Intégration de l'approche bayésienne pour évaluer l'incertitude des estimateurs de régression par noyau associé discret

Tristan Senga Kiessé, Etienne Rivot

Intégration de l'approche bayésienne pour évaluer l'incertitude des estimateurs de régression par noyau associé discret

Tristan SENGA KIESSÉ[1] & Etienne RIVOT[2]

[1] *UMR SAS, INRAE, Institut Agro, Rennes, France, tristan.senga-kiesse@inrae.fr*
[2] *DECOD, Dynamique et durabilité des écosystèmes, Institut Agro, INRAE, Ifremer, Rennes, France, etienne.rivot@institut-agro.fr*

Résumé. L'approche de régression non-paramétrique à noyau discret fournit généralement des estimations de distributions discrètes, qui ne tiennent pas compte de l'incertitude sur les paramètres sous jacents de ces distributions. Dans ce travail, nous avons développé une approche bayésienne pour estimer l'incertitude autour des paramètres d'un modèle de régression discrète par noyau. L'estimation des densités a posteriori des paramètres est réalisée par des simulations de Monte Carlo par Chaînes de Markov à l'aide du logiciel JAGS. Une illustration est proposée sur des données simulées d'un modèle de régression discrète. L'intervalle de crédibilité bayésien des estimations est donné, et les performances de l'estimateur de régression non-paramétrique sont comparées selon différents noyaux associés discrets. La qualité d'ajustement est mesurée en termes de biais et de variance d'estimation. Le noyau associé discret qui permet d'obtenir les estimations les moins baisées ne fournit pas toujours la meilleure précision. La démarche développée dans ce travail permet notamment d'identifier les points sur lesquels il y a une plus grande incertitude d'estimation, et d'évaluer la probabilité de sur- ou sous-estimer au niveau de ces points.

Mots-clés. Distribution a posteriori, estimateur de régression, noyau discret,

Abstract. The nonparametric discrete kernel regression generally provides estimations of discrete distributions, which do not account for uncertainty about underlying parameters of these distributions. We developed a Bayesian approach to quantify the uncertainty of parameters involved in discrete kernel regression model. The estimation of posterior density of parameters was performed by using Markov Chain Monte Carlo simulations by using JAGS software. An illustration was proposed on simulated data from a discrete reression model. Bayesian credible interval of estimates was given, and performances of nonparametric discrete regression estimator were compared as a function of discrete kernels used. The goodness of fit was assessed in terms of accuracy and precision of estimates. Discrete kernels that fitted the simulated distribution well did not still provide the better precision of estimates. That approach allows to identify points for which uncertainty of the estimates was high, and to evaluate the probability to under- or over-estimate at these points.

Keywords. Posterior density, regression estimator, discrete kernel

1 Introduction

L'approche d'estimation à noyau a connu de nombreux développements sur ces dernières années, notamment pour lisser des ditributions de probabilité discrètes (Racine et al., 2020, Kokonendji et Senga Kiessé, 2011). Ces développements sont essentiellements motivés par la nécessité de proposer un estimateur de même nature (c'est-à-dire, discrète) que les distributions des variables auxquelles il est appliqué. Par exemple, O'Neill (2022) a recomandé l'usage d'un estimateur à noyau discret pour estimer les contours de densité pour une distribution de probabilité discrète.

Soit f une fonction de masse de probabilité d'une v.a. X sur un support discret $\mathcal{S}$, qui est inclus dans $\mathbb{Z}$. Les *noyaux associés discrets* ont été introduits pour contribuer à fournir une estimation en un point $x \in \mathcal{S}$ en attribuant la probabilité la plus importante (c'est-à-dire, proche de 1) en ce point, tout en utilisant une fenêtre de lissage $h > 0$ pour prendre en compte les observations dans le voisinage de x. Plus précisement, $K_{x,h}$ est une fonction de masse de probabilité de v.a. $\mathcal{K}_{x,h}$ telle que $\sum_{y \in \mathcal{S}_x} K_{x,h}(y) = 1$, avec $0 \leq K_{x,h}(y) = \Pr(\mathcal{K}_{x,h} = y) \leq 1$. Les noyaux associés discrets $K_{x,h}$ de support $\mathcal{S}_x$ et de v.a. $\mathcal{K}_{x,h}$ construits se distinguent selon que leur probabilité modale se comporte asymptotiquement comme suit quand $h \to 0$:

