty$$

an let

$$T_{\sigma_l}(g) = (\int_0^{x_1} g(s)ds, ..., \int_{x_{l-1}}^{x_l} g(s)ds) \in U_l$$

for any positive $g \in$B such that $\int_0^1 g(s)ds = 1$.

The finite-dimensional distributions $P_{X_j}^{\sigma_l} = P_{X_j} o T_{\sigma_l}^{-1}$ of X_j are probability measures on U_l defined by

$$P_{X_j}^{\sigma_l}(A) = P_{X_j}(T_{\sigma_l}^{-1}(A))$$

for any Borel set $A \subseteq U_l$.

For any $u = (u_1, ..., u_r ..., u_l) \in U_l$ define $h_{\sigma_l, u}$ as the polygonal function

. taking the value $\frac{u_i}{x_{l+1} - x_i}$ at x_i for $i = 1, ..., l-1$,

. taking the value $\frac{u_l}{x_l - x_{l-1}}$ at $x_l = 1$,

. affine between the points x_i.

Then the weak convergence result can be stated as follows :

Proposition 1 *If* $\mathrm{B} = C[0,1]$ *or* $L_q[0,1]$, the for *any bounded continuous positive* $\Psi : \mathrm{B} \to \mathrm{C}$ *(the complex field), we have*

$$\int_{U_l} \Psi(h_{\sigma_l, u}) dP_{X_j}^{\sigma_l}(u) \to \int_{g \in B} \Psi(g) dP_{X_j}(g) as\ \ell \to \infty.$$

A similar result holds if $\mathrm{B} = D[0,1]$, the usual Skorohod space (see the proof).

6. Continous time process of random distributions

6.1. **Continuous time process of Ferguson RDs.** Consider any standard polish space $\mathcal{F}$ of real functions defined on an interval $I \subset [0, \infty)$, e.g. the space $\mathcal{C}(I)$ (resp. $\mathcal{D}(I)$) of continuous (resp. cadlag) functions. For any time $t \in I$, let π_t : $x \longrightarrow x(t)$ denote the usual projection at time t from the space $\mathcal{F}$ to R. Recall that π_t maps any measure μ on $\mathcal{F}$ on a measure $\pi_t \mu$ on R defined as $\pi_t \mu(A) = \mu(\pi_t^{-1}(A))$ for any Borel subset A of R.

The following theorem defines a continous time process (X_t) such that each X_t is a Ferguson-Dirichlet random distribution.

Theorem 4 *Let α be any finite measure on $\mathcal{F}$, let X be a Ferguson-Dirichlet random distribution $\mathcal{D}(\alpha)$ and let $X_t = \pi_t X$. Then the time continuous process $(X_t)_{t \in I}$ is such that for each $t \in I$, X_t is a Ferguson-Dirichlet random distribution $\mathcal{D}(\alpha_t)$ where $\alpha_t = \pi_t \alpha$.*

If $V^{(i)}$ is any i.i.d. sequence on $\mathcal{F}$ such that $V^{(i)} \sim \frac{\alpha}{\alpha(\mathcal{F})}$ and $X(\omega) = \sum_{i=1}^{\infty} p_i(\omega) \delta_{V^{(i)}(\omega)}$ where (p_i) has a Poisson-Dirichlet distribution $\mathcal{PD}(\alpha(\mathcal{F}))$, then $X_t(\omega) = \sum_{i=1}^{\infty} p_i(\omega) \delta_{V^{(i)}(t)(\omega)}$.

Proof : Let $k \in \{1, 2, 3, ...\}$ and $A_1, ..., A_k$ a measurable partition of R. Then $\pi_t^{-1}(A_1), ..., \pi_t^{-1}(A_k)$ is a measurable partition of $\mathcal{F}$ so that, by definition of X, the joint distribution of the random vector $(X(\pi_t^{-1}(A_1)), ..., X(\pi_t^{-1}(A_k)))$ is Dirichlet with parameters $(\alpha(\pi_t^{-1}(A_1)), ..., \alpha(\pi_t^{-1}(A_k))$. In other words $(X_t(A_1)), ..., X_t(A_k))$ is Dirichlet with parameters $(\alpha_t(A_1), ..., \alpha_t(A_k))$ and $X_t \sim \mathcal{D}(\alpha_t)$.

A consequence of the definition of π_t is that $\pi_t(\sum_{i=1}^{\infty} \mu_i) = \sum_{i=1}^{\infty} \pi_t \mu_i$ for any sequence of positive measures on $\mathcal{F}$ and $\pi_t(\lambda \mu) = \lambda \pi_t(\mu)$ for any positive real number λ. Hence if $V^{(i)}$ is any i.i.d. sequence on $\mathcal{F}$ such that $V^{(i)} \sim \frac{\alpha}{\alpha(\mathcal{F})}$ and $X(\omega) = \sum_{i=1}^{\infty} p_i(\omega) \delta_{V^{(i)}(\omega)}$ where (p_i) has a Poisson-Dirichlet distribution $\mathcal{PD}(\alpha(\mathcal{F}))$, then $X_t(\omega) = \pi_t(X(\omega)) = \sum_{i=1}^{\infty} p_i(\omega) \pi_t(\delta_{V^{(i)}(\omega)}) = \sum_{i=1}^{\infty} p_i(\omega) \delta_{V^{(i)}(t)(\omega)}$, the last equality being due to the fact that $\pi_t(\delta_f) = \delta_{f(t)}$ as easily seen. In addition the sequence $V^{(i)}(t)$ is i.i.d. with $V^{(i)}(t) = \pi_t(Vt) \sim \pi_t(\frac{\alpha}{\alpha(\mathcal{F})}) = \frac{1}{\alpha(\mathcal{F})} \pi_t(\alpha) = \frac{1}{\alpha_t(R)} \alpha_t$

