Le vecteur de meilleur rang moyen : une statistique pour l'analyse de données multidimensionnelles - Application au filtrage d'images couleurs

Cyril DE RUNZ[1], Michel HERBIN[2], Frédéric BLANCHARD[1], Laurent HUSSENET[1], Valeriu VRABIE[2], Philippe VAUTROT[1]

[1]CReSTIC-SIC, IUT de Reims Champagne-Ardenne,
Rue des Crayères, BP 1035, 51687 Reims cedex 2, tél : +33 3 26 91 84 58

[2]CReSTIC, Université de Reims Champagne-Ardenne, IUT,
chaussée du port, BP 541, 51012 Châlons-en-Champagne Cedex, France, tél : +33 3 26 91 81 98
michel.herbin@univ-reims.fr

Résumé – L'obtention du meilleur représentant d'un échantillon est une question ouverte et dépend de la statistique utilisée. Cette communication propose une statistique, qui repose sur le calcul des rangs, et qui permet en outre d'introduire la notion de *données représentant le mieux son échantillon*. La donnée représentant le mieux son échantillon est nommée *donnée de meilleur rang moyen*. En appliquant notre statistique au filtrage d'images couleurs, nous illustrons l'efficacité et l'intérêt de notre procédure, visuellement et numériquement.

Abstract – The obtention of the most representative element of a dataset is an open question and depends on the statistic used. This paper proposes a statistic, which is computed using rank statistics, and which allows to introduce to notion of the *most representative data* of a dataset. The so defined best representant of the dataset is called *best mean rank data*. The use of this statistic in colour image filtering illustrates the efficiency and the interest of our procedure both visually and numerically.

1 Introduction

L'utilisation des statistiques non paramétriques et plus particulièrement celle des statistiques de rangs, connaît depuis quelques années un regain d'intérêt [3]. L'utilisation de ces dernières en analyse de données permet notamment de ne pas être assujetti à l'hypothèse de normalité et apporte de la robustesse aux méthodes les utilisant [2].

Nous proposons ici une nouvelle statistique, dont le calcul repose sur les statistiques de rangs, et qui hérite ainsi des avantages cités précédemment. Elle permet en outre d'introduire la notion de *données représentant le mieux son échantillon*. Autrement dit, et plus précisément, nous définissons une fonctionnelle de statistiques de rangs qui permet de déterminer, dans un échantillon de données, quelle est celle qui représente le mieux, au sens de l'ordre que nous proposons, ledit échantillon.

Les statistiques de rangs sont classiquement utilisées en filtrage d'images [5]. Ces statistiques donnent lieu à de nombreuses applications en débruitage d'images couleurs [4, 6]. Le filtre médian [1] est sans doute le plus connu de ces outils. Nous proposons d'appliquer notre statistique au filtrage d'images couleurs affectées d'un bruit impulsionnel, et de comparer nos résultats à ceux obtenus avec le filtre médian.

Ce document présente, après l'exposé du principe théorique de notre concept, une illustration de son utilisation pour le débruitage d'images. Une discussion et des perspectives sont enfin proposées avant de conclure.

2 L'élément de meilleur rang moyen

Considérons un échantillon de données multidimensionnelles $S = \{x_1, x_2, \ldots, x_n\}$ dans un espace E de dimension $p \in \mathbb{N}$. On suppose que l'on dispose, sur cet échantillon, d'un indice de dissimilarité (ou, de façon duale, d'un indice de similarité). Autrement dit, on suppose que l'on dispose d'un moyen de quantifier la dissimilarité entre deux éléments quelconques de notre échantillon S. On notera $\delta(x_i, x_j)$ la dissimilarité entre x_i et x_j $(i, j \in [1..n])$. La distance euclidienne est un exemple d'indice de dissimilarité.

2.1 Statistiques de rangs marginales

On notera $X_1, X_2, ..., X_n$ les variables (vecteurs) aléatoires dont les éléments de l'échantillon S sont les observations. Les statistiques d'ordres associées sont les $X_{(1)}$, $X_{(2)}$, ..., $X_{(n)}$ triés par ordre croissants (et on notera $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ les observations ordonnées associées). Par définition, les statistiques d'ordres sont donc intrinsèquemet liées à la façon dont sont triées les variables aléatoires. Dans le cas multidimensionnel, le tri n'est pas trivial, on se reportera à [2] pour une étude des différentes techniques pour trier des vecteurs.