$$\Pr(\mathcal{K}_{x,h} = x) \to \Pr(\mathcal{D}_x = x) = 1, \tag{1}$$

avec les propriétés suivantes de leur moyenne et variance :

$$\mathbb{E}(\mathcal{K}_{x,h}) \to x \text{ et } \mathrm{Var}(\mathcal{K}_{x,h}) \to 0, \tag{2}$$

qui correspondent asymptotiquement aux propriétés du noyau de type Dirac de v.a. $\mathcal{D}_x$ sur le support $\mathcal{S}_x = \{x\}$ (par exemple, le noyau associé discret Conway-Maxvell-Poisson de Huang et al., 2022).

Les noyaux associés discrets sont utilisés pour construire des estimateurs de fonction de masse de probabilité (Chee, 2017) et de fonction de régression discrète (Abdous et al. 2012). Néanmoins, les estimateurs à noyau discret traditionnel ne fournissent pas d'information qui permettent d'évaluer l'exactitude et la précision des estimations obtenues. Par exemple, il peut être utile d'évaluer la probabilité de sur- ou sous-estimer une observation en fonction du noyau et de la fenêtre de lissage utilisés. De même, il peut être utile de fournir un intervalle de confiance des estimations, dû à de l'incertitude ou de la variabilité dans les données ou les paramètres ; cela diffère des estimations ponctuelles traditionnellement obtenues. Cette information sur l'incertitude des résultats peut être particulièrement plus cruciale quand nous estimons la distribution d'un échantillon de taille relativalement petite, plutôt que celle d'un échantillon de grande taille qui reflète mieux la vraie distribution de la population étudiée.

Pour conduire une analyse d'incertitude sur un estimateur à noyau associé discret, nous étudions l'approche bayésienne qui utilise les informations fournies par les données pour améliorer la connaissance a priori des paramètres (Li et Wang, 2017). Par exemple, les modèles hiérarchiques bayésiens sont utilisés pour prendre en compte la dépendance dans les données longitididunales multivariées aquatiques (Parent et Rivot, 2013). Dans le cadre de l'estimation à noyau associé discret, l'approche bayésienne a été principalement intégrée pour la sélection

de la fenêtre de lissage afin d'améliorer la qualité d'ajustement des observations (par exemple, Zhang et al. 2016 dans le cas de données mixtes).

Nous intégrons l'approche bayésienne dans l'estimateur non-paramétrique de régression à noyau discret de type Nadaraya-Watson. Nous étudions particulièrement les intervalles de crédibilité et la variabilité ponctuelle des estimations en fonction du noyau et de la fenêtre de lissage utilisés. Nous analysons ainsi les performances de l'estimateur en termes de biais et de variance des estimations. Cela nous a permis une comparaison avec les intervalles de confiance d'autres modèles de régression tels que les GAM (*Generalized additive models*) (Wood, 2006).

2 Estimateur de régression à noyau discret

Nous considérons des séquences i.i.d. $(X_i, Y_i)_{i=1,2\cdots n}$ d'une paire de v.a. $(X, Y) \in \mathcal{S}\times\mathbb{R}$ liées par le modèle m tel que $Y_i = m(X_i)+\epsilon_i$, où les résidus ϵ_i satisfont $\mathbb{E}(\epsilon_i) = 0$ et $\mathrm{Var}(\epsilon_i) < +\infty$. L'estimateur de régression non-paramétrique à noyau discret de type Nadaraya-Watson a été développé pour modéliser la fonction de régression discrète $m(\cdot) = \mathbb{E}(Y|X = \cdot)$ tel que

$$\widehat{m}_{K,h}(x) = \sum_{i=1}^{n} \frac{Y_i K_{x,h}(X_i)}{\sum_{i=1}^{n} K_{x,h}(X_i)}, \quad x \in \mathcal{S}, \tag{3}$$

