C. R. Acad. Sci. Paris, Ser. I 335 (2002) 189–193

Statistique/*Statistics*
(Probabilités/*Probability Theory*)

Classification et mélanges de processus

Richard Emilion [a,b]

[a] **UFR SEGMI, Modalx, Université Paris X, 200, avenue de la République, 92001 Nanterre, France**

[b] **Lise-Ceremade, Université Paris IX–Dauphine, place du Maréchal de Lattre de Tassigny, 75775 Paris cedex 16, France**

Reçu le 2 février 2001 ; accepté après révision le 29 avril 2002

Note présentée par Paul Deheuvels.

Résumé Nous proposons une méthode de classification basée sur l'estimation de mélanges de lois, le point nouveau étant que les unités statistiques sont décrites par des lois de probabilité. Les composantes du mélange sont des processus de Dirichlet, des processus Gamma pondérés normalisés ou des processus de Kraft utilisés en satististique non paramétrique Bayesienne. Les mélanges obtenus par des algorithmes appliqués aux marginales des composantes en dimension finie convergent vers le mélange souhaité lorsque la dimension augmente car les composantes sont orthogonales grâce à un théorème de Kakutani et leur support sont alors les classes recherchées. ***Pour citer cet article : R. Emilion, C. R. Acad. Sci. Paris, Ser. I 335 (2002) 189–193.*** © 2002 Académie des sciences/Éditions scientifiques et médicales Elsevier SAS

Clustering and mixtures of processes

Abstract We propose a clustering method based on the estimation of mixtures of probability distributions, the new point being that the statistical units are described by probability distributions. The components of the mixtures are Dirichlet processes, normalized weighted Gamma processes, and Kraft processes. Mixtures obtained by applying some algorithms to the finite dimensional distributions of the components converge to the desired mixture as the dimension increases, since the components are mutually singular due to a theorem of Kakutani. The desired clusters are then the support of these components. ***To cite this article: R. Emilion, C. R. Acad. Sci. Paris, Ser. I 335 (2002) 189–193.*** © 2002 Académie des sciences/Éditions scientifiques et médicales Elsevier SAS

Abridged English version

There are many situations where one is faced with statistical units described by probability distributions rather than real vectors: descriptions of data categories in huge datasets, inaccurate observations, probabilistic models in physics, etc. We propose, for such units, a clustering method based on the estimation of mixtures of probability distributions. This generalizes well-known results obtained when the descriptions are real vector and answers a question posed by E. Diday in the framework of symbolic data analysis (*see* [3]).

We consider a random variable $X : \Omega \to \mathbf{P}(V)$ where $(\Omega, \mathcal{F}, \mathcal{P})$ is a probability space and $\mathbf{P}(V)$ denotes the set of all probability measures defined on a measurable space $(V, \mathcal{V})$. In the terminology of some authors using Bayesian analyses of nonparametric problems (*see* [6,7,5,10]), the stochastic process

S1631-073X(02)02432-9/FLA

$\{X_A(\cdot) = X(\cdot)(A),\ A \in \mathcal{V}\}$ is a *random distribution*. In most of concrete situations V will be a product space $\prod_{j=1}^{p} V_j$ but for sake of simplicity we assume here that $V = [0, 1)$.

Let $f_i = X^i(\omega)$ be the i-th observation from a sample $X^{(i)},\ i = 1, \ldots, n$, of X.

Real vector case. – Given $x_1, x_2, \ldots, x_n \in R^p$ n observations of a sample of size n from a random vector $X : (\Omega, \mathcal{P}) \to R^p$, the mixture problem of estimating the distribution $\mathcal{P}_X$ of X as a simple mixture (that is a convex combination) $\sum_{k=1,\ldots,K} p_k P_k$ of distributions P_k belonging to a specific parametric family, has been widely studied. An efficient method is given for example by S.A.E.M. algorithm [1], a stochastic approximation version of the popular E.M. algorithm. There is also a clustering approach of this problem given for example by DS algorithm [4].

Discrete distributions. – If $X(\omega)$ is a discrete distribution for a.a. ω, we will assume that $\mathcal{P}_X = \sum_{s=1}^{K} p_s \mathcal{D}_s(\alpha_s; \beta_s)$ is a mixture of normalized weighted Gamma processes (*see* [5,10]).

