Approche non-paramétrique par noyaux associés discrets des données de dénombrement

Tristan Senga Kiessé

ACADÉMIE DE BORDEAUX |_|_|_|_|_|_|_|_|_|_|

$\mathcal{T}$ $\mathcal{H}$ $\mathcal{È}$ $\mathcal{S}$ $\mathcal{E}$

présentée à

L'UNIVERSITÉ DE PAU ET DES PAYS DE L'ADOUR

ÉCOLE DOCTORALE DES SCIENCES ET DE LEURS APPLICATIONS

par

$\mathcal{T}$**ristan** $\mathcal{S}$**ENGA** $\mathcal{K}$**IESSÉ**

pour obtenir le grade de

DOCTEUR D'UNIVERSITÉ

Spécialité : **MATHÉMATIQUES APPLIQUÉES** (option : **Statistique**)

sous la direction de

M. $\mathcal{C}$**élestin** $\mathcal{C}$**.** $\mathcal{K}$**OKONENDJI, HDR**

Approche non-paramétrique par noyaux associés discrets des données de dénombrement

Soutenue **le 15 Octobre 2008**

Après avis des Rapporteurs :

- $\mathcal{B}$elkacem $\mathcal{A}$BDOUS, Professeur Titulaire à l'Université Laval, Canada
- $\mathcal{H}$ervé $\mathcal{C}$ARDOT, Professeur à l'Université de Bourgogne, France

Devant le Jury composé des Rapporteurs, du Directeur de thèse et des Messieurs :

- $\mathcal{F}$lorin $\mathcal{A}$VRAM, Professeur à l'Université de Pau et des Pays de l'Adour, France
- $\mathcal{B}$ernard $\mathcal{G}$AREL, Professeur à l'INP-ENSEEIHT de Toulouse, France
- $\mathcal{P}$ascal $\mathcal{S}$ARDA, Professeur à l'Université Toulouse-Le-Mirail, France

Remerciements

Je remercie ici toutes les personnes qui ont contribué à la réalisation de cette thèse.

Je remercie très sincèrement chaque membre du Jury d'avoir accepté de juger mon travail : Monsieur F. AVRAM que j'ai eu l'honneur de cotoyer ; Monsieur B. GAREL pour les chaleureux échanges au cours du 2ème congrès des Jeunes Statisticiens à Aussois ; Monsieur P. SARDA qui m'a fait l'honneur de prendre part à ce Jury.

Je veux remercier aussi chacun des rapporteurs qui ont écrit des rapports encourageants : Monsieur B. ABDOUS pour ses remarques et propositions intéressantes, et Monsieur H. CARDOT pour ses précieuses orientations.

Je tiens à rendre hommage à Monsieur C. C. KOKONENDJI, mon directeur de thèse, qui m'a accompagné dans mes premiers pas de chercheur avec détermination et amabilité. Il a toujours été disponible, m'a constamment soutenu et encouragé à donner le meilleur de moi tout au long de ces années de thèse.

Le Laboratoire de Mathématiques et de leurs Applications de Pau m'a mis dans les meilleures conditions de travail, grâce à Monsieur M. AMARA qui en était le directeur à mon arrivée ; grâce aussi à Monsieur L. BORDES, directeur depuis cette année ; grâce enfin aux autres chercheurs, en particulier de l'Équipe de Statistique et Probabilités, et à l'ensemble du département de Mathématiques Enseignement. Que Monsieur S. DOSSOU-GBÉTÉ et Monsieur D. MIZÈRE soient également remerciés.

Merci à toutes celles et à tous ceux qui m'ont fidèlement soutenu par leurs prières, leur amitié et leur affection. La République du Congo m'a accordé le financement pour cette thèse au travers de l'Office de Gestion des Étudiants et Stagiaires Congolais.

Merci du fond du coeur à ma famille : mes frères et soeurs et leurs familles respectives. Merci Papa, merci Maman, vous avez «instruit l'enfant dans la voie qu'il doit suivre ; afin que quand il sera grand, il ne s'en détourne pas» (Proverbes 22 : 6).

Article publié

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

Résumé : Nous introduisons une nouvelle approche non-paramétrique, par noyaux associés discrets, pour les données de dénombrement. Pour cela, nous définissons la notion de noyaux associés discrets à partir d'une loi de probabilité discrète donnée et nous étudions leurs propriétés. De là, nous construisons l'estimateur à noyau discret lequel est l'analogue de certains estimateurs à noyau continu de cette dernière décennie. Nous examinons ses propriétés fondamentales ; en particulier, nous montrons la convergence ponctuelle en moyenne quadratique de l'estimateur. Le choix de fenêtre du lissage discret s'effectue essentiellement par validation croisée et excès de zéros. Nous étudions également le comportement des lois classiques de dénombrement comme noyau associé, par exemple, Poisson, binomiale et binomiale négative. Ainsi, il s'est révélé nécessaire de construire une nouvelle famille de lois discrètes dites triangulaires pour servir de noyaux associés symétriques. Cette méthode des noyaux associés discrets est utilisée dans l'estimation semi-paramétrique des distributions de données de dénombrement, ainsi que pour la régression non-paramétrique sur une variable explicative de dénombrement. Tout au long de ce travail, nous illustrons les résultats à travers des simulations et des jeux de données réelles. Dans le cas d'échantillons de tailles petites et modérées, l'importance et les très bonnes performances des noyaux associés discrets sont mises en évidence, en comparaison avec le noyau du type Dirac et parfois les noyaux continus.

Mots clés : Biais de bordure, différence finie, estimation non-paramétrique, noyau variable, loi discrète, loi triangulaire discrète, noyau asymétrique, proportion de zéros, régression non-paramétrique, risque quadratique intégré, validation croisée.

Nonparametric approach by discrete associated-kernel for count data

Abstract : This work introduces a new nonparametric approach by discrete associated-kernels for count data. First, we define the discrete kernel associated to a discrete probability distribution and we examine its basical properties. Furthermore, we construct the discrete associated-kernel estimator which is the analog of some one in the continuous case of the last decade. We investigate their properties ; in particular, we show the pointwise convergence of the estimator in the sense of mean squared error. The choice of bandwidth is mainly done through cross-validation and excess of zeros. For illustrating, we study some discrete probability distributions such that Poisson, binomial, negative binomial, that we consider as associated-kernels. Thus, we need to improve it by introducing a new discrete probability distribution, called triangular, in order to serve as symmetric associated-kernel. The discrete associated-kernel method is then used for a semiparametric estimation of count distributions and, also, for nonparametric regression on a count explanatory variable. This discrete associated-kernel method is illustrated through simulations and real examples of count data. For a sample size not so large, the importance and the performance of discrete associated-kernels are pointed out compared with the Dirac type kernel and, sometimes, the continuous ones.

Key words : Asymmetric kernel, boundary bias, cross-validation, discrete distribution, discrete triangular distribution, finite difference, mean integrated squared error, nonparametric estimation, nonparametric regression, variable kernel, zero-proportion.

Table des matières

Liste des tableaux

Table des figures

Introduction générale

Nous présentons ici les motivations à l'origine de ce travail. Par la suite, nous donnons le contenu de la thèse.

Motivations

L'approche traditionnelle pour estimer une distribution de données de dénombrement ou de comptage a été jusqu'à récemment entièrement paramétrique. En effet, dans cette approche, la structure de distribution discrète est précisée dès le départ : tel que le modèle de Poisson. Elle peut être, ensuite, modifiée (par exemple, par pondération, mixturisation, lagrangénisation) en une autre appartenant à une famille de lois discrètes ayant le même support. Cela permet de tenir compte des phénomènes spécifiques des données de dénombrement. On peut se référer, par exemple, à Greenwood & Yule (1920), Gupta & Ong (2005), Johnson *et al.* (2005), Kokonendji *et al.* (2007a), Shmueli *et al.* (2005), Kokonendji *et al.* (2008), Mizère *et al.* (2006) et leurs références. Parallèlement à la loi normale pour les données continues, la loi de Poisson est la distribution standard pour l'analyse des données de comptage bien qu'elle ait un seul paramètre qui fait aussi office de paramètre de dispersion. Il s'est révélé nécessaire, dans plusieurs situations, de construire des lois discrètes adaptées aux phénomènes étudiés en utilisant des critères pour détecter des écarts par rapport à une loi de Poisson. En pratique, pour de nombreuses données de comptage, il est commun d'avoir la variance égale, plus petite ou plus grande que la moyenne. Ceci correspond aux phénomènes d'*équidispersion*, de *sousdispersion* ou de *surdispersion*, respectivement. Les écarts par rapport à une loi de Poisson les plus étudiés sont d'une part la surdispersion et d'autre part l'*excès de zéros*. Les phénomènes opposés tels que la sousdispersion et le *défaut de zéros* sont aussi possibles, mais moins communs.

Précisons ici, par rapport à la loi de Poisson, les notions de (sur/sous) dispersion et de (excès/défaut) proportion de zéros pour une loi discrète sur l'ensemble des entiers naturels $\mathbb{N}$ de moyenne $\mu > 0$, de variance $\sigma^2 > 0$ et de proportion de zéros $p_0 > 0$. Nous disons qu'une telle loi discrète est surdispersée (respectivement sousdispersée) si sa variance σ^2 est supérieure (respectivement inférieure) à la variance μ de la loi de Poisson. De là, l'indice de dispersion peut être défini par $D = (\sigma^2 - \mu)/\mu$. De manière similaire, une loi de dénombrement est dite en excès de zéros (respectivement

en défaut de zéros) si sa proportion de zéros est plus importante (respectivement moins importante) que la proportion de zéros de la loi de Poisson de même moyenne μ, c'est-à-dire $\exp(-\mu)$. L'indice de proportion de zéros de cette loi est alors défini par $Z = 1 + \log(p_0)/\mu$. Ces deux indices sont nuls pour la loi de Poisson. L'indice D est positif dans le cas surdispersé et négatif dans la cas sousdispersé. L'indice Z est positif pour la situation d'excès de zéros et négatif pour celle de défaut de zéros. Ces situations de (sur/sous) dispersion et de (excès/défaut) proportion de zéros peuvent se retrouver dans une même loi de dénombrement mais elles ont des effets indépendants ; voir, par exemple, Puig (2003), Puig & Valero (2006) et Nikoloulopoulos & Karlis (2008).

Une procédure plus générale pour modifier la loi de Poisson est de la multiplier par une fonction poids ; voir Kokonendji *et al.* (2008) et leurs références ainsi que Balakrishnan & Kozubowski (2008). Les *lois de Poisson pondérées* permettent de prendre en compte, entre autres, les phénomènes de dispersion et de proportion de zéros. Il s'agit là d'un outil de choix d'un modèle approprié pour des données de dénombrement observées sans un cadre propre. Nous donnons maintenant la définition des lois de Poisson pondérées. Soit X une variable aléatoire de Poisson standard de probabilité individuelle $p(x;\theta) = \Pr(X = x)$, où $\theta \in \mathbb{R}$ est le paramètre canonique et $x \in \mathbb{N}$, et soit $w(x)$ une fonction positive sur $\mathbb{N}$. La nouvelle variable aléatoire X^ω de probabilité individuelle $p_\omega(x;\theta) = \Pr(X^\omega = x)$ est dite *version pondérée de* X si sa fonction de masse de probabilité est donnée par :

$$p_\omega(x;\theta) = \frac{\omega(x)\, p(x;\theta)}{\sum_{x\in\mathbb{N}} \omega(x) p(x;\theta)}, \quad x \in \mathbb{N}, \tag{1}$$

où le dénominateur est la constante de normalisation dépendant de θ. De plus, à partir de la relation (1), il est clair que $0 < \sum_{x\in\mathbb{N}} \omega(x)p(x;\theta) < \infty$. La fonction poids peut dépendre d'un paramètre ϕ tel que $\omega(x) \equiv \omega(x;\phi)$. Ce paramètre ϕ, qui peut être lié au paramètre θ de Poisson, représente le mécanisme d'enregistrement du phénomène de l'écart par rapport à une loi de Poisson. Ainsi, toute loi de dénombrement peut être écrite comme une loi de Poisson pondérée. Il vient clairement que la loi de Poisson classique est une loi de Poisson pondérée de fonction poids $\omega(x) = 1$, pour tout $x \in \mathbb{N}$. Finalement, la fonction poids d'une loi de dénombrement peut être considérée comme une mesure uniforme pour détecter des départs de la loi de Poisson initiale.

Lorsqu'aucune information n'est disponible sur le processus sous-jacent aux données de dénombrement (de tailles moins importantes), l'approche non-paramétrique est la plus appropriée pour un traitement statistique. Cette approche est complémentaire au cas paramétrique. Parmi les estimateurs non-paramétriques (ondelettes, splines, noyaux, etc.), on s'intéresse ici à 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. Il va de soi de «lisser» cette distribution de manière discrète par des estimations équivalentes à des bâtons. L'approche non-paramétrique doit permettre de détecter une multimodalité et d'autres phénomènes

tels qu'une variation brutale, une absence d'observations ou des données parsemées.

Un objectif de cette thèse est de fournir un premier cadre approprié d'étude de l'estimateur à noyau discret. Très utile et simple dans la mise en oeuvre, ce travail introduit les fondements des estimateurs à noyau discret pour des données de dénombrement. Nous insisterons également sur l'importance de l'utilisation des lois discrètes usuelles pour construire des noyaux discrets que nous définissons dans le premier chapitre, ainsi que sur les choix de fenêtre. Les travaux de Aitchison & Aitken (1976) font partie des premiers sur le lissage discret de variables ou de fonctions de masse. Cependant, le noyau discret qu'ils proposent a une forme unique et est essentiellement conçu pour des données catégorielles ou des distributions discrètes finies (voir Section A.5 de l'Annexe A pour quelques détails). On peut également se référer pour le cas catégoriel à l'ouvrage de Li & Racine (2007) et leurs références, pour des applications en économétrie.

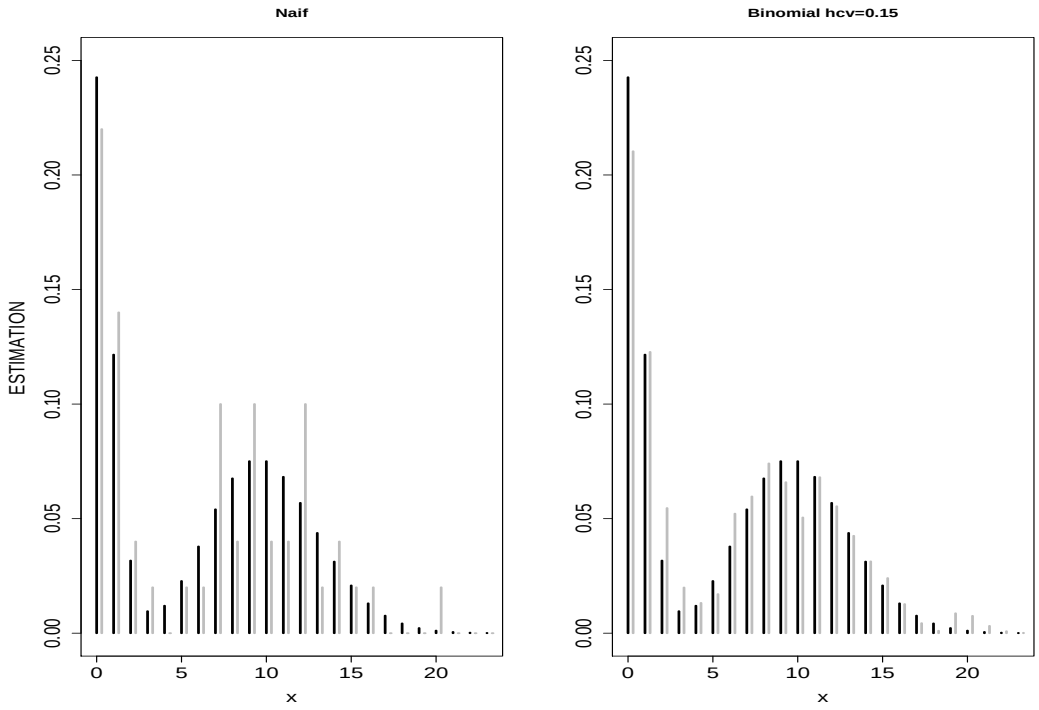

FIG. 1 – Lissages discrets par des estimateurs (bâtons gris) empirique et à noyau binomial des données simulées de tailles $n = 50$ de la distribution (bâtons noirs) du mélange de Poisson $f = 0.4\mathcal{P}(0.5) + 0.6\mathcal{P}(10)$

Au final, les motivations de ce travail sur des données de dénombrement sont principalement de deux ordres. La première motivation est le passage de l'approche paramétrique (quasi systématique) à une approche non-paramétrique, par noyau discret, laquelle se justifie facilement pour des échantillons de petites et de moyennes tailles.

En effet, dans le cas discret, l'estimateur empirique ou naïf est généralement suffisant pour des échantillons de grande taille. La Figure 1 met en évidence l'intérêt du lissage discret par un autre noyau en dehors du noyau de type Dirac, en l'occurence il s'agit ici du binomial. La seconde motivation, et donc la principale, est une estimation semi-paramétrique de la loi de Poisson pondérée définie en (1). En fait, dans l'expression (1), $p(x;\theta)$ est la partie paramétrique et $\omega(x)$ est la partie non-paramétrique que l'on va estimer dans le Chapitre 3 en appliquant la méthode des noyaux associés discrets. Une démarche analogue est également proposée en considérant des pondérations de la loi binomiale lorsque le support de la distribution est finie. Cependant, cela nécessite d'estimer auparavant la fonction de masse de probabilité $f(x) = \Pr(X = x)$ par la méthode appropriée des noyaux associés discrets. Une autre application présentée dans le Chapitre 4 est d'étudier une fonction discrète de régression $m(x) = \mathbb{E}(Y|X = x)$, où Y est une variable aléatoire réelle à expliquer.

Contenu de la thèse

Les quatres chapitres qui constituent ce travail sont placés dans un ordre chronologique et peuvent être considérés comme indépendants de manière à faire l'objet de quatres articles.

Dans le Chapitre 1, nous définissons d'abord le noyau associé discret qui est ensuite utilisé pour construire l'estimateur à noyau discret d'une fonction de masse. Nous y montrons quelques propriétés fondamentales de cet estimateur. Nous illustrons les résultats et la difficulté à construire des noyaux associés à partir des lois discrètes standards de dénombrement équidispersée (Poisson), sousdispersée (binomiale) et surdispersée (binomiale négative).

Le Chapitre 2 introduit de nouvelles lois de probabilités discrètes dites triangulaires. Nous les utilisons pour construire une classe de noyaux associés discrets symétriques. Ces derniers améliorent les noyaux discrets standards du Chapitre 1. Par conséquent, les estimateurs à noyaux associés discrets triangulaires ainsi construits seront plus performants que ceux à noyaux discrets asymétriques. L'essentiel de ce chapitre est déjà publié ; voir Kokonendji, Senga Kiessé & Zocchi (2007b).

Dans le Chapitre 3, nous répondons à la question de l'estimation de la fonction discrète poids $\omega(x)$ de (1). Ceci nous conduit à l'estimation semi-paramétrique des distributions des données de dénombrement. De plus, nous mettons en place des modèles de diagnostique permettant d'orienter le choix entre une approche semi-paramétrique et paramétrique, voire non-paramétrique.

Le Chapitre 4 est une application à l'estimation non-paramétrique d'une fonction discrète de régression sur une variable explicative de dénombrement. Nous adaptons ici l'estimateur de Nadaraya-Watson du cas continu et nous comparons leurs performances. Des perspectives y sont données pour les extensions de l'estimateur proposé à la régression semi-paramétrique univariée ainsi qu'au cas multivarié mixte.

Nous terminons par une conclusion générale suivi de quelques perspectives à ce travail. Notons que dans chacun des chapitres, les résultats seront illustrés à travers des simulations et appliqués sur des des données réelles et simulées.

Pour rendre ce document aussi compréhensible que possible, nous présentons dans l'Annexe A une approche unifiée de la notion de noyau associé à une loi de probabilité quelconque, continue ou discrète catégorielle. Nous y étudions des exemples de noyaux associés continus symétriques, asymétriques et discret catégoriel ainsi que leurs estimateurs respectifs. L'Annexe B contient des résultats graphiques et numériques supplémentaires. Enfin, l'Annexe C donne quelques détails des programmes mis en place sous le logiciel R*.

*R Development Core Team, 2006, *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.

Chapitre 1

Estimateur à noyau discret standard

1.1 Introduction

Soit $X_1, X_2, \ldots, X_n$ un n-échantillon aléatoire indépendant et identiquement distribué (i.i.d.) de densité de probabilité continue et inconnue f sur $\mathbb{R}$. Un estimateur à noyau continu $\widetilde{f}_n$ de f peut être défini de deux manières suivantes :

$$\widetilde{f}_n(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x-X_i}{h}\right) \tag{1.1}$$

$$= \frac{1}{n}\sum_{i=1}^{n}K_{x,h}\left(X_i\right), \quad x\in\mathbb{R}, \tag{1.2}$$

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 le «noyau associé continu» de cible x et de fenêtre h (cf. Définition 1.2.1 pour le cas discret). D'après l'expression (1.1) bien connue pour des supports non bornés de f, K est classiquement symétrique et donc le noyau associé s'écrit $K_{x,h}(\cdot) = (1/h)K\{(x-\cdot)/h\}$; mais le passage de (1.2) à (1.1) n'est pas toujours possible, par exemple pour des noyaux associés asymétriques par rapport à la cible. Dans les deux écritures (1.1) et (1.2), la fenêtre h joue le rôle de paramètre de dispersion autour de la cible ; ceci s'illustre simplement à travers le noyau associé gaussien symétrique $N_{x,h}$ de moyenne x (la cible) et d'écart-type h (la fenêtre) avec $K = N_{0,1}$. L'écriture (1.1) est connue depuis les travaux de Rosenblatt (1956) puis de Parzen (1962). Pour de récentes références, on peut consulter Berlinet & Biau (2002) et Tsybakov (2004). Les ouvrages couramment cités de Devroye (1987), Scott (1992) et Silverman (1986) sont pour des généralités sur des données (supposées) continues. Pour des données fonctionnelles, on peut se référer à l'ouvrage de Ferraty & Vieu (2006). Les travaux de Simonoff (1996) et Simonoff & Tutz (2000) sont dédiés aux données catégorielles ordonnées et discrètes utilisant *toujours* les noyaux continus. La seconde écriture (1.2) dont on fera usage dans ce travail est due à Chen (1999, 2000a) dans le but d'adapter au support pas nécessairement non borné de f un «type

de noyau continu» généralement asymétrique tels bêta et gamma ; voir aussi Scaillet (2004) pour les types de noyaux gaussien-inverse et gaussien-inverse-réciproque. Le cas d'un support borné, au moins d'un côté, de f à estimer induit un choix de type de noyau asymétrique. Tandis que les noyaux continus symétriques K n'ont pas d'effet propre majeur et peuvent s'utiliser indifféremment pour lisser des fonctions f à supports non bornés.

Pour estimer une *fonction de masse de probabilité* sur $\aleph$ (*e.g.* $\mathbb{N} + p\mathbb{N}$ pour $p \geq 0$, $\mathbb{Z}^d$, $\{0, 1, \ldots, N\}^d$, $d \in \mathbb{N}^*$) par une méthode de noyau discret, l'estimateur empirique ou naïf est souvent utilisé en raison de ses bonnes propriétés asymptotiques. Cependant, cet estimateur à noyau du type Dirac n'est pas approprié quand il s'agit d'échantillons de petites tailles. De plus, il a le défaut majeur de ne pas tenir compte des observations autour de la cible car sa fenêtre est nulle ou inexistante. En dehors de l'estimateur naïf, Aitchison & Aitken (1976) ont été les pionniers des estimateurs à noyau discret dans le sens (1.2). Mais le noyau discret utilisé n'a qu'une seule forme et est approprié pour des données catégorielles et des distributions discrètes finies ; voir aussi Li & Racine (2007) et leurs références. Ainsi, mis à part une première tentative de Marsh & Mukhopadhyay (1999), le cas des noyaux discrets pour les données de dénombrement est encore inexploré. La tentative de Marsh & Mukhopadhyay (1999) est uniquement expérimentale et est appliquée sur des données de comptage univariées (*i.e.* $\aleph = \mathbb{N}$).

Ce chapitre est organisé comme suit. Dans la Section 1.2, nous définissons un «noyau associé discret » et l'estimateur à noyau discret correspondant $\widetilde{f}_n$ d'une fonction de masse de probabilité f. Nous donnons aussi les propriétés globales de $\widetilde{f}_n$. Ensuite, nous montrons les convergences ponctuelle puis globale en moyenne quadratique de $\widetilde{f}_n$. La Section 1.3 illustre théoriquement certains aspects de ces estimateurs à l'aide des noyaux discrets standards de Poisson, binomial et binomial négatif. Ces derniers sont des prototypes de noyaux équidispersé, sousdispersé et surdispersé, respectivement (voir, par exemple, Mizère *et al.*, 2006). Dans la Section 1.4, nous proposons un indicateur de performance relative des noyaux discrets. La Section 1.5 propose des critères de choix de fenêtre de lissage discret. La Section 1.6 présente des illustrations de ces estimateurs pour le lissage discret des données de comptage simulées et réelles. Des comparaisons sont faites entre les trois types de noyaux discrets : Poisson, binomial et binomial négatif. Enfin, nous donnons une conclusion dans la Section 1.7.

1.2 Méthode des noyaux associés discrets

Pour simplifier, nous considérons que le support $\aleph$ de la fonction de masse de probabilité f à estimer est un ensemble discret (souvent $\mathbb{N}$) inclus dans $\mathbb{R}$ et, ainsi, la topologie induite est celle de $\mathbb{R}$. Pour cela, nous précisons d'abord les notions d'intégrale, de continuité et de dérivée d'une fonction f sur $\aleph$ avant de présenter la méthode des noyaux associés discrets.

Tout d'abord, l'ensemble discret $\aleph$ est muni de la mesure de dénombrement $\mu = \sum_{y\in\aleph} \eta_y$, où η_y est la masse de Dirac en y. Ainsi, pour toute fonction mesurable f, l'intégrale sur $\aleph_1$ inclus dans $\aleph$ n'est autre que la sommation :

$$\int_{\aleph_1} f(x)\mu(dx) = \sum_{x\in\aleph_1} f(x).$$

Ensuite, pour la fonction discrète f définie de $\aleph$ dans $\mathbb{R}$, la notion de continuité induite par la topologie de $\mathbb{R}$ se traduit par :

$$\forall\, \epsilon > 0,\ \exists\, \eta > 0 : \forall\, y \in]x-\eta, x+\eta[\cap\aleph \Rightarrow |f(y) - f(x)| < \epsilon. \tag{1.3}$$

Ainsi, nous ne sommes pas limités à une forme particulière de prolongement par continuité d'une fonction discrète. De plus, nous pouvons admettre que toute fonction discrète et bornée est continue au sens de (1.3), en particulier les fonctions de masse de probabilité. Notons que, pour $\eta > 0$ dans (1.3), la notion de voisinage discret $]x-\eta, x+\eta[\cap\aleph$ de x peut se réduire au singleton $\{x\}$. Dans la suite, nous utilisons cette propriété pour établir la convergence ponctuelle de l'estimateur à noyau discret.

Enfin, les dérivées sur $\mathbb{R}$ sont remplacées sur $\aleph$ par les différences finies de f en $x \in \aleph$; voir, par exemple, Schumaker (1981, page 343), Agarwal & Bohner (1999) pour d'autres définitions. Dans le cas particulier $\aleph = \mathbb{N}$ qui est une hypothèse de travail pour toute la suite, pour $k \geq 1$ on écrit par récurrence :

$$f^{(k)}(x) = \left\{f^{(k-1)}(x)\right\}^{(1)}$$

avec

$$f^{(1)}(x) = \begin{cases} \{f(x+1) - f(x-1)\}/2 & \text{si } x \in \mathbb{N}\setminus\{0\} \\ f(1) - f(0) & \text{si } x = 0. \end{cases} \tag{1.4}$$

Les $f^{(k)}(x)$ existent toujours et sont des combinaisons linéaires de $f(x \pm j)$ pour $j \in \{0, 1, \cdots, k\}$ et $x \pm j \in \mathbb{N}$. Ainsi, à partir de (1.4), nous déduisons par exemple

$$f^{(2)}(x) = \begin{cases} \{f(x+2) - 2f(x) + f(x-2)\}/4 & \text{si } x \in \mathbb{N}\setminus\{0,1\} \\ \{f(3) - 3f(1) + 2f(0)\}/4 & \text{si } x = 1 \\ \{f(2) - 2f(1) + f(0)\}/2 & \text{si } x = 0 \end{cases} \tag{1.5}$$

qui nous sera utile par la suite. En fait, nous utilisons ces différences finies dans le développement limité discret de Taylor de $f(x)$ en un point $a \in \aleph$ tel que

$$f(x) = \sum_{j=0}^{k} \frac{f^{(j)}(a)}{j!}(x-a)^j + o\{(x-a)^k\} \tag{1.6}$$

(voir, par exemple, Schumaker, 1981, Théorème 8.61 - page 351, pour une formulation précise). En général, nous n'irons pas au delà de l'ordre 2 ou 3 dans ce travail.

Notons que si $a \notin \aleph$ alors on a naturellement $f(a) = 0$. Cependant, en désignant par $\lfloor a \rfloor \in \aleph$ la valeur la plus proche (au sens de la topologie usuelle de $\mathbb{R}$) de $a \in \mathbb{R} \setminus \aleph$ tels que $a = \lfloor a \rfloor \pm \eta$ avec $\eta > 0$, on peut définir un prolongement de $f(a)$ par

$$\begin{aligned} f(a) &= f(\lfloor a \rfloor \pm \eta) \\ &= f(\lfloor a \rfloor) \pm \eta f^{(1)}(\lfloor a \rfloor) + o(\eta) \\ &= f(\lfloor a \rfloor) \pm \epsilon, \quad \epsilon > 0. \end{aligned} \tag{1.7}$$

Ainsi, le développement (1.6) de $f(x)$ est prolongeable en un point $a \notin \aleph$ en utilisant (1.7) qui peut être développée au delà de l'ordre 1. Cette définition va nous servir dans le développement limité discret de Taylor en des points qui n'appartiennent pas nécessairement à $\aleph$ comme la moyenne d'une loi discrète sur $\aleph = \mathbb{N}$.

1.2.1 Définition

Toute loi discrète de probabilité ne permet pas de définir un noyau associé discret (1.2). 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 associé discret pour le lissage discret des données tant en position qu'en échelle. Pour simplifier, on suppose pour toute la suite que le support de f est $\aleph = \mathbb{N}$.

Définition 1.2.1 *Soit $\aleph$ le support d'une fonction de masse de probabilité f à estimer. Étant donné $x \in \aleph$ et $h > 0$, on appelle « noyau associé discret » $K_{x,h}(\cdot)$ toute fonction de masse de probabilité liée à la variable aléatoire discrète $\mathcal{K}_{x,h}$, de support $\aleph_x$ contenant au moins x et indépendant de h, vérifiant les quatre conditions :*

$$\bigcup_{x \in \aleph} \aleph_x \supseteq \aleph, \tag{1.8}$$

$$\mathbb{E}(\mathcal{K}_{x,h}) \sim x \quad lorsque\ h \to 0, \tag{1.9}$$

$$var(\mathcal{K}_{x,h}) < +\infty, \tag{1.10}$$

$$var(\mathcal{K}_{x,h}) \to 0 \quad lorsque\ h \to 0. \tag{1.11}$$

Pour construire un noyau associé discret $K_{x,h}$ à partir d'une loi de probabilité discrète paramétrique K_θ, $\theta \in \Theta \subset \mathbb{R}^d$, de support $\aleph_\theta$ tel que $\aleph_\theta \cap \aleph \neq \emptyset$, on doit établir une correspondance entre $(x, h) \in \aleph \times \mathbb{R}_+^*$ et $\theta \in \Theta$. Par la suite, on appellera $K \equiv K_\theta$ le *type de noyau discret* pour différencier de la notion de «noyau» dans (1.1). Ainsi,

le choix du noyau associé discret sera tout aussi important que celui de la fenêtre. De plus, on distingue les noyaux associés discrets dits parfois du «second ordre» qui vérifient les quatres conditions ci-dessus de ceux dits du «premier ordre» qui vérifient toutes les conditions sauf (1.11).

REMARQUE 1.2.1 :

a. Etant donné un type de noyau discret K, le noyau associé discret construit n'est évidemment pas unique.

b. La condition (1.8) impose que le noyau associé discret $K_{x,h}$ doit tenir compte du support $\aleph$ de la fonction de masse de probabilité f à estimer. De plus, si $\cup_{x\in\aleph}\aleph_x$ n'est pas égale à $\aleph$ alors on a un problème naturel de biais de bordure qui n'est pas abordé ici. Il est parfois simple de supposer $\aleph_x \subseteq \aleph$ pour tout $x \in \aleph$.

c. La condition (1.9) exprime la prise en compte d'information autour de la cible de telle sorte que si $h \to 0$ nous retrouvons en moyenne le noyau de l'estimateur naïf. Cette condition 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 associés discrets à partir d'un type de noyau discret K ; par exemple, $\mathbb{E}(\mathcal{K}_{x,h}) = x+h$ ou $\mathbb{E}(\mathcal{K}_{x,h}) = x$. Ceci a été implicitement utilisé dans les cas continus asymétriques par Chen (1999, 2000a) puis Scaillet (2004) sans pour autant qu'une définition claire soit présentée. Ainsi, tous les noyaux associés discrets vérifiant (1.9) ont en commun une forme qui s'adapte selon la valeur de la cible x où ils sont calculés. La qualité de lissage discret obtenue change selon le comportement de leur variance var$(\mathcal{K}_{x,h})$ par rapport à la cible x. Ainsi, nous distinguons des noyaux associés discrets sousdispersés (var$(\mathcal{K}_{x,h}) < \mathbb{E}(\mathcal{K}_{x,h})$), équidispersés (var$(\mathcal{K}_{x,h}) = \mathbb{E}(\mathcal{K}_{x,h})$) ou surdispersés (var$(\mathcal{K}_{x,h}) > \mathbb{E}(\mathcal{K}_{x,h})$).

d. Le paramètre de lissage discret h permet de tenir compte des observations X_i qui sont proches (au sens de l'écart stochastique du type de noyau discret K) de la cible $x \in \aleph$. La dispersion locale var$(\mathcal{K}_{x,h})$ en tout $x \in \aleph$ traduit l'importance du noyau associé discret $K_{x,h}$ choisi qui impose sa propriété de variance pour une convergence.

e. Le comportement asymptotique recherché en (1.11) nous conduit à rechercher des noyaux associés discrets ayant la plus petite variance possible en chaque cible. Ceci conditionne les noyaux associés discrets à tendre vers le noyau de l'estimateur naïf qui est un noyau du type Dirac.

f. Toutes les conditions du noyau associé discret sont vérifiées par le noyau du type Dirac $D_{x,0}$ lié à la variable aléatoire $\mathcal{D}(x)$, pour $x \in \mathbb{N}$ et $h = 0$, et tel que

$$D_{x,0}(y) = \mathbf{1}_{y=x} = \begin{cases} 1 & \text{si } y = x \\ 0 & \text{sinon,} \end{cases} \tag{1.12}$$

avec $\mathbb{E}\{\mathcal{D}(x)\} = x$ et var$\{\mathcal{D}(x)\} = 0$.

Nous proposons ici trois exemples de construction de noyaux discrets asymétriques dits standards lesquels sont des noyaux associés discrets du premier ordre (*i.e.*, ne vérifiant pas la condition (1.11) de Définition 1.2.1). Ils sont aussi utiles pour estimer une fonction de masse de probabilité f sur $\aleph = \mathbb{N}$ ou pour lisser des distributions des données de comptage de petites tailles.

Pour tout $x \in \mathbb{N}$ et $h > 0$, la variable aléatoire discrète $\mathcal{K}_{x,h}$ des *noyaux discrets standards* satisfait, entre autre, les deux conditions

$$\mathbb{E}(\mathcal{K}_{x,h}) = x + h \tag{1.13}$$

$$\text{var}(\mathcal{K}_{x,h}) \nrightarrow 0 \text{ quand } h \to 0, \tag{1.14}$$

lesquelles remplacent (1.9) et (1.11), respectivement. En fait, la condition (1.13) s'adapte mieux au bord $x = 0$ et, de manière générale, la cible x n'est évidemment pas la moyenne de $\mathcal{K}_{x,h}$ qui est asymétrique mais plutôt son mode.

EXEMPLE 1.2.1 : Pour une loi de Poisson $\mathcal{P}(\lambda)$, on considère le noyau discret $P_{x,h}$ de loi $\mathcal{P}(x+h)$ sur $\aleph_x = \mathbb{N}$ avec $x \in \mathbb{N}$ et $h > 0$, tels que :

$$P_{x,h}(y) \;=\; \frac{(x+h)^y e^{-(x+h)}}{y!}, \quad y \in \mathbb{N}. \tag{1.15}$$

Signalons que le noyau discret proposé par Marsh & Mukhopadhyay (1999) interéchange x en y dans (1.15). De plus, il ne permet mathématiquement aucune étude des propriétés. Notons qu'en une cible $x \in \mathbb{N}$ et pour tout $h > 0$, le noyau discret $P_{x,h}$ est de support $\mathbb{N}$, équidispersé de moyenne égale à la variance $x + h$, et de mode compris entre $x + h - 1$ et $x + h$. Comme nous l'avons souligné précédement, la relation (1.11) n'est pas vérifiée ici car la variance tend vers x quand h tend vers 0.

EXEMPLE 1.2.2 : Si on considère une loi binomiale $\mathcal{B}(N, p)$, on lui associe le noyau $B_{x,h}$ de loi $\mathcal{B}\{x+1, (x+h)/(x+1)\}$ sur $\aleph_x = \{0, 1, \cdots, x+1\}$ pour tout $x \in \mathbb{N}$ et $h \in]0, 1]$ avec $\cup_x \aleph_x = \mathbb{N}$, de sorte que :

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

Ce noyau discret binomial $B_{x,h}$ est à support $\{0, 1, \cdots, x+1\}$ (dépendant de x), sous-dispersé de moyenne $x + h$ et de variance $(x+h)(1-h)/(x+1) < x + h$, et de mode autour de $x + h$. Quand h tend vers 0 la variance tend vers $x/(x+1) < 1$; donc, la condition (1.11) d'un noyau associé discret n'est pas satisfaite.

EXEMPLE 1.2.3 : Dans le cas de la loi binomiale négative $\mathcal{BN}(\lambda, p)$, on considère le noyau discret $BN_{x,h}$ de loi $\mathcal{BN}\{x+1, (x+1)/(2x+1+h)\}$ sur $\aleph_x = \mathbb{N}$ 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 $BN_{x,h}$ 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 $x+h$. De plus, lorsque h tend vers 0, la variance tend vers $x(2x+1)/(x+1) \neq 0$; donc, la condition (1.11) n'est pas vérifiée.

REMARQUE 1.2.2 : 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 serait $U(x,a)$, de loi $\mathcal{U}(x,a)$ sur $\aleph_x = \{x, x \pm 1, \cdots, x \pm a\}$ avec $\cup_x \aleph_x = \{-a, \cdots, -1\} \cup \mathbb{N} \supsetneq \mathbb{N}$. On obtiendrait :

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

D'après la Définition 1.2.1 du noyau associé discret, il apparaît qu'on ne peut pas établir de correspondance entre (x,a) et (x,h). Le paramètre réel de lissage habituel $h > 0$ ne peut se substituer ici à l'entier naturel non nul $a \in \mathbb{N}^*$. 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 la loi de Dirac $\mathcal{D}(x)$ en x.

La Figure 1.1 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. Dans la Section B.1 de l'Annexe B, nous ajoutons des graphiques du noyau binomial qui illustrent la forme variable des noyaux discrets standards en fonction de x et de h.

1.2.2 Propriétés élémentaires

Nous sommes en mesure de donner une définition précise d'un estimateur à noyau discret pour une fonction de masse de probabilité f inconnue sur un ensemble discret $\aleph$. Par la suite, nous présentons des propriétés fondamentales.

Définition 1.2.2 *Soit $X_1, X_2, \ldots, X_n$ un n-échantillon aléatoire i.i.d. de fonction de masse de probabilité inconnue f sur $\aleph$. Etant donné un type de noyau discret K, un estimateur à noyau discret $\widetilde{f}_n(x) = \widetilde{f}_{n,h,K}(x)$ de $f(x) := \Pr(X_1 = x)$ est défini par :*

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

où $h > 0$ est le paramètre de lissage discret (ou fenêtre) et $K_{x,h}$ (dépendant de x et de h) est le noyau associé discret sur $\aleph_x$.