où $K_{x,h}$ est un noyau associé discret de support $\mathcal{S}_x$ et $h = h(n)$ est la fenêtre de lissage telle que $\lim_{n\to\infty} h(n) = 0$. Les propriétés principales des noyaux $K_{x,h}$ sont reflétées à travers la variance et la moyenne de la v.a. $\mathcal{K}_{x,h}$ donnée par :

$$\mathbb{E}(\mathcal{K}_{x,h}) = x + a(x, h) \ \text{ et } \mathrm{Var}(\mathcal{K}_{x,h}) = b(x, h). \tag{4}$$

Noyaux associés discrets. On distingue les noyaux discrets associés de second ordre dont $a(x, h)$ et $b(x, h)$ tendent vers 0 quand h tend vers 0, et qui satisfont l'équation (2). Par exemple, nous pouvons citer les noyaux discrets triangulaires généralisés (Kokonendji et Zocchi, 2010) et le noyau discret Conway-Maxwell-Poisson (Huang et al., 2022). De plus, nous avons des noyaux discrets associés de premier ordre qui ne satisfont que la première condition sur la moyenne dans l'équation (2) ; pour la variance de ces noyaux, il est uniquement requis qu'elle soit dans un voisinage de 0 (par exemple, les noyaux Poisson, binomial et binomial négatif cités dans Kokonendji et Senga Kiessé, 2011).

Pour $(x, p) \in \mathcal{S}_x \times \mathbb{N}$ et $h > 0$, nous pouvons établir les comparaisons suivantes entre la probabilité modale et la variance des noyaux associés discrets triangulaires $T_{p;x,h}$ (de second ordre) et les noyaux associés discrets standards $K_{x,h}$ (de premier ordre) quand h tend vers 0 :

$$\Pr(\mathcal{T}_{p;x,h} = x) \geq \Pr(\mathcal{K}_{x,h} = x) \text{ et } \mathrm{Var}(\mathcal{T}_{p;x,h}) \leq \mathrm{Var}(\mathcal{K}_{x,h}).$$

Des relations similaires peuvent être établies en ne comparant que les noyaux discrets standards entre eux, et aussi les noyaux associés discrets triangulaires entre eux en fonction du paramètre p. Dans le paragraphe ci-dessous nous montrons que la probabilité modale et la variance des noyaux influencent les performances des estimateurs correspondants.

Propriétés de l'estimateur. Les propriétés de l'estimateur non-paramétrique de régression à noyau discret (par exemple, biais, variance, erreur quadratiques ponctuelles et globales, convergences) sont essentiellement présentées dans la littérature en utilisant les noyaux discrets associées de second ordre ayant pour moyenne $\mathbb{E}(\mathcal{K}_{x,h}) = x$ (c'est-à-dire $a(x,h) = 0$), comme dans les travaux de Abdous et al. (2012). En adaptant les résultats des travaux précédents, nous présentons une expression plus générale de ces propriétés, en particulier le biais de $\widehat{m}_{K,h}$ tel que :

$$\text{Biais}\{\widehat{m}_{K,h}(x)\} = a(x,h)m^{(1)}(x) + \frac{1}{2}\Big\{a^2(x,h) + b(x,h)\Big\}\Big\{m^{(2)}(x) + 2m^{(1)}(x)\frac{f^{(1)}(x)}{f(x)}\Big\} + o(h^2) + O\big(\frac{1}{n}\big),$$

où $f > 0$ est la fonction de masse de probabilité de X et $m^{(1)}$, $m^{(2)}$ et $f^{(1)}$ sont les différences finies de m et f. L' expression de la variance de $\widehat{m}_{K,h}$ est donnée par

$$\text{Var}\{\widehat{m}_{K,h}(x)\} = \frac{1}{nf(x)}\text{Var}(Y|X = x)\{\Pr(\mathcal{K}_{x,h} = x)\}^2 + \text{R}_n,$$

où R_n est une quantité qui dépend de n. De manière directe, le biais et la variance conduisent à l'erreur quadratique moyenne intégrée (*mean integrated squared error*, MISE) :

$$\text{MISE}\{\widehat{m}_{K,h}(x)\} = \sum_{x\in\mathcal{S}} \text{Biais}^2\{\widehat{m}_{K,h}(x)\} + \sum_{x\in\mathcal{S}} \text{Var}\{\widehat{m}_{K,h}(x)\}. \tag{5}$$

