Reconnaissance de loi via une estimation de densités par histogramme pour la modélisation de canaux de transmission.

Guilhem COQ[1,2], Olivier ALATA[2], Christian OLIVIER[2], Xiang LI[2], Yannis POUSSET[2]

[1]Laboratoire de Mathématiques et Applications, UMR CNRS 6086
Téléport 2, BP 30179, 86962 Futuroscope Chasseneuil cedex

[2]Laboratoire XLIM-SIC (Signal, Image et Communications), UMR CNRS 6172
Téléport 2, BP 30179, 86962 Futuroscope Chasseneuil cedex

coq@math.univ-poitiers.fr, alata,olivier,li,pousset@sic.univ-poitiers.fr

Résumé – Nous nous intéressons à la reconnaissance de la loi d'un échantillon parmi une famille de lois données. Les critères d'information sont utilisés pour déterminer, à partir de l'échantillon, un histogramme estimant la loi inconnue, ici, celle modélisant un canal de propagation radioélectrique. Cet histogramme permet de calculer des distances de type Kullback-Leibler entre densités de probabilités. Ces distances sont ensuite utilisées pour la reconnaissance de la loi. Cette méthode diffère de l'approche de type Kolmogorov-Smirnov, utilisée habituellement dans ce contexte applicatif, qui utilise des distances sur les fonctions de répartition empiriques. Les perfomances de l'outil de reconnaissance de loi par histogramme sont comparées à celles de l'outil classique de Kolmogorov-Smirnov par une étude statistique appliquée aux radiocommunications, cette étude portant sur les sensibilité, spécificité, précision, valeurs prédictives positive et négative des stratégies de reconnaissance.

Abstract – We are interested in the recognition of the law of a sample among a family of given laws. Information Criteria are used in order to determine, from the sample, an histogram estimating the unknown law, here the one modelizing a propagation channel. This histogram allows to compute Kullback-Leibler distances between probability densities. Those distances are then used for the recognition of the law. This method differs from the Kolmogorov-Smirnov approach, classically used in this applicative context, that uses distances on empirical cumulative functions. The efficiency ot our tool of law recognition via histogram are compared to the ones of the classical Kolmogorov-Smirnov method by a statistical study applied to radiocommunications. This study involves sensibility, specificity, precision, positive and negative predictive values of the tools of recognition.

1 Introduction

Le problème de reconnaissance de loi est d'un grand intérêt dans beaucoup de domaines tels que le traitement de l'image, la reconnaissance de forme ou la télécommunication qui nous intéresse ici. Dans ce cadre, l'outil le plus largement utilisé pour résoudre ce problème est le test d'adéquation de Kolmogorov-Smirnov. Il est basé sur une estimation de la fonction de répartition empirique de la loi inconnue. Dans cet article, nous choisissons d'estimer la densité elle-même par un histogramme. Cette estimation nous permet de calculer une distance de type Kullback-Leibler sur laquelle est basée la reconnaissance.

Les critères d'informations, ou critères d'entropie pénalisés, sont utilisés en vue d'estimer la densité inconnue par un histogramme. Birgé [2] et Birgé et al. [3] étudient ce problème dans des travaux récents. Ils justifient l'utilisation d'un critère d'information par la minimisation du risque résultant de l'estimation. D'un autre point de vue, Rissanen développe dans [8] la notion de "Minimum Description Length" (MDL) et dans [9] celle de complexité stochastique, fortement reliée à la théorie du codage. A partir de ces notions, nous justifions dans [6] et [5] l'utilisation d'un critère d'information par des arguments de type codage.

Dans cet article, nous utilisons ce critère pour estimer la densité par un histogramme et ainsi résoudre le problème de reconnaissance de loi avec une distance de type Kullback-Leibler. Notre but est de montrer, via des simulations et une étude statistique, que cette stratégie de reconnaissance de loi est significativement plus efficace que celle de Kolmogorov-Smirnov.

La section 2 présente le critère d'information (eq. 2) ainsi que son utilisation pour l'estimation d'une densité inconnue par un histogramme. Dans la section 3, nous présentons les différentes stratégies de reconnaissance de loi que nous allons comparer. Enfin, la section 4 est consacrée aux simulations et aux comparaisons des performances de ces stratégies de reconnaissance.

