Estimation imprécise de densité de probabilité par transfert imprécis de comptage.

Olivier Strauss

LIRMM, Université Montpellier II, 161, rue Ada, 34392 Montpellier Cedex 5, France
e-mail : Olivier.Strauss@lirmm.fr

Résumé — Un histogramme *quasi-continus* (HQC) est un outil de représentation et d'analyse statistique apparenté aux méthodes du noyau. Il permet de réaliser des estimations statistiques de moments, de modes ou de fractiles. Cette aptitude est due à la possibilité de reconstruire une estimation de la densité de probabilité en n'importe quel point de la droite réelle, comme l'aurait permise une méthode du noyau. Cependant, la discrétisation de la droite, propre aux histogrammes, créé une altération de la répartition locale des données se traduisant par une imprécision de la densité estimée. Nous proposons, dans cet article, une méthode d'estimation de cette densité imprécise basée sur une intégrale de Choquet.

Abstract — A *quasi-continuous* histogram (QCH) is a statistical tool related to the kernel density estimation. It allows statistical estimates of moments, modes or quartils. This property is ensued from the possibility to estimate the density of probability in any point like would have allowed a kernel estimate. However, the discretization creates a deterioration of the local distribution of the data. It results an inaccuracy of the estimated density. We propose, in this article, a method to estimate the imprecision of this density by using a Choquet integral.

1. Introduction.

Un histogramme quasi-continu (HQC) est un histogramme construit sur une partition floue de la droite réelle. C'est un outil de représentation et d'analyse statistique dont les fondements reposent sur la théorie des sous-ensembles flous grossiers [1]. Les HQC permettent de réaliser des opérations statistiques de base telles que l'estimation de modes, de moments et de fractiles [2,3]. Les estimations produites par les HQC sont à la fois rapides et robustes.

La possibilité de réaliser des opérations d'estimation avec des HQC provient du lien étroit qui existe entre histogramme et densité de probabilité, ou plus exactement entre comptage et densité de probabilité. Ce lien étroit est une des bases des travaux de B. Silverman sur l'estimation de densité par la méthode du noyau [4].

Le principe des HQC est très proche de celui de la méthode du noyau en ce sens que des noyaux sont utilisés pour estimer une grandeur statistique locale. Ces deux méthodes diffèrent cependant en quatre points essentiels :

- les noyaux utilisés pour construire un histogramme quasi-continu sont des noyaux flous donc non-sommatifs,
- la représentation des données par HQC ne se limite pas à la reconstruction de la densité de probabilité sous-jacente mais permet aussi des manipulations statistiques des données [5],
- les noyaux sont répartis, comme pour un histogramme classique, de façon arbitraire sur la droite mais représentation par HQC répercute de façon explicite ce caractère arbitraire,
- la représentation par HQC permet de prendre en compte une connaissance a priori sur l'imprécision ou l'incertitude relative des données traitées.

Parce qu'il est réalisé sur une partition floue de la droite réelle, un HQC peut être vu comme un emboîtement de plusieurs histogrammes binaires ayant le même nombre de cellules mais des résolutions différentes. De cette propriété découle qu'il est possible de réaliser une estimation de la densité des données qui serait accumulé sur un noyau différent de ceux utilisés pour réaliser la partition. Ce que l'on appelle le transfert de comptage est un des outils de base de l'estimation statistique utilisant des HQC.

La technique de transfert de comptage que nous avons utilisé jusqu'à présent est inspirée de la méthode de transfert de croyance pignistique proposé par Smets [6]. Cette technique présente de nombreux avantages car elle peut être représentée localement par une fonctionnelle. C'est sur cette propriété que nous nous appuyons pour réaliser la détection des modes d'une distribution empirique [3]. Cependant, cette méthode de transfert ne permet pas de répercuter l'altération produite par le partitionnement de la droite réelle sur l'estimation de la densité. Nous en proposons une nouvelle généralisation basée sur une intégrale de Choquet transformant l'altération dû à l'échantillonnage en imprécision sur la densité estimée. Nous lui avons donné le nom de transfert pignistique imprécis. Dans le cadre de cet article, les données sont supposées précises, ou d'imprécision inconnue.

