Estimateur à noyau discret standard pour une densité de probabilité discrète

Célestin Kokonendji, Tristan Senga Kiessé

Estimateur à noyau discret standard pour une densité de probabilité discrète*

Célestin C. Kokonendji, Tristan Senga Kiessé

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

RÉSUMÉ

Dans cet article, nous introduisons un estimateur à noyau discret pour lequel le paramètre de lissage discret est adapté à la cible pour un type de noyau discret donné. Nous démontrons des propriétés fondamentales pour cet estimateur en utilisant les différences finies à la place des dérivées. Le biais, la variance et le risque quadratique intégré de l'estimateur sont, en particulier, examinés sous des considérations générales. Quelques aspects des résultats obtenus sont illustrés à travers les types de noyaux standards de Poisson ou équidispersé, binomial ou sousdispersé, et binomial négatif ou surdispersé. Un critère d'efficacité relative de ces types de noyaux est proposé. Des choix de fenêtres de lissage discret sont étudiés, notamment par validation croisée et par une autre méthode en cas d'excès de zéros. Leurs performances sont enfin comparées à l'aide des données simulées et réelles. Pour des échantillons de petites et moyennes tailles, certains estimateurs à noyau discret (sousdispersé) sont plus performants que l'estimateur naïf ou limite.

Mots-clés : *Différence finie, estimation non-paramétrique, excès de zéros, loi discrète, risque quadratique intégré, validation croisée.*

ABSTRACT

In this paper we introduce a discrete kernel estimator for which the discrete smoothing parameter is adapted to the target for a given type of discrete kernel. We show some basical properties of this estimator using finite difference approximation instead of the derivation. In particular, bias, variance and mean integrated squared error are investigated under general considerations. Some aspects of obtained results are illustrated through the standard discrete kernels like Poisson or equidispersed, binomial or underdispersed, and negative binomial or overdispersed. We provide a simple measure of the relative efficiency between two discrete kernels. Choices of discrete smoothing bandwidths are studied according to several usual criteria as the cross-validation method and a method in the situation of zeros excess. Their performances are finally compared by using simulated and real count data. For a sample size not so large, certains discrete kernel estimators are more efficient than the empirical estimator which is the limit one.

Key words: *Cross-validation, discrete distribution, excess of zeros, finite difference, mean integrated squared error, nonparametric estimation.*

*Prépublication du LMA, Rapport Technique No. 0632 révisée le 2007.06.22
Email: celestin.kokonendji@univ-pau.fr (Célestin C. Kokonendji), tristan.sengakiesse@univ-pau.fr (Tristan Senga Kiessé)

1 Introduction

Soit $X_1, \cdots, X_n$ un n-échantillon aléatoire i.i.d. de densité de probabilité inconnue f sur $\Re$. Un estimateur à noyau continu de f peut être défini par

$$\widehat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \tag{1}$$

$$= \frac{1}{n} \sum_{i=1}^{n} K_{x,h}\left(X_i\right), \quad x \in \Re, \tag{2}$$

où K est la fonction noyau continu (*i.e.* $K(t) \geq 0$ et $\int K(t)dt = 1$), $h > 0$ est le paramètre de lissage (ou fenêtre) et $K_{x,h}$ sera dit alors "noyau associé" de cible x et de fenêtre h. D'après (1), le noyau associé du cas continu s'écrit $K_{x,h}(\cdot) = (1/h)K((x - \cdot)/h)$. Cette écriture (1) est la plus connue depuis les travaux de Rosenblatt (1956), puis de Parzen (1962). Pour de récentes références en français, on peut consulter Berlinet & Biau (2002) et Tsybakov (2004) ; sinon, on a Devroye (1987), Scott (1992) et Silverman (1986) pour des généralités sur des données (supposées) continues, Ferraty & Vieu (2006) pour des données fonctionnelles, Simonoff (1996) et Simonoff & Tutz (2000) pour des données catégorielles ordonnées et discrètes utilisant *toujours* les noyaux continus. Quant à la seconde notation (2), elle est due à Chen (1999, 2000) dans le but d'adapter la fenêtre à la cible pour un "type de noyau" donné tels que beta et gamma généralement asymétrique (voir aussi Scaillet, 2004).

L'estimateur à noyau continu (1) a été développé primairement pour les densités à supports continus et non-bornés (*e.g.* $\Re = \mathbb{R}^d$). La fonction noyau K est classiquement symétrique (*i.e.* $K(-x) = K(x)$) et est considérée comme moins importante que le paramètre de lissage h. Bien qu'un noyau symétrique soit approprié pour ajuster des densités à supports non-bornés, il ne l'est pas pour des densités à supports compacts ou bornés d'un côté (voir Chen, 1999, 2000) et, a fortiori, à supports discrets qu'on désigne par $\aleph$ (*e.g.* $\mathbb{Z}^d$, $\{0, 1, \cdots, N\}^d$ et $\mathbb{N} + p\mathbb{N}$ pour $p \geq 0$) ; d'où la seconde écriture (2). Le cas des noyaux discrets est encore inexploré. On peut noter qu'une première tentative, uniquement de manière expérimentale, a été proposée par Marsh & Mukhopadhyay (1999), en économie où l'on rencontre aussi de nombreuses données de comptage univariées (*i.e.* $\aleph = \mathbb{N}$). Cependant, la définition d'estimateur à noyau discret poissonnien de Marsh & Mukhopadhyay (1999) ne permet pas une étude théorique des propriétés (voir notre Exemple 2.1).

Les perspectives qui motivent ce travail sont principalement de deux ordres. La première est le passage de l'approche paramétrique (quasi systématique) des données de comptage ou dénombrement à une approche non-paramétrique laquelle peut-être justifiée facilement pour des échantillons de petites et moyennes tailles. En fait, l'estimateur empirique ou naïf de densité dans le cas discret est généralement suffisant pour des échantillons de plus grande taille (voir Figure 6). La seconde qui est à l'origine immédiate du travail est une estimation semi-paramétrique de la loi de Poisson pondérée de la forme

$$p_\omega(x, \theta) = \omega(x) p^*(x; \theta), \quad x \in \mathbb{N},$$

où $\omega(.)$ est une fonction discrète (Kokonendji *et al.*, 2007b). Aussi, des extensions dans le cas multidimensionnel des estimateurs à noyau discret univarié (avec ou sans combinaison avec les noyaux continus) seront finalement envisageables pour "lisser" les données (clairsemées) multivariées telles que les tableaux de contingence (*e.g.* Simonoff, 1996).

De manière précise, pour de nombreuses données de comptage observées, il est commun d'avoir la variance qui est égale, plus petite, ou plus grande que la moyenne. Ce qui correspond aux phénomènes d'équidispersion, de sousdispersion ou de surdispersion respectivement. L'approche traditionnelle de ces données commence par une

structure spécifique de distribution paramétrique telle que la loi de Poisson qui est équidispersée. Par la suite et selon le phénomène étudié, il a été question de modifier la loi initiale (*e.g.* par pondération, mixturisation, lagrangénisation) en une autre famille de lois de même support. Voir depuis Greenwood & Yule (1920) jusqu'à Johnson *et al.* (2005), ainsi que Kokonendji *et al.* (2007, 2008) et Mizère *et al.* (2006) pour de récents modèles paramétriques et analyses des données de comptage.

Cependant, si aucune information n'est disponible sur le processus sous-jacent aux données (de tailles moins importantes), l'approche non-paramétrique est la plus appropriée pour un traitement statistique, même dans le cas discret. Cette approche est complémentaire au cas paramétrique. De manière plus naturelle, elle nécessite un estimateur à noyau discret pour estimer une distribution discrète. En effet, une distribution de probabilité discrète se représente par un diagramme à bâtons et non par une courbe. Ainsi, il va de soi de la "lisser" de manière discrète par des estimations équivalentes à des bâtons. L'approche non-paramétrique doit permettre entre autre de détecter une multimodalité et autres phénomènes tels une variation brutale, une absence d'observations ou de données clairsemées. Similairement au cas continu, l'objectif de ce travail est de fournir un premier cadre approprié d'étude des estimateurs à noyau discret. En particulier, les différences finies prendront la place des dérivées du cas continu. Très utile aussi et simple dans la mise en oeuvre, cet estimateur n'est donc qu'une introduction aux fondements des estimateurs non-paramétriques des données de comptage. Nous insisterons également sur la partie pratique de choix de fenêtre et de l'importance des noyaux discrets associés (2).

Cet article est organisé comme suit. La section 2 définit un "noyau discret associé", un estimateur à noyau discret $\widehat{f}$ d'une fonction de probabilité discrète f et présente des exemples ainsi qu'un premier résultat de convergence ponctuelle. On y étudie ponctuellement les comportements asymptotiques exacts du biais, de la variance et, par conséquent, du risque quadratique de $\widehat{f}$. Puis, on y donne les propriétés globales de $\widehat{f}$. La section 3 illustre théoriquement certains aspects de ces estimateurs à l'aide des noyaux discrets standards de Poisson, binomial et binomial négatif, lesquels sont des prototypes de modèles équidispersé, sousdispersé et surdispersé, respectivement (voir, par exemple, Mizère *et al.*, 2006). Dans la section 4, nous établissons une mesure relative de performance de type des noyaux discrets standards. La section 5 propose certains critères de choix de fenêtres de lissage discret. Les sections 6 et 7 présentent des illustrations de ces estimateurs pour le lissage discret des données de comptage simulées et réelles. Des comparaisons sont faites entre ces trois types de noyaux discrets standards ainsi qu'entre les différentes fenêtres associées. Tous les résultats numériques et graphiques sont effectués à l'aide du logiciel R[1].

2 Estimateur à noyau discret

Sans perte de généralité, nous travaillons essentiellement sur un ensemble discret quelconque $\aleph$ inclus dans $\mathbb{R}$ et muni de la mesure de dénombrement μ. Autrement dit, pour toute fonction mesurable g sur $\aleph$, on écrit simplement

$$\int_{\aleph} g(x)\mu(dx) = \sum_{x\in\aleph} g(x).$$

L'estimateur à noyau discret est défini de manière analogue à la relation (2) (Chen 1999, 2000). Dans la suite, nous donnons une définition du noyau discret associé, de l'estimateur à noyau discret et, nous montrons certaines propriétés ponctuelles et globales.

[1] R Development Core Team, 2007, *A Language and Environment for Statistical Computing*. Vienna - Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0, URL http://www.r-project.org.

2.1 Noyau discret associé

Toute loi discrète de probabilité n'est pas potentiellement un noyau discret associé (2). Les lois de Poisson $\mathcal{P}(\lambda)$, binomiale $\mathcal{B}(N,p)$, binomiale négative $\mathcal{BN}(\lambda,p)$ et uniforme discrète $\mathcal{U}(c,a)$ sont des exemples de lois de probabilités discrètes sur $\mathbb{N}$, où les paramètres sont $\lambda > 0, p \in [0,1]$, $N, c, a \in \mathbb{N}$ (*e.g.* Johnson *et al.*, 2005). Ces paramètres vont servir dans la définition intrinsèque du noyau discret associé pour le lissage discret des données tant en position qu'en échelle. Pour simplifier ici, on suppose que le support de f est $\aleph = \mathbb{N}$.

Définition 2.1 *Soit $x \in \mathbb{N}$ et $h > 0$. Etant donnée une loi de probabilité discrète paramétrique K_θ, $\theta \in \Theta \subset \mathbb{R}^d$, de support $\aleph_\theta \subseteq \mathbb{N}$. On appelle "noyau discret associé" $K_{x,h}$ à K_θ, de cible x et de paramètre de lissage discret h, s'il existe une correspondance entre θ et (x,h) telle que $K_{x,h}$ est une loi de probabilité sur le support $\aleph_{x,h}$ de même famille que K_θ :*

$$K_{x,h} \geq 0 \quad \text{et} \quad \sum_{y \in \aleph_{x,h}} K_{x,h}(y) = 1 \tag{3}$$

et que

$$\sum_{y \in \aleph_{x,h}} y K_{x,h}(y) \sim x \quad \text{lorsque} \;\; h \to 0. \tag{4}$$

Par la suite, on appellera $K \equiv K_\theta$ de la Définition 2.1 le "type de noyau discret" pour différencier à la notion classique de "noyau" dans (1). Ainsi, le choix du noyau (discret) associé sera d'autant plus important que celui de la fenêtre.

REMARQUE 2.1 : La relation (4) permet d'assurer la convergence ponctuelle de l'estimateur à noyau discret (cf. Proposition 4). Elle traduit la prise en compte d'un maximum d'information autour de la cible et dans son entourage immédiat, de telle sorte qu'asymptotiquement nous retrouvons l'estimateur naïf. On peut remplacer (4) par

$$\sum_{y \in \aleph_{x,h}} y K_{x,h}(y) = x + h + o(h). \tag{5}$$

La condition (4) (ou (5)) est fondamentale et met en évidence que l'estimateur à noyau discret défini dans la suite est à noyau variable. Cela nous autorise aussi une plus grande flexibilité dans la construction des différents noyaux discrets associés à une loi de probabilité discrète K. Ce qui est implicitement utilisé par Chen (1999, 2000) puis Scaillet (2004). Ainsi, tous les noyaux discrets associés vérifiant la relation (4) ont en commun d'avoir une forme qui s'adapte selon la valeur de la cible x où ils sont calculés. La qualité de lissage obtenue en appliquant ces noyaux discrets associés (ou noyau à forme variable) change selon le comportement de leur variance par rapport à la cible x. Ce qui nous amène à distinguer trois types de noyaux discrets associés : sousdispersés (var$[\mathcal{K}_{x,h}] < \mathbb{E}(\mathcal{K}_{x,h})$), equidispersés (var$[\mathcal{K}_{x,h}] = \mathbb{E}(\mathcal{K}_{x,h})$) et surdispersés (var$[\mathcal{K}_{x,h}] > \mathbb{E}(\mathcal{K}_{x,h})$).

Pour fixer les idées, on propose trois exemples de noyaux discrets associés asymétriques standards sur $\mathbb{N}$ tels qu'ont ait exactement

$$\sum_{y \in \aleph_{x,h}} y K_{x,h}(y) = x + h. \tag{6}$$

Notons que pour ces noyaux discrets, la cible x n'est pas la moyenne mais plutôt le mode.

EXEMPLE 2.1 : Pour un type de noyau poissonnien $\mathcal{P}(\lambda)$, on considère le noyau discret associé $P_{x,h}$ de loi $\mathcal{P}(x+h)$ sur $\aleph_{x,h} = \mathbb{N}$ avec $x \in \mathbb{N}$ et $h > 0$, tels que :

$$P_{x,h}(y) = \frac{(x+h)^y e^{-(x+h)}}{y!}, \quad y \in \mathbb{N}. \tag{7}$$

On signale que le noyau discret associé proposé dans Marsh & Mukhopadhyay (1999) inter-échange x en y dans (7). Notons qu'en une cible $x \in \mathbb{N}$ et pour tout $h > 0$, le noyau associé $P_{x,h}$ est de support $\mathbb{N}$, *équidispersé* de moyenne égale à la variance $x+h$, et de mode compris entre $x+h-1$ et $x+h$.

EXEMPLE 2.2 : Si on considère un type de noyau binomial $\mathcal{B}(N,p)$, on lui associe $B_{x,h}$ de loi $\mathcal{B}(x+1,(x+h)/(x+1))$ sur $\aleph_{x,h} = \{0,1,\cdots,x+1\}$ pour tout $x \in \mathbb{N}$ et $h \in]0,1]$ avec $\cup_x \aleph_{x,h} = \mathbb{N}$, de telle sorte :

$$B_{x,h}(y) = \frac{(x+1)!}{y!(x+1-y)!}\left(\frac{x+h}{x+1}\right)^y \left(\frac{1-h}{x+1}\right)^{x+1-y}, \quad y \in \mathbb{N}.$$

Ce noyau discret associé binomial $B_{x,h}$ est à support $\{0,1,\cdots,x+1\}$ (dépendant uniquement de x), *sousdispersé* de moyenne $x+h$ et de variance $(x+h)(1-h)/(x+1) < x+h$, et de mode autour de $x+h$.