Considérons maintenant les n classements (i.e. les n tris) obtenus en utilisant les dissimilarités par rapport à chaque x_i. Autrement dit, pour chaque élément x_i, nous classons l'ensemble de l'échantillon par ordre de dissimilarité croissante avec x_i.

Soit $Rg_{x_i}(x_j)$ le rang de la donnée x_j dans le classement par dissimilarité croissante à x_i. On a naturellement : $\forall i, j \in [1.n], Rg_{x_i}(x_j) \in [1..n]$ et $\forall i \in [1..n], Rg_{x_i}(x_i) = 1$. La valeur de $Rg_{x_i}(x_j)$ représente ainsi la position de x_j dans le classement des données les plus similaires à x_i. Par exemple, $Rg_{x_i}(x_j) = k$ ($k \in [1..n]$) signifie que x_j est la k-ième donnée de S la plus similaire à x_i, c'est à dire $x_{(k)}$ dans l'ordre pour la donnée x_i.

Ainsi, les n individus de l'échantillon (x_i avec i entre 1 et n) induisent n classements différents.

2.2 Rang moyen d'une donnée

On calcule ensuite, pour chaque donnée x_j de S, la moyenne des rangs $Rg_{x_i}(x_j)$ qu'il a obtenus au cours de ces n classements. On note ce rang moyen : $\overline{Rg}(x_j)$ et on a $\overline{Rg}(x_j) = 1/n * \sum_i Rg_{x_i}(x_j)$.

Le rang moyen est un critère qui nous permet alors d'évaluer le potentiel d'une donnée à représenter l'échantillon auquel elle appartient. En effet, cette valeur moyenne traduit la façon dont une donnée est *la plus similaire* de l'ensemble des autres. On appelle cette notion la *représentativité d'une donnée dans son échantillon.*

2.3 Statistique de meilleur rang moyen

Nous terminons notre processus par la recherche dans l'échantillon, de la donnée ayant le plus petit rang moyen, c'est à dire la donnée de l'échantillon la plus représentative dudit échantillon.

Finalement, nous avons donc, au cours de ces étapes, défini une statistique exprimée comme une fonctionnelle des statistiques de rangs marginales et notée *MRM* (Meilleur Rang Moyen) :

$$MRM : (X_1, X_2, \dots X_n) \mapsto \underset{X_i, i=1..n}{argmin} \left(\overline{Rg}(X_j)\right)$$

Cette statistique associe à un échantillon l'élément qui le représente le mieux. Cette notion de meilleur représentant d'un échantillon rejoint, d'un point de vue sémantique, la notion de représentant de classes en classification automatique des données. De plus, au même titre que la médiane, notre statistique est un estimateur robuste de position de l'échantillon. En effet, la donnée de meilleur rang moyen est un élément typique et représentatif de l'échantillon.

Nous allons maintenant illustrer l'utilisation de cette nouvelle statistique à travers un exemple numérique simple avec une donnée aberrante pour montrer la robustesse de MRM.

2.4 Exemple

Soit l'échantillon $S = \{x_1, x_2, x_3, x_4, x_5, x_6\}$ constitué de 6 vecteurs de $\mathbb{R}^4$. Les valeurs numériques des 6 vecteurs correspondants sont présentées dans le tableau 1a. En prenant la distance euclidienne comme indice de dissimilarité entre les données, on obtient les rangs présentés dans le tableau 1b. La colonne Rg_{x_i} de ce tableau représente les statistiques de rangs induites par les dissimilarités avec la donnée x_i. Autrement dit, si l'élément de la $j^{\text{ième}}$ ligne et de la colonne Rg_{x_i} vaut k alors cela signifie que x_j est le $k$ième plus proche vecteur de x_i dans S. Les rangs moyens de chaque donnée de l'échantillon sont calculés à partir des rangs dans le tableau 1c.

La statistique MRM de meilleur rang moyen appliqué à l'échantillon S donne finalement l'élément x_4 dont le rang moyen est minimal. On peut remarquer que la donnée x_5 est aberrante mais qu'elle ne perturbe pas le résultat. En effet, on peut facilement vérifier que le vecteur x_4 est plus proche de la moyenne de l'ensemble des vecteurs privé de la donnée aberrante que le Vecteur Médian qui est ici x_2. Cette propriété de robustesse est une caractéristique des statistiques de rangs.

