APPROCHE SEMI-PARAMÉTRIQUE POUR L'ESTIMATION DE LA FONCTION DE MASSE DE PROBABILITÉ MULTIVARIÉE

Nawal BELAID [1] , Célestin C. KOKONENDJI [2] & Smail ADJABI [3]

[1] *Université de Bejaia, Unité de Recherche LaMOS, Algérie. belaidnawelro@hotmail.fr*
[2] *Université Bourgogne Franche-Comté, Laboratoire de Mathématiques de Besançon, UMR 6623 CNRS-UFC, 16 route de Gray, 25030 Besançon cedex, France. celestin.kokonendji@univ-fcomte.fr*
[3] *Université de Bejaia, Unité de Recherche LaMOS, Algérie. adjabi@hotmail.com*

Résumé. Dans ce travail, nous proposons un estimateur semi-paramétrique d'une fonction de masse de probabilité multidimensionnelle. Cet estimateur est composé d'une partie paramétrique dirigée par la distribution multivariée de Poisson, et d'une partie non-paramétrique qui est une fonction discrète inconnue de poids à estimer par la méthode du noyau associé discret multivarié. La sélection de la matrice des fenêtres de lissage est effectuée essentiellement par une approche bayésienne. Un modèle de diagnostic est présenté afin d'orienter le choix entre les approches semi-paramétrique, paramétrique et non-paramétrique. Les performances de la méthode proposée sont illustrées sur des données réelles.

Mots-clés. Approche bayésienne, distribution multivariée de Poisson, noyau binomial, validation croisée

Abstract. This work treats the semiparametric estimator of multivariate probability mass function. This estimator is composed by a parametric part directed by a multivariate Poisson distribution, and a nonparametric part which is an unknown weight discrete function to be estimated through discrete associated kernels. The selection of the matrix of bandwidths is carried out essentially by a Bayesian approach. The diagnostic model is discussed to make an appropriate choice between the parametric, semiparametric and nonparametric approaches. Performances of the proposed method are illustrated on real count data.

Keywords. Bayesian approach, binomial kernel, cross-validation, multivariate Poisson distribution

1 Introduction

En multivariés, trouver une distribution multidimensionnelle (paramétrée) adéquate pour modéliser un jeu des données de comptage est encore un challenge. Surtout que les notions essentielles de sur-, equi- et sous-dispersion en multivarié pour ces données ne sont pas encore bien établies dans la littérature ; voir le récent travail de Kokonendji & Puig (2017). Par ailleurs, sans hypothèse spécifiée de loi paramétrée on peut aussi laisser parler les données de comptage à l'aide d'une approche non-paramétrique pour estimer la fonction de masse de probabilité (fmp)

multidimensionnelle ; voir par exemple Belaid et al. (2016ab). Soient $\mathbf{X}_1, \ldots, \mathbf{X}_n$ des vecteurs aléatoires indépendants et identiquement distribués (iid) de fmp multivariée commune inconnue f à estimer sur $\mathbb{T}_d \subseteq \mathbb{N}^d$ pour $d \geq 2$. L'estimateur à noyau associé discret multivarié $\widetilde{f}_n$ de f s'écrit de la forme :

$$\widetilde{f}_n(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} K_{\mathbf{x},H}(\mathbf{X}_i), \qquad \mathbf{x} \in \mathbb{T}_d, \tag{1}$$