EXEMPLE 2.3 : Dans le cas d'un type de noyau binomial négatif $\mathcal{BN}(\lambda,p)$, on considère le noyau discret associé $BN_{x,h}$ de loi $\mathcal{BN}(x+1,(x+1)/(2x+1+h))$ sur $\aleph_{x,h} = \mathbb{N}$ pour tout $x \in \mathbb{N}$ et $h > 0$, tels que :

$$BN_{x,h}(y) = \frac{(x+y)!}{x!y!}\left(\frac{x+h}{2x+1+h}\right)^y \left(\frac{x+1}{2x+1+h}\right)^{x+1}, \quad y \in \mathbb{N}.$$

Ce noyau discret associé $BN_{x,h}$ est de support $\mathbb{N}$, *surdispersé* de moyenne $x+h$ et de variance $(x+h)[1+(x+h)/(x+1)] > x+h$, et de mode autour de $x+h$.

REMARQUE 2.2 : Le rôle du paramètre de lissage discret $h > 0$ reste alors semblable au cas continu, car il permet de tenir compte des $y \equiv X_i$ qui sont proches de la cible $x \in \mathbb{N}$ (dans le sens de l'écart stochastique du type de noyau discret K utilisé) lorsque $h = h_n \to 0$. Cependant, la dispersion locale en tout point d'estimation $x \in \mathbb{N}$ se traduit, pour l'instant, par l'importance du noyau discret associé $K_{x,h}$ choisi, lequel impose intrinséquement sa propriété de variance. Ainsi, le choix d'un type de noyau discret K s'oriente vers des distributions de $K_{x,h}$ qui soient moins dispersées autour de la cible $x \in \mathbb{N}$, à $h > 0$ fixé.

REMARQUE 2.3 : Il existe des lois discrètes qui ne peuvent être associées à aucun noyau discret. En effet, si on considère la loi uniforme discrète $\mathcal{U}(c,a)$ centrée en $c \in \mathbb{N}$ et de bras $a \in \mathbb{N}^*$, le noyau discret associé serait $U(x,a)$, de loi $\mathcal{U}(x,a)$ sur $\aleph_{x,a} = \{x, x \pm 1, \cdots, x \pm a\}$ avec $\cup_x \aleph_{x,a} = \{-a,\cdots,-1\} \cup \mathbb{N} \supsetneq \mathbb{N}$. Le noyau discret associé correspondant s'écrirait :

$$U_{x,a}(y) = \frac{1}{2a+1}\mathbf{1}_{\{x,x\pm1,\cdots,x\pm a\}}(y), \quad y \in \mathbb{N}.$$

D'après la Définition 2.1 du noyau discret associé, il apparaît qu'on ne peut pas établir de correspondance entre (x,a) et (x,h). Le paramètre de lissage habituel $h > 0$ ne peut se substituer ici à $a \in \mathbb{N}^*$. Cette remarque est aussi valable pour une loi triangulaire discrète malgré sa propriété de symétrie autour de x. Cependant si $a = 0$, la loi discrète

uniforme $\mathcal{U}(x,0)$ correspond à une loi de Dirac $\mathcal{D}(x)$ en x. On construit alors le noyau discret associé de Dirac (noyau "naïf") $D_{x,0}$ de loi $\mathcal{D}(x)$, pour tout $x \in \mathbb{N}$ et $h = 0$:

$$D_{x,0}(y) = \delta_x(y), \quad y \in \mathbb{N}.$$

La Figure 1 donne l'allure des troix types de noyaux discrets, dits standards, donnés en exemple ainsi que du noyau "naïf" pour une cible $x \in \mathbb{N}$ et une fenêtre $h > 0$ fixées.

2.2 Propriétés élémentaires

Nous sommes en mesure de donner une définition précise d'un estimateur à noyau discret pour une densité de probabilité f sur un ensemble discret $\aleph$ et de présenter les propriétés fondamentales.

Définition 2.2 *Soit $X_1, \cdots, X_n$ un n-échantillon aléatoire i.i.d. de densité de probabilité discrète inconnue f sur $\aleph$. Un estimateur à noyau discret $\hat{f}(x) = \widehat{f}_{n,h,K}(x)$ de $f(x) := \Pr(X_1 = x)$ est défini par*

$$\widehat{f}_{n,h,K}(x) = \frac{1}{n}\sum_{i=1}^{n} K_{x,h}\left(X_i\right), \quad x \in \aleph,$$

où $h > 0$ est le paramètre de lissage discret (ou fenêtre) et $K_{x,h}$ (dépendant de x et de h) est le noyau discret associé sur $\aleph_{x,h}$ tel que $\aleph_{x,h} \bigcap \aleph \neq \emptyset$ et $\cup_x \aleph_{x,h} \supseteq \aleph$.

Pour illustrer par des exemples, on peut considérer les estimateurs standards $\hat{f}$ liés aux noyaux discrets associés $K_{x,h}$ donnés en exemple au paragraphe précédent (Exemples 2.1 à 2.3).

La proposition suivante est élémentaire mais fondamentale pour l'étude des estimateurs à noyaux discrets.

Proposition 2.3 *Soit $\underline{X} = (X_1, \cdots, X_n)$ un n-échantillon aléatoire i.i.d. de densité de probabilité discrète inconnue f sur $\aleph$. Soit $\widehat{f}_{n,h,K}(x)$ un estimateur de $f(x)$ à partir d'un noyau discret associé $K_{x,h}$ sur $\aleph_{x,h}$ tel que $\aleph_{x,h} \bigcap \aleph \neq \emptyset$ et $\cup_x \aleph_{x,h} \supseteq \aleph$. Alors, pour tout $x \in \aleph$ et $h > 0$,*

$$\mathbb{E}[\widehat{f}_{n,h,K}(x)] = \mathbb{E}[f(\mathcal{K}_{x,h})], \tag{8}$$

où $\mathcal{K}_{x,h}$ est la variable aléatoire sur $\aleph_{x,h}$ de loi $K_{x,h}$. De plus, on a $\widehat{f}_{n,h,K}(x) \in [0,1]$ et

$$\sum_{x \in \aleph} \widehat{f}_{n,h,K}(x) = C, \tag{9}$$

où $C = C(\underline{X}; h, K)$ est une constante strictement positive et finie.

DÉMONSTRATION : Il suffit d'écrire (8) comme ci-dessous, car le reste est trivial. En effet, on a successivement

$$\mathbb{E}[\widehat{f}_{n,h,K}(x)] = \sum_{y \in \aleph_{x,h}} K_{x,h}(y) f(y) = \sum_{y \in \aleph_{x,h}} f(y) \Pr(\mathcal{K}_{x,h} = y) = \mathbb{E}[f(\mathcal{K}_{x,h})]. \square$$

À partir de (8), la connaissance des propriétés de l'estimateur $\hat{f}(x) = \widehat{f}_{n,h,K}(x)$ de $f(x)$ est gouvernée en partie par une construction judicieuse du noyau discret associé $K_{x,h}$. Si $\cup_x \aleph_{x,h} \supsetneq \aleph$ alors il est nécessaire de corriger le biais de bordure (en anglais "edge effect") de $\widehat{f}_{n,h,K}(x)$ autour du bord ; ce qui n'est pas le cas pour nos trois noyaux discrets associés standards.

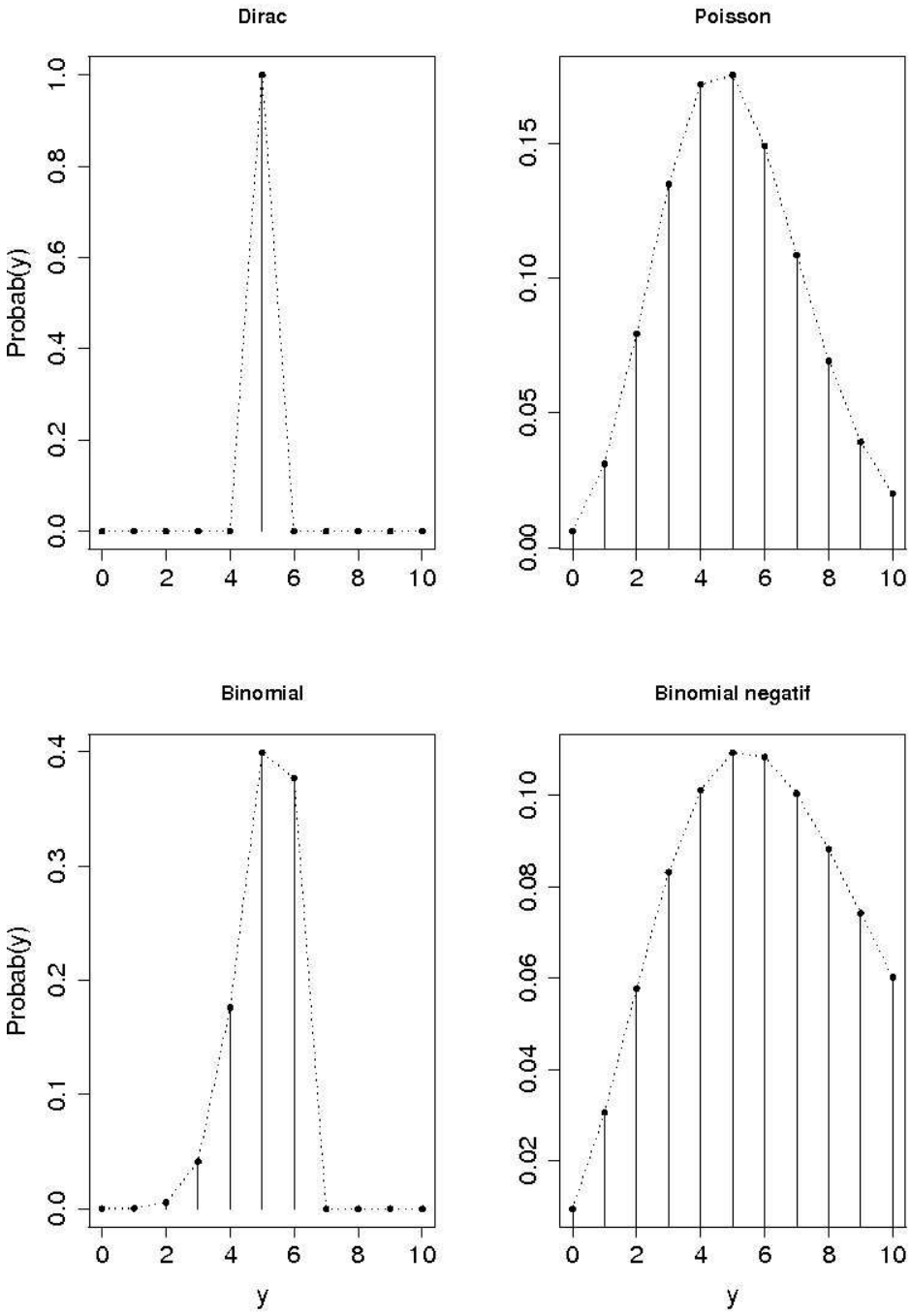

Figure 1: Noyaux discrets associés de Dirac (noyau "naïf") $\mathcal{D}(x)$, de Poisson $\mathcal{P}(x+h)$, binomial $\mathcal{B}(x+1,(x+h)/(x+1))$ et binomial négatif $\mathcal{BN}(x+1,(x+1)/(2x+1+h))$ avec $y=x=5$, $h=0.1$ (sauf pour Dirac où $h=0$)

À travers (9), on suppose désormais que la fonction

$$x \mapsto \widehat{f}_{n,h,K}(x) \equiv \frac{1}{n}\sum_{i=1}^{n}\frac{K_{x,h}\left(X_i\right)}{C}$$

est une densité de probabilité sur $\aleph$. Dans la pratique, nous calculons cette constante C avant la normalisation de l'estimation.

Le résultat suivant garantit qu'un estimateur à noyau discret converge, au moins ponctuellement vers l'estimateur naïf. Pour cela, on a besoin de considérer une version continue ou une interpolation par morceaux de toute fonction $g : \aleph \to \mathbb{R}$ en plongeant à la fois $\aleph \subset \mathbb{R}$ et $g(\aleph) \subset \mathbb{R}$ dans la topologie usuelle de $\mathbb{R}$. Autrement dit, toutes les valeurs successives $g(x)$ et $g(x+s)$ de $x \in \aleph$ et de $x+s \in \aleph$ $(s > 0)$ par g, respectivement, sont reliées par une fonction réelle continue. Par exemple, en interpolant les valeurs consécutives (ou les sommets des bâtons consécutifs) par une droite, on obtient une version continue dite interpolation linéaire par morceaux (voir Figure 1).

Proposition 2.4 *Soit x fixé dans le support $\aleph$ de f dont on considère une version continue $\tilde{f}$ dans $\mathbb{R}$. Soit $\mathcal{K}_{x,h}$ la v.a. (8) du noyau discret associé $K_{x,h}$ de $\widehat{f}_{n,h,K}(x)$. Alors, on a :*

$$\lim_{n\to\infty}\mathbb{E}[\widehat{f}_{n,h,K}(x)] = \tilde{f}(x).$$

DÉMONSTRATION : En posant $m_{x,h} = \mathbb{E}(\mathcal{K}_{x,h}) = \sum_{y\in\aleph_{x,h}} yK_{x,h}(y)$, on a successivement :

$$\begin{aligned}
\lim_{n\to+\infty}\mathbb{E}[\widehat{f}_{n,h,K}(x)] &\overset{(a)}{=} \lim_{h\to 0}\mathbb{E}[f(\mathcal{K}_{x,h})] \\
&\overset{(b)}{=} \lim_{h\to 0}\tilde{f}(m_{x,h}) \\
&\overset{(c)}{=} \tilde{f}(\lim_{h\to 0} m_{x,h}) \\
&\overset{(d)}{=} \tilde{f}(x).
\end{aligned}$$

En fait, (a) découle de (8) et du fait que $h = h_n \to 0$ lorsque $n \to +\infty$; (b) est possible car la fonction f est discrète ou somme des indicatrices dont n'importe quelle version continue $\tilde{f}$ le permette ; (c) est alors un simple calcul de limite ; et, (d) découle de (4). □

La Proposition 2.4 est facilement vérifiable pour les trois exemples standards de Poisson, binomial et binomial négatif, car la condition (4) des noyaux discrets associés est réalisée grâce à (6).

2.3 Risque asymptotique exact en un point

Avec des hypothèses supplémentaires à (4) de la Définition 2.1, nous établissons dans ce paragraphe une version classique du comportement asymptotique exact du risque quadratique en un point fixé. Ainsi, on considère d'abord l'extension de (4) par (5) qu'on réécrit ici

$$m_{x,h} = \mathbb{E}(\mathcal{K}_{x,h}) = x + h + o(h),$$

à laquelle on ajoute l'hypothèse générale de variance du type

$$\sigma^2_{x,h} = Var(\mathcal{K}_{x,h}) = V(x,h) + o(h), \tag{10}$$

avec $V(x,h) \geq 0$. Le *risque quadratique ponctuel* (en anglais "Mean Squared Error") de $\widehat{f}_{n,h,K(x)}$ défini par

$$\begin{aligned} MSE &:= \mathbb{E}\left[\left(\widehat{f}_{n,h,K}(x) - f(x)\right)^2\right] \\ &= Var\left[\widehat{f}_{n,h,K}(x)\right] + Biais^2\left[\widehat{f}_{n,h,K}(x)\right] \end{aligned}$$

permet de contrôler à la fois le biais (ou l'erreur systématique) et la variance (ou l'erreur stochastique) de $\widehat{f}_{n,h,K}(x)$ en fonction de $h = h_n$, pour un type de noyau discret K donné.