Tab. 1 – Exemple sur un échantillon de données

a : Echantillon exemple

x_1	10	-15	10	30
x_2	15	-60	20	50
x_3	55	-10	20	30
x_4	15	-25	15	25
x_5	30	-200	500	200
x_6	15	-60	20	75

b : Rangs

	Rg_{x_1}	Rg_{x_2}	Rg_{x_3}	Rg_{x_4}	Rg_{x_5}	Rg_{x_6}
x_1	1	4	2	2	5	4
x_2	4	1	4	3	3	2
x_3	3	5	1	4	6	5
x_4	2	3	3	1	4	3
x_5	6	6	6	6	1	6
x_6	5	2	5	5	2	1

c : Rangs moyens

	$\overline{Rg}$
x_1	3
x_2	2.83
x_3	4
x_4	2.67
x_5	5.17
x_6	3.33

Comme nous l'avons déjà rappelé, les statistiques d'ordre dont le Vecteur Médian sont fréquemment utilisées en filtrage d'images. Nous allons illustrer l'intérêt de notre contribution en l'utilisant dans ce contexte et en montrant sa supériorité par rapport au Vecteur Médian.

3 Application au filtrage d'images couleurs

Dans le contexte du filtrage d'images couleurs, il suffit de considérer chaque voisinage où l'on opère le filtrage, comme l'échantillon S dont on cherche l'image par notre statistique. Autrement lorsque l'on applique le filtre à l'aide d'un voisinage $t \times t$, on considère l'échantillon dont les éléments sont les t^2 pixels de dimension 3 (R,G,B) de ce voisinage, avec t entier naturel impair. On remplace alors le pixel central de cette fenêtre par le pixel le plus représentatif, c'est à dire le pixel de meilleur rang moyen

dans cet échantillon. On nommera par FMRM le filtre ainsi construit.

Le filtre Vecteur Median [1] est classiquement utilisé pour l'élimination du bruit d'images couleurs. Nous avons utilisé un jeu classique d'images sur lesquelles nous avons rajouté du bruit impulsionnel. Pour des fenêtre 5×5, l'efficacité de notre filtrage est similaire à celle issue du filtre Vecteur Median tant du point visuel (Fig. 3c et Fig. 3d), que numérique (cf. Tab. 2) en calculant l'erreur quadratique moyenne (MSE : Mean Square Error), l'erreur absolue moyenne (MAE : Mean Absolute Error), ainsi que l'erreur maximale (ME : max error).

4 Conclusion

Nous avons construit une nouvelle statistique utilisant les statistiques de rangs. Ce nouvel outil permet de désigner, dans un échantillon, l'élément représentant le mieux cet échantillon. Cette notion de représentativité est caractérisée par la valeur du rang moyen de chaque donnée. Le rang moyen est calculé par agrégation des rangs marginaux obtenus lors des classements induits par chaque donnée. La robustesse due à l'insensibilité aux valeurs aberrantes est une autre caractéristique de notre méthode. Par ailleurs, comme toutes les méthodes non paramétriques, elle ne nécessite pas d'a priori sur la distribution des données initiales.

Pour illustrer l'intérêt de notre méthode, nous avons utilisé la statistique de meilleur rang moyen dans le contexte du filtrage d'images couleurs. Le filtre obtenu se révèle efficace dans la réduction de bruit tout en préservant les arrêtes et contours. Notre filtre obtient à titre de comparaison des résultats similaires à ceux du filtre Vecteur Médian, qui est un outil classique basé, lui aussi, sur les rangs.

Références

[1] J. Astola, P. Haavisto, and Y. Neuvo. Vector median filters. In *Proceedings of the IEEE*, volume 78, pages 678–689, 1990.

