Construction d'histogrammes irréguliers par maximum de vraisemblance pénalisé

Valentina ZELAYA MENDIZABAL [1] & Marc BOULLE [2] & Fabrice ROSSI [3]

[1] *Orange Labs, Université Panthéon-Sorbonne,*
valentina.zelayamendizabal@orange.com
[2] *Orange Labs,* marc.boulle@orange.com
[3] *Université Paris-Dauphine,* Fabrice.Rossi@dauphine.psl.eu

Résumé. Nous proposons dans cette communication une méthode de construction totalement automatique d'histogrammes irréguliers basée sur le principe du *Minimum Description Length*. Couplée à une heuristique d'optimisation, notre approche permet une construction en $\mathcal{O}(n \log n)$ pour n observations ce qui la rend applicable à des données très volumineuses, contrairement aux méthodes existantes de même nature. Une évaluation expérimentale sur des données synthétiques et réelles montre les atouts de notre approche par rapport à celles de l'état de l'art.

Mots-clés. Histogramme, estimation de densité, sélection de modèle

Abstract. We present in this paper a new fully automated method for irregular histogram construction based on the *Minimum Description Length* principle. Associated to a greedy search heuristic, our method scales in $\mathcal{O}(n \log n)$ for n observations and can applied to large scale data sets, contrarily to existing work. An experimental evaluation on synthetic and real data shows the strengths and limitations of our approach compared to state-of-the-art methods.

Keywords. Histograms, density estimation, model selection

1 Construction automatique d'histogrammes

Malgré leur défauts, les histogrammes restent un outil populaire d'estimation de densité, notamment en raison de leur interprétation visuelle simple. Leur emploi systématique est notamment facilité par l'utilisation de règles simples pour les construire : beaucoup de logiciels statistiques se contentent de proposer des histogrammes réguliers, avec des intervalles de longueurs égales, et de choisir le nombre d'intervalles au moyen de règles simples comme la vénérable règle de Sturges [7] ou celle de Freedman-Diaconis [3]. Cependant pour des distributions complexes, comme celles à queues lourdes, des histogrammes à intervalles de longueurs variables sont plus adaptés et il existe assez peu de méthodes complètement automatique pour adapter aux données le nombre d'intervalles et les points de coupure (cf [2, 5] pour deux états de l'art sur ces méthodes).

Nous proposons dans cette communication une nouvelle méthode de ce type. Comme dans la plupart des travaux s'attaquant à ce problème, nous considérons la construction

d'un histogramme comme un problème de sélection de modèle. Ce type de problème est généralement résolu par une stratégie de maximum de vraisemblance pénalisé. Les pénalités retenues dans les méthodes les plus efficaces comme [5] s'appuient sur des résultats théoriques mais aussi sur des considérations heuristiques et expérimentales.

Nous utilisons dans cette communication une approche similaire basée sur le principe du *Minimum Description Length* (MDL) et notamment sur le « *maximum de vraisemblance normalisé* » utilisé dans [4] pour dériver un critère de sélection de modèle (`NML`) adapté aux histogrammes irréguliers. Ce critère donne des résultats convaincants sur des distributions simples mais présente deux limitations : il dépend d'un paramètre de précision et son optimisation est coûteuse en temps de calcul.

Nous proposons dans cette communication un nouveau critère inspiré de `NML` et de type MDL, qui l'améliore sur deux points. Tout d'abord, notre critère permet d'automatiser le choix de la précision de représentation des données. D'autre part, il est plus rapide à calculer que le `NML`. Nous proposons en outre une heuristique d'optimisation qui conduit à une construction d'un histogramme en $\mathcal{O}(n \log n)$ pour n observations (contre un temps polynomial pour `NML`).

Nous consacrons la suite de cette communication à la présentation du critère proposé et à celle d'une évaluation comparative de ses performances comparées à l'état de l'art.

2 Histogrammmes G-Enum

2.1 Formulation du problème

Nous considérons un échantillon de n observations $x^n = (x_1, ..., x_n)$ sur l'intervalle $[x_{\min}, x_{\max}]$. On note ϵ la précision de représentation des données : il s'agit du paramètre à régler dans le `NML`. Chaque observation $x_j \in x^n$ est approchée par un élément de $\widetilde{x}_j \in \mathcal{X} = \{x_{\min} + t\epsilon; t = 0, ..., E\}$ où $E = \dfrac{x_{\max} - x_{\min}}{\epsilon}$.

