Estimation semi-paramétrique de distribution des données de dénombrement

Célestin Kokonendji, Tristan Senga Kiessé

Estimation semi-paramétrique de distribution des données de dénombrement

Célestin C. Kokonendji & Tristan Senga Kiessé

Université de Pau et des Pays de l'Adour
Laboratoire de Mathématiques Appliquées - UMR 5142 CNRS
Département STID - IUT des Pays de l'Adour
Avenue de l'Université - 64000 Pau, France
Email : celestin.kokonendji@univ-pau.fr ;
tristan.sengakiesse@univ-pau.fr

RÉSUMÉ : Dans cette communication, nous proposons un estimateur semi-paramétrique de distribution des données de dénombrement sous l'hypothèse de modèle de Poisson pondéré ou binomial pondéré. L'estimation non-paramétrique de la fonction discrète de poids correspondant au modèle est effectuée en utilisant la méthode des noyaux associés discrets. Des modèles de diagnostique sont mis en place pour aider au choix d'une approche purement paramétrique, semi-paramétrique ou non-paramétrique.

ABSTRACT: We propose a semiparametric estimator for count data distribution under the hypothesis of weighted Poisson or binomial model. The nonparametric estimation for the discrete weight function of the model is realized by using discrete associated-kernel method. Some model diagnostics enable us to make an appropriate choice among parametric, semiparametric and nonparametric approaches.

Mots clés : Loi discrète pondérée ; estimateur à noyau associé discret.

Dans ce travail, nous partons du constat que toute distribution des données de dénombrement peut s'écrire comme une loi de Poisson pondérée

$$\begin{aligned} f(x) &= \frac{\omega(x)\ p(x;\mu)}{\sum_{x\in\mathbb{N}} \omega(x) p(x;\mu)} \\ &= \omega(x;\mu) p(x;\mu) \\ &=: f_\omega(x;\mu), \quad x \in \mathbb{N} \end{aligned} \tag{1}$$

(voir Kokonendji *et al.*, 2008). La partie paramétrique $p(x;\mu)$ de (1) est la fonction de masse de probabilité (f.m.p.) de Poisson de paramètre inconnu μ à estimer. La partie non-paramétrique $\omega(x)$ de (1) est une fonction discrète de poids inconnue de telle sorte que la fonction normalisée $\omega(x;\mu) := \omega(x) \left\{\sum_{x\in\mathbb{N}} \omega(x) p(x;\mu)\right\}^{-1}$ soit à estimer de manière non-paramétrique connaissant μ. Dans le cas particulier où la f.m.p. inconnue $f(\cdot) = f_\omega(\cdot;\mu)$ est définie sur un ensemble fini $\{0, 1, \ldots, N\}$ de $\mathbb{N}$, nous considérons $f_\omega(\cdot;\mu)$ comme une loi binomiale pondérée ; voir, par exemple, Johnson *et al.* (2005) [pages 149–150], Chakraborty et Das (2006), et leurs références.

Tout d'abord, nous rappelons que l'estimateur non-paramétrique par noyau associé discret d'une f.m.p. f s'écrit

$$\widetilde{f}_n(x) = \frac{1}{n}\sum_{i=1}^{n} K_{x,h}(X_i), \quad x \in \mathbb{N}, \tag{2}$$

où $X_1, X_2, \ldots, X_n$ est une suite de variables aléatoires i.i.d de f.m.p. $f(x) := \Pr(X_i = x)$, $h = h(n) > 0$ est le paramètre de lissage et $K_{x,h}(\cdot)$ désigne un noyau associé discret (Kokonendji *et al.*, 2007, et Senga Kiessé, 2008). Précisons qu'une fonction noyau associé discret $K_{x,h}(\cdot)$ est elle même une f.m.p. sur le support discret $\aleph_x$ (ne dépendant pas de h et contenant au moins x). En plus, on impose les deux conditions suivantes :

$$\lim_{h\to 0}\mathbb{E}(\mathcal{K}_{x,h}) = x \quad \text{et} \quad \lim_{h\to 0}\mathrm{Var}(\mathcal{K}_{x,h}) = 0,$$

où $\mathcal{K}_{x,h}$ est la v.a. discrète de f.m.p. $K_{x,h}(\cdot)$. Le biais ponctuel de $\widetilde{f}_n$ est donné par