and (p_i) has a Poisson-Dirichlet distribution $\mathcal{PD}(\alpha(\mathcal{F})) = \mathcal{PD}(\alpha_t(R))$ so that the preceeding expression of $X_t(\omega)$ is exactly the expression of a Ferguson-Dirichlet random distribution $\mathcal{D}(\alpha_t)$ as a random mixture of random Dirac masses.

6.2. **Continuous time process of Kraft RDs.** Let $X_t(\omega)$ be Dirichlet as above. Then $X_t(\omega)(A)$ is Beta for any subset A. As Kraft construction depends on Beta distributions, we see that we can generalize the construction.

7. Application to Internet traffic

In his section we present an application of the preceding clustering method which has been developped with K. Salamatian and A. Soule at the lip6 laboratory of Paris VI university. The complete details can be found in Sigmetrics'05

Internet is nowadays a large highway network crossed everyday by the informations of millions of users throughout the world. Network users send packets of information using various protocols such as UDP or TCP over IP.

7.1. **Flows.** Packet transmissions induce flows that are mixed up at routers to create larger and larger aggregated flows that run from source to destination through links. We define our flows as the sequence of packets going from a network prefix announced through BGP (Border Gateway Protocole) to another BGP prefix. Therefore every flow is characterized by a source BGP prefix and a destination BGP prefix.

7.2. **Classification purposes.** At each instant of time, several thousands of such flows may cross Internet backbone links and each one will have its specific behaviour and characteristics. The objective here is to present a way of classifying these flows for managing them.

The litterature have used widely of animal name as elephant, mice, tortoise, dragon, etc, for addressing flows belonging to each class. We will not derogate from this tradition and we consider our research as a safari where we want to hunt different type of flow behaviour.

Classification can be used for several purposes. We describe here some obvious ones.

First, classification enables us to give a concise and simple description of the otherwisely random like traffic flows, in term of behavioural classes. Traffic flows observed at specific point of network are described by means of characteristics of classes rather than by characteristics of each particular flow.

Next, concerning more application oriented purposes, classification can ease the burden of traffic engineering, but it can also concerns the distributed denial of services (DDOS) by making easier intrusion detection. Indeed it suffices to apply the principle of *divide & conquer* and to use the characteristic of each class. An example of such classification is based on the so called *Elephant and mice* phenomenon. Studies of the Internet traffic at the level of network prefixes, fixed length prefixes, TCP flows, Autonomous System's, and WWW traffic, have shown that in all cases a very small percentage of flows carries the largest part of the information. This phenomenon can be the base for a classification that will make traffic engineering easier by using the fact that one need to manage only a small number of elephant flows to solve most of network problems.

7.3. **Stable classes.** Our aim is to investigate the existence of stable and meaningful classes of flow with similar behaviour inside a network link. By stable we mean that these classes should persistent for a large period of time, how large the period of time being determined by an application of the classification.

7.4. **Empirical histograms.** Previous studies have tried to characterize a flow by its mean rate over a period of time. However the mean rate is not a sufficient parameter to characterize the behaviour of a flow. In this work we use the empirical histogram as a criteria of classification of flows. We believe that this give a better characterization of flow behaviour.

7.5. **Dataset.** The data used in this paper comes from packet traces collected in the core of a major Tier-1 ISP network. Optical splitters are used in conjunction with passive monitoring equipment to collect 44-byte headers from every IP packet traversing monitored link. Monitoring equipment has been installed in three major POPs in the USA. We use data two different OC-12 links, one in an east coast POP and the other in a west coast POP, collected on July 24, 2001. The links used are two hops away from the periphery of the network so that traffic towards specific destinations exhibit sufficient level of aggregation. Our traces constitute 3 1/2 days of continuous data.

Packet trace collection was accompanied by the collection of the BGP routing tables at the corresponding POPs. Those BGP tables are default-free and contain approximatively 120K entries. We calculate the volume of traffic headed towards each BGP destination and computed the average bandwidth of each flow over 5 minute time intervals. We found that in any given measurement interval, approximatively 90% of the network prefixes had no traffic travelling towards them. We thus define a flow to be *active* if it transmits at least one packet during the measurement interval.At each instant of time we have roughly 2000 flows in the observed OC-12 links. Before classifying the data, the histograms are obtained over 24 hours period (corresponding to 288 five minutes samples) for each flow observed during the monitoring period.