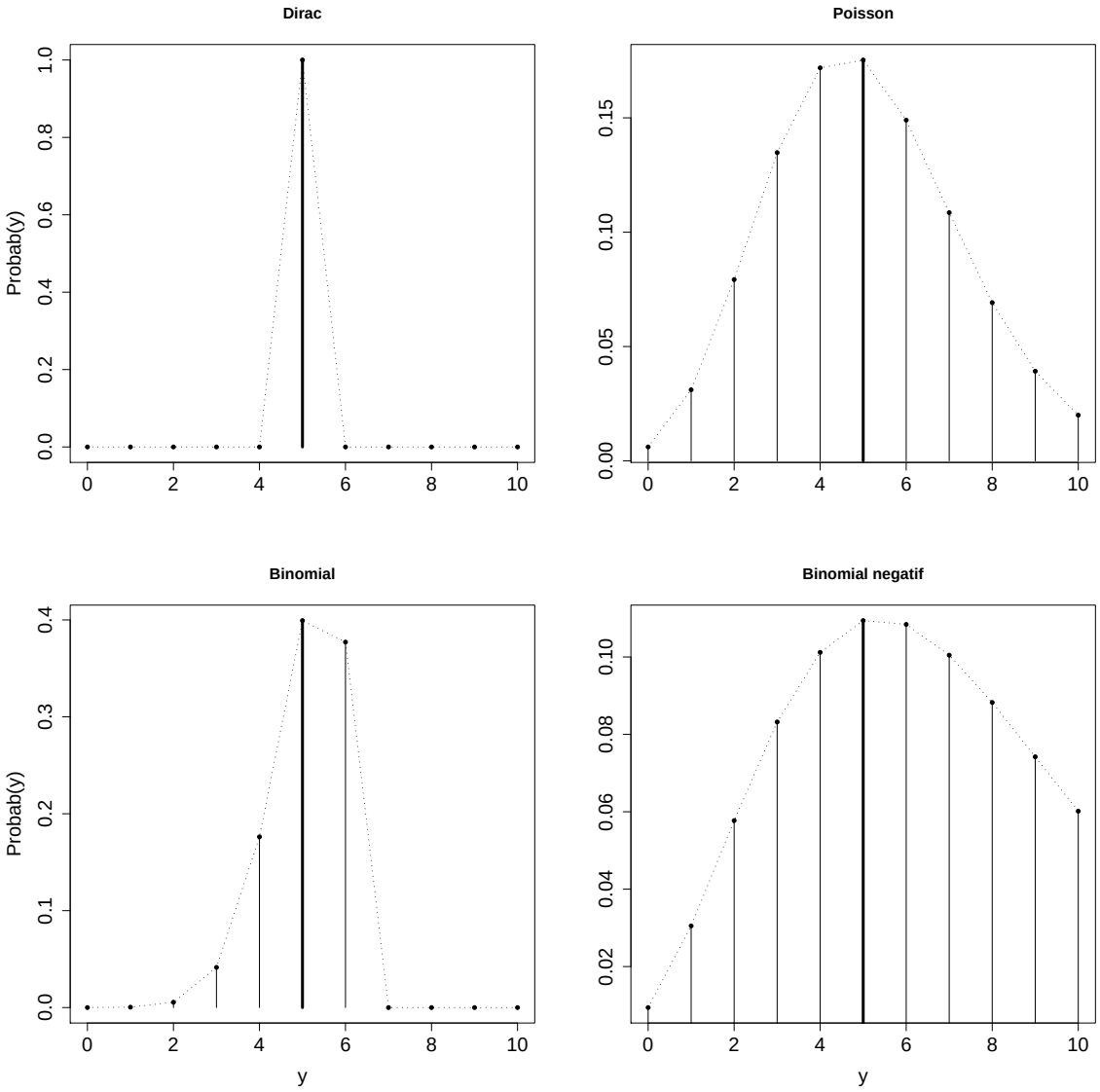

FIG. 1.1 – Noyaux discrets du type Dirac $\mathcal{D}(x)$, de Poisson $\mathcal{P}(x+h)$, binomial $\mathcal{B}\{x+1,(x+h)/(x+1)\}$ et binomial négatif $\mathcal{BN}\{x+1,(x+1)/(2x+1+h)\}$ avec $y=x=5$, $h=0.1$ (sauf pour Dirac)

Pour illustrer par des exemples, on peut étendre cette définition aux estimateurs standards $\widetilde{f}_n$ liés aux noyaux discrets $K_{x,h}$ donnés en exemple au paragraphe précédent (Exemples 1.2.1 à 1.2.3).

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

Proposition 1.2.3 *Soit $\underline{X} = (X_1, X_2, \ldots, X_n)$ un n-échantillon aléatoire i.i.d. de fonction de masse de probabilité inconnue f sur $\aleph$. Soit $\widetilde{f}_n = \widetilde{f}_{n,h,K}$ un estimateur à noyau associé discret de f. Alors, pour tout $x \in \aleph$ et $h > 0$, on a*

$$\mathbb{E}\{\widetilde{f}_n(x)\} = \mathbb{E}\{f(\mathcal{K}_{x,h})\}, \tag{1.16}$$

où $\mathcal{K}_{x,h}$ est la variable aléatoire de loi $K_{x,h}$ sur $\aleph_x$. De plus, on a $\widetilde{f}_n(x) \in [0,1]$ pour tout $x \in \aleph$ et

$$\sum_{x \in \aleph} \widetilde{f}_n(x) = C, \tag{1.17}$$

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

DÉMONSTRATION : Pour tout $x \in \aleph$, on a successivement

$$\mathbb{E}\{\widetilde{f}_n(x)\} = \sum_{y \in \aleph \cap \aleph_x} K_{x,h}(y) f(y) = \sum_{y \in \aleph \cap \aleph_x} f(y) \Pr(\mathcal{K}_{x,h} = y) = \mathbb{E}\{f(\mathcal{K}_{x,h})\},$$

car $f(y)K_{x,h}(y) = 0$ pour $y \notin \aleph \cap \aleph_x$; et la formule est ainsi montrée. Ensuite, $\widetilde{f}_n(x) \in [0,1]$ découle immédiatement de $K_{x,h}(X_i) \in [0,1]$ pour tout X_i. Enfin, en écrivant $C = n^{-1} \sum_{i=1}^{n} \left\{ \sum_{x \in \aleph} K_{x,h}(X_i) \right\}$ pour tout $h > 0$ et en observant que

$$\sum_{x \in \aleph} K_{x,h}(y) = \sum_{x \in \aleph \cap \aleph_x} K_{x,h}(y)$$

pour tout $y \in \aleph$ =support(X_i), il en résulte que :

- d'une part $C > 0$ car $K_{x,h}(x) > 0$ pour $y = x \in \aleph \cap \aleph_x$ et

- d'autre part $C < +\infty$ car $0 \leq K_{x,h}(y) < 1$ pour tout $y \in \aleph \cap \aleph_x$ et $K_{x,h}(y) \to 0$ quand $x \to +\infty$.■

De manière pratique, nous calculons cette constante C, dépendant des observations, avant la normalisation de l'estimateur $\widetilde{f}_n$ pour en faire une fonction de masse de probabilité. Sans perte de généralité, nous considérons désormais $C = 1$.

1.2.3 Convergence ponctuelle

Le résultat suivant garantit que l'estimateur à noyau discret est asymptotiquement sans biais en tout point. C'est une adaptation au cas discret du Lemme de Hille (1948) dont une démonstration a été donnée par Feller (1966, pages 219–220). Nous présentons dans la Section A.1 de l'Annexe A la version continue de ce résultat.

Proposition 1.2.4 *Soit $f : \aleph \to \mathbb{R}$ une fonction de masse de probabilité et soit $x \in \aleph$ fixé. Si $\widetilde{f}_n(x)$ est l'estimateur de $f(x)$ à noyau associé discret $K_{x,h}$ sur $\aleph_x$, alors*

$$\mathbb{E}\{\widetilde{f}_n(x)\} = \sum_{y \in \aleph \cap \aleph_x} f(y) K_{x,h}(y) \to f(x) \quad \textit{quand} \quad h = h(n) \to 0 \ \textit{ si } \ n \to +\infty.$$

DÉMONSTRATION : Puisque $f(y)K_{x,h}(y) = 0$ pour tout $y \notin \aleph \cap \aleph_x$, on suppose $\aleph_x \subseteq \aleph$ pour tout $x \in \aleph$. Pour tout $\eta > 0$ on note $\aleph_{x,\eta} = \{y \in \aleph_x : |y - x| < \eta\}$ et $\overline{\aleph}_{x,\eta} = \{y \in \aleph_x : |y-x| \geq \eta\}$ son complémentaire. Sachant que $f(x) = f(x) \sum_{y \in \aleph_x} K_{x,h}(y)$, nous exprimons

$$\begin{aligned} |\mathbb{E}\{\widetilde{f}_n(x)\} - f(x)| &= \Big| \sum_{y \in \aleph_x \subseteq \aleph} \{f(y) - f(x)\} K_{x,h}(y) \Big| \\ &\leq \sum_{y \in \aleph_{x,\eta}} |f(y) - f(x)| K_{x,h}(y) + \sum_{y \in \overline{\aleph}_{x,\eta}} |f(y) - f(x)| K_{x,h}(y), \end{aligned}$$

où $\eta > 0$ est une constante arbitraire. Puisque f est une fonction de masse de probabilité et par conséquent continue d'après (1.3), pour tout $\epsilon > 0$ il existe un $\eta = \eta(\epsilon) > 0$ pour lequel

$$\sum_{y \in \aleph_{x,\eta}} |f(y) - f(x)| K_{x,h}(y) \leq \epsilon. \tag{1.18}$$

Considérons la variable aléatoire $\mathcal{K}_{x,h}$. Pour $a > 0$ et pour tout $\epsilon > 0$, l'inégalité de Tchebychev-Markov s'écrit ici :

$$\Pr\left(|\mathcal{K}_{x,h} - a| \geq \epsilon\right) \leq \frac{\mathbb{E}\{(\mathcal{K}_{x,h} - a)^2\}}{\epsilon^2}.$$

Ainsi, puisque f est une fonction de masse de probabilité et donc $f \leq 1$, nous pouvons majorer le second terme par :

$$\begin{aligned} \sum_{y \in \overline{\aleph}_{x,\eta}} |f(y) - f(x)| K_{x,h}(y) &\leq 2 \sum_{y \in \overline{\aleph}_{x,\eta}} K_{x,h}(y) = 2 \Pr\{|\mathcal{K}_{x,h} - x| > \eta\} \\ &\leq \frac{2}{\eta^2} \mathbb{E}\{(\mathcal{K}_{x,h} - x)^2\} \\ &\leq \frac{2}{\eta^2} [\text{var}(\mathcal{K}_{x,h}) + \{\mathbb{E}(\mathcal{K}_{x,h}) - x\}^2]. \end{aligned} \tag{1.19}$$

Finalement, sous les conditions (1.9) et (1.11) de la définition d'un noyau associé discret, les inégalités (1.18) et (1.19) permettent d'aboutir au résultat recherché. ■

Nous sommes en mesure de montrer le résultat de la convergence ponctuelle en moyenne quadratique de l'estimateur à noyau discret.

Proposition 1.2.5 *Soit $f : \aleph \to \mathbb{R}$ une fonction de masse de probabilité et soit $x \in \aleph$ fixé. Si $\widetilde{f}_n(x)$ est un estimateur à noyau associé discret de $f(x)$ alors*

$$\mathbb{E}\{\widetilde{f}_n(x) - f(x)\}^2 \to 0 \quad \textit{quand} \quad n \to +\infty \ \textit{et} \ h = h(n) \to 0.$$

DÉMONSTRATION : Sans perte de généralité, on suppose $\aleph_x \subseteq \aleph$ pour tout $x \in \aleph$. D'après la Proposition 1.2.4 nous avons, pour tout $x \in \aleph$,

$$\text{biais}\{\widetilde{f}_n(x)\} = \mathbb{E}\{\widetilde{f}_n(x)\} - f(x) \to 0 \ \text{ quand } \ n \to +\infty \ \text{ et } \ h = h(n) \to 0.$$

La variance ponctuelle peut être majorée successivement par

$$\begin{aligned} n \ \text{var}\{\widetilde{f}_n(x)\} &= \text{var}\{K_{x,h}(X_1)\} \\ &\leq \mathbb{E}\left\{K_{x,h}\left(X_1\right)\right\}^2 \\ &\leq \sum_{y \in \aleph_x} f(y)\{\Pr(\mathcal{K}_{x,h} = y)\}^2 \\ &\leq \sum_{y \in \aleph_x} f(y)\{\Pr(\mathcal{K}_{x,h} = y)\} \\ &\leq 1. \end{aligned}$$

Les dernières inégalités s'obtiennent facilement grâce à $\Pr(\mathcal{K}_{x,h} = y) \leq 1$ et à $f(y) \leq 1$. De là, puisqu'on a $0 \leq \text{var}\{\widetilde{f}_n(x)\} \leq (1/n)$ alors $\text{var}\{\widetilde{f}_n(x)\}$ tend vers 0 pour tout $x \in \mathbb{N}$ lorsque $n \to +\infty$ et $h = h(n) \to 0$. Par conséquent, le résultat découle de la décomposition $\mathbb{E}\{\widetilde{f}_n(x) - f(x)\}^2 = \text{var}\{\widetilde{f}_n(x)\} + \text{biais}^2\{\widetilde{f}_n(x)\}$.■

Finalement, sous l'hypothèse des noyaux associés discrets, on peut déduire la convergence globale de l'estimateur à noyau discret par

$$\sum_{x \in \aleph} \mathbb{E}\{\widetilde{f}_n(x) - f(x)\}^2 \to 0 \ \text{ quand } n \to +\infty \ \text{ et } \ h = h(n) \to 0,$$

pour tout f telle que $\lim_{x \to +\infty} f(x) = 0$ (voir, par exemple, Théorème 1.2.6).

REMARQUE 1.2.5 : Le développement discret de Taylor en (1.24) permet d'exprimer

$$\begin{aligned} \lim_{n \to +\infty} \mathbb{E}\{\widetilde{f}_{n,h,K}(x)\} &\overset{(1.16)}{=} \lim_{h \to 0} \mathbb{E}\{f(\mathcal{K}_{x,h})\} \\ &\overset{(1.24)}{=} \lim_{h \to 0} [f\{\mathbb{E}(\mathcal{K}_{x,h})\} + \frac{1}{2}\text{var}(\mathcal{K}_{x,h}) f^{(2)}(x)] \\ &\overset{(a)}{=} f(x) + \alpha, \end{aligned}$$

avec $\alpha = \lim_{h \to 0}\{(1/2)\text{var}(\mathcal{K}_{x,h}) f^{(2)}(x)\}$. La première égalité découle de (1.16) et du fait que $h = h_n \to 0$ lorsque $n \to +\infty$. Dans le résultat final (a), $\alpha = 0$ sous les

conditions (1.9) et (1.11) du noyau associé discret. Tandis que, pour les noyaux discrets standards, la condition (1.11) n'est pas vérifiée et donc $\alpha \neq 0$. Ceci induit un biais qui incite à valoriser les noyaux discrets standards ayant la plus petite variance. Dans le cas des noyaux associés discrets de premier ordre ne vérifiant pas la condition (1.11), nous pouvons obtenir une convergence ponctuelle faible dite également de premier ordre en arrêtant le développement limité discret à l'ordre 1 tel que $f(\mathcal{K}_{x,h}) \doteq f(m_{x,h}) + (\mathcal{K}_{x,h} - m_{x,h}) f^{(1)}(m_{x,h})$.

1.2.4 Convergence globale

Le critère à utiliser pour cette convergence est le *risque quadratique intégré* (en anglais «Mean Integrated Squared Error») de $\widetilde{f}_n = \widetilde{f}_{n,h,K}$ défini par

$$\begin{aligned} MISE &= \mathbb{E}\left[\sum_{x\in\aleph}\left\{\widetilde{f}_n(x) - f(x)\right\}^2\right] && (1.20)\\ &= \sum_{x\in\aleph}\text{var}\left\{\widetilde{f}_n(x)\right\} + \sum_{x\in\aleph}\text{biais}^2\left\{\widetilde{f}_n(x)\right\} && (1.21)\\ &=: MISE(n,h,K,f). \end{aligned}$$

Nous commençons par l'analyse de la variance ponctuelle. Sans perte de généralité, on suppose $\aleph_x \subseteq \aleph$ pour tout $x \in \aleph$. La variance $\text{var}\left\{\widetilde{f}_n(x)\right\}$ se décompose autour de la cible x (qui réalise la probabilité modale de $\mathcal{K}_{x,h}$) par :

$$\begin{aligned} \text{var}\left\{\widetilde{f}_n(x)\right\} &= \frac{1}{n}\text{var}\{K_{x,h}(X_1)\}\\ &= \frac{1}{n}\left[\mathbb{E}K_{x,h}^2(X_1) - \{\mathbb{E}K_{x,h}(X_1)\}^2\right]\\ &= \frac{1}{n}\left[\sum_{y\in\aleph_x} f(y)\left\{\Pr(\mathcal{K}_{x,h}=y)\right\}^2 - \left\{\sum_{z\in\aleph_x} f(z)\Pr(\mathcal{K}_{x,h}=z)\right\}^2\right]\\ &= \frac{1}{n}\left[f(x)\left\{\Pr(\mathcal{K}_{x,h}=x)\right\}^2 - \left\{f(x)\Pr(\mathcal{K}_{x,h}=x)\right\}^2\right] + R_n(x;h)\\ &= \frac{1}{n}f(x)\left\{1-f(x)\right\}\left\{\Pr(\mathcal{K}_{x,h}=x)\right\}^2 + R_n(x;h), \qquad (1.22) \end{aligned}$$

où le reste

$$\begin{aligned} R_n(x;h) &= \frac{1}{n}\left[\sum_{y\in\aleph_x\setminus\{x\}} f(y)\left\{\Pr(\mathcal{K}_{x,h}=y)\right\}^2 + \left\{f(x)\Pr(\mathcal{K}_{x,h}=x)\right\}^2\right]\\ &\quad -\frac{1}{n}\left\{\sum_{z\in\aleph_x} f(z)\Pr(\mathcal{K}_{x,h}=z)\right\}^2 \qquad (1.23) \end{aligned}$$

devient négligeable sous l'hypothèse de noyau associé discret ; *i.e.* pour tout $x \in \aleph$, $R_n(x;h) \to 0$ quand $n \to +\infty$ et $h = h(n) \to 0$. En effet, pour les $y \in \aleph_x \setminus \{x\}$ il existe $\eta = \eta(y) > 0$ tel que

$$\begin{aligned} 0 &\leq \Pr(\mathcal{K}_{x,h} = y) = \Pr(|\mathcal{K}_{x,h} - x| > \eta) \\ &\leq \frac{1}{\eta^2}\mathbb{E}\{(\mathcal{K}_{x,h} - x)^2\} = \frac{1}{\eta^2}[\text{var}(\mathcal{K}_{x,h}) + \{\mathbb{E}(\mathcal{K}_{x,h}) - x\}^2] \to 0 \text{ quand } h \to 0, \end{aligned}$$

et pour les $y = x$ on a la probabilité modale $\Pr(\mathcal{K}_{x,h} = x) \to 1$ quand $h \to 0$; d'où, le résultat en découle. Concernant le biais ponctuel de $\widetilde{f}_n$, par le développement discret de Taylor (1.6) de $f(\mathcal{K}_{x,h})$ prolongé au point $m_{x,h}$ satisfaisant (1.26), nous obtenons :

$$f(\mathcal{K}_{x,h}) \doteq f(m_{x,h}) + (\mathcal{K}_{x,h} - m_{x,h})f^{(1)}(m_{x,h}) + \frac{1}{2}(\mathcal{K}_{x,h} - m_{x,h})^2 f^{(2)}(m_{x,h})$$

où le symbole $\doteq$ indique un équivalent asymptotique et $f^{(1)}$ et $f^{(2)}$ sont les différences finies données en (1.4) et (1.5), respectivement. Puis, en prenant l'espérance mathématique et en utilisant (1.7) uniquement pour $f^{(2)}(m_{x,h}) \doteq f^{(2)}(x+h) = f^{(2)}(x) + hf^{(3)}(x) + o(h)$, nous arrivons à :

$$\mathbb{E}\{f(\mathcal{K}_{x,h})\} \doteq f\{\mathbb{E}(\mathcal{K}_{x,h})\} + \frac{1}{2}\text{var}(\mathcal{K}_{x,h})f^{(2)}(x).$$

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. De là, nous aboutissons à :

$$\begin{aligned} \text{biais}\{\widetilde{f}_n(x)\} &= \mathbb{E}\{f(\mathcal{K}_{x,h})\} - f(x) \\ &= f\{\mathbb{E}(\mathcal{K}_{x,h})\} - f(x) + \frac{1}{2}\text{var}(\mathcal{K}_{x,h})f^{(2)}(x) + o(h). \end{aligned} \quad (1.24)$$

Nous pouvons donner le résultat général suivant qui permet au cas par cas des noyaux associés discrets d'aboutir à la convergence de l'estimateur $\widetilde{f}_n$ au sens du $MISE$.

Théorème 1.2.6 *Soit f une fonction de masse de probabilité sur $\aleph$. Alors, l'estimateur à noyau associé discret $\widetilde{f}_n = \widetilde{f}_{n,h,K}$ de f est tel que, pour $n \to +\infty$ et $h = h(n) \to 0$, on ait le comportement*

$$\begin{aligned} MISE(n,h,K,f) &\doteq \frac{1}{n}\sum_{x\in\aleph}\{\Pr(\mathcal{K}_{x,h} = x)\}^2 f(x)\{1 - f(x)\} \\ &\quad + \sum_{x\in\aleph}\left[f\{\mathbb{E}(\mathcal{K}_{x,h})\} - f(x) + \frac{1}{2}\text{var}(\mathcal{K}_{x,h})f^{(2)}(x)\right]^2, \end{aligned} \quad (1.25)$$

où $f^{(2)}$ est la différence finie d'ordre 2 donnée en (1.5).

DÉMONSTRATION : Sans perte de généralité, on suppose $\aleph_x \subseteq \aleph$ pour tout $x \in \aleph$. Le théorème découle de (1.22) et (1.24) dans le critère du $MISE$ (1.21). ■

REMARQUE 1.2.3 : Quand $n \to +\infty$ et $h = h(n) \to 0$, on pourrait obtenir $R_n(x;h) = o(1/n)$ avec un noyau associé discret de moyenne x, de variance tendant vers 0 et qui pourrait être également symétrique et unimodal en x.

REMARQUE 1.2.4 : Si $\text{var}(\mathcal{K}_{x,h}) \nrightarrow 0$ quand $h = h(n) \to 0$ et $n \to +\infty$, le reste $R_n(x;h)$ est difficilement négligeable. Les calculs peuvent être affinés en tenant compte d'un certain nombre de points dans le voisinage de la cible.

Pour les noyaux discrets standards qui ne sont pas des noyaux associés discrets, on considère d'abord l'extension de la condition (1.9) de la manière suivante :

$$m_{x,h} = \mathbb{E}(\mathcal{K}_{x,h}) = x + h + o(h), \tag{1.26}$$

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

$$\sigma^2_{x,h} = \text{var}(\mathcal{K}_{x,h}) = V(x,h) + o(h), \tag{1.27}$$

où $V(x,h)$ est positive, finie et ne tend pas nécessairement vers 0 quand $h \to 0$. Sachant que $m_{x,h} = \mathbb{E}(\mathcal{K}_{x,h})$ satisfait (1.26), en utilisant (1.7) pour

$$f(m_{x,h}) \doteq f(x+h) = f(x) + hf^{(1)}(x) + o(h),$$

il s'en suit donc de (1.27) que :

$$\text{biais}\{\widetilde{f}_{n,h,K}(x)\} = hf^{(1)}(x) + \frac{1}{2}V(x,h)f^{(2)}(x) + o(h).$$

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

$$\begin{aligned} AMISE^*(n,h,K,f) &= \frac{1}{n}\sum_{x\in\aleph} f(x)\{1-f(x)\}\{\Pr(\mathcal{K}_{x,h}=x)\}^2 \\ &\quad + \sum_{x\in\aleph}\left\{hf^{(1)}(x) + \frac{1}{2}V(x,h)f^{(2)}(x)\right\}^2. \end{aligned} \tag{1.28}$$

Par conséquent, la minimisation en h de la borne supérieure de l'approximation $AMISE(n,h,K,f)$ en (1.25) pourrait conduire à la vitesse de convergence globale $O(n^{-1})$ de $\widetilde{f}_{n,h,K}$ vers f lorsque $n \to +\infty$. Aussi, pour un type de noyau discret K spécifié, la valeur optimale h^* de h est donnée par

$$h^* = \arg\min_{h>0} AMISE(n,h,K,f) = h^*(n,K,f). \tag{1.29}$$

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

1.3 Noyau du type Dirac et noyaux discrets standards

Nous détaillons ici les expressions des risques quadratiques intégrés exacts $MISE$ de (1.21) pour les estimateurs à noyaux discrets associés aux lois de Dirac, de Poisson, binomiale et binomiale négative. Nous étudions aussi les limites des majorations des $MISE$ des estimateurs à noyaux discrets standards et présentons des simulations de leurs $MISE$ comparés avec celui de l'estimateur empirique pour une fonction de masse probabilité f ; ceci dans le but de valider les nouvels estimateurs mis en place et de déterminer leurs limites pratiques.

1.3.1 Dirac

Dans le cas particulier de l'estimateur à noyau du type Dirac (1.12) tel que $\Pr(\mathcal{D}_{x,0} = x) = 1$ et, donc, pour $y \in \aleph_x \setminus \{x\}$ on a $\Pr(\mathcal{D}_{x,0} = y) = 0$. Il s'en suit que la variance de $\widetilde{f}_{n,0,D}$ en x est explicitement donnée par

$$\text{var}\left\{\widetilde{f}_{n,0,D}(x)\right\} = \frac{1}{n} f(x)\left\{1 - f(x)\right\},$$

donc le reste $R_n(x;h) = 0$ (1.23), et l'estimateur $\widetilde{f}_{n,0,D}$ est sans biais :

$$\begin{aligned}
\text{biais}\left\{\widetilde{f}_{n,0,D}(x)\right\} &= \mathbb{E}\left\{\widetilde{f}_{n,0,D}(x)\right\} - f(x) \\
&= \mathbb{E}\left\{D_{x,0}(X_1)\right\} - f(x) \\
&= \sum_{y \in \aleph_x} f(y)\Pr(\mathcal{D}_{x,0} = y) - f(x) \\
&= 0.
\end{aligned}$$

Par conséquent, le risque quadratique intégré devient

$$MISE(n,0,D,f) = \frac{1}{n}\sum_{x\in\mathbb{N}} f(x)\left\{1 - f(x)\right\} = \frac{1}{n}\left\{1 - \sum_{x\in\mathbb{N}} f^2(x)\right\}. \qquad (1.30)$$

Ce résultat exact sert de référence pour la comparaison avec le $MISE$ des autres estimateurs à noyaux associés discrets, car $0 \leq \sum_{x\in\aleph} f^2(x) < 1$ et donc on a la convergence globale de l'estimateur naïf par

$$MISE(n,0,D,f) \to 0 \text{ quand } n \to +\infty. \qquad (1.31)$$

Type de noyau	$\mathbb{E}(\mathcal{K}_{x,h})$	var$(\mathcal{K}_{x,h})$	$V(x,h)$	$\lim_{h\to 0}$var$(\mathcal{K}_{x,h})$
Dirac	x	0	0	0
Poisson	$x+h$	$x+h$	$x+h$	$x \in \mathbb{N}$
Binomial	$x+h$	$(x+h)\left(\frac{1-h}{x+1}\right)$	$(x+h)\left(\frac{1}{x+1}\right) - \frac{xh}{x+1}$	$\frac{x}{x+1} \in [0,1[$
Binomial négatif	$x+h$	$(x+h)\left(1+\frac{x+h}{x+1}\right)$	$(x+h)\left(1+\frac{x}{x+1}\right) + \frac{xh}{x+1}$	$\frac{x(2x+1)}{x+1} \geq 0$

TAB. 1.1 – Résumé des propriétés de quelques noyaux discrets

1.3.2 Poisson

Risque quadratique intégré exact

Les calculs directs du biais et de la variance de l'estimateur à noyau de Poisson (Exemple 1.2.1) donnent, en tout point $x \in \mathbb{N}$,

$$\text{biais}\{\widetilde{f}_{n,h}(x)\} = f(x)\left\{\frac{(x+h)^x e^{-(x+h)}}{x!} - 1\right\} + \sum_{y\in\mathbb{N}\setminus\{x\}} f(y)P_{x,h}(y)$$

et

$$\begin{aligned}\text{var}\left\{\widetilde{f}_{n,h}(x)\right\} = \frac{1}{n}&\left[f(x)\left\{\frac{(x+h)^x e^{-(x+h)}}{x!}\right\}^2 + \sum_{y\in\mathbb{N}\setminus\{x\}} f(y)P_{x,h}^2(y)\right] \\ &-\frac{1}{n}\left\{f(x)\frac{(x+h)^x e^{-(x+h)}}{x!} + \sum_{y\in\mathbb{N}\setminus\{x\}} f(y)P_{x,h}(y)\right\}^2.\end{aligned}$$

Nous pouvons en déduire l'expression de $MISE$ à partir de (1.21).

Pour n et x fixés, nous exprimons la limite du biais de $\widetilde{f}_{n,h}(x)$ quand $h \to 0$ par

$$\lim_{h\to 0}\text{biais}\{\widetilde{f}_{n,h}(x)\} = f(x)\left\{\frac{x^x e^{-x}}{x!} - 1\right\} + e^{-x}\sum_{y\in\mathbb{N}\setminus\{x\}} f(y)\frac{x^y}{y!} \neq 0,$$

puis celle de la variance de $\widetilde{f}_{n,h}(x)$ par

$$\begin{aligned}\lim_{h\to 0} \text{var}\left\{\widetilde{f}_{n,h}(x)\right\} &= \frac{1}{n}\left[f(x)\{1-f(x)\}\left\{\frac{x^x e^{-x}}{x!}\right\}^2 + e^{-2x}\sum_{y\in\mathbb{N}\setminus\{x\}} f(y)\frac{x^{2y}}{(y!)^2}\right] \\ &\quad -\frac{2}{n}f(x)\frac{x^x e^{-2x}}{x!}\sum_{y\in\mathbb{N}\setminus\{x\}} f(y)\frac{x^y}{y!} - \frac{1}{n}e^{-2x}\left\{\sum_{y\in\mathbb{N}\setminus\{x\}} f(y)\frac{x^y}{y!}\right\}^2.\end{aligned}$$

Ainsi, pour n fixé, nous pouvons calculer la limite du $MISE$ quand $h \to 0$ de l'estimateur à noyau de Poisson, laquelle est différente du $MISE$ de l'estimateur naïf en (1.30).

Dans la Figure 1.2, pour $h > 0$ fixé, nous comparons les $MISE$ exacts des estimateurs à noyau de Poisson et à noyau du type Dirac en fonction de n et pour une fonction de masse de probabilité f qui est un mélange de Poisson. Le point d'intersection des courbes indique la limite supérieure de n pour laquelle l'estimateur à noyau de Poisson est plus performant que l'estimateur empirique, pour cette distribution f. Au delà de cette limite, l'estimateur empirique devient meilleur et son $MISE$ tend vers 0, ce qui n'est pas le cas du $MISE$ de l'estimateur à noyau de Poisson même si h est très petit.

1.3.3 Binomial

Risque quadratique intégré exact

Les calculs directs permettent aussi d'exprimer le biais ainsi que la variance de l'estimateur à noyau binomial (Exemple 1.2.2), en tout point $x \in \mathbb{N}$, de la manière suivante :

$$\text{biais}\{\widetilde{f}_{n,h}(x)\} = f(x)\left\{(1-h)\left(\frac{x+h}{x+1}\right)^x - 1\right\} + \sum_{y\in\aleph_x\setminus\{x\}} f(y)B_{x,h}(y)$$

et

$$\begin{aligned}\text{var}\left\{\widetilde{f}_{n,h}(x)\right\} &= \frac{1}{n}\left\{f(x)(1-h)^2\left(\frac{x+h}{x+1}\right)^{2x} + \sum_{y\in\aleph_x\setminus\{x\}} f(y)B_{x,h}^2(y)\right\} \\ &\quad -\frac{1}{n}\left\{f(x)(1-h)\left(\frac{x+h}{x+1}\right)^x + \sum_{y\in\aleph_x\setminus\{x\}} f(y)B_{x,h}(y)\right\}^2,\end{aligned}$$

avec $\aleph_x = \{0, 1, \ldots, x+1\}$. De là, nous pouvons déduire aisément le $MISE$ exact de l'estimateur à noyau binomial.

Pour n et x fixés, les calculs des limites quand $h \to 0$ conduisent à :

$$\lim_{h\to 0} \text{biais}\{\widetilde{f}_{n,h}(x)\} = f(x)\left\{\left(\frac{x}{x+1}\right)^x - 1\right\} + \frac{x!}{(x+1)^x} \sum_{y\in\aleph_x\setminus\{x\}} f(y)\frac{x^y}{y!(x+1-y)!} \neq 0$$

et

$$\begin{aligned}
\lim_{h\to 0} \text{var}\left\{\widetilde{f}_{n,h}(x)\right\} &= \frac{1}{n}f(x)\{1-f(x)\}\left(\frac{x}{x+1}\right)^{2x} \\
&\quad + \frac{1}{n}\left\{\frac{x!}{(x+1)^x}\right\}^2 \sum_{y\in\aleph_x\setminus\{x\}} f(y)\left\{\frac{x^y}{y!(x+1-y)!}\right\}^2 \\
&\quad - \frac{2}{n}f(x)\frac{x^{x+1}(x-1)!}{(x+1)^{2x}} \sum_{y\in\aleph_x\setminus\{x\}} f(y)\frac{x^y}{y!(x+1-y)!} \\
&\quad - \frac{1}{n}\left\{\frac{x!}{(x+1)^x}\right\}^2 \left\{\sum_{y\in\aleph_x\setminus\{x\}} f(y)\frac{x^y}{y!(x+1-y)!}\right\}^2 .
\end{aligned}$$

Ainsi, pour n fixé, la limite du $MISE$ quand $h \to 0$ de l'estimateur à noyau binomial est différente du $MISE$ de l'estimateur naïf (1.30).

Dans la Figure 1.3, pour $h \in]0,1]$ fixé et la distribution f donnée, l'estimateur à noyau binomial est meilleur que l'estimateur naïf au sens de $MISE$ jusqu'à la limite supérieure de n située à l'intersection des courbes. Au delà de cette limite, l'estimateur naïf devient meilleur et son $MISE$ tend vers 0 ; ce que ne vérifie pas le $MISE$ de l'estimateur à noyau binomial, même pour h petit, bien que graphiquement cela semble être le cas pour $h = 0.1$.

1.3.4 Binomial négatif

Risque quadratique intégré exact

Les calculs directs du $MISE$ (1.21) de l'estimateur à noyau binomial négatif permettent d'obtenir le biais puis la variance ponctuels par

$$\begin{aligned}
\text{biais}\{\widetilde{f}_{n,h}(x)\} &= f(x)\left\{\frac{(2x)!}{(x!)^2}\left(\frac{x+h}{2x+1+h}\right)^x\left(\frac{x+1}{2x+1+h}\right)^{x+1} - 1\right\} \\
&\quad + \sum_{y\in\mathbb{N}\setminus\{x\}} f(y)BN_{x,h}(y)
\end{aligned}$$

et

$$\begin{aligned}\text{var}\left\{\widetilde{f}_{n,h}(x)\right\} &= \frac{1}{n}f(x)\left\{\frac{(2x)!}{(x!)^2}\left(\frac{x+h}{2x+1+h}\right)^x\left(\frac{x+1}{2x+1+h}\right)^{x+1}\right\}^2 \\ &\quad +\frac{1}{n}\sum_{y\in\mathbb{N}\setminus\{x\}} f(y)BN^2_{x,h}(y) \\ &\quad -\frac{1}{n}\left\{f(x)\frac{(2x)!}{(x!)^2}\left(\frac{x+h}{2x+1+h}\right)^x\left(\frac{x+1}{2x+1+h}\right)^{x+1}+\sum_{y\in\mathbb{N}\setminus\{x\}} f(y)BN_{x,h}(y)\right\}^2.\end{aligned}$$

Pour n et x fixés, quand $h \to 0$, nous obtenons

$$\begin{aligned}\lim_{h\to 0}\text{biais}\{\widetilde{f}_{n,h}(x)\} &= f(x)\left\{\frac{(2x)!}{(x!)^2}\frac{x^x(x+1)^{x+1}}{(2x+1)^{2x+1}}-1\right\} \\ &\quad +\left(\frac{x+1}{2x+1}\right)^{x+1}\sum_{y\in\mathbb{N}\setminus\{x\}} f(y)\frac{(x+y)!}{x!y!}\left(\frac{x}{2x+1}\right)^y \neq 0\end{aligned}$$

et

$$\begin{aligned}\lim_{h\to 0}\text{var}\left\{\widetilde{f}_{n,h}(x)\right\} &= \frac{1}{n}f(x)\{1-f(x)\}\left\{\frac{(2x)!}{(x!)^2}\frac{x^x(x+1)^{x+1}}{(2x+1)^{2x+1}}\right\}^2 \\ &\quad +\frac{1}{n}\left\{\frac{1}{x!}\left(\frac{x+1}{2x+1}\right)^{x+1}\right\}^2\sum_{y\in\mathbb{N}\setminus\{x\}} f(y)\left\{\frac{(x+y)!}{y!}\left(\frac{x}{2x+1}\right)^y\right\}^2 \\ &\quad -\frac{2}{n}f(x)\frac{(2x)!}{(x!)^3}\frac{x^x(x+1)^{2(x+1)}}{(2x+1)^{3x+2}}\sum_{y\in\mathbb{N}\setminus\{x\}} f(y)\frac{(x+y)!}{y!}\left(\frac{x}{2x+1}\right)^y \\ &\quad -\frac{1}{n}\left(\frac{x+1}{2x+1}\right)^{2(x+1)}\left\{\sum_{y\in\mathbb{N}\setminus\{x\}} f(y)\frac{(x+y)!}{x!y!}\left(\frac{x}{2x+1}\right)^y\right\}^2.\end{aligned}$$

Ainsi, pour n fixé, la limite du $MISE$ lorsque $h \to 0$ de l'estimateur à noyau binomial négatif est différente du $MISE$ l'estimateur empirique.

Dans la Figure 1.4, pour $h > 0$ fixé et la distribution f choisie, le $MISE$ de l'estimateur à noyau binomial négatif est meilleur que celui du naïf en dessous d'une taille limite n assez petite qui est indiquée par l'intersection des courbes. Au delà de cette limite, l'estimateur naïf devient meilleur et son $MISE$ tend vers 0, ce n'est pas le cas du $MISE$ de l'estimateur à noyau binomial négatif même quand h tend 0 (très petit).