2. Accumulation de données précises d ans un histogramme quasi-continu.

Soient n observations réelles précises $x_1 \ldots x_n$. Construire un histogramme a partir de ces observations consiste à partitionner un intervalle de référence réel $\Omega=[x_{min}, x_{max}]$ en p cellules C_k et à compter le nombre a_k d'observations appartenant à chaque cellule. a_k, est l'accumulateur associé à C_k. Lorsque toutes les cellules de l'histogramme sont de même largeur, l'histogramme est dit régulier. La largeur d'une cellule C_k est Δ, appelée pas ou granularité de l'histogramme.

En substituant, à la partition binaire de la droite réelle, une partition floue, on obtient un histogramme quasi-continu (HQC). Nous nous plaçons, dans cet article, dans le cas d'une partition floue forte régulière de fonction d'appartenance triangulaire, c'est à dire un ensemble de nombres flous (ou intervalles ou cellules), $(C_k)_{k=1\ldots p}$ tel que :

$$\forall x \in \mathbb{R}, \sum_k \mu_{C_k}(x) = 1 \qquad (1)$$

$\Delta = \int \mu_{C_k}(x)dx$ est la granularité de l'histogramme, m_k est le mode de l'intervalle C_k et $\mu_{C_k}(x)$ est l'appartenance de la valeur réelle x à la cellule C_k. La première et la dernière cellule de la partition sont de granularité infinie pour respecter la propriété (1) (Fig. 1).

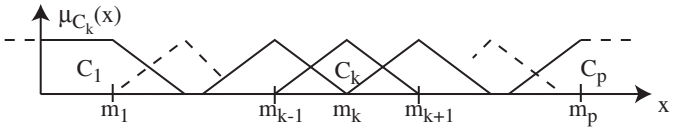

Figure 1 : Partition floue forte de $\mathbb{R}^2$

L'accumulation des données dans un histogramme quasi-continu n'est autre qu'une généralisation de l'accumulation des données dans un histogramme classique. Soit a_k l'accumulateur associé à la cellule C_k :

$$a_k = \sum_{i=1}^{n} \mu_{C_k}(x_i) \tag{2}$$

3. Estimation de densité de probabilité.

3.1 Comptage, probabilité et densité.

La probabilité d'un événement peut être approchée par le rapport du nombre de cas favorables à cet événement au nombre total d'événements considéré. Dans le cas binaire précis, soient (x_i) les observations et une quantité $W \subset \mathbb{R}$ de granularité Γ, $P(W;(x_i))$, la probabilité de W basée sur les observations (x_i) est donnée par :

$$P(W;(x_i)) = \frac{nb(x_i \in W)}{nb(x_i)} = \frac{nb(W;(x_i))}{nb(\Omega;(x_i))} \tag{3}$$

Si W est un sous-ensemble flou, alors on utilise la formule (2) pour estimer $nb(W;(x_i))$.

Le principe de la méthode du noyau consiste à déplacer un noyau de granularité fixé sur la droite réelle et à réaliser une estimation locale de la densité. Si on veut réaliser une estimation de la densité en chaque point de la droite réelle à partir d'un histogramme quasi-continu, il faut être capable de transférer le comptage réalisé sur la partition (C_k) sur tout sous-ensemble W de la droite réelle.

3.2 Principe du transfert de comptage.

La méthode pignistique précise consiste à transférer l'accumulateur de chaque cellule de la partition vers le sous-ensemble W au prorata du recouvrement de W et C_k. Le transfert pignistique s'écrit :

$$nb(W;(x_i)) = \sum_{k=1}^{p} a_k \left(\frac{|W \cap C_k|}{|C_k|} \right) \tag{4}$$

Ce type de transfert considère que l'échantillonnage réalise un mélange uniforme des votes à chaque niveau de confiance. Si a_k est le comptage associé à C_k, alors la densité locale supposée dans la cellule C_k est a_k/Δ. L'utilisation de cette méthode repose sur une hypothèse d'indépendance des cellules voisines. Or, si on considère la partition floue forte comme un emboîtement d'intervalles binaires associés à des mesures de confiance croissantes [7], cette hypothèse n'est vraie que si le niveau de confiance requis est inférieur à 0,5 (coupes de niveau $\alpha > 0{,}5$). A partir de la valeur $\alpha = 0{,}5$, les coupes de niveau α des cellules de la partition ont des intersections non-nulles ce qui invalide l'hypothèse d'indépendance. Il est souhaitable, pour éviter un biais des estimations, de prendre en compte ce couplage pour estimer le transfert de comptage.