Les propriétés de $\widehat{m}_{K,h}$ dépendent de la probabilité modale et de la variance des noyaux discrets utilisés. Par exemple, le MISE de $\widehat{m}_{K,h}$ tend vers 0 quand on considère les noyaux associés discrets de second ordre pour lesquelles, entre autres, $\text{R}_n \to 0$ quand $h \to 0$ et $n \to +\infty$. De même, on peut établir d'autres propriétés asymptotiques $\widehat{m}_{K,h}$ en utilisant les noyaux associés discrets de second ordre, telles que la convergence presque-sure

$$\widehat{m}_{K,h}(x) \xrightarrow{p.s.} m(x)$$

et la normalité asymptotique

$$\sqrt{n}\{\widehat{m}_{K,h}(x) - m(x)\} \xrightarrow{d} \mathcal{N}(\mu, \Lambda),$$

avec $\mu = 0$ et $\Lambda = \sigma^2\{\Pr(\mathcal{K}_{x,h} = x)\}^2/f(x)$, quand $n \to +\infty$. L'estimateur utilisant les noyaux associés discrets de premier ordre ne vérifie pas de telles propriétés asymptotiques mais cependant nous illustrons dans les résultats que ces noyaux peuvent ponctuellement fournir des résultats compétitifs pour des échantillons de tailles relativement petites.

3 Inférence bayésienne

Nous considérons les paires de réalisations $(x_i, y_i)_{i=1,2\cdots n}$ liées par le modèle $y_i = m(x_i) + e_i$, où e_i sont les résidus de distribution Gaussienne, avec une moyenne nulle et une variance

finie σ^2. La fonction de vraisemblance de $y_1, y_2, \ldots, y_n$ étant donné les paramètres h et σ^2 de l'estimateur $\widehat{m}_{K,h}$ est donnée par :

$$\mathrm{L}(y_1, y_2, \ldots, y_n | h, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\{y_i - \widehat{m}_{K,h}(x_i)\}^2\right]. \tag{6}$$

Dans l'analyse de l'estimateur $\widehat{m}_{K,h}$ en intégrant l'approche bayésienne, nous considérons les paramètres h et σ^2 comme des variables aléatoires ayant des fonctions de densité de probabilité a priori. La fonction de densité de probabilité a posteriori jointe de h et σ^2, en tenant en compte l'information apportée par les observations, s'exprime par :

$$\pi(h, \sigma^2 | y_1, y_2, \ldots, y_n|) \propto \pi(h)\pi(\sigma^2)\mathrm{L}(y_1, y_2, \ldots, y_n | h, \sigma^2). \tag{7}$$

Dans un souci de simplification, nous considérons que le paramètre h suit une loi beta $\mathcal{B}(\alpha, \beta)$ de paramètres $\alpha = \beta = 1$, qui correspond a une loi a priori uniforme $\pi(h)$ non-informative. De plus, nous considérons un paramètre de précision $\tau = 1/\sigma^2$ de loi a priori $\mathcal{G}(g_1, g_2)$ $g_1 = g_2 = 0.001$, qui est un conjugué naturel de la variance dans la vraisemblance du modèle Gaussien. L'intégration de l'approche d'échantillonnage bayésienne avec l'estimateur non-paramétrique de régression s'est faite sous le logiciel "OpenBUGS" qui permet d'implémenter les modèles bayésiens de manière relativement simple, en utilisant le package R "BRugs" comme interface (Marley et Wand, 2010). Nous utilisons aussi le package R "HRW" pour résumer les résultats de l'inférence bayésienne pour les paramètres h et la variance des résidus σ^2 (Harezlak et al. 2021). La distribution a posteriori des paramètres est estimée par des simulations de Monte Carlo par Chaînes de Markov (MCMC) avec 5000 iterations.