Nous commençons par l'analyse de la variance ponctuelle. Soit $x \in \aleph$ fixé. La variance $Var\left[\hat{f}_{n,h,K}(x)\right] = \mathbb{E}\left[\hat{f}_{n,h,K}(x) - \mathbb{E}\left(\hat{f}_{n,h,K}(x)\right)\right]^2$ est donnée de manière successive par

$$\begin{aligned} Var[\widehat{f}_{n,h,K}(x)] &= \frac{1}{n}Var[K_{x,h}(X_1)] \\ &= \frac{1}{n}\mathbb{E}\left[K_{x,h}\left(X_1\right)\right]^2 - \frac{1}{n}[E\{K_{x,h}\left(X_1\right)\}]^2 \\ &= \frac{1}{n}\left[\sum_{y\in\aleph_{x,h}} f(y)\{\Pr(\mathcal{K}_{x,h}=y)\}^2 - \left\{\sum_{y\in\aleph_{x,h}} f(y)\Pr(\mathcal{K}_{x,h}=y)\right\}^2\right] \\ &= \frac{1}{n}[f(x)\Sigma(\mathcal{K}_{x,h}^2) - f^2(x)] + O(\frac{h}{n}) \qquad (11) \\ &\doteq \frac{1}{n}f(x)\Pr(\mathcal{K}_{x,h}=x), \qquad (12) \end{aligned}$$

où le terme $\Sigma(\mathcal{K}_{x,h}^2) := \sum_{y\in\aleph_{x,h}}\{\Pr(\mathcal{K}_{x,h} = y)\}^2$ de (11) est majoré par 1 et l'approximation finale (12) dépend de la condition (4) à travers la probabilité modale $\Pr(\mathcal{K}_{x,h} = x)$ de $\mathcal{K}_{x,h}$.

Ainsi, on a : $Var\left[\widehat{f}_{n,h,K}(x)\right] \to 0$ quand $n \to +\infty$, pour tout $x \in \aleph$ et $h > 0$. On peut observer que cette conclusion est obtenue sans aucune condition ni sur la fonction de probabilité discrète f [alors que $f(x) \leq f_{\max} < +\infty$ pour la densité f continue], ni sur le type de noyau discret K [alors que $\int K^2(t)dt < +\infty$ pour le noyau continu], et ni sur la fenêtre associée h [alors que $nh \to +\infty$ par rapport au cas continu].

Concernant le biais ponctuel, on considère que la v.a. $\mathcal{K}_{x,h}$ associée à $K_{x,h}$ est de moyenne $m_{x,h} = \mathbb{E}(\mathcal{K}_{x,h})$ et de variance $\sigma_{x,h}^2 = Var(\mathcal{K}_{x,h}) < +\infty$. Par le développement discret de Taylor (*e.g.* Mangasarian & Schumaker, 1973) à l'ordre 2, on a :

$$f(\mathcal{K}_{x,h}) = f(m_{x,h}) + (\mathcal{K}_{x,h} - m_{x,h})f'(x) + \frac{(\mathcal{K}_{x,h} - m_{x,h})^2}{2}f''(x) + o(h).$$

Puis, en prenant l'espérance mathématique, on obtient le biais ponctuel

$$\begin{aligned} Biais[\widehat{f}_{n,h,K}(x)] &= \mathbb{E}[f(\mathcal{K}_{x,h})] - f(x) \\ &= \left(f(\mathbb{E}[\mathcal{K}_{x,h}]) - f(x) + \frac{1}{2}Var[\mathcal{K}_{x,h}]f''(x)\right)(1 + o(1)), \end{aligned}$$

où $o(1)$ ne dépend pas de n et tend vers 0 quand $h \to 0$.

Notons qu'il est raisonnable de s'arrêter à l'ordre 1 en h, car dépasser l'ordre 2 dépend des propriétés des moments centrés de $\mathcal{K}_{x,h}$ d'ordre supérieur à 2. Les dérivées

$f^{(k)}(x)$ d'ordre $k \geq 1$ de f en $x \in \aleph$ sont remplacées par les différences finies. En particulier, si $\aleph = \mathbb{N}$ alors on écrit par récurrence

$$f^{(k)}(x) = \left[f^{(k-1)}(x)\right]' \text{ et } f'(x) = \begin{cases} [f(x+1) - f(x-1)]/2 & \text{si } x \in \mathbb{N}^* \\ f(1) - f(0) & \text{si } x = 0. \end{cases} \tag{13}$$

En fait, les $f^{(k)}(x)$ existent toujours et sont des combinaisons linéaires de $f(x \pm j)$ pour $j \in \{0, 1, \cdots, k\}$ et $x \pm j \in \aleph$. Par conséquent, l'expression de MSE obtenu par les approximations de la variance et du biais est donné par :

$$\begin{aligned} AMSE(x; n, h, K, f) \doteq & \frac{f(x)\Pr(\mathcal{K}_{x,h} = x)}{n} \\ & + \left(f(\mathbb{E}[\mathcal{K}_{x,h}]) - f(x) + \frac{Var[\mathcal{K}_{x,h}]}{2} f''(x)\right)^2. \end{aligned} \tag{14}$$

En réalisant le second développement discret de Taylor

$$f(m_{x,h}) \doteq f(x+h) = f(x) + hf'(x) + o(h),$$

il s'en suit

$$Biais[\widehat{f}_{n,h,K}(x)] \doteq h\left(f'(x) + \frac{V(x,h)}{2h} f''(x)\right)$$

Puisqu'il est facile de vérifier que $f'(x) + f''(x)V(x,h)/2h < +\infty$ et $h = h_n$ tend vers 0 lorsque $n \to +\infty$, alors $Biais\left[\widehat{f}_{n,h,K}(x)\right]$ tend vers 0 quand $n \to +\infty$ pour tout $x \in \aleph$. Précisons qu'à partir de (13), on a clairement

$$f''(x) = \begin{cases} [f(x+2) - 2f(x) + f(x-2)]/4 & \text{si } x \in \mathbb{N} \setminus \{0,1\} \\ [f(3) - f(2) - f(1) + f(0)]/2 & \text{si } x = 1 \\ f(2) - 2f(1) + f(0) & \text{si } x = 0. \end{cases} \tag{15}$$

Nous sommes donc en mesure de formuler un théorème donnant le comportement asymptotique exact du MSE de $\widehat{f}_{n,h,K}(x)$.

Théorème 2.5 *Supposons que, pour tout $x \in \aleph$ et $h > 0$,*
(i) le noyau discret associé $K_{x,h}$ vérifie (5) et (10) avec $V(x,h) > 0$ et $K_{x,h}(x) > 0$,
(ii) la densité de probabilité discrète inconnue f vérifie $f(x) > 0$ et $f'(x) + f''(x)V(x,h)/2h \neq 0$. Alors, pour tout $n \geq 1$,

$$MSE = \left[\frac{1}{n} f(x)\Pr(\mathcal{K}_{x,h} = x) + h^2\left(f'(x) + \frac{V(x,h)}{2h} f''(x)\right)^2\right](1 + o(1)), \tag{16}$$

où $o(1)$ ne dépend pas de n et tend vers 0 quand $h \to 0$.

Le terme principal de MSE est donné par l'expression entre crochets dans (16) lequel s'écrit :

$$AMSE^*(x; n, h, K, f) \doteq \frac{1}{n} f(x)\Pr(\mathcal{K}_{x,h} = x) + h^2\left(f'(x) + \frac{V(x,h)}{2h} f''(x)\right)^2. \tag{17}$$

Pour $x \in \aleph$ fixé, la minimisation en h de la borne supérieure $AMSE(x; n, h, K, f)$ conduit à la vitesse de convergence ponctuelle $O(n^{-1})$ de $\widehat{f}_{n,h,K}(x)$ vers $f(x)$ lorsque $n \to +\infty$.

2.4 Risque quadratique intégré

De manière similaire, nous évaluons le risque global de $\widehat{f}_{n,h,K}$. Le *risque quadratique intégré* (en anglais "Mean Integrated Squared Error") de $\widehat{f}_{n,h,K}$ peut être défini directement par

$$MISE = \mathbb{E} \sum_{x \in \aleph} \left[\widehat{f}_{n,h,K}(x) - f(x) \right]^2 \tag{18}$$

$$= \sum_{x \in \aleph} Var \left[\widehat{f}_{n,h,K}(x) \right] + \sum_{x \in \aleph} Biais^2 \left[\widehat{f}_{n,h,K}(x) \right] \tag{19}$$

L'approximation asymptotique de $MISE$ s'obtient par la somme des approximations de la variance et du carré du biais (14) par

$$\begin{aligned} AMISE(n,h,K,f) \doteq & \frac{1}{n} \sum_{x \in \aleph} f(x) \Pr(\mathcal{K}_{x,h} = x) \\ & + \sum_{x \in \aleph} \left[f(\mathbb{E}[\mathcal{K}_{x,h}]) - f(x) + \frac{Var[\mathcal{K}_{x,h}]}{2} f''(x) \right]^2 \end{aligned} \tag{20}$$

Il découle du Théorème 2.5 que nous pouvons donner aussitôt le principal résultat sur le $MISE$.

Théorème 2.6 *Supposons que :*
(i) le noyau discret K satisfaisant (5) et (10) est tel que

$$\sum_{x \in \aleph} K_{x,h}(x) < +\infty \ ;$$

(ii) la densité de probabilité discrète inconnue f vérifie

$$\sum_{x \in \aleph} \left[f'(x) + \frac{V(x,h)}{2h} f''(x) \right]^2 < +\infty.$$

Alors, pour tout $h > 0$,

$$MISE \leq \frac{C_2}{n} \sum_{x \in \aleph} K_{x,h}(x) + h^2 \sum_{x \in \aleph} \left[f'(x) + \frac{V(x,h)}{2h} f''(x) \right]^2, \tag{21}$$

où $C_2 = f_{\max} \leq 1$.

DÉMONSTRATION : Puisque $f(x) \leq f_{\max} \leq 1$ pour tout $x \in \aleph$, le résultat découle immédiatement des hypothèses annoncées et de la majoration triviale de l'expression intégrée de $AMSE(x; n, h, K, f)$ en (17). □

Le terme principal de $MISE$ est donné par l'expression :

$$\begin{aligned} AMISE^*(n,h,K,f) \doteq & \frac{1}{n} \sum_{x \in \aleph} f(x) \Pr(\mathcal{K}_{x,h} = x) \\ & + h^2 \sum_{x \in \aleph} \left[f'(x) + \frac{V(x,h)}{2h} f''(x) \right]^2 \end{aligned} \tag{22}$$

Par conséquent, la minimisation en h de la borne supérieure $AMISE(n,h,K,f)$ conduit cette fois-ci à la vitesse de convergence globale $O(n^{-1})$ de $\widehat{f}_{n,h,K}$ vers f lorsque

$n \to +\infty$. Aussi, pour un type de noyau discret K spécifié, la valeur optimale h^* de h est donnée par

$$h^* = \arg\min_{h>0} AMISE(n, h, K, f) = h^*(n, K, f). \tag{23}$$

Des études plus fines sont à faire au cas par cas des noyaux discrets associés standards, lesquels ont encore une importance primordiale dans cette théorie des estimateurs à noyau discret.

3 Exemples standards

On donne des exemples standards des estimateurs à noyau discret relativement aux cas des noyaux discrets associés de Poisson, binomial et binomial négatif avec les expressions de $AMISE^*$ (22). Mais les simulations sont présentées dans les sections suivantes.

Dans le cas particulier du noyau naïf présenté dans la Remarque 2.3, la v.a. $\mathcal{K}_{x,h}$ de loi de Dirac $\mathcal{D}_x$, du noyau discret associé $D_{x,0}$, vérifie $\mathbb{E}(\mathcal{K}_{x,h}) = x$ et $Var(\mathcal{K}_{x,h}) = 0$, pour $x \in \mathbb{N}$. A partir de l'expression de $AMISE$ (20), pour une taille d'échantillon n on a :

$$AMISE(n, 0, D, f) \doteq \frac{1}{n}.$$

Pour le mesurer avec ceux des cas standrards vérifiant (5) et (10), on comparera les ordres de grandeur de leur $AMISE$ à $1/n$.

Type de noyau	$\mathbb{E}(\mathcal{K}_{x,h})$	$Var(\mathcal{K}_{x,h})$	$V(x,h)$
Poisson	$x+h$	$x+h$	$x+h$
Binomial	$x+h$	$(x+h)\left(\frac{1-h}{x+1}\right)$	$(x+h)\left(\frac{1}{x+1}\right) - \frac{xh}{x+1}$
Binomial négatif	$x+h$	$(x+h)\left(1+\frac{x+h}{x+1}\right)$	$(x+h)\left(1+\frac{x}{x+1}\right) + \frac{xh}{x+1}$

Table 1: Résumé des propriétés des noyaux discrets associés standards

3.1 Poisson

À partir de l'Exemple 2.1 du noyau discret associé $P_{x,h}$ de la loi de Poisson $\mathcal{P}(x+h)$ sur $\mathbb{N}$ pour tout $x \in \mathbb{N}$ et $h > 0$, les conditions *(i)* du Théorème 2.6 sont ici satisfaites. En effet, on a facilement (5) et (10) par la Table 1 et

$$\sum_{x\in\mathbb{N}} P_{x,h}(x) = \sum_{x\in\mathbb{N}} \frac{(x+h)^x e^{-(x+h)}}{x!} = \sum_{x\in\mathbb{N}} p(x,h) < +\infty,$$

car $p(x,h) = \Pr(\mathcal{K}_{x,h} = x)$ est une probabilité individuelle de $\mathcal{P}(x+h)$, donc $p(x,h) < 1$ pour tout $x \in \mathbb{N}$ et $h > 0$. La condition *(ii)* du Théorème 2.6 portant essentiellement sur la densité discrète f s'écrit ici comme suit :

$$\sum_{x\in\mathbb{N}} \left[f'(x) + \frac{1}{2}\left(\frac{x}{h}+1\right) f''(x) \right]^2 < +\infty,$$

où $f'(x)$ et $f''(x)$ sont respectivement données en (13) et en (15). Ainsi, l'expression du terme principal de $MISE$ dans le cas du noyau de type poissonnien est donnée à l'ordre 2 par :

$$AMISE^*(n,h,f) \doteq \frac{1}{n}\sum_{x\in\mathbb{N}} \frac{(x+h)^x e^{-(x+h)}}{x!} f(x) + h^2 \sum_{x\in\mathbb{N}} \left[f'(x) + \frac{1}{2}\left(\frac{x}{h}+1\right) f''(x) \right]^2.$$

3.2 Binomial

Les conditions *(i)* du Théorème 2.6 sont aussi vérifiées pour le noyau discret associé $B_{x,h}$ de la loi binomiale $\mathcal{B}(x+1,(x+h)/(x+1))$, présenté dans l'Exemple 2.2, de support $\{0,1,\cdots,x+1\}$ pour tout $x\in\mathbb{N}$ et $h\in]0,1]$. De manière similaire au cas de Poisson, on a :

$$\sum_{x\in\mathbb{N}} B_{x,h}(x) = (1-h)\sum_{x\in\mathbb{N}} \left(\frac{x+h}{x+1}\right)^x < +\infty.$$

La condition *(ii)* du Théorème 2.6 se traduit dans le cas du noyau de type binomial par

$$\sum_{x\in\mathbb{N}} \left[f'(x) + \frac{1}{2(x+1)}\left(\frac{x}{h}+1-x\right) f''(x) \right]^2 < +\infty,$$

où les approximations des dérivées $f'(x)$ et $f''(x)$ de la densité discrète f sont respectivement données en (13) et en (15). Par conséquent, le terme principal du $MISE$ à l'ordre 2 s'écrit ici par :

$$\begin{aligned} AMISE^*(n,h,f) \doteq\ & \frac{1-h}{n}\sum_{x\in\mathbb{N}} \left(\frac{x+h}{x+1}\right)^x f(x) \\ & +\ h^2 \sum_{x\in\mathbb{N}} \left[f'(x) + \frac{1}{2(x+1)}\left(\frac{x}{h}+1-x\right) f''(x) \right]^2. \end{aligned}$$

3.3 Binomial négatif

De manière similaire, les conditions *(i)* du Théorème 2.6 sont également vérifiées pour le noyau discret associé $BN_{x,h}$ de la loi binomiale négative $\mathcal{BN}(x+1,(x+1)/(2x+1+h))$ donné à l'Exemple 2.3, de support $\mathbb{N}$ pour tout $x\in\mathbb{N}$ et $h>0$ et tels que

$$\sum_{x\in\mathbb{N}} BN_{x,h}(x) = \sum_{x\in\mathbb{N}} \frac{(2x)!}{(x!)^2}\left(\frac{x+h}{2x+1+h}\right)^x \left(\frac{x+1}{2x+1+h}\right)^{x+1} < +\infty.$$

Aussi, la condition *(ii)* du Théorème 2.6 s'écrit dans le cas du noyau de type binomial négatif par

$$\sum_{x\in\mathbb{N}} \left[f'(x) + \frac{1}{2}\left\{ \left(\frac{x}{h}+1\right)\left(1+\frac{x}{x+1}\right) + \frac{x}{x+1} \right\} f''(x) \right]^2 < +\infty,$$

où les approximations des dérivées $f'(x)$ et $f''(x)$ de la densité discrète f sont respectivement données en (13) et en (15). D'où, le terme principal du $MISE$ à l'ordre 2 se présente de la manière suivante :

$$\begin{aligned} AMISE^*(n,h,f) \doteq\ & \frac{1}{n}\sum_{x\in\mathbb{N}} \frac{(2x)!}{(x!)^2}\left(\frac{x+h}{2x+1+h}\right)^x \left(\frac{x+1}{2x+1+h}\right)^{x+1} f(x) \\ & +\ h^2 \sum_{x\in\mathbb{N}} \left[f'(x) + \frac{1}{2}\left\{ \left(\frac{x}{h}+1\right)\left(1+\frac{x}{x+1}\right) + \frac{x}{x+1} \right\} f''(x) \right]^2. \end{aligned}$$

4 Efficacité relative des noyaux discrets associés

Pour une cible $x \in \mathbb{N}$ et une fenêtre $h > 0$, nous donnons une mesure relative de performance entre les noyaux discrets associés standards.