$$\mathrm{biais}\{\widetilde{f}_n(x)\} = f\{\mathbb{E}(\mathcal{K}_{x,h})\} - f(x) + \frac{1}{2}\mathrm{Var}(\mathcal{K}_{x,h})f^{(2)}(x) + o(h), \tag{3}$$

où $f^{(2)}$ est la différence finie du second ordre

$$f^{(2)}(x) = \begin{cases} \{f(x+2) - 2f(x) + f(x-2)\}/4 & \text{si } x \in \mathbb{N}\setminus\{0,1\} \\ \{f(3) - 3f(1) + 2f(0)\}/4 & \text{si } x = 1 \\ \{f(2) - 2f(1) + f(0)\}/2 & \text{si } x = 0. \end{cases} \tag{4}$$

Pour la variance ponctuelle, on a

$$\mathrm{Var}\{\widetilde{f}_n(x)\} = \frac{1}{n}f(x)\{\Pr(\mathcal{K}_{x,h} = x)\}^2 - \frac{1}{n}f^2(x) + R_n(x;h), \tag{5}$$

où le dernier terme

$$\begin{aligned} R_n(x;h) &= \frac{1}{n}\sum_{y\in\aleph_x\setminus\{x\}} f(y)\left\{\Pr(\mathcal{K}_{x,h} = y)\right\}^2 + \frac{1}{n}f^2(x) \\ &\quad -\frac{1}{n}\left[f(x) + \sum_{y\in\aleph_x}\{f(y) - f(x)\}\Pr(\mathcal{K}_{x,h} = y)\right]^2 \end{aligned} \tag{6}$$

tend vers 0 quand $n \to +\infty$ et $h = h(n) \to 0$ sous les conditions d'un noyau associé discret.

Par la suite, la méthode d'estimation non-paramétrique par noyau associé discret s'applique à la fonction discrète de poids $\omega(.;\mu)$ et permet de définir l'estimateur semi-paramétrique $\widehat{f}_n$ de f de la manière suivante :

$$\begin{aligned} \widehat{f}_n(x) &= p(x;\widehat{\mu}_n)\ \widetilde{w}_n(x;\widehat{\mu}_n) \\ &= \frac{1}{n}\sum_{i=1}^{n} K_{x,h}(X_i)\frac{p(x;\widehat{\mu}_n)}{p(X_i;\widehat{\mu}_n)}, \quad x \in \mathbb{N}, \end{aligned}$$

où $\widehat{\mu}_n = \overline{X}_n = n^{-1}(X_1 + \cdots + X_n)$ est la moyenne empirique qui est l'estimateur du maximum de vraisemblance de la moyenne μ de la loi de Poisson.

Dans le cas où la loi de Poisson initiale est connue et fixée $p_0(x) = p(x;\mu_0)$, nous écrivons $f = p_0\omega$ et nous avons le résultat suivant.

Théorème 1 *Pour $n \to +\infty$ et $h = h(n) \to 0$, l'estimateur semi-paramétrique $\widehat{f}_n$ de $f = p_0\omega$ possède le biais et la variance suivants :*

$$biais\{\widehat{f}_n(x)\} = p_0(x)\left[w\{\mathbb{E}(\mathcal{K}_{x,h})\} - \frac{f(x)}{p_0(x)} + \frac{1}{2}\,Var(\mathcal{K}_{x,h})w^{(2)}(x)\right]\{1 + o(1)\} \qquad (7)$$

et

$$Var\big\{\widehat{f}_n(x)\big\} = \frac{1}{n}f(x)\{\Pr(\mathcal{K}_{x,h} = x)\}^2 - \frac{1}{n}f^2(x) + R_n(x;h), \qquad (8)$$

où $w^{(2)}$ est la différence finie d'ordre 2 définie comme dans l'expression (4), $o(1)$ ne dépend pas de n et tend vers 0 quand $h = h(n) \to 0$ et $R_n(x;h)$ est le reste comme en (6).