où H est la matrice des fenêtres de lissage tels que $H \equiv H_n \to 0_d$ quand $n \to \infty$ (0_d est la matrice carrée nulle d'ordre d) et $K_{\mathbf{x},H}(\cdot)$ est le noyau associé discret multivarié qui est une fmp de support $\mathbb{S}_{\mathbf{x},H} \subseteq \mathbb{Z}^d$ vérifiant les conditions : $\mathbb{S}_{\mathbf{x},H} \supseteq \mathbf{x}$, $\lim\limits_{H\to 0_d} \mathbb{E}(\mathcal{Z}_{\mathbf{x},H}) = \mathbf{x}$, $\lim\limits_{H\to 0_d} Cov(\mathcal{Z}_{\mathbf{x},H}) = 0_d$ ou $\mathbf{Diag}([0,1)^d)$ i.e. d'éléments diagonaux dans $[0,1)$, avec $\mathcal{Z}_{\mathbf{x},H}$ un vecteur aléatoire de loi $K_{\mathbf{x},H}$. Puisqu'il n'est pas encore immédiat d'obtenir un noyau associé discret multivarié, nous utilisons le produit des noyaux associés discrets univariés, $K_{\mathbf{x},H}(\cdot) = \prod_{j=1}^{d} K_{x_j,h_j}(\cdot)$, où K_{x_j,h_j} peuvent être de même famille. En général, le choix du noyau associé est adapté au support de la fonction inconnue à estimer (p.ex. Kokonendji et Senga Kiessé, 2011).

Comme un compromis entre l'estimation non-paramétrique (1) et l'estimation paramétrique, nous proposons l'approche semi-paramétrique pour estimer f. Cette approche a été introduite par Hjort et Glad (1995) dans le cas continu multivarié et utilisée par Kokonendji et al. (2009) dans le cas discret univarié ; ce dernier a été récemment amélioré par Senga Kiessé et al. (2016) essentiellement pour le choix de fenêtre à la bayésienne. Dans ce travail, nous partons du constat que toute distribution d'un vecteur de dénombrement peut s'écrire comme la distribution d'une loi multidimensionnelle de Poisson pondérée par une fonction de poids appropriée, c'est-à-dire :

$$f(\mathbf{x}) = \omega(\mathbf{x}) p(\mathbf{x}, \boldsymbol{\theta}) =: f_{\omega,\boldsymbol{\theta}}(\mathbf{x}), \tag{2}$$

où $p(\mathbf{x}, \boldsymbol{\theta})$ est la partie paramétrique suivant la loi multidimensionnelle de Poisson, et $\omega(\mathbf{x})$ est la fonction poids inconnue à estimer par une méthode de noyau associé. Nous présentons aussi les approches bayésienne locale et validation croisée (globale) pour le choix de la matrice des fenêtres de lissage en utilisant simplement le noyau binomial produit. Enfin, on effectue une illustration sur des données réelles.

2 Méthodologie et résultats

Pour le point de départ paramétrique dans (2), on considére la forme explicite de fmp jointe de distribution multivariée de Poisson (MP) avec une corrélation positive commune pour toutes les paires (X_i, X_j) de $\boldsymbol{X} = (X_1, \ldots, X_d)^\top$:

$$MP(\mathbf{X} = \mathbf{x}|\boldsymbol{\mu}; \sigma_0^2) = \exp\left(-\sigma_0^2 - \sum_{i=1}^{d} \mu_i\right) \sum_{k=0}^{\min(x_1,\ldots,x_d)} \prod_{j=1}^{d} \frac{\mu_j^{x_j}}{x_j} C_k^{x_j} (k!)^{d-1} \left(\frac{\sigma_0^2}{\mu_1 \cdots \mu_d}\right)^j, \tag{3}$$

avec $\sigma_0^2 = \sigma_{ij} = Cov(X_i, X_j) \geq 0$ pour tout $i \neq j$ et $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)^\top$; voir, par exemple, Johnson et al. (1997). Selon la valeur du paramètre σ_0^2, on ne distingue que deux situations

intéressantes : la première est le modèle MP indépendant avec $\sigma_0^2 = 0$, où la distribution (3) est réduite au produit de d fmp de Poisson univariés ; la seconde est le MP ayant la même covariance $\sigma_0^2 > 0$ et dépendant de $d+1$ paramètres. La partie paramétrique $p(\cdot, \boldsymbol{\theta})$ de l'équation (2) est la distribution $MP(\cdot, \boldsymbol{\theta})$ avec $\boldsymbol{\theta} = (\boldsymbol{\mu}, \sigma_0^2)$. Ainsi, toutes les autres situations sont passées dans la pondération $\omega_{\boldsymbol{\theta}}(\mathbf{x}) = \omega(\mathbf{x}, \boldsymbol{\theta})$, y compris $\sigma_{ij} < 0$ pour lesquels $\sigma_0^2 = \sup(0, \sigma_{ij})$.