Soient $K^0_{x,h}$ et $K^1_{x,h}$ deux noyaux discrets associés à $\mathcal{K}^0_{x,h}$ et $\mathcal{K}^1_{x,h}$ respectivement, qui vérifie chacun la relation $AMISE$ (20). Pour comparer leur efficacité dans l'estimation non-paramétrique d'une densité de probabilité discrète f, nous supposons d'abord qu'ils sont comparables en imposant la condition de comparabilité :

$$\mathbb{E}\left[\mathcal{K}^0_{x,h}\right] = \mathbb{E}\left[\mathcal{K}^1_{x,h}\right], \tag{24}$$

pour tout $x \in \mathbb{N}$ et $h > 0$ petit. Ensuite, toujours à partir de (20), l'efficacité relative entre $\mathcal{K}^0_{x,h}$ et $\mathcal{K}^1_{x,h}$ peut être grossièrement mesurée par le rapport de leur variance :

$$\textit{eff}(\mathcal{K}^0, \mathcal{K}^1) = \frac{Var\left[\mathcal{K}^1_{x,h}\right]}{Var\left[\mathcal{K}^0_{x,h}\right]}, \tag{25}$$

pour tout $x \in \mathbb{N}$ et $h > 0$ petit. En d'autres termes, si nous avons

$$Var\left[\mathcal{K}^0_{x,h}\right] < Var\left[\mathcal{K}^1_{x,h}\right], \tag{26}$$

pour tout $x \in \mathbb{N}$ et $h > 0$ petit, alors le noyau discret associé $K^0_{x,h}$ est plus efficace que $K^1_{x,h}$.

Le noyau discret associé le plus performant entre $K^0_{x,h}$ et $K^1_{x,h}$ peut être alors comparé au noyau naïf. Toutefois, l'égalité (24) entre la moyenne du noyau naïf et celle du noyau discret associé n'est pas vérifiée, on ne peut donc pas utiliser le critère d'efficacité définie par la relation (25). Il faut donc évaluer l'ordre des grandeurs des expressions de $AMISE$ respectives. Ce qui revient à choisir le noyau discret associé le plus adapté selon la taille n de l'échantillon.

Il y a deux façons heuristiques de voir que le critère (25) suffise à comparer l'efficacité relative entre $\mathcal{K}^0_{x,h}$ et $\mathcal{K}^1_{x,h}$ en utilisant l'expression de $AMISE$ (20) sous la condition de comparabilité (24). La première façon est de considérer n grand ($\to \infty$) et par conséquent $h = h_n$ est petit ($\to 0$). Ainsi, l'approximation de l'expression de la variance intégrée dans $AMISE$ (20) peut être majorée par :

$$\frac{1}{n}\sum_{x\in\mathbb{N}} f(x)\Pr(\mathcal{K}_{x,h} = x) \leq \frac{1}{n}\sum_{x\in\mathbb{N}} \Pr(\mathcal{K}_{x,h} = x),$$

laquelle devient très petite (tend vers 0) pour n grand car

$$\sum_{x\in\mathbb{N}} \Pr(\mathcal{K}_{x,h} = x) < \infty.$$

La seconde façon utilise (24) qui implique, pour toute densité de probabilité discrète f,

$$\frac{1}{n}\sum_{y\in\mathbb{N}} f(y)\Pr(\mathcal{K}^0_{x,h} = y) = \frac{1}{n}\sum_{y\in\mathbb{N}} f(y)\Pr(\mathcal{K}^1_{x,h} = y).$$

En remplaçant y par x dans les deux membres de l'égalité ci-dessus, on obtient

$$\frac{1}{n}\sum_{x\in\mathbb{N}} f(x)\Pr(\mathcal{K}^0_{x,h} = x) = \frac{1}{n}\sum_{x\in\mathbb{N}} f(x)\Pr(\mathcal{K}^1_{x,h} = x).$$

Par conséquent, la comparaison de $AMISE$ (20) de $\mathcal{K}^0_{x,h}$ et $\mathcal{K}^1_{x,h}$ sous la condition (24) revient au rapport (25) à travers ces deux chemins.

Maintenant, considérons la condition (24) modifiée comme

$$\mathbb{E}\left[\mathcal{K}_{x,h}^{j}\right] = x + h + o(h), \;\; j = 0, 1 \tag{27}$$

et l'hypothèse suivante sur les variances

$$Var\left[\mathcal{K}_{x,h}^{j}\right] = V^{j}(x,h) + o(h), \;\; j = 0, 1, \tag{28}$$

avec $V^{j}(x,h) \geq 0$. Ainsi, sous les conditions modifiées (27) et (28), la nouvelle mesure d'efficacité relative entre $\mathcal{K}_{x,h}^{0}$ et $\mathcal{K}_{x,h}^{1}$ est obtenu par le rapport suivant :

$$\text{eff}^{*}(\mathcal{K}^{0}, \mathcal{K}^{1}) = \frac{V^{1}(x,h)}{V^{0}(x,h)}, \tag{29}$$

pour tout $x \in \mathbb{N}$ et $h > 0$ petit.

D'après le critère (29), parmi les types des noyaux discrets standards, un noyau discret associé est d'autant plus efficace que si sa variance est petite. Donc, les noyaux sousdispersés sont plus performants que les noyaux équidispersés ou surdispersés.

La Figure 2 représente sur une même échelle les différents noyaux discrets associés déjà présentés en Figure 1. On peut mieux y observer les différents comportements de dispersion autour de la cible choisie.

5 Choix de fenêtres

Nous présentons trois méthodes de choix de fenêtres pour approcher la valeur idéale de la fenêtre h définie par

$$h_{id} = \arg\min_{h>0} MISE(n,h,K,f) = h_{id}(n,K,f). \tag{30}$$

La première approche consiste à minimiser le ISE ou encore les approximations $AMISE^{*}$ du $MISE$ obtenues en utilisant les différences finies de f. En effet, l'existence d'un minimum par rapport à h est garantie par la décroissance de la variance intégrée et la croissance du carré du biais intégré dans le risque quadratique global (19). Pour une petite valeur de h, le biais est également petit mais la variance est grande. A l'inverse, si h est grand, c'est la variance qui devient petite et le biais plus grand. Pour trouver la fenêtre optimale, on doit balancer les approximations du carré du biais et de la variance. Autrement dit, il existe $\varepsilon > 0$ telle que la fonction $h \mapsto AMISE(n,h,K,f)$ soit décroissante sur $]0,\varepsilon[$ et croissante sur $]\varepsilon,+\infty[$ pour tout $h > 0$.

La deuxième méthode est la validation croisée qui est bien connue. Enfin, la dernière procédure est adaptée au cas courant d'une proportion importante de zéros dans les données de comptage.

5.1 Minimisation des erreurs quadratiques

Du point vue purement pratique où $\underline{X} = (X_1, \cdots, X_n)$ est un n-échantillon fixé de f et, donc, associé à la distribution empirique f_0 de f, nous proposons maintenant quelques types de fenêtres liées aux erreurs d'estimations. La première est déduite de l'*erreur quadratique intégrée* (en anglais "Integrated Squared Error") définie par

$$ISE := \sum_{x \in \aleph} \left[\widehat{f}_{n,h,K}(x) - f(x)\right]^{2} = ISE(\underline{X}; h, K, f), \tag{31}$$

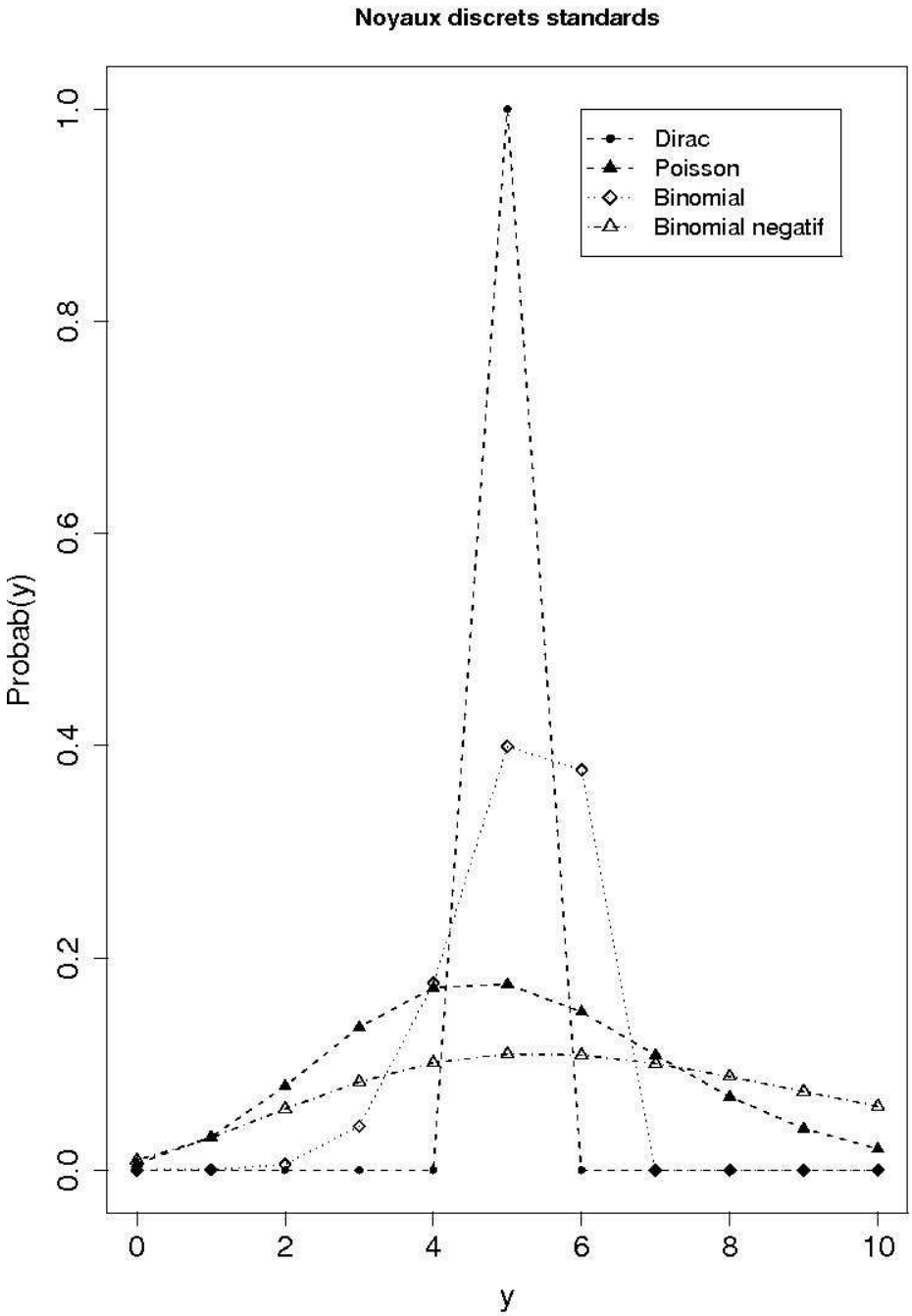

Figure 2: Noyaux discrets associés de Dirac (noyau "naïf") (x), de Poisson $\mathcal{P}(x+h)$, binomial $\mathcal{B}(x+1,(x+h)/(x+1))$ et binomial négatif $\mathcal{BN}(x+1,(x+1)/(2x+1+h))$ avec $y=x=5$, $h=0.1$ (sauf pour Dirac où $h=0$)

laquelle mesure sur un seul échantillon $\underline{X}$ l'écart (au sens quadratique) entre $\hat{f}$ et f. Par conséquent, la minimisation en h de l'ISE (31) conduit à choisir une *fenêtre adéquate*

$$h^{**} = \arg\min_{h>0} ISE(\underline{X}; h, K, f) = h^{**}(n, K, f). \tag{32}$$

En remplaçant f par f_0 dans (32), on utilisera $h_0^{**} = h^{**}(n, K, f_0)$ pour le lissage discret d'un f_0 de f. Autrement dit, on a

$$h_0^{**} = \arg\min_{h>0} ISE(\underline{X}; h, K, f_0) = h_0^{**}(n, K, f_0). \tag{33}$$

Basé sur la convergence de f_0 vers f quand $n \to +\infty$, on a immédiatement

$$\lim_{n\to+\infty} h_0^{**}(n, K, f_0) = \lim_{n\to+\infty} h^{**}(n, K, f), \tag{34}$$

pour un type de noyau discret K donné. L'importance de la fenêtre adéquate h^{**} (32) de h est due, en partie, aux relations suivantes :

$$MISE = \mathbb{E}(ISE) = \sum_{x\in\aleph} MSE(x). \tag{35}$$

La procédure fournissant h_0^{**} par (33) en se basant sur (34) est illustrée à travers des données simulées dans la prochaine section.

Quant à la fenêtre h^* obtenue par (23), on peut procéder de manière similaire qu'en (33) et (34) pour se donner une estimation $h_0^* = h^*(n, K, f_0)$ de $h^* = h^*(n, K, f)$ par f_0. Cependant, cette fenêtre h^* de (23) est loin d'être la plus appropriée dans le cas discret si l'allure (ou la régularité) de la densité discrète f n'est pas sympathique ; c'est à dire si l'approximation des dérivées de f faite grossièrement par les différences finies (13) et (15) n'est pas satisfaisante. Dans ce cas, il faudrait réduire l'ordre $k \geq 1$ des $f^{(k)}(x)$ apparaissant dans $AMISE(n, h, K, f)$ de (21), voire (17). Finalement, d'après (23) et (35), la fenêtre optimale h^* de h dans ce cas discret peut être obtenue à travers

$$h^* = \arg\min_{h>0} \mathbb{E}[ISE(\underline{X}; h, K, f)], \tag{36}$$

où les approximations des dérivées de f n'y interviennent pas.