2 Estimation de densités

Soit $x = x^n = x_1, \ldots, x_n$ un n-échantillon d'une densité de probabilité inconnue f dont le support est un intervalle $I \subset \mathbb{R}$. Soit également $P = (I_j)_{j \in [\![1,m]\!]}$ une partition de I en m intervalles. On sait que l'estimée au sens du maximum de vraisemblance de f par une fonction constante par morceaux sur P est donnée par l'histogramme :

$$h_P = \sum_{j=1}^{m} \frac{n_j}{n\ell_j} \mathbb{1}_{I_j} \tag{1}$$

où ℓ_j est la longueur de l'intervalle I_j et n_j le nombre d'échantillons contenus dans I_j.

Le choix de la partition P sur laquelle est construit cet histogramme est crucial lors d'une telle estimation. Il s'agit d'un problème de sélection de modèle non-paramétrique que nous

nous proposons de résoudre à l'aide d'un critère d'information. Nous appuyant sur les travaux de Rissanen et al. [9, 10, 1], nous avons proposé dans [6, 4] un critère reposant sur la théorie de l'information. Chaque partition P de I est utilisée pour encoder les données x. La longueur d'un tel codage est estimée par le critère

$$\text{CRIT}(x,P) = -\sum_{j=1}^{m} n_j \log \frac{n_j}{n\ell_j} + (m-1)\frac{\log n}{2}. \quad (2)$$

Le principe du MDL [1] nous conduit alors à choisir la partition $\widehat{P}$ qui minimise ce critère et à estimer f par $h_{\widehat{P}}$ comme en (1).

L'intervalle I est initiallement découpé en R intervalles de longueurs égales r. Nous nous limitons aux partitions de I dont les bornes des intervalles coïncident avec ce découpage. Un algorithme de programmation dynamique détaillé en annexe A de [4] permet de minimiser (2) sur les 2^{R-1} partitions en question en un nombre d'opérations de l'ordre de seulement R^2. La partition obtenue, dénommée dynamique, présente des intervalles de longueurs variables.

3 Reconnaissance de loi

3.1 Position du problème

Donnons nous une famille de densités de probabilités paramétrées. Dans ce travail nous considérons la famille constituée des lois de Rayleigh, Nakagami et Weibull :

$$\begin{array}{ll} \text{Rayleigh} & f_{R,\sigma}(t) = \frac{t}{\sigma^2}\exp\left(-\frac{t^2}{2\sigma^2}\right)\mathbb{1}_{t\geq 0} \\ \text{Nakagami} & f_{N,\mu,\Omega}(t) = \frac{2\mu^\mu t^{2\mu-1}}{(\mu-1)!\Omega^\mu}\exp\left(-\frac{\mu t^2}{\Omega}\right)\mathbb{1}_{t\geq 0} \\ \text{Weibull} & f_{W,k,\lambda}(t) = \frac{kt^{k-1}}{\lambda^k}\exp\left(-\frac{t^k}{\lambda^k}\right)\mathbb{1}_{t\geq 0} \end{array} \quad (3)$$

où $\sigma, \mu, \Omega, k, \lambda$ sont les paramètres de forme de ces lois. Notons que, pour $\mu = 1$ ou $k = 2$, les lois de Nakagami ou Weibull correspondent à celle de Rayleigh.

Ces lois sont classiquement utilisées pour la modélisation des canaux de propagation radio (voir [11, 12]). Nous nous intéresserons aux cas où les émetteurs et récepteurs sont en position de visibilité (Line Of Sight, LOS) ou de non-visibilité (Non Line Of Sight, NLOS). Le cas NLOS est le plus délicat. En effet, une simulation de la propagation des ondes radio en configuration NLOS par un logiciel [7], mis au point par le laboratoire XLIM-SIC, montre que les densités candidates (3) sont plus proches visuellement que dans la configuration LOS.

Nous nous proposons, à partir d'un échantillon x^n d'une loi inconnue de ce type, de déterminer si cette loi est du type Rayleigh, Weibull, ou Nakagami.

3.2 Les différentes méthodes étudiées

Les paramètres de forme $\sigma, \mu, \Omega, k, \lambda$ dans (3) sont fixés *a priori*. Les lois $f_{R,\sigma}, f_{N,\mu,\Omega}$ et $f_{W,k,\lambda}$ deviennent les lois en compétition. On peut considérer qu'elles ont été déterminées au préalable par apprentissage. La reconnaissance s'effectue ensuite par l'une des deux méthodes suivantes :

La méthode de Kolmogorov-Smirnov (KS) : c'est celle classiquement utilisée. La distance de Kolmogorov-Smirnov entre deux fonctions de répartition F et G est définie par

$$\text{KS}(F,G) = \sup_{t\in I} |F(t) - G(t)|.$$

Notons $\widehat{F}$ la fonction de répartition empirique de l'échantillon. La loi est choisie par minimisation de la distance KS entre $\widehat{F}$ et les fonctions de répartition des lois en compétition.