7.6. **Bins.** The main point while transforming data from measured values to random distributions is the choice of the bins center $\mathcal{B}$. If one choose poorly the centers then some bin could be empty and prevent the classification algorithm to work. With this in mind we try to implement a good algorithm to find the center which will guarantee that all bins will have a relatively reasonable number of members. We use all measurement observed over the measurement period to find the bin centers by an iterative process. First we set the two bin center to be the min and max values of observed bandwidth among all flows. Then we cut the most populated bin in two and recompute the cardinal of each bin. We continue splitting the most populated bin as long as we arrive to the desired number of bins. We have used in this paper 12 bins. This leads to quasi-logarithmic bin centers. We have observed that the classification remains unchanged when the number of bins is greater than 12, confirming the theoretical result obtained in a preceding section.

These bin centers are used to derivate histogram for each flow observed over the network. The classification procedure is applied to these histograms with a predefined mixture of order K.

At each instant of time we have roughly 2000 flows in the observed OC-12 links. Before classifying the data, the histograms are obtained over 24 hours period (corresponding to 288 five minutes samples) for each flow observed during the monitoring period.

7.7. **Classification with two classes.** For two classes ($K = 2$), the loglikelihood behaviour shows that globally the likelihood goes increasing and reach a maximum, meaning that the algorithm converges. The oscillations are due to the stochastic behaviour of the SAEM algorithm and their amplitudes go decreasing with γ_q going to 0. Each run of the estimation algorithm takes around a couple of minutes, with most of the time spent in the histogram generation phase.

At the end of the algorithm we find that 1142 (64%) are classified as class 1 and 658 (36%) flows belongs to class 2. The mean behaviour of the first class has an exponential like behaviour that is seen by the linear alignement we have after the second point. In that class, flows have with large probability (around 40%) a bandwidth close to zero and the probability falls exponentially. On the other hand flows that are in the second class experience larger values and are almost never close to zero. This empirical classification is related to the well known elephant and mice phenomenon. We therefore call the first class of flows the class of mices and the second one the class of elephants. We have therefore find a way of detecting elephant and mices without any *a priori*.

7.8. **Classification with more classes.** We have calibrate a 3 classes model as well as a 4 classes model to our monitored network link. In the 3 classes scenario, classes 2 and 3 results from splitting class 2 in the two classes scenario but class 3 is of interest. In the 4 classes scenario, the new class hase a very small number of members and is meaningless. We conclude that the three classes classification is sufficient.

8. Proofs

PROOF OF THEOREM 1 i).

Let $X : (\Omega, \mathcal{F}, \mathcal{P}) \to \mathrm{P}([0,1])$ be measurable, and let $\mathcal{G}$ be the $\sigma-$algebra on $\mathrm{P}([0,1])$ generated by the mappings

$$\varphi_A : \mathbf{P}([0,1]) \to [0,1]$$

defined by

$$\varphi_A(P) = P(A)$$

for any Borel set A in $[0,1]$ (i.e. $\mathcal{G}$ is the smallest $\sigma-$algebra for which all the φ_A are measurable).

For any partition

$$\sigma_l = ([\frac{k-1}{2^l}, \frac{k}{2^l}), k = 1, ..., 2^l)$$

of $[0,1]$, let $\mathcal{G}_l$ be the $\sigma-$algebra generated by the mappings

$$\varphi_k = \varphi_{[\frac{k-1}{2^l}, \frac{k}{2^l})}, k = 1, ..., 2^l$$

so that we have

$$\mathcal{G}_l \subseteq \mathcal{G}_{l+1} \subseteq \mathcal{G}$$

and

$$\mathcal{G} = \vee_{l=1}^{+\infty} \mathcal{G}_l.$$

For any $B \in \mathcal{G}_l$ there exists Borel sets O_k in $[0,1]$ such that

$$\begin{aligned} B &= \{P \in \mathbf{P}([0,1]) : \varphi_k(P) \in O_k \ , \ k = 1, ..., 2^l\} \\ &= \{P \in \mathbf{P}([0,1]) : P([\frac{k-1}{2^l}, \frac{k}{2^l})) \in O_k \ , \ k = 1, ..., 2^l\}. \end{aligned}$$

Let

$$\alpha_k = \alpha([\frac{k-1}{2^l}, \frac{k}{2^l})), \ k = 1, ..., 2^l$$

and

$$d_l(\alpha, y) = d((\alpha_1, ..., \alpha_{2^l})|(y_1, ..., y_{2^l-1}))$$

where d is defined by (**??**) in Section 1.
If X is a Dirichlet process $\mathcal{D}(\alpha)$ then

$$\begin{aligned} \mathcal{D}(\alpha)(B) &= \mathcal{P}\{\omega \in \Omega : X(\omega) \in B\} \\ &= \mathcal{P}\{\omega \in \Omega : X(\omega)[\frac{k-1}{2^l}, \frac{k}{2^l}) \in O_k \ , \ k = 1, ..., 2^l\} \\ &= \mathcal{D}(\alpha_1, ..., \alpha_{2^l-1})(O_1 \times ... \times O_{2^l}) \\ &= \int_{O_0 \times ... \times O_{2^l-1} \cap S_{2^l}} d_l(\alpha, y) dy_1 ... dy_{2^l-1}. \end{aligned}$$