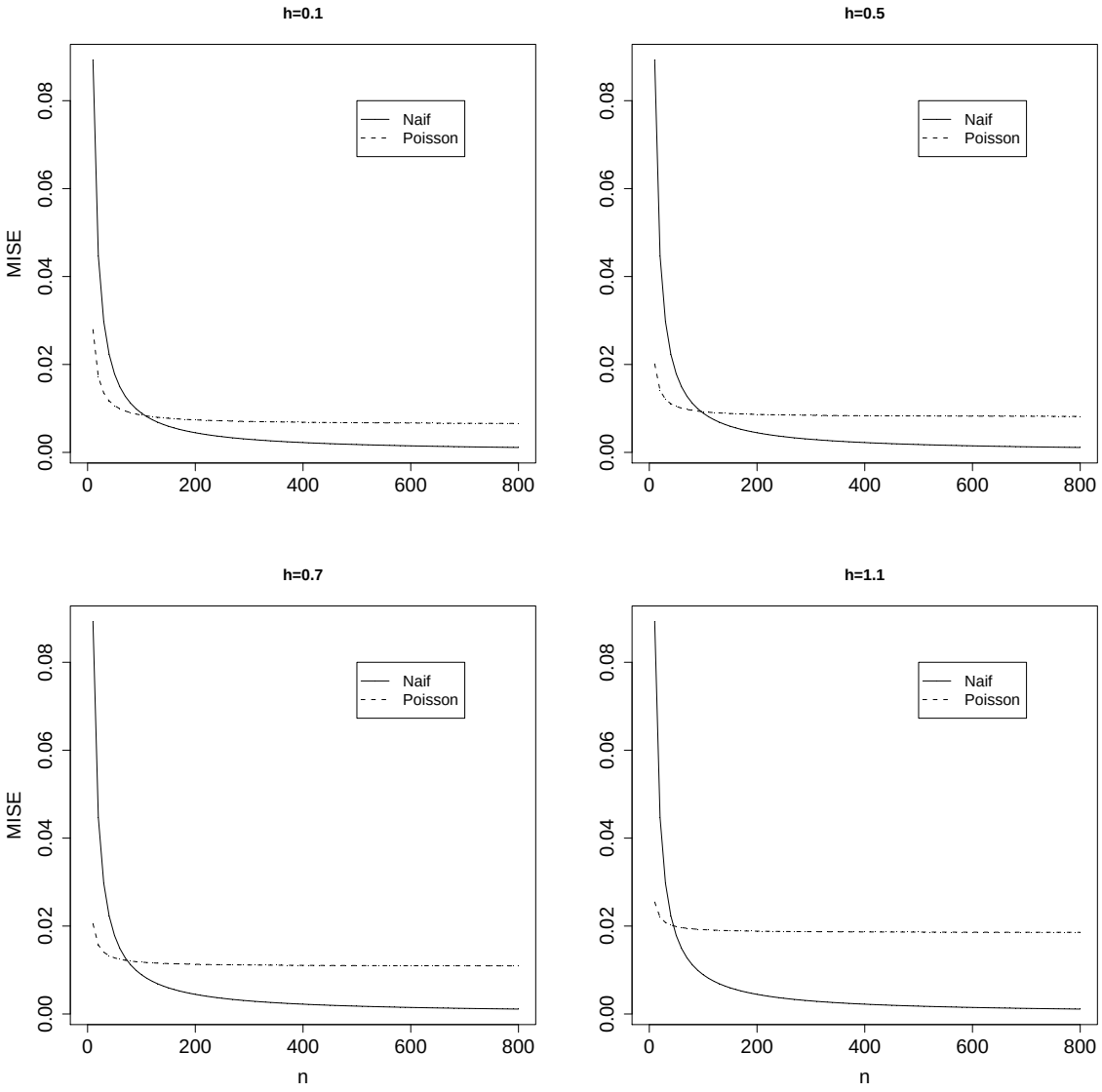

FIG. 1.2 – Comparaison entre $MISE(n, 0, D, f)$ de l'empirique et $MISE(n, h, f)$ de Poisson où f est la distribution du mélange de Poisson $f = 0.4\mathcal{P}(0.5) + 0.6\mathcal{P}(10)$

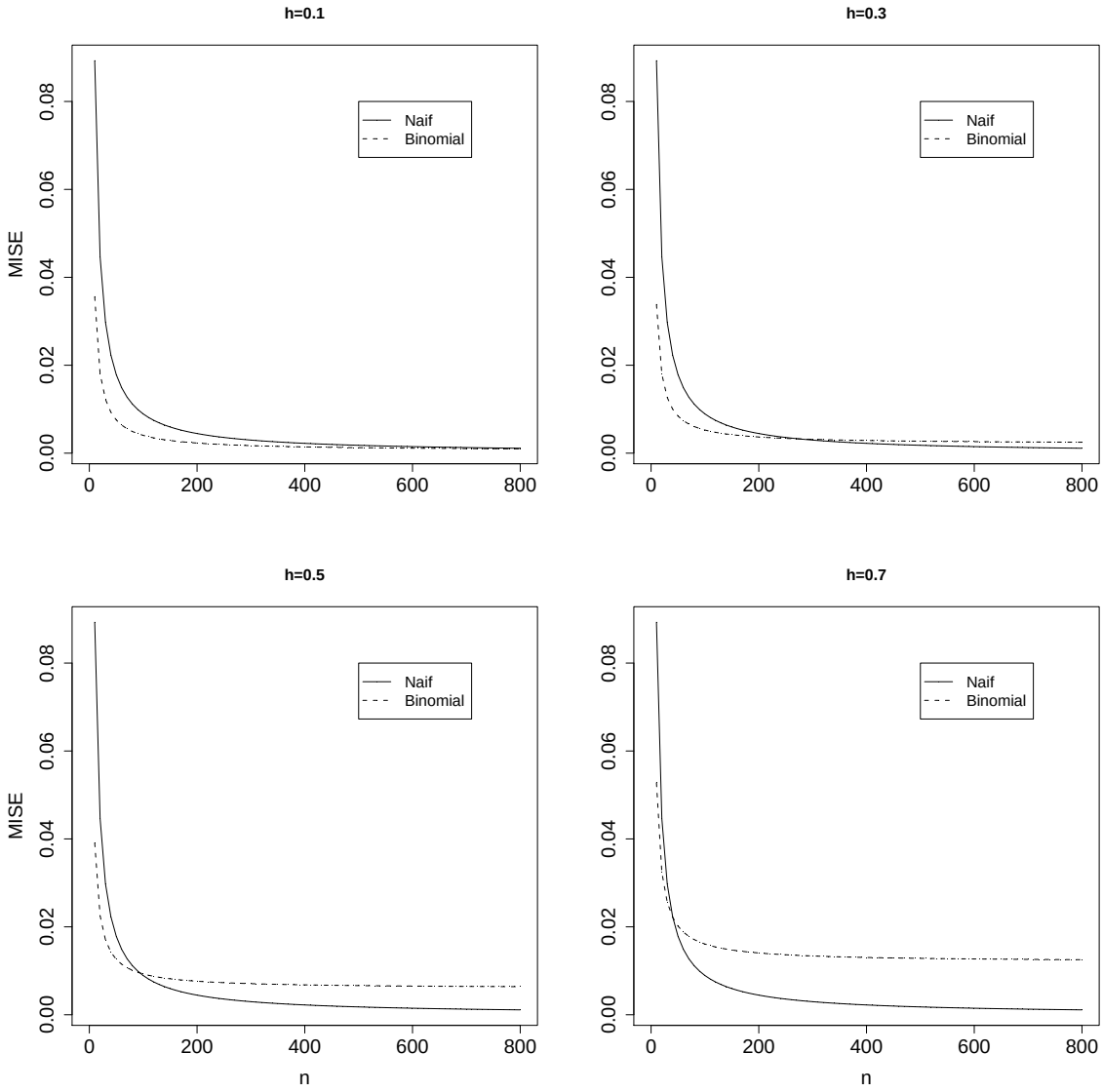

FIG. 1.3 – Suite de Figure 1.2 pour le noyau binomial

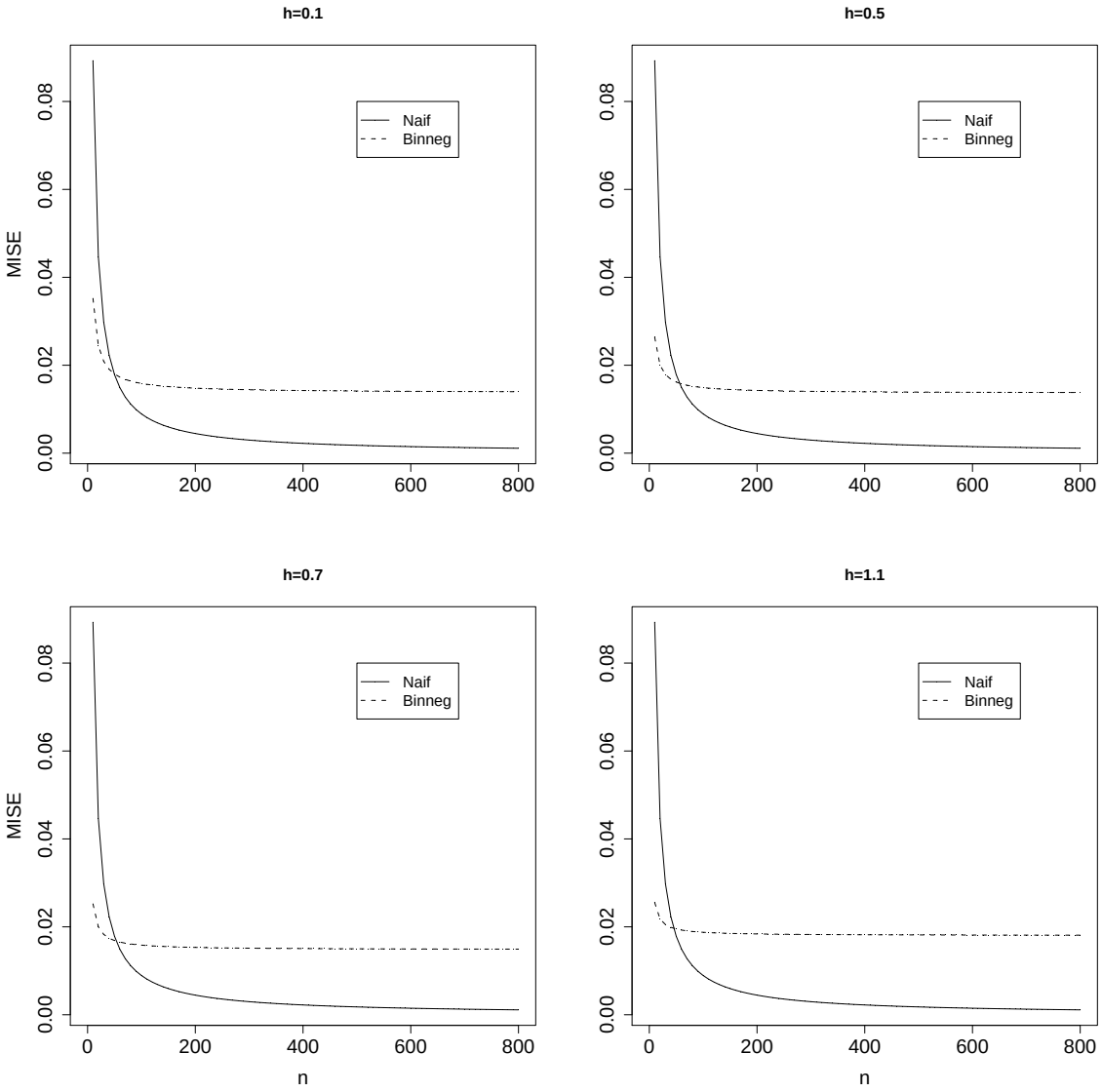

FIG. 1.4 – Suite et fin de Figure 1.2 pour le noyau binomial négatif

1.4 Performance relative des noyaux associés discrets

Nous proposons un indicateur de performance relative pour le noyau discret idéal satisfaisant

$$K_{id} = \arg\min_{K} MISE(n, h, K, f) = K_{id}(n, h, f).$$

Pour faire l'étude de $MISE$ (1.21) essentiellement dans le cas des noyaux discrets standards $\mathcal{K}_{x,h}$ pour lesquels var$(\mathcal{K}_{x,h}) \nrightarrow 0$ quand $h \to 0$, on pourrait faire un développement limité de var$\{\widetilde{f}_{n,h,K}(x)\}$ et de biais$^2\{\widetilde{f}_{n,h,K}(x)\}$ en fonction de n et h. Ceci dans le but de déduire des hypothèses sur h, n, $f(x)$ et K qui permettraient entre autres de sélectionner les noyaux les plus performants. Ici, nous nous appuyons sur le fait que l'étude de $MISE$ montre une différence importante au niveau de la partie biais (1.24) lorsque var$(\mathcal{K}_{x,h}) \nrightarrow 0$ quand $h \to 0$. Dans cette situation, le reste $R_n(x; h)$ n'est pas négligeable mais n'est pas facile à calculer. De plus, quand la variance var$(\mathcal{K}_{x,h})$ devient plus importante, il en va de même pour les probabilités individuelles $\Pr(\mathcal{K}_{x,h} = y)$, $y \in \aleph_x \setminus \{x\}$, et le reste $R_n(x; h)$. Alors, il faudrait faire une étude du comportement des noyaux discrets au cas par cas. De manière simple, nous ne présentons ici qu'un indicateur de performance. Ainsi, sous l'hypothèse d'égalité des espérances des noyaux discrets standards $\mathcal{K}_{x,h}$, nous pouvons nous intéresser essentiellement à leurs variances pour réduire le critère $MISE$.

Puisque les noyaux (associés) discrets $K_{x,h}$ correspondant aux estimateurs dépendent du support $\aleph$ de f à estimer ainsi que de chacune de cible $x \in \aleph$, on doit se restreindre dans une classe spécifique de noyaux discrets pour réaliser l'optimisation.

Ainsi, sans perte de généralité, on considère $\mathcal{K}^1_{x,h}$ et $\mathcal{K}^2_{x,h}$ deux variables aléatoires liées aux noyaux associés discrets (du premier ou du second ordre) $K^1_{x,h}$ et $K^2_{x,h}$, de supports comparables $\aleph^1_x$ et $\aleph^2_x$ respectivement. Sous la base de comparaison

$$\mathbb{E}\left(\mathcal{K}^1_{x,h}\right) = \mathbb{E}\left(\mathcal{K}^2_{x,h}\right), \quad \forall x \in \aleph \text{ et } h > 0, \tag{1.32}$$

le noyau discret K^1 est dit meilleur que le noyau discret K^2 si et seulement si

$$\text{var}\left(\mathcal{K}^1_{x,h}\right) \leq \text{var}\left(\mathcal{K}^2_{x,h}\right), \quad \forall x \in \aleph \text{ et } h > 0. \tag{1.33}$$

Le noyau (associé) discret le plus performant entre $K^0_{x,h}$ et $K^1_{x,h}$ peut être alors comparé au noyau naïf. Toutefois, l'égalité (1.32) entre la moyenne du noyau naïf et celle d'un noyau discret standard (1.13) n'est pas vérifiée. On ne peut donc pas utiliser le critère d'efficacité défini par la relation (1.33) pour un échantillon de petite taille n. Il faut alors évaluer l'ordre des grandeurs des expressions de $AMISE$ respectives. Ceci revient à choisir le noyau (associé) discret le plus adapté selon la taille n de l'échantillon (cf. Figures 1.2, 1.3 et 1.4).

Maintenant, considérons la condition (1.32) modifiée comme

$$\mathbb{E}\left(\mathcal{K}^j_{x,h}\right) = x + h + o(h), \quad j = 0, 1, \tag{1.34}$$

et l'hypothèse suivante sur les variances

$$\text{var}\left(\mathcal{K}_{x,h}^{j}\right) = V^{j}(x,h) + o(h), \;\; j = 0, 1, \tag{1.35}$$

avec $V^{j}(x,h) \geq 0$. Ainsi, sous les conditions modifiées (1.34) et (1.35), K^{1} est dit meilleur que K^{2} si et seulement si

$$V^{1}(x,h) \leq V^{2}(x,h), \;\; \forall x \in \aleph \text{ et } h > 0. \tag{1.36}$$

D'après le critère (1.36), parmi les types des noyaux discrets standards, un noyau associé discret est d'autant plus performant que si sa variance est petite. Donc, les noyaux sousdispersés sont plus performants que les noyaux équidispersés ou surdispersés. Ceci confirme la bonne qualité du noyau binomial. En effet, quand $h \to 0$, la variance du noyau binomial est $x/(x+1) < 1$, pour tout $x \in \mathbb{N}$. Tandis que sous la même condition celle du noyau Poisson est x et celle du noyau binomial négatif est $x(2x+1)/(x+1)$ (voir Table 1.1).

La Figure 1.5 représente sur une même échelle les noyaux associés discrets de premier ordre déjà présentés en Figure 1.1. Nous pouvons mieux y observer les différents comportements de dispersion autour de la cible choisie. Pour bien distinguer les distributions des différents noyaux discrets standards, nous ne les représentons pas ici avec des bâtons.

1.5 Choix de fenêtre

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

$$h_{id} = \arg\min_{h>0} MISE(n,h,K,f) = h_{id}(n,K,f). \tag{1.37}$$

Parmi ces méthodes, principalement trois d'entres elles seront mises en oeuvre dans ce travail.

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 serait garantie par la décroissance de la variance intégrée et la croissance du carré du biais intégré dans le risque quadratique global (1.21). 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. Ainsi, pour trouver la fenêtre optimale, on devrait balancer les approximations de la variance et du carré du biais. Autrement dit, il existerait $\varepsilon > 0$ telle que la fonction $h \mapsto AMISE(n,h,K,f)$ serait décroissante sur $]0,\varepsilon[$ et croissante sur $]\varepsilon,+\infty[$ pour tout $h > 0$.

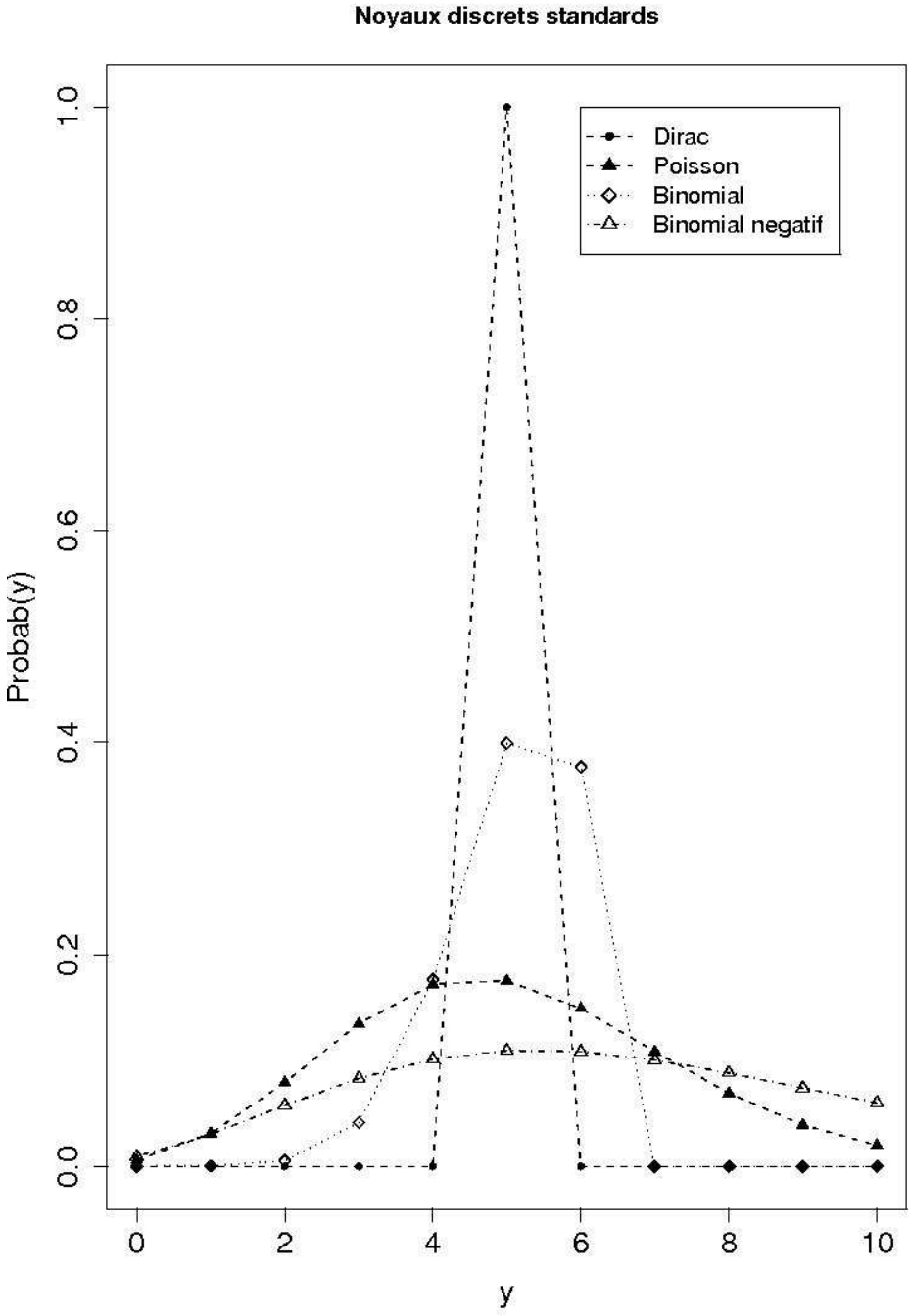

FIG. 1.5 – Noyaux discrets de type Dirac $\mathcal{D}(x)$, de Poisson $\mathcal{P}(x+h)$, binomial $\mathcal{B}\{x+1,(x+h)/(x+1)\}$ et binomial négatif $\mathcal{BN}\{x+1,(x+1)/(2x+1+h)\}$ avec $y=x=5$, $h=0.1$ (sauf pour Dirac)

La deuxième méthode est la validation croisée par moindres carrées, laquelle méthode est bien connue.

Enfin, la troisième procédure est la méthode des excés de zéros adaptée au cas fréquent d'une proportion importante de zéros dans les données de comptage.

D'autres choix de fenêtre sont possibles tels que le «Plug-in» (à adapter) et la validation croisée par maximum de vraisemblance où l'on cherche à minimiser la distance L de Kulleback-Leibler. Nous présenterons la dernière sans pour autant les mettre en oeuvre dans la section 1.6.

1.5.1 Minimisation des erreurs quadratiques

Soit $\underline{X} = (X_1, X_2, \ldots, X_n)$ un n-échantillon fixé i.i.d. de f et, donc, associé à la distribution empirique f_0 de f. Du point de vue purement pratique, nous proposons maintenant quelques types de fenêtres liées aux erreurs d'estimations. Le premier type de fenêtre est déduit de l'*erreur quadratique intégrée* (en anglais «Integrated Squared Error») définie par

$$ISE := \sum_{x\in\mathbb{N}} \left\{ \widetilde{f}_{n,h,K}(x) - f(x) \right\}^2 = ISE(\underline{X}; h, K, f), \tag{1.38}$$

laquelle mesure sur un seul échantillon $\underline{X}$ l'écart (au sens quadratique) entre $\widetilde{f}_n$ et f. Par conséquent, la minimisation en h de l'ISE (1.38) conduit à choisir une *fenêtre adéquate*

$$h^{**} = \arg\min_{h>0} ISE(\underline{X}; h, K, f) = h^{**}(n, K, f). \tag{1.39}$$

En remplaçant f par f_0 dans (1.38), on utilisera $h_0^{**} = h^{**}(n, K, f_0)$ pour le lissage discret d'un f_0 de f. Autrement dit, on a

$$h_0^{**} = \arg\min_{h>0} ISE(\underline{X}; h, K, f_0) = h_0^{**}(n, K, f_0). \tag{1.40}$$

Basé sur la convergence empirique de f_0 vers f quand $n \to +\infty$, on peut arriver à

$$\lim_{n\to+\infty} h_0^{**}(n, K, f_0) = \lim_{n\to+\infty} h^{**}(n, K, f), \tag{1.41}$$

pour un type de noyau discret K donné. La fenêtre adéquate h^{**} donnée en (1.39) de h est liée à la fenêtre optimale par la relation suivante :

$$MISE = \mathbb{E}(ISE) = \sum_{x\in\mathbb{N}} MSE(x). \tag{1.42}$$

La procédure fournissant h_0^{**} par (1.40) en se basant sur (1.41) est illustrée à travers des données simulées dans la prochaine section.

Quant à la fenêtre h^* obtenue par (1.29), on peut procéder de manière similaire qu'en (1.40) et (1.41) 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 (1.29) est loin d'être la plus appropriée dans le cas discret si l'allure (ou la régularité) de la fonction de masse f n'est pas sympathique ; c'est à dire si l'approximation des dérivées de f faite grossièrement par les différences finies (1.4) et (1.5) 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 (1.25). Finalement, d'après (1.29) et (1.42), la fenêtre optimale h^*_{opt} de h dans ce cas discret peut être obtenue à travers

$$h^*_{opt} = \arg\min_{h>0} \mathbb{E}\{ISE(\underline{X}; h, K, f)\}, \tag{1.43}$$

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

On peut aussi déterminer une valeur optimale moyenne de la fenêtre de lissage par l'approche de Monte Carlo. Pour cela, on simule un nombre n_{sim} d'échantillons j, $j = 1, 2, \ldots, n_{sim}$ de tailles finies n. Pour chaque échantillon simulé j, la valeur optimale est

$$\widehat{h}_j = \arg\min_{h>0} ISE_j(h),$$

qui minimise

$$ISE_j(h) = \sum_{x\in\mathbb{N}} \{\widetilde{f}_{(j)}(x) - f(x)\}^2.$$

La fenêtre optimale moyenne et sa variance s'obtiennent, respectivement, par

$$\overline{\widehat{h}} = \frac{1}{n_{sim}} \sum_{j=1}^{n_{sim}} \widehat{h}_j \text{ et } \operatorname{var}\big(\widehat{h}\big) = \frac{1}{n_{sim}-1} \sum_{j=1}^{n_{sim}} \big(\widehat{h}_j - \overline{\widehat{h}}\big)^2.$$

Nous n'avons pas établi de relation formelle entre $\overline{\widehat{h}}$ et h^*. Mais, ces deux types de fenêtres sont liées à travers la relation (1.42). Des résultats de simulations obtenues par cette méthode sont présentées en Annexe B.

1.5.2 Validation croisée par les moindres carrées

Considérons un noyau (associé) discret $K_{x,h}$, $x \in \mathbb{N}$ 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 estimer la valeur idéale h_{id} de h donnée en (1.37). La fenêtre optimale s'obtient par

$$h_{cv} = arg\min_{h>0} CV(h) \tag{1.44}$$

avec

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

où $\widetilde{f}_{n,h,K,-i}$ est calculé à partir de $\widetilde{f}_{n,h,K}$ 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, le critère $MISE$ de (1.20) peut être développé comme suit :

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

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

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

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 $\sum_{x\in\mathbb{N}} \widetilde{f}^2_{n,h,K}(x)$ qui est un estimateur sans biais de $\mathbb{E}\left\{\sum_{x\in\mathbb{N}} \widetilde{f}^2_{n,h,K}(x)\right\}$. Ensuite, soit

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

Par construction,

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

est un estimateur de $\mathbb{E}\left\{\sum_{x\in\mathbb{N}} \widetilde{f}_{n,h,K}(x)f(x)\right\}$ 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

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

D'autre part, on a successivement

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

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 (1987), Rudemo (1982), Stone (1984) et leurs références.

REMARQUE 1.5.1 : Pour des échantillons de tailles finies, la performance de la méthode validation croisée peut être examinée par la procédure des simulations de Monte Carlo (voir les résultats de simulations dans la Section B.5 de l'Annexe B).

1.5.3 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 $\underline{X} = (X_1, X_2, \ldots, X_n)$. Pour ce phénomène bien connu (voir, par exemple, Kokonendji *et al.*, 2007a, et leurs références) et étant donné un noyau discret $K_{x,h}$, on peut choisir une *fenêtre adaptée* $h_0 = h_0(\underline{X}; K)$ de h satisfaisant

$$\sum_{i=1}^{n}\Pr\left(\mathcal{K}_{X_i,h_0}=0\right) = n_0, \tag{1.45}$$

où $n_0 = \sharp\left(X_i = 0\right)$ désigne le nombre des zéros dans l'échantillon $\underline{X}$; voir Marsh & Mukhopadhyay (1999) pour leur noyau du type poissonnien différent du notre (cf. Exemple 1.2.1). L'équation (1.45) s'obtient à partir de l'expression

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

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

Cette fenêtre h_0 est un critère de choix de fenêtre comparable à la validation croisée. Selon l'importance de la proportion des zéros dans l'échantillon et du noyau discret retenu, la valeur de la fenêtre adaptée h_0 obtenue par (1.45) peut être plus ou moins proche de celles de la fenêtre idéale h_{id} de (1.37), de la fenêtre de validation croisée (1.44) ou de la fenêtre adéquate h_0^{**} de (1.40) (voir exemples illustrés en section 1.6). Nous n'avons pas encore établi de résultat de convergence, mais ce critère mériterait d'être approfondi.

Dans le cas du noyau de type de Poisson (Exemple 1.2.1), la fenêtre adaptée h_0 est connue explicitement. Tandis que dans le cas des noyaux de type binomial (Exemple 1.2.2) et binomial négatif (Exemple 1.2.3), la fenêtre h_0 est obtenue par la résolution numérique d'une équation non-linéaire (voir Table 1.2).

Type de noyau	h_0 tel que $\sum_{i=1}^{n} \Pr\left(\mathcal{K}_{X_i,h_0} = 0\right) = n_0$
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	$\sum_{i=1}^{n}\left(\frac{X_i+1}{2X_i+1+h_0}\right)^{X_i+1} = n_0$

TAB. 1.2 – Solutions h_0 pour les noyaux discrets standards

1.5.4 Minimisation de la distance de Kulleback-Leibler

Dans cette partie, on cherche à minimiser la distance L de Kulleback-Leibler entre f et $\widetilde{f}_n$ telle que

$$\begin{aligned} L(f,\widetilde{f}_n) &:= \sum_{x\in\mathbb{N}} f(x)\log\{f(x)/\widetilde{f}_n(x)\} \\ &= \sum_{x\in\mathbb{N}} f(x)\log\{f(x)\} - \sum_{x\in\mathbb{N}} f(x)\log\{\widetilde{f}_n(x)\}. \end{aligned} \tag{1.46}$$

Notons déjà que la distance L n'est pas une métrique et les critères définis en la minimisant ne sont pas appropriés dans le sens d'obtenir un lissage discret adéquat. Nous verrons que la convergence des critères est influencée par des noyaux et des distributions inconnues à longues queues.

Validation croisée par le maximum de vraisemblance

Nous déterminons une fenêtre optimale h_{LCV} telle que

$$h_{LCV} = \arg\max_{h>0} LCV(h),$$

qui maximise le critère

$$LCV(h) = \frac{1}{n}\sum_{i=1}^{n} \log \widetilde{f}_{n,h,K,-i}(X_i), \tag{1.47}$$

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

En effet, comme la distance $L(f, \widetilde{f}_n)$ dépend de la fonction inconnue f à estimer, il est nécessaire d'en déterminer un estimateur qui est une fonction de h. Le premier terme de l'expression (1.46) ne dépend pas de h ; il faut uniquement estimer le second terme.

Par construction, soit

$$I_n = \frac{1}{n}\sum_{i=1}^{n} \log \widetilde{f}_{n,h,K,-i}(X_i).$$

Comme les v.a. $X_1, X_2, \ldots, X_n$ sont i.i.d., d'une part, nous obtenons successivement

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

D'autre part, nous avons

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

Ainsi, on vient de montrer que $LCV(h)$ est un estimateur asymptotiquement sans biais de $\sum_{x\in\mathbb{N}} f(x)\log\{\widetilde{f}_n(x)\}$.

Validation croisée par Kullback-Leibler

De manière similaire à la procédure précédente, on cherche à minimiser la distance $L(f, \widetilde{f}_n)$ par rapport à h, en utilisant comme critère l'estimateur

$$\frac{1}{n}\sum_{i=1}^{n} \log\left[f(X_i)\left\{\frac{1}{n-1}\sum_{j\neq i} K_{X_i,h}(X_j)\right\}^{-1}\right] \tag{1.48}$$

de cette distance.

Les estimateurs construits sont très sensibles aux valeurs aberrantes. En effet, les points x situés dans la queue de la distribution à estimer ont de petites valeurs de masses de probabilité $f(x)$, ce qui conduit à de petites valeurs des estimations $\widetilde{f}_n(x)$ correspondantes. La présence du log dans les expressions (1.47) et (1.48), combinée aux petites valeurs de masses de probabilité au niveau des queues posent des problèmes de convergence de ces critères. Pour plus de détails sur ces trois dernières procédures de choix de fenêtre, on peut se référer à Bowman (1984), Hall (1987) et leurs références.

1.6 Illustrations

Dans cette section, nous illustrons certains résultats à travers des jeux de données simulées et réelles.

1.6.1 Données simulées

A l'aide des données de dénombrement simulées, nous mettons en évidence certains aspects des estimateurs à noyaux discrets standards. D'une part, nous comparons les trois types de noyaux discrets, à savoir Poisson (ou equidispersé), binomial (ou sousdispersé) et binomial négatif (ou surdispersé). D'autre part, nous examinons les comportements des approximations des dérivées $f^{(k)}(x)$, $k \in \{1, 2\}$, par les différences finies (1.4) et (1.5) dans les $AMISE^*$ en comparaison avec le $MISE$ dans les cas des noyaux discrets standards. De plus, nous étudions les performances des estimateurs pour un lissage discret selon la fenêtre idéale h_{id} de (1.37), la fenêtre h_{cv} de validation croisée (1.44) et la fenêtre h_0 des excès de zéros (1.45). 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 (1.38) et reliée à $MISE$ par (1.42). Dans le cas (1.40) où f est remplacée par la distribution empirique f_0, le critère ISE devient alors

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

Plusieurs types de lois discrètes ont été considérés pour réaliser ces études par données simulées. Pour ce document, nous retenons la fonction de masse

$$f(x) = 0.4\, e^{-0.5}\, 0.5^x/x! + 0.6\, e^{-10}\, 10^x/x!, \quad x \in \mathbb{N}, \tag{1.49}$$

qui est un mélange de deux lois de Poisson de moyennes $\mu_1 = 0.5$ et $\mu_2 = 10$. Cette distribution 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 $1 - \sum_{x=0}^{21} f(x) = 4.198 \times 10^{-4}$.

1.6.2 Approximation du risque quadratique intégré

Pour f définie en (1.49), les Figures 1.6, 1.7 et 1.8 rapportent quelques allures de $MISE(f)$, $AMISE^*(f, f^{(1)}, f^{(2)})$, $AMISE^*(f, f^{(1)})$ 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 $AMISE^*(f, f^{(1)})$ est obtenu en supprimant les termes de second ordre $f^{(2)}$ dans $AMISE^*(f, f^{(1)}, f^{(2)}) := AMISE^*(n, h, f)$. Pour les données simulées, on calcule $f^{(1)}$ et $f^{(2)}$ à partir des formules des différences finies (1.4) et (1.5) en utilisant la fonction f qui ici est connue.

On remarque alors que les approximations des dérivées de f par des différences finies sont globalement satisfaisantes ; les approximations $AMISE^*$ du $MISE$ 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 masse de probabilité f. Les ordres de grandeur des $AMISE^*$ et ISE correspondent à celui du vrai $MISE$; de plus, ils sont comparables à la fois pour une taille d'échantillon donnée et pour un type de noyau discret standard choisi. Ces résultats valident l'utilisation des approximations $AMISE^*$.

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 définie en (1.49) 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 des répliques est aussi appliqué pour les observations qui vont suivre.

Dans la Table B.4 de la Section B.2 (Annexe B), nous présentons des simulations de $AMISE^*$ pour les estimateurs à noyaux discrets standards et l'estimateur naïf quand n augmente.

1.6.3 Validation croisée et excès de zéros

Dans cette partie, nous illustrons entre autre la nouveauté (1.45) des excès de zéros parmi les nombreux critères de choix de fenêtre $h > 0$ ainsi que les performances comparatives des estimateurs à noyaux discrets standards. Pour cela, nous utilisons des données simulées de f définie en (1.49) de taille $n \in \{50, 100, 300, 1000\}$.

Les Tables 1.3, 1.4 et 1.5 présentent les différentes qualités de lissage discret par des estimations à noyaux discrets selon la fenêtre idéale h_{id} de (1.37) qui minimise le $MISE$, la fenêtre adaptée h_0 de proportion de zéros (1.45), la fenêtre optimale de validation croisée (1.44) et les fenêtres h_0^{**} et h^{**}. On y insère la constante de normalisation $C = \sum_{x\in\mathbb{N}} \widetilde{f}_{n,h,K}(x)$ introduite en (1.17), laquelle a tendance à surestimer pour ces données ($C > 1$). Puisque la proportion de zéros est assez importante dans les

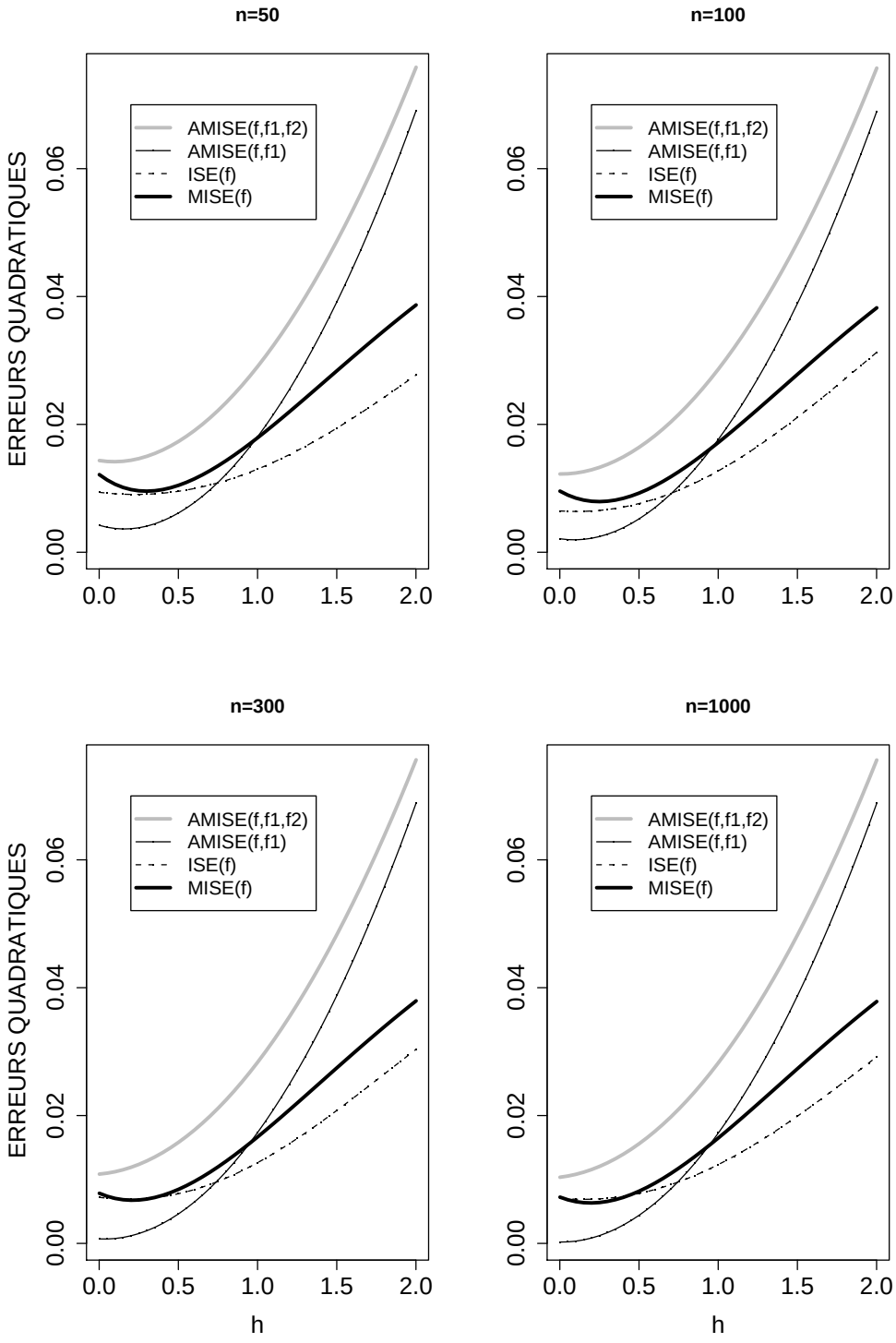

FIG. 1.6 – Graphiques des erreurs quadratiques de l'estimateur à noyau de Poisson pour la distribution du mélange de Poisson $f = 0.4\mathcal{P}(0.5) + 0.6\mathcal{P}(10)$

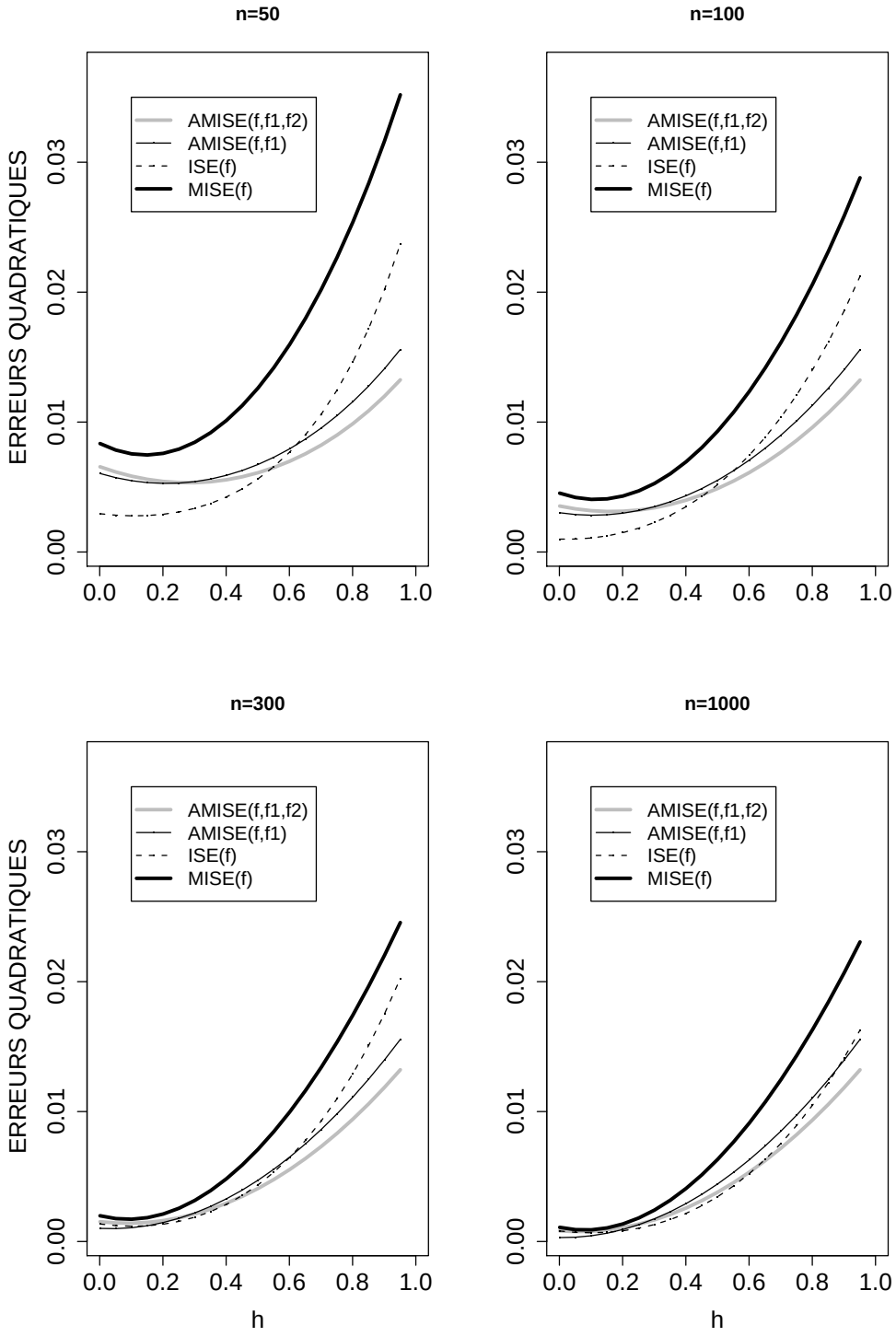

FIG. 1.7 – Suite de Figure 1.6 pour le noyau binomial

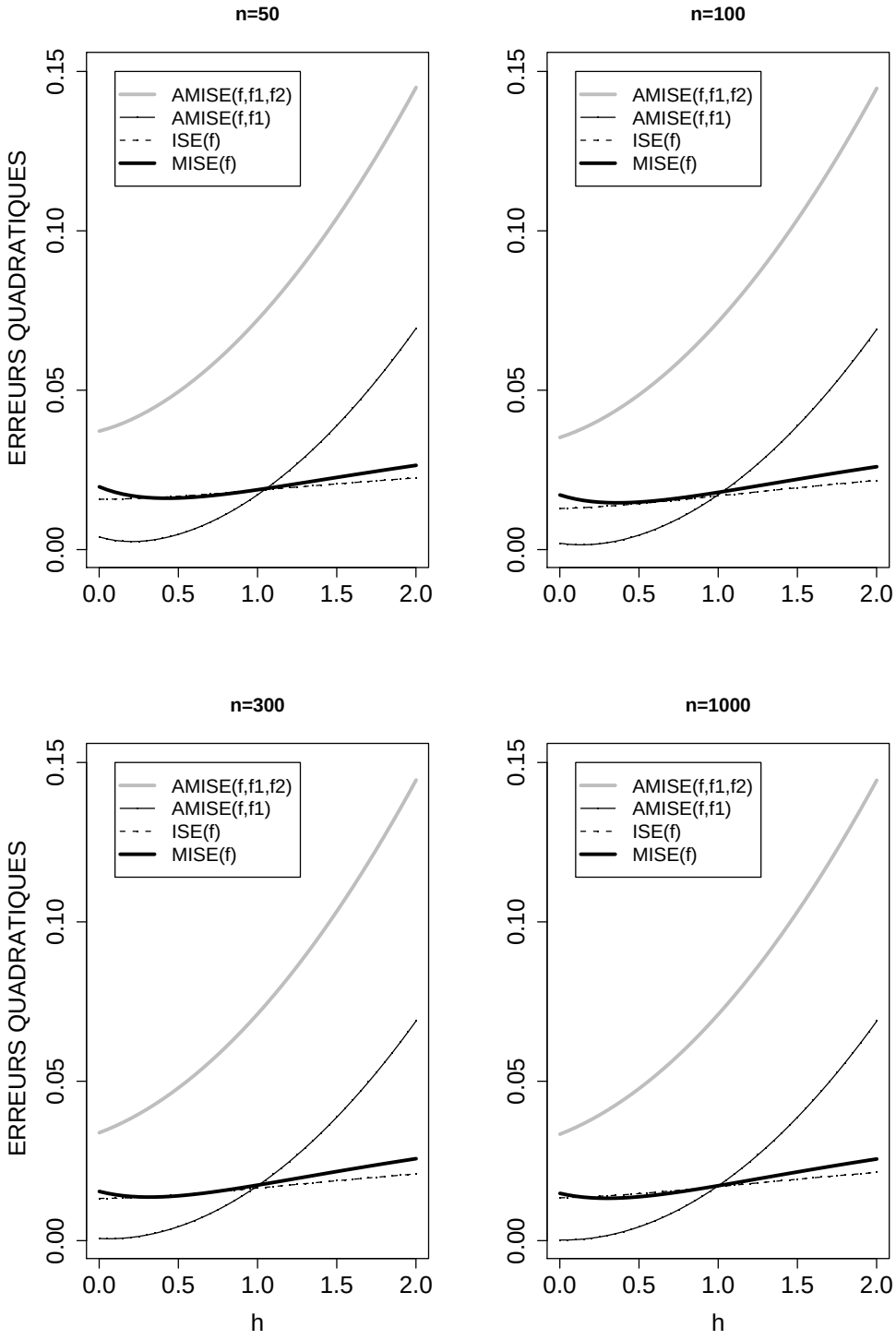

FIG. 1.8 – Suite et fin de Figure 1.6 pour le noyau binomial négatif

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_{id}, 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 h^{**} dans la Table 1.3. De plus, nous retrouvons l'ordre de préférence parmi ces trois types de noyaux standards : binomial, Poisson et binomial négatif.