On considère des histogrammes construits à partir de la grille $\mathcal{C}$ obtenue à partir des milieux de paires de valeurs consécutives de $\mathcal{X}$, comme dans [4]

$$\mathcal{C} = \{x_{\min} + \epsilon/2 + t\epsilon; t = 0, ..., E-1\},$$

avec $c_0 = x_{\min} - \epsilon/2$ et $c_K = x_{\max} + \epsilon/2$. Ces points de coupure définissent E *cellules élémentaires* de longueur ϵ, que nous appelons ϵ-bins (voir figure 1). Ils constituent les éléments de base des intervalles d'un histogramme : chaque combinaison de ϵ-bins en K intervalles, avec K allant de 1 à E, définit un modèle d'histogramme. Dans cet éventail de possibilités, notre objectif est de sélectionner un ensemble de $K-1$ points de coupure $C = (c_1, ..., c_{K-1}),\ c_k \in \mathcal{C}$ tels que $[x_{\min} - \epsilon/2, x_{\max} + \epsilon/2]$ soit partitionné en K intervalles $\{[c_0, c_1],]c_1, c_2], ...,]c_{K-1}, c_K]\}$ adaptés à la distribution réelle des données (voir figure 1). Chaque intervalle k a un effectif de h_k observations et une longueur $L_k = c_k - c_{k-1}$, qui est un multiple de ϵ : $\forall k,\ \exists\ E_k \in \mathbb{N}^*$ tel que $L_k = E_k \cdot \epsilon$.

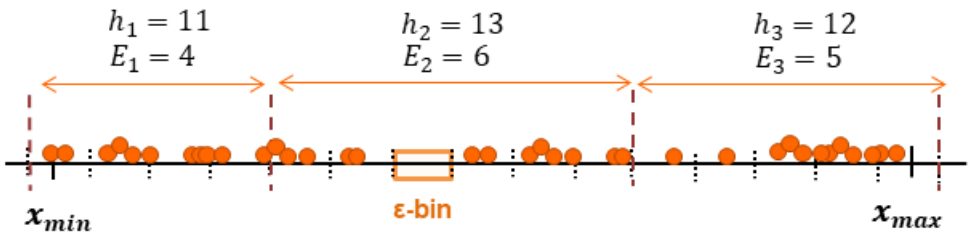

FIGURE 1 – Un choix possible d'intervalles, avec leurs effectifs et longueurs

Un histogramme est entièrement défini par le choix du nombre d'intervalles, l'ensemble des points de coupure qui les définissent et les effectifs associés.

2.2 Formalisation d'un critère granularisé pour les histogrammes

Les critères proposés sont donnés sans justification dans la table 1 pour des raisons de place. Ils sont obtenus en interprétant l'approche MDL comme une recherche de maximum a posteriori. Chaque critère comprend un terme de vraisemblance et un terme de codage ou d'a priori sur les paramètres du modèle (ici l'histogramme). Dans la table, ces deux termes sont réorganisés pour faire apparaître les différences entre les trois critères.

Le critère `Enum` est directement issu de ce point de vue bayésien sur le MDL et préserve certaines propriétés d'optimalité du Maximum de vraisemblance Normalisé [1]. Par rapport au critère `NML`, il évite un terme complexe à calculer, log $\mathcal{R}^n_{\mathcal{M}}$ la complexité paramétrique [4]. En outre la pénalisation induite est croissante avec le nombre d'intervalles.

Afin d'automatiser le choix du seul paramètre des méthodes `NML` et `Enum`, nous introduisons une version avec une granularité G à optimiser : un intervalle d'histogramme est maintenant constitué de G_k g-bins, chaque g-bins étant elle même constituée de $g = \frac{E}{G}$ ϵ-bins. On découple ainsi la précision de représentation des données (les ϵ-bins) de la précision de représentation des histogrammes (les g-bins). On peut ainsi fixer ϵ autour de la précision machine, g étant optimisé dans la procédure de sélection de modèle. Le critère obtenu est baptisé `G-Enum`.

2.3 Heuristiques de recherche

L'algorithme de programmation dynamique proposé pour optimiser le critère `NML` est optimal, mais il nécessite un temps de calcul en $\mathcal{O}(E^3)$ [4]. Nous utilisons divers heuristiques de recherche, s'appuyant notamment sur l'additivité des critères et sur une stratégie gloutonne de fusion d'intervalles adjacents. La complexité est ainsi réduite en $\mathcal{O}(n \log n)$. Pour limiter l'impact de cette approche sur la qualité des histogrammes obtenus, nous utilisons des post-optimisations heuristiques comme des découpages et combinaisons des intervalles une fois un optimum local atteint.