Also observed that, if 1_B denotes the indicator of B, then

$$1_B = 1_{O_1 \times ... \times O_{2^l}}(\varphi_1, ..., \varphi_{2^l}) = 1_{O_1 \times ... \times O_{2^l-1} \cap S_{2^l}}(\varphi_1, ..., \varphi_{2^l-1})$$

since $\varphi_{2^l} = 1 - \varphi_1 - ... - \varphi_{2^l-1}$.
Thus

$$\begin{aligned} \mathcal{D}(\alpha)(B) &= \int 1_{O_1 \times ... \times O_{2^l-1} \cap S_{2^l}}(\varphi_1, ..., \varphi_{2^l-1})(P) d\mathcal{D}(\alpha)(P) \\ &= \int_{S_{2^l}} 1_{O_1 \times ... \times O_{2^l-1}} d_l(\alpha, y) dy_1 ... dy_{2^l-1}. \end{aligned}$$

More generally for any $\mathcal{G}_l$-measurable positive function $f(\varphi_1, ..., \varphi_{2^l-1})$, with f mesurable, positive and defined on S_{2^l}, we have

$$\begin{aligned} & \int f(\varphi_1, ..., \varphi_{2^l-1})(P) d\mathcal{D}(\alpha)(P) \\ &= \int_{S_{2^l}} f(y_1, ..., y_{2^l-1}) d_l(\alpha, y) dy_1 ... dy_{2^l-1}. \end{aligned}$$

This implies that

$$\begin{aligned} \mathcal{D}(\alpha')(B) &= \int_{S_{2^l}} 1_{O_1 \times ... \times O_{2^l-1}} d_l(\alpha', y) dy_1 ... dy_{2^l-1} \\ &= \int_{S_{2^l}} 1_{O_1 \times ... \times O_{2^l-1}} \frac{d_l(\alpha', y)}{d_l(\alpha, y)} d_l(\alpha, y) dy_1 ... dy_{2^l-1} \\ &= \int (1_{O_1 \times ... \times O_{2^l-1}} \frac{d_l(\alpha', .)}{d_l(\alpha, .)})(\varphi_1, ..., \varphi_{2^l-1})(P) d\mathcal{D}(\alpha)(P) \\ &= \int (1_B)(P)(\frac{d_l(\alpha', .)}{d_l(\alpha, .)})(\varphi_1, ..., \varphi_{2^l-1})(P) d\mathcal{D}(\alpha)(P). \end{aligned}$$

This clearly shows that if α and α' are equivalent to Lebesgue measure on $[0,1]$ then the restriction $\mathcal{D}_l(\alpha)$ of $\mathcal{D}(\alpha)$ to $\mathcal{G}_l$ is equivalent to the restriction $\mathcal{D}_l(\alpha')$ of $\mathcal{D}(\alpha')$ to $\mathcal{G}_l$ and

$$
\begin{aligned}
(8.1) \qquad \frac{d\mathcal{D}_l(\alpha')}{d\mathcal{D}_l(\alpha)}(P) &= \frac{d_l(\alpha',.)}{d_l(\alpha,.)})(\varphi_1,...,\varphi_{2^l-1})(P) \\
&= \frac{\Gamma(\alpha'_1+...+\alpha'_{2^l})}{\Gamma(\alpha'_1)...\Gamma(\alpha'_{2^l})}\frac{\Gamma(\alpha_1)...\Gamma(\alpha_{2^l})}{\Gamma(\alpha_1+...+\alpha_{2^l})} \\
&\quad P([0,\frac{1}{2^l}))^{\alpha'_1-\alpha_1}...P([\frac{2^l-2}{2^l},\frac{2^l-1}{2^l}))^{\alpha'_{2^l-1}-\alpha_{2^l-1}} \\
&\quad (1-\sum_{k=1,...,2^l} P([\frac{k-1}{2^l},\frac{k}{2^l})))^{\alpha'_{2^l}-1} \\
&\quad (1-\sum_{k=1,...,2^l} P([\frac{k-1}{2^l},\frac{k}{2^l})))^{1-\alpha_{2^l}}.
\end{aligned}
$$

This yields the equivalence on $\mathcal{G}_l$ of $\mathcal{D}_l(\alpha;\beta)$ and $\mathcal{D}_l(\alpha';\beta')$, if β and β' are strictly positive functions. □

PROOF OF THEOREM 1 ii).

It is well-known and easy to prove that $\frac{d\mathcal{D}_l(\alpha';\beta')}{d\mathcal{D}_l(\alpha;\beta)}$ is a $\mathcal{D}(\alpha;\beta)-$martingale w.r.t. $(\mathcal{G}_l)_l$.