Let $l \geqslant 2$ be an integer and let $\sigma_l = ([\frac{k-1}{2^l}, \frac{k}{2^l}),\ k = 1, \ldots, 2^l)$ be a dyadic partition of $[0, 1)$. Define the random vector $X(\sigma_l)$ by $X(\sigma_l)(\omega) = (X(\omega)(V_{1l}), \ldots, X(\omega)(V_{2^l l}))$ where $V_{kl} = [\frac{k-1}{2^l}, \frac{k}{2^l})$.

Noticing that if $X \sim \mathcal{D}(\alpha)$ is a Dirichlet process with parameter a probability measure α (*see* [6]), then $X(\sigma_l)$ is a standard finite dimensional Dirichlet $\mathcal{D}(\alpha(V_{1l}), \ldots, \alpha(V_{2^l l})) = \mathcal{D}^{\sigma_l}(\alpha)$, we see that $\mathcal{P}_X = \sum_{s=1}^{K} p_s \mathcal{D}_s(\alpha_s; \beta_s)$ implies that the distribution $\mathcal{P}_X^{\sigma_l}$ of the random vector $X(\sigma_l)$ is $\sum_{s=1}^{K} p_s \mathcal{D}_s^{\sigma_l}(\alpha_s; \beta_s)$.

Then S.A.E.M. algorithm applied to $f_i(V_{1l}), \ldots, f_i(V_{2^l l})$ estimates $\sum_{s=1}^{K} p_s \mathcal{D}_s^{\sigma_l}(\alpha_s; \beta_s)$. As the number of iterations increases, the algorithm yields mixing coefficients $p_s(\sigma_l)$ (the prior probability of cluster s), normalized weighted Gamma distributions $G_s(\sigma_l)$, and numbers $t_{is}(\sigma_l)$ (the posterior probability of cluster s given f_i). The behaviour of these finite-dimensional mixtures is given by the following:

THEOREM. – *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 finite probability measures equivalent to the Lebesgue measure on $[0, 1]$ and the β_s's are strictly positive functions. Let S_s be the support of $\mathcal{D}_s(\alpha_s; \beta_s)$ so that the support of $\mathcal{P}_X$ is $\bigcup_{s=1}^{K} S_s$. Then*

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

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

(iii) *There exists measurable sets $S'_s \subset S_s$ with $\mathcal{P}_X(S_s \setminus S'_s) = 0$ such that if $f_i \in \bigcup_{s=1}^{K} S'_s$, $i = 1, \ldots, n$, 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*

$$\lim_{l\to\infty} t_{is}(\sigma_l) = 1 \quad \text{if } f_i \in S'_s,$$

$$\lim_{l\to\infty} t_{is}(\sigma_l) = 0 \quad \text{if } f_i \notin S'_s,$$

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

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

Continuous distributions. – When the probability measures $X^{(i)}(\omega)$ are absolutely continuous w.r.t. Lebesgue measure with density f_i, we replace the above Gamma processes by processes defined by Kraft [9]. We prove that these ones are mutually singular too and we obtain a similar result to the above theorem for Kraft process mixtures. We also prove that the finite dimensional distributions converge weakly if the f_i's lie in some standard function spaces.

1. Lois aléatoires

Classifier n réalisations d'un vecteur aléatoire X par estimation de sa loi $\mathcal{P}_X$ comme mélange de lois classiques est une méthode bien connue quand il s'agit d'un vecteur réel. Nous nous proposons d'étendre la méthode quand les unités statistiques sont décrites par des lois de probabilités, répondant à une question de

E. Diday dans le cadre de l'analyse de données symboliques (*voir* [2]). Ce type d'unités apparaissent dans plusieurs situations concrètes : description de catégories de données issues de grandes bases de données, observations imprécises, modèles probabilistes en physique, etc.

Soit donc une v.a. $X : \Omega \to \mathbf{P}(V)$, où $(\Omega, \mathcal{F}, \mathcal{P})$ est un espace probabilisé et $\mathbf{P}(V)$ l'ensemble des probabilités définies sur un espace mesurable $(V, \mathcal{V})$. Selon la terminologie de certains auteurs utilisant des analyses bayesiennes de problèmes non paramétriques (*voir* [6,7,5,10]), le processus stochastique $\{X_A(\cdot) = X(\cdot)(A),\ A \in \mathcal{V}\}$ est une loi aléatoire.