La Figure 1.9 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 définie en (1.49) n'est pas satisfaisant. Ce n'est que dans le cas $n = 1000$, donc n grand, que le lissage discret à l'aide du noyau naïf est entièrement convenable.

Les Figures 1.10, 1.11, 1.12 et 1.13 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 et celui à noyau binomial donnent des qualités d'ajustements très proches. 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. La faible convergence des noyaux discrets standards poissonnien et binomial négatif explique en partie ces mauvais ajustements. Le noyau binomial fournit de meilleurs résultats car il a la variance la plus petite et qui est toujours infèrieure à 1.

Dans la pratique, notons que concernant le choix de fenêtre, les fonctions des erreurs $h \mapsto ISE(h)$ et $h \mapsto ISE^0(h)$ ou de la validation croisée $h \mapsto CV(h)$ sont croissantes pour certains jeux de données. Dans ces situations particulières, nous avons fixé la valeur de la fenêtre qui minimise ces fonctions à 10^{-3}.

		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

TAB. 1.3 – Qualités de lissages discrets par les noyaux de type binomial, Poisson et binomial négatif des données simulées de la distribution du mélange de Poisson $f = 0.4\mathcal{P}(0.5) + 0.6\mathcal{P}(10)$ avec la fenêtre h_0^{**} (1.40) et h^{**} (1.39)

	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
$AMISE^*(f)$		**0.00449**	0.01586	0.04987
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
$AMISE^*(f)$		**0.00459**	0.01456	0.04527
h_{id}		0.146	0.298	0.426
C		0.99001	1.03384	1.07848
$MISE(f)$	0.01786	**0.00747**	0.00955	0.01611
$n = 100$				
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
$AMISE^*(f)$		**0.00255**	0.01327	0.04379
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
$AMISE^*(f)$		**0.00260**	0.01282	0.04162
h_{id}		0.116	0.248	0.362
C		1.00395	1.04940	1.09878
$MISE(f)$	0.00893	**0.00404**	0.00792	0.01468

TAB. 1.4 – Qualités de lissages discrets par les noyaux de type Dirac, binomial, Poisson et binomial négatif des données simulées de la distribution du mélange de Poisson $f = 0.4\mathcal{P}(0.5) + 0.6\mathcal{P}(10)$ avec les fenêtres h_{cv} (1.44), h_0 (1.45) et h_{id} (1.37)

		Naïf	Binomial	Poisson	Binomial négatif
$n = 300$					
	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
	$AMISE^*(f)$		**0.00120**	0.01181	0.04116
	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
	$AMISE^*(f)$		**0.00120**	0.01172	0.04019
	h_{id}		0.090	0.209	0.316
	C		1.01354	1.06690	1.11693
	$MISE(f)$	0.00298	**0.00171**	0.00675	0.01366
$n = 1000$					
	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
	$AMISE^*(f)$		**0.00068**	0.01137	0.04087
	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
	$AMISE^*(f)$		**0.00072**	0.01147	0.04063
	h_{id}		0.080	0.195	0.299
	C		1.01503	1.07077	1.12481
	$MISE(f)$	0.00089	**0.00088**	0.00632	0.01329

TAB. 1.5 – Suite et fin de Table 1.4 pour $n \in \{300, 1000\}$

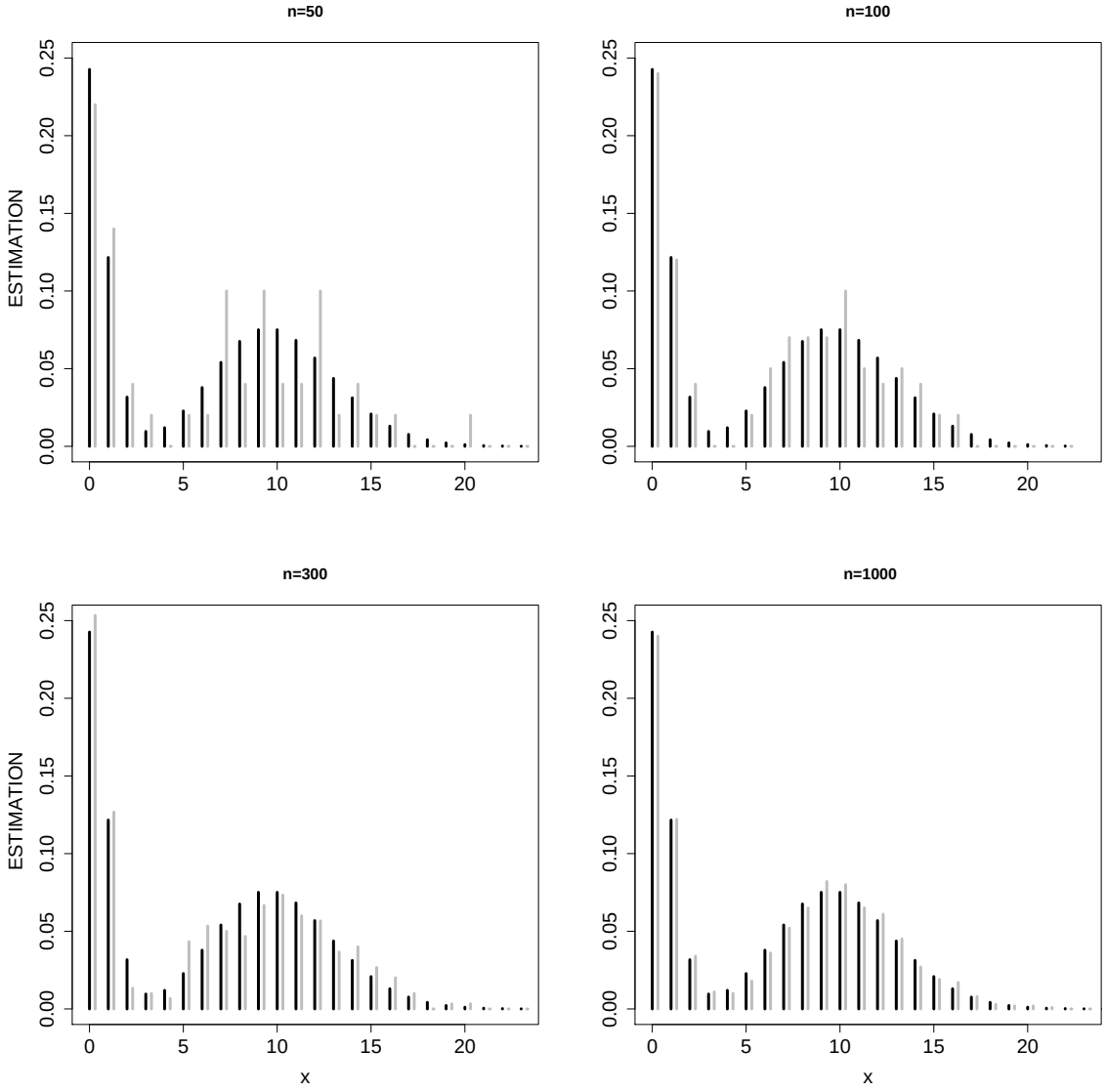

FIG. 1.9 – Lissages discrets par l'estimateur empirique des données simulées pour $n \in \{50, 100, 300, 1000\}$ de la distribution du mélange de Poisson $f = 0.4\mathcal{P}(0.5) + 0.6\mathcal{P}(10)$. Les barres en noir correspondent aux valeurs de la vraie distribution et celles en gris représentent les estimations discrètes obtenues

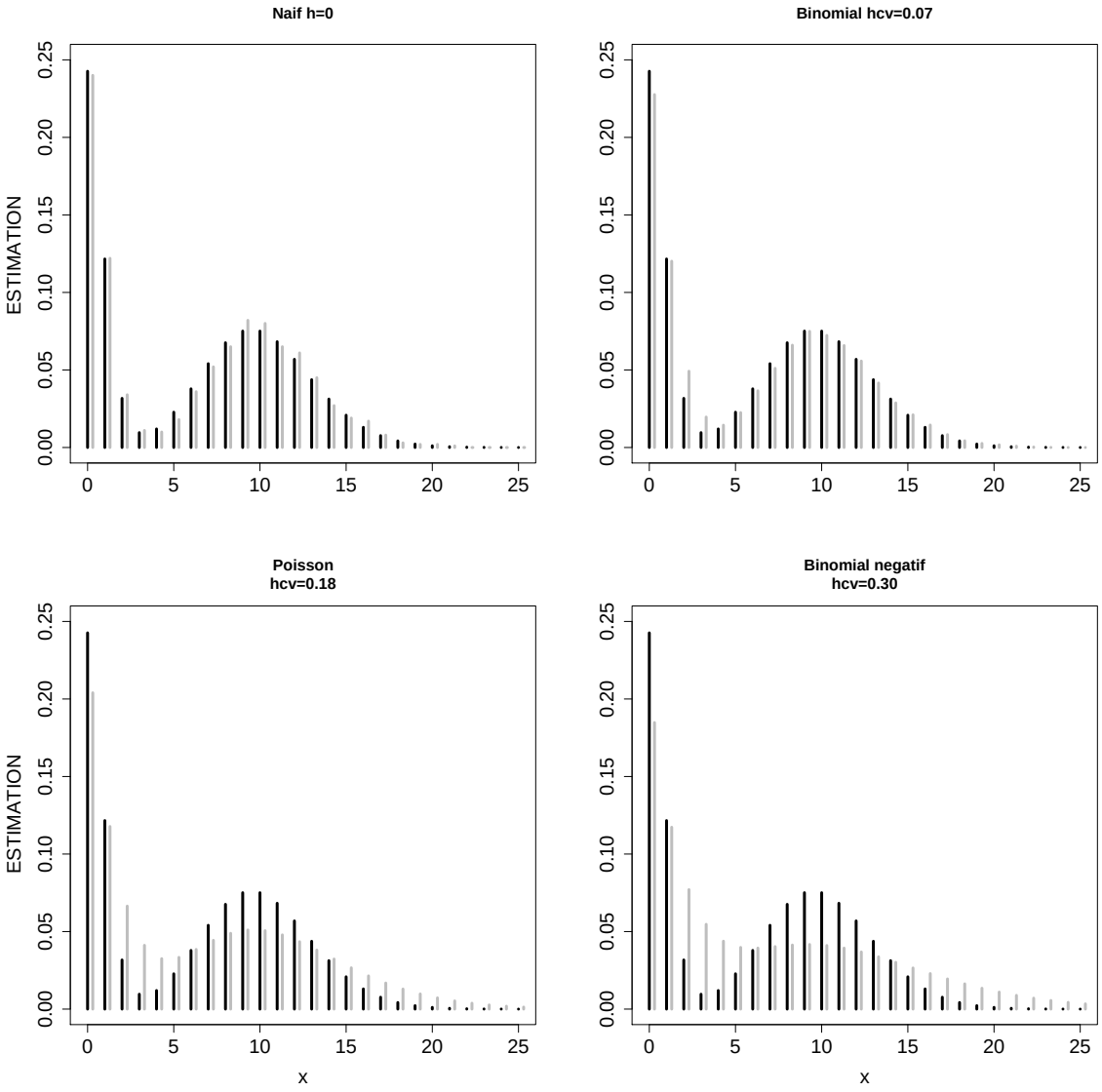

FIG. 1.10 – Lissages discrets par les noyaux de type Dirac, binomial, Poisson et binomial négatif des données simulées ($n = 1000$) de la distribution du mélange de Poisson $f = 0.4\mathcal{P}(0.5) + 0.6\mathcal{P}(10)$. Les barres en noir correspondent aux valeurs de la vraie distribution et celles en gris représentent les estimations discrètes obtenues

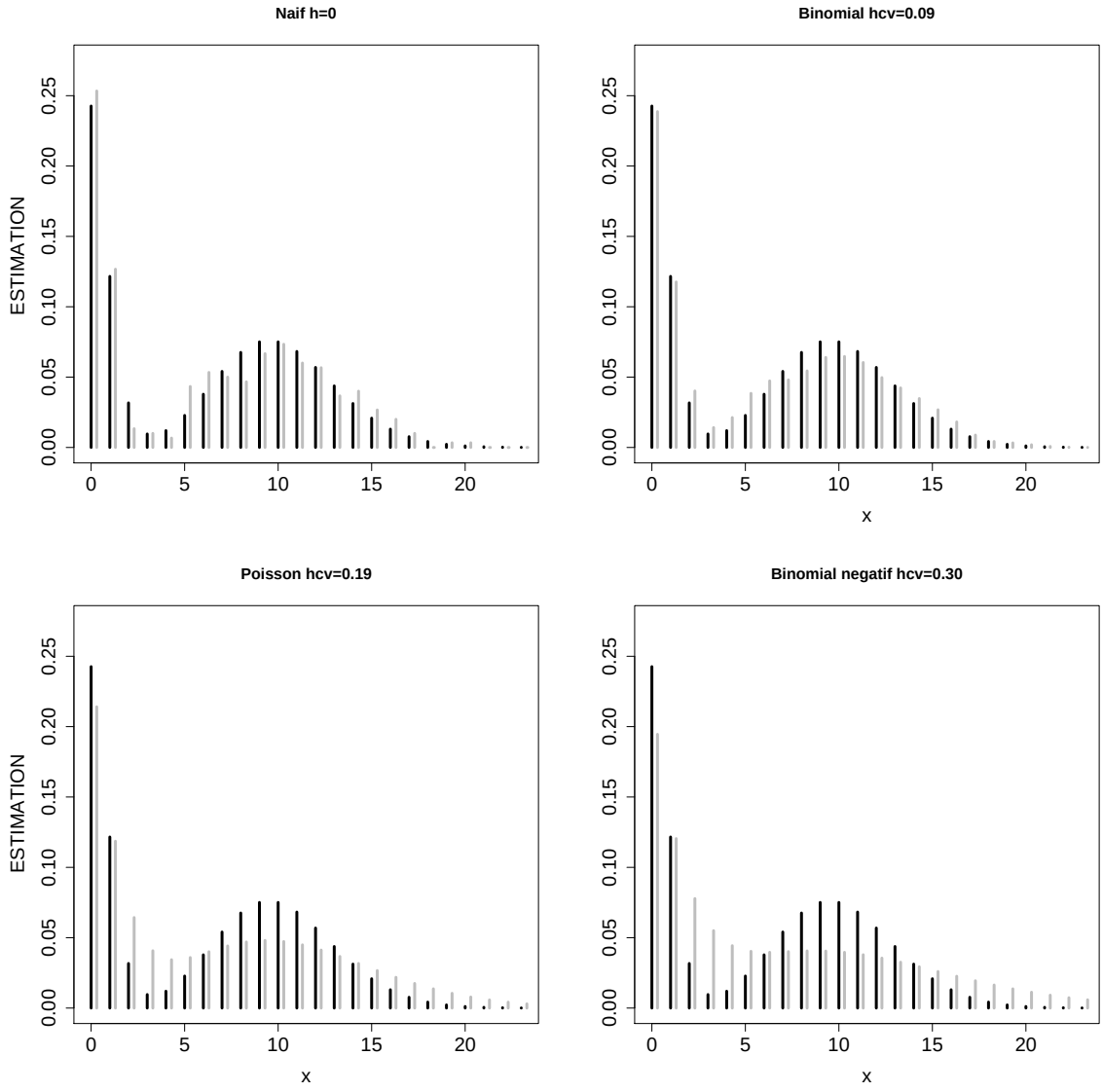

FIG. 1.11 – Suite de Figure 1.10 pour $n = 300$

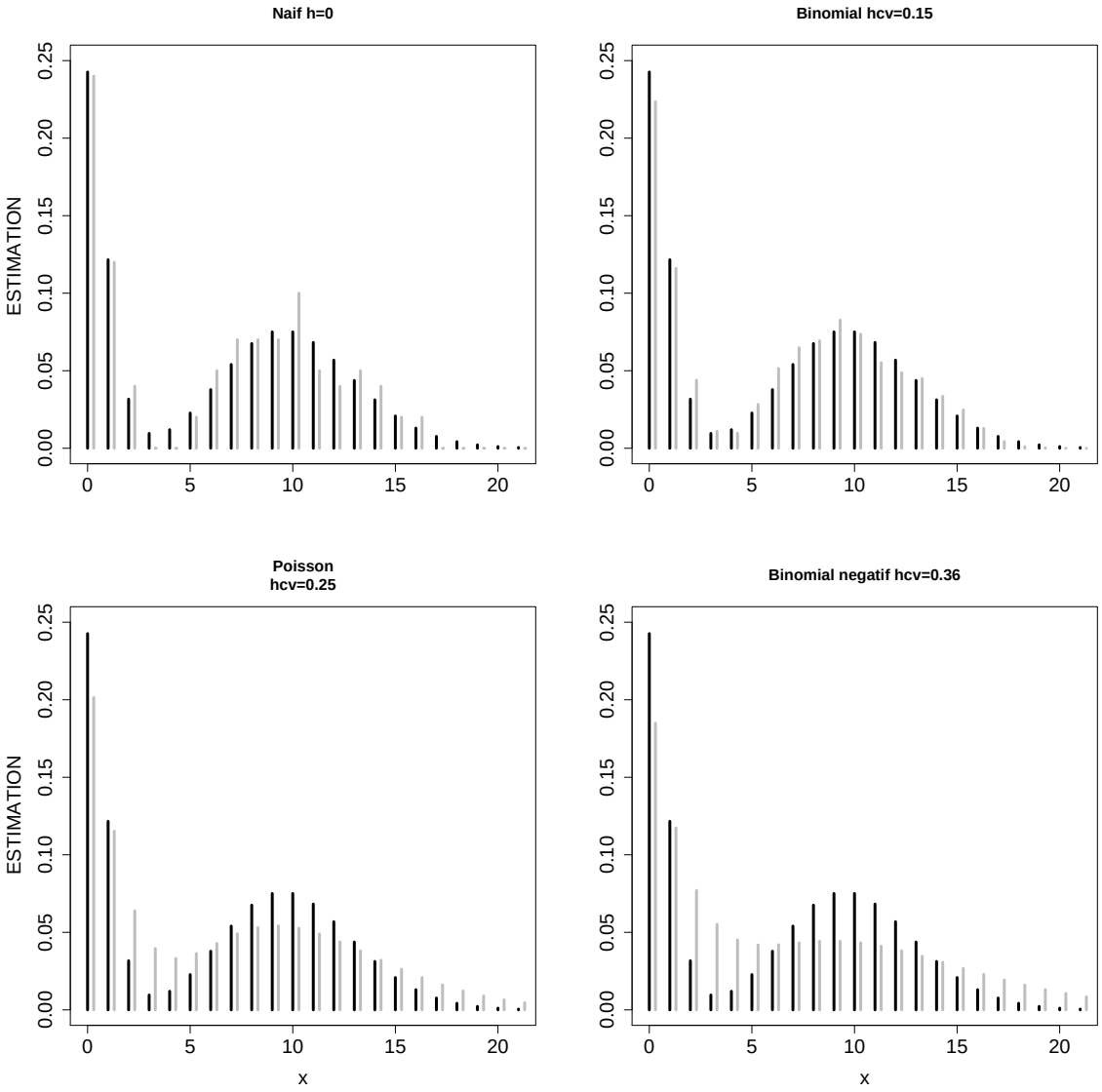

FIG. 1.12 – Suite de Figure 1.10 pour $n = 100$

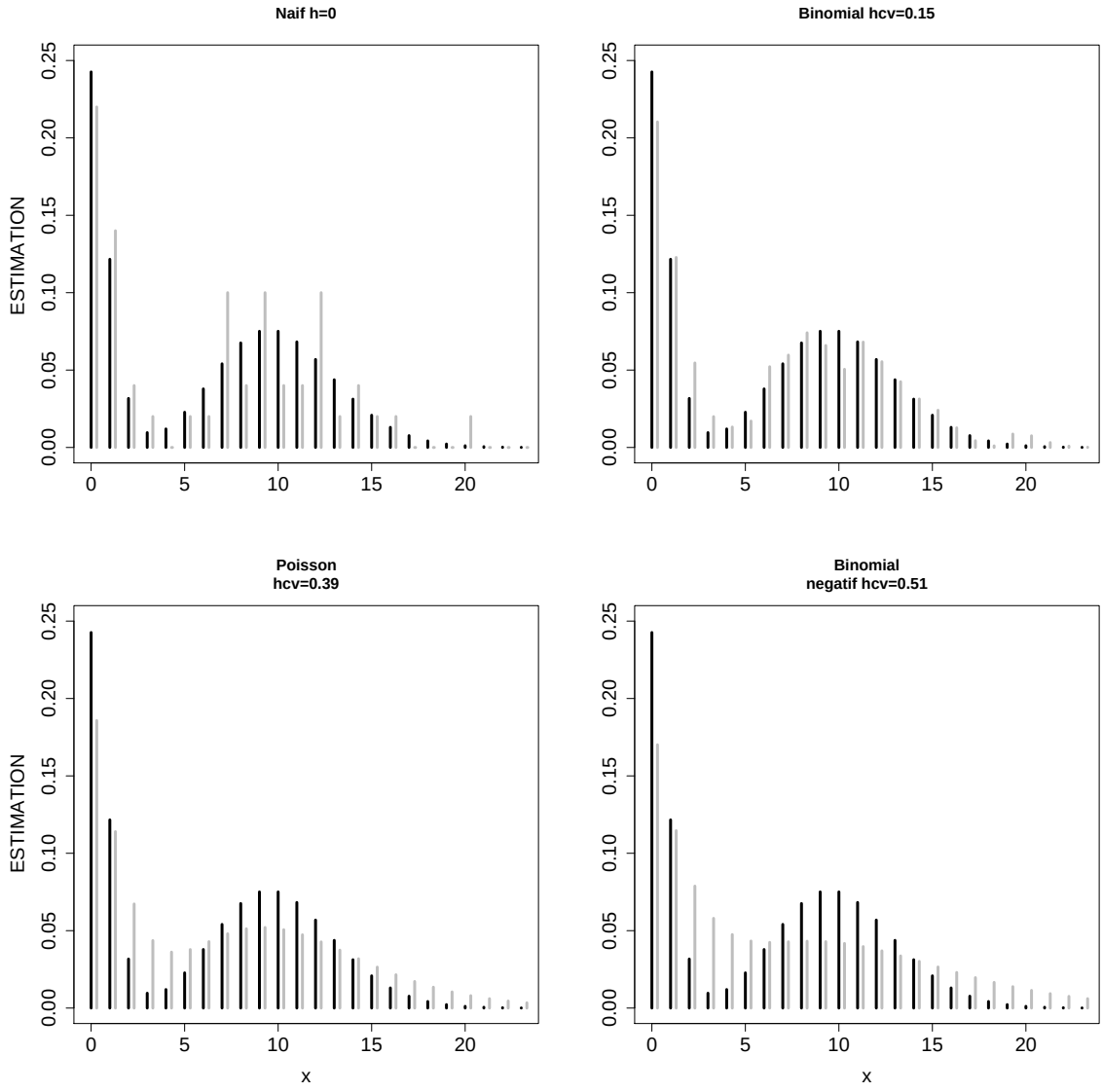

FIG. 1.13 – Suite et fin de Figure 1.10 pour $n = 50$

1.6.4 Données de buts

Dans la nature, il existe de nombreux exemples de données de dénombrement 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 la distribution de buts par matchs $\{0, 1, 2, \cdots\}$ d'un championnat de football. Les observations relatives aux résultats numériques ci-dessous complètent celles des données simulées du précédent paragraphe. Notons au passage qu'on peut se referer à Karlis & Ntzoufras (2003) pour des modèles paramétriques relatifs à l'influence des buts marqués par les équipes durant une rencontre.

Données

Nous considérons les données de la Table 1.6 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 dans la saison. En comparant ces deux championnats, on peut alors justifier le nouveau classement de 2006-2007 de la Ligue Professionnelle de Football en France, appelé classement de l'offensive*. 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

TAB. 1.6 – 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

En fait, d'après aussi la Table 1.7, la Ligue 1 française possède un déficit de 123 buts par rapport à la Liga espagnole. On remarque également que les données de buts

*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

	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

TAB. 1.7 – Résumé des statistiques de la Table 1.6 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)

de Ligue 1 sont surdispersées ($s_g^2/\overline{g} = 1.113 > 1$) alors que celles de Liga sont sous-dispersées ($s_g^2/\overline{g} = 0.904 < 1$). 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 est presque le double de celle de Liga $f_{02}(0) = 0.071$. Dans le championnat italien, cette proportion est probablement plus grande à cause de la technique défensive connue sous le nom de «catenaccio». Dans la suite, nous donnons uniquement les résultats relatifs au lissage discret des données de la Ligue 1 française.

Résultats

Les Table 1.8 et Figure 1.14 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 modérée $n = 380$, nous avons présenté les résultats avec la fenêtre h_0 de proportion de zéros (1.45) ainsi que les fenêtres h_{cv} de validation croisée (1.44) et h_0^{**} définie en (1.41).

Les meilleurs ajustements de f_{01} au sens du critère ISE^0 minimal par rapport aux différents choix de fenêtre 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 une estimation non-paramétrique $\widetilde{f}_{n,h,K}$ de chaque championnat, on peut alors estimer le nombre $n \times \widetilde{f}_{n,h,K}(g)$ de rencontres étant soldées par un nombre g donné de buts dans une saison régulière à n matchs.

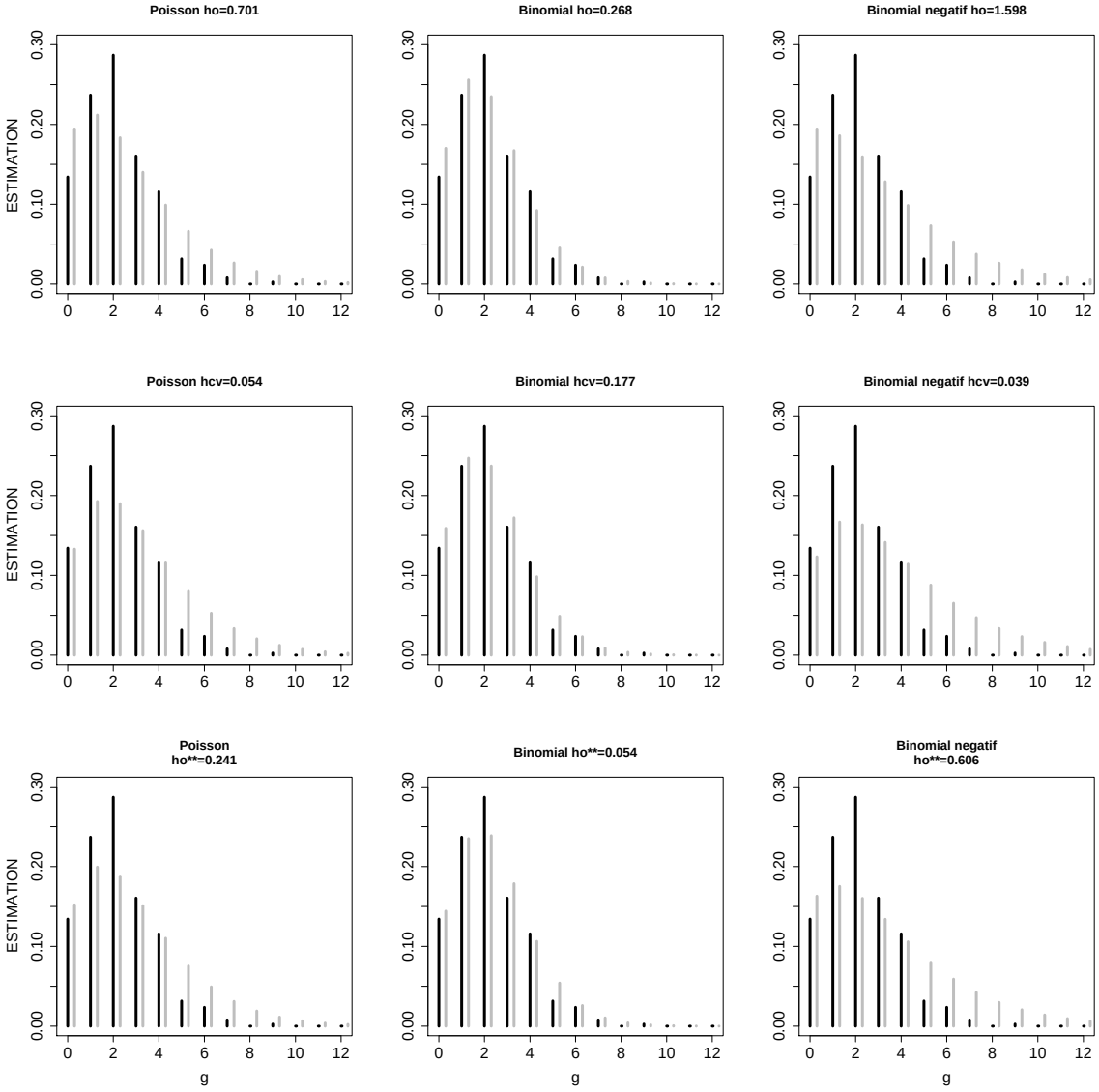

FIG. 1.14 – Lissages discrets (bâtons gris) par les noyaux de type de Poisson, binomial et binomial négatif pour les données réelles (bâtons noirs) de football de Ligue 1 française $n = 380$

	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

TAB. 1.8 – Qualités de lissages discrets par les trois types de noyaux discrets standards pour les données réelles de football de Ligue 1 française pour $n = 380$ avec les fenêtres h_0 (1.45), h_{cv} (1.44) et h_0^{**} (1.40)

1.7 Conclusion

L'estimateur à noyau discret ouvre une voie pour une approche non-paramétrique des données discrètes dans divers domaines d'application. Il fournit un large degré de flexibilité. En effet, contrairement au cas paramétrique, le modèle basé sur l'estimateur à noyau discret ne dépend pas des phénomènes de surdispersion ou de sousdispersion des données de comptage (*e.g.* Mizère *et al.*, 2006). Il possède des propriétés satisfaisantes et moins contraignantes que le dans le cas d'un estimateur à noyau continu. L'approximation des dérivées du cas continu par des différences finies dans le cas discret en est un exemple.

L'étude des noyaux discrets standards révèle une nette préférence pour le noyau binomial grâce à la propriété de sousdispersion. En effet, cette propriété induit un biais plus petit pour l'estimateur à noyau discret comparativement au cas d'équidispersion (loi de Poisson) ou de surdispersion (loi binomiale négative). Cependant, les résultats obtenus par le noyau binomial sont à améliorer notamment par la recherche d'un noyau associé discret vérifiant toutes les hypothèses de Définition 1.2.1, assurant au moins la convergence en moyenne quadratique de l'estimateur. Toutefois, même si la variance des noyaux proposés ne tend pas vers 0, ce n'est pas une condition superflue. Elle

prend plus d'importance pour des échantillons de petite taille, d'où la nécessité d'avoir la variance du noyau associé la plus petite. Ainsi, contrairement à la situation continue et symétrique, le noyau associé discret joue un rôle décisif.

Pour un noyau discret donné, le choix de la fenêtre de lissage discret est aussi important. Parmi de nombreuses techniques existantes pour ce choix, la méthode de validation croisée, souvent conseillée dans le cas continu, demeure applicable dans le cas discret. Par ailleurs, dans la situation où l'échantillon présente une importante proportion de zéros, on peut utiliser la fenêtre adaptée h_0 des excès de zéros (1.45) pour un noyau associé discret (standard) ne présentant pas de problème de biais de bordure en zéro. En fait, dans tout ce chapitre, les noyaux proposés ne posent pas de problème de biais de bordure au sens où en chaque point x leur support $\aleph_x$ est tel que $\aleph_x \subseteq \mathbb{N}$.

Chapitre 2

Noyaux associés discrets triangulaires

2.1 Introduction

Dans le chapitre précédent, nous avons introduit les estimateurs à noyau discret définis de la manière suivante. Soit X_1, X_2, ..., X_n un n-échantillon i.i.d. de fonction de masse de probabilité inconnue f sur un ensemble discret $\aleph$. Etant donné un type de noyau discret K, un estimateur à noyau discret de f est donné par

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

où $h \geq 0$ est le paramètre de lissage (ou fenêtre) et $K_{x,h}$ est le noyau associé discret lié à une variable aléatoire (v.a.) discrète $\mathcal{K}_{x,h}$ sur $\aleph_x$. Rappelons la définition du noyau associé discret $K_{x,h}$ [Chapitre 1, Section 1.2, Définition 1.2.1] tel que $\cup_{x\in\aleph}\aleph_x \supseteq \aleph$,

$$\mathbb{E}(\mathcal{K}_{x,h}) \sim x \quad \text{quand} \quad h \to 0, \tag{2.1}$$

$$\text{var}(\mathcal{K}_{x,h}) < +\infty, \tag{2.2}$$

$$\text{var}(\mathcal{K}_{x,h}) \to 0 \quad \text{quand} \quad h \to 0. \tag{2.3}$$

De plus, on dispose de la relation suivante :

$$\mathbb{E}\{\widetilde{f}_{n,h,K}(x)\} = \mathbb{E}\{f(\mathcal{K}_{x,h})\}. \tag{2.4}$$

Dans le cas des lois de probabilités discrètes usuelles telles que Poisson, binomiale et binomiale négative (voir aussi dans Johnson *et al.*, 2005), il n'est pas facile de construire des noyaux associés discrets qui vérifient toutes les conditions précédentes. Dans le chapitre précédent, nous avons construit des noyaux discrets standards ayant les propriétés (asymptotiques) générales suivantes :

$$\mathbb{E}(\mathcal{K}_{x,h}) = x + h + o(h) \quad \text{et} \quad \text{var}(\mathcal{K}_{x,h}) = V_K(x,h) + o(h),$$

où $V_K(x,h)$ est positive et finie mais ne tend pas nécessairement vers 0. Il se révèle nécessaire de définir judicieusement les paramètres du noyau associé discret de manière à obtenir la condition (2.3), *i.e.*, $V_K(x,h) \to 0$ quand $h \to 0$. Pour le lissage des fonctions de masses de probabilité, les estimateurs à noyaux discrets sont plus appropriés que ceux à noyaux continus. Cependant, le choix du noyau associé discret joue un rôle crucial dans cette approche contrairement au cas bien connu des noyaux continus symétriques ; voir, par exemple, Devroye (1987) et Tsybakov (2004) pour une généralité des données (supposées) continues, Scott (2000) pour le cas multivarié, Ferraty & Vieu (2006) pour des données fonctionnelles, Simonoff (1996) pour des données catégorielles ordonnées. Entre autre, la fonction noyau continu possède classiquement la propriété de symétrie. Cependant, si la densité continue à estimer est à support compact ou borné d'un côté (*e.g.*, $\mathbb{N}$, $[0,+\infty[$), la symétrie de la fonction noyau continu crée le phénomène de *biais de bordure* (en anglais «edge effect»). Dans le cas de données (supposées) continues, il existe des solutions pour y remédier (voir Lejeune & Sarda, 1992 ; Chen 1999, 2000a, ainsi que leurs références).

Dans ce chapitre, nous introduisons une nouvelle famille de lois discrètes. Les v.a. associées à ces nouvelles lois ont une propriété de symétrie autour de la moyenne qui n'est autre que la cible x de l'estimation. De plus, leurs variances ne dépendent que de la fenêtre de lissage h. Cette famille de lois dites *triangulaires discrètes* permettra de définir de vrais noyaux associés discrets symétriques vérifiant toutes les conditions d'un noyau associé. Ainsi, ils améliorent les noyaux discrets standards asymétriques construits dans le chapitre précédent [Section 1.2]. L'étude des propriétés des noyaux associés discrets triangulaires facilite le choix optimal des paramètres. De plus, leur comportement asymptotique les rend comparables au noyau de Dirac associé à l'estimateur empirique (ou naïf). Une solution originale est proposée pour le biais de bordure posé par ce type de noyau discret.

Ce chapitre est structuré comme suit. Section 2.2 définit la famille des lois triangulaires discrètes et donne quelques propriétés (asymptotiques) élémentaires. Section 2.3 étudie les estimateurs à noyaux discrets triangulaires. Elle présente le risque quadratique intégré ainsi que la fenêtre optimale et celle adpatée pour l'excès de zéros. Une transformation est présentée pour résoudre le problème de biais de bordure. Section 2.4 illustre les résultats théoriques des estimateurs à noyaux discrets triangulaires sur des données simulées et présente une application pour lisser la distribution de buts dans le championnat de football français. Puis les résultats obtenus sont comparés à ceux du noyau binomial standard. Section 2.5 conclut ce chapitre.

2.2 Famille de lois triangulaires discrètes

Soit $a, c \in \mathbb{N}$. Une variable aléatoire (v.a.) discrète $\mathcal{T}_{a,c}$ est dite *triangulaire symétrique* de centre c et de bras a, si pour tout y dans son support $\aleph_c = \{c, c\pm 1, \ldots, c\pm a\}$

sa probabilité individuelle s'écrit :

$$\Pr(\mathcal{T}_{a,c} = y) = \frac{a + 1 - |y - c|}{(a + 1)^2}.$$

Sa représentation graphique présente évidemment une forme triangulaire pyramidale sur le support $\aleph_c$ et symétrique autour de sa moyenne $c = \mathbb{E}(\mathcal{T}_{a,c})$. Une extension de la loi triangulaire symétrique sur le même support $\aleph_c$ est la suivante.

Définition 2.2.1 *Soit $h > 0$ et $(a, c) \in \mathbb{N}^2$. Une v.a. discrète $\mathcal{T}_{a;c,h}$ est dite triangulaire d'ordre h, de centre c et de bras a, si pour tout y dans son support $\aleph_c = \{c, c \pm 1, \ldots, c \pm a\}$ sa fonction de masse de probabilité s'écrit :*

$$\Pr(\mathcal{T}_{a;c,h} = y) = \frac{(a + 1)^h - |y - c|^h}{P(a, h)},$$

où $P(a, h)$ est la constante de normalisation telle que

$$P(a, h) = (2a + 1)(a + 1)^h - 2\sum_{k=0}^{a} k^h.$$

On remarque que le cas $h = 1$ correspond à la v.a. triangulaire pyramidale $\mathcal{T}_{a,c}$. Les cas $h \leq 0$ ne sont pas définis en c et, en particulier, ils donnent la v.a. nulle si $h = 0$. Pour des entiers non nuls $h \in \mathbb{N}^*$, on peut noter au passage que la constante de normalisation $P(a, h)$ peut s'écrire de manière explicite :

$$P(a, h) = (2a + 1)(a + 1)^h - 2\sum_{k=0}^{a} \frac{(-1)^{h-k+1} h! B_{h-k+1}}{k!(h - k + 1)!} a^k,$$

où B_j est le nombre de Bernoulli (voir, par exemple, Bouvier *et al.*, 2005). Figure 2.1 présente quelques graphiques des lois triangulaires discrètes. Ici, nous ne représentons pas ces distributions discrètes avec des bâtons ; pour bien distinguer leurs formes, nous relions simplement les points.

Les propriétés élémentaires mais fondamentales des v.a. triangulaires discrètes sont données dans la proposition suivante.