L'approche classique considérerait comme connue la probabilité de couplage des cellules voisines. Ce couplage pourrait alors être pris en compte en utilisant une régularisation probabiliste bayésienne. Le transfert obtenu resterait précis mais cette précision serait arbitraire puisque le couplage dépend des données.

La nouvelle généralisation de la méthode pignistique que nous proposons exploite la représentation emboîtée de l'histogramme grâce à une intégrale de Choquet [8]. Cette méthode provoque un transfert imprécis qui suppose l'uniformité pour chaque sous-ensemble issus du découpage emboîté en α-coupe (base du transfert pignistique) mais suppose inconnu le couplage probabiliste entre les intervalles. L'estimation de comptage produite est donc imprécise. A la valeur précise $\widehat{nb}(W;(x_i))$ de l'estimation de $nb(W;(x_i))$, on doit substituer l'intervalle :

$$\widehat{Nb}(W;(x_i)) = [\underline{\widehat{nb}}(W;(x_i)), \widehat{nb}(W;(x_i))] . \tag{5}$$

Pour réaliser une estimation supérieure de $\widehat{nb}(W;(x_i))$ par une intégrale de Choquet on définit $a_{(k)}$ la série des accumulateurs triés, telle que $a_{(1)} \geq a_{(2)} \geq \ldots \geq a_{(p)}$. On en déduit les coalitions[1] floues $E_{(k)}$:

$$E_{(k)} = C_{(1)} \cup \ldots \cup C_{(k)} \tag{6}$$

et la mesure de confiance (ou capacité) $\upsilon(E_{(k)};W)$:

$$\upsilon(E_{(k)};W) = \frac{|E_{(k)} \cap W|}{|UC \cap W|} \frac{|W|}{\Delta} \text{ avec } UC = \bigcup_k C_k \tag{7}$$

Alors le transfert flou supérieur s'écrit :

$$\widehat{nb}(W;(x_i)) = \sum_{k=1}^{p} a_k (\upsilon(E_{(k)};W) - \upsilon(E_{(k-1)};W)) \tag{8}$$

On pose par convention $\upsilon(E_{(0)};W) = 0$.

L'estimation inférieure utilise les propriété de complémentation de l'intégrale de Choquet :

$$\underline{\widehat{nb}}(W;(x_i)) = -\widehat{nb}(W;-(x_i)) \tag{9}$$

1. Union d'ensembles élémentaires recouvrant une partie de Ω dont les mesures associées sont cohérentes.

Pour transformer un comptage imprécis en probabilité imprécise, il suffit d'utiliser la généralisation de l'équation (3) proposé dans [9] :

$$\overline{P}(W;(x_i)) = \frac{\overline{nb}(W;(x_i))}{\overline{nb}(W;(x_i)) + \underline{nb}(W^c;(x_i))},$$

$$\underline{P}(W;(X_i)) = \frac{\underline{nb}(W;(x_i))}{\underline{nb}(W;(x_i)) + \overline{nb}(W^c;(x_i))} \qquad (10)$$

4. Expérimentations.

Nous avons procédé à un grand nombre d'expérimentations dont nous proposons ici deux illustrations.

La première série d'illustrations utilise un lot de données classiques extrait du livre de Silverman [4] et concernant les éruptions du geyser Old Faithful. Nous comparons la reconstruction imprécise de la densité de probabilité obtenue par notre méthode à des reconstructions obtenues avec 6 types de noyaux monomodaux symétriques ayant la même granularité que l'histogramme. Nous avons choisi ici de créer un histogramme de 30 cellules (ce choix influe surtout sur la séparation des modes [3]) donc la granularité est Δ=2 minutes. Les noyaux que nous utilisons pour cette expérimentation sont les suivants :

- noyau d'Epanechnikov : $k(u) = \frac{3}{4}(1-u^2)$
- noyau cosinus : $k(u) = \frac{\pi}{4}\cos\left(u\frac{\pi}{2}\right)$
- noyau hyperlisse : $k(u) = \exp((-1)/(1-u^2))$
- noyau triangle : $k(u) = 1-|u|$
- noyau exponentiel : $k(u) = \frac{1}{2}\exp(-|u|)$
- noyau gaussien : $k(u) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{u^2}{2}\right)$.