Soit $f_i = X^i(\omega)$ la i-ème observation d'un échantillon $X^{(i)},\ i = 1, \ldots, n$, de X.

Dans la plupart des situations concrètes V sera un espace produit $\prod_{j=1}^{p} V_j$ mais pour simplifier nous supposons ici que $V = [0, 1)$.

L'idée de notre méthode est de partitionner V en une partition finie $(V_k)_{k=1,\ldots,\upsilon}$ et d'appliquer des algorithmes connus aux vecteurs réels $f_i(V_k)$. Le problème est de trouver des composantes telles que le mélange obtenu converge lorsque les partitions se raffinent. Nous distinguons deux cas.

2. $X(\omega)$ est une loi discrète

2.1. *Lois Gamma pondérées normalisées.* – Soit $\gamma(a, b)(x) = \frac{1}{\Gamma(a)} b^a \, \mathrm{e}^{-bx} x^{a-1} I_{(x>0)}$ la densité d'une loi Gamma. Soit β une fonction strictement positive définie sur R_+, $\beta\gamma(\alpha, 1)$ étant integrable. Soit $c_\alpha(\beta) > 0$ telle que $\gamma^\beta(\alpha, 1)(x) = c_\alpha(\beta)\beta(x)\gamma(\alpha, 1)(x)$ soit une densité de probabilité. La loi Gamma pondérée normalisée $\mathcal{D}(\alpha_1, \ldots, \alpha_l; \beta)$ est la loi du vecteur de probabilité $(\frac{Z_1}{Z}, \ldots, \frac{Z_l}{Z})$ pour des $Z_i \sim \gamma^\beta(\alpha_i, 1)$ indépendantes avec $Z = Z_1 + \cdots + Z_l$. Pour $\beta = 1$ c'est la loi classique de Dirichlet.

2.2. *Processus de Dirichlet et processus Gamma.* – Soit α une mesure sur $[0, 1]$. Une loi aléatoire $X : \Omega \to \mathbf{P}(V)$ est dit processus de Dirichlet $\mathcal{D}(\alpha)$ si pour toute partition finie $(B_1, \ldots, B_k)$ de $[0, 1]$, le vecteur aléatoire $(X(B_1), \ldots, X(B_k))$ suit une loi de Dirichlet $\mathcal{D}(\alpha(B_1), \ldots, \alpha(B_k))$. Les processus Gamma $\mathcal{D}(\alpha; \beta)$ se définissent de manière identique. Ces processus construits par Ferguson (*voir* [6,7]) (resp. Dykstra et Laud [5], et Lo [10]) sont tels que $X(\omega)$ est une probabilité discrète pour p.t. ω.

2.3. *Mélanges de processus.* – Soit $X : \Omega \to \mathbf{P}(V)$ mélange de processus Gamma au sens suivant : $X = \sum_{s=1,\ldots,K} 1_{(U=s)} Z_s$ où les Z_s sont des processus Gamma sur $\mathbf{P}(V)$ et U une v.a.r. indépendante des Z_s et prenant la valeur s avec probabilité p_s.

On considère des partitions dyadiques $\sigma_l = (V_{kl} = [\frac{k-1}{2^l}, \frac{k}{2^l}),\ k = 1, \ldots, 2^l)$ de plus en plus fine. L'algorithme S.A.E.M. appliqué aux vecteurs réels $(f_i(V_{1l}), \ldots, f_i(V_{2^l l}))$, nous estime les paramètres des lois Gamma fini-dimensionnelles dépendant de σ_l ainsi que $1_{(U=s)} = p_s(\sigma_l)$ (resp. $t_{is}(\sigma_l)$) la probabilité à priori (resp. à postériori) de la classe recherchée C_s (resp. sachant f_i).

Notre principal résultat montre que ces estimations approchent le mélange qui donne la loi de X et que les classes recherchées sont les supports disjoints des composantes de ce mélange.