Proposition 2.2.2 *Soit $\mathcal{T}_{a;c,h}$ la v.a. triangulaire discrète d'ordre $h > 0$, de centre $c \in \mathbb{N}$ et de bras $a \in \mathbb{N}$. Alors $\mathcal{T}_{a;c,h}$ est symétrique autour de sa moyenne $c = \mathbb{E}(\mathcal{T}_{a;c,h})$ et sa variance var$(\mathcal{T}_{a;c,h}) = V(a, h) = aC_h(a)$ ne dépend pas de c avec*

$$V(a, h) = \frac{1}{P(a, h)} \left\{ \frac{a(2a + 1)(a + 1)^{h+1}}{3} - 2\sum_{k=0}^{a} k^{h+2} \right\}.$$

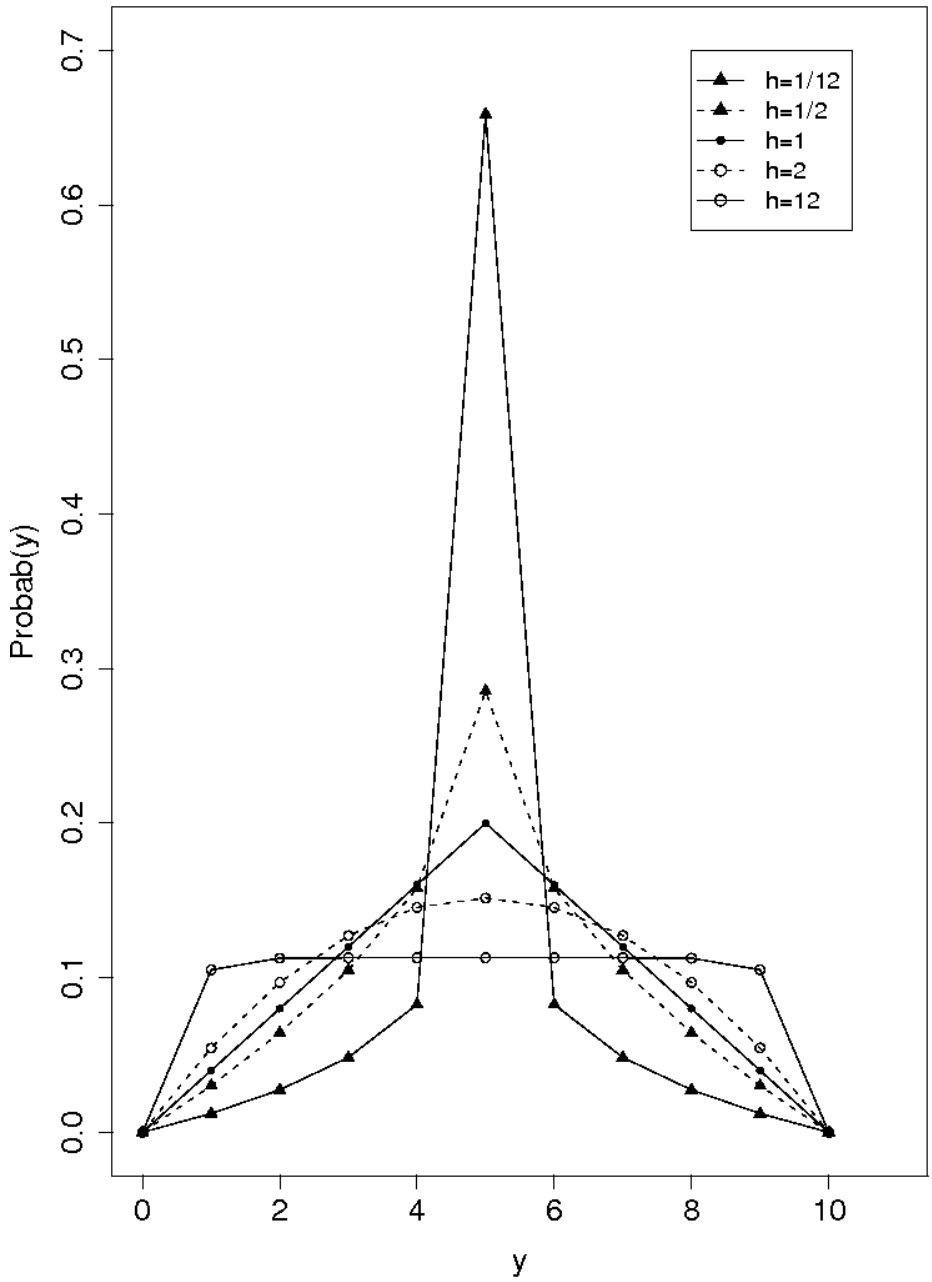

FIG. 2.1 – Quelques distributions triangulaires discrètes centrées en $y = c = 5$, de bras $a = 4$ et selon des valeurs de l'ordre h

DÉMONSTRATION : A partir de la Définition 2.2.1, on a successivement

$$\begin{aligned}
\mathbb{E}(\mathcal{T}_{a;c,h}) &= \sum_{y\in\aleph_c} y \Pr(\mathcal{T}_{a;c,h}=y) \\
&= \frac{1}{P(a,h)}\left\{(a+1)^h \sum_{y\in\aleph_c} y - \sum_{y\in\aleph_c} y|y-c|\right\} \\
&= c \times \frac{(a+1)^h(2a+1) - 2\sum_{k=0}^{a} k^h}{P(a,h)} \\
&= c,
\end{aligned}$$

et on a triviallement, pour $y \in \aleph_c$,

$$\begin{aligned}
\Pr(\mathcal{T}_{a;c,h}=c-y) &= \frac{(a+1)^h - |y|^h}{P(a,h)} \\
&= \Pr(\mathcal{T}_{a;c,h}=c+y).
\end{aligned}$$

Enfin, on obtient :

$$\begin{aligned}
\text{var}(\mathcal{T}_{a;c,h}) &= \sum_{y\in\aleph_c} y^2 \Pr(\mathcal{T}_{a;c,h}=y) - \{\mathbb{E}(\mathcal{T}_{a;c,h})\}^2 \\
&= \frac{1}{P(a,h)}\left\{(a+1)^h \sum_{y\in\aleph_c} y^2 - \sum_{y\in\aleph_c} y^2|y-c|^h\right\} - c^2 \\
&= \frac{1}{P(a,h)}\left\{\frac{2a(a+1)^{h+1}(2a+1)}{6} - 2\sum_{k=0}^{a} k^{h+2} + c^2 P(a,h)\right\} - c^2 \\
&= \frac{1}{P(a,h)}\left\{\frac{a(a+1)^{h+1}(2a+1)}{3} - 2\sum_{k=0}^{a} k^{h+2}\right\} \\
&= aC_h(a). \blacksquare
\end{aligned}$$

Tables 2.1 et 2.2 rapportent quelques expressions de la constante de normalisation $P(a,h)$ et de la variance $\text{var}(\mathcal{T}_{a;c,h})$, respectivement. Ces expressions sont tirées des expressions explicites des nombres de Bernoulli (Bouvier *et al.*, 2005).

Une partie de la remarque suivante est facile à vérifier à partir de Définition 2.2.1, Proposition 2.2.2 et le fait que le support $\aleph_c = \{c, c\pm 1, \ldots, c\pm a\}$ de $\mathcal{T}_{a;c,h}$ dépend du bras $a \in \mathbb{N}$, mais pas de l'ordre $h > 0$. En général, cela peut aussi être observé à l'aide d'un logiciel de calculs formels tel que Matlab ou Maple.

REMARQUE 2.2.3 : On a les deux assertions suivantes :

(i) Pour tout $h > 0$ fixé, la première fonction variance partielle $a \mapsto V(a,h) = \text{var}(\mathcal{T}_{a;c,h})$ de $\mathcal{T}_{a;c,h}$ est croissante et non bornée par rapport au bras $a \in \mathbb{N}$.

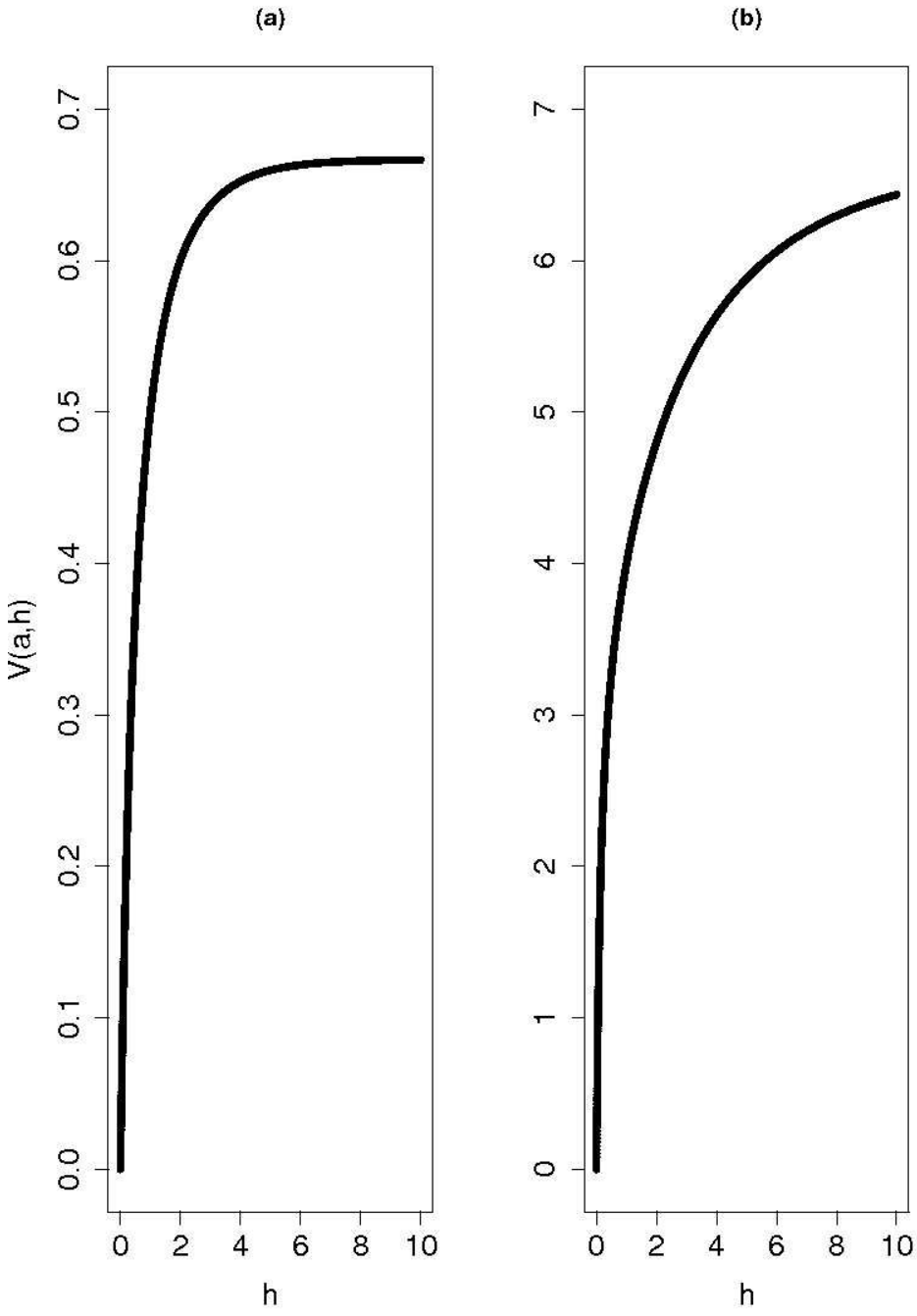

FIG. 2.2 – Graphes de $h \mapsto V(a,h) = \text{var}(\mathcal{T}_{a;c,h})$ pour $a = 1$ (a) et $a = 4$ (b)

h	$P(a,h)$
1	$(a+1)^2$
2	$(2a+3)(2a+1)(a+1)/3$
3	$(a+1)^2(3a^2+6a+2)/2$
4	$(2a+3)(2a+1)(a+1)(6a^2+12a+5)/15$
5	$(a+1)^2(10a^4+40a^3+55a^2+30a+6)/6$
6	$(2a+3)(2a+1)(a+1)(3a^2+9a+7)(3a^2+3a+1)/21$
7	$(a+1)^2\{12(2a+1)(a+1)^5-a^2(3a^4+6a^3-a^2-4a+2)\}/12$
8	$(2a+1)(a+1)\{45(a+1)^7-a(5a^6+15a^5+5a^4-15a^3-a^2+9a-3)\}/45$

TAB. 2.1 – Expressions de $P(a,h)$ pour $h=1,2,\ldots,8$ (Définition 2.2.1)

(ii) Pour tout $a \in \mathbb{N}$ fixé, la seconde fonction variance partielle $h \mapsto V(a,h) = \text{var}(\mathcal{T}_{a;c,h})$ de $\mathcal{T}_{a;c,h}$ est croissante de $0 = \lim_{h\to 0} V(a,h)$ à $a(a+1)/3 = \lim_{h\to+\infty} V(a,h)$ par rapport à l'ordre $h > 0$.

Afin de déduire les bornes de la fonction $h \mapsto V(a,h) = \text{var}(\mathcal{T}_{a;c,h})$ de la partie (ii) de Remarque 2.2.3, la proposition suivante donne les lois limites de la v.a. $\mathcal{T}_{a;c,h}$ par rapport à $h > 0$.

Proposition 2.2.3 *Soit $\mathcal{T}_{a;c,h}$ une v.a. triangulaire discrète d'ordre $h > 0$, de centre $c \in \mathbb{N}$ et de bras $a \in \mathbb{N}$. Alors, pour tout $(a,c) \in \mathbb{N}^2$, on a :*
(i) quand $h \to 0$, la v.a. limite $\mathcal{T}_{a;c,0}$ suit la loi de Dirac en c ;
(ii) quand $h \to +\infty$, la v.a. limite $\mathcal{T}_{a;c,+\infty}$ suit la loi uniforme discrète sur $\aleph_c = \{c, c\pm 1, \ldots, c\pm a\}$.

DÉMONSTRATION : Puisque le support de $\mathcal{T}_{a;c,h}$ est l'ensemble discret $\aleph_c$, il suffit d'étudier les limites des probabilités individuelles par rapport à h.

h	$\text{var}(\mathcal{T}_{a;c,h}) = aC_h(a)$
1	$a(a+2)/6$
2	$a(a+2)/5$
3	$a(2a^3+8a^2+9a+2)/3(3a^2+6a+2)$
4	$5a\{7(a+1)^4-(3a^4+6a^3-3a+1)\}/7(2a+3)(6a^2+12a+5)$
5	$a\{4(a+1)^4(2a+1)-a(3a^4+6a^3-a^2-4a+2)\}/2(10a^4+40a^3+55a^2+30a+6)$
6	$7a\{15(a+1)^6-(5a^6+15a^5+5a^4-15a^3-a^2+9a-3)\}$ $\div 15(2a+3)(3a^2+9a+7)(3a^2+3a+1)$
7	$2a\{10(a+1)^6(2a+1)-3a(a^2+a-1)(2a^4+4a^3-a^2-3a+3)\}$ $\div 5\{12(2a+1)(a+1)^5-a^2(3a^4+6a^3-a^2-4a+2)\}$
8	$15h\{11(a+1)^8-(a^2+a-1)(3a^6+9a^5+2a^4-11a^3+3a^2+10a-5)\}$ $\div 11\{45(a+1)^7-h(5a^6+15a^5+5a^4-15a^3-a^2+9a-3)\}$

TAB. 2.2 – Expressions de la variance de $\mathcal{T}_{a;c,h}$ pour $h = 1, 2, \ldots, 8$ (Proposition 2.2.2)

(i) D'une part, on a :

$$\begin{aligned}\lim_{h\to 0}\Pr(\mathcal{T}_{a;c,h}=c) &= \lim_{h\to 0}\frac{1}{2a+1-2\sum_{k=0}^{a}\{k/(a+1)\}^h}\\ &= \frac{1}{2a+1-2a}\\ &= 1.\end{aligned}$$

D'autre part, pour tout $y \neq c$, il existe $j \in \{1, 2, \ldots, h\}$ tel que $y = c \pm j$; et donc, on obtient

$$\begin{aligned}\lim_{h\to 0}\Pr(\mathcal{T}_{a;c,h}=y) &= \lim_{h\to 0}\frac{(a+1)^h-j^h}{(2a+1)(a+1)^h-2\sum_{k=0}^{a}k^h}\\ &= \frac{1-1}{2a+1-2a}\\ &= 0.\end{aligned}$$

(ii) Soit $y \in \aleph_c$. On peut toujours écrire $y = c \pm j$, où $j \in \{0, 1, \ldots, a\}$. Par conséquent, on a :

$$\begin{aligned} \Pr(\mathcal{T}_{a;c,h} = y) &= \frac{(a+1)^h - j^h}{(2a+1)(a+1)^h - 2\sum_{k=0}^{a} k^h} \\ &= \frac{1}{2a+1} \times \frac{1 - \{j/(a+1)\}^h}{1 - \{2/(2a+1)\}\sum_{k=0}^{a}\{k/(a+1)\}^h}, \end{aligned}$$

pour $j \in \{0, 1, 2, \cdots, a\}$. Il en découle : $\lim_{h\to+\infty} \Pr(\mathcal{T}_{a;c,h} = y) = 1/(2a+1)$, pour tout $y \in \aleph_c$. ■

Ainsi, les bornes 0 et $a(a+1)/3$ de Remarque 2.2.3 (ii) sont obtenues par $\lim_{h\to 0} V(a,h) = \text{var}(\mathcal{T}_{a;c,0}) = 0$ et $\lim_{h\to+\infty} V(a,h) = \text{var}(\mathcal{T}_{a;c,+\infty}) = a(a+1)/3$.

2.3 Estimateurs à noyaux discrets triangulaires

Maintenant, nous sommes en mesure d'introduire les estimateurs à noyaux discrets triangulaires. Pour cela, nous vérifions d'abord que nous disposons bien de noyaux associés discrets construits à partir des lois triangulaires discrètes.

Pour $a \in \mathbb{N}$ fixé et connu, soient $x \in \mathbb{N}$ et $h > 0$. Le noyau discret triangulaire $T_{a;x,h}$ associé à la variable aléatoire $\mathcal{T}_{a;x,h}$ définie sur $\aleph_x = \{x, x \pm 1, \ldots, x \pm a\}$ est donné par :

$$T_{a;x,h}(y) = \frac{(a+1)^h - |y-x|^h}{(2a+1)(a+1)^h - 2\sum_{k=0}^{a} k^h}, \quad y \in \aleph_x. \tag{2.5}$$

Il vérifie les quatre conditions d'un noyau associé discret. En effet, la première condition $\cup_{x\in\mathbb{N}}\aleph_x \supseteq \mathbb{N}$ est triviale ainsi que la deuxième condition (2.1) car $\mathbb{E}(\mathcal{T}_{a;x,h}) = x$. Les deux dernières conditions sur la variance de $\mathcal{T}_{a;x,h}$ (2.2) et (2.3) découlent immédiatement des Proposition 2.2.2, Proposition 2.2.3 et Remarque 2.2.3 (ii).

Soit $X_1, X_2, \ldots, X_n$ un n-échantillon aléatoire i.i.d. de fonction de masse de probabilité inconnue f sur $\mathbb{N}$. Soit $a \in \mathbb{N}$. Un estimateur à noyau associé discret triangulaire de bras a de f est défini par

$$\begin{aligned} \widetilde{f}_{n,h,a}(x) &= \frac{1}{n}\sum_{i=1}^{n} T_{a;x,h}(X_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} \frac{(a+1)^h - |X_i - x|^h}{(2a+1)(a+1)^h - 2\sum_{k=0}^{a} k^h}, \quad x \in \mathbb{N}, \end{aligned} \tag{2.6}$$

où $h > 0$ est le paramètre de lissage (ou fenêtre) et $T_{a;x,h}$ est le noyau associé à la loi de la v.a. triangulaire discrète $\mathcal{T}_{a;x,h}$ sur $\aleph_x = \{x, x \pm 1, \ldots, x \pm a\}$. On a $\widetilde{f}_{n,h,a}(x) \in [0,1]$ pour tout $x \in \mathbb{N}$ et, à constante de normalisation $C = \sum_{x\in\mathbb{N}} \widetilde{f}_{n,h,a}(x)$ près, on suppose

que la fonction $x \mapsto \widetilde{f}_{n,h,a}(x)$ est une fonction de masse de probabilité sur $\mathbb{N}$. Notons que l'équivalent de la relation (2.4) est

$$\mathbb{E}\{\widetilde{f}_{n,h,a}(x)\} \quad = \quad \mathbb{E}\{f(\mathcal{T}_{a;x,h})\}. \tag{2.7}$$

On rappelle que le support $\aleph_x$ de la v.a. $\mathcal{T}_{a;x,h}$ ne dépend pas de h et on a $\cup_{x\in\mathbb{N}}\aleph_x \supseteq \mathbb{N}$. Si $a = 0$, on obtient $\cup_{x\in\mathbb{N}}\aleph_x = \mathbb{N}$. Dans ce cas, l'estimateur $\widetilde{f}_{n,h,0}$ de f n'est autre que la distribution empirique f_0 des observations. Tandis que si $a \neq 0$ on a :

$$\bigcup_{x\in\mathbb{N}} \aleph_x = \{-a, \ldots, -1\} \cup \mathbb{N}. \tag{2.8}$$

Le fait que le support $\cup_{x\in\mathbb{N}}\aleph_x$ (2.8) du noyau discret triangulaire à $a \neq 0$ fixé contienne strictement le support $\mathbb{N}$ de f induit un biais de bordure à gauche du support de f. Nous y remédions en fin de section en modifiant le bras a par a_0 de sorte que, pour tout a_0, on ait

$$\bigcup_{x\in\mathbb{N}} \aleph_x = \mathbb{N}. \tag{2.9}$$

Signalons ici que si le support de f est $\mathbb{Z}$ (non-borné) alors on a $\cup_{x\in\mathbb{Z}}\aleph_x = \mathbb{Z}$ et, donc, il n'y a pas de problème de biais de bordure. Tandis que si le support de f est $\{0, 1, \ldots, N\}$ (fini), alors on a $\cup_{x\in\{0,1,\ldots,N\}}\aleph_x = \{-a, \ldots, -1\} \cup \{0, 1, \cdots, N\} \cup \{N + 1, \ldots, N + a\}$; ce qui conduirait à des biais de bordure à la fois à gauche et à droite du support de f.

2.3.1 Risque quadratique intégré

De la relation (2.7) d'un estimateur à noyau discret triangulaire $\widetilde{f}_{n,h,a}(x)$, nous examinons les effets du paramètre de lissage discret h et du bras $a \in \mathbb{N}$ à travers l'erreur quadratique intégrée

$$\begin{aligned} MISE(n, h, T_a, f) &= \mathbb{E}\left[\sum_{x\in\mathbb{N}} \left\{\widetilde{f}_{n,h,a}(x) - f(x)\right\}^2\right] && (2.10)\\ &= \sum_{x\in\mathbb{N}} MSE_a(x; n, h) && (2.11)\\ &= \sum_{x\in\mathbb{N}} \text{var}\{\widetilde{f}_{n,h,a}(x)\} + \sum_{x\in\mathbb{N}} \text{biais}^2\{\widetilde{f}_{n,h,a}(x)\} && (2.12) \end{aligned}$$

de l'estimateur $\widetilde{f}_{n,h,a}$ de f défini en (2.6), où $MSE_a(n, h, x)$ est l'erreur quadratique ponctuelle au point $x \in \mathbb{N}$.

Risque quadratique intégré exact

Nous exprimons ici le risque quadratiqué intégré à partir des calculs directs du biais et de la variance de l'estimateur $\widetilde{f}_{n,h,a}$.

Le calcul du biais ponctuel de $\widetilde{f}_{n,h,a}$ conduit à

$$\text{biais}\{\widetilde{f}_{n,h,a}(x)\} = f(x)\left\{\frac{(a+1)^h}{P(a,h)}-1\right\}+\sum_{y\in\aleph_x\setminus\{x\}} f(y)T_{a;x,h}(y),$$

avec $P(a,h)=P(a,h)=(2a+1)(a+1)^h-2\sum_{k=0}^{a}k^h$. La variance ponctuelle de $\widetilde{f}_{n,h,a}$ s'obtient par

$$\begin{aligned}\text{var}\left\{\widetilde{f}_{n,h,a}(x)\right\} &= \frac{1}{n}\left\{f(x)\frac{(a+1)^{2h}}{P^2(a,h)}+\sum_{y\in\aleph_x\setminus\{x\}} f(y)T^2_{a;x,h}(y)\right\}\\ &\quad -\frac{1}{n}\left\{f(x)\frac{(a+1)^h}{P(a,h)}+\sum_{y\in\aleph_x\setminus\{x\}} f(y)T_{a;x,h}(y)\right\}^2.\end{aligned}$$

Ainsi, l'expression du $MISE(n,h,T_a,f)$ découle aisément de (2.12).

En particulier, pour $a=1$, l'erreur quadratique ponctuelle est donnée explicitement par

$$\begin{aligned}MSE_{a=1}(x;n,h) &= \left[f(x)\left\{\frac{2^h}{P(1,h)}-1\right\}+\{f(x-1)+f(x+1)\}\frac{2^h-1}{P(1,h)}\right]^2\\ &\quad +\frac{1}{n}f(x)\{1-f(x)\}\left\{\frac{2^h}{P(1,h)}\right\}^2\\ &\quad +\{f(x-1)+f(x+1)\}\frac{(2^h-1)^2}{P^2(1,h)}\\ &\quad -\frac{2}{n}f(x)\{f(x-1)+f(x+1)\}\frac{2^h(2^h-1)}{P^2(1,h)}\\ &\quad -\frac{1}{n}\{f(x-1)+f(x+1)\}^2\frac{(2^h-1)^2}{P^2(1,h)},\end{aligned}$$

avec $P(1,h)=3.2^h-2$. Pour n et x fixés, quand $h\to 0$, nous avons biais$\{\widetilde{f}_{n,h,1}(x)\}\to 0$ et var$\{\widetilde{f}_{n,h,1}(x)\}\to(1/n)f(x)\{1-f(x)\}$ car $2^h\to 1$ et $P(1,h)\to 1$. Il s'en suit :

$$\begin{aligned}\lim_{h\to 0} MISE(n,h,T_{a=1},f) &= \frac{1}{n}\sum_{x\in\mathbb{N}} f(x)\{1-f(x)\} \qquad (2.13)\\ &=: MISE_{\text{naïf}}(n,0,f),\end{aligned}$$

où $MISE_{\text{naïf}}(n, 0, f)$ est le risque quadratique intégré de l'estimateur empirique (ou à noyau du type Dirac). Ce résultat (2.13) peut être étendu à $a \in \mathbb{N}^*$ avec, pour tout x,

$$\lim_{h\to 0} \text{biais}\{\widetilde{f}_{n,h,a}(x)\} = 0 \text{ et } \lim_{h\to 0} \text{var}\{\widetilde{f}_{n,h,a}(x)\} = \frac{1}{n} f(x)\{1 - f(x)\}.$$

Par conséquent, pour $a \in \mathbb{N}^*$ et $h > 0$ fixés, nous pouvons réaliser une comparaison avec l'estimateur naïf en fonction de n pour valider l'intérêt du nouvel estimateur mais aussi déterminer ses limites pratiques. Par exemple, dans la Figure 2.3, le point d'intersection des courbes indique la limite supérieure de n pour laquelle l'estimateur à noyau discret triangulaire est plus performant que l'estimateur naïf. Au delà de cette limite, l'estimateur naïf devient meilleur et son $MISE$ tend vers 0, ce qui est aussi le cas du $MISE$ de l'estimateur à noyau discret triangulaire à condition que $h = h(n) \to 0$ (très petit). Finalement, on peut déduire la convergence globale de l'estimateur à noyaux discrets triangulaires par

$$MISE(n, h, T_a, f) \to 0 \text{ quand } n \to +\infty \text{ et } h = h(n) \to 0,$$

pour tout $a \in \mathbb{N}^*$ et f telle que $\lim_{x\to+\infty} f(x) = 0$.

Risque quadratique intégré approché

Dans cette partie, nous proposons un calcul du risque quadratique intégré en utilisant le développement limité de Taylor discret pour le calcul du biais comme cela a été fait dans le chapitre 1.

Soit $x \in \mathbb{N}$. Puisque $\mathbb{E}(\mathcal{T}_{a;x,h}) = x$ et $\text{var}(\mathcal{T}_{a;x,h}) = V(a, h)$ (voir Proposition 2.2.2), nous appliquons aux estimateurs $\widetilde{f}_{n,h,a}(x)$ à noyaux discrets triangulaires les résultats qui précédent et ceux plus généraux du Chapitre 1. Nous obtenons d'une part le biais de $\widetilde{f}_{n,h,a}(x)$ par le développement de Taylor discret [Equation (1.24)]

$$\begin{aligned}
\text{biais}\{\widetilde{f}_{n,h,a}(x)\} &= \mathbb{E}\{\widetilde{f}_{n,h,a}(x)\} - f(x) \\
&= \mathbb{E}\{f(\mathcal{T}_{a;x,h})\} - f(x) \\
&= f\{\mathbb{E}(\mathcal{T}_{a;x,h})\} - f(x) + \frac{1}{2}\text{var}(\mathcal{T}_{a;x,h}) f^{(2)}(x) + o(h) \\
&= \frac{1}{2}\text{var}(\mathcal{T}_{a;x,h}) f^{(2)}(x) + o(h) \\
&= \frac{1}{2} V(a, h) f^{(2)}(x) + o(h), \qquad (2.14)
\end{aligned}$$

où la différence finie d'ordre 2 de f en tout point $x \in \mathbb{N}$ s'écrit :

$$f^{(2)}(x) = \begin{cases} \{f(x+2) - 2f(x) + f(x-2)\}/4 & \text{si } x \in \mathbb{N} \setminus \{0, 1\} \\ \{f(3) - 3f(1) + 2f(0)\}/4 & \text{si } x = 1 \\ \{f(2) - 2f(1) + f(0)\}/2 & \text{si } x = 0. \end{cases} \qquad (2.15)$$

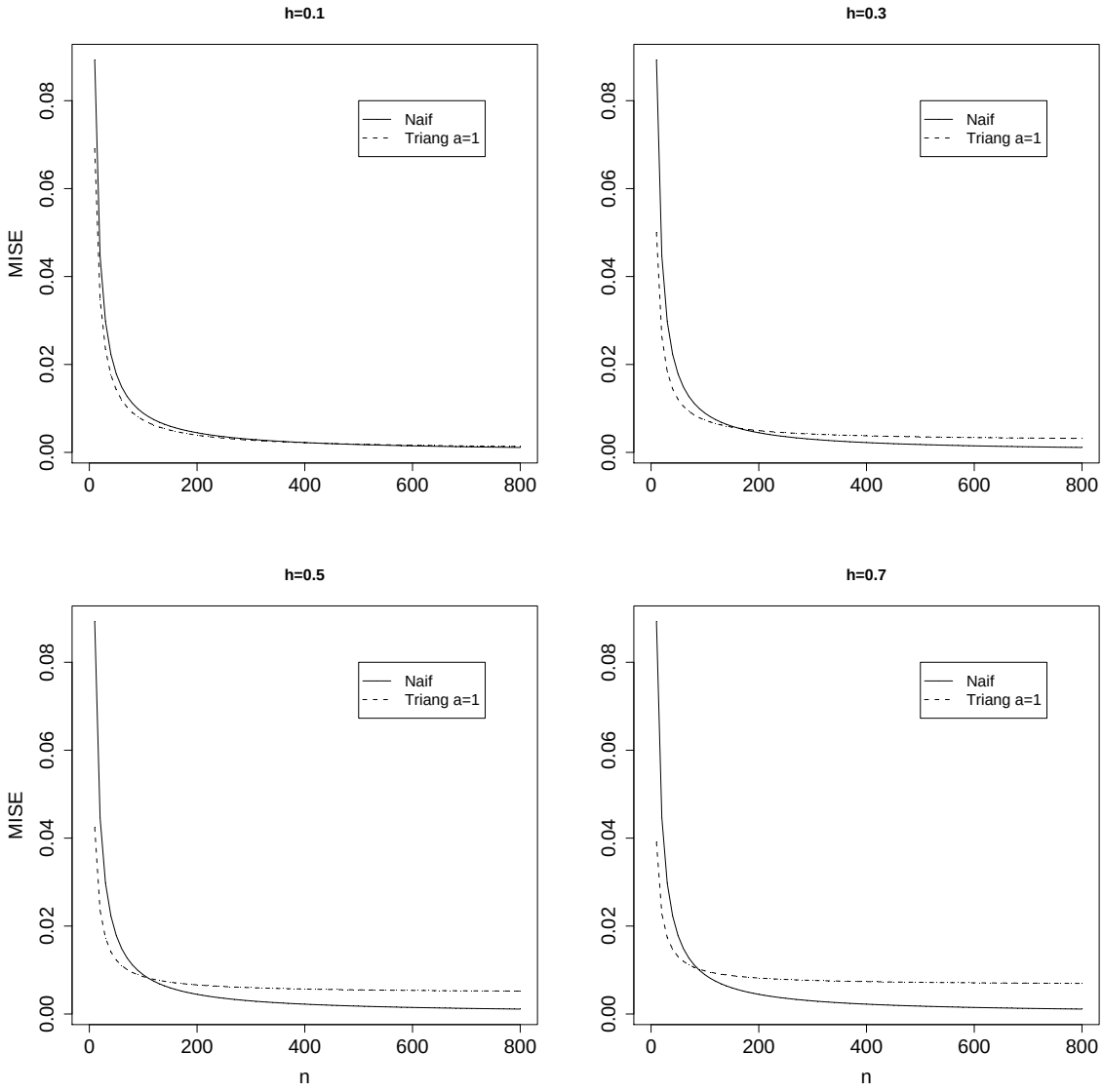

FIG. 2.3 – Comparaison entre $MISE_{\text{naïf}}(n, 0, f)$ et $MISE_{a=1}(n, h, f)$ où f est la distribution du mélange de Poisson $f = 0.4\mathcal{P}(0.5) + 0.6\mathcal{P}(10)$

D'autre part, la variance de $\widetilde{f}_{n,h,a}(x)$ peut s'écrire successivement comme

$$\begin{aligned}\text{var}\left\{\widetilde{f}_{n,h,a}(x)\right\} &= \frac{1}{n}\text{var}\left\{T_{a;x,h}(X_1)\right\} \\ &= \frac{1}{n}f(x)\left\{1-f(x)\right\}\left\{\Pr(\mathcal{T}_{a;x,h}=x)\right\}^2+o\left(\frac{1}{n}\right) \\ &= \frac{1}{n}f(x)\left\{1-f(x)\right\}\left\{\frac{(a+1)^h}{P(a,h)}\right\}^2+o\left(\frac{1}{n}\right), \qquad (2.16)\end{aligned}$$

où $P(a,h)=(2a+1)(a+1)^h-2\sum_{k=0}^{a}k^h$ (voir Définition 2.2.1). Au final, nous pouvons exprimer l'erreur quadratique $MSE(x)$ de $\widetilde{f}_{n,h,a}(x)$ par :

$$MSE(x;n,h)=\frac{1}{n}f(x)\left\{1-f(x)\right\}\left\{\frac{(a+1)^h}{P(a,h)}\right\}^2+\frac{1}{4}\{V(a,h)f^{(2)}(x)\}^2+o\left(\frac{1}{n}+h^2\right).$$

A partir de (2.16), la variance intégrée est donnée par

$$\sum_{x\in\mathbb{N}}\text{var}\{\widetilde{f}_{n,h,a}(x)\}=\frac{1}{n}\left\{\frac{(a+1)^h}{P(a,h)}\right\}^2\sum_{x\in\mathbb{N}}f(x)\left\{1-f(x)\right\}+o\left(\frac{1}{n}\right). \qquad (2.17)$$

En remarquant que $(a+1)^h/P(a,h)=\Pr(\mathcal{T}_{a;x,h}=x)$ est la probabilité modale de la v.a. $\mathcal{T}_{a;x,h}$ (voir Définition 2.2.1 et aussi Figure 2.1), il en découle que la fonction $h\mapsto n^{-1}[(a+1)^h\{P(a,h)\}^{-1}]^2$ est décroissante de n^{-1} à $n^{-1}\left(2a+1\right)^{-2}$ pour tout n et a dans $\mathbb{N}^*$.

Du biais ponctuel (2.14), le carré du biais intégré s'écrit

$$\sum_{x\in\mathbb{N}}\text{biais}^2\{\widetilde{f}_{n,h,a}(x)\}=\frac{1}{4}\left\{V(a,h)\right\}^2\sum_{x\in\mathbb{N}}\left\{f^{(2)}(x)\right\}^2+o(h^2). \qquad (2.18)$$

Comme $f^{(2)}(x)$ est une combinaison linéaire finie de $f(x\pm j)\in[0,1]$ pour $j\in\{0,1,2\}$ (voir formule (2.15)), on a $\sum_{x\in\mathbb{N}}\left\{f^{(2)}(x)\right\}^2<+\infty$. Il suit de Remarque 2.2.3 (ii) et de Proposition 2.2.3 que la fonction $h\mapsto\left\{V(a,h)\right\}^2/4=\left\{\text{var}(\mathcal{T}_{a;x,h})/2\right\}^2$ est croissante de 0 à $a^2(a+1)^2/36$ pour tout $a\in\mathbb{N}$.

Par conséquent, en remplaçant (2.17) et (2.18) dans (2.12), à partir du Théorème 1.2.6 on trouve

$$\begin{aligned}MISE(n,h,T_a,f) &\doteq \frac{1}{n}\left\{\frac{(a+1)^h}{P(a,h)}\right\}^2\sum_{x\in\mathbb{N}}f(x)\left\{1-f(x)\right\} \\ &\quad+\frac{1}{4}\left\{V(a,h)\right\}^2\sum_{x\in\mathbb{N}}\left\{f^{(2)}(x)\right\}^2, \qquad (2.19)\end{aligned}$$

laquelle $MISE(n,h,T_a,f)$ conduit aux choix optimaux du paramètre de lissage discret h et du bras a. Ainsi, on a : $MISE(n,h,T_a,f)\to 0$ quand $n\to+\infty$ et

$h = h(n) \to 0$; car $\lim_{h\to 0}(a+1)^h/P(a,h) = 1$, $0 \leq \sum_{x\in\mathbb{N}} f(x)\{1-f(x)\} < 1$, $\lim_{h\to 0} V(a,h) = 0$ et

$$\sum_{x\in\mathbb{N}} \left\{f^{(2)}(x)\right\}^2 < +\infty$$

d'après (1.5). D'où, les estimateurs à noyaux associés discrets triangulaires convergent au sens du $MISE$.

2.3.2 Choix optimaux des paramètres

La valeur optimale de $h > 0$ et celle de $a \in \mathbb{N}^*$ sont obtenues (si elles existent) en minimisant le terme principal dans (2.19) :

$$AMISE(n,h,T_a,f) = \frac{1}{n}\left\{\frac{(a+1)^h}{P(a,h)}\right\}^2 \sum_{x\in\mathbb{N}} f(x)\{1-f(x)\} + \frac{1}{4}\{V(a,h)\}^2 \sum_{x\in\mathbb{N}} \left\{f^{(2)}(x)\right\}^2. \tag{2.20}$$

Tout d'abord, comme $a \in \mathbb{N}^*$ et que la fonction $a \mapsto AMISE(n,h,T_a,f)$ est croissante pour $h > 0$ fixé, le bras optimal a^* est simplement

$$a^* = arg\min_{a\in\mathbb{N}^*} AMISE(n,h,T_a,f) = 1, \tag{2.21}$$

pour tout $h > 0$. Ensuite, pour un bras $a \in \mathbb{N}$ donné, la fenêtre optimal pour un estimateur à noyau discret triangulaire est définie par

$$h^* = arg\min_{h>0} AMISE(n,h,T_a,f) = h^*(n,a,f). \tag{2.22}$$

L'existence de h^* serait garantie par la décroissance du terme principal de la variance intégrée (2.17) et la croissance du terme principal du carré de biais intégré (2.18) dans le terme principal du risque quadratique global (2.20). 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 devrait balancer les approximations du carré du biais et de la variance. Autrement dit, il existerait $\varepsilon > 0$ telle que la fonction $h \mapsto AMISE(n,h,T_a,f)$ serait décroissante sur $]0,\varepsilon[$ et croissante sur $]\varepsilon,+\infty[$ pour tout $h \in \mathbb{N}^*$.

Dans la pratique, comme $h^* = h^*(n,a,f)$, on ne peut pas le déterminer car il dépend de f qui est inconnue. On pourrait se faire une idée particulière du h^* de (2.22) en utilisant

$$h_0^* = arg\min_{h>0} AMISE(n,h,T_a,f_0) = h_0^*(n,a,f_0), \tag{2.23}$$

où f_0 est la distribution empirique des observations. Puisque f_0 tend vers f quand $n \to +\infty$, on devrait avoir

$$\lim_{n\to+\infty} h_0^*(n,a,f_0) = \lim_{n\to+\infty} h^*(n,a,f). \tag{2.24}$$

Une alternative au choix de la fenêtre optimale est la méthode de validation croisée bien connue dans la littérature (Bowman, 1984 ; Marron, 1987 ; Yang, 2007). Dans le cas discret, cette procédure présente encore l'avantage de ne pas utiliser l'approximation de $MISE(n, h, T_a, f)$ et par suite les différences finies de f. Nous développons autrement l'expression de $MISE$ (2.10) par :

$$MISE(n, h, T_a, f) = \mathbb{E}\left\{\sum_{x\in\mathbb{N}} \widetilde{f}^2_{n,h,a}(x)\right\} - 2\mathbb{E}\left\{\sum_{x\in\mathbb{N}} \widetilde{f}_{n,h,a}(x) f(x)\right\} + \sum_{x\in\mathbb{N}} f^2(x).$$

La fenêtre optimale est obtenue par

$$h_{cv} = arg\min_{h>0} CV_a(h) = h_{cv}(a), \tag{2.25}$$

où

$$CV_a(h) = \sum_{x\in\mathbb{N}} \left\{\frac{1}{n}\sum_{i=1}^{n} T_{a;x,h}(X_i)\right\}^2 - \frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i} T_{a;X_i,h}(X_j)$$

est un estimateur sans biais de $\mathbb{E}\left\{\sum_{x\in\mathbb{N}} \widetilde{f}^2_{n,h,a}(x)\right\} - 2\mathbb{E}\left\{\sum_{x\in\mathbb{N}} \widetilde{f}_{n,h,a}(x) f(x)\right\}$.

Selon la méthode de proportion de zéros développée dans le Chapitre 1, Section 1.5.3, nous pouvons chercher ici une *fenêtre adaptée* h_0 dans la situation d'excès de zéros pour les données de dénombrement. Pour cela, il faut résoudre l'équation $\sum_{i=1}^{n} \Pr\left(\mathcal{T}_{a;X_i,h_0} = 0\right) = n_0$, laquelle est équivalente à

$$\left\{n - (2a+1)\, n_0\right\}(a+1)^h + 2n_0 \sum_{k=0}^{a} k^h - \sum_{i=1}^{n} X_i^h = 0, \tag{2.26}$$

où $n_0 = \natural(X_i = 0)$ est le nombre d'observations égal à 0 dans l'échantillon. Mais cette équation n'admet pas de solution. En effet, dans le premier membre de l'équation (2.26), on a

$$\begin{aligned}\left\{n - (2a+1)\, n_0\right\}(a+1)^h + 2n_0 \sum_{k=0}^{a} k^h &< \left\{n - (2a+1)\, n_0\right\}(a+1)^h + 2n_0 a(a+1)^h \\ &= (n - n_0)(a+1)^h.\end{aligned}$$

Puisque $\sum_{i=1}^{n} X_i^h = \sum_{j=0}^{k} n_j X_j^h$ avec $\sum_{j=0}^{k} n_j = n$ et $X_j \geq a+1$, $j = 1, 2, \cdots, k$, on voit que

$$\sum_{i=1}^{n} X_i^h \geq (n - n_0)(a+1)^h.$$

Ces deux dernières inégalités permettent de conclure que l'équation (2.26) n'admet pas de solution et, finalement, on a :

$$\{n-(2a+1)\,n_0\}\,(a+1)^h+2n_0\sum_{k=0}^{a}k^h-\sum_{i=1}^{n}X_i^h<0.$$

Maintenant, nous proposons une solution au problème de biais de bordure pour les estimateurs à noyaux discrets triangulaires. En cas d'observations importantes au bord $\{0,1,\cdots,k\}$ du support $\mathbb{N}$ de f (k très petit, de l'ordre de 0, 1 ou 2), on considère le *bras modifié* a_0 de a satisfaisant (2.9). Notre solution est telle que, pour $k\in\mathbb{N}\setminus\{0\}$ donné et $x\in\mathbb{N}$,

$$a_0=k\Longleftrightarrow a_0=\begin{cases} j & \text{si } x=j,\ \ j\in\{0,1,\ldots,k-1\}\\ k & \text{si } x\in\{k,k+1,\cdots\}.\end{cases} \tag{2.27}$$

Ainsi, le noyau associé discret triangulaire correspondant à ce bras modifié a_0 préserve la symétrie locale autour de chaque point d'estimation. En pratique, cette opération est effectuée avant la normalisation de l'estimateur $\widetilde{f}_{n,h,a^*}(x)$ par la constante $C^*=\sum_{x\in\mathbb{N}}\widetilde{f}_{n,h,a^*}(x)$. Si bien que $\widetilde{f}_{n,h,0}(0)$ n'est généralement pas égal à l'estimation empirique $f_0(0)$ de f en $x=0$ en cas d'excès de zéros dans l'échantillon observé. Ceci résoud implicitement le problème d'excès de zéros dans la distribution. Notons au passage ici qu'une alternative à cette solution (3.13) serait de considérer un noyau discret triangulaire asymétrique.