If X is a Dirichlet process $\mathcal{D}(\alpha)$ then it is proved in [Ferguson (1973)] that there exists an i.i.d sequence $V_{n,\alpha}$ of random variables

$$V_{n,\alpha} \sim \alpha$$

such that the support of the discrete probability measure $X(\omega)$ is contained in the random set

$$\{V_{1,\alpha}(\omega), V_{2,\alpha}(\omega), ..., V_{n,\alpha}(\omega), ...\}.$$

But if α and α' are different and equivalent to the uniform distribution λ on $[0,1]$, then

$$\int \sqrt{\frac{d\alpha}{d\lambda}\frac{d\alpha'}{d\lambda}} < 1$$

by Cauchy-Schwarz inequality and

$$\left(\int \sqrt{\frac{d\alpha}{d\lambda}\frac{d\alpha'}{d\lambda}}\right)^n \to 0 \text{ as } n \to +\infty.$$

This implies that the product measures $\otimes_{n=1}^{\infty}\alpha$ and $\otimes_{n=1}^{\infty}\alpha'$ are mutually singular by a theorem of Kakutani (see e.g., Hewitt and Stromberg p. 453).

Hence, if

$$V_{n,\alpha'} \sim \alpha'$$

is an i.i.d sequence, there exists two disjoint sets S and T in $[0,1]^{IN}$ such that

$$
\begin{aligned}
\mathcal{P}\{\omega &\in \Omega : \{V_{1,\alpha}(\omega), V_{2,\alpha}(\omega), ..., V_{n,\alpha}(\omega), ...\} \in S\} = 1 \\
\mathcal{P}\{\omega &\in \Omega : \{V_{1,\alpha'}(\omega), V_{2,\alpha'}(\omega), ..., V_{n,\alpha'}(\omega), ...\} \in T\} = 1.
\end{aligned}
$$

Thus if Y is a Dirichlet process $\mathcal{D}(\alpha')$, we have

$$X(\omega) \in \{P \in \mathbf{P}([0,1]) : \text{support of } P \in S\}$$

and

$$Y(\omega) \in \{P \in \mathbf{P}([0,1]) : \text{support of } P \in T\}$$

for a.a. ω. But as these two sets of probability measures are disjoint, $\mathcal{D}(\alpha)$ and $\mathcal{D}(\alpha')$ are mutually singular. The same holds for $\mathcal{D}(\alpha;\beta)$ and $\mathcal{D}(\alpha';\beta')$ for strictly positive functions β and β'. □

PROOF OF THEOREM 2.

$F_{j,l}$ has, except at $\frac{k}{2^l}$, a derivative $g_{j,l}$. If x is written as $x = \sum_{r=1}^{\infty} \frac{\varepsilon_r(x)}{2^r}$ and $k_r(x)$ is defined by $\frac{k_r(x)}{2^r} < x < \frac{k_r(x)+1}{2^r}$, then we have

$$g_{j,l}(x) = 2^l \prod_{r=1}^{l} Z_{\frac{k_r(x)+1}{2^r}}{}^{1-\varepsilon_r(x)} (1 - Z_{\frac{k_r(x)}{2^r}})^{\varepsilon_r(x)}.$$

This implies that

$$Z_{\frac{1}{2}} = \frac{1}{2} g_{j,1}(x) \text{ for } 0 < x < \frac{1}{2} \tag{8.2}$$

$$Z_{\frac{1}{4}} = \frac{1}{2} \frac{g_{j,2}}{g_{j,1}}(x) \text{ for } 0 < x < \frac{1}{4} \tag{8.3}$$

$$Z_{\frac{3}{4}} = \frac{1}{2} \frac{g_{j,2}(x)}{g_{j,1}(x)} \text{ for } \frac{1}{2} < x < \frac{3}{4} \tag{8.4}$$

and so on.

Moreover, $g_{j,l}$ is a $\mathcal{P} \otimes \lambda-$ martingale w.r.t. $(\mathcal{F} \otimes \mathcal{B}_l)_l$, where $\mathcal{B}_l$ is the finite $\sigma-$algebra generated by the dyadic intervals $[\frac{k}{2^l}, \frac{k+1}{2^l}), k = 0, ..., 2^l - 1$.

As $l \to +\infty$, condition (**??**) implies that

$$F_{j,l}(x) \to F_j(x) \text{ and } g_{j,l} = F'_{j,l}(x) \to g_j(x) = F'_j(x).$$

Observe now that F_j is a function of $Z_j = (Z_{\frac{k}{2^r},j})$, say :

$$F_j = \Psi_j(Z_j)$$

Moreover, if $Z'_j = (Z'_{\frac{k}{2^r},j}) \neq Z_j$ is a sequence having the same properties as Z_j, then

$$\Psi_j(Z_j) = \Psi_j(Z'{}_j) \Rightarrow (Z_j) = (Z'{}_j). \tag{8.5}$$

Indeed $F_j = \Psi_j(Z_j)$ determines g_j and thus determines $g_{j,l} = E^{\mathcal{F} \otimes \mathcal{B}_l}(g_j)$ and (**??** - **??**- **??**) imply that the $g_{j,r}$'s , $r = 1, ...l$ completely determine $Z_{\frac{k}{2^l}}, k = 0, ..., 2^l$.