Nous définissons alors l'estimateur semi-paramétrique de f de (2) via (3) par :

$$\widehat{f}(\mathbf{x}) = MP(\mathbf{x}, \widehat{\boldsymbol{\theta}}_n)\, \widetilde{\omega}(\mathbf{x}, \widehat{\boldsymbol{\theta}}_n) = \frac{1}{n} \sum_{i=1}^{n} \prod_{j=1}^{d} \frac{MP(\mathbf{x}, \widehat{\boldsymbol{\theta}}_n)}{MP(\mathbf{X}_i, \widehat{\boldsymbol{\theta}}_n)} K_{x_j, h_j}(X_{ij}),\ \mathbf{x} \in \mathbb{T}_d, \tag{4}$$

où $\widehat{\boldsymbol{\theta}}_n = (\widehat{\boldsymbol{\mu}}_n, \widehat{\sigma}_0^2)$ avec $\widehat{\boldsymbol{\mu}}_n = \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i = (\overline{X}_1, \ldots, \overline{X}_d)^\top$ et $\widehat{\sigma}_0^2 = \frac{2}{(n-1)d(d-1)} \sum_{j \neq k}^{d} \sum_{i=1}^{n} (X_{ij} - \overline{X}_j)(X_{ik} - \overline{X}_k)$. Entre autre, on montrera que $\widehat{\boldsymbol{\theta}}_n \to \widehat{\boldsymbol{\theta}}_0 := \arg\min_{\boldsymbol{\theta}} \sum_{\mathbf{x} \in \mathbb{T}_d} f(\mathbf{x}) \log[f(\mathbf{x})/MP(\mathbf{x}, \boldsymbol{\theta})]$ quand $n \to +\infty$. Le noyau associé multivarié utilisé est le produit des noyaux binomiaux univariés $B_{x,h}(\cdot)$ suivants la loi binomiale $\mathcal{B}(x, h)$ avec sa fmp $B_{x,h}(y) = \frac{(x+1)!}{y!(x+1-y)!} \left(\frac{x+h}{x+1}\right)^y \left(\frac{1-h}{x+1}\right)^{x+1-y}$. La matrice des fenêtres de lissage est sélectionnée par les deux approches suivantes.

Approche par validation croisée globale

La matrice des fenêtres optimale, notée H_{LSCV}, est donnée par : $\widehat{H}_{LSCV} = \arg\min_{H \in \mathcal{M}} LSCV(H)$, avec

$$\begin{aligned} LSCV(H) \;=\; & \frac{1}{n^2} \sum_{i=1}^{n} \sum_{\ell=1}^{n} \frac{1}{MP(\mathbf{X}_i, \widehat{\boldsymbol{\theta}}_n) MP(\mathbf{X}_\ell, \widehat{\boldsymbol{\theta}}_n)} \sum_{\mathbf{x} \in \mathbb{T}^d} MP^2(\mathbf{x}, \widehat{\boldsymbol{\theta}}_n) \\ & \times \prod_{j=1}^{d} B_{x_j, h_j}(X_{ij}) B_{x_j, h_j}(X_{\ell j}) - \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{\ell \neq i} \prod_{j=1}^{d} B_{X_{ij}, h_j}(X_{\ell j}) \frac{MP(\mathbf{X}_i, \widehat{\boldsymbol{\theta}}_{n,-i})}{MP(\mathbf{X}_\ell, \widehat{\boldsymbol{\theta}}_{n,-i})}, \end{aligned}$$

où $\mathcal{M}$ est l'espace des matrices de lissage symétriques définies positives, et $\widehat{\boldsymbol{\theta}}_{n,-i}$ est calculé à partir de $\widehat{\boldsymbol{\theta}}_n$ sans l'observation $\mathbf{X}_i$.