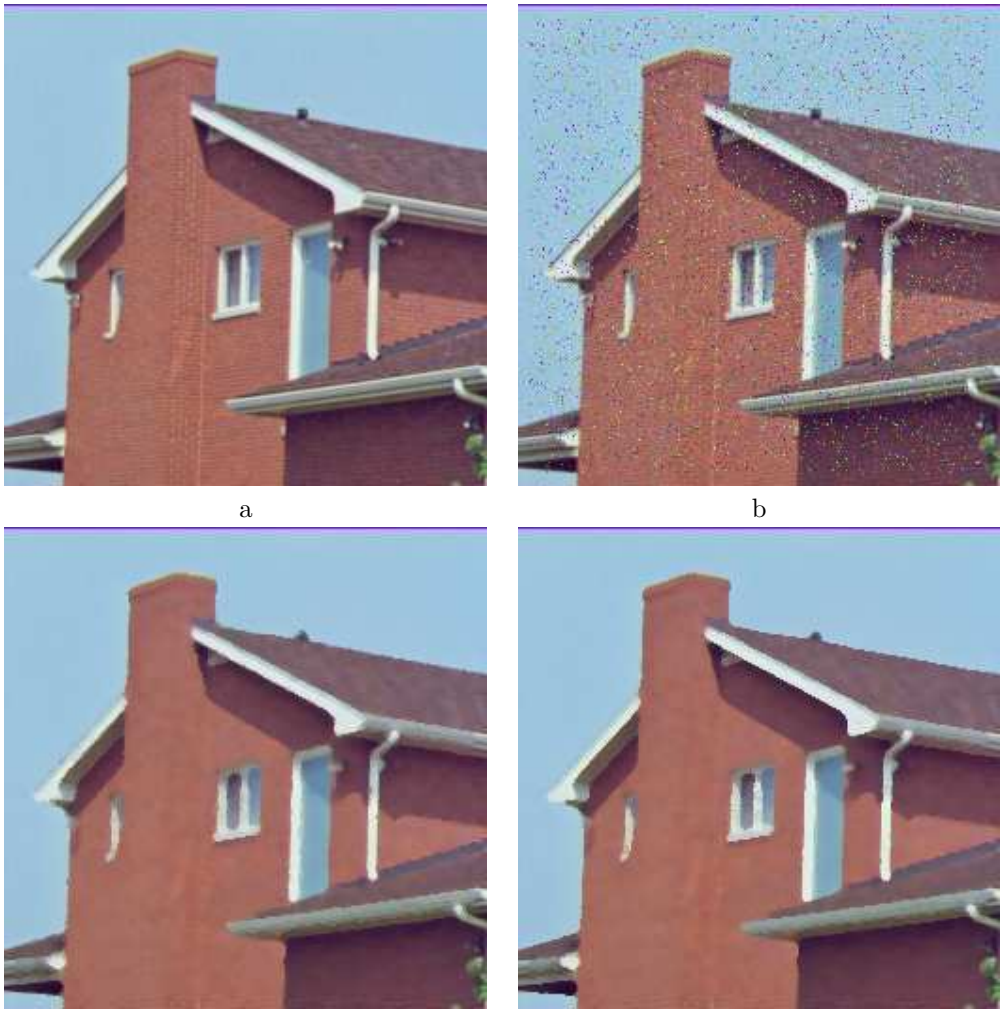
[2] V. Barnett. The ordering of multivariate data. *Journal of the Royal Statistical Society, Series A (General)*, 139(3) :318–355, 1976.

[3] H. A. David and H. N. Ngaraja. *Order Statistics*. Wiley, third edition, 2003.

[4] R. Lukac, B. Smolka, K.N. Plataniotis, and A.N. Venetsanopoulos. Vector sigma filters for noise detection and removal in color images. *J. Vis. Commun. Image R.*, 17 :1–26, 2006.

[5] I. Pitas and P. Tsakalides. Multivariate ordering in color image filtering. *IEEE Transactions on Circuits and Systems for Video Technology*, 1(3) :247–259, September 1991.

[6] P. Vautrot, L. Hussenet, and M. Herbin. A robust filtering method using owa filters : Application to color images. In *3rd European Conference on Colour in Graphics, Imaging and Vision*, University of Leeds, UK, jun 2006.



a b

c d

FIG. 1 – a : image originale (House), b : image bruitée (bruit impulsionnel 5%), c : image filtrée avec FMRM, d : image filtrée avec le Filtre Vecteur Median

TAB. 2 – Comparaison du Vecteur Median et FMRM sur un voisinage $3 * 3$

Image	Bruit (%)	MAE Median	MAE FMRM	MSE Median	MSE FMRM	ME Median	ME FMRM
House	1	5.16	5.10	97.41	90.62	107.67	105.33
House	5	5.17	5.08	95.87	86.95	107.67	105.33
House	10	5.18	5.32	96.32	97.16	109.67	109.67
House	20	5.28	6.01	98.19	120.77	115.00	115.00
Lena	1	8.11	8.03	223.03	212.36	146.67	146.67
Lena	5	8.16	8.15	223.81	212.13	146.67	146.67
Lena	10	8.24	8.39	224.93	216.80	146.67	146.67
Lena	20	8.45	9.18	229.08	243.74	151.00	120.00