Pour des bras a et a_0, nous illustrons le comportement des noyaux associés $T_{a;x,h}$ et $T_{a_0;x,h}$ selon de différentes valeurs de la cible x. Dans Figure 2.4, pour $a=4$ et $h=1$, le noyau associé $T_{a;x,h}$ est symétrique et garde la même forme pour les valeurs de x choisies. La distribution de $T_{a;x,h}$ s'étend à des valeurs de $x\in\{-a,\ldots,-1\}$ (en pointillés) en dehors de $\mathbb{N}$. Ceci induit le biais de bordure que nous résolvons en modifiant le bras a en a_0. Ainsi, dans Figure 2.5 et pour $a_0=4$ et $h=1$, nous obtenons un noyau $T_{a_0;x,h}$ qui conserve la propriété de symétrie et tel qu'on ait la relation (2.9). Précisons ici que, d'après la relation (2.9), on ait $\aleph_x\subseteq\mathbb{N}$ pour tout $x\in\mathbb{N}$.

Pour $a=1$ et $a_0=1$, on a respectivement :

$$\begin{aligned} AMISE(n,h,T_{a_0=1},f) &= AMSE_{a=0}(n,h,0)+\sum_{x\in\mathbb{N}^*}AMSE_{a=1}(n,h,x)\\ &= \frac{1}{n}f(0)\{1-f(0)\}+\sum_{x\in\mathbb{N}^*}AMSE_{a=1}(n,h,x)\end{aligned}$$

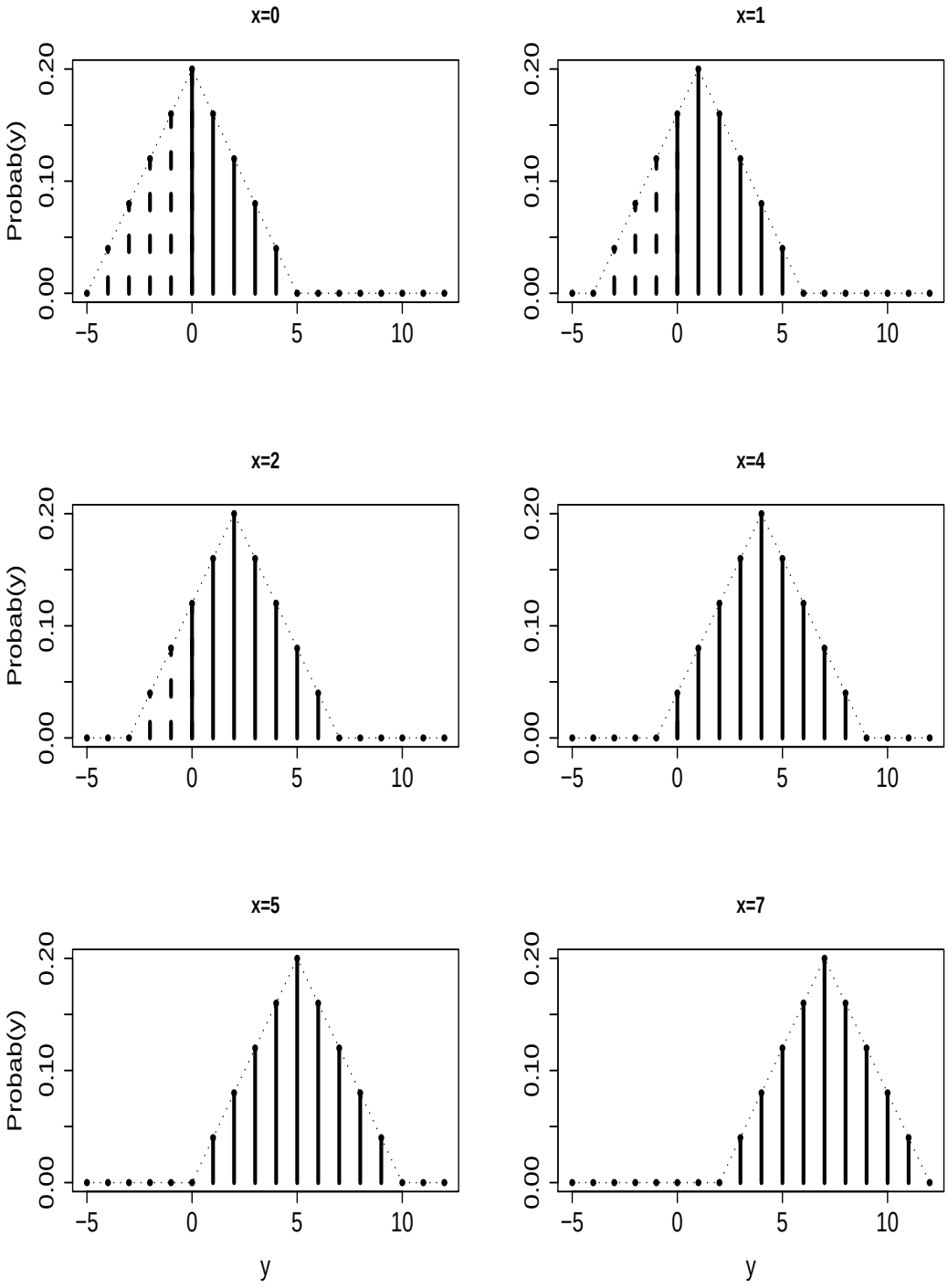

FIG. 2.4 – Noyau associé discret triangulaire avec $a = 4$, $h = 1$ et selon des valeurs de la cible x

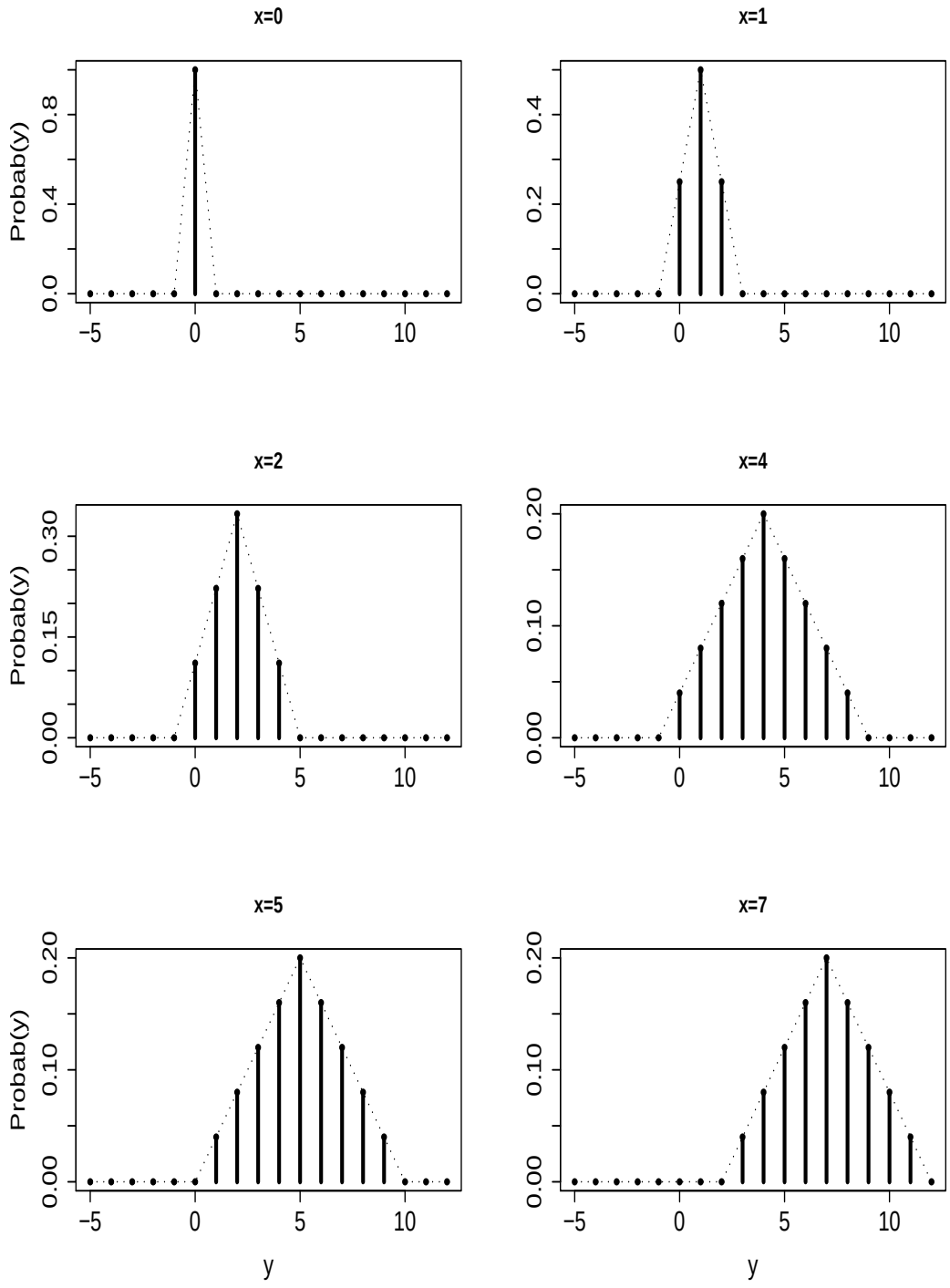

FIG. 2.5 – Suite et fin de la Figure 2.4 avec le bras modifié $a_0 = 4$

et

$$\begin{aligned} AMISE(n,h,T_{a=1},f) &= AMSE_{a=1}(n,h,0) + \sum_{x\in\mathbb{N}^*} AMSE_{a=1}(n,h,x) \\ &= \frac{2^{2h}}{nP^2(1,h)} f(0)\{1-f(0)\} + \frac{1}{4}\{V(1,h)\}^2\{f^{(2)}(0)\}^2 \\ &\quad + \sum_{x\in\mathbb{N}^*} AMSE_{a=1}(n,h,x), \end{aligned}$$

où $AMSE_a$ est l'approximation de MSE_a de (2.11). Par conséquent, chercher à comparer $AMISE(n,h,T_{a_0},f)$ et $AMISE(n,h,T_a,f)$ revient à comparer le rapport $f(0)\{1-f(0)\}/n$ avec $2^{2h}f(0)\{1-f(0)\}/\{nP^2(1,h)\} + \{V(1,h)\}^2\{f^{(2)}(0)\}^2/4$. Cependant, nous ne pouvons pas conclure sur cette comparaison parce que cela dépend du paramètre de lissage discret h, de la taille n de l'échantillon et de la proportion de zéros $f(0)$ dans l'échantillon. Nous présentons dans la section suivante des graphiques pour illustrer ce comportement de $AMISE(n,h,T_{a_0},f)$ et $AMISE(n,h,T_a,f)$ à travers Figure 2.8.

En conclusion, le choix des fenêtres optimales h_0^* ou h_{cv} et celui des bras a^* ou a_0^* suffisent largement pour un lissage discret très satisfaisant d'une fonction de masse de probabilité. L'utilisation de a_0^* par rapport à a^* est essentiel pour les données ayant une proportion non-négligeable de zéros au bord. De plus, il serait intéressant d'approfondir les résultats de convergence.

2.4 Illustrations

Nous illustrons ici les résultats précédents à travers des jeux de données simulées et réelles.

2.4.1 Données simulées

Dans cette section, nous mettons en évidence les différents choix et effets de la fenêtre de lissage discret h, des bras a et a_0 sur des données de comptage simulées. Afin de mieux évaluer la performance des estimateurs à noyaux discrets triangulaires, nous reconsidérons la fonction de masse de probabilité (1.49) du Chapitre 1 :

$$f(x) = 0.4\, e^{-0.5}\, 0.5^x/x! + 0.6\, e^{-10}\, 10^x/x!, \quad x \in \mathbb{N}, \tag{2.28}$$

laquelle est un mélange de deux lois de Poisson de moyennes $\mu_1 = 0.5$ et $\mu_2 = 10$. Cette fonction de masse de probabilité f sur $\mathbb{N}$ ne produit pas de distribution parsemée. Elle admet sa plus grande valeur en $x = 0$ ($f(0) = 0.243$), un minimum local en $x = 3$ ($f(3) = 9.594 \times 10^{-3}$), un maximum local en $x = 9$ ($f(9) = 7.506603 \times 10^{-2}$) et une queue à partir de $x = 22$ ($1 - \sum_{x=0}^{21} f(x) = 4.198 \times 10^{-4}$).

Connaissant explicitement f avec $\sum_{x\in\mathbb{N}}\left\{f^{(2)}(x)\right\}^2 = 0.00489$ [voir (2.15)], le noyau discret triangulaire optimal associé à h^* (2.22) ne dépend que de la taille n de l'échantillon et du bras $a \in \mathbb{N}^*$. Pour $n = 700$, les parties (a) et (c) de Figure 2.6 représentent les courbes de $AMISE(n, h, T_a, f)$ théoriques avec $a = 1$ et $a = 2$, respectivement. Les valeurs de $AMISE(n, h, T_a, f)$ passent du simple (de l'ordre de 10^{-3}) pour $a = 1 = a^*$ avec $h^* = 0.61$ à au moins le double pour $a = 2$ avec $h^* = 0.05$. On peut aussi observer que ces valeurs théoriques de $AMISE(n, h, T_a, f)$ sont bornées.

En simulant un seul échantillon de f définie en (2.28) de taille $n = 700$, nous avons obtenu $\sum_{x\in\mathbb{N}}\left\{f_0^{(2)}(x)\right\}^2 = 0.00683$ et les parties (b) et (d) de Figure 2.6 donnent les courbes de $AMISE(n, h, T_a, f)$ correspondantes avec $a = 1$ et $a = 2$, respectivement. Cette fois-ci, les valeurs de $AMISE(n, h, T_a, f)$ sont toujours de l'ordre de 10^{-3} pour $a = 1 = a^*$ avec $h_0^* = 0.32$ mais les $AMISE(n, h, T_a, f)$ de $a = 2$ sont au moins le triple de celles de $a = 1$ avec son correspondant $h_0^* = 0.03$. De plus, les $AMISE(n, h, T_a, f)$ observées sont aussi bornées.

Pour un échantillon de f de taille n donnée, 1000 repliques indépendantes de f_0 de f sont effectuées et nous avons constaté les phénomènes suivants dans plus de 98% des cas. A savoir, les $AMISE(n, h, T_a, f)$ observées correspondantes sont bornées, de même ordre de grandeur (très petit) que les $AMISE(n, h, T_a, f)$ théoriques et augmentent en fonction de $a \in \mathbb{N}^*$. Les noyaux discrets triangulaires optimaux associés à a^* sont obtenus avec la fenêtre optimale $a = 1 = a^*$ et les h_0^* sont proches de h^* réalisé avec $a = 1 = a^*$. Ainsi, l'approximation h_0^* est un bon indicateur de h^*. Voir aussi Figure 2.7 pour $n = 1000$.

Pour des tailles différentes $n \in \{300, 1000\}$ (Figure 2.8), nous traçons les fonctions $h \mapsto MISE(n, h, T_{a=1}, f)$, $h \mapsto AMISE(n, h, T_{a_0=1}, f)$ et $h \mapsto AMISE(n, h, T_{a=1}, f)$. Pour $n = 300$, la courbe de $MISE$ présente un minimum tandis que celles de $AMISE_{a_0=1}$ et $AMISE_{a=1}$ décroissent avec la même allure. Pour $n = 1000$, les trois courbes ont la même allure et présentent chacune une valeur minimale, mais celle de $MISE$ est la plus proche de 0.

Nous observons que pour un échantillon avec une proportion assez importante de zéros ($f_0(0) = 0.240$), nous pouvons estimer le paramètre h^* par h_0^*. L'utilisation du bras modifié a_0 ou du bras d'origine a est possible pour toute distribution f_0, mais le bras a_0 est approprié avec un nombre important de zéros.

Dans la section suivante, Figure 2.10 présente le comportement des deux fonctions $AMISE(n, h, T_{a_0=1}, f)$ et $AMISE(n, h, T_{a=1}, f)$ mis en évidence dans le cas de données réelles pour un échantillon de taille modérée $n = 380$.

Nous présentons en Annexe B des simulations des erreurs $AMISE_a$ pour les estimateurs à noyaux discrets triangulaires qui sont comparées avec celles des estimateurs à noyaux discrets standards et de l'estimateur fréquence (voir Table B.4).

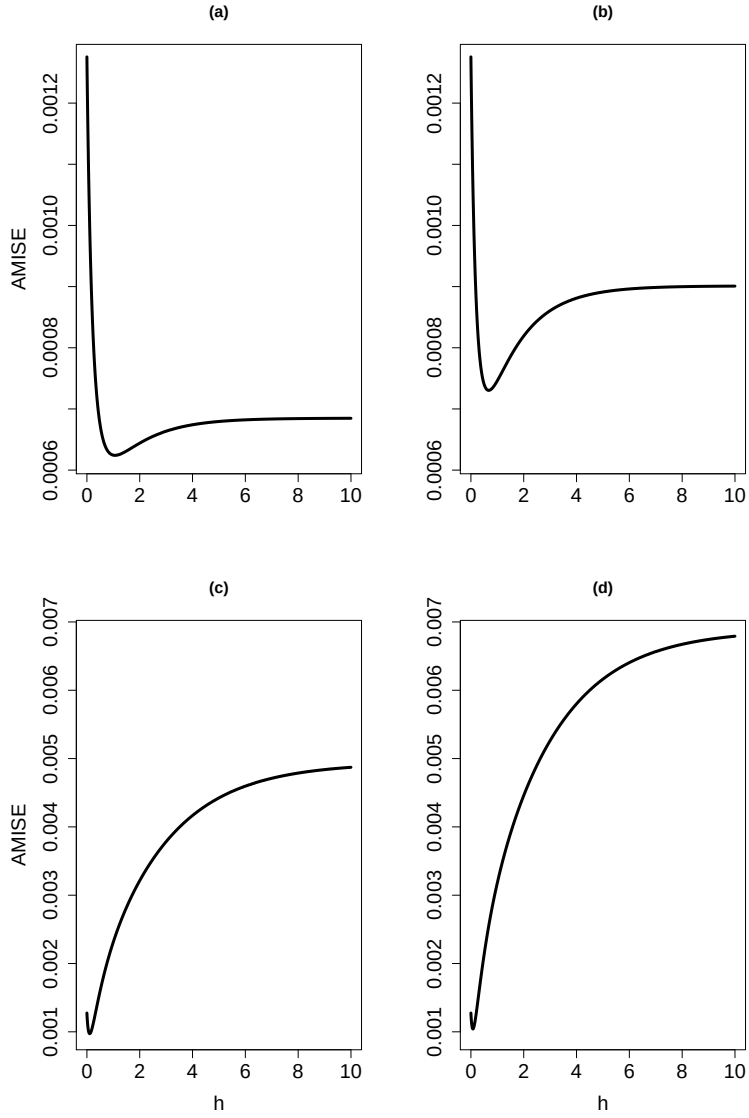

FIG. 2.6 – Graphiques de $AMISE(n, h, T_a, f)$ et $AMISE(n, h, T_a, f_0)$ pour $n = 700$ avec $\sum_{x\in\mathbb{N}}\{f^{(2)}(x)\}^2 = 0.00489$ si $a = 1$, donc $h^* = 0.61$ (a) et $a = 2$, donc $h^* = 0.05$ (c) ; puis $\sum_{x\in\mathbb{N}}\{f_0^{(2)}(x)\}^2 = 0.00683$ si $a = 1$, donc $h_0^* = 0.32$ (b) et $a = 2$, donc $h_0^* = 0.03$ (d). En agrandissant (d) au voisinage de $h = 0$, la courbe décroit puis croit comme dans les autres graphes.

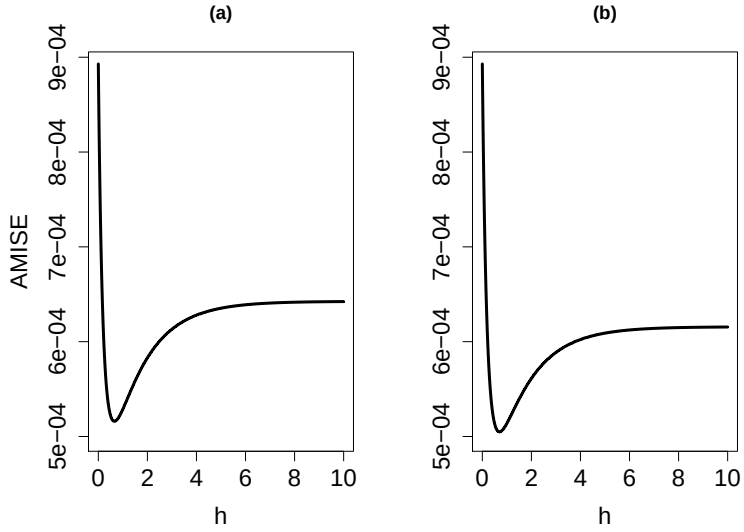

FIG. 2.7 – Suite et fin de Figure 2.6 pour $n = 1000$ avec $a = 1$: (a) $\sum_{x\in\mathbb{N}}\{f_0^{(2)}(x)\}^2 = 0.00465$, donc $h_0^* = 0.35$; (b) $\sum_{x\in\mathbb{N}}\{f^{(2)}(x)\}^2 = 0.00489$, donc $h^* = 0.32$

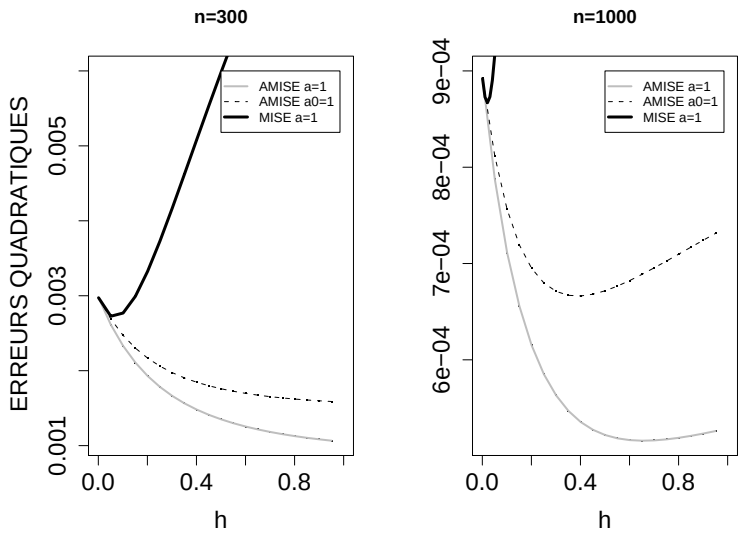

FIG. 2.8 – Graphique des erreurs quadratiques de l'estimateur à noyau triangulaire discret pour la distribution du mélange de Poisson $f = 0.4\mathcal{P}(0.5) + 0.6\mathcal{P}(10)$

2.4.2 Données de buts

Sans trop chercher l'approche d'une modélisation pour l'instant (*e.g.* Karlis & Ntzoufras, 2003, et quelques unes de leurs références pour les modèles paramétriques), nous utilisons les estimateurs à noyaux discrets triangulaires pour lisser de manière discrète la distribution des buts dans $n = 380$ matchs du championnat de football de la Ligue 1 française (saison 2005-2006). Puis, nous comparons les résultats obtenus à ceux analysés à l'aide d'un noyau binomial déjà présenté en Section 1.6.4.

Par rapport à son équivalent espagnol (Liga) de la même saison et de nombre identique de matchs, la Ligue 1 française possède un déficit de 123 buts. Tables 2.3 et 2.4 rappellent les données du nombre de buts de ces deux championnats européens de football ainsi qu'un résumé des statistiques élémentaires. On peut remarquer qu'entre autre 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 Kokonendji *et al.* (2008) pour des modèles (semi-)paramétriques.

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

TAB. 2.3 – Données du nombre de buts des championnats de football de la saison 2005-2006 avec un total de $n = 380$ matchs pour la Ligue 1 française et pour la Liga espagnole (cf. Table 1.6)

	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

TAB. 2.4 – Résumé des statistiques de Table 2.3 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 (cf. Table 1.7)

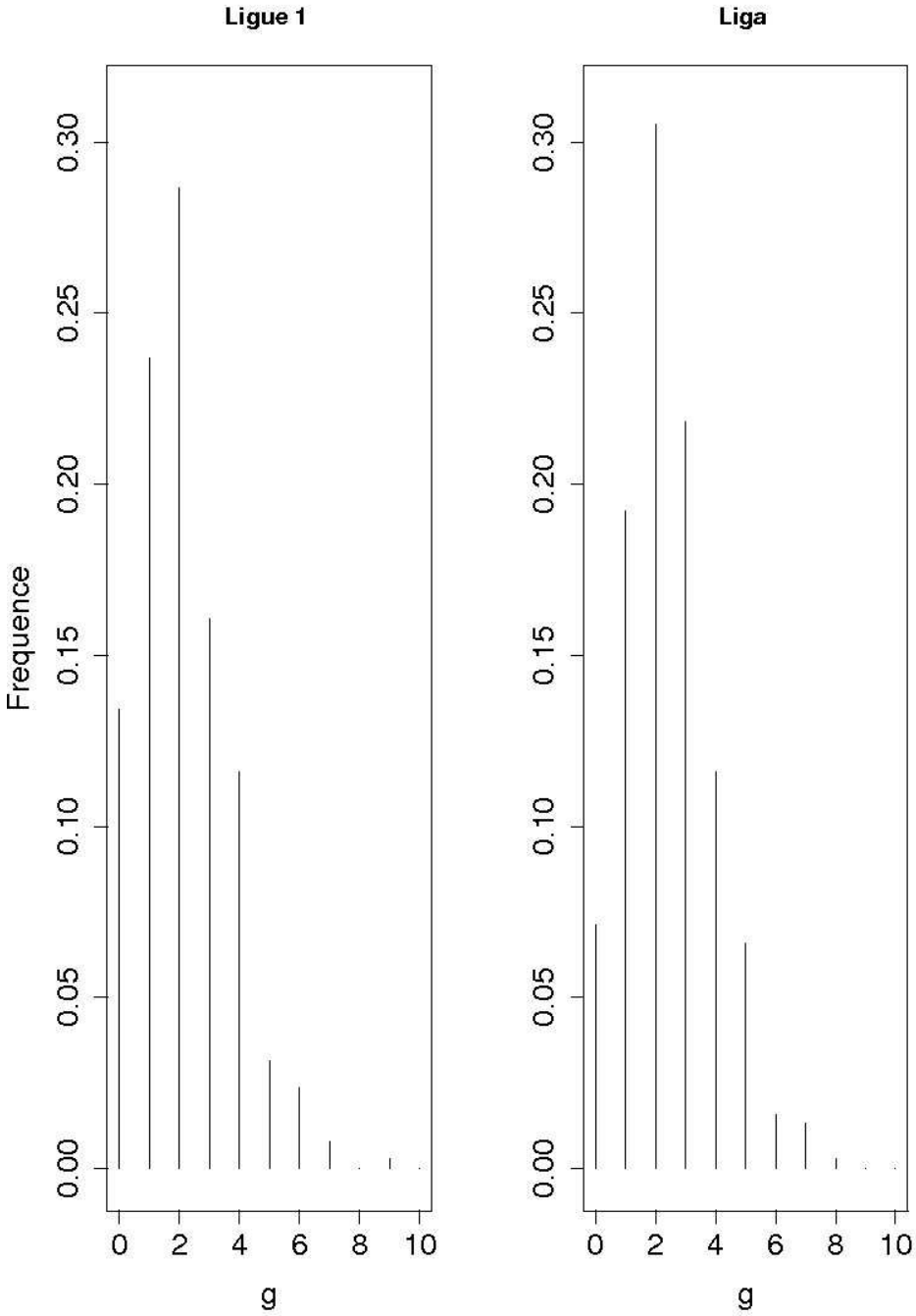

FIG. 2.9 – Distribution de buts des championnats de football de Ligue 1 française et de Liga espagnole (saison 2005-2006) avec $n = 380$ matchs

	Triang($a=1$)	Triang($a=2$)	Triang($a_0=1$)	Binomial
h_{cv}	0.204	0.055	0.028	0.177
C	0.98600	0.98618	0.99810	0.95872
ISE^0	0.00042	0.00023	0.00001	0.00395
$\widetilde{AMISE}$	0.00147	0.00177	0.00168	0.00163

TAB. 2.5 – Resultats du lissage discret par les noyaux discrets triangulaires et binomial des données de football de Ligue 1 française avec $n = 380$

Figure 2.9 représente les deux distributions empiriques f_0 de buts de Ligue 1 et de Liga. Contrairement à la Liga espagnole, on constate une forme moins régulière de f_0 de la Ligue 1 française avec une importante proportion de matchs nuls 0-0 ($f_0(0) = 0.134$ pour la Ligue 1 et $f_0(0) = 0.071$ pour la Liga). Ainsi, nous n'appliquons ces estimateurs qu'aux données de la Ligue 1. Rappelons ici que, depuis la saison 2006-2007 en France, un nouveau classement dans le championnat de Ligue 1 a fait son apparition. Il est appelé Classement de l'offensive* dont l'objectif est de mettre fin à la pénurie de buts et d'améliorer le spectacle dans le football français.

Dans Figure 2.10, nous avons tracé les fonctions $h \mapsto AMISE(n, h, T_{a_0}, f)$ et $h \mapsto AMISE(n, h, T_a, f)$ pour les bras $a \in \{1, 2\}$ et $a_0 \in \{1, 2\}$. Pour $a = 1$ et $a_0 = 1$, la courbe de $AMISE(n, h, T_{a_0}, f_0)$ est au-dessus de celle de $AMISE(n, h, T_a, f_0)$. Pour le bras $a = 2$ et le bras modifié $a_0 = 2$, la courbe de $AMISE(n, h, T_{a_0}, f_0)$ est au-dessus puis en dessous de celle de $AMISE(n, h, T_a, f_0)$ mais le point d'intersection des courbes est proche de 0. Dans ce qui suit nous n'utilisons pas les valeurs de h qui minimise $AMISE(n, h, T_a, f_0)$ mais le paramètre optimal de lissage discret h_{cv} [voir (2.25)] déterminé par la méthode de validation croisé.

Pour évaluer la performance des estimateurs, nous utilisons simplement l'erreur quadratique intégrée

$$ISE^0 = \sum_{x \in \mathbb{N}} \left\{ \widetilde{f}_{n,h_{cv}(a)}(x) - f_0(x) \right\}^2$$

(laquelle a l'avantage de pouvoir être directement observée sur les graphiques ; voir Marron & Padgett, 1987) et aussi une estimation $\widetilde{AMISE}(f_{n,h_{cv}(a)})$ de $AMISE(n, h, T_a, f)$ pour $a \in \mathbb{N}$ fixé. Table 2.5 présente les résultats de lissages discrets des données

*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

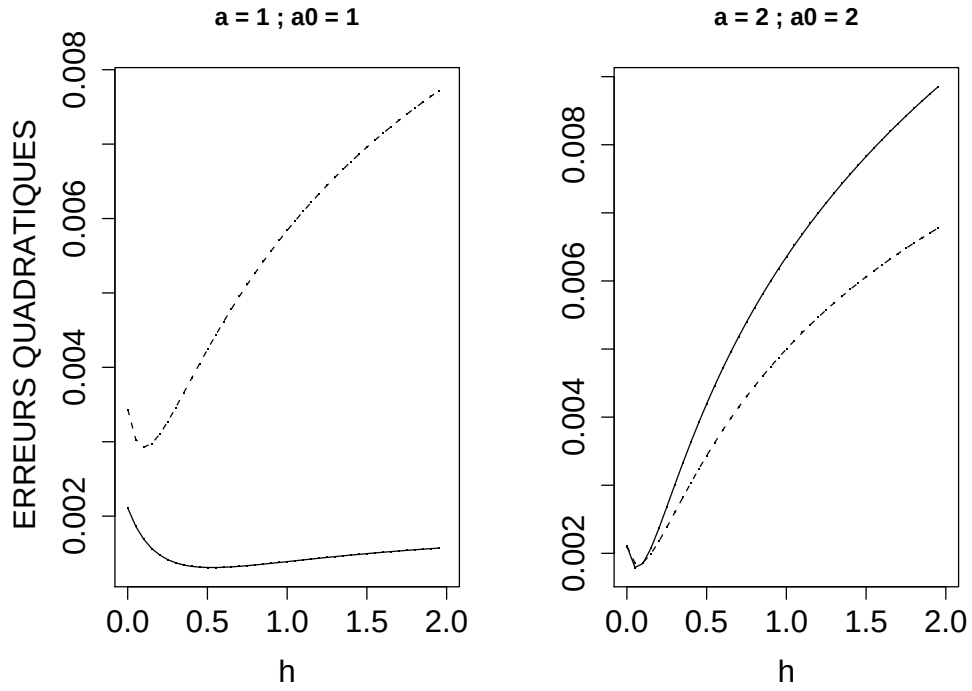

FIG. 2.10 – Graphique de $AMISE(n, h, T_a, f_0)$ [continu] et $AMISE(n, h, T_{a_0}, f_0)$ [tiret] de l'estimateur à noyau triangulaire discret pour la distribution de buts de Ligue 1 (saison 2005-2006) avec $n = 380$.

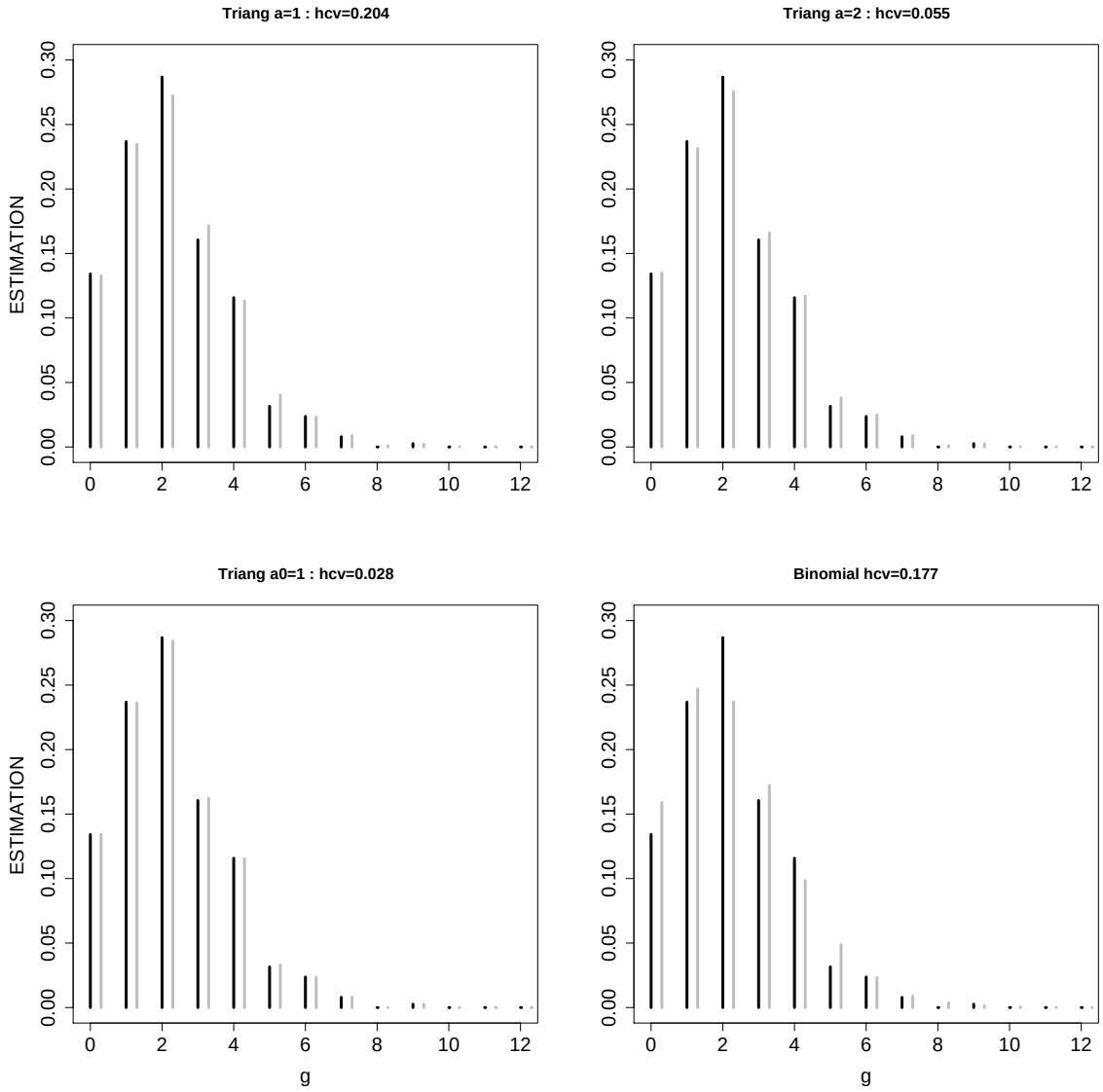

FIG. 2.11 – Lissages discrets [bâtons gris] par des noyaux discrets triangulaires $a_0 = 1$, $a \in \{1, 2\}$ et binomial de la distribution empirique [bâtons noirs] des données de football de la Ligue 1 française avec $n = 380$

de Ligue 1 par les estimateurs à noyaux discrets triangulaires pour $a \in \{1,2\}$ et $a_0 = 2$ avec les valeurs $h_{cv}(a)$ et la constante de normalisation C correspondantes. En utilisant la mesure $\widetilde{AMISE}(f_{n,h_{cv}(a)})$, la meilleure performance de l'estimateur à noyau associé discret triangulaire est obtenu avec $a = 1$. Ceci est du à la proportion $f_0(0) = 0.134$ de zéros dans l'échantillon qui est plus petite que la proportion $f_0(0) = 0.243$ dans l'échantillon simulé de (2.28). Par conséquent, le bras $a_0 = 1$ ne donne pas de meilleur résultat que $a = 1$. Par contre, en utilisant le critère ISE^0, c'est le bras modifié $a_0 = 1$ qui est plus performant que $a = 1$, bien que tous les deux donnent des résultats du même ordre. La différence entre ces deux mesures s'expliquent par le fait que $\widetilde{AMISE}(f_{n,h_{cv}(a)})$ soit une mesure approchée théorique qui est biaisée par l'usage des différences finies de f dans son calcul. Les noyaux associés discrets triangulaires améliorent les performances des noyaux discrets standards. On le met en évidence par une comparaison avec le meilleur d'entre eux : le noyau binomial qui a pour estimateur associé

$$\begin{aligned}\widetilde{f}_{n,h}(x) &= \frac{1}{n}\sum_{i=1}^{n} B_{x,h}(X_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} \frac{(x+1)!}{X_i!(x+1-X_i)!}\left(\frac{x+h}{x+1}\right)^{X_i}\left(\frac{1-h}{x+1}\right)^{x+1-X_i}, \quad x \in \mathbb{N},\end{aligned}$$

où $h \in]0,1]$ est le paramètre de lissage discret et $B_{x,h}$ est le noyau associé à la loi binomiale $\mathcal{B}\{x+1,(x+h)/(x+1)\}$ de support $\aleph_x = \{0,1,\cdots,x+1\}$; voir l'Exemple 1.2.2 du Chapitre 1 pour des détails assez comparables aux noyaux asssociés discrets triangulaires (2.5). Les résultats correspondants à l'estimation par le noyau binomial pour la valeur optimale $h_{cv} = 0.177$ sont comparés dans Figure 2.11 et Table 2.5 avec ceux des noyaux associés discrets triangulaires. Les valeurs du critère d'erreur $\widetilde{AMISE}(f_{n,h_{cv}(a)})$ pour les noyaux associés discrets triangulaires ($\widetilde{AMISE} = 1.47 \times 10^{-3}$ pour $a = 1$) sont meilleures que celles du noyau binomial ($\widetilde{AMISE} = 1.63 \times 10^{-3}$) bien que très voisines. Quant au critère ISE^0 généralement utilisée en pratique, on peut voir clairement que les meilleurs résultats sont aussi obtenus avec les noyaux associés discrets triangulaires ($ISE^0 = 10^{-5}$ for $a_0 = 1$) en comparaison avec le noyau binomial ($ISE^0 = 3.95 \times 10^{-3}$). Le critère ISE^0 étant plus approprié que $\widetilde{AMISE}$, nous concluons que l'estimateur à noyau associé discret triangulaire ($a_0 = 1$) est meilleur que celui à noyau binomial pour ces données de comptage.

2.5 Conclusion

Tout d'abord, les lois triangulaires discrètes $\mathcal{T}_{a;c,h}$ d'ordre $h > 0$ complètent la panoplie des familles de lois discrètes existantes (voir, par exemple, Johnson *et al.*, 2005). En outre, elles peuvent être étendues au cas multivarié.

Ensuite, les noyaux discrets triangulaires vérifient toutes les conditions d'un noyau associé discret et améliorent ainsi les noyaux discrets standards. Ils présentent de très satisfaisantes propriétés, à tailles finies ainsi qu'asymptotiques. En effet, les estimateurs à noyaux discrets triangulaires contiennent l'estimateur empirique ou naïf ; ce qui les rend souvent comparables pour des petites valeurs du bras a. Asymptotiquement par rapport à n, ils sont plus performants que les estimateurs à noyau discret standard. La qualité de lissage par les estimateurs à noyaux discrets triangulaires s'améliore lorsque les valeurs du paramètre de lissage discret h, des bras a et a_0 sont au voisinage de leurs valeurs optimales respectives. Cependant, l'effet du bras modifié a_0 sur la qualité d'estimation par rapport à celui du bras normal a est nettement plus perceptible dans le cas de données de comptage avec un excès de zéros. À travers le bras modifié, on a résolu indirectement à la fois le problème du biais de bordure et l'excès de zéros.

Enfin, les estimateurs à noyaux discrets triangulaires sont appropriés pour quasiment tous les types de données discrètes (parsemées ou non), en particulier les données de comptages telles qu'en économie, finance, écologie, agriculture, sports, épidémiologie, médecine, assurance, etc (voir aussi Table B.4 de l'Annexe B, pour des simulations). De plus, la mise en oeuvre est à la portée de tout utilisateur de la statistique (voir Section C.2 de l'Annexe C pour quelques codes sous R). Les diverses applications des noyaux associés discrets (triangulaires ou standards) sont maintenant possibles telles que des estimations non-paramétriques des fonctions discrètes de poids $\omega(x)$ ou de régression $m(x) = \mathbb{E}(Y|X = x)$, où Y est une variable aléatoire réelle à expliquer.