La méthode à histogrammes (KL) : elle utilise la divergence de Kullback-Leibler (KL) symétrisée qui est définie entre deux densités de probabilités f et g par

$$\text{KL}(f,g) = \frac{1}{2}\int_I (f-g)\log(f/g)dt.$$

avec dt la mesure de Lebesgue.
Nous notons $h_{\widehat{P}}$ l'histogramme dynamique choisi par minimisation du critère (2). La loi est choisie par minimisation de la distance KL entre $h_{\widehat{P}}$ et les densités des lois en compétition.

4 Simulations

Nous travaillons dans un cadre théorique visant à valider les méthodes à histogramme KL et à les comparer avec les méthodes usuelles KS.

4.1 Génération des échantillons

Nous envisageons deux cas. Le cas LOS, où les lois génératrices sont relativement éloignées visuellement, et le cas NLOS où ces lois sont plus proches. Pour chacun de ces cas et pour

TAB. 1 – Paramètres des différentes lois appris à partir d'échantillons réels

Loi	Cas LOS	Cas NLOS
Rayleigh σ	1.5788	1.7108
Nakagami $(\mu;\Omega)$	0.6755 ; 4.985	0.9565 ; 4.8536
Weibull $(k;\lambda)$	1.5291 ; 2.0505	1.8447 ; 2.3674

n variant de 100 à 1000, nous générons 450 n-échantillons de chaque loi. Nous appliquons à chacun de ces échantillons les deux stratégies de reconnaissance étudiées.

4.2 Matrices de confusion

A chaque loi de (3), et à chacune des stratégies de reconnaissance étudiées, nous associons la matrice de confusion donné dans la table 2. Cette matrice est présentée ici, pour illustration, en particularisant la loi de Weibull.

TAB. 2 – Matrice de confusion pour la loi de Weibull

loi génératrice \ loi reconnue	Weibull	Autre
Weibull	a	$c = 450 - a$
Autre	b	$d = 900 - b$

4.3 Statistiques étudiées

Pour chacune des matrices de confusion, nous étudions les statistiques suivantes :

- Sensibilité : la probabilité d'être reconnu Weibull lorsque l'on est généré par Weibull.

$$\mathrm{S} = a/450$$

- Spécificité : la probabilité de ne pas être reconnu Weibull lorsque l'on n'est pas généré par Weibull.

$$\mathrm{SP} = d/(b+d)$$

- Précision : la probabilité d'être bien classé par le test au regard de la loi de Weibull.

$$\mathrm{P} = (a+d)/1350$$

- Valeur Prédictive Positive : la probabilité d'avoir été généré par Weibull lorsqu'on est reconnu Weibull.

$$\mathrm{VPP} = a/(a+b)$$

- Valeur Prédictive Négative : la probabilité d'avoir été généré par une autre loi lorsque l'on n'est pas reconnu Weibull.

$$\mathrm{VPN} = d/(c+d)$$

Plus ces statistiques sont proches de 1, plus la stratégie étudiée est efficace pour la reconnaissance de la loi de Weibull.

La figure 1 présente l'évolution de la sensibilité pour les méthodes KL et KS et les différentes lois en fonction de la taille n des échantillons. Nous nous sommes placés ici dans le cas LOS.