Approche bayésienne locale

L'estimateur bayésien local de H est défini comme suit :

$$\widehat{H}_{BayesL}(\mathbf{x}) = \int_{\mathcal{M}} H \widehat{\pi}(H | \mathbf{x}, \mathbf{X}_1, \ldots, \mathbf{X}_n) dH,$$

où $\widehat{\pi}(H | \mathbf{x}, \mathbf{X}_1, \ldots, \mathbf{X}_n) = \widehat{f}_n(\mathbf{x}) \pi(H) \left[\int_{\mathcal{M}} \widehat{f}_n(\mathbf{x}) \pi(H) dH \right]^{-1}$ est l'estimateur de la loi a posteriori et $\pi(H)$ est la loi a priori bêta de paramètres positifs α et β. En exploitant la conjugalité,

nous obtenons la forme exacte des éléments diagonaux de l'estimateur de H comme suit :

$$\widehat{h}_j(x_j) = \mathbf{D}_B^{-1} \sum_{i=1}^{n} \frac{MP(\mathbf{x}, \widehat{\boldsymbol{\theta}}_n)}{MP(\mathbf{X}_i, \widehat{\boldsymbol{\theta}}_n)} \left(\sum_{k=0}^{X_{ij}} \mathbf{A}_{ijk} \mathbf{B}(X_{ij} - k + \alpha + 1, x_j - X_{ij} + \beta + 1) \right) \times \left(\prod_{\substack{m=1 \\ m \neq j}}^{d} \sum_{k=0}^{X_{im}} \mathbf{A}_{imk} \mathbf{B}(X_{im} - k + \alpha, x_m - X_{im} + \beta + 1) \right),$$

avec $\mathbf{B}(\alpha, \beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1} dt$ pour $\alpha, \beta > 0$,

$$\mathbf{D}_B = \sum_{i=1}^{n} \frac{MP(\mathbf{x}, \widehat{\boldsymbol{\theta}}_n)}{MP(\mathbf{X}_i, \widehat{\boldsymbol{\theta}}_n)} \prod_{j=1}^{d} \sum_{k=0}^{X_{ij}} \frac{(x_j+1)! x_j^k \mathbf{B}(X_{ij} - k + \alpha, x_j - X_{ij} + \beta + 1)}{(x_j + 1 - X_{ij})! k! (X_{ij} - k)! (x_j + 1)^{x_j + 1}}$$

et

$$\mathbf{A}_{ijk} = \frac{(x_j+1)! x_j^k}{(x_j + 1 - X_{ij})! k! (X_{ij} - k)! (x_j + 1)^{x_j + 1}}.$$

Résultats

Les performances de l'estimateur (4) sont mesurées via l'erreur quadratique intégrée empirique $ISE_0 = \sum_{\mathbf{x} \in \mathbb{N}^d} [\widehat{f}_n(\mathbf{x}) - \widehat{f}_0(\mathbf{x})]^2$, où $\widehat{f}_0$ est l'estimation empirique de f. Le modèle de diagnostic consiste à examiner le graphe de la fonction poids
$\mathbf{x} \mapsto \widetilde{\omega}(\mathbf{x}, \widehat{\boldsymbol{\theta}}_n) = (1/n) \sum_{i=1}^{n} \prod_{j=1}^{d} B_{x_j, h_j}(X_{ij}) / MP(\mathbf{X}_i, \widehat{\boldsymbol{\theta}}_n)$, ou bien la fonction $\mathbf{x} \mapsto Z(\mathbf{x}) := \log \widetilde{\omega}(\mathbf{x}, \widehat{\boldsymbol{\theta}}_n) = \log[\widehat{f}(\mathbf{x}) / MP(\mathbf{x}, \widehat{\boldsymbol{\theta}}_n)]$, avec une bande de confiance ± 1.96 pour vérifier si $\omega(\mathbf{x}) = 1$ est acceptable ou non. En pratique, si le taux des points se trouvant à l'intérieur de la bande est $< 5\%$ alors il est préférable de considérer l'estimation non-paramétrique ; si le taux appartient à $[5\%, 95\%]$ c'est pour l'estimation semi-paramétrique ; et, si le taux est $> 95\%$ alors il faudrait considérer une estimation paramétrique (donc poissonnienne en occurrence).