TABLE 1 – Comparaison des termes entre critères

Critère	Termes d'indexation	Termes multinomiaux	Termes d'indexation des bins
NML [4]	$\log \binom{E}{K-1}$	$\log\ \mathcal{R}^n_{\mathcal{M}} + \log \frac{n^n}{{h_1}^{h_1} \cdots {h_K}^{h_K}}$	$\sum_{k=1}^{K} h_k \log E_k$
Enum	$\log^* K + \log \binom{E+K-1}{K-1}$	$\log \binom{n+K-1}{K-1} + \log \frac{n!}{h_1!...h_K!}$	$\sum_{k=1}^{K} h_k \log E_k$
G-Enum	$\log^* K\ +\ \log^* G\ + \log \binom{G+K-1}{K-1}$	$\log \binom{n+K-1}{K-1} + \log \frac{n!}{h_1!...h_K!}$	$\sum_{k=1}^{K} h_k \log G_k + n \log \frac{E}{G}$

3 Évaluation expérimentale

3.1 Protocole et métriques

Nous évaluons notre stratégie par comparaison avec une sélection de méthodes concurrentes automatiques (règles de Sturges et Freedman-Diaconis [3], taut strings [2], RMG [5] et Bayesian blocks [6]) sur plusieurs échantillons de 6 types de distribution. Ces méthodes sont testées sur des échantillons de tailles croissantes, allant de $n = 10$ à $n = 10^5$ ou $n = 10^6$.

Les résultats sont comparés selon trois métriques : le nombre d'intervalles, le temps de calcul et la distance de Hellinger par rapport au modèle de distribution. Afin d'évaluer la variabilité des résultats, nous présentons les moyennes et les écarts types calculés sur un total de 10 d'expériences pour une distribution donnée et une taille d'échantillon donnée.

3.2 Résultats clés

Comparaison entre méthodes MDL

Les histogrammes `NML` et `Enum` sont interchangeables en termes de nombre d'intervalles et de distance de Hellinger, et ce quel que soit l'algorithme choisi pour optimiser les critères. Toutefois, en termes de temps de calcul, il y a un avantage significatif à préférer l'heuristique de recherche et les critères énumératifs plus simples. Les histogrammes `G-Enum` prennent un peu plus de temps à calculer mais produisent des estimations légèrement meilleures en termes de distance de Hellinger et ne nécessitent pas de fixer ϵ de façon spécifique.

Comparaison avec d'autres méthodes de l'état de l'art (tables 2 et 3)

Les autres méthodes ont des meilleurs résultats dans les cas spécifiques pour lesquels elles ont été conçues. Bien qu'ils soient rarement en première place pour chaque type de distribution, les histogrammes `G-Enum` sont toujours parmi les meilleurs estimateurs, et ce sans la forte variabilité des autres méthodes. Parmi les histogrammes irréguliers, `G-Enum` se distingue comme le plus parcimonieux en nombre d'intervalles. Il s'agit d'une qualité importante pour l'analyse exploratoire car cela facilite l'interprétation des résultats. `G-Enum` est également de loin la plus rapide des méthodes irrégulières, ce qui la rend adaptée aux jeux de données de grande taille.

Application à des jeux de données réels de grande taille

Une application à quelques jeux de données de grande taille (entre 1 à 25 millions d'entrées) et de différents domaines (de tailles de cratères lunaires jusqu'à durées d'appels des clients Orange) montre que les histogrammes `G-Enum` constituent une méthode automatique, efficace et robuste pour l'analyse exploratoire de données dont la distribution est inconnue.

Références

[1] M. Boullé, F. Clérot, and C. Hue. Revisiting enumerative two-part crude MDL for Bernoulli and multinomial distributions (extended version). Technical report, arXiv, abs/1608.05522, 2016.

[2] Davies, Laurie, Gather, Ursula, Nordman, Dan, and Weinert, Henrike. A comparison of automatic histogram constructions. *ESAIM : PS*, 13 :181–196, 2009.

[3] David Freedman and Persi Diaconis. On the histogram as a density estimator :l2 theory. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 57(4) :453–476, Dec 1981.

[4] Petri Kontkanen and Petri Myllymäki. Mdl histogram density estimation. In Marina Meila and Xiaotong Shen, editors, *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, volume 2 of *Proceedings of Machine Learning Research*, pages 219–226, San Juan, Puerto Rico, 21–24 Mar 2007. PMLR.

[5] Yves Rozenholc, Thoralf Mildenberger, and Ursula Gather. Combining regular and irregular histograms by penalized likelihood. *Computational Statistics and Data Analysis*, 54(12) :3313 – 3323, 2010.

[6] Jeffrey D. Scargle, Jay P. Norris, Brad Jackson, and James Chiang. Studies in Astronomical Time Series Analysis. VI. Bayesian Block Representations. *Astrophysical Journal*, 764(2) :167, February 2013.

[7] Herbert A. Sturges. The choice of a class interval. *Journal of the American Statistical Association*, 21(153) :65–66, 1926.