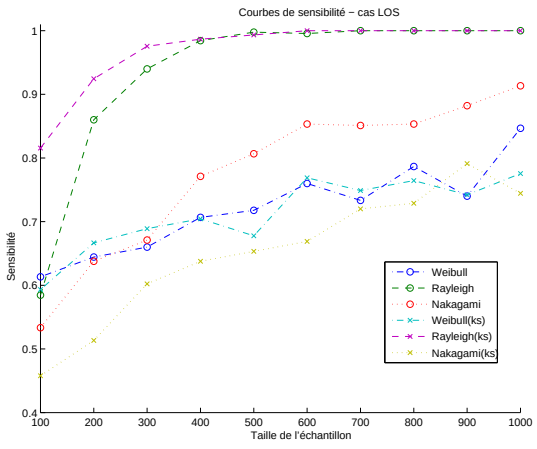


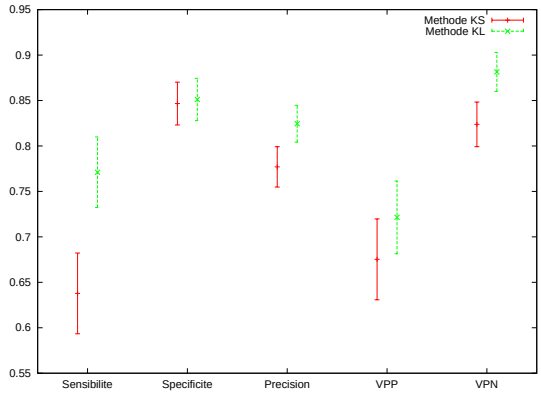
FIG. 1 – Evolution, en fonction de la taille des échantillons, de la sensibilité des stratégies de reconnaissance dans le cas LOS.

Pour $n \geq 400$, nous remarquons que la méthode KL présente une sensibilité au moins aussi bonne que la méthode KS, voire une sensibilité meilleure dans le cas de la loi de Nakagami. Pour des tailles d'échantillons plus faibles ($n \leq 300$), la méthode KS reste privilégiée pour la reconnaissance de la loi de Rayleigh. Cependant les différences observées ici ne portent que sur l'estimation brute des statistiques de reconnaissance, elle ne permettent donc pas de décider si les méthodes KL et KS ont des comportements significativement différents. Dans la prochaine section nous apportons des éléments de réponse à ce problème.

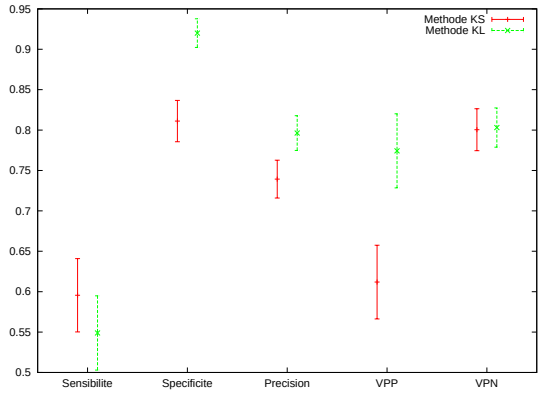
4.4 Intervalles de confiance

Nous nous intéressons ici aux intervalles de confiance à 95% des statistiques de reconnaissance. Calculés de manière classique à partir des matrices de confusion, ces intervalles contiennent la valeur inconnue des statistiques avec un risque d'erreur de 5%.

Pour illustration, nous présentons en figure 2 les intervalles de confiance des statistiques étudiés pour les méthodes KL et KS dans deux des cas de notre étude.



Reconnaissance de la loi de Nakagami, cas LOS, n=400



Reconnaissance de la loi de Rayleigh, cas NLOS, n=1000

FIG. 2 – Comparaison des intervalles de confiance des statistiques de reconnaissance obtenues par les méthodes KL et KS.

On observe, par exemple, que l'intervalle de confiance de la sensibilité de la reconnaissance de la loi de Nakagami à $n = 400$ par la méthode KL est disjoint de celui obtenu par la méthode KS. Cela indique que la méthode KL présente un comportement significativement meilleur, en terme de sensibilité, pour la reconnaissance de cette loi.

De manière plus exhaustive, nous présentons, dans la table 3, des tableaux de comparaison. Pour chacune des lois (Rayleigh, Nakagami et Weibull), chacun des cas (LOS et NLOS) et chacune des statistiques (sensibilité S, spécificité SP, précision P, valeur prédictive positive VPP et valeur prédictive négative VPN) de notre étude, nous indiquons quelle méthode de reconnaissance est significativement la plus efficace en terme de disjonction des intervalles de confiance. Un $+$ indique que la méthode à histogramme KL est meilleure que la méthode KS, un $-$ indique le contraire. Les cases sont laissées vides lorsque les intervalles de confiance à 95% ont une intersection.

Notons que la première colonne de ces tableaux correspond

en fait à la figure (1). On observe que, globalement, les performances de la méthode KL sont significativement meilleures que celles de la méthode KS, particulièrement pour les échantillons de tailles moyennes ou grandes ($n \geq 400$).