Dans le but d'illustrer cette méthodologie, nous étudions trois jeux de données réelles pour $d = 2$. Le premier (i) représente le nombre de plantes de type *Lacistema aggregatum* (y_1) et *Protium guianense* (y_2) observées dans la forêt de Trinidad, voir Holgate (1966). Les données (ii) concernent le nombre de céphalophes bleus (y_1) et d'autres petits animaux (y_2) chassés dans un village en Guinée, voir Bonat et al. (2016). Le dernier jeu de données (iii) est relatif au nombre de consultations chez un médecin (y_1) et le nombre de médicaments (y_2) consommés dans cette période ; ces données ont été récoltées dans une enquête sur la santé en Australie, voir Cameron et Trivedi (1998). La Table 1 résume les caractéristiques statistiques de ces données ainsi que la valeur de l'indice généralisé de dispersion $\widehat{GDI}$, donné comme suit :

$$\widehat{GDI} = \frac{\widehat{\sigma}_1^2 \widehat{\mu}_1 + \widehat{\sigma}_2^2 \widehat{\mu}_2 + 2\widehat{\rho}\widehat{\sigma}_1 \widehat{\sigma}_2 \sqrt{\widehat{\mu}_1 \widehat{\mu}_2}}{\widehat{\mu}_1^2 + \widehat{\mu}_2^2}.$$

TABLE 1 – Résumé des statistiques des données.

Données	n	$\widehat{\mu}_1$	$\widehat{\mu}_2$	$\widehat{\sigma}_1^2$	$\widehat{\sigma}_2^2$	$\widehat{\rho}$	$\widehat{GDI}$	$\widehat{\sigma}_{12}$
(i)	100	0.9500	0.6000	1.4217	0.6667	0.1141	1.5195	0.1111
(ii)	1216	4.4498	0.7113	19.3802	2.2927	-0.0432	4.2766	-0.2882
(iii)	5190	0.3017	0.8626	0.6370	2.0032	0.3077	2.7240	0.3476

TABLE 2 – Valeurs des ISE_0 selon la méthode de fenêtres et celles de diagnostic (en $\%$).

Modèles $\rightrightarrows$	Semi-	param.	$\sigma_0^2 = 0$	Semi-	param.	$\sigma_0^2 > 0$	Non-param
Données $\downdownarrows$	LSCV	Bayes	Diagn.	LSCV	Bayes	Diagn.	Bayes
(i)	1.32e-04	4.02e-05	44	1.26e-04	**2.98e-05**	38	5.86e-05
(ii)	5.52e-09	**2.04e-10**	36	–	–	–	7.22e-10
(iii)	6.79e-03	8.51e-06	4	5.12e-03	8.23e-06	8	**2.78e-09**

Note : Le gras est pour les meilleurs résultats ; le cas $\sigma_0^2 > 0$ n'est pas applicable pour les données (ii).

Voir Kokonendji et Puig (2017) qui disent que la distribution des données bivariées de comptage est surdispersée (équidispersée et sousdispersée) si $\widehat{GDI} > 1$ ($=1$ et $\in [0, 1)$, respectivement).

Les résultats obtenus sont illustrés dans la Table 2, qui rapporte les ISE_0 obtenus par les deux approches : semi-paramétrique et non-paramétrique, en utilisant le même noyau binomial produit . Les choix de fenêtres sont réalisés avec la méthode bayésienne locale (Bayes) pour les deux approches et aussi par LSCV pour le semi-paramétrique qui apparaît moins performant que Bayes. En semi-paramétrique, on a considéré deux départs poissonniens MP avec : $\sigma_0^2 = 0$ puis $\sigma_0^2 > 0$; sauf pour les données (ii) où l'on a $\widehat{\sigma}_{12} = -0.2882 < 0$ (Table 1) et, donc, on ne considère que l'approche semi-paramétrique avec $\sigma_0^2 = 0$.