5.2 Validation croisée

Considérons un noyau discret associé $K_{x,h}$, $x \in \mathbb{N}$ et $h > 0$. La méthode classique de *validation croisée* (en anglais "Cross-Validation") ne faisant pas usage des approximations des dérivées de f est toujours applicable dans le contexte des estimateurs à noyau discret pour mieux estimer la valeur idéale h_{id} (30) de h. La fenêtre optimale s'obtient par

$$h_{cv} = arg \min_{h>0} CV(h) \tag{37}$$

avec

$$\begin{aligned} CV(h) &= \sum_{x\in\mathbb{N}} \hat{f}^2_{n,h,K}(x) - \frac{2}{n}\sum_{i=1}^{n} \widehat{f}_{n,h,K,-i}\left(X_i\right) \\ &= \sum_{x\in\mathbb{N}} \left[\frac{1}{n}\sum_{i=1}^{n} K_{x,h}\left(X_i\right)\right]^2 - \frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i} K_{X_i,h}\left(X_j\right), \end{aligned}$$

où $\widehat{f}_{n,h,K,-i}$ est calculé sans l'observation X_i.

Le principe de cette méthode est de minimiser par rapport à h un estimateur de $MISE$ pour trouver le paramètre optimal. Pour cela, $MISE$ (18) peut être développé comme suit :

$$MISE = \mathbb{E}\left[\sum_{x\in\mathbb{N}} \hat{f}^2_{n,h,K}(x)\right] - 2\mathbb{E}\left[\sum_{x\in\mathbb{N}} \widehat{f}_{n,h,K}(x)f(x)\right] + \sum_{x\in\mathbb{N}} f^2(x).$$

Le terme $\sum_{x\in\mathbb{N}} f^2(x)$ n'est pas aléatoire, et ne dépend pas de h. On note alors,

$$MISE_{cv} = \mathbb{E}\left[\sum_{x\in\mathbb{N}} \hat{f}^2_{n,h,K}(x)\right] - 2\mathbb{E}\left[\sum_{x\in\mathbb{N}} \widehat{f}_{n,h,K}(x)f(x)\right] = MISE_{cv}(h),$$

le terme de $MISE$ qui dépend de h. Dans la suite, nous déterminons un estimateur $CV(h)$ de $MISE_{cv}$.

D'abord, on a évidemment $\sum_{x\in\mathbb{N}} \hat{f}^2_{n,h,K}(x)$ qui est un estimateur sans biais de $\mathbb{E}\left[\sum_{x\in\mathbb{N}} \hat{f}^2_{n,h,K}(x)\right]$.

Ensuite, soit

$$\widehat{f}_{n,h,K,-i}(x) = \frac{1}{n-1}\sum_{j\neq i} K_{x,h}\left(X_j\right).$$

Par construction,

$$\begin{aligned}\widehat{G} &= \frac{1}{n}\sum_{i=1}^{n} \widehat{f}_{n,h,K,-i}\left(X_i\right)\\ &= \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i} K_{X_i,h}\left(X_j\right)\end{aligned}$$

est un estimateur de $\mathbb{E}\left[\sum_{x\in\mathbb{N}} \widehat{f}_{n,h,K}(x)f(x)\right]$ et on vérifie de plus qu'il est sans biais. En effet, d'une part, comme les v.a. $X_1, \cdots, X_n$ sont i.i.d., on a

$$\begin{aligned}\mathbb{E}[\widehat{G}] &= \mathbb{E}\left[\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq 1} K_{X_1,h}\left(X_j\right)\right]\\ &= \mathbb{E}\left[\frac{1}{n-1}\sum_{j\neq 1} K_{X_1,h}\left(X_j\right)\right]\\ &= \mathbb{E}\left[K_{X_1,h}\left(X_2\right)\right].\end{aligned}$$

D'autre part, on a successivement :

$$\begin{aligned}\mathbb{E}\left[\sum_{x\in\mathbb{N}} \widehat{f}_{n,h,K}(x)f(x)\right] &= \mathbb{E}\left[\sum_{x\in\mathbb{N}} f(x)\frac{1}{n}\sum_{i=1}^{n} K_{x,h}(X_i)\right]\\ &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} K_{X_1,h}(X_i)\right]\\ &= \mathbb{E}\left[K_{X_1,h}(X_2)\right].\end{aligned}$$

Finalement, on vient de montrer que $CV(h)$ est un estimateur sans biais de $MISE_{cv}$. Pour quelques détails, on peut se référer à de nombreux auteurs tels Bowman (1984), Marron (1984), Rudemo (1982), Stone (1984) et leurs références.

5.3 Excès de zéros

Pour cette section, le choix de la fenêtre repose sur une particularité des données de comptage avec $\aleph = \mathbb{N}$ qui n'est autre que l'excès des zéros dans l'échantillon $\underline{X} = (X_1, \cdots, X_n)$. Pour ce phénomène bien connu (voir, par exemple, Kokonendji *et al.*, 2007a, et leurs références) et étant donné un noyau discret associé $K_{x,h}$, on peut choisir une *fenêtre adaptée* $h_0 = h_0(\underline{X}; K)$ de h satisfaisant

$$\sum_{i=1}^{n} \Pr\left[\mathcal{K}_{X_i,h_0} = 0\right] = n_0, \tag{38}$$

où $n_0 = \sharp\,(X_i = 0)$ désigne le nombre des zéros dans $\underline{X}$; voir Marsh & Mukhopadhyay (1999) pour leur noyau du type poissonnien. L'équation (38) s'obtient à partir de l'expression

$$\mathbb{E}[\widehat{f}_{n,h,K}(x)] = \sum_{y\in\mathbb{N}} \Pr(\mathcal{K}_{x,h} = y) f(y),$$

dans laquelle on prend $y = 0$ et $f(0) = 1$ afin d'identifier le nombre de zéros théoriques au nombre de zéros empiriques n_0.

Cette fenêtre h_0 ajuste le nombre de zéros théorique au nombre de zéros observé. Selon la proportion des zéros dans l'échantillon et du noyau discret associé retenu, la fenêtre adaptée h_0 obtenue par (38) devient alors comparable à la fenêtre de validation croisée (37) ou à la fenêtre adéquate h_0^{**} de (33).

Dans le cas du noyau de type de Poisson (Exemple 2.1), la fenêtre adaptée h_0 est connue explicitement. Tandis que dans le cas des noyaux de type binomial (Exemple 2.2) et binomial négatif (Exemple 2.3), la fenêtre h_0 est obtenue par la résolution numérique d'une équation non-linéaire (voir Table 2).

Type de noyau	h_0 : (38)
Poisson	$h_0 = \log\left[\frac{1}{n_0}\sum_{i=1}^{n} e^{-X_i}\right]$
Binomial	$\sum_{i=1}^{n}\left(\frac{1-h_0}{X_i+1}\right)^{X_i+1} = n_0$
Binomial négatif	$\sum_{i=1}^{n}\left(\frac{X_i+1}{2X_i+1+h_0}\right)^{X_i+1} = n_0$

Table 2: Solutions h_0 pour les noyaux discrets associés standards

6 Etudes par des données simulées

À l'aide des données de comptage simulées, nous mettons en évidence certains aspects des estimateurs à noyaux discrets standards. Entre autre la comparaison entre les trois types de noyaux discrets, à savoir Poisson (ou equidispersé), binomial (ou sousdispersé) et binomial négatif (ou surdispersé), nous examinons dans les cas des noyaux discrets standards les comportements des approximations des dérivées $f^{(k)}(x)$, $k \in \{1, 2\}$, par différences finies (13) et (15) dans les $AMISE$. Nous étudions également les performances des estimateurs pour un lissage discret selon la fenêtre h_{cv} de validation croisée (37) et la fenêtre h_0 des excès de zéros (38). Notons qu'en pratique et sans perte de généralité, la qualité des estimations ou des lissages est mesurée classiquement par l'erreur ISE définie en (31) et reliée à $MISE$ par (35). Dans le cas

(33) où f est remplacée par la distribution empirique f_0, l'ISE devient alors

$$ISE^0 = \sum_{x \in \aleph} \left[\widehat{f}_{n,h,K}(x) - f_0(x) \right]^2 .$$

Plusieurs modèles de lois discrètes peuvent être considérés pour réaliser ces études. Nous retenons ici la fonction de probabilités

$$f(x) = 0.4\, e^{-0.5}\, 0.5^x/x! + 0.6\, e^{-10}\, 10^x/x!, \quad x \in \mathbb{N},$$

qui est un mélange de deux lois de Poisson de moyennes $\mu_1 = 0.5$ et $\mu_2 = 10$. Cette fonction f a la particularité d'admettre sa plus grande valeur en $x = 0$ avec $f(0) = 0.243$, un minimum local en $x = 3$ où $f(3) = 9.594 \times 10^{-3}$, un maximum local en $x = 9$ où $f(9) = 7.506603 \times 10^{-2}$, et une queue à partir de $x = 22$ tel que $1 - \sum_{x=0}^{21} f(x) = 4.198 \times 10^{-4}$.

6.1 Approximation du risque quadratique intégré

Les Figures 3, 4 et 5 rapportent quelques allures de $AMISE^*(f, f', f'')$, $AMISE^*(f, f')$ et $ISE(f)$ en fonction de la fenêtre h selon la taille d'échantillon $n \in \{50, 100, 300, 1000\}$ et les trois types de noyaux discrets standards. Précisons que $AMISE^*(f, f')$ est obtenu en supprimant les termes en f'' dans $AMISE^*(f, f', f'') := AMISE^*(n, h, f)$.

On remarque alors que les approximations des dérivées de f par des différences finies sont globalement satisfaisantes ; elles sont d'autant meilleures quand la taille d'échantillon augmente. L'effet de réduction de l'ordre d'approximation des différences finies dans $AMISE^*$ semble négligeable pour cette fonction de probabilité f. Les ordres de grandeur des $AMISE^*$ et ISE fournissent des indicateurs corrects du vrai $MISE$; de plus, elles sont comparables à la fois pour une taille d'échantillon donnée et pour un type de noyau discret standard choisi.

Ainsi, pour un même critère de choix de fenêtre de lissage discret, la qualité d'estimation par un noyau binomial est bien meilleure que par un noyau de Poisson et, enfin, par un noyau binomial négatif. Cette comparaison entre les trois types de noyaux discrets standards est confirmée dans d'autres études à travers des données de comptage réelles et simulées. En particulier, pour un échantillon de taille n donnée, 1000 répliques indépendantes f_0 de f sont effectuées et nous avons constaté, entre autre, cette hiérarchie entre les trois types de noyaux discrets standards dans plus de 99% des cas. Ce travail long et fastidieux des répliques des simulations est aussi appliqué pour les observations qui vont suivre.

6.2 Validation croisée et excès de zéros

Sur des données simulées de f de taille $n \in \{50, 100, 300, 1000\}$, nous illustrons entre autre la nouveauté (38) dans les nombreux critères de choix de fenêtres $h > 0$ ainsi que les performances comparatives des estimateurs à noyaux discrets standards.

Les Tables 3 et 4 présentent les différentes qualités de lissage discret par des estimations à noyaux discrets selon la fenêtre adaptée h_0 de proportion de zéros (38), la fenêtre optimale de validation croisée (37) et les fenêtres h_0^{**} et h^{**}. On y insère la constante de normalisation $C = \sum_{x \in \mathbb{N}} \widehat{f}_{n,h,K}(x)$ introduite en (9), laquelle a tendance à surestimer pour ces données ($C > 1$). Puisque la proportion de zéros est assez importante dans les échantillons (ici 24% environ), la qualité de lissage discret par la fenêtre adaptée h_0 est quasiment la meilleure pour chacun des trois types de noyaux discrets standards. Cependant, les autres fenêtres h_{cv} et h_0^{**} sont des alternatives valables pour ces jeux de données. On peut aussi observer le bon comportement de h_0^{**} par rapport à

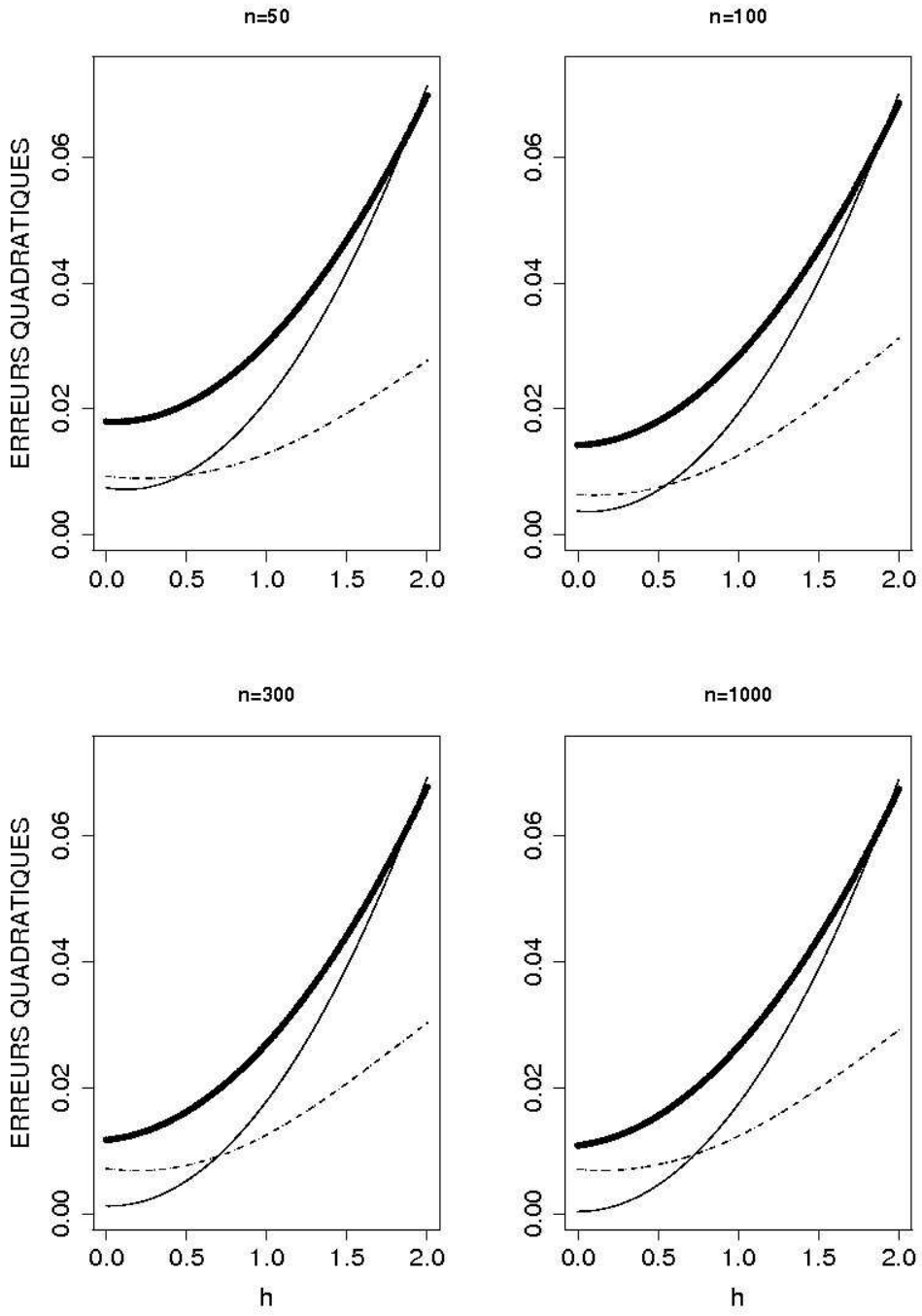

Figure 3: Graphiques de $AMISE^*(f, f', f'')$ [en gras], $AMISE^*(f, f')$ [en fin] et $ISE(f)$ [en pointillé] de l'estimateur à noyau de Poisson pour la distribution du mélange de Poisson $f = 0.4\mathcal{P}(0.5) + 0.6\mathcal{P}(10)$