Les noyaux sont bien sûr évalués sur un support [-1, 1].

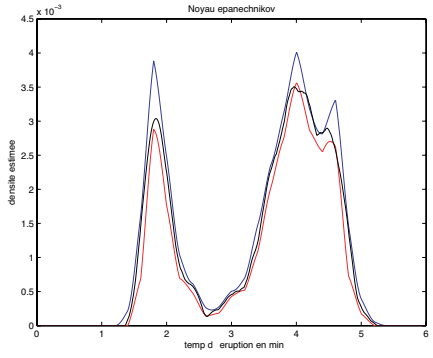

Figure 2 : comparaison avec le noyau d'Epanechnikov.

Sur les résultats présentés, l'estimation par noyau précis est donnée par la courbe noire continue tandis que les estimations supérieures et inférieures sont données respectivement par les courbes bleues (sup.) et rouges (inf.).

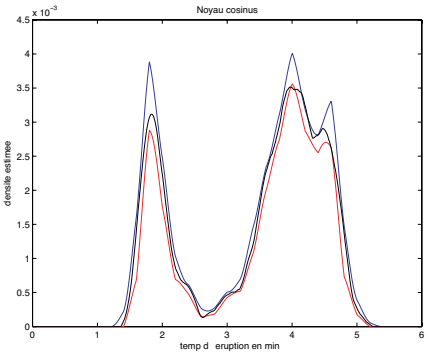

Figure 3 : comparaison avec le noyau cosinus

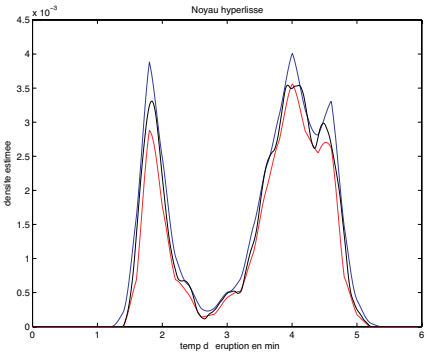

Figure 4 : comparaison avec le noyau hyperlisse

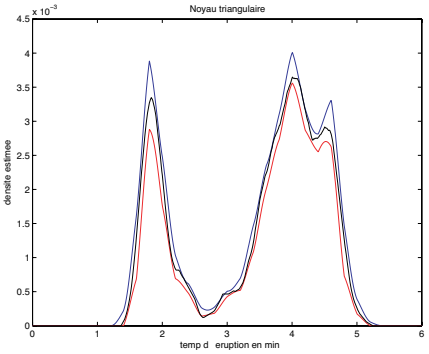

Figure 5 : comparaison avec le noyau triangle

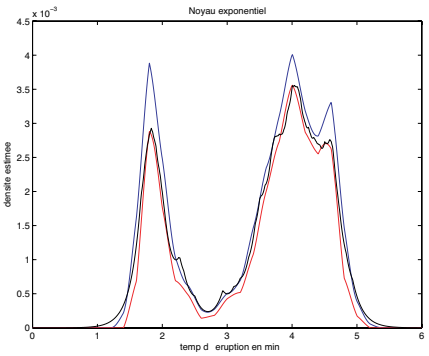

Figure 6 : comparaison avec le noyau exponentiel

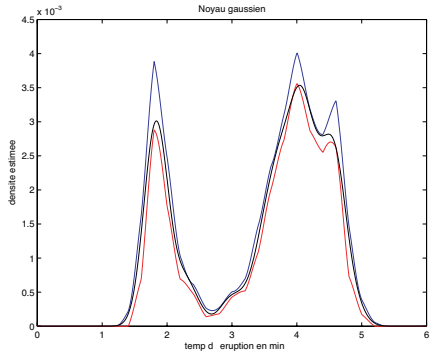

Figure 7 : comparaison avec le noyau gaussien

Cette série d'expérience montre que l'estimation obtenue par transfert de comptage est tout à fait cohérente avec celle qu'on obtiendrait avec un noyau monomodal symétrique de même granularité.