4 Résultats

Nous avons illustré l'estimateur non-paramétrique de régression à noyau discret qui intègre l'approche d'échantillonnage bayésienne sur des données $(x_i, y_i)_{i=1,2\ldots n} \in \mathbb{N}\times\mathbb{R}$, $n = \{10, 20, 50\}$, simulées à partir d'un modèle discret $m(x) = 2^x/x!$ tel que $y_i = m(x_i) + e_i$. Nous avons utilisé des noyaux associés de premier ordre (Poisson, binomial) et second ordre (symmétrique triangulaire). Nous avons aussi appliqué un modèle additif généralisé basé les fonctions splines.

L'échantillonnage des paramètres h and σ^2 par les méthodes MCMC n'indiquaient pas d'anomalies (Figure 1). Plus particlulièrement, les fonctions de densité a posteriori des paramètres avaient des distributions différentes en fonction du noyau utilisé. Par exemple, la densité a posteriori était asymétrique de valeurs moyennes 0.293 et 0.064, en considérant les noyaux Poisson et binomial, respectivement. En comparaison, la densité a posteriori de h était symétrique de moyenne 0.551, en considérant le noyau discret symétrique triangulaire. De plus, la densité a posteriori de σ^2 avait une distribution symmétrique pour les trois noyaux discrets utilisés. Les intervalles de crébilité bayésiens de h et σ^2 étaient plus larges en considérant le noyau Poisson et plus petits en considérant le noyau discret symétrique triangulaire.

Lorsque l'on compare les intervalles de crédibilité bayésien des estimations, peu de points simulés étaient à l'intérieur de ces intervalles lorsque l'on considérait les noyaux Poisson et

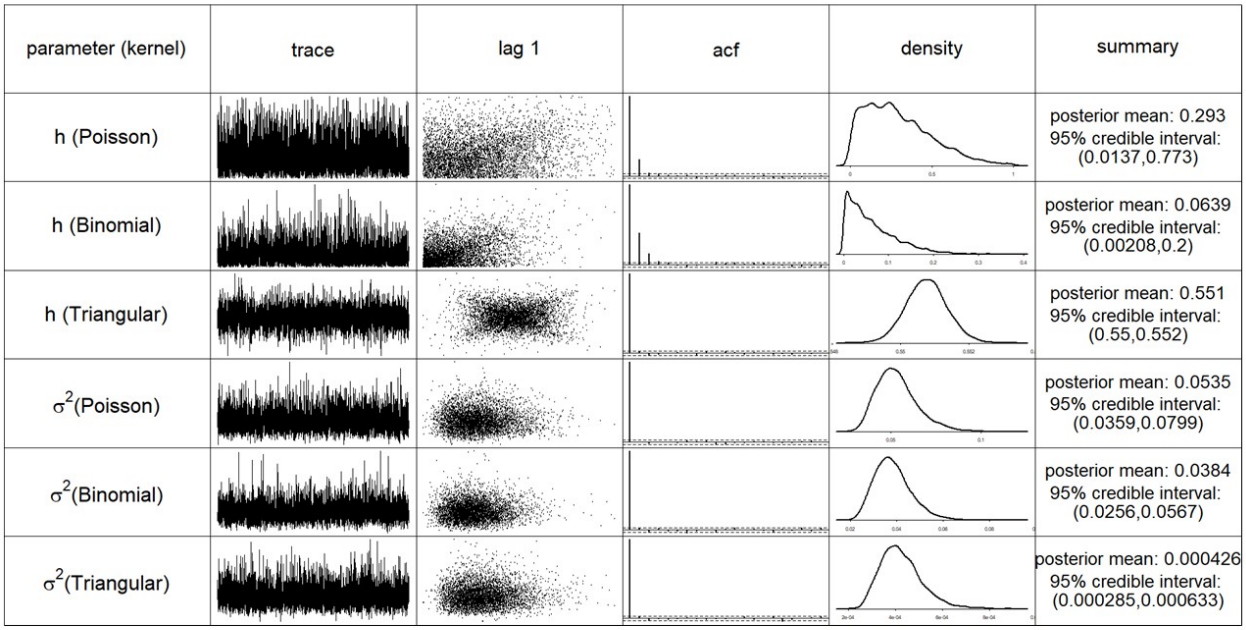

FIGURE 1 – Résultats de l'inférence bayésienne pour les paramètres h et σ^2 de l'estimateur non-paramétrique discret de régression appliqué sur les données simulées (de taille $n = 50$)