Now by (**??**) the distribution of $Z_{\frac{k}{2^r},j}$ is equivalent to that of $Z'_{\frac{k}{2^r},j}$.

Therefore, if their derivatives w.r.t. Lebesgue mesure on $[0,1]$ are choosen so that they satisfy Kakutani's theorem condition, then the distribution of Z_j (which is the product of that of $Z_{\frac{k}{2^r},j}$) and the distribution of Z'_j are mutually singular : their support, say $S(Z_j)$ and $S(Z'{}_j)$ respectively, are disjoint.

Thus the support of the distribution of $\Psi_j(Z_j)$, which is included in $\Psi_j(S(Z_j))$, is disjoint from the support of the distribution of $\Psi_j(Z'{}_j)$, since (**??**) implies that

$$\Psi_j(S(Z_j)) \cap \Psi_j(S(Z'{}_j)) = \Psi_j(S(Z_j) \cap S(Z'{}_j)) = \emptyset.$$

Thus two distinct Kraft processes $\Psi_j(Z_j)$ and $\Psi_j(Z'{}_j)$ are mutually singular.

On the other hand the distribution of $E^{\mathcal{F}\otimes\mathcal{B}_l}(\Psi_j(Z_j))$ and that of $E^{\mathcal{F}\otimes\mathcal{B}_l}(\Psi_j(Z'_j))$ are equivalent since $E^{\mathcal{F}\otimes\mathcal{B}_l}(\Psi_j(Z_j)) = g_{j,l}$ is a function of $(Z_{j,\frac{k}{2^l}}, k = 0, ..., 2^l)$ and $E^{\mathcal{F}\otimes\mathcal{B}_l}(\Psi_j(Z'_j))$ is a function of $(Z'_{j,\frac{k}{2^l}}, k = 0, ..., 2^l)$.□

PROOF OF THEOREM 3.

In the S.A.E.M. algorithm, the random numbers t^o_{ik} , wich represent the initial probability that individual i belongs to class k, does not depend on the dimension l. Also note that the partition C_0 and the numbers $p_{0k} > 0$ in iteration 0 does not depend on l.

As G_r and G_k are mutually singular if $r \neq k$, the martingale theorem implies that

$$\frac{dG_{rl}}{dG_{kl}}(P) \to 0 \text{ as } l \to +\infty$$

for all $P \in S'_k \subset S_k$ with

$$G_k(S_k \backslash S'_k) = 0.$$

Observing that $\mathcal{P}_X(S_k\backslash S'_k) = \sum_{s=1}^K p_s G_s(S_k\backslash S'_k) = 0$, and replacing P with f_i we see by (**??**) that for $\mathcal{P}_X$−almost all f_i in S_k :

$$\lim_{l\to+\infty} \frac{G_{0rl}(f_i^*(\sigma_l))}{G_{0kl}(f_i^*(\sigma_l))} = 0.$$

Then by step E of S.A.E.M algorithm 3.3.1.

$$\frac{1}{t^1_{ik}(\sigma_l)} = 1 + \sum_{r\neq k} \frac{p_{0r}G_{0rl}(f_i^*(\sigma_l))}{p_{0k}G_{0kl}(f_i^*(\sigma_l))} \to 1.$$

Since $t^1_{ik}(\sigma_l) + \sum_{r\neq k} t^1_{ir}(\sigma_l) = 1$ and $t^1_{ir}(\sigma_l) \geq 0$ we also get

$$t^1_{ik}(\sigma_l) \to 1,\ t^1_{ir}(\sigma_l) \to 0 \text{ if } r \neq k \text{ for } \mathcal{P}_X - \text{almost all } f_i \text{ in } S_k.$$

Interverting r and k we arrive at the announced result :

$$\lim_{l\to+\infty} t^1_{ik}(\sigma_l) = 1 \text{ for } \mathcal{P}_X - \text{almost all } f_i \text{ in } S_k$$

and

$$\lim_{l\to+\infty} t^1_{ik}(\sigma_l) = 0 \text{ for } \mathcal{P}_X - \text{almost all } f_i \text{ in } S_r \text{ , if } r \neq k.$$

By step S of S.A.E.M. algorithm, we see that the partition C_1 and the numbers p_{1r} are only determined by the numbers $t^1_{ik}(\sigma_l)$.

Since $\frac{\sum_{i=1,...,n} e^k_{1i}}{n} \geq c(n)$ (otherwise S.A.E.M. algorithm is reinitialized),we also have $\frac{\sum_{i=1,...,n} t^1_{ki}}{n} \geq c(n)$ by taking expectations. Hence

$$p_{2k} = \frac{1}{n}[(1-\gamma_q)\sum_{i=1,...,n} t^1_{ik} + \gamma_q \sum_{i=1,...,n} e^k_{1i}] \geq c(n)$$

and therefore $\frac{p_{2r}}{p_{2k}} \leq \frac{1}{c(n)}$. Hence the above limits also hold for $t^2_{ik}(\sigma_l)$ and more generally for $t_{ik}(\sigma_l) = t^Q_{ik}(\sigma_l)$ (recall that $p_{qk}(\sigma_l)$ and $t_{qk}(\sigma_l)$ converge a.s. as $q \to \infty$).