	Loi de Rayleigh									
	Cas LOS					Cas NLOS				
n	S	SP	P	VPP	VPN	S	SP	P	VPP	VPN
100	−		−	−	−	−				
200	−				−	−	+			
300			−			−	+			
400						−	+			
500						−	+			
600							+			
700							+		+	
800						−	+		+	
900							+	+	+	
1000							+	+	+	

	Loi de Nakagami									
	Cas LOS					Cas NLOS				
n	S	SP	P	VPP	VPN	S	SP	P	VPP	VPN
100		−								
200	+									
300										
400	+		+		+	+				
500	+		+		+					
600	+		+		+					
700	+				+	+				
800	+		+		+	+				
900	+				+	+		+		+
1000	+		+	+	+	+				

	Loi de Weibull									
	Cas LOS					Cas NLOS				
n	S	SP	P	VPP	VPN	S	SP	P	VPP	VPN
100							−	−		
200						+	−			
300							−			
400		+		+		+	−			
500		+	+	+		+				+
600		+	+	+		+				+
700		+		+		+				+
800		+	+	+		+				+
900		+		+		+		+		+
1000		+	+	+	+	+				+

TAB. 3 – Tableaux résumant la comparaison entre les méthodes KL et KS suivant les statistiques considérées : un $+$ indique que la méthode à histogramme KL est meilleure que la méthode KS, un $-$ indique le contraire

5 Conclusion

Nous avons proposé une stratégie de reconnaissance de loi basée sur l'estimation directe de la densité par un histogramme déterminé par le critère d'information (2). Cette stratégie est comparée à celle, classiquement utilisée dans la modélisation de canaux de transmission, de Kolmogorov-Smirnov. A travers des simulations et un test statistique au risque 5%, nous avons montré que notre stratégie est, dans la majorité des cas, la plus efficace. L'histogramme que nous utilisons peut être vu comme un résumé de l'information apportée par l'échantillon en un nombre restreint de classes correspondant aux intervalles de l'histogramme. Par opposition, la fonction de répartition empirique utilisé dans la procédure de Kolmogorov-Smirnov conserve toute l'information de l'échantillon. En ce sens, on peut dire que l'information résumée par l'histogramme permet une meilleure reconnaissance de ses caractéristiques que l'information totale.

Références

[1] A. Barron, J. Rissanen, and B. Yu. The minimum description length principle in coding and modeling. *IEEE Trans. Inform. Theory*, 44(6) :2743–2760, 1998.

[2] L. Birgé. Statistical estimation with model selection. *Indag. Math. (N.S.)*, 17(4) :497–537, 2006.

[3] L. Birgé and Y. Rozenholc. How many bins should be put in a regular histogram. *ESAIM Probab. Stat.*, 10 :24–45 (electronic), 2006.

[4] G. Coq. *Utilisation d'approches probabilistes basées sur les critères entropiques pour la recherche d'information sur support multimédia*. Thèse de doctorat, Université de Poitiers, France, Décembre 2008.

[5] G. Coq, X. Li, O. Alata, Y. Pousset, and C. Olivier. Law recognition via histogram-based estimation. In *ICASSP-IEEE, Taipei (Taiwan)*, pages 3425–3428, April 2009.

[6] G. Coq, C. Olivier, O. Alata, and M. Arnaudon. Information criteria and arithmetic codings : an illustration on raw image. In *15th European Signal Processing Conference proceedings*, pages 634–638, Poznan, Poland, 2007.

[7] F. Escarieu, Y. Pousset, R. Vauzelle, and L. Aveneau. Outdoor and indoor channel characterization by a 3d simulation software. Septembre 2001. PIMRC '2001, San Diego, USA.

[8] J. Rissanen. Modeling by the shortest data description. *Automatica*, 14 :465–471, 1978.

[9] J. Rissanen. Stochastic complexity and modeling. *Ann. Statist.*, 14(3) :1080–1100, 1986.

[10] J. Rissanen, T. P. Speed, and B. Yu. Density estimation by stochastic complexity. *IEEE Transactions on Information Theory*, 38(2) :315–323, 1992.

[11] N. C. Sagias and G. K. Karagiannidis. Gaussian class multivariate Weibull distributions : Theory and applications in fading channels. *IEEE Transactions on Information Theory*, 51(10) :3608–3619, Oct 2005.

[12] T.K. Sarkar, Zhong Ji, Kyungjung Kim, A. Medouri, and M. Salazar-Palma. A survey of various propagation models for mobile communication. *Antennas and Propagation Magazine, IEEE*, 45(3) :51–82, 2003.