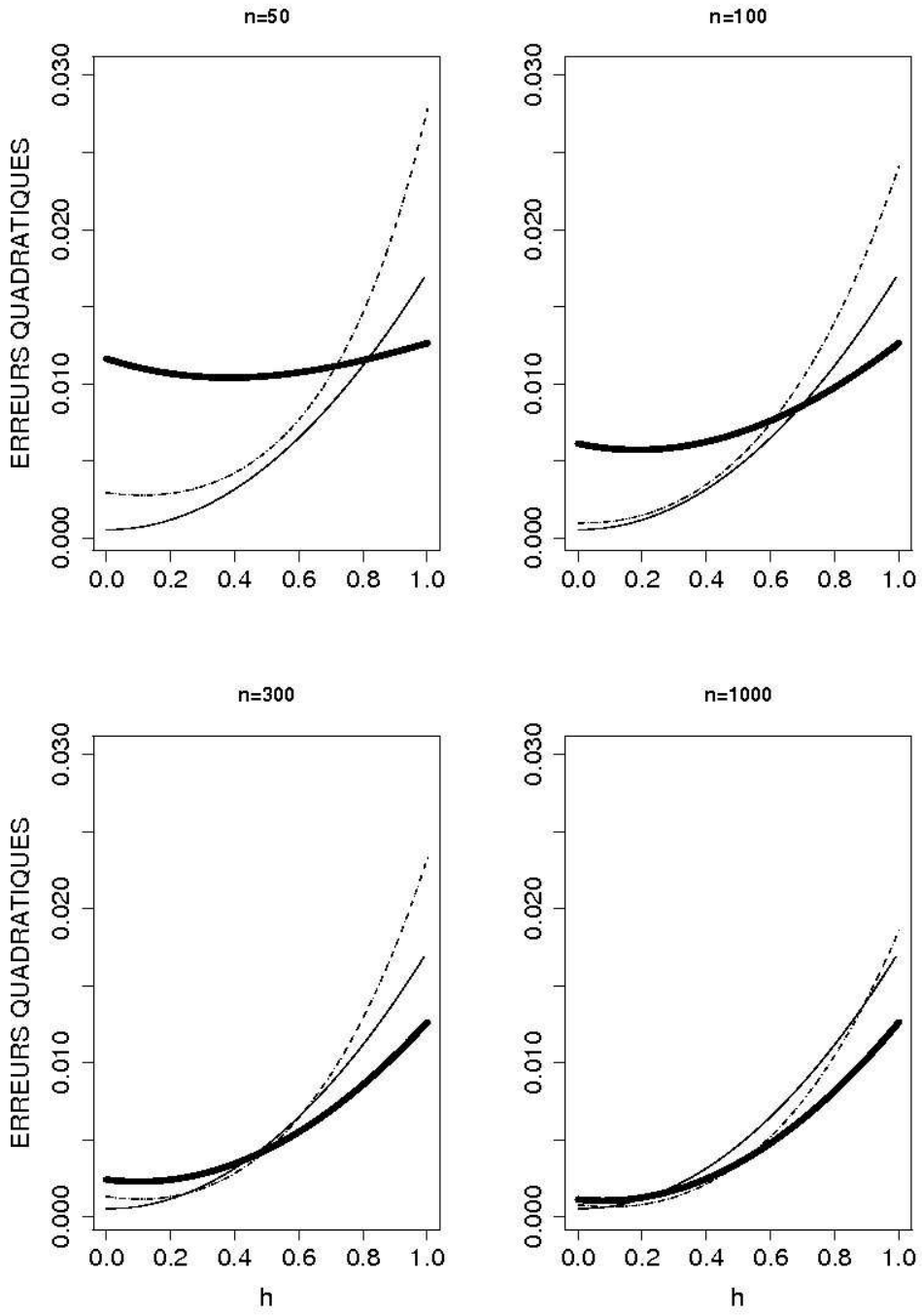

Figure 4: Graphiques de $AMISE^*(f, f', f'')$ [en gras], $AMISE^*(f, f')$ [en fin] et $ISE(f)$ [en pointillé] de l'estimateur à noyau binomial pour la distribution du mélange de Poisson $f = 0.4\mathcal{P}(0.5) + 0.6\mathcal{P}(10)$

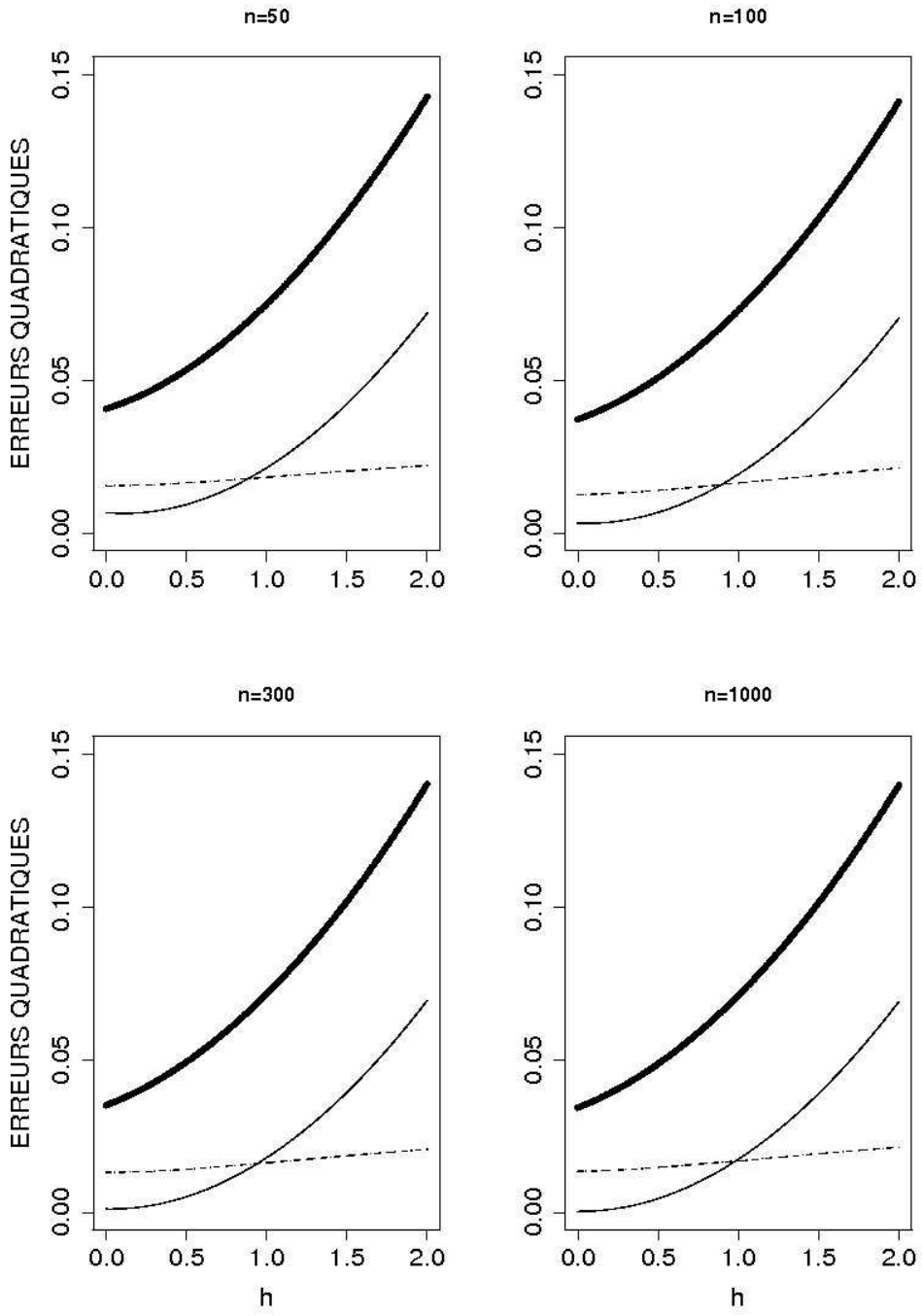

Figure 5: Graphiques de $AMISE^*(f, f', f'')$ [en gras], $AMISE^*(f, f')$ [en fin] et $ISE(f)$ [en pointillé] de l'estimateur à noyau binomial négatif pour la distribution du mélange de Poisson $f = 0.4\mathcal{P}(0.5) + 0.6\mathcal{P}(10)$

h^{**} dans la Table 3. De plus, nous retrouvons l'ordre de préférence parmi ces types de noyaux standards : binomial, Poisson et binomial négatif.

La Figure 6 présente le lissage discret des données simulées par le noyau naïf ($h = 0$). Pour les tailles d'échantillons $n \in \{50, 100, 300\}$, l'ajustement discret de f n'est pas satisfaisant. Ce n'est que dans le cas $n = 1000$ que le lissage discret à l'aide du noyau naïf est entièrement convenable.

Les Figures 7, 8, 9 et 10 représentent graphiquement les lissages discrets de f à travers les noyaux discrets naïf et standards en choisissant les deux fenêtres h_{cv} et h_0. Pour $n = 1000$, de manière générale les lissages discrets sont réguliers en suivant l'allure de f_0 bien que le noyau binomial négatif ne soit pas très performant par rapport aux autres. Dans ce cas, l'estimateur empirique ou naïf réalise le meilleur ajustement. Pour des échantillons de petites et moyennes tailles ($n \in \{50, 100, 300\}$), l'estimateur à noyau binomial devient le plus approprié. Cette dernière constatation est nettement visible pour une petite taille d'échantillon ($n = 50$), situation dans laquelle l'estimateur empirique n'est plus adéquat.

		Binomial	Poisson	Binomial négatif
$n = 50$				
	h_0^{**}	0.010	0.196	0.034
	C	1.02772	1.05812	1.16468
	ISE^0	**0.00897**	0.01486	0.02061
	h_0^{**}	0.010	0.196	0.034
	C	1.02772	1.05812	1.16468
	ISE	**0.00292**	0.00895	0.01566
	h^{**}	0.110	0.232	0.020
	C	0.99987	1.04946	1.16800
	ISE	**0.00278**	0.00895	0.01566
$n = 100$				
	h_0^{**}	0.088	0.176	0.010
	C	1.01280	1.06888	1.17971
	ISE^0	**0.00167**	0.00905	0.01613
	h_0^{**}	0.088	0.176	0.010
	C	1.01280	1.06888	1.17971
	ISE	**0.00105**	0.00627	0.01261
	h^{**}	0.021	0.115	0.010
	C	1.03428	1.08592	1.17971
	ISE	**0.00098**	0.00625	0.01261
$n = 300$				
	h_0^{**}	0.092	0.129	0.010
	C	1.01286	1.09038	1.20003
	ISE^0	**0.00159**	0.00808	0.01429
	h_0^{**}	0.092	0.129	0.010
	C	1.01286	1.09038	1.20003
	ISE	**0.00118**	0.00690	0.01315
	h^{**}	0.104	0.190	0.020
	C	1.00884	1.07240	1.19708
	ISE	**0.00117**	0.00686	0.01315
$n = 1000$				
	h_0^{**}	0.099	0.158	0.001
	C	1.00901	1.08104	1.20242
	ISE^0	**0.00066**	0.007226	0.01391
	h_0^{**}	0.099	0.158	0.001
	C	1.00901	1.08104	1.20242
	ISE	**0.00067**	0.00689	0.01345
	h^{**}	0.100	0.152	0.001
	C	1.00870	1.08272	1.20242
	ISE	**0.00067**	0.00689	0.01345

Table 3: Qualités de lissages discrets par les noyaux de type binomial, Poisson et binomial négatif des données simulées de la distribution du mélange de Poisson $f = 0.4\mathcal{P}(0.5) + 0.6\mathcal{P}(10)$

		Naïf	Binomial	Poisson	Binomial négatif
$n = 50$					
	h_{cv}		0.150	0.390	0.510
	C	1.00000	0.98892	1.01263	1.06201
	ISE	0.01001	**0.00280**	0.00911	0.01658
	$AMISE^*(f)$	0.02000	**0.01084**	0.01954	0.05390
	h_0		0.126	0.235	0.371
	C	1.00000	0.99548	1.04874	1.08966
	ISE	0.01001	**0.00278**	0.00895	0.01618
	$AMISE^*(f)$	0.02000	**0.01094**	0.01841	0.04940
$n = 100$					
	h_{cv}		0.150	0.250	0.360
	C	1.00000	0.99330	1.04887	1.09198
	ISE	0.00224	**0.00124**	0.00639	0.01343
	$AMISE^*(f)$	0.01000	**0.00574**	0.01529	0.04630
	h_0		0.105	0.189	0.287
	C	1.00000	1.00742	1.06531	1.10869
	ISE	0.00224	**0.00110**	0.00629	0.01319
	$AMISE^*(f)$	0.01000	**0.00580**	0.01488	0.04413
$n = 300$					
	h_{cv}		0.090	0.190	0.300
	C	1.00000	1.01354	1.07240	1.12089
	ISE	0.00200	**0.00118**	0.00685	0.01353
	$AMISE^*(f)$	0.00333	**0.00231**	0.01267	0.04260
	h_0		0.102	0.179	0.267
	C	1.00000	1.00951	1.07561	1.12920
	ISE	0.00200	**0.00117**	0.00686	0.01345
	$AMISE^*(f)$	0.00333	**0.00230**	0.01259	0.04163
$n = 1000$					
	h_{cv}		0.070	0.180	0.300
	C	1.00000	1.01820	1.07491	1.12457
	ISE	**0.00020**	0.00068	0.00690	0.01406
	$AMISE^*(f)$	**0.00100**	0.00107	0.01190	0.04193
	h_0		0.111	0.191	0.292
	C	1.00000	1.00523	1.07187	1.12647
	ISE	**0.00020**	0.00067	0.00690	0.01404
	$AMISE^*(f)$	**0.00100**	0.00109	0.01182	0.04169

Table 4: Qualités de lissages discrets par les noyaux de type naïf, binomial, Poisson et binomial négatif des données simulées de la distribution du mélange de Poisson $f = 0.4\mathcal{P}(0.5) + 0.6\mathcal{P}(10)$