Since $\lim_{q\to\infty} \gamma_q = 0$ the above limits imply that

$$\lim_{l\to\infty} p_k(\sigma_l) = \frac{\sum_{i=1,...,n} \lim_{l\to\infty} t_{ik}(\sigma_l)}{n} = \frac{\sum_{i=1,...,n} I_{(X^{(i)}\in S_k)}}{n}.$$

But

$$X^{(1)}, ..., X^{(n)} \ i.i.d. \sim X$$

implies that

$$\frac{1}{n}\sum_{i=1}^{n}\mathcal{E}(I_{(X^{(i)}\in S_k)}) = \frac{1}{n}\sum_{i=1}^{n}\mathcal{P}(X^{(i)}\in S_k) = \mathcal{P}_X(S_k) = p_k$$

and even that, for almost all observations,

$$\lim_{n\to\infty}\lim_{l\to\infty} p_k(\sigma_l) = \lim_{n\to\infty}\frac{\sum_{i=1,...,n} I_{(X^{(i)}\in S_k)}}{n} = \mathcal{P}_X(S_k) = p_k.$$

We conclude that the finite-dimensional mixture estimates well estimate a mixture of gamma processes.□

PROOF OF THEOREM 3 for DS algorithm.

A similar result holds for the second algorithm 3.3.2.

Indeed step 1 trivially does not depend on l.

Following the preceding notations, let $G_k(\sigma_l)$ be the density estimated in step 2.

By step 3 individual i is affected to class j if $G_{jl}(f_i(\sigma_l)^*) \geq G_{kl}(f_i(\sigma_l)^*)$ for all $k = 1, ..., K$, or equivalently if $\frac{G_{jl}(f_i(\sigma_l)^*)}{G_{kl}(f_i(\sigma_l)^*)} \geq 1$.

But as seen above, for l large enough, we can consider that $\frac{G_{rl}(f_i(\sigma_l)^*)}{G_{kl}(f_i(\sigma_l)^*)}$ does not depend on l since it is closed to 0 or $+\infty$. Hence the clusters determined in step 3 can be considered as independent of l. The same trivially holds for steps 4 and 5. □

PROOF OF PROPOSITION 1.

Clearly, the mapping $t \to h_{\sigma_l,t}$ is continuous on U_l and therefore $t \to \Psi(h_{\sigma_l\ ,t})$ is bounded continuous.

By definition of $P^{\sigma_l}_{X_j}$ we then have

$$\int_{F_l}\Psi(h_{\sigma_l,t})d\mathcal{P}^{\sigma_l}_{X_j}(t) = \int_{g\in\mathcal{C}^+[0,1]}\Psi(h_{\sigma_l,T_{\sigma_l}(g)})d\mathcal{P}_{X_j}(g).$$

But $h_{\sigma_l,T_{\sigma_l}(g)}$ is nothing but the polygonal function, say g_{σ_l} ,

. taking the values $\frac{\int_{x_i}^{x_{i+1}} g(s)ds}{x_{i+1}-x_i}$ at the point x_i of the subdivision, for $i = 1, ..., l-1$,

. taking the value $\frac{\int_{x_{l-1}}^{x_l} g(s)ds}{x_l-x_{l-1}}$ at $x_l = 1$,

. affine between the points x_i.

So, we obtain

$$\int_{F_l}\Psi(h_{\sigma_l,t})d\mathcal{P}^{\sigma_l}_{X_j}(t) = \int_{g\in\mathcal{C}^+[0,1]}\Psi(g_{\sigma_l})d\mathcal{P}_{X_j}(g).$$

If B $= \mathcal{C}[0,1]$, then the uniform continuity of any fixed $g \in \mathcal{C}^+[0,1]$ implies

$$\lim_{l\to\infty} g_{\sigma_l} = g$$

for the usual supremum norm in $\mathcal{C}^+[0,1]$, since $|\sigma_l| = \max_{i=1,...,l-1}|x_{i+1}-x_i| \to 0$ as $l \to \infty$.

Hence $\lim_{l\to\infty} \Psi(g_{\sigma_l}) = \Psi(g)$, and the announced result is a consequence of Lebesgue dominated convergence theorem, since Ψ is bounded and P_{X_j} is a probability measure.

To prove the result if $\mathrm{B} = L_q$, observe that the mapping

$$g \to g_{\sigma_l} \text{ is linear and positive } (g \geq 0 \Rightarrow g_{\sigma_l} \geq 0)$$

and hence continuous in L_q. Moreover, for g continuous, $\lim_{l\to\infty} g_{\sigma_l} = g$ for the supremum norm and thus for the $L_q[0,1]-$norm. Then, a standard approximation argument yields $\lim_{l\to\infty} g_{\sigma_l} = g$ in $L_q[0,1]-$norm, for $g \in L_q$. The other points of the proof in $C[0,1]$ apply in $L_q[0,1]$. □

PROOF OF PROPOSITION 1 FOR SKOROHOD SPACES.