Chapitre 3

Estimation semi-paramétrique

3.1 Introduction

Soit $X_1, X_2, \ldots, X_n$ une suite de variables aléatoires i.i.d de fonction de masse de probabilité inconnue $f(x) := \Pr(X_i = x)$ sur l'ensemble des entiers naturels $\mathbb{N}$. La version discrète de l'estimateur à noyau continu pour f (voir, par exemple, Chen 1999, 2000a ; Rosenblatt, 1956 ; Scaillet, 2004) peut être exprimée par

$$\widetilde{f}_n(x) = \frac{1}{n}\sum_{i=1}^{n} K_{x,h}(X_i) = \frac{1}{n}\sum_{i=1}^{n} K(X_i; x, h), \quad x \in \mathbb{N}, \tag{3.1}$$

où $h > 0$ est la fenêtre ou paramètre de lissage discret, K est le type de noyau discret et $K_{x,h}$ le noyau associé discret (à repréciser dans la Section 3.2). En dehors du contexte des données de dénombrement, l'estimateur non-paramétrique (3.1) est bien connu pour être complètement impartial d'une forme particulière de la fonction sous-jacente f à estimer ; voir, par exemple, Izenman (1991) pour une revue de fonction de densité de probabilité. Cependant, la robustesse de cet estimateur à un prix. De manière générale, le choix de la fonction noyau n'est vraiment pas très important asymptotiquement, comme le montre l'usage de l'estimateur fréquence. Mais, si nous disposons d'échantillons de petites tailles, la structure du noyau peut alors jouer un rôle crucial dans l'approximation de la distribution recherchée, en particulier pour les variables aléatoires de dénombrement. De là, le choix du noyau associé discret devient aussi primordial que celui de la fenêtre de lissage optimal pour des échantillons de tailles petites et modérées. De plus, la vitesse de convergence des estimateurs à noyaux continus est en général plus lente que celle des estimateurs paramétriques. Le biais induit par la procédure du lissage discret peut être substantiel même pour des échantillons de tailles modérées. Comme les distributions de dénombrement ont pour support l'ensemble $\mathbb{N} = \{0, 1, 2, \cdots\}$ borné à gauche, il s'avère alors nécessaire de résoudre l'éventuel problème de biais de bordure. Ce dernier dépend du type de noyau discret utilisé ou de phénomènes particuliers des données de comptage. Toutefois, dans ce chapitre, nous

ne visons pas la résolution du problème de biais de bordure. Il faut noter qu'en dehors de l'estimateur fréquence ou naïf et des estimateurs à noyaux discrets présentés dans les deux premiers chapitres, Aitchison & Aitken (1976) ont été les pionniers des estimateurs à noyau discret dans le sens donné en (3.1). Cependant, le noyau discret d'Aitchison & Aitken (1976) associé à leur estimateur non-paramétrique n'a qu'une seule forme et est essentiellement pour des données catégorielles ou des distributions à support fini ; voir la Section A.5 de l'Annexe A pour des précisions. En ce qui concerne les données de dénombrement traitées ici, nous avons présenté dans l'Introduction générale quelques unes de leurs propriétés, leurs particularités et les difficultés associées. De plus, notons que pour l'analyse de certaines données de comptage, Böhning (2000) a développé une approche dite non-paramétrique au sens d'un mélange fini de distributions.

Dans ce chapitre, nous proposons un estimateur semi-paramétrique d'une distribution de données de dénombrement analogue à celui proposé dans le cas continu par Hjort & Glad (1995). Pour ce faire, nous partons du constat que toute distribution d'une variable de dénombrement peut s'écrire comme la distribution d'une loi de Poisson pondérée par une fonction de poids appropriée telle que

$$\begin{aligned} f(x) &= \frac{\omega(x)\, p(x;\theta)}{\sum_{x\in\mathbb{N}} \omega(x) p(x;\theta)} \\ &= \omega(x;\theta) p(x;\theta) \\ &=: f_\omega(x;\theta), \quad x \in \mathbb{N} \end{aligned} \tag{3.2}$$

(voir Kokonendji *et al.*, 2008, pour le cas poissonnien). La partie paramétrique $p(x;\theta)$ de (3.2) est une fonction de masse de probabilité sur $\mathbb{N}$ de forme connue et de paramètre inconnu θ à estimer. La partie non-paramétrique $\omega(x)$ de (3.2) est une fonction discrète de poids inconnue de telle sorte que la fonction normalisée $\omega(x;\theta) := \omega(x) \left\{\sum_{x\in\mathbb{N}} \omega(x) p(x;\theta)\right\}^{-1}$ de $\omega(x)$ soit à estimer de manière non-paramétrique connaissant θ. Pour simplifier la présentation, nous supposons que la partie paramétrique $p(x;\theta)$ de (3.2) n'est autre qu'une Poisson comme cela a été largement discuté et justifié dans les Motivations de l'Introduction générale autour de l'expression (1). Cette approche d'estimation semi-paramétrique a été évoquée à la fin du travail de Kokonendji *et al.* (2008). Ce sera un estimateur naturellement compétitif aussi bien avec l'estimateur non-paramétrique (3.1) qu'avec l'estimateur paramétrique (3.2) dans le cas où la loi Poisson pondérée est bien explicitée. De plus, en supposant que la fonction de masse de probabilité inconnue $f(\cdot) = f_\omega(\cdot;\theta)$ est définie sur un ensemble fini $\{0, 1, \ldots, N\}$ de $\mathbb{N}$, nous considérons $f_\omega(\cdot;\theta)$ comme une distribution binomiale pondérée ; voir, par exemple, Johnson *et al.* (2005) [pages 149–150], Chakraborty & Das (2006), et leurs références. Cette estimation semi-paramétrique $f(\cdot) = f_\omega(\cdot;\theta)$ est un compromis entre la pure estimation non-paramétrique (3.1) et l'estimation paramétrique courante de Poisson/binomiale modifiée. Quand la fonction poids discrète ω ne représente pas théoriquement le vrai mécanisme dans (3.2) ou n'est pas bien

spécifiée, il est plus approprié d'utiliser l'approche semi-paramétrique pour estimer $f(\cdot) = f_\omega(\cdot;\theta)$ afin de «laisser faire les données» pour l'estimation non-paramétrique de ω. L'estimateur proposé sera, à la fois, l'analogue discret et une version particulière de l'estimateur proposé par Hjort & Glad (1995) pour les données continues. Dans le cas qui nous concerne, la fonction poids en chaque point discret peut être considérée comme le coefficient multiplicatif de correction locale. Ceci vise à accomoder un départ ponctuel à partir de la loi paramétrique de référence $p(x;\theta)$: Poisson ou binomiale. Cette méthode d'estimation est, par dessus tout, simple et efficace pour estimer une fonction de masse de probabilité f inconnue. De plus, elle reste performante même si la fonction de masse de probabilité inconnue f ne peut pas être bien approchée par la loi paramétrique de référence. Nous étudions les propriétés statistiques fondamentales de cette procédure d'estimation et nous la comparons aux estimateurs (non-paramétriques) des distributions de dénombrement.

La suite de ce chapitre est organisée de la façon suivante. Dans la Section 3.2, nous rappelons les principaux résultats des deux chapitres précédents dont on a besoin ici, essentiellement la mise en oeuvre des estimateurs à noyaux discrets asymétriques standards et symétriques. La Section 3.3 définit l'estimateur semi-paramétrique d'une fonction de masse de probabilité sur $\mathbb{N}$ en la considérant comme une Poisson pondérée. Elle présente aussi quelques propriétés de cet estimateur. En outre, nous évaluons la performance de cet estimateur par rapport à celle de l'estimateur à noyau discret classique (3.1). Dans la Section 3.4, nous discutons des effets des phénomènes de dispersion et de proportion de zéros sur cette méthode d'estimation. La Section 3.5 présente une extension de l'approche semi-paramétrique sur la base de binomiale pondérée pour l'estimation d'une fonction de masse de probabilité à support fini. Dans la Section 3.6, nous illustrons l'ajustement de modèle à travers trois jeux de données : le premier jeu de données est celui des nombres de buts de la Ligue 1 française de 2005-2006 ; le deuxième jeu de données provient d'une étude sociologique menée sur la consommation d'alcool sur les jours d'une semaine (Alanko & Lemmens, 1996) ; le troisième est un jeu de données simulées d'un mélange de Poisson. Dans la Section 3.7, nous présentons un modèle de diagnostique lequel est très important pour notre approche d'analyse des données de dénombrement. Nous concluons par quelques remarques dans la Section 3.8.

3.2 Récapitulatif de la méthode des noyaux discrets

Cette section résume et précise les principaux résultats sur les estimateurs à noyaux discrets présentés dans les chapitres précédents afin de faciliter la comparaison avec le nouvel estimateur.

Soit $x \in \mathbb{N}$ et $h > 0$. Nous avons défini le *noyau associé discret* $K_{x,h}(\cdot)$ comme étant une fonction de masse de probabilité sur $\aleph_x$ liée à la variable aléatoire discrète

$\mathcal{K}_{x,h}$ telle que $\cup_{x\in\mathbb{N}}\aleph_x \supseteq \mathbb{N}$, var$(\mathcal{K}_{x,h}) < +\infty$ et on a les deux dernières conditions :

$$\lim_{h\to 0}\mathbb{E}(\mathcal{K}_{x,h}) = x, \tag{3.3}$$

$$\lim_{h\to 0}\text{var}(\mathcal{K}_{x,h}) = 0. \tag{3.4}$$

Rappelons que le support $\aleph_x$ contient au moins x et ne dépend pas de h. De plus, la première condition sur le support $\mathbb{N}$ de la fonction de masse de probabilité inconnue et $\aleph_x$ du noyau associé discret $K_{x,h}$ peut être remplacée par $\aleph_x \subseteq \mathbb{N}$. Ensuite, nous avons souligné l'importance des conditions (3.3) et (3.4) qui garantissent la convergence ponctuelle de l'estimateur à noyau discret (3.1) ; voir Propositions 1.2.4 et 1.2.5 du Chapitre 1. En fait, la condition (3.3) reflète l'une des principales différences avec le cas continu symétrique où les noyaux correspondants sont centrés autour de 0 tels que $K_{x,h}(\cdot) = (1/h)K\{(x-\cdot)/h\}$. De cette manière, cette condition exprime clairement que l'estimateur à noyau discret $\widetilde{f}_n$ de f est à noyau variable. Nous distinguons deux familles de noyaux (associés) discrets tels que $\mathbb{E}(\mathcal{K}_{x,h}) = x + h$ ou $\mathbb{E}(\mathcal{K}_{x,h}) = x$. Ces noyaux associés discrets sont analogues aux noyaux continus asymétriques utilisés par Chen (1999, 2000a) et plus récemment par Scaillet (2004). La dernière condition (3.4) garantit que l'estimateur à noyau discret $\widetilde{f}_n$ de f défini en (3.1) se comporte asymptotiquement comme l'estimateur fréquence (voir Proposition 1.2.5 pour la convergence ponctuelle en moyenne quadratique).

Nous rappelons brièvement quelques propriétés de l'estimateur à noyau discret $\widetilde{f}_n$ défini en (3.1). Notons d'abord que la fonction $x \mapsto \widetilde{f}_n(x)$ est une fonction de masse de probabilité à une constante de normalisation près : $\widetilde{C} = \sum_{x\in\mathbb{N}}\widetilde{f}_n(x)$. Une propriété fondamentale pour les calculs est donnée par :

$$\mathbb{E}\{\widetilde{f}_n(x)\} = \mathbb{E}\{f(\mathcal{K}_{x,h})\}. \tag{3.5}$$

Par exemple, on exprime le biais ponctuel de $\widetilde{f}_n$ par :

$$\begin{aligned}\text{biais}\{\widetilde{f}_n(x)\} &= \mathbb{E}\{f(\mathcal{K}_{x,h})\} - f(x)\\ &= f\{\mathbb{E}(\mathcal{K}_{x,h})\} - f(x) + \frac{1}{2}\text{var}(\mathcal{K}_{x,h})f^{(2)}(x) + o(h),\end{aligned} \tag{3.6}$$

où $f^{(2)}$ est la différence finie d'ordre 2 telle que

$$f^{(2)}(x) = \begin{cases}\{f(x+2) - 2f(x) + f(x-2)\}/4 & \text{si } x \in \mathbb{N}\setminus\{0,1\}\\ \{f(3) - 3f(1) + 2f(0)\}/4 & \text{si } x = 1\\ \{f(2) - 2f(1) + f(0)\}/2 & \text{si } x = 0.\end{cases} \tag{3.7}$$

De l'expression (3.6), il vient que le biais dépend de la moyenne $\mathbb{E}(\mathcal{K}_{x,h})$ et de la variance var$(\mathcal{K}_{x,h})$ du noyau associé discret $K_{x,h}$.

En ce qui concerne la variance ponctuelle, elle peut s'exprimer par :

$$\begin{aligned}\text{var}\left\{\widetilde{f}_{n,h,K}(x)\right\} &= \frac{1}{n}\text{var}\{K_{x,h}(X_1)\} \\ &= \frac{1}{n}\left[\mathbb{E}K_{x,h}^2(X_1) - \{\mathbb{E}K_{x,h}(X_1)\}^2\right] \\ &= \frac{1}{n}f(x)\left\{1-f(x)\right\}\left\{\Pr(\mathcal{K}_{x,h}=x)\right\}^2 + R_n(x;h), \quad (3.8)\end{aligned}$$

avec

$$\begin{aligned}R_n(x;h) &= \frac{1}{n}\left[\sum_{y\in\aleph_x\setminus\{x\}} f(y)\left\{\Pr(\mathcal{K}_{x,h}=y)\right\}^2 + \left\{f(x)\Pr(\mathcal{K}_{x,h}=x)\right\}^2\right] \\ &\quad -\frac{1}{n}\left\{\sum_{z\in\aleph_x} f(z)\Pr(\mathcal{K}_{x,h}=z)\right\}^2. \quad (3.9)\end{aligned}$$

Sous l'hypothèse d'un noyau associé discret, le reste vérifie $R_n(x;h) = o(1/n)$.

REMARQUE 3.2.1 Pour l'estimateur non-paramétrique (3.1), la performance relative entre deux noyaux (associés) discrets $\mathcal{K}_{x,h}^1$ et $\mathcal{K}_{x,h}^2$ tels que $\mathbb{E}(\mathcal{K}_{x,h}^1) = \mathbb{E}(\mathcal{K}_{x,h}^2)$ peut se mesurer par la différence de leur variance $\text{var}(\mathcal{K}_{x,h}^1) - \text{var}(\mathcal{K}_{x,h}^2)$ pour les noyaux discrets standards ; autrement dit, ceci est intéressant pour le cas où $\text{var}(\mathcal{K}_{x,h}^i) \nrightarrow 0$ quand $h \to 0$, $i = 1, 2$.

Maintenant, nous rappelons les deux exemples de familles de noyaux discrets : les asymétriques standards (dont binomial) et les symétriques (triangulaires discrets). La Figure 3.1 présente un graphique comparatif de leurs allures pour une cible x et une fenêtre h fixées ; leurs principales propriétés sont résumées dans la Table 3.1. Notons qu'asymptotiquement le choix du noyau associé discret n'est pas très important comme le montre l'utilisation de l'estimateur fréquence. Toutefois, lorsque l'on est en présence d'échantillons de tailles petites ou modérées, la structure du noyau joue un rôle crucial pour approcher la distribution de l'échantillon pour les variables aléatoires de dénombrement.

EXEMPLE 3.2.1 (Binomial). Considérons une loi binomiale $\mathcal{B}(N,p)$, $N \in \mathbb{N}, p \in [0,1]$. Le noyau binomial $B_{x,h}$ suit la loi $\mathcal{B}\{x+1, (x+h)/(x+1)\} =: \mathcal{B}_{x,h}$ avec $h \in]0,1]$ et $\aleph_x = \{0,1,\ldots,x+1\}$. À partir de Remarque 3.2.1, il est l'un des plus performants dans la famille des noyaux discrets standards asymétriques $\mathcal{K}_{x,h}$ dont font partie le Poisson et le binomial négatif de moyenne $\mathbb{E}(\mathcal{K}_{x,h}) = x + h$. Le bon comportement à taille finie du noyau binomial vient de sa propriété de sousdipsersion : $\text{var}(\mathcal{B}_{x,h}) = (x+h)(1-h)/(x+1) < x+h$. Rappelons que le noyau binomial $B_{x,h}$ est du premier ordre car $\lim_{h\to 0}\text{var}(\mathcal{B}_{x,h}) = x/(x+1) \neq 0$ pour $x \in \mathbb{N}\setminus\{0\}$; autrement dit la condition (3.4) n'est pas vérifiée. L'estimateur à noyau binomial s'écrit

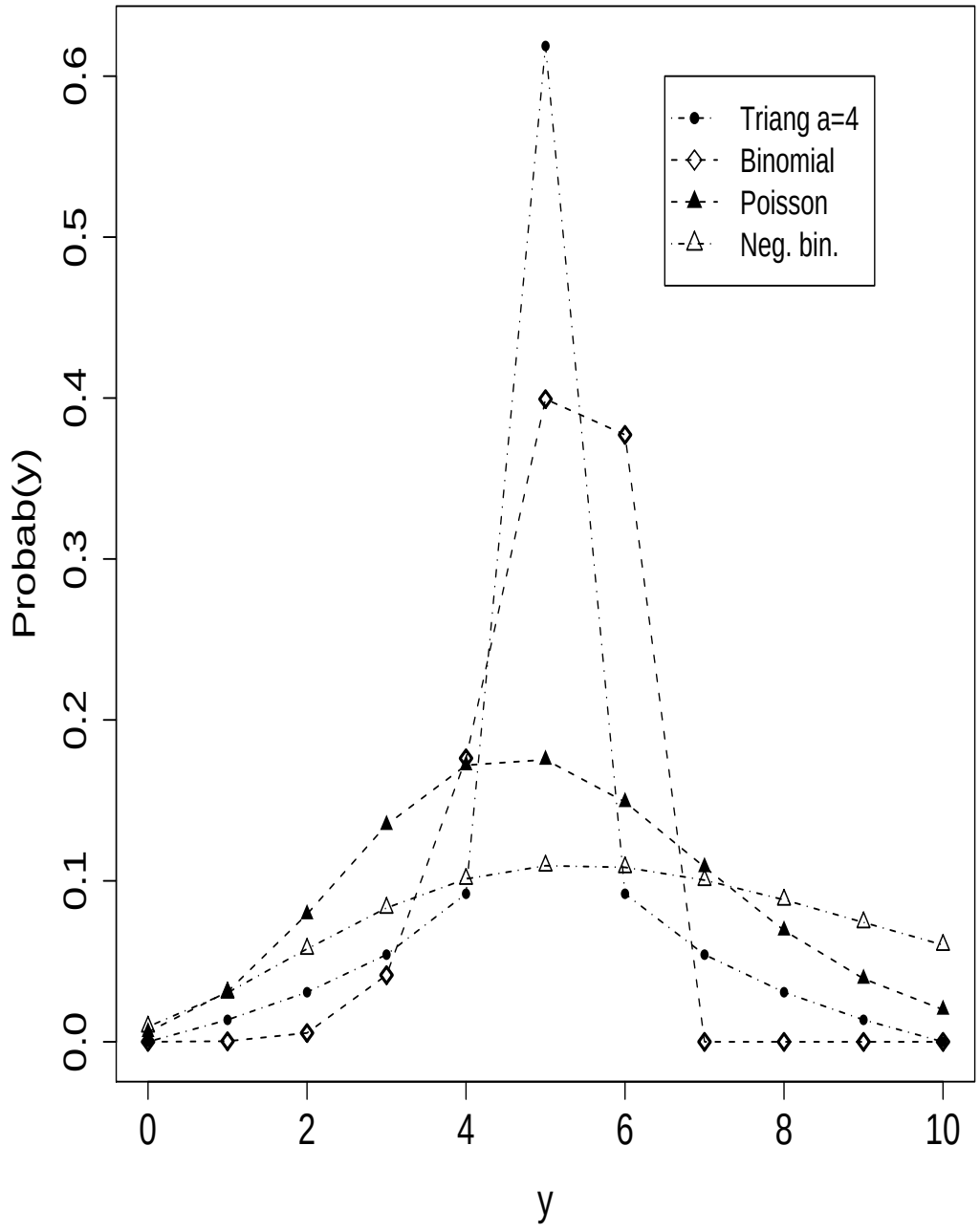

FIG. 3.1 – Noyaux discrets triangulaire, binomial, de Poisson et binomial négatif pour la cible $y = x = 5$ et le paramètre de lissage $h = 0.1$

ici :

$$\begin{aligned}\widetilde{f}_n^B(x) &= \frac{1}{n}\sum_{i=1}^{n} B_{x,h}(X_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} \frac{(x+1)!}{X_i!(x+1-X_i)!}\left(\frac{x+h}{x+1}\right)^{X_i}\left(\frac{1-h}{x+1}\right)^{x+1-X_i}, \quad x \in \mathbb{N},\end{aligned}$$

avec $X_i \leq x+1$. Une expression de sa variance ponctuelle se déduit de (3.8) par

$$\text{var}\{\widetilde{f}_n^B(x)\} = \frac{(1-h)^2}{n} f(x)\{1-f(x)\}\left(\frac{x+h}{x+1}\right)^{2x} + R_n^B(x;h).$$

Pour obtenir le biais, nous réalisons deux améliorations dans l'expression (3.6) : on effectue d'abord un second développement de Taylor discret pour $f\{\mathbb{E}(\mathcal{K}_{x,h})\} = f(x+h) = f(x) + hf^{(1)}(x) + o(h)$, puis on considère la partie principale de la variance $\text{var}(\mathcal{B}_{x,h}) = (x+h)/(x+1) - xh/(x+1) + o(h)$. Ainsi, on obtient :

$$\text{biais}\{\widetilde{f}_n^B(x)\} = hf^{(1)}(x) + \frac{1}{2}\left(\frac{x+h-xh}{x+1}\right) f^{(2)}(x) + o(h),$$

avec

$$f^{(1)}(x) = \begin{cases} \{f(x+1) - f(x-1)\}/2 & \text{si } x \in \mathbb{N}\setminus\{0\} \\ f(1) - f(0) & \text{si } x = 0. \end{cases} \tag{3.10}$$

Les estimateurs à noyaux asymétriques ne posent pas de problème de biais de bordure, mais le biais dépend des différences finies d'ordre 1 et 2 de la fonction de masse inconnue f. Ceci est dû au fait que la cible x n'est pas la moyenne du noyau binomial mais plutôt le mode. Par comparaison à la famille des estimateurs à noyaux triangulaires discrets symétriques que nous présentons par la suite, le biais ne dépend que de la différence finie d'ordre 2 de f. Finalement, la sélection de la fenêtre h de lissage discret pour les estimateurs à noyaux discrets standards se fait de manière classique par la méthode de validation croisée. Pour la situation particulière d'excès de zéros, on peut choisir la fenêtre h_0 de lissage discret en résolvant l'équation

$$\sum_{i=1}^{n} \Pr\left(\mathcal{K}_{X_i,h_0} = 0\right) = n_0, \tag{3.11}$$

où $n_0 = \sharp\left(X_i = 0\right)$ est le nombre d'observations égales à zéros. Pour l'estimateur à noyau binomial, la fenêtre adaptée h_0 est telle que $\sum_{i=1}^{n}\{(1-h_0)(X_i+1)^{-1}\}^{X_i+1} = n_0$.

EXEMPLE 3.2.2 (Triangulaires discrets). Dans le Chapitre 2, nous avons introduit les lois triangulaires discrètes utiles, notamment, pour la construction des *noyaux associés discrets et symétriques*. Soit $(a,x,h) \in \mathbb{N}\times\mathbb{N}\times]0,+\infty[$. La variable aléatoire

$\mathcal{T}_{a;x,h}$ de support $\aleph_x = \{x, x \pm 1, \ldots, x \pm a\}$ est dite triangulaire discrète si sa probabilité individuelle est donnée par :

$$\Pr(\mathcal{T}_{a;x,h} = y) = \frac{(a+1)^h - |y-x|^h}{P(a,h)},\ y \in \aleph_x,$$

où $P(a,h) = (2a+1)(a+1)^h - 2\sum_{k=0}^{a} k^h$ est la constante de normalisation. Les trois paramètres (a, x, h) sont tels que a désigne le bras et est fixé dans $\mathbb{N}$, $x = \mathbb{E}(\mathcal{T}_{a;x,h})$ est le centre et représente la cible dans le problème d'estimation (3.1), et $h > 0$ est l'ordre (de la loi triangulaire discrète) lequel correspond au paramètre de lissage. Le bras nul $a = 0$ dans $\mathcal{T}_{a;x,h}$ correspond à la variable aléatoire de Dirac en x. Pour $a \neq 0$ fixé, l'estimateur à noyau discret triangulaire est donnée par

$$\begin{aligned} \widetilde{f}_n^{T_a}(x) &= \frac{1}{n}\sum_{i=1}^{n} T_{a;x,h}(X_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} \frac{(a+1)^h - |X_i - x|^h}{(2a+1)(a+1)^h - 2\sum_{k=0}^{a} k^h}, \quad x \in \mathbb{N}. \end{aligned}$$

En utilisant l'expression (3.8), leur variance peut s'exprimer par

$$\text{var}\{\widetilde{f}_n^{T_a}(x)\} = \frac{1}{n} f(x)\{1 - f(x)\}\left\{\frac{(a+1)^h}{P(a,h)}\right\}^2 + o\left(\frac{1}{n}\right).$$

Puis, à partir de la relation (3.6), leur biais est

$$\text{biais}\{\widetilde{f}_n^{T_a}(x)\} = \frac{1}{2} V(a,h) f^{(2)}(x) + o(h),$$

où $f^{(2)}$ est définie en (3.7) et

$$V(a,h) = \text{var}(\mathcal{T}_{a;x,h}) = \frac{1}{P(a,h)}\left\{\frac{a(2a+1)(a+1)^{h+1}}{3} - 2\sum_{k=0}^{a} k^{h+2}\right\} \tag{3.12}$$

ne dépend pas de la cible x. Signalons que la première fonction variance partielle $a \mapsto \mathrm{V}(a,h)$ est croissante et non bornée, et la seconde fonction $h \mapsto \mathrm{V}(a,h)$ croît de 0 à $a(a+1)/3$. Ainsi, la condition (3.4) est vérifiée pour $\mathcal{T}_{a;x,h}$. Remarque 3.2.1 permet de comparer les noyaux associés triangulaires discrets entre eux de telle sorte que $\mathcal{T}_{a_1;x,h}$ est plus performant que $\mathcal{T}_{a_2;x,h}$ quand $a_1 < a_2$. Les biais des estimateurs à noyaux discrets triangulaires ne dépendent pas de la différence finie $f^{(1)}$ d'ordre 1 comme dans le cas des noyaux continus symétriques. Cependant, pour $a \neq 0$, les estimateurs à noyaux discrets triangulaires induisent un biais de bordure à gauche car le support $\mathbb{N}$ de la fonction de masse inconnue f est contenu dans l'ensemble $\cup_x \aleph_x = \{-a, \ldots, -1\} \cup \mathbb{N}$. Pour des observations au bord $\{0, 1, \ldots, m\}$ (m petit comme 0, 1

ou 2), une solution originale à ce problème est de considérer un bras modifié a_0 de a tel que pour $k \in \mathbb{N} \setminus \{0\}$ et $x \in \mathbb{N}$ on ait

$$a_0 = k \Longleftrightarrow a_0 = \begin{cases} j & \text{si } x = j,\ j \in \{0, 1, \ldots, k-1\} \\ k & \text{si } x \in \{k, k+1, \ldots\}. \end{cases} \tag{3.13}$$

Cette procédure préserve la propriété de symétrie locale du noyau associé triangulaire discret autour de chaque cible. En ce qui concerne le choix de la fenêtre optimale, on ne retient ici que la méthode de validation croisée. Nous n'avons pas recours à la méthode de proportion de zéros (3.11) qui n'admet pas de solution pour les noyaux associés discrets triangulaires. En comparaison avec les noyaux discrets asymétriques, notons que le biais des noyaux associés discrets triangulaires dépend uniquement de la différence finie d'ordre 2. Finalement, par rapport à l'estimateur à noyau binomial, les estimateurs à noyaux discrets triangulaires sont convergents au sens du $MISE$.

REMARQUE 3.2.2 L'estimateur empirique pour des données de dénombrement correspond à l'estimateur à noyau discret triangulaire avec $a = 0$. Ainsi, son risque quadratique intégré asymptotique est de $(1/n)\left\{1 - \sum_{x \in \mathbb{N}} f^2(x)\right\}$. Il peut être pris comme un point de référence pour la convergence du $MISE$.

Type de noyau discret	$\mathbb{E}(\mathcal{K}_{x,h})$	$\text{var}(\mathcal{K}_{x,h})$	$\lim_{h\to 0}\text{var}(\mathcal{K}_{x,h})$	Convergence du $MISE$	Validation croisée	Excès de zéros	Symétrie de $\mathcal{K}_{x,h}$	Remarques
Dirac	x	0	0	OUI $(n \nearrow +\infty)$	$--$	$--$	OUI	Pas de fenêtre
Poisson	$x+h$	$x+h$	$x \in \mathbb{N}$	NON	OUI	OUI	NON	Équi-dispersion
Binomial	$x+h$	$(x+h)\left(\frac{1-h}{x+1}\right)$	$0 \leq \frac{x}{x+1} < 1$	NON	OUI	OUI	NON	Sous-dispersion
Binomial négatif	$x+h$	$(x+h)\left(1+\frac{x+h}{x+1}\right)$	$\frac{x(2x+1)}{x+1} \geq 0$	NON	OUI	OUI	NON	Sur-dispersion
Triangulaire $a \in \mathbb{N}\setminus\{0\}$	x	$V(a,h)$: voir (3.12)	0	OUI $(n \nearrow +\infty$ et $h \searrow 0)$	OUI	NON	OUI	Biais de bordure
Aitchison-Aitken	$x+o(h)$	voir formule (A.3)	0	OUI $(n \nearrow +\infty$ et $h \searrow 0)$	OUI	$--$	NON	Données catégorielles

TAB. 3.1 – Résumé des propriétés de quelques estimateurs à noyaux discrets

3.3 Estimateur semi-paramétrique

Nous rappelons à travers (3.2) que toute distribution ou variable aléatoire X de dénombrement de fonction de masse de probabilité inconnue $f(x) = \Pr(X = x)$ peut être considérée comme une distribution de Poisson pondérée :

$$f(x) = \omega(x;\mu)\, p(x;\mu) =: f_w(x;\mu), \quad \forall x \in \mathbb{N}, \tag{3.14}$$

où $p(x;\mu) := \mu^x e^{-\mu}/x! > 0$ est la fonction de masse de Poisson de moyenne $\mu > 0$ et $x \mapsto \omega(x;\mu) := \omega(x) \left\{\sum_{x\in\mathbb{N}} \omega(x) p(x;\mu)\right\}^{-1}$ est la fonction poids normalisée de Poisson. Pour estimer $f(\cdot) \equiv f_\omega(\cdot;\mu)$ dans (3.14), une procédure naturelle est de considérer le problème d'estimation sous l'angle semi-paramétrique. Ainsi, $p(x;\mu)$ est la partie paramétrique relative au paramètre μ et la fonction $x \mapsto \omega(x;\mu)$ est la partie non-paramétrique sur $\mathbb{N}$. Notons, cependant, que le rapport $\omega(x;\mu) = f(x)/p(x;\mu)$ peut ne pas être bien défini, notamment quand le dénominateur est proche de 0. De plus, nous considérons que la moyenne μ de la loi de Poisson de référence est la même que celle de la variable aléatoire Poisson pondérée X, *i.e.*, $\mathbb{E}(X) = \mu$; voir, par exemple, Kokonendji *et al.* (2008).

Considérons une suite de variables aléatoires $X_1, \ldots, X_n$, i.i.d. à X, de fonction de masse de probabilité inconnue f sur $\mathbb{N}$ présentée en (3.14). Suivant les travaux de Hjort & Glad (1995), nous définissons l'estimateur semi-paramétrique de f par

$$\begin{aligned}
\widehat{f}_n(x) &= p(x;\widehat{\mu}_n)\, \widetilde{w}_n(x;\widehat{\mu}_n) \\
&= p(x;\widehat{\mu}_n)\, \frac{1}{n}\sum_{i=1}^{n} \frac{K_{x,h}(X_i)}{p(X_i;\widehat{\mu}_n)} && (3.15)\\
&= \frac{1}{n}\sum_{i=1}^{n} K_{x,h}(X_i)\, \frac{p(x;\widehat{\mu}_n)}{p(X_i;\widehat{\mu}_n)} && (3.16)\\
&= \frac{1}{n}\sum_{i=1}^{n} K_{x,h}(X_i)\, \frac{X_i!(\widehat{\mu}_n)^{x-X_i}}{x!}, \quad x \in \mathbb{N}, && (3.17)
\end{aligned}$$

où $\widehat{\mu}_n = \overline{X}_n = n^{-1}(X_1 + \cdots + X_n)$ est la moyenne empirique qui est l'estimateur du maximum de vraisemblance de la moyenne μ de la loi de Poisson, $h > 0$ est la fenêtre ou le paramètre de lissage discret et $K_{x,h}$ est un noyau (associé) discret donné comme cela est rappelé dans la section précédente. De manière similaire à l'estimateur non-paramétrique défini en (3.1), l'estimateur semi-paramétrique $\widehat{f}_n$ de f ci-dessus est défini à constante de normalisation près $\widehat{C} = \sum_{x\in\mathbb{N}} \widehat{f}_n(x)$.

Examinons maintenant les différentes expressions de l'estimateur $\widehat{f}_n$ de f dans (3.14). Tout d'abord, à partir de l'expression (3.15), nous déduisons l'estimateur non-paramétrique naturel

$$\widetilde{w}_n(x;\widehat{\mu}_n) = \frac{1}{n}\sum_{i=1}^{n} \frac{K_{x,h}(X_i)}{p(X_i;\widehat{\mu}_n)}, \quad x \in \mathbb{N}, \tag{3.18}$$

de la fonction poids de Poisson $\omega(x;\mu)$ qui est le coefficient multiplicatif de correction local dépendant du paramètre μ. Cet estimateur «interne» (3.18) de $\omega(x;\mu)$ peut être considéré comme la moyenne empirique du rapport $K_{x,h}(\cdot)/p(\cdot;\widehat{\mu}_n)$. Il est évidemment plus approprié qu'un estimateur «externe» $\widetilde{w}_n^E(x;\widehat{\mu}_n) = \{p(x;\widehat{\mu}_n)\}^{-1} n^{-1} \sum_{i=1}^n K_{x,h}(X_i)$ qui conduit directement à l'estimateur non-paramétrique $\widetilde{f}_n$ de f défini par (3.1). On peut se référer, par exemple, à Mack & Müller (1989), Jones (1992), Patil, Wells & Marron (1994), dans d'autres contextes. Ainsi, nous estimons par (3.18) les mesures ponctuelles de départ d'une distribution de Poisson. L'estimation de la fonction correction ou poids $\widetilde{w}_n(x;\widehat{\mu}_n)$ est uniformément égale à 1 si la Poisson de référence est bien choisie. Par conséquent, si le degré de non-spécification n'est pas trop sévère, il est intuitivement plus naturel de modéliser le facteur de correction localement que la fonction de masse de probabilité inconnue elle-même. En utilisant (3.16) puis (3.17), nous notons que la nouvelle partie paramétrique

$$\frac{p(x;\widehat{\mu}_n)}{p(X_i;\widehat{\mu}_n)} = \frac{X_i!(\widehat{\mu}_n)^{x-X_i}}{x!}, \quad x \in \mathbb{N} \tag{3.19}$$

représente l'apport qui améliore l'estimateur à noyau discret ordinaire (3.1). De plus, si le rapport (3.19) est constant, on obtient alors l'estimateur à noyau discret traditionnel (3.1) avec un départ paramétrique implicite donné par une loi discrète uniforme impropre.

Nous examinons ci-dessous le biais et la variance de l'estimateur proposé et nous les comparons à ceux de l'estimateur non-paramétrique traditionnel. Pour cela, nous travaillons sous deux hypothèses : celle d'un départ Poisson connu puis celle d'un départ Poisson de paramètre inconnu.

3.3.1 Départ Poisson connu

Soit $p_0(x) = p(x;\mu_0)$ une fonction de masse de Poisson fixée comme départ dans (3.14). Nous écrivons donc $f = p_0\omega$ et, ensuite, nous estimons la fonction poids non-paramétrique ω par

$$\widetilde{w}_n(x) = \frac{1}{n}\sum_{i=1}^n \frac{K_{x,h}(X_i)}{p_0(X_i)},$$

aboutissant ainsi à l'estimateur

$$\widehat{f}_n(x) = \; p_0(x)\; \widetilde{w}_n(x) = \frac{1}{n}\sum_{i=1}^n K_{x,h}(X_i)\; \frac{p_0(x)}{p_0(X_i)}, \quad x \in \mathbb{N}. \tag{3.20}$$

Proposition 3.3.1 *Pour $n \to +\infty$ et $h = h(n) \to 0$, l'estimateur semi-paramétrique de (3.20) avec un départ Poisson fixé possède le biais et la variance suivants :*

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

et

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

où $w^{(2)}$ *est la différence finie d'ordre 2 définie comme dans l'expression (3.7),* $o(1)$ *ne dépend pas de* n *et tend vers* 0 *quand* $h = h(n) \to 0$ *et* $R_n(x;h)$ *est le reste comme en (3.9).*

Par conséquent, le nouvel estimateur $\widehat{f}_n$ en (3.20) peut être meilleur que l'estimateur ordinaire $\widetilde{f}_n$ en (3.1). Cette comparaison est faite au sens classique de l'approximation (des termes principaux) du risque quadratique intégré donné par

$$MISE(f_n^*) = \sum_{x\in\mathbb{N}} \text{var}\{f_n^*(x)\} + \sum_{x\in\mathbb{N}} [\text{biais}\{f_n^*(x)\}]^2$$

avec le même noyau (associé) discret $K_{x,h}$ et la même fenêtre h pour les deux estimateurs. Bien qu'il est facile de vérifier que la variance var$\{\widehat{f}_n(x)\}$ est égale à var$\{\widetilde{f}_n(x)\}$ donnée en (3.8), la différence provient du biais. En effet, selon le type de noyau discret (Exemples 3.2.1 et 3.2.2), l'influence des termes dans la comparaison vient de

$$f^{(1)} = (p_0 w)^{(1)} = p_0 w^{(1)} + p_0^{(1)} w \;\lesseqgtr\; p_0 w^{(1)} \tag{3.21}$$

et de

$$f^{(2)} = (p_0 w)^{(2)} = p_0 w^{(2)} + 2p_0^{(1)} w^{(1)} + p_0^{(2)} w \;\lesseqgtr\; p_0 w^{(2)}, \tag{3.22}$$

où le symbole $\lesseqgtr$ signifie $\leq$ ou $\geq$. Le sens des inégalités (3.21) et (3.22) dépendrait de la forme et, donc, des variations de la loi choisie comme loi de départ ; par exemple, selon qu'elle soit unimodale ou multimodale. En fait, les estimateurs semi-paramétriques se révèlent plus efficaces si la loi à estimer est dans un voisinage de la loi de départ. Pour l'approximation de second ordre dans le biais de $\widehat{f}_n$, les noyaux discrets standards asymétriques (comme le binomial) utilisent les deux inégalités (3.21) et (3.22). Tandis que les noyaux associés discrets symétriques (comme les triangulaires) n'ont besoin que de l'inégalité (3.22) comme pour les noyaux continus symétriques.

Selon Hjort & Glad (1995) pour le cas continu, ceci decrit un certain voisinage de la distribution de dénombrement autour du Poisson fixé p_0 où la méthode d'estimation proposée est meilleure que la traditionnelle non-paramétrique.

DÉMONSTRATION DE LA PROPOSITION 3.3.1 : À partir de (3.20), il suffit de calculer $\mathbb{E}\{\widetilde{w}_n(x)\}$ et var$\{\widetilde{w}_n(x)\}$ du fait que biais$\{\widehat{f}_n(x)\} = p_0(x)\mathbb{E}\{\widetilde{w}_n(x)\} - f(x)$ et

$\text{var}\{\widehat{f}_n(x)\} = p_0^2(x)\text{var}\{\widetilde{w}_n(x)\}$. Maintenant, à partir de (3.5)–(3.7), nous obtenons

$$\begin{aligned}
\mathbb{E}\{\widetilde{w}_n(x)\} &= \sum_{y\in\aleph_x} K_{x,h}(y)p_0^{-1}(y)f(y) \\
&= \sum_{y\in\aleph_x} f(y)p_0^{-1}(y)\Pr(\mathcal{K}_{x,h}=y) \\
&= \mathbb{E}\{f(\mathcal{K}_{x,h})p_0^{-1}(\mathcal{K}_{x,h})\} = \mathbb{E}\{w(\mathcal{K}_{x,h})\} \\
&= \left[w\{\mathbb{E}(\mathcal{K}_{x,h})\} + \frac{1}{2}\text{var}(\mathcal{K}_{x,h})w^{(2)}(x)\right]\{1+o(1)\}.
\end{aligned}$$

En procédant de la même manière qu'en (3.8), nous avons

$$\begin{aligned}
\text{var}\{\widetilde{w}_n(x)\} &= \frac{1}{n}\text{var}\left\{K_{x,h}(X_1)p_0^{-1}(X_1)\right\} \\
&= \frac{1}{n}\sum_{y\in\aleph_x} f(y)p_0^{-2}(y)\left\{\Pr(\mathcal{K}_{x,h}=y)\right\}^2 - \frac{1}{n}\left\{\sum_{z\in\aleph_x} f(z)p_0^{-1}(z)\Pr(\mathcal{K}_{x,h}=z)\right\}^2 \\
&= \frac{1}{n}p_0^{-2}(x)f(x)\left\{1-f(x)\right\}\left\{\Pr(\mathcal{K}_{x,h}=x)\right\}^2 + r_n(x;h),
\end{aligned}$$

avec

$$\begin{aligned}
r_n(x;h) &= \frac{1}{n}\left[\sum_{y\in\aleph_x\setminus\{x\}} f(y)p_0^{-2}(y)\left\{\Pr(\mathcal{K}_{x,h}=y)\right\}^2 + \left\{f(x)p_0^{-1}(x)\Pr(\mathcal{K}_{x,h}=x)\right\}^2\right] \\
&\quad -\frac{1}{n}\left\{\sum_{z\in\aleph_x} f(z)p_0^{-1}(z)\Pr(\mathcal{K}_{x,h}=z)\right\}^2 = p_0^{-2}(x)R_n(x;h).
\end{aligned}$$

De là, nous obtenons le résultat recherché. ■

3.3.2 Départ Poisson inconnu

Plus généralement, nous considérons notre estimateur semi-paramétrique $\widehat{f}_n$ présenté en (3.15)–(3.17) de f donnée en (3.14) tel que l'estimateur $\widehat{\mu}_n$ de μ est obtenu par la méthode du maximum de vraisemblance ; voir Hjort & Glad (1995) pour un estimateur général de μ. Il est bien connu que quand le modèle paramétrique est mal spécifié, l'estimateur $\widehat{\mu}_n$ du maximum de vraisemblance converge en probabilité vers la pseudo vraie valeur μ_0 qui minimise la distance de Kulleback-Leibler

$$\sum_{x\in\mathbb{N}} f(x)\log\frac{f(x)}{p(x;\mu)} =: d\left\{f(\cdot), p(\cdot;\mu)\right\}$$

de $p(x;\mu)$ à partir de la vraie fonction de masse $f(x)$; voir, par exemple, White (1982).