TABLE 2 – Comparaison des distances de Hellinger sur différents jeux de taille $n = 10^4$

Distribution	G-Enum	NML [4]	BB [6]	TS [2]	RMG [5]	FD [3]	Sturges
Normale	$0.045 \pm 6 \cdot 10^{-4}$	0.046 ± 0.002	0.047 ± 0.002	0.040 ± 0.002	0.034 ± 0.002	0.033 ± 0.002	0.055 ± 0.002
Cauchy	0.061 ± 0.004	0.074 ± 0.003	0.064 ± 0.002	0.045 ± 0.005	0.064 ± 0.001	0.138 ± 0.002	0.862 ± 0.036
Uniforme	0.024 ± 0.001	0.050 ± 0.005	0.025 ± 0.004	0.031 ± 0.015	0.029 ± 0.011	0.082 ± 0.011	0.028 ± 0.002
Triangle	0.039 ± 0.002	0.038 ± 0.0025	0.039 ± 0.001	0.084 ± 0.024	0.084 ± 0.029	0.032 ± 0.002	$0.049 \pm 9 \cdot 10^{-4}$
4 triangles	0.037 ± 0.002	0.038 ± 0.003	0.040 ± 0.003	0.078 ± 0.029	0.069 ± 0.026	0.032 ± 0.002	$0.043 \pm 4 \cdot 10^{-4}$
6 gaussiennes	0.057 ± 0.002	0.059 ± 0.002	0.060 ± 0.002	0.040 ± 0.001	0.052 ± 0.002	0.060 ± 0.001	0.142 ± 0.013

TABLE 3 – Comparaison des temps de calcul (en secondes) sur différents jeux de taille $n = 10^4$

Distribution	G-Enum	NML [4]	BB [6]	TS [2]	RMG [5]	FD [3]	Sturges
Normale	0.014 ± 0.003	2.724 ± 0.283	5.785 ± 0.479	0.014 ± 0.002	1.239 ± 0.085	0.002 ± 2.10^{-4}	$6.10^{-4} \pm 2.10^{-4}$
Cauchy	0.028 ± 0.006	121.60 ± 4.99	3.250 ± 0.112	0.116 ± 0.205	0.906 ± 0.107	0.009 ± 0.014	0.001 ± 0.003
Uniforme	0.015 ± 0.002	0.168 ± 0.012	5.989 ± 0.167	0.011 ± 0.002	1.387 ± 0.139	0.002 ± 3.10^{-4}	$6.10^{-4} \pm 2.10^{-4}$
Triangle	0.014 ± 0.005	0.169 ± 0.005	5.962 ± 0.113	0.015 ± 0.002	1.291 ± 0.091	0.002 ± 2.10^{-4}	$6.10^{-4} \pm 2.10^{-4}$
4 triangles	0.012 ± 0.006	0.103 ± 0.027	3.004 ± 0.245	0.013 ± 0.006	0.954 ± 0.138	0.002 ± 0.005	0.0 ± 0.0
6 gaussiennes	0.017 ± 0.002	1.91 ± 0.085	4.165 ± 0.369	0.048 ± 0.005	1.056 ± 0.077	0.006 ± 0.008	0.002 ± 0.005

TABLE 4 – Comparaison du nombre d'intervalles sur différents jeux de taille $n = 10^4$

Distribution	G-Enum	NML [4]	BB [6]	TS [2]	RMG [5]	FD [3]	Sturges
Normale	16.30 ± 0.46	15.60 ± 1.02	15.90 ± 1.04	72.50 ± 4.41	39.70 ± 7.34	62.40 ± 2.29	15.0 ± 0.0
Cauchy	30.90 ± 2.43	23.60 ± 1.02	29.40 ± 1.56	144.90 ± 9.26	29.50 ± 1.57	110711.90 ± 132580.43	15.0 ± 0.0
Uniforme	1.0 ± 0.0	2.80 ± 0.60	1.30 ± 0.90	3.70 ± 5.44	1.70 ± 1.80	22.0 ± 0.0	15.0 ± 0.0
Triangle	12.50 ± 0.92	13.60 ± 0.66	12.60 ± 0.66	48.0 ± 5.85	33.70 ± 8.25	32.10 ± 0.30	15.0 ± 0.0
4 triangles	11.20 ± 0.75	12.00 ± 0.77	10.90 ± 0.70	42.30 ± 4.90	27.60 ± 8.0	30.80 ± 0.40	15.0 ± 0.0
6 gaussiennes	28.90 ± 1.22	27.30 ± 1.49	27.00 ± 2.00	134.90 ± 9.37	100.40 ± 11.60	66.40 ± 2.42	15.0 ± 0.0