If $\mathrm{B} = D[0,1]$, we need a different definition for g_{σ_l}. Given $g \in D[0,1]$ and $l \geq 1$, there exists a minimal subdivsion of $[0,1]$, say $\tau_l = (0 = t_1 < t_2 < ... < t_{k_l} = 1)$, such that

$$w_g[t_i, t_{i+1}) = \sup\{|g(s) - g(t)|, s, t \in [t_{i-1}, t_i)\} < \frac{1}{l}, i = 1, ..., k_l - 1,$$

(apply lemma 1 p.110 in [Billingsley (1968)]).

Then, let $g_{\sigma_l} = g_{\tau_l}$ be the function in $D[0,1]$ which takes the value $\frac{\int_{t_i}^{t_{i+1}} g(s)ds}{t_{i+1}-t_i}$ over the interval $[t_i, t_{i+1})$. As $\sup\{|g_{\sigma_l}(s) - g(t)|, s, t \in [0,1]\} < \frac{1}{l}$ we have $g_{\sigma_l} \to g$ in the Skorohod topology and therefore $\Psi(g_{\sigma_l}) \to \Psi(g)$ as $l \to \infty$. □

REFERENCES

Aboa, J. P. and Emilion, R. (2000). Decision tree for probabilistic data, *Dawak 2000, Lect. N. in comp. Sci.* N°**1874**, 393-398.

Agrawala (1970). Learning with a probabilistic teacher. *IEEE, Information theory,* **16***, 4.*

Antoniak, C. E. (1974). Mixtures of Dirichlet processes. *Ann. Statist.* **2**, 6, 1152-1174.

Billingsley, P. (1968). *Probability metric spaces.* Wiley.

Cooper (1964). Nonsupervised adaptative signal detection and pattern recognition. *Information and Control.* **7**.

Celeux, G. and Diebolt, J. (1992). A stochastic approximation type EM algorithm for the mixture problem. *Stochastics Stochastics*
Rep. **41** 119-134.

Darwich, A. (2001). About the absolute continuity and orthogonality of two probability measures. *Stat. and prob. letters* **52**, 1, 1-8.

Dempster, Laird and Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm, *JRSS.* B. **39**.

Diday, E. (1989). Introduction à l'approche symbolique en analyse de données. *RAIRO,* **23**, 2.

Diday, E. and Schroeder, A. (1976). A new approach in mixing distributions detection. *RAIRO, operational research*, **10**, 6

Dykstra, R. L. and Laud, P. (1981). A Bayesian nonparametric approach to reliability. *Ann. Statist.* **9** 356-367.

Doss, H. (1984). Bayesian nonparametric estimation. *Ann. Statist.* **22** 763-1786.

FERGUSON, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1** 209-230.

FERGUSON, T.S. (1974). Prior distributions on spaces of probability measures. *Ann. Statist.* **2** 615-629.

HEWITT, E. and STROMBERG, K. (1969). *Real and abstract analysis.* Springer Verlag.

KINGMAN, J. F. C. (1975). Random discrete distributions. *J. Roy. Statist. Soc. B,* **37**, 1-22.

KRAFT, C. H. (1964). A class of distribution function processes which have derivatives. *J. Appl. Prob.* **1**, 385-388.

LO, A. Y. (1982). Bayesian nonparametric statistical inference for Poisson point process. *Z. Wahrsch. Verw. Gebiete* **59** 55-66.

MAKOV and SMITH (1976). Quasi Bayes procedures for unsupervised learning. *Proc. IEEE. Conf. on Decision and Control.*

PEARSON (1894) Contribution to the mathematic theory of evolution. *Philos. Trans. Soc.* **185**.

PITMAN, J. and YOR, M. (1996). Random discrete distributions derived from self-similar random sets. *EJP,* **1**, 1-28.

QUANDT and RAMSEY (1978). Estimating mixtures of normal distributions and switching regression. *JASA,* **73**.

SCHROEDER, A. (1976). Analyse d'un mélange de distributions. *Rev. statist. appli*, **XXIV**, 1

SCHWEIZER, B. (1985). Distribution functions : numbers of the future. *Proceed. of the 2nd Napoli meeting on "The mathematics of fuzzy systems".* Instit. Mat., Univ. Napoli. A. di Nola - A. Ventre (Eds.), 137-149.

SCHWEIZER, B. (1985). Distribution functions : numbers of the future. *Proceed. of the 2nd Napoli meeting on "The mathematics of fuzzy systems".* Instit. Mat., Univ. Napoli. A. di Nola - A. Ventre (Eds.), 137-149.

SCHWEIZER, B. (2002). Commentary on Probabilistic Geometry. *Karl Mengen, Selecta Matematica.* **Vol. II**, B. Schweizer, K. Sigmund, A. Sklar (Eds.).

SCOTT and SYMONS (1971). Clustering methods based on likelyhood ratio criteria. *Biometrics,* **27**.