Influence de la forme de la fenêtre de scan sur la distribution des statistiques de scan bidimensionnelles discrètes

Michaël Genin[1] & Cristian Preda[2] & Alain Duhamel[3]

[1] *Université de Lille, Faculté de Médecine (Pôle Recherche),*
Département de Biostatistiques
1, place de Verdun, 59000 Lille, France
michael.genin@univ-lille2.fr

[2] *Inria Lille-Nord-Europe, Equipe-projet MΘDAL*
Parc scientifique de la Haute Borne
40, avenue Halley - Bât A - Park Plaza
59650 Villeneuve d'Ascq, France
cristian.preda@polytech-lille.fr

[3] *Université de Lille, Faculté de Médecine (Pôle Recherche),*
Département de Biostatistiques
1, place de Verdun, 59000 Lille, France
alain.duhamel@univ-lille2.fr

Résumé. Les statistiques de scan bidimensionnelles discrètes sont usuellement définies avec une fenêtre de scan de forme rectangulaire. Cependant, elles peuvent être définies pour toute forme convexe de fenêtre de scan. Aussi, dans ce travail, nous nous intéressons à l'influence de la forme de la fenêtre de scan sur la distribution de probabilité des statistiques de scan bidimensionnelles discrètes.. Nous montrons que la forme de la fenêtre de scan a une influence sur la distribution. Ceci est réalisé par une adaptation d'une méthode d'approximation de la distribution basée sur les propriétés des *extremums* de suites de variables aléatoires 1-dépendantes aux statistiques de scan bi-dimensionnelles à fenêtre de forme convexe. Ce résultat est illustré par une étude de simulation pour les modèles de Poisson et binomiaux, dans laquelle nous avons considéré les cas des formes carrées, rectangulaires et circulaires (cercle discret).

Mots-clés. Statistiques de scan discrètes, Forme de fenêtre de scan, Simulation

Abstract. The two-dimensional discrete scan statistics are usually defined with a rectangular scanning window shape. However, they can be defined for all convexe shape of scanning window. In this work, we study the influence of the shape of the scanning window on the distribution of the two-dimensional discrete scan statistics distribution. We show that the scanning window shape has an influence on the distribution. This is

shown by the adaptation to general convex shape of an approximation method of the distribution based on the property of the extremums of 1-dependent random variable sequences. This result is highlighted by means of a simulation study for binomial and Poisson models in which, we take into consideration the cases of square, rectangular and circular (discrete circle) shape of the scanning window.

Keywords. Discrete scan statistics, Scanning window shape, Simulation

Texte long

Soient N_1, N_2 des entiers positifs, $\mathcal{R} = [0, N_1] \times [0, N_2]$ une région rectangulaire et $\{X_{ij}\}$, $1 \leq i \leq N_1$, $1 \leq j \leq N_2$, une famille de variables aléatoires à valeurs entières non négatives, *i.i.d.* et issues d'une distribution spécifique(Bernoulli, binomial, Poisson, etc ...). En pratique, les X_{ij} représentent le nombre d'évènements observés dans le carré élémentaire $\mathcal{R}_{i,j} = [i-1, i] \times [j-1, j]$.

Soit $\mathcal{W}$ un ensemble convexe dans $\mathbb{R}^2$. Nous dénotons par $[\mathcal{W}]$ l'ensemble de tous les carrés élémentaires adjacents $\mathcal{R}_{i,j}$ contenus dans $\mathcal{W}$,

$$[\mathcal{W}] = \bigcup_{\mathcal{R}_{ij} \subseteq \mathcal{W},\, (i,j) \in \mathbb{Z}^2} \mathcal{R}_{i,j}. \tag{1}$$

Soit $T([\mathcal{W}])$ l'ensemble de toutes les translations de $[\mathcal{W}]$ pour tout vecteur $\vec{u} \in \mathbb{Z}^2$. Ensuite, nous définissons la statistique de scan bi-dimensionnelle discrète sur la région $\mathcal{R}$ avec une fenêtre de scan $[\mathcal{W}]$ comme le nombre maximum d'évènements observés dans tout tout ensemble de $T([\mathcal{W}])$ contenu dans $\mathcal{R}$,

$$S = S([\mathcal{W}], \mathcal{R}) = \max_{W \in T([\mathcal{W}]),\, W \subseteq \mathcal{R}} \sum_{(i,j):\mathcal{R}_{i,j} \in W} X_{ij}. \tag{2}$$