binomial (Figure 2 (a)). L'estimateur non-paramétrique utilisant ces deux noyaux discrets n'ajustait pas bien la distribution des données simulées, bien que les intervalles de crédillité bayésiens étaient plus petits en utilisant le noyau binomial (Figure 2 (b)). À l'inverse, les intervalles de crédibilité des estimations contenaient tous les points lorsque l'on considérait l'estimateur non-paramétrique utilisant le noyau discret symmétrique triangulaire. L'estimateur utilisant ce noyau fournissait aussi des estimations plus proches des données simulées. Cependant, les intervalles de crédibilité des estimations pouvaient être plus petits en utilisant le noyau binomial.

Les intervalles de confiance du GAM était généralement plus larges que les intervalles de crédibilité obtenus par l'estimateur non-paramétrique à noyau discret. Les intervalles de ces deux types de modèles se superposaient notamment quand on utilisait les noyaux discrets binomial et symétrique triangulaire.

Ce travail propose un nouvel angle d'étude de l'estimation à noyau discret en intégrant l'approche bayésienne, pour permettre notamment d'identifier les points pour lesquelles il y a une plus grande incertitude d'estimation et d'évaluer la probabilité de sur- ou sous-estimer en ces points.

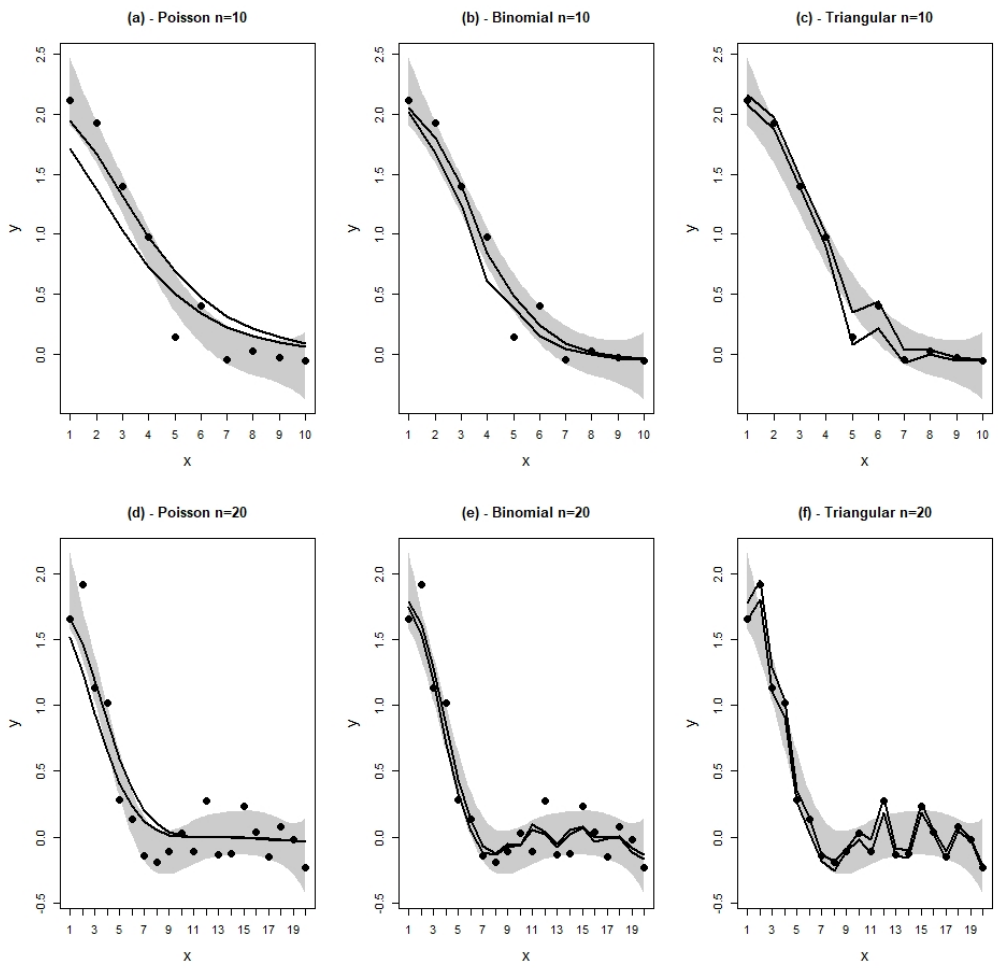

FIGURE 2 – Intervalles bayésiens de crédibilité des données simulées de taille $n = \{10, 20\}$ (points noirs) en utilisant en utilisant l'estimateur non-paramétrique discret de régression avec les noyaux (a,d) Poisson, (b,e) binomial et (c,f) symétriques triangulaires. Les zones grisées correspondent aux intervalles de confiance d'un modèle additif généralisé

Bibliographie

Abdous, C, Kokonendji, C.C. et Senga Kiessé, T. (2012). On semiparametric regression for count explanatory variables. *Journal of Statistical Planning and Inference*, 142, pp. 1537–1548.

Chee, C-S. (2017). A mixture model-based nonparametric approach to estimating a count distribution. *Computational Statistics & Data Analysis*, 109, pp. 34-44.

Dunson, D.B. (2009). Bayesian nonparametric hierarchical modeling. *Biometrical Journal*, 51, pp. 273–284.