THÉORÈME 1. – *Supposons que $\mathcal{P}_X = \sum_{s=1}^{K} p_s \mathcal{D}_s(\alpha_s; \beta_s)$ soit un mélange simple de processus Gamma normalisés avec des probabilités α_s distinctes absolument continues sur $[0, 1]$. Soit S_s le support de $\mathcal{D}_s(\alpha_s; \beta_s)$ (le support de $\mathcal{P}_X$ est donc $\bigcup_{s=1}^{K} S_s$). Alors*

(i) *Les marginales finies dimensionnelles des lois $\mathcal{D}_s(\alpha_s, \beta_s)$ sont équivalentes.*

(ii) *Les lois $\mathcal{D}_s(\alpha_s; \beta_s)$ sont mutuellement singulières et les S_s sont disjoints.*

(iii) *Il esite des ensembles mesurables $S'_s \subset S_s$ tels que $\mathcal{P}_X(S_s \backslash S'_s) = 0$ et si $f_i \in \bigcup_{s=1}^{K} S'_s$, $i = 1, \ldots, n$, alors les nombres $t_{is}(\sigma_l)$ et $p_s(\sigma_l)$ vérifient*

$$\lim_{l\to\infty} t_{is}(\sigma_l) = 1 \quad \textit{si } f_i \in S'_s,$$

$$\lim_{l\to\infty} t_{is}(\sigma_l) = 0 \quad \textit{si } f_i \notin S'_s,$$

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

Un résultat analogue se démontre pour l'algorithme DS.

3. $X(\omega)$ est une loi absolument continue

3.1. *Processus de Kraft.* – Soit $Z = \{Z_{k/2^r};\ r = 1, 2, \ldots,\ k = 1, 3, \ldots, 2^r - 1\}$ un ensemble de v.a.r. indépendantes telles que $0 \leqslant Z_{k/2^r} \leqslant 1$et $E(Z_{k/2^r}) = \frac{1}{2}$. Soit F_l une suite de fonctions de répartition sur $[0, 1]$ définies par récurrence

$$F_1(0) = 0, \qquad F_1\left(\frac{1}{2}\right) = Z_{1/2}, \qquad F_1(1) = 1,$$

F_1 est affine sur $\left[0, \frac{1}{2}\right)$ et $\left[\frac{1}{2}, 1\right]$,

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

F_l est affine sur les intervalles dyadiques.

Soit g_l la dérivée de F_l. Sous certaines conditions

$$F_l(x) \to F(x) \quad \text{et} \quad g_l(x) \to g(x) = F'(x) \quad \text{si } l \to +\infty. \tag{1}$$

On note alors $g = \text{Kraft}(Z)$.

On supposera que la loi $d_{k,l}$ de $Z_{k/2^l}$ a une densité de type exponentielle à support dans $[k/2^l, (k+1)/2^l)$.

3.2. *Mélanges de processus de Kraft.* – On dira que X est un mélange de processus de Kraft s'il exite des suites $Z^{(s)}$ comme Z ci-dessus ainsi qu'une v.a. U, indépendante des $Z^{(s)}$, prenant K valeurs distinctes $u_1, \ldots, u_K$, telles que

$$X = \sum_{s=1}^{K} 1_{(U=u_s)}\text{Kraft}\big(Z^{(s)}\big),$$
$$\mathcal{P}(U = u_s) = p_s.$$

On obtient alors un résultat analogue au Théorème 1 :

THÉORÈME 2. –
(i) *Les marginales finies-dimensionnelles des processus de Kraft sont équivalentes.*
(ii) $\text{Kraft}(Z^{(s)})$ *et* $\text{Kraft}(Z^{(t)})$ *sont mutuellement singulières si* $s \neq t$.
(iii) $\lim_{l\to\infty} t_{is}(\sigma_l) = t_{is}$ $(= 1$ *ou* $0)$ *et* $\lim_{n\to\infty} \lim_{\ell\to\infty} p_s(\sigma_l) = p_s$.

3.3. *Convergence faible.* – Si les observations f_i sont dans des espaces fonctionnels tels que $\mathcal{C}[0, 1]$, $L_q[0, 1]$ $(1 \leqslant q < \infty)$, ou l'espace $D[0, 1]$ de Skorohod, on montre une convergence de type faible de la loi marginale $P_X^{\sigma_l}$ vers P_X.