À partir de ces résultats de la Table 2, on obtient des performances (par rapport à ISE_0) qui sont conformes à la nouvelle approche semi-paramétrique et au diagnostic résultant. En effet, les diagnostics (44% et 38%) pour les données (i) suggèrent l'utilisation d'une approche semi-paramétrique avec $\sigma_0^2 > 0$ comme préférence. Pour les données (ii) avec un diagnostic à 36%, on obtient aussi la suggestion d'une modélisation semi-paramétrique mais avec $\sigma_0^2 = 0$, car $\widehat{\sigma}_{12} = -0.2882 < 0$. Quant aux données (iii), les résultats de diagnostics (4% et 8%) montrent qu'il est préférable de laisser parler ces données à travers l'approche non-paramétrique.

3 Conclusion

Dans ce travail, nous avons présenté l'approche semi-paramétrique pour l'estimation des fonctions de masse de probabilités multivariées. En plus d'être simple, facile à mettre en œuvre et efficace, l'estimateur proposé est bien approprié dans le cadre des données multivariées de

comptage pour des échantillons de tailles petites et moyennes ; c'est dans cette situation où des modèles paramétriques font souvent défaut. Aussi, nous soulignons le grand intérêt du modèle de diagnostic pour l'orientation vers la méthode d'estimation adaptée : paramétrique, non-paramétrique et semi-paramétrique. Les résultats numériques basés essentiellement sur le noyau binomial produit et une sélection de fenêtres par la méthode bayésienne locale sont déjà prometteurs et ont vocation à être améliorés très prochainement selon les résultats théoriques en cours d'élaboration.

Bibliographie

[1] Belaid, N. Adjabi, S. Zougab, N & Kokonendji, C. C. (2016a), Bayesian bandwidth selection in discrete multivariate associated kernel estimators for probability mass functions, *Journal of the Korean Statistical Society*, 45, 557–567.

[2] Belaid, N. Adjabi, S. Kokonendji, C. C. & Zougab, N. (2016b), Bayesian local bandwidth selector in multivariate associated kernel estimator for joint probability mass functions, *Journal of Statistical Computation and Simulation*, 86, 3667–3681.

[3] Bonat, W .H. Oliveiro, J. Grande-Vega, M. Farfán, M. A. & Fa, J. E. (2016), Modelling the covariance structure in multivariate count models : Hunting in Bioko Island, arXiv :1608.05428.

[4] Cameron, A. & Trivedi, P. (1998), Regression Analysis of Count Data. Oxford University Press, Oxford.

[5] Johnson, N. L. Kotz, S. & Balakrishnan, N. (1997), *Discrete Multivariate Distributions*, Wiley, New York.

[6] Holgate, P. (1966), Bivariate generalizations of Neyman's type A distribution, *Biometrika*, 53, 241–245.

[7] Hjort, N. L & Glad, I. K. (1995), Nonparametric density estimation with a parametric start, *The Annals of Statistics*, 23, 882–904.

[8] Kokonendji, C. C. & Puig, P. (2017), Fisher dispersion index for multivariate count distributions : a review and a new proposal. Submitted for publication.

[9] Kokonendji, C. C. & Senga Kiessé, T. (2011). Discrete associated kernels method and extensions. *Statistical Methodology*, 8, 497–516.

[10] Kokonendji, C. C. Senga Kiessé, T. & Balakrishnan, N. (2009), Semiparametric estimation for count data through weighted distributions *Journal of Statistical Planning and Inference*, 139, 3625–3638.

[11] Sellers, K. F., Morris, D. S. & Balakrishnan, N. (2016), Bivariate Conway-Maxwell-Poisson distribution : formulation, properties, and inference, *Journal of Multivariate Analysis*, 150, 152–168.

[12] Senga Kiessé, T. Zougab, N. & Kokonendji, C. C. (2016), Bayesian estimation of bandwidth in semiparametric kernel estimation of unknown probability mass and regression functions of count data, *Computational Statistics*, 31 , 189-206.