Harezlak, J., Ruppert, D. et wand, M.P. (2021). HRW : Datasets, functions and scripts for semiparametric regression. R package version 1.0-5.

Huang, A, Sippel, L et Fung, T. (2022). Consistent second-order discrete kernel smoothing using dispersed conway–maxwell–poisson kernels. *Computational Statistics*, 37, pp. 551–563.

Kokonendji, C.C. et Senga Kiessé, T. (2011). Discrete associated kernel method and extensions. *Statistical Methodology*, 8, pp. 497–516.

Kokonendji, C.C., Senga Kiessé, T. et Demétrio, C.G.B. (2009). Appropriate kernel regression on a count explanatory variable and applications. *Advances and Applications in Statistics*, 12, pp. 99–125.

Kokonendji, C.C. et Somé, S.M. (2021). Bayesian bandwidths in semiparametric modelling for nonnegative orthant data with diagnostics. *Stats*, 4, pp. 162–183.

Kokonendji, C.C. et Zocchi, S.S. (2010). Extensions of discrete triangular distribution and boundary bias in kernel estimation for discrete functions. *Statistical and Probability Letters*, 80, pp. 1655–1662.

Li, S. et Wang, M (2017). Bayesian estimation of the generalized lognormal distribution using objective priors. *Journal of Statistical Computation and Simulation*, 87, pp. 1323–1341.

Marley, J.K. et Wand, M.P. (2010). Non-standard semiparametric regression via brugs. *Journal of Statistical Software*, 37, pp. 1-30.

O'Neill, B. (2022). Smallest covering regions and highest density regions for discrete distributions. *Computational Statistics*, pp. 1–26.

Parent, E. et Rivot, E (2013). *Introduction to Hierarchical Bayesian Modelling for Ecological Data*. Chapman/CRC.

Racine, S. Li, Q. et Yan, K.X. (2020). Kernel smoothed probability mass functions for ordered datatypes. *Journal of Nonparametric Statistics*, 32, pp. 563–586.

Senga Kiessé, T., Zougab, N. et 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, pp. 189–206.

Wood, S.N. (2006). *Generalized additive models : an introduction with R*. Chapman and hall/CRC.

Zhang, X., King, M.L. et Shang, H.L. (2016). Bayesian bandwidth selection for a nonparametric regression model with mixed types of regressors. *Econometrics*, 4, pp. 24.