Notons que la définition donnée en (2) étend celle usuellement considérée pour une fenêtre de scan de forme rectangulaire $\mathcal{W} = [0, m_1] \times [0, m_2]$, avec m_1, m_2 des entiers positifs (voir Glaz et *al.* (2001)),

$$S = S(m_1, m_2, N_1, N_2) = \max_{\substack{1 \leq t \leq N_1 - m_1 + 1 \\ 1 \leq s \leq N_2 - m_2 + 1}} \nu_{ts}, \tag{3}$$

avec $\nu_{ts} = \nu_{ts}(m_1, m_2) = \sum_{i=t}^{t+m_1-1} \sum_{j=s}^{s+m_2-1} X_{ij}$.

La statistique S est utilisée pour tester l'hypothèse nulle $(\mathcal{H}_0)$ assumant que les X_{ij} sont i.i.d. selon une distribution spécifique, contre une hypothèse alternative $(\mathcal{H}_1)$ qui spécifie l'existence d'un cluster d'évènements.

Par exemple, pour le modèle de Bernoulli, l'hypothèse nulle assume que les X_{ij} sont i.i.d., $X_{ij} \sim \mathcal{B}(1,p)$, avec p la probabilité de succès. L'hypothèse alternative présume l'existence d'une sous-région convexe $\mathcal{W}$ de $\mathcal{R}$ de taille fixe, telle que pour tout $(i,j) : \mathcal{R}_{i,j} \in [\mathcal{W}]$, les variables aléatoires X_{ij} sont indépendantes et identiquement distribuées selon une loi de Bernoulli $\mathcal{B}(1,q)$ avec $q > p$ et $q = p$ en dehors. Ensuite, il a été montré (voir Chen et Glaz (1996)) que pour une fenêtre de scan de forme rectangulaire de taille fixe, le test de rapport de vraisemblance généralisé est utilisé pour rejeter la précédente hypothèse nulle lorsque S est plus grande que son quantile d'ordre $1-\alpha$, où α est l'erreur de type 1.

Les statistiques de scan bi-dimensionnelles ont été largement utilisées dans plusieurs champs d'application tels que la cosmologie (Darling (1986)), la fiabilité (Barbour (1996)), l'épidémiologie et la santé publique (Viel (2000)). Etant donné qu'il n'existe pas de formules exactes pour la distribution de probabilité de S, cet élément a motivé les chercheurs pour mettre en place des techniques précises d'approximation. De nos jours, l'étude de la distribution des statistiques de scan est un sujet de recherche actif en statistique et plusieurs approximations et bornes ont été fournies dans la littérature (Chen (1996), Boutsikas (2003), Haiman (2006)).

Selon la définition des statistiques de scan bidimensionnelles discrètes, la forme de la fenêtre de scan est rectangulaire. Cette définition est la plus commune et la plus facile à utiliser parmi d'autres choix possibles. A notre connaissance, aucun travail n'a considéré d'autre forme de fenêtre et une explication possible réside dans la nature discrète des données. Notons que pour les statistiques de scan continues, il existe des travaux considérant différentes formes de fenêtres. En effet, Naus (1965), Loader (1991) ont utilisé des rectangles, Alm (1997), Alm (1998), Anderson (1997) ont considéré des rectangles et des cercles tandis que Alm (1997), Alm (1998) ont considéré des triangles, des ellipses et d'autres formes convexes. Intuitivement, la forme de la fenêtre de scan a une influence sur la distribution des statistiques de scan bidimensionnelles discrètes sous l'hypothèse nulle. A titre d'exemple, considérons une région rectangulaire $\mathcal{R}$ de taille $[0,N_1] \times [0,N_2]$ et posons $\{X_{ij}\}$, $1 \leq i \leq N_1$, $1 \leq j \leq N_2$, un champ aléatoire discret associé á $\mathcal{R}$ tel que les variables aléatoires X_{ij} sont indépendantes et identiquement distribuées selon une loi de Bernoulli $\mathcal{B}(1,0.5)$. Concernant la forme de la fenêtre de scan, considérons deux cas de figure : une fenêtre rectangulaire de taille $m_1 \times m_2 = 1 \times 4$ et une fenêtre carré de taille $m_1 \times m_2 = 2 \times 2$.