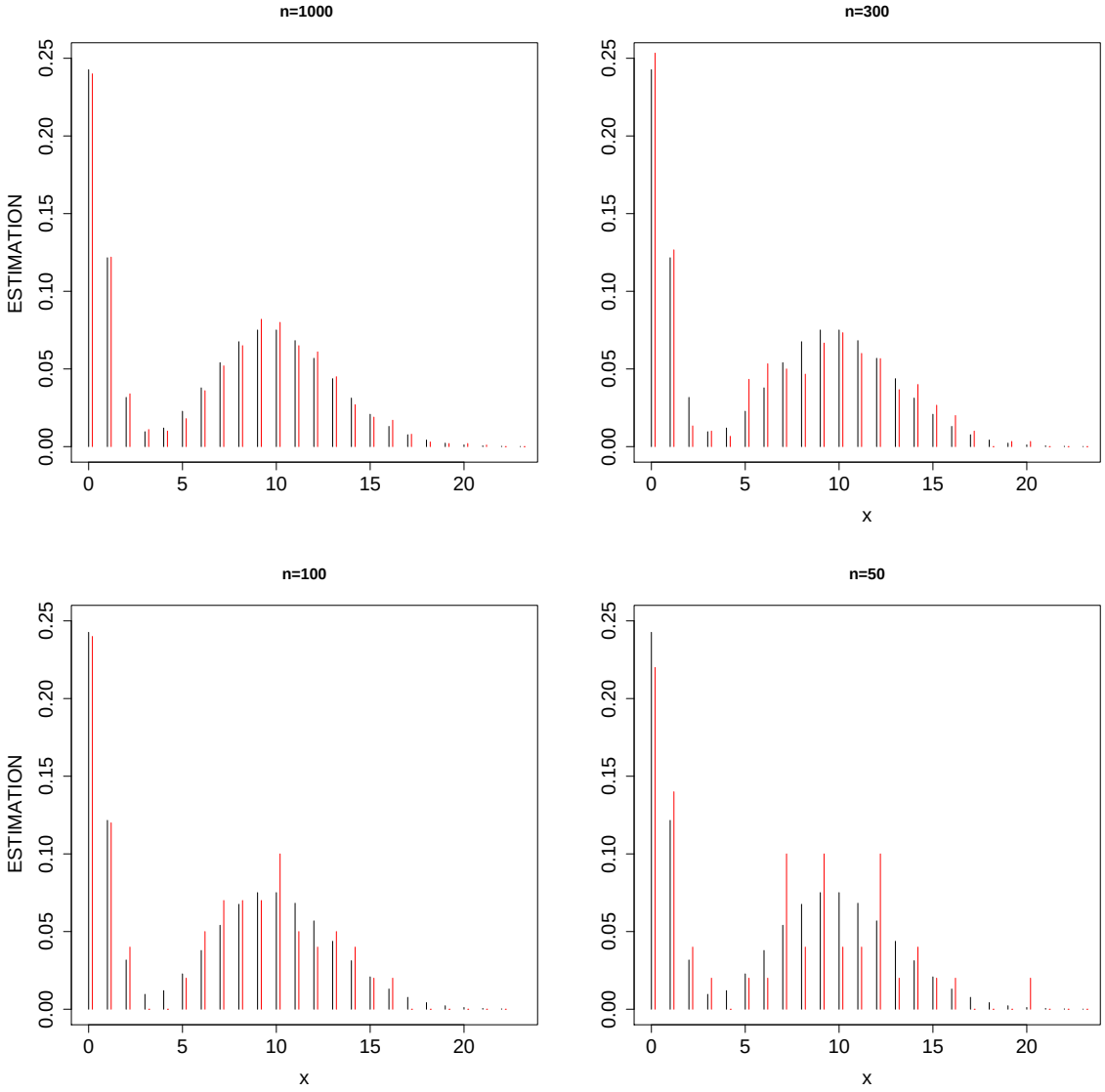

Figure 6: Lissages discrets par un estimateur empirique ("naïf") des données simulées pour $n \in \{50, 100, 300, 1000\}$ de la distribution du mélange de Poisson $f = 0.4\mathcal{P}(0.5) + 0.6\mathcal{P}(10)$

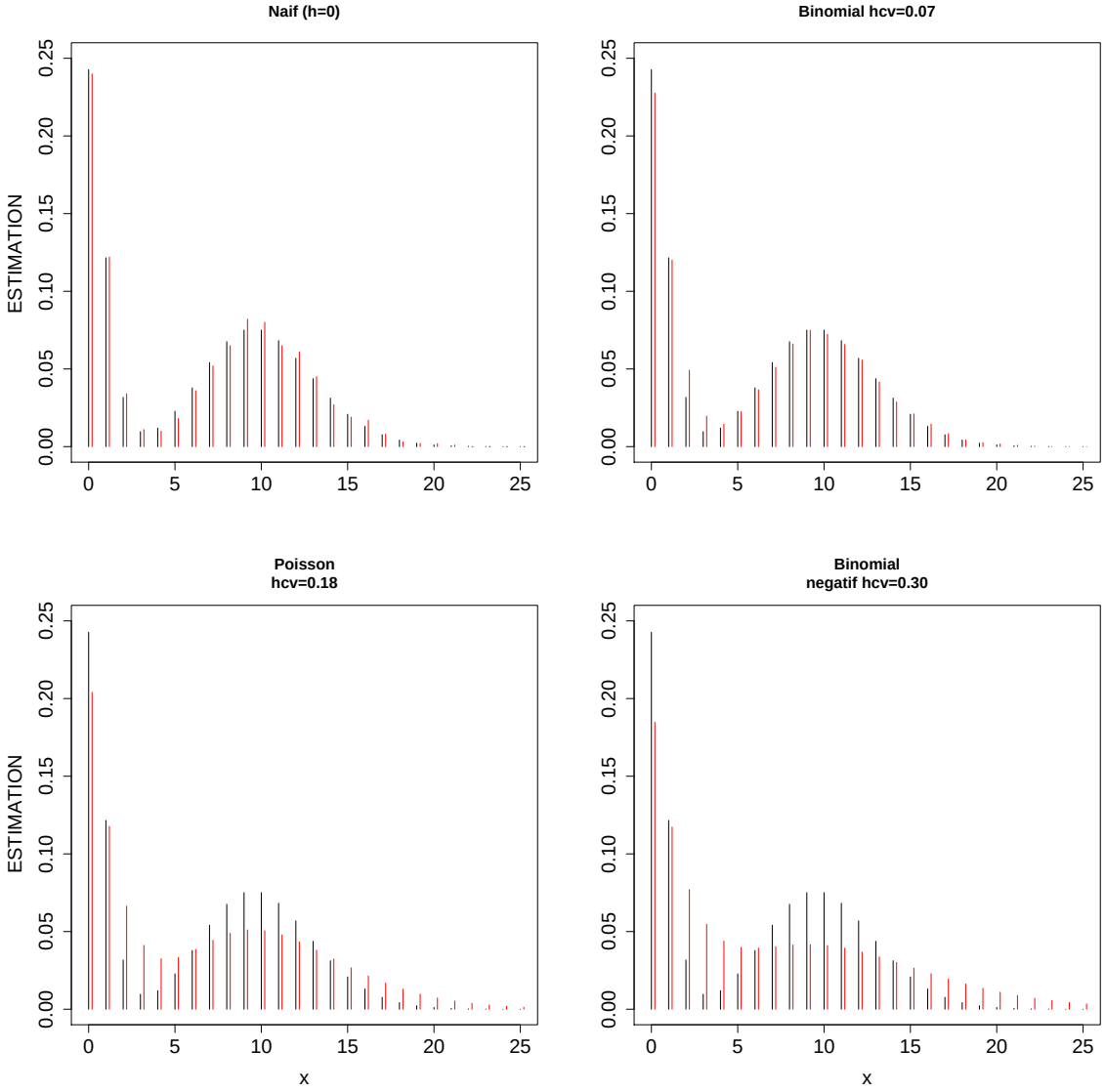

Figure 7: Lissages discrets par les noyaux de type naïf, binomial, Poisson et binomial négatif des données simulées ($n = 1000$) de la distribution du mélange de Poisson $f = 0.4\mathcal{P}(0.5) + 0.6\mathcal{P}(10)$

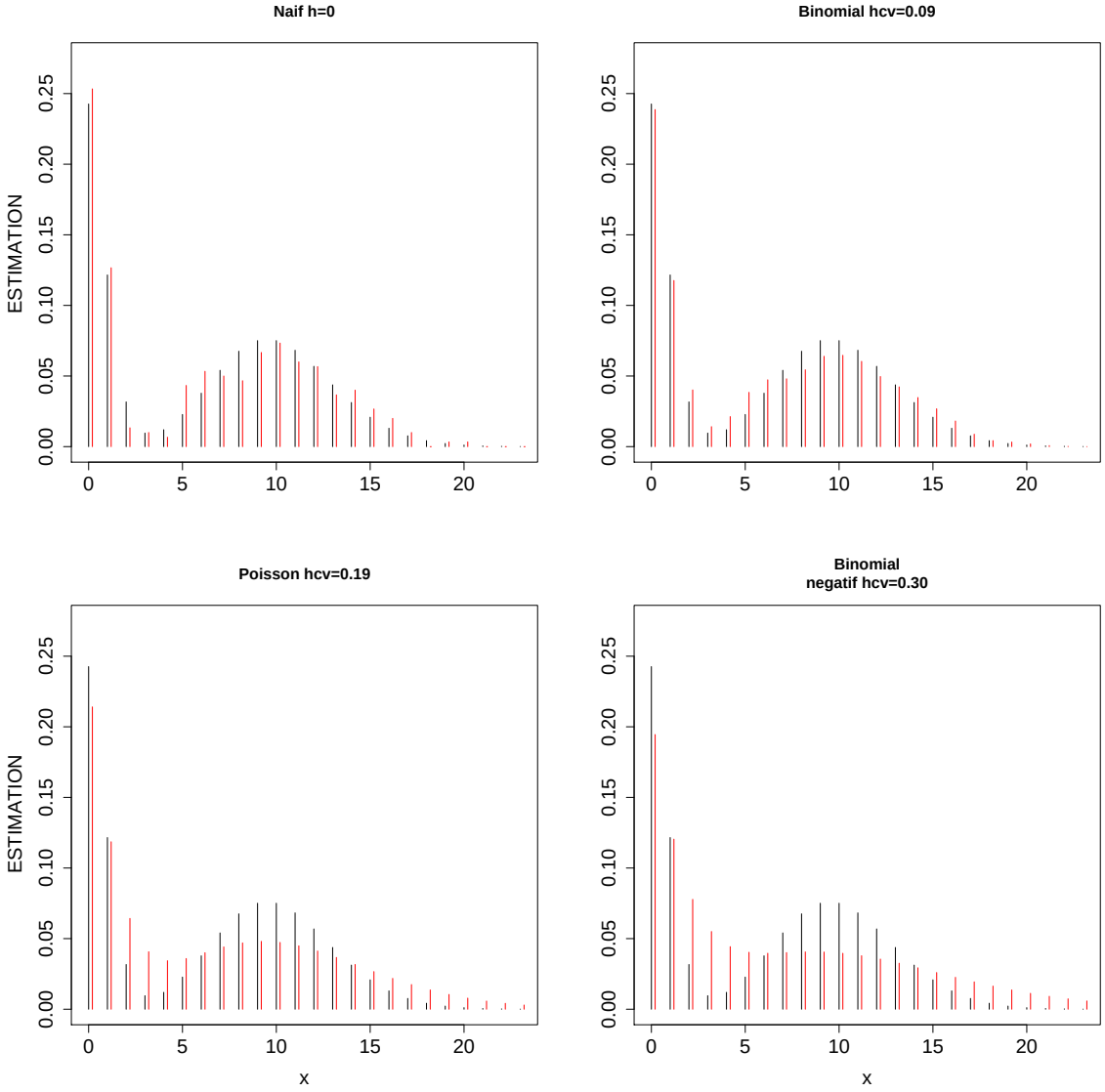

Figure 8: Lissages discrets par les noyaux de type naïf, binomial, Poisson et binomial négatif des données simulées ($n = 300$) de la distribution du mélange de Poisson $f = 0.4\mathcal{P}(0.5) + 0.6\mathcal{P}(10)$

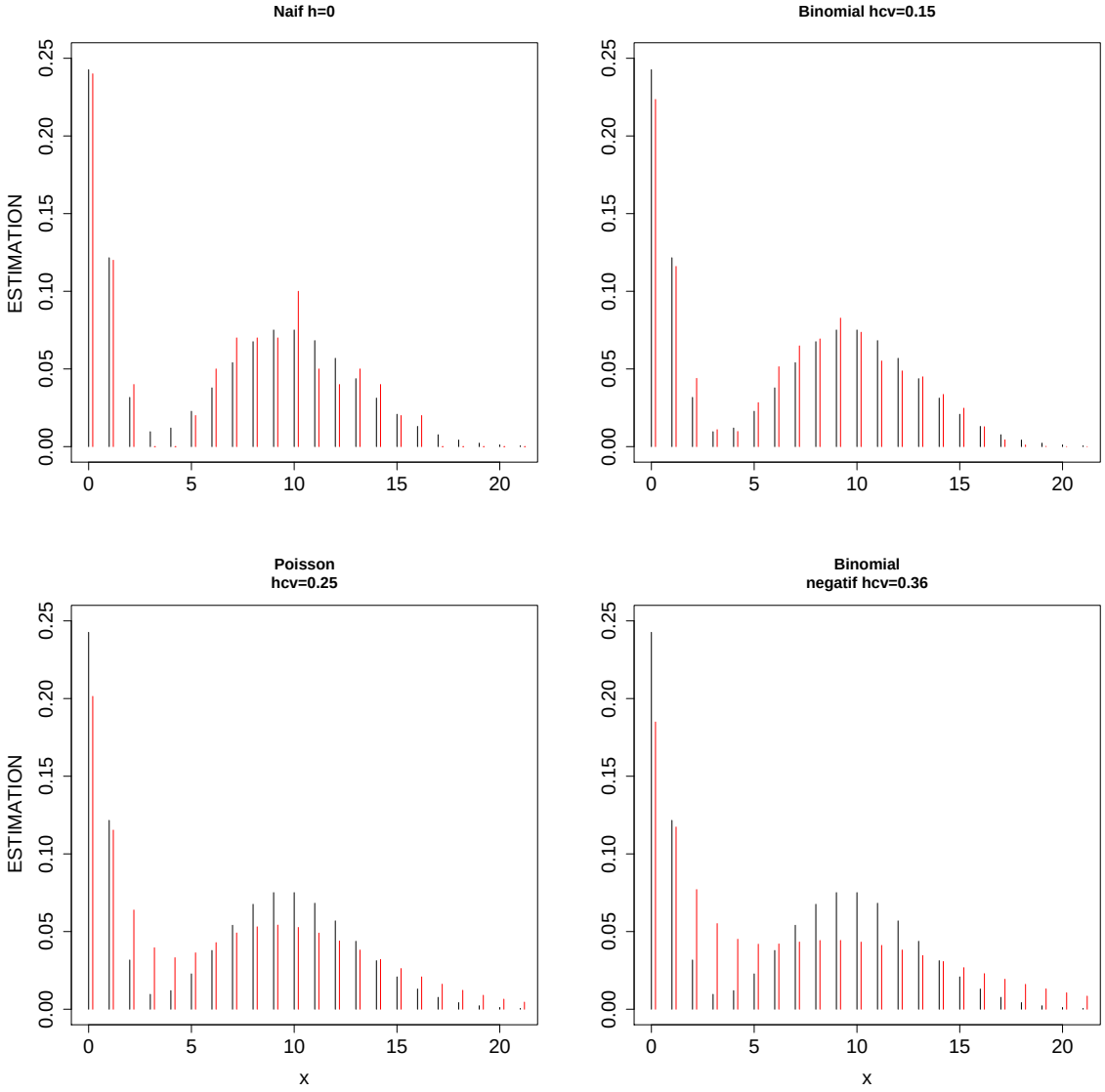

Figure 9: Lissages discrets par les noyaux de type naïf, binomial, Poisson et binomial négatif des données simulées ($n = 100$) de la distribution du mélange de Poisson $f = 0.4\mathcal{P}(0.5) + 0.6\mathcal{P}(10)$