4. Principaux points des démonstrations

Théorème 1. – Il est bien connu que $\mathrm{d}\mathcal{D}_l(\alpha'; \beta')/\mathrm{d}\mathcal{D}_l(\alpha; \beta)$ est une $\mathcal{D}(\alpha; \beta)$-martingale. Si X est un processus de Dirichlet $\mathcal{D}(\alpha)$ alors il existe une suite i.i.d. $V_{n,\alpha}$ de loi α telle que le support de la probabilité discrète $X(\omega)$ soit une partie de l'ensemble aléatoire $\{V_{1,\alpha}(\omega), V_{2,\alpha}(\omega), \ldots, V_{n,\alpha}(\omega), \ldots\}$ [6].

Mais si α et α' sont distincts et équivalents à la loi uniforme λ sur $[0, 1]$, alors $\int \sqrt{\frac{\mathrm{d}\alpha}{\mathrm{d}\lambda}\frac{\mathrm{d}\alpha'}{\mathrm{d}\lambda}} < 1$ par l'inégalité de Cauchy–Schwarz et $\big(\int \sqrt{\frac{\mathrm{d}\alpha}{\mathrm{d}\lambda}\frac{\mathrm{d}\alpha'}{\mathrm{d}\lambda}}\big)^n \to 0$ si $n \to +\infty$. Un théorème de Kakutani (*voir* [8],

p. 453) montre alors que les mesures produits $\bigoplus_{n=1}^{\infty}\alpha$ et $\bigoplus_{n=1}^{\infty}\alpha'$ sont mutuellement étrangères. On en déduit que $\mathcal{D}(\alpha)$ et $\mathcal{D}(\alpha')$ sont mutuellement étrangers ainsi que $\mathcal{D}(\alpha;\beta)$ et $\mathcal{D}(\alpha';\beta')$.

À l'itération q dans l'algorithme S.A.E.M., on observe alors que

$$\frac{1}{t_{ik}^{q+1}(\sigma_l)} = 1 + \sum_{r\neq k} \frac{p_{qr}G_{0rl}(f_i(\sigma_l))}{p_{qk}G_{0kl}(f_i(\sigma_l))} \to 1$$

car d'une part $p_{qr}/p_{qk} \leqslant c(n)$ pour une constante $c(n)$ ne dépendant que de la taille de l'échantillon et telle que $\lim_{n\to\infty} c(n) = 0$, et d'autre part le théorème des martingales implique que $\lim_{l\to+\infty} G_{0rl}(f_i(\sigma_l))/G_{0kl}(f_i(\sigma_l)) = 0$. On obtient alors

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

et comme $X^{(1)},\dots,X^{(n)}$ est i.i.d. $\sim X$, $\lim_{n\to\infty}\lim_{l\to\infty} p_k(\sigma_l) = \mathcal{P}_X(S_k) = p_k$.

Références bibliographiques

[1] G. Celeux, J. Diebolt, S.A.E.M. algorithm, Stochastics Stochastics Rep. 41 (1992) 119–134.
[2] A. Darwich, About the absolute continuity and orthogonality of two probability measures, Statist. Probab. Lett. 52 (1) (2001) 1–8
[3] E. Diday, in: Bock, Diday (Eds.), Analysis of Symbolic Data, Springer-Verlag, 2000.
[4] E. Diday, A. Schroeder, A new approach in mixing distributions detection, RAIRO Operational Research 10 (6) (1976).
[5] R.L. Dykstra, P. Laud, A Bayesian nonparametric approach to reliability, Ann. Statist. 9 (1981) 356–367.
[6] T.S. Ferguson, A Bayesian analysis of some nonparametric problems, Ann. Statist. 1 (1981) 209–230.
[7] T.S. Ferguson, Prior distributions on spaces of probability measures, Ann. Statist. 2 (1974) 615–629.
[8] E. Hewitt, K. Stromberg, Real and Abstract Analysis, Springer-Verlag, 1969.
[9] C.H. Kraft, A class of distribution function processes which have derivatives, J. Appl. Probab. 1 (1964) 385–388.
[10] A.Y. Lo, Bayesian nonparametric statistical inference for Poisson point process, Z. Wahrsch. Verw. Gebiete 59 (1982) 55–66.