L'estimateur proposé $\widehat{f}_n$ de f pourrait être encore plus approprié que l'estimateur traditionnel $\widetilde{f}_n$ en (2) à condition que le départ soit adapté. Il est facile de vérifier que la variance $\mathrm{Var}\{\widehat{f}_n(x)\}$ en (8) est égale à $\mathrm{Var}\{\widetilde{f}_n(x)\}$ donnée en (5). Tandis que la différence entre les deux estimateurs $\widehat{f}_n$ et $\widetilde{f}_n$ de f provient de leurs biais. En effet, pour un noyau associé discret donné, la comparaison vient des termes des différences finies d'ordre k apparaissant dans l'expression même de leurs biais ponctuels (3) et (7). Par exemple, on a :

$$f^{(1)} = (p_0w)^{(1)} = p_0w^{(1)} + p_0^{(1)}w \;\lesseqgtr\; p_0w^{(1)}$$

et

$$f^{(2)} = (p_0w)^{(2)} = p_0w^{(2)} + 2p_0^{(1)}w^{(1)} + p_0^{(2)}w \;\lesseqgtr\; p_0w^{(2)},$$

où le symbole $\lesseqgtr$ signifie $\leq$ ou $\geq$.

Dans le cas où la loi de Poisson initiale est inconnue (et donc à estimer), on trouve quasiment le même résultat que le Théorème 1 en considérant $p_0(x) = p(x;\mu_0)$ comme étant la meilleure f.m.p. de Poisson approchant la f.m.p. inconnue f au sens de la distance de Kullback-Leibler.

Cependant, si la distribution des données de dénombrement a pour support $\{0, 1, \ldots, N\}$ avec $N \in \mathbb{N} \setminus \{0\}$ fixé, nous écrivons f comme

$$f(x) = w(x;\theta)\; b(x;\theta) =: f_w(x;\theta), \quad \forall x \in \{0, 1, \ldots, N\},$$

où $b(x;\theta) := N!\{x!(N-x)!\}^{-1}\theta^x(1-\theta)^{N-x}$ est la f.m.p. de binomiale $\mathcal{B}(N,\theta)$, $\theta \in [0,1]$, et $x \mapsto w(x;\theta) := \omega(x) \times \left\{\sum_{x\in\{0,1,\ldots,N\}} \omega(x)b(x;\theta)\right\}^{-1}$ est la fonction poids normalisée de binomiale. L'estimateur semi-paramétrique de f est défini par :

$$\widehat{f}_n(x) \;=\; b(x;\widehat{\theta}_n)\;\widetilde{w}_n(x;\widehat{\theta}_n), \quad x \in \{0, 1, \ldots, N\},$$

où $\widehat{\theta}_n = N^{-1}\overline{X}_n = (nN)^{-1}(X_1 + \cdots + X_n)$ est la proportion de succès dans l'échantillon.

Enfin, l'estimation de la fonction discrète de poids procure des informations utiles pour les modèles de diagnostique. Pour l'exemple d'un départ poissonnien, un simple test graphique se construit en représentant $(x, Z(x))$ avec

$$Z(x) = \frac{\log \widetilde{w}_n(x;\widehat{\mu}_n) + (2n)^{-1}\{p(x;\widehat{\mu}_n)\}^{-1}\Pr(\mathcal{K}_{x,h} = x)}{\left[n^{-1}\{p(x;\widehat{\mu}_n)\}^{-1}\Pr(\mathcal{K}_{x,h} = x)\right]^{1/2}} \rightsquigarrow \mathcal{N}(0,1).$$

Ceci permettra le choix entre les trois approches suivantes : paramétrique, semi-paramétrique et non-paramétrique. Des illustrations faites sur des données et des perspectives seront présentées.

Bibliographie

[1] Chakraborty, S. et Das, K.K. (2006). On some properties of a class of weighted quasi-binomial distributions. *Journal of Statistical Planning and Inference* 136, 156–182.

[2] Johnson, N.L., Kemp, A.W. et Kotz, S. (2005). *Univariate Discrete Distributions* (3rd ed.). John Wiley & Sons, Hoboken, New Jersey.

[3] Kokonendji, C.C., Senga Kiessé, T. et Zocchi, S.S. (2007). Discrete triangular distributions and non-parametric estimation for probability mass function. *Journal of Nonparametric Statistics* 19, 241–254.

[4] Kokonendji, C.C., Mizère, D. et Balakrishnan, N. (2008). Connections of the Poisson weight function to overdispersion and underdispersion. *Journal of Statistical Planning and Inference* 138, 1287–1296.

[5] Senga Kiessé, T. (2008). Approche non-paramétrique par noyaux associés discrets des données de dénombrement. Thèse de Doctorat de Statistique de l'Université de Pau.