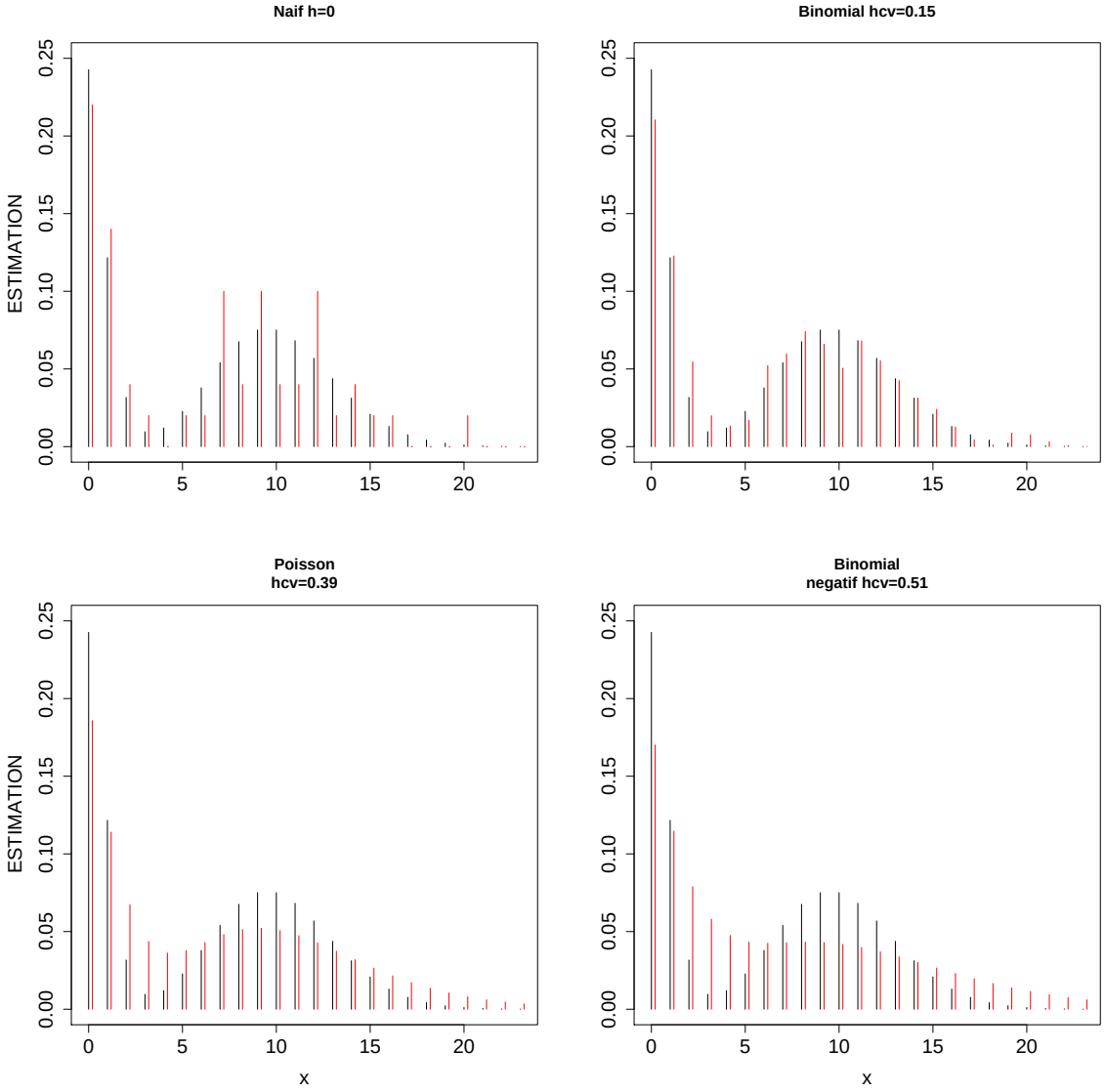

Figure 10: Lissages discrets par les noyaux de type naïf, binomial, Poisson et binomial négatif des données simulées ($n = 50$) de la distribution du mélange de Poisson $f = 0.4\mathcal{P}(0.5) + 0.6\mathcal{P}(10)$

7 Application

Dans la nature, il existe de nombreux exemples de données de comptage provenant de domaines très divers comme l'agriculture, l'économie, la médecine, l'assurance, le sport, etc. La diversité des données ouvre un large champ d'application des estimateurs à noyau discret et renforce l'intérêt de ce travail. Nous appliquons maintenant ces estimateurs à noyaux discrets standards aux lissages discrets de nombre (ou fréquence) de buts par matchs d'un championnat de football en fonction des buts marqués $(0, 1, 2, \cdots)$. Les observations relatives aux résultats numériques ci-dessous complètent celles des données simulées du précédent paragraphe. Notons au passage qu'on peut se reférer à Karlis & Ntzoufras (2003) pour des modèles paramétriques relatifs à l'influence des buts marqués par les équipes durant une rencontre.

7.1 Données

Nous considérons les données de la Table 5 décrivant les buts marqués par match dans chacun des championnats de football français (ou Ligue 1) et espagnol (ou Liga) pour la saison 2005-2006. Nous avons un nombre total de $n = 380$ matchs. En comparant ces deux championnats, on peut alors justifier le nouveau classement 2006-2007 de la Ligue Professionnelle de Football en France, appelé classement de l'offensive[2]. Ce classement qui pourrait être généralisé à des matchs de poules de la Coupe du Monde ou de la Champions League a pour objectif de réhausser le nombre de buts marqués et donc d'améliorer le spectacle.

Buts (g)	0	1	2	3	4	5	6	7	8	9	Total
Ligue 1	51	90	109	61	44	12	9	3	0	1	380
Liga	27	73	116	83	44	25	6	5	1	0	380
Ligue 1 − Liga	24	17	−7	−22	0	−13	3	−2	−1	1	0

Table 5: Données du nombre de buts par match des championnats de football de Ligue 1 française et de Liga espagnole pour la saison 2005-2006 avec $n = 380$ rencontres

	Total de buts ($n\overline{g}$)	$\overline{g}$	s_g^2	$s_g^2/\overline{g}$
Ligue 1	811	2.134	2.375	1.113
Liga	934	2.458	2.222	0.904

Table 6: Résumé des statistiques de la Table 5 où $\overline{g}$ est la moyenne de buts par match, s_g^2 et $s_g^2/\overline{g}$ sont respectivement la variance et l'indice de dispersion de Fisher associé (*e.g.* Mizère *et al.*, 2006)

En fait, d'après aussi la Table 6, la Ligue 1 française possède un déficit de 123 buts par rapport à la Liga espagnole. On remarque également que les données de buts de Ligue 1 sont surdispersées ($s_g^2/\overline{g} = 1.113 > 1$) alors que celles de Liga sont sousdispersées ($s_g^2/\overline{g} = 0.904 < 1$) ; voir, par exemple, Kokonendji *et al.* (2006b) et Mizère *et al.* (2006). En représentant graphiquement les distributions empiriques f_{01} et f_{02} des fréquences des matchs de Ligue 1 et de Liga, respectivement, en fonction de buts marqués, on peut constater que f_{01} est moins régulière que f_{02}. Enfin, la proportion $f_{01}(0) = 0.134$ des zéros ou des matchs nuls sans but (0-0) de Ligue 1

[2]Ce classement encore parallèle à l'officiel accorde 3 points pour une victoire par plus d'un but d'écart, 2 points pour un succès par un but d'écart et 1 point pour un match nul (avec ou sans but) : www.lpf.fr/ligue1/classementOffensive.asp

est presque le double de celle de Liga $f_{02}(0) = 0.071$. Dans la suite nous donnons uniquement les résultats relatifs au lissage discret des données de la Ligue 1 française.

7.2 Résultats

Les Table 7 et Figure 11 affichent les résultats numériques et graphiques de différents lissages discrets par les noyaux discrets standards de la distribution empirique f_{01} des fréquences de matchs de Ligue 1 en fonction de buts marqués. Malgré une proportion non-négligeable de zéros (matchs nuls sans aucun but) dans ce jeux de données de taille moyenne $n = 380$, nous avons présenté les résultats avec la fenêtre h_0 de (38) ainsi que les fenêtres h_{cv} de validation croisée (37) et h_0^{**} définie en (34).

	Poisson	Binomial	Binomial négatif
h_0	0.701	0.268	1.598
C	0.97266	0.95042	0.88442
ISE^0	0.01789	**0.00515**	0.02837
h_{cv}	0.054	0.177	0.039
C	1.05082	0.95872	1.12028
ISE^0	0.01580	**0.00395**	0.02904
h_0^{**}	0.241	0.054	0.606
C	1.03322	0.97279	1.05378
ISE^0	0.01522	**0.00337**	0.02781

Table 7: Qualités de lissages discrets par les trois types de noyaux discrets standards pour les données réelles de football de Ligue 1 française avec $n = 380$

Les meilleurs ajustements de f_{01} au sens de l'ISE^0 minimal sont obtenus une fois de plus par le noyau binomial. En particulier, le noyau binomial associé à la fenêtre h_0^{**} est préférable à son association aux fenêtres h_{cv} puis h_0. Pour ce jeu de données, le noyau binomial a tendance à sousestimer ($C < 1$) alors que les noyaux de Poisson et binomial négatif sousestiment quand ils sont associés à h_0 et surestiment lorsqu'ils sont associés à h_{cv} et h_0^{**}.

Avec le meilleur modèle non-paramétrique $\widehat{f}_{n,h,K}$ de chaque championnat, on peut alors estimer le nombre $n \times \widehat{f}_{n,h,K}(g)$ de rencontres étant soldées par un nombre g donné de buts dans une saison régulière à n matchs.

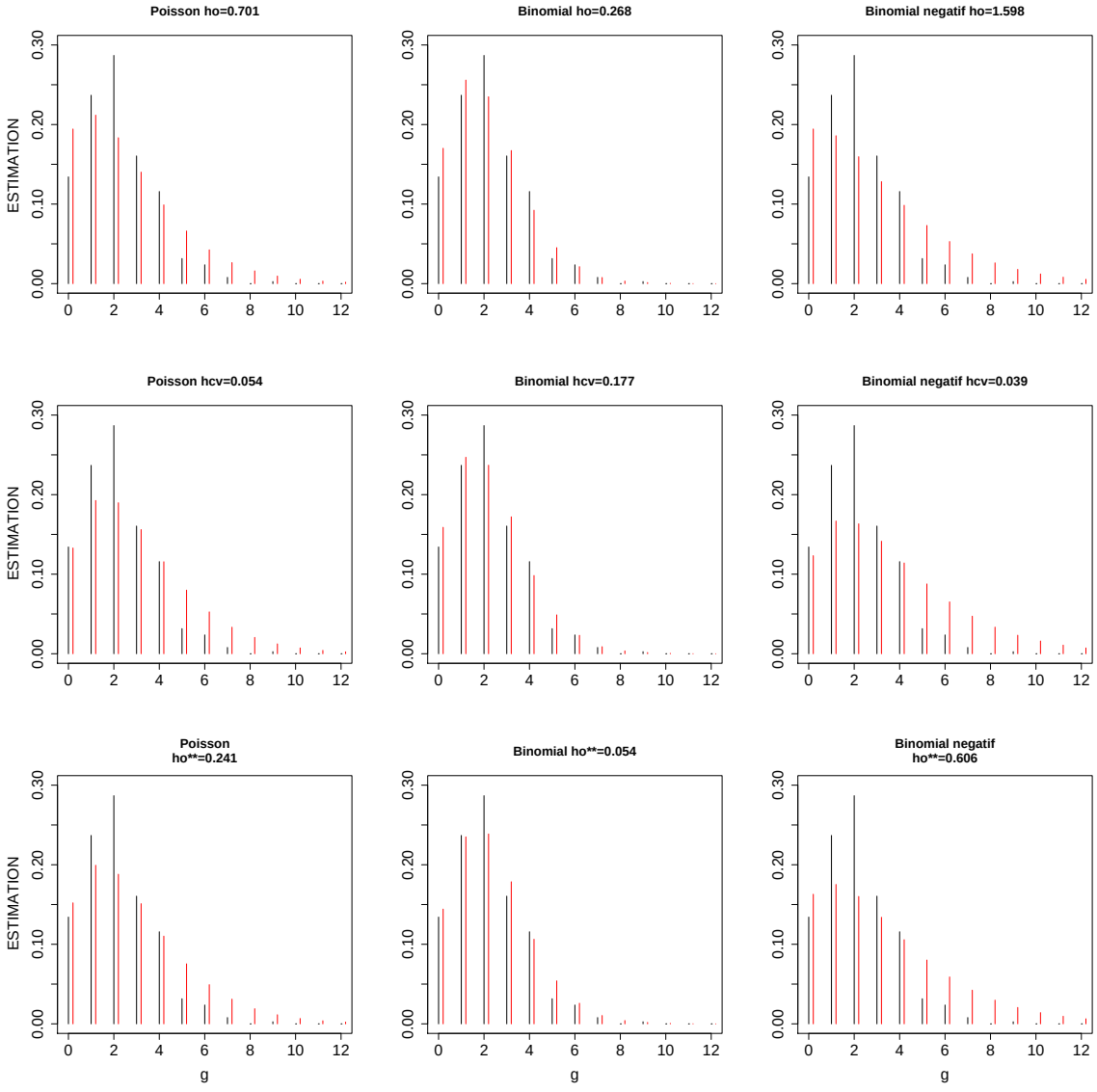

Figure 11: Lissages discrets par les noyaux de type de Poisson, binomial et binomial négatif pour les données réelles de football de Ligue 1 française $n = 380$

Remerciements

Nous remercions Laurent Bordes, Frédéric Ferraty, Pascal Sarda et Philippe Vieu pour les premiers commentaires sur ce travail.

Références

BERLINET A., BIAU G. (2002), Estimation de densité et prise de décision, in *Décision et Reconnaissance de Formes en Signal*, ed. R. Lengellé, pp.141–179, Hermès, Paris.

BOWMAN A. (1984), An alternative method of cross-validation for the smoothing of density estimates, *Biometrika* 71, 352-360.

CHEN S.X. (1999), Beta kernels estimators for density functions, *Computat. Statist. Data Anal.* 31,131-145.

CHEN S.X. (2000), Gamma kernel estimators for density functions, *Ann. Instit. Statist. Math.* 52, 471-480.

DEVROYE L. (1987), *A Course in Density Estimation,* Birkhäuser, Boston.

FERRATY F., VIEU P. (2006), *Nonparametric Functional Data Analysis: Theory and Practice,* Springer, Berlin.

GREENWOOD, M., YULE G.U., (1920), An inquiry into the nature of frequency distributions representative of multiple happenings with particular referee to the occurrence of multiple attacks of disease or of repeated accidents, *J. R. Statist. Soc. Ser. A* 83, 255-279.

JOHNSON N.L. , KEMP A.W., KOTZ S. (2005), *Univariate Discrete Distributions*, Third Edition, Wiley, New York.

KARLIS D., NTZOUFRAS L. (2003), Analysis of sport data by using bivariate Poisson models, *The Statistician* 52, 381-393.

KOKONENDJI C.C., DEMÉTRIO C.G.B, ZOCCHI S.S. (2007), On Hinde-Demétrio regression models for oversdispersed count data. *Statistical Methodology* 4, 277-291.

KOKONENDJI C.C., MIZÈRE D., BALAKRISHNAN N. (2007b), Connections of the Poisson weight function to overdispersion and underdispersion, *J. Statist. Plann. Inference* [sous presse].

MANGASARIAN O.L., SCHUMAKER L.L. (1973), Best Summation Formulae and Discrete Splines, *SIAM J. Numer. Anal.* 10, 448-459.

MARRON J.S. (1987), A comparison of cross-validation techniques in density estimation, *Ann. Statist.* 15, 152-162.

MARSH L.C., MUKHOPADHYAY K. (1999), Discrete Poisson kernel density estimation with an application to wildcat coal strikes, *Applied Economics Letters* 6, 393-396.

MIZÈRE D., KOKONENDJI C.C., DOSSOU-GBÉTÉ S. (2006), Quelques tests de la loi de Poisson contre des alternatives générales basés sur l'indice de dispersion de Fisher, *Rev. Statistique Appliquée* LIV (4), 61-84.

PARZEN E., (1962), On estimation of a probability density function and mode, *Ann. Math. Statist.* 33, 1065-1076.

ROSENBLATT M., (1956), Remarks on some nonparametric estimates of a density function, *Ann. Math. Statist.* 27, 832-837.

RUDEMO M., (1982), Empirical choice of histograms and kernel density estimators, *Scandinavian J. Statist.* 9, 65-78.

SCAILLET, O., (2004), Density estimation using inverse and reciprocal inverse Gaussian kernels, *Journal of Nonparametric Statistics* 16, 217-226.

SIMONOFF J.S. (1996), *Smoothing Methods in Statistics*, Springer, New York.

SIMONOFF J.S., TUTZ G. (2000), Smoothing methods for discrete data, in *Smoothing and Regression: Approaches, Computation, and Application,* Ed. M.G. Schimek, pp.193-228, Wiley, New York.

SCOTT D.W. (1992), *Multivariate Density Estimation - Theory, Practice, and Visualization,* Wiley, New York.

SILVERMAN B.W. (1986), *Density Estimation for Statistics and Data Analysis,* Chapman & Hall, London.

STONE C.J. (1984), An asymptotically optimal window selection rule for kernel density estimates, *Ann. Statist.* 12, 1285-1297.

TSYBAKOV A.B. (2004), *Introduction à l'Estimation Non-Paramétrique*, Springer, Paris.