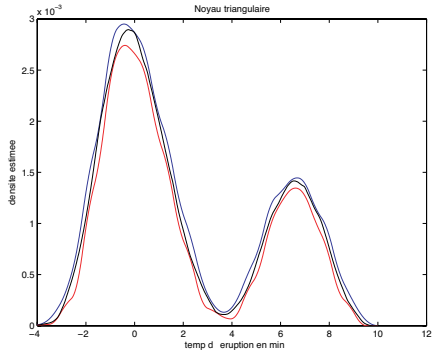

Figure 8 : estimation avec 15 cellules.

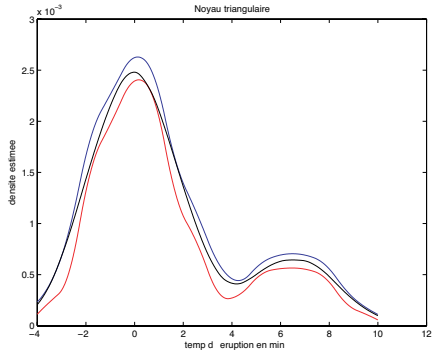

Figure 9 : estimation avec 9 cellules.

Dans la seconde expérience, nous avons simulé une distribution bimodale de 400 échantillons. Nous avons réalisé une estimation de la densité par un noyau triangulaire d'une part et un transfert pignistique imprécis issus d'un histogramme quasi-continu de 15 cellules et 9 cellules d'autre part (ce choix est arbitraire). Comme on peut le voir sur les figure 8 et 9, l'estimations par noyau triangulaire est en effet toujours incluse entre les deux bornes données par l'estimation par transfert pignistique imprécis. Cet encadrement découle d'une exploitation correcte des propriétés des quantité floues triangulaires [10]. De plus, on peut voir que le fait d'augmenter la granularité du partitionnement a pour effet d'augmenter l'imprécision d'estimation de la densité de probabilité et aussi de rendre plus difficile la séparation des deux modes.

5. Conclusion et prospectives.

Nous avons présenté une nouvelle méthode d'estimation de densité de probabilité basée sur la technique des histogrammes quasi-continus. Cette méthode permet de répercuter l'altération produite par le partitionnement (flou) de la droite réelle sur la précision de l'estimation de la densité. De nombreux points n'ont pas été abordés dans cet article dont le traitement spécifique des première et dernière cellules (de granularité infinie), l'estimation de densité de probabilité basées sur des données dont on connaît l'imprécision et/ou l'incertitude ainsi que l'utilisation des distributions de probabilité imprécises pour l'estimation de grandeurs statistiques (fractile imprécis, mode imprécis, moment imprécis, test statistique imprécis, etc.).

6. Références.

[1] D. Dubois, H. Prade, "Rough fuzzy sets and fuzzy rough sets", International Journal of General Systems, n°17, pp. 191-200, 1990.

[2] O. Strauss, F. Comby, M.J. Aldon. Rough histograms for robust statistics. Proceedings of the International Conference on Pattern Recognition, vol2, pp. 688-691. Sept. 2000.

[3] F. Comby, O. Strauss, M.J. Aldon, "Possibility theory and rough histograms for motion estimation in a video sequence", IWVF'O1: 4th International Workshop on Visual Form, Capri, Italy, May 28-30 2001.

[4] B.W. Silverman, "Density estimation for statistics and data analysis", Chapman & Hall/CRC, London 1986.

[5] T. Runkler, J. Bezdek, "Alternating cluster estimation: A new tool for clustering and function approximation". IEEE Trans. on Fuzzy Systems, 7(4):377--393, 1999.

[6] P. Smets, "The transferable belief model", Artificial Intelligence (66), pp. 191-243, 1994.

[7] D. Dubois, H. Prade, "Possibility theory: an approach to computerized rocessing of uncertainty", Plenum Press, London, 1985

[8] J.-L. Marichal, An axiomatic approach of the discrete Choquet integral as a tool to aggregate interacting criteria, IEEE Transactions on Fuzzy Systems 8 (6) pp. 800-807, 2000.

[9] O. Strauss, F. Comby, "Les histogrammes quasi-continus", XXXVèmes Journées de Statistiques, Lyon, 2-6 juin 2003, pp. 847-850..

[10] D. Dubois, L. Foulloy, G. Mauris, H. Prade, Probability-possibility transformation, triangular fuzzy sets, and probabilistic inequalities, Reliable Computing, vol. 10, pp. 273-297, 2004.