Balayons la région de petite taille $\mathcal{R} = [0,4] \times [0,4]$. Dans cette configuration simple, la distribution exacte de la statistique de scan est présentée dans la Table 1. Nous observons des différences entre les deux distributions.

Dans ce travail, nous étudions l'influence de la forme de la fenêtre de scan sur la distribution des statistiques de scan bidimensionnelles discrètes. Nous prenons en considération les cas des formes carrées, rectangulaires et circulaires (cercle discret). La puissance du test basé sur les statistiques de scan bidimensionnelles est également utilisé comme mesure de comparaison entre les balayages avec différentes formes de fenêtres de

Table 1: Distribution sous $\mathcal{H}_0$ de la statistique de scan bidimensionnelle discrète dans le cas de fenêtres de tailles $m_1 \times m_2 = 1 \times 4$ et $m_1 \times m_2 = 2 \times 2$ sur la région $\mathcal{R} = [0,4] \times [0,4]$.

	$\mathbb{P}(S \leq k)$	$\mathbb{P}(S \leq k)$
k	$m_1 = 1,\ m_2 = 4$	$m_1 = m_2 = 2$
0	1.5e-05	1.8e-05
1	0.0095	0.0048
2	0.2234	0.1368
3	0.7725	0.6435
4	1.0000	1.0000

scan et différentes formes de clusters d'évènements. Dans un premier temps, nous adaptons la méthode proposée dans Haiman et Preda (2006) pour approximer la distribution des statistiques de scan bidimensionnelles discrètes pour toute forme convexe de fenêtre et montrons que la forme de la fenêtre a une influence sur la distribution. Dans un second temps, nous définissons le cercle discret au regard d'une fenêtre de scan rectangulaire de même surface. Dans un troisième temps, nous réalisons une étude de simulation. La distribution des statistiques de scan est présentée pour des modèles binomiaux et Poisson en considérant différentes formes de fenêtre de scan. La puissance du test basé sur les statistiques de scan afin de détecter différentes formes de clusters est présentée au travers de simulations et de comparaisons.

Bibliographie

[1] Glaz, J., Naus, J., Wallenstein, S., (2001), Scan statistics, Springer Series in Statistics.
[2] Darling, R. and Waterman, M. S., (1986), Extreme value distribution for the largest cube in a random lattice, SIAM Journal on Applied Mathematics, 46(1),118132.
[3] Barbour, A.,Chryssaphinou,O., and Roos,M.,(1996), Compound poisson ap- proximation in systems reliability, Naval Research Logistics (NRL), 43(2):251264.
[4] Viel, J. F., Arveux, P., Baverel, J., and Cahn, J. Y., (2000), Soft-tissue sarcoma and non-hodgkins lymphoma clusters around a municipal solid waste incinerator with high dioxin emission levels, Am J Epidemiol, 152(1):139.
[5] Chen, J., Glaz, J.,1996, Two-dimensional discrete scan statistics, Statistics and probability letters, 31,59-68.
[6] Boutsikas, M. V. and Koutras, M. V., (2003), Bounds for the distribution of two-dimensional binary scan statistics, Probability in the Engineering and Informational Sciences, 17:509-525.
[7] Haiman, G., Preda, C., (2006), Estimation for the distribution of two dimensional discrete scan statistics, Methodology and Computing in Applied Probability, Vol. 8, No.

3, 373-382.
[8] Naus, J. I., (1965). Clustering of random points in two dimensions, Biometrika, 52(1/2):263 267.
[9] Loader, C. R., (1991). Large-deviation approximations to the distribution of scan statistics, Advances in Applied Probability, pages 751-771.
[10] Alm, S. E., (1997). On the distributions of scan statistics of a two-dimensional poisson process, Advances in Applied Probability, pages 118.
[11] Alm, S. E., (1998). Approximation and simulation of the distributions of scan statistics for poisson processes in higher dimensions, Extremes, 1(1):111126.
[12] Anderson, N. H. and Titterington, D. M., (1997). Some methods for investi- gating spatial clustering, with epidemiological applications, Journal of the Royal Statistical Society : Series A (Statistics in Society), 160(1):87105.