Ecrivons $p_0(x) = p(x;\mu_0)$ pour cette meilleure approximation paramétrique de μ par μ_0, mais ce p_0 ne s'exprime pas explicitement comme celui de la relation (3.20). Notons

$$u_0(x) = \frac{\partial \log p(x;\mu_0)}{\partial \mu} = \frac{x}{\mu_0} - 1$$

et

$$v_0(x) = \frac{\partial^2 \log p(x;\mu_0)}{\partial \mu^2} = \frac{-x}{\mu_0^2}.$$

Un développement de Taylor de second ordre permet d'obtenir

$$\begin{aligned}
\frac{p(x;\widehat{\mu}_n)}{p(X_i;\widehat{\mu}_n)} &= \exp\{\log p(x;\widehat{\mu}_n) - \log p(X_i;\widehat{\mu}_n)\} \\
&\doteq \frac{p_0(x)}{p_0(X_i)}\left[1 - \{u_0(X_i) - u_0(x)\}(\widehat{\mu}_n - \mu_0) + \frac{1}{2}\tau(x,X_i)(\widehat{\mu}_n - \mu_0)^2\right] \\
&= \frac{p_0(x)}{p_0(X_i)}\left[1 - \frac{X_i - x}{\mu_0}(\widehat{\mu}_n - \mu_0) + \frac{(x - X_i)(x - X_i - 1)}{2\mu_0^2}(\widehat{\mu}_n - \mu_0)^2\right],
\end{aligned}$$

avec $\tau(x,X_i) = v_0(x) - v_0(X_i) + \{u_0(x) - u_0(X_i)\}^2$.

De là, nous pouvons représenter l'estimateur proposé $\widehat{f}_n$ dans (3.16)–(3.17) comme

$$\begin{aligned}
\widehat{f}_n(x) &\doteq \frac{1}{n}\sum_{i=1}^{n} K_{x,h}(X_i)\,\frac{p_0(x)}{p_0(X_i)} \\
&\quad \times\left\{1 + \frac{x - X_i}{\mu_0}(\widehat{\mu}_n - \mu_0) + \frac{(x - X_i)(x - X_i - 1)}{2\mu_0^2}(\widehat{\mu}_n - \mu_0)^2\right\}.
\end{aligned}$$

La proposition suivante fournit des approximations du biais et de la variance pour l'estimateur $\widehat{f}_n$. Nous omettons la preuve de cette proposition parce qu'elle est analogue au résultat du cas continu de Hjort & Glad (1995) [Proposition 1] et à notre Proposition 3.3.1 montrée ci-dessus.

Proposition 3.3.2 *Soit $p_0(x) = p(x;\mu_0)$ la meilleure fonction de masse de Poisson approchant la fonction de masse inconnue f au sens de la distance de Kullback-Leibler. Soit $w = f/p_0$ la fonction poids de Poisson correspondante. Quand $n \to +\infty$ et $h = h(n) \to 0$, l'estimateur semi-paramétrique en (3.16) possède le biais et la variance suivants :*

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

et

$$\textit{var}\{\widehat{f}_n(x)\} = \frac{1}{n}f(x)\left\{1 - f(x)\right\}\left\{\Pr(\mathcal{K}_{x,h} = x)\right\}^2 + R_n(x;h),$$

où $w^{(2)}$ représente la différence finie de second ordre comme en (3.7), $o(1)$ ne dépend pas de n et tend vers 0 quand $h = h(n) \to 0$ et $R_n(x;h)$ est le reste comme en (3.9).

Similairement au cas de p_0 connu, l'estimateur proposé $\widehat{f}_n$ en (3.15)–(3.17) de f en (3.14) pourrait être encore plus approprié que l'estimateur traditionnel $\widetilde{f}_n$ en (3.1) à condition que le départ soit adapté.

Aussi, considérant les deux estimateurs $\widehat{f}_n$ et $\widetilde{f}_n$ avec le même noyau (associé) discret et la même fenêtre et suivant Hjort & Glad (1995), il y a un voisinage non-paramétrique de la fonction de masse autour de la famille de Poisson tel que si f se trouve dans ce voisinage alors $\widehat{f}_n$ est meilleur que $\widetilde{f}_n$. Pour le voir, nous écrivons $f = \exp(g)$ et $p_0 = \exp(q_0)$ puis nous examinons les différences finies (logarithmiques) de premier et second ordre

$$f^{(1)} = fg^{(1)} \text{ tandis que } p_0 w^{(1)} = f\left\{g^{(1)} - q_0^{(1)}\right\}$$

et

$$f^{(2)} = f\left\{g^{(2)} + (g^{(1)})^2\right\} \text{ tandis que } p_0 w^{(2)} = f\left\{g^{(2)} - q_0^{(2)} + (g^{(1)} - q_0^{(1)})^2\right\}.$$

Par conséquent, nous remplaçons les relations (3.21) et (3.22) par les deux inégalités $|g^{(1)} - q_0^{(1)}| \leqslant |g^{(1)}|$ et $|g^{(2)} - q_0^{(2)}| \leqslant |g^{(2)}|$, qui peuvent être exprimées de manière équivalente comme $0 \leqslant q_0^{(1)}/g^{(1)} \leqslant 2$ et $0 \leqslant q_0^{(2)}/g^{(2)} \leqslant 2$.

Finalement, un fait important pour une distribution de données de dénombrement est que l'effet de var$(\mathcal{K}_{x,h})$ dans cette expression du biais$\{\widehat{f}_n(x)\}$ est minimisé à l'intérieur d'une classe de noyaux discrets, telle que celle des noyaux discrets standards asymétriques contenant des noyaux discrets sur-, équi- ou sousdispersé. C'est parce que le départ Poisson p_0 est un coefficient multiplicatif de l'expression du biais dans les Propositions 3.3.1 et 3.3.2. Ceci se produit quand l'approximation de Poisson est bien spécifiée pour les données de dénombrement ; voir notre premier exemple illustratif presenté dans la Section 3.6. A cause de cela, l'estimateur proposé peut être encore bien meilleur en choisissant une fenêtre appropriée.

3.4 Choix de fenêtres

Dans cette section, nous examinons deux méthodes classiques de sélections de fenêtre pour l'estimateur à noyau discret traditionnel (3.1) utilisées pour le nouvel estimateur semi-paramétrique de (3.15)–(3.17) : les méthodes de validation croisée et d'excès de zéros. Nous montrons ici que la validation croisée est appplicable dans ce nouveau contexte tandis que la méthode de l'excès de zéros ne l'est pas. De plus, en pratique, une autre alternative pour le choix de fenêtre est la minimisation de l'erreur quadratique intégrée ISE^0 que l'on a redéfini dans la section 3.6.

3.4.1 Validation croisée

Pour l'estimateur $\widehat{f}_n$ défini en (3.16), la fenêtre optimale h est obtenue par la procédure populaire de validation croisée par

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

où

$$\begin{aligned} CV(h) &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{1}{p(X_i;\widehat{\mu}_n)p(X_j;\widehat{\mu}_n)}\sum_{x\in\mathbb{N}}p^2(x;\widehat{\mu}_n)K_{x,h}(X_i)K_{x,h}(X_j) \\ &\quad -\frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}K_{X_i,h}(X_j)\frac{p(X_i;\widehat{\mu}_{n,-i})}{p(X_j;\widehat{\mu}_{n,-i})} \end{aligned} \tag{3.23}$$

avec $\widehat{\mu}_{n,-i} = n^{-1}\sum_{j\neq i}X_j$.

En effet, pour déterminer la fenêtre optimale h, nous devons miminiser (par rapport à h) un estimateur de $MISE$ qui peut être écrit comme

$$MISE(h) = \mathbb{E}\left\{\sum_{x\in\mathbb{N}}\widehat{f}_n^2(x)\right\} - 2\mathbb{E}\left\{\sum_{x\in\mathbb{N}}\widehat{f}_n(x)f(x)\right\} + \mathbb{E}\left\{\sum_{x\in\mathbb{N}}f^2(x)\right\}.$$

Comme le dernier terme $\sum_{x\in\mathbb{N}}f^2(x)$ ne dépend pas de h, nous minimisons uniquement un estimateur de l'expression

$$\mathbb{E}\left\{\sum_{x\in\mathbb{N}}\widehat{f}_n^2(x)\right\} - 2\mathbb{E}\left\{\sum_{x\in\mathbb{N}}\widehat{f}_n(x)f(x)\right\} \tag{3.24}$$

qui dépend de la fonction de masse inconnue f. Rappelons ici que h intervient bien dans l'expression de $\widehat{f}_n$. Un estimateur sans biais du premier terme $\mathbb{E}\{\sum_{x\in\mathbb{N}}\widehat{f}_n^2(x)\}$ de (3.24) est

$$\widehat{H}_n = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{1}{p(X_i;\widehat{\mu}_n)p(X_j;\widehat{\mu}_n)}\sum_{x\in\mathbb{N}}p^2(x;\widehat{\mu}_n)K_{x,h}(X_i)K_{x,h}(X_j).$$

En fait, puisque les variables aléatoires $X_1,\ldots,X_n$ sont i.i.d., nous avons directement

$$\begin{aligned} \sum_{x\in\mathbb{N}}\widehat{f}_n^2(x) &= \sum_{x\in\mathbb{N}}\left\{p(x;\widehat{\mu}_n)\frac{1}{n}\sum_{i=1}^{n}\frac{K_{x,h}(X_i)}{p(X_i;\widehat{\mu}_n)}\right\}^2 \\ &= \frac{1}{n^2}\sum_{x\in\mathbb{N}}p^2(x;\widehat{\mu}_n)\left\{\sum_{i=1}^{n}\frac{K_{x,h}(X_i)}{p(X_i;\widehat{\mu}_n)}\right\}^2 \\ &= \widehat{H}_n. \end{aligned}$$

Puis, le second terme $\mathbb{E}\{\sum_{x\in\mathbb{N}} \widehat{f}_n(x)f(x)\}$ de (3.24) est estimé par

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

où $\widehat{f}_{n,-i}$ et $\widehat{\mu}_{n,-i}$ sont calculés comme $\widehat{f}_n$ et $\widehat{\mu}_n$, respectivement, en excluant X_i. Comme précédemment, nous avons d'abord

$$\begin{aligned}\mathbb{E}(\widehat{G}_n) &= \mathbb{E}\left\{\frac{1}{n(n-1)} \sum_{i=1}^{n}\sum_{j\neq i} K_{X_i,h}(X_j)\frac{p(X_i;\widehat{\mu}_{n,-i})}{p(X_j;\widehat{\mu}_{n,-i})}\right\} \\ &= \mathbb{E}\left\{\frac{1}{n-1} \sum_{j\neq 1} K_{X_1,h}(X_j)\frac{p(X_1;\widehat{\mu}_{n,-1})}{p(X_j;\widehat{\mu}_{n,-1})}\right\} \\ &= \mathbb{E}\left\{K_{X_1,h}(X_2)\frac{p(X_1,\widehat{\mu}_{n,-1})}{p(X_2;\widehat{\mu}_{n,-1})}\right\}\end{aligned}$$

et, ensuite,

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

L'estimateur ci-dessus $\widehat{G}_n$ de $\mathbb{E}\{\sum_{x\in\mathbb{N}} \widehat{f}_n(x)f(x)\}$ n'est pas nécessairement sans biais car $p(X_1,\widehat{\mu}_{n,-1})/p(X_2;\widehat{\mu}_{n,-1})$ et $p(X_1,\widehat{\mu}_n)/p(X_2;\widehat{\mu}_n)$ n'ont pas la même distribution. Pour cela, nous notons que les estimateurs $\widehat{\mu}_{n,-i}$ et $\widehat{\mu}_n$ de μ sont tous les deux convergents et non biaisés, mais $\text{var}(\widehat{\mu}_{n,-i}) = (n-1)^{-1}\mu = \{n/(n-1)\}\,\text{var}(\widehat{\mu}_n) \neq \text{var}(\widehat{\mu}_n)$. Ainsi, il suit que $CV(h)$ défini en (3.23) est un estimateur asymptotiquement sans biais de l'expression en (3.24).

3.4.2 Excès de zéros

Considérons la situation particulière de proportion importante de zéros pour des données de dénombrement. Pour l'estimateur non-paramétrique à noyau discret (3.1), l'équation (3.11) de proportion de zéros peut être utilisée pour trouver une fenêtre adaptée avec les noyaux discrets standards mais pas avec les noyaux associés discrets

triangulaires. Tous les effets d'excès de zéros sont pris en compte par la partie poissonnienne de l'estimateur semi-paramétrique présenté en (3.15)–(3.17).

En effet, nous montrons ici que la seule solution h_0 de l'équation qui correspond à la proportion de zéros dans le cas semi-paramétrique est $h_0 = 0$ pour le noyau binomial. Des résultats similaires sont obtenus aussi bien pour les noyaux Poisson que binomial négatif. Pour les noyaux associés discrets triangulaires, il n'y a pas de solution.

À partir de (3.17), nous obtenons

$$\mathbb{E}\{\widehat{f}_n(x)\} = \int_{z\geq 0} \sum_{y\in\mathbb{N}} \Pr(\mathcal{K}_{x,h} = y) f(y) \frac{y!(z)^{x-y}}{x!} \varphi(z) dz, \tag{3.25}$$

où φ_n est la densité de probabilité de $\widehat{\mu}_n$ inclus dans $[0, +\infty[$. Dans le cas semi-paramétrique, la proportion de zéros peut être obtenue en identifiant le nombre de zéros théoriques en prenant $y = z = 0$ dans l'expression (3.25) et $f(0) = \varphi(0) = 1$ avec le nombre de zéros empiriques $n_0 = \sharp\,(X_i = 0)$ d'observations égales à zéro dans l'échantillon. Nous obtenons

$$\begin{aligned} n_0 &= \sum_{i=1}^{n} \Pr(\mathcal{K}_{X_i,h_0} = 0) \frac{0!0^{X_i}}{X_i!} \\ &= \sum_{i=1|X_i=0}^{n} \Pr(\mathcal{K}_{X_i,h_0} = 0) 0^0 + \sum_{i=1|X_i\neq 0}^{n} \Pr(\mathcal{K}_{X_i,h_0} = 0) \frac{0^{X_i}}{X_i!} \\ &= \sum_{i=1|X_i=0}^{n} \Pr(\mathcal{K}_{X_i,h_0} = 0), \end{aligned} \tag{3.26}$$

avec $0^0 = 1 = 0!$ et $0^{X_i} = 0$ pour $X_i \neq 0$. Par conséquent, il est facile de montrer que l'équation (3.26) pour le noyau binomial (voir Exemple 3.2.1) s'écrit :

$$\sum_{i=1|X_i=0}^{n} \frac{(X_i+1)!}{0!(X_i+1)!} \left(\frac{X_i+h_0}{X_i+1}\right)^0 \left(\frac{1-h_0}{X_i+1}\right)^{X_i+1} = n_0$$

qui se réduit simplement à $n_0(1-h_0) = n_0$ et, ainsi, $h_0 = 0$.

3.5 Restriction à un support fini

Dans cette section, nous supposons que la distribution ou la variable aléatoire X de dénombrement de fonction de masse de probabilité inconnue $f(x) = \Pr(X = x)$ a pour support $\{0, 1, \ldots, N\}$ avec $N \in \mathbb{N} \setminus \{0\}$ fixé. De là, dans la même approche que pour les lois de Poisson pondérées vues plus haut en (3.14), nous écrivons f comme

la *distribution binomiale pondérée* (voir, par exemple, Johnson *et al.*, 2005 [pp. 149–150]) :

$$f(x) = w(x;q)\, b(x;q) =: f_w(x;q), \ \ \forall x \in \{0,1,\ldots,N\}, \tag{3.27}$$

où $b(x;q) := \frac{N!}{x!(N-x)!} q^x (1-q)^{N-x}$ est la fonction de masse d'une binomiale de probabilité de succés q, de moyenne $Nq > 0$, $q \in [0,1]$, et $x \mapsto w(x;q) := \omega(x) \times \left\{\sum_{x\in\{0,1,\ldots,N\}} \omega(x) b(x;q)\right\}^{-1}$ est la fonction poids normalisée de binomiale.

Soit $X_1, \ldots, X_n$ des variables aléatoires i.i.d. de fonction de masse de probabilité inconnue f sur $\{0,1,\ldots,N\}$ présentée en (3.27). L'estimateur semi-paramétrique de f est défini différemment par :

$$\begin{aligned}
\widehat{f}_n(x) &= b(x;\widehat{q}_n)\, \widetilde{w}_n(x;\widehat{q}_n), \quad x \in \{0,1,\ldots,N\} \\
&= b(x;\widehat{q}_n)\, \frac{1}{n}\sum_{i=1}^{n} \frac{K_{x,h}(X_i)}{b(X_i;\widehat{q}_n)} \\
&= \frac{1}{n}\sum_{i=1}^{n} K_{x,h}(X_i)\, \frac{b(x;\widehat{q}_n)}{b(X_i;\widehat{q}_n)} \\
&= \frac{1}{n}\sum_{i=1}^{n} K_{x,h}(X_i) \frac{X_i!(N-X_i)!}{x!(N-x)!} \widehat{q}_n^{x-X_i}(1-\widehat{q}_n)^{X_i-x},
\end{aligned} \tag{3.28}$$

où $\widehat{q}_n = N^{-1}\overline{X}_n = (nN)^{-1}(X_1+\cdots+X_n)$ est la proportion de succés de l'échantillon, $h > 0$ est le paramêtre de lissage discret, et $K_{x,h}$ est un noyau associé discret donné.

Comme dans les cas précédents, cet estimateur $\widehat{f}_n$ de f est défini à constante de normalisation près $\widehat{C} = \sum_{x\in\{0,1,\ldots,N\}} \widehat{f}_n(x)$. Toutes les propriétés vues plus tôt dans les Sections 3.3 et 3.4 sont encore valables pour cet estimateur $\widehat{f}_n$ à support fini $\{0,1,\ldots,N\}$. Cependant, nous devons aussi prendre en compte le biais de bordure à droite de $\{0,1,\ldots,N\}$. Pour cela, nous ne pouvons utiliser que deux types de noyaux discrets existant : binomial et triangulaire, ayant aussi des supports finis. Le noyau binomial modifié est le même que celui présenté en Exemple 3.2.1, si ce n'est qu'à la dernière cible $x = N$ où nous devons considérer le noyau binomial $\mathcal{B}\{N,(N+h)/(N+1)\}$ avec $h \in]0,1]$. Toutefois, ce noyau binomial modifié est instable pour lisser de manière discrète une distribution de probabilité discrète sur un support compact puisqu'il prend en compte toute les informations à gauche de la dernière cible. Pour les noyaux associés discrets triangulaires modifiés, nous appliquons aussi la modification du bras utilisée en (3.13) à droite du support $\{0,1,\ldots,N\}$ (cela se fait au voisinage du point $x = N$) comme suit : pour $k \in \{1,2,\ldots,N\}$ donné et $x \in \{0,1,\ldots,N\}$,

$$a_N = k \Longleftrightarrow a_N = \begin{cases} j & \text{si } x = N-j,\ j \in \{0,1,\ldots,k-1\} \\ k & \text{si } x \in \{k,k+1,\ldots,N-k\}. \end{cases} \tag{3.29}$$

Les modifications simultanées à gauche (3.13) et à droite (3.29) des noyaux associés discrets triangulaires sont plus appropriées pour des distributions de probabilités dis-

crètes à support fini. Elles sont aussi utilisables pour des données catégorielles ordonnées.

3.6 Illustrations

Pour évaluer la perfomance d'un estimateur semi-paramétrique ou non-paramétrique f_n^* de f, nous utilisons simplement le critère empirique de l'erreur quadratique intégrée défini par

$$ISE^0 = \sum_{x \in \mathbb{N}} \{f_n^*(x) - f_0(x)\}^2,$$

où f_0 est l'estimateur fréquence empirique. Ce critère peut être directement observé à travers les représentations graphiques ; voir, par exemple, Marron & Padgett (1987).

Dans la situation particulière des données de dénombrement, nous pouvons aussi mesurer la performance en utilisant la distance du khi-deux :

$$\chi_0^2 = \sum_{x' \in \{0,1,\ldots,N_0\}} \frac{n\{f_n^*(x') - f_0(x')\}^2}{f_n^*(x')},$$

où $N_0 + 1$ est le nombre de classes valides au sens du test du khi-deux ; voir, par exemple, Greenwood & Nikulin (1996). Ainsi, la statistique χ_0^2 peut être convenablement approchée par la distribution du χ^2 avec $N_0 - r$ degrés de liberté (ddl), où r est le nombre de paramètres estimés (h, μ ou q) dans f_n^*. C'est une mesure d'adéquation bien connue et qui a pour avantage de permettre une comparaison directe avec un modèle paramétrique utilisant aussi le test d'ajustement du khi-deux ; voir les deux sections suivantes.

3.6.1 Données de buts

Reconsidérons les données présentées dans Table 3.2, déjà utilisées dans les chapitres précédents. L'étude des statistiques élémentaires révèle que ces données de dénombrement sont légérement surdispersées ($D = 0.113$) et en excès de zéros ($Z = 0.059$) par rapport à la distribution de Poisson de moyenne estimée à $2.13421 = \widehat{\mu}_n$. Bien que la loi de Poisson puisse ajuster ces données ($\chi^2 = 8.13326$ avec 6 degrés de liberté correspondant à la p-valeur égale à 0.2285), nous montrons ici que l'approche semi-paramétrique proposée procure de meilleurs résultats que les approches paramétriques et non-paramétriques pour certains noyaux associés discrets. Notons que, pour un nombre donné de buts $g \in \mathbb{N}$, l'entier le plus proche de $n \times f_n^*(g)$ est l'estimation semi-paramétrique/non-paramétrique du nombre correspondant de matchs. Cependant, on peut toujours lisser modérément en choisissant une autre fenêtre de lissage discret que la fenêtre optimale.

Table 3.3 présente les résultats numériques de la comparaison entre l'estimation non-paramétrique traditionnelle (3.1) et l'estimation semi-paramétrique (3.17) basée

Buts (g)	0	1	2	3	4	5	6	7	8	9	Total
Matches	51	90	109	61	44	12	9	3	0	1	380

TAB. 3.2 – Données du nombre de buts par match des championnats de football de Ligue 1 française pour la saison 2005-2006 avec $n = 380$ rencontres (cf. Table 2.3)

sur un départ Poisson avec différents noyaux associés discrets. Bien que les résultats par les méthodes semi-paramétrique et non-paramétrique semblent similaires en utilisant le noyau binomial, la différence devient plus visible entre les deux méthodes semi-paramétrique et non-paramétrique dans le cas des noyaux de Poisson, binomial négatif et triangulaires discrets modifiés $a_0 \in \{1, 2\}$. Pour ces données de dénombrement, l'estimation semi-paramétrique (3.17) procure une bonne amélioration par rapport à l'estimation non-paramétrique. De plus, nous n'observons aucune différence entre les estimations semi-paramétriques par les noyaux discrets standards asymétriques qui sont sur-, équi- ou sousdispersés. On peut dire que l'effet de la dispersion des noyaux discrets est pris en compte par la partie paramétrique dans la méthode semi-paramétrique. Ainsi, pour les lissages semi-paramétriques, le choix entre les noyaux discrets triangulaires (modifiés) et asymétriques est nettement en faveur des noyaux associés discrets symétriques (modifiés) triangulaires. Toutes ces conclusions sont confirmées à travers les deux critères ISE^0 et χ^2_0 calculés en utilisant la fenêtre optimale de lissage discret obtenu par validation croisée.

3.6.2 Consommation journalière d'alcool

Une expérience a été menée dans le contexte sociologique sur le nombre de jours dans une semaine pendant lesquels de l'alcool ait été consommé ; ceci est d'après les travaux de Alanko & Lemmens (1996). Un échantillon de taille $n = 399$ personnes choisies de façon aléatoire a été intérrogé sur deux semaines consécutives pendant lesquels ils ont noté leur consommation journalière d'alcool. Pour ajuster ces données présentées dans Table 3.4, Alanko & Lemmens (1996) ont utilisé une loi bêta-binomiale qui est un mélange d'une loi binomiale avec une probabilité de réussite dans un essai individuel suivant une loi bêta ; voir, par exemple, Johnson *et al.* (2005) [pages 253–256]. Les résultats obtenus par ces auteurs n'ont pas été très satisfaisants ; en effet, les p-valeurs obtenues par le test d'ajustement du khi-deux sont, respectivement, de 0.086 et de 0.082 pour la semaine 1 et pour la semaine 2. Ainsi, la loi bêta-binomiale et *a fortiori* la loi binomiale ne semblent pas être des modèles convenables pour ces données de dénombrement. De plus, nous observons que les données ont deux valeurs

Noyau discret :	Triang $a_0 = 1$	Triang $a_0 = 2$	Binomial	Poisson	Bin. nég.
Semi-param.					
h_{cv}	0.001	0.001	0.001	0.540	0.820
$\widehat{C}$	1.00004	1.00007	1.01650	1.07392	1.12218
ISE^0	$1.65\ 10^{-8}$	$3.50\ 10^{-8}$	0.00250	0.00301	0.00305
χ^2_0 avec 5 ddl	0.0036	0.0062	10.29	13.99	15.81
p-valeur	1.0	1.0	0.0674	0.0157	0.0074
Non-param.					
h_{cv}	0.028	0.019	0.177	0.054	0.039
$\widetilde{C}$	0.99810	0.99780	0.95872	1.05082	1.12028
ISE^0	0.00001	0.00004	0.00395	0.01580	0.02904
χ^2_0avec 6 ddl	0.1396	0.2816	11.34	63.33	116.80
p-valeur	0.999946	0.999581	0.0784	$9.45\ 10^{-12}$	0.0

TAB. 3.3 – Résultats comparatifs des estimations semi-paramétriques et non-paramétriques basées sur les données de Table 3.2

modales : la première se trouve pour le nombre de jour 1 de semaine 1 ou 2 jours de semaine 2 et la seconde se situe pour 7 jours pour les deux semaines. En s'appuyant sur ces faits, il convient d'ajuster ces données par la procédure semi-paramétrique proposée (3.28) avec un départ binomial.

Tables 3.5 et 3.6 présentent les résultats numériques des ajustements obtenus par l'intermédiaire de l'estimateur semi-paramétrique (3.28) avec un départ binomial. Pour ces deux jeux de données, uniquement les noyaux associés discrets triangulaires modifiés ($a_0 = a_N \in \{1, 2\}$) ont été utilisés avec leur fenêtre optimal ($h_{cv} = 0.001$). Dans le sens du khi-deux, les ajustements sont clairement meilleurs en comparaison des résultats obtenus par le modèle bêta-binomiale dans la Table 3.4. Pour des bras modifiés fixés $a_0 = a_N \in \{1, 2, \ldots, 7\}$ du noyau associé discret triangulaire, nous pouvons choisir une fenêtre $h(\neq h_{cv})$ pour obtenir de nouveaux lissages (ou ajustements) discrets ayant une meilleur p-valeur que celle de la loi bêta-binomiale. Pour une fenêtre $h > 0$ fixée dans le noyau discret triangulaire modifié, il est facile d'observer que si $a_0 = a_N$ augmente alors simultanément ISE^0 et χ^2_0 augmentent aussi. Ainsi, les p-valeurs associées diminuent.

Nombre de jours par semaine	Fréquences observées (semaine 1)	Fréquences bêta-binomial attendues (semaine 1)	Fréquences observées (semaine 2)	Fréquences bêta-binomial attendues (semaine 2)
0	47	54.6	42	47.9
1	54	42.0	47	42.9
2	43	38.9	54	41.9
3	40	38.5	40	42.5
4	40	40.1	49	44.3
5	41	44.0	40	47.8
6	39	53.1	43	54.9
7	95	87.8	84	76.7
Total	399	399.0	399	399.0
ISE^0		$3.02\ 10^{-3}$		$3.03\ 10^{-3}$
χ^2		9.6		9.7
ddl		5		5
p-valeur		0.086		0.082

TAB. 3.4 – Nombre de jours de consommation d'alcool pour les semaines 1 et 2, Alanko & Lemmens (1996).

Nombre de jours par semaine	Fréquences observées (semaine 1)	Fréquences attendues SP par (3.28) Triang $a_0 = 1 = a_N$	Fréquences attendues SP par (3.28) Triang $a_0 = 2 = a_N$
0	47	46.87	46.63
1	54	54.13	54.02
2	43	43.03	43.47
3	40	39.97	39.99
4	40	39.98	39.89
5	41	40.98	41.39
6	39	39.29	39.35
7	95	94.75	94.26
Total	399	399.00	399.00
$\widehat{C}$		1.00131	1.00487
ISE^0		$1.15\ 10^{-6}$	$7.29\ 10^{-6}$
χ_0^2		0.0035	0.0204
ddl		5	5
p-valeur		1.0	0.999997

TAB. 3.5 – Estimations semi-paramétriques par (3.28) en utilisant les noyaux associés discrets triangulaires modifiés avec $h_{cv} = 0.001$ pour le nombre de jour de consommation d'alcool pendant la semaine 1 dans Table 3.4

Nombre de jours par semaine	Fréquences observées (semaine 2)	Fréquences attendues SP par (3.28) Triang $a_0 = 1 = a_N$	Fréquences attendues SP par (3.28) Triang $a_0 = 2 = a_N$
0	42	41.90	41.70
1	47	47.13	47.05
2	54	53.99	54.32
3	40	40.00	40.02
4	49	48.94	48.87
5	40	40.00	40.38
6	43	43.24	43.26
7	84	83.80	83.40
Total	399	399.00	399.00
$\widehat{C}$		1.00117	1.00431
ISE^0		$8.41\ 10^{-7}$	$4.29\ 10^{-6}$
χ^2_0		0.0026	0.0139
ddl		5	5
p-valeur		1.0	0.999999

TAB. 3.6 – Suite et fin de Table 3.5 pour la semaine 2

3.6.3 Données simulées

Nous présentons dans la Table 3.7 un jeu de données de dénombrement simulées de taille $n = 300$ d'un mélange de Poisson $f = 0.4\mathcal{P}(0.5) + 0.6\mathcal{P}(10)$; voir la Section 1.6 du Chapitre 1.

x_i	0	1	2	3	4	5	6	7	8	9	10
n_i	76	38	4	3	2	13	16	15	14	20	22
x_i	11	12	13	14	15	16	17	18	19	20	
n_i	18	17	11	12	8	6	3	0	1	1	

TAB. 3.7 – Données simulées de $f = 0.4\mathcal{P}(0.5) + 0.6\mathcal{P}(10)$ avec $n = 300$.

Les résultats numériques obtenus au moyen de l'estimateur semi-paramétrique avec un départ Poisson (3.17) sont présentés dans la Table 3.8. Les estimations non-paramétriques donnent même de meilleurs résultats avec les noyaux discrets standards et des résultats parfois comparables à ceux des estimations semi-paramétriques pour les noyaux associés discrets triangulaires. Ceci peut s'expliquer par le fait que les données proviennent d'un mélange de Poisson. Par conséquent, il faut envisager une estimation semi-paramétrique ayant un départ qui soit aussi un mélange de Poisson. Dans cette situation, l'estimation non-paramétrique a l'avantage de laisser parler directement les données sans *a priori* sur la distribution de départ.

Noyau :	Triang $a_0 = 1$	Triang $a_0 = 2$	Binomial	Poisson	Bin. nég.
Semi-par.					
h_{cv}	0.001	0.001	0.470	0.680	0.559
$\widehat{C}$	1.00128	1.00405	1.17831	4.36825	9.13831
ISE^0	7.70×10^{-7}	6.97×10^{-6}	0.02058	0.06977	0.08882
χ^2_0 (ddl)	0.021 (12)	1.18 (12)	61.29 (13)	573.85 (13)	1105.75 (12)
p-value	1.0	0.9999646	3.08×10^{-8}	0	0
Non-par.					
h_{cv}	0.001	0.001	0.090	0.190	0.300
C	1.00009	1.00028	1.01354	1.07240	1.12089
ISE^0	9.13×10^{-9}	7.24×10^{-8}	0.00158	0.00811	0.01502
χ^2_0 (ddl)	0.001 (13)	0.00186 (13)	10.91 (15)	51.31 (18)	84.02 (18)
p-value	1	1	0.7589478	4.77×10^{-5}	1.68×10^{-10}

TAB. 3.8 – Résultats comparatif des estimations semi-paramétriques et non-paramétriques sur des données simulées de $f = 0.4\mathcal{P}(0.5) + 0.6\mathcal{P}(10)$ avec $n = 300$

3.7 Modèles de diagnostique

L'estimation de la fonction poids (3.18) procure des informations utiles pour les modèles de diagnostique. La fonction poids binomiale ou de Poisson doit être égale à 1 si la loi paramétrique de départ binomial ou de Poisson correspond à la vraie distribution discrète de probabilité. Hjort & Glad (1995) [Section 8.2] proposent de montrer l'adéquation du modèle en examinant le graphe de la fonction poids ω avec une bande de confiance ponctuelle pour vérifier si $\omega(x) = 1$ est acceptable ou non pour plusieurs modèles potentiels. Ce graphe permet de désigner clairement où l'ajustement est localement le plus mauvais. Pour l'estimation de la fonction poids (3.18), le biais et la variance peuvent être directement déduit de la Proposition 3.3.2 par

$$\begin{aligned}\mathbb{E}\{\widetilde{w}_n(x)\} &\doteq w\{\mathbb{E}(\mathcal{K}_{x,h})\} + \frac{1}{2}\text{var}(\mathcal{K}_{x,h})w^{(2)}(x) \\ &\quad -\frac{1}{n}\omega(x)u_0(x)\{1 + u_0(x)/J\}\end{aligned}$$

et

$$\text{var}\{\widetilde{w}_n(x)\} \doteq \frac{1}{n}\frac{\omega(x)}{p_0(x)}\{\Pr(\mathcal{K}_{x,h} = x)\}^2 - \frac{1}{n}\omega^2(x)\{\Pr(\mathcal{K}_{x,h} = x)\}^2\{1 + u_0^2(x)/J\},$$

avec $J = -1/\mathbb{E}\{v_0(X_i)\} = \mu/\mu_0^2$ et $u_0(x) = x/\mu_0 - 1$.

Une autre possibilité, qui a été aussi discutée par Hjort & Glad (1995) [Section 8.2], est de tracer la fonction poids logarithme

$$\log \widetilde{w}_n(x; \widehat{\mu}_n) = \log\{\widehat{f}_n(x)/p(x; \widehat{\mu}_n)\}$$

afin de voir sa distance par rapport à zéro. Un simple test graphique d'ajustement se construit en représentant $(x, Z(x))$ avec

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

Quand le départ poissonnien est en effet la vraie loi de Poisson, la statistique $Z(x)$ est approximativement distribuée comme une loi normale centrée et réduite pour chaque cible x (voir Hjort & Glad, 1995, pour le cas continu). Cela signifie que près de 95% des valeurs de $Z(x)$ devraient être contenues dans la bande ± 1.96. Ainsi, notons que le comportement attendu de la fonction $x \mapsto Z(x)$ se résume à un ensemble approprié $\{x \in \mathbb{N}; p(x; \widehat{\mu}_n) \geq \alpha > 0\}$ du support $\mathbb{N}$; cela est du à de très petites valeurs des probabilités individuelles de Poisson $p(x; \widehat{\mu}_n)$.

Par rapport au départ binomial en (3.28), une expression similaire de $Z(x)$ est

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

x	0	1	2	3	4	5	6	7	8	9	$Z(x) \in [\pm 1.96]$
Tr.$a_0 = 1$	0.92	-0.58	0.68	-1.46	0.85	-1.19	1.23	1.06	-2.70	2.21	**80%**
Tr.$a_0 = 2$	0.92	-0.58	0.68	-1.45	0.85	-1.19	1.23	1.06	-2.34	2.21	**80%**
Binom.	0.81	0.06	-0.65	-0.38	-0.78	0.49	1.33	0.21	1.81	1.60	**100%**
Poisson	-0.04	-1.04	-1.31	-0.15	1.65	2.64	2.44	1.70	1.07	0.85	**80%**
Bin.nég.	-0.43	-1.28	-0.73	0.76	1.86	2.01	1.54	0.97	0.60	0.56	**90%**

TAB. 3.9 – Les valeurs de $Z(x)$ associées aux résultats dans Table 3.3

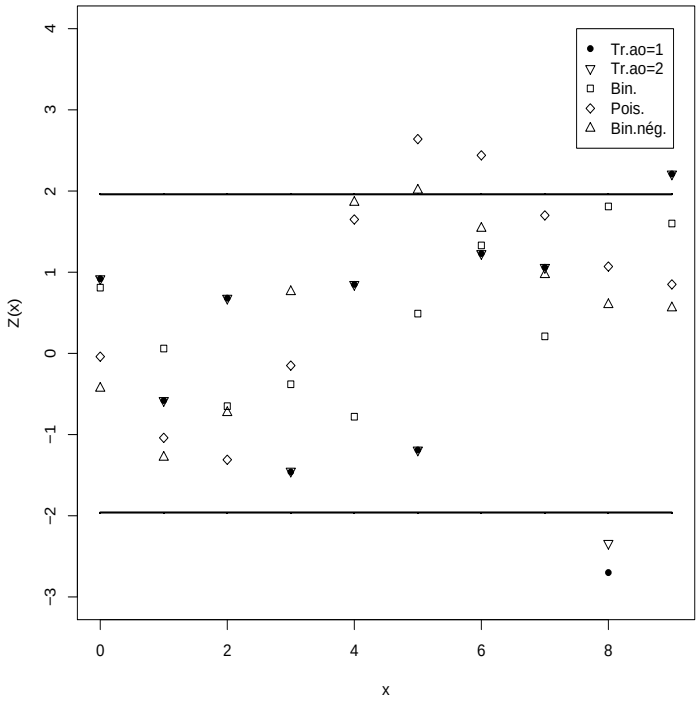

FIG. 3.2 – Valeurs de $Z(x)$ associées aux résultats dans Table 3.9

x	0	1	2	3	4	5	6	7	$Z(x) \in [\pm 1.96]$
Tr.$a_0 = 1 = a_N$ (sem. 1)	4.67	5.25	-0.73	-8.64	-11.45	-6.48	1.01	6.93	**25%**
Tr.$a_0 = 2 = a_N$ (sem. 1)	4.66	5.24	-0.66	-8.63	-11.47	-6.39	1.02	6.92	**25%**
Tr.$a_0 = 1 = a_N$ (sem. 2)	4.55	4.70	0.72	-8.75	-9.24	-6.55	1.68	6.61	**25%**
Tr.$a_0 = 2 = a_N$ (sem. 2)	4.54	4.70	0.78	-8.74	-9.26	-6.46	1.68	6.60	**25%**

TAB. 3.10 – Valeurs de $Z(x)$ associées aux résultats dans Tables 3.5 et 3.6

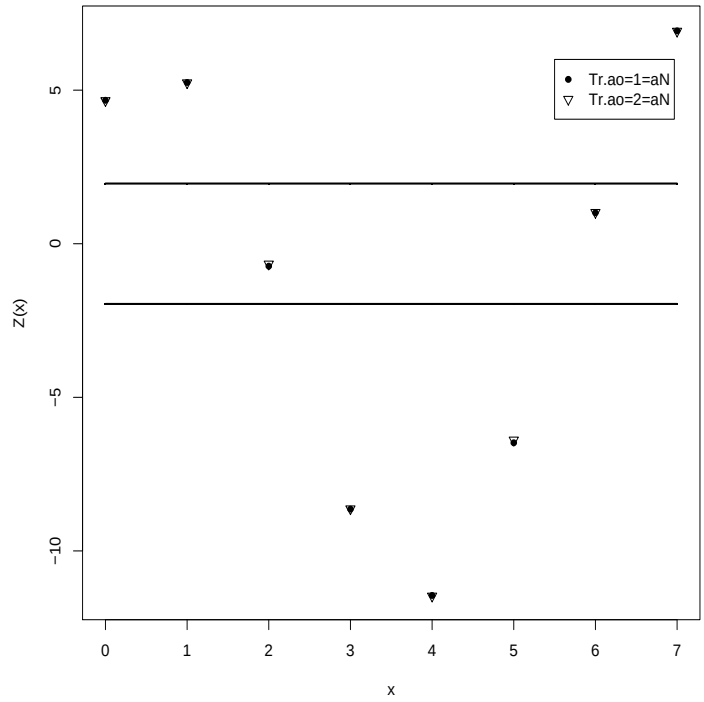

FIG. 3.3 – Valeurs de $Z(x)$ associées aux résultats de la semaine 1 dans Table 3.10

Elle sera nécessaire dans le cas où on examine la fonction $x \mapsto Z(x)$ sur le support $\{0, 1, \ldots, N\}$.

De manière concrête, en faisant premièrement une application aux différents noyaux discrets de la Table 3.3 sur l'ensemble $\{0, 1, \ldots, 9\}$ de $\mathbb{N}$ avec $0.9998 = \sum_{x\in\{0,1,\ldots,9\}} p(x; \widehat{\mu}_n)$, au moins 80% des valeurs de $Z(x)$ restent à l'intérieur de l'intervalle ± 1.96 (voir Table 3.9 et Figure 3.2). Cela peut suggèrer aussi qu'il est intéressant de considérer une distribution paramétrique pure de Poisson pour modéliser ces données plutôt que l'estimation semi-paramétrique avec le départ poissonnien associé, en particulier, aux noyaux discrets binomial et binomial négatif. De là, nous validons ici le choix d'un modèle paramétrique de Poisson pour les données de dénombrement de la Table 3.2 avec une p-valeur de 22.85%.

Cependant, pour le deuxième exemple (voir les Tables 3.5 et 3.6), seulement 25% des valeurs de $Z(x)$ associés avec l'estimation semi-paramétrique à départ binomial se situent dans la bande de confiance ± 1.96 (voir Table 3.10 et Figure 3.3). Par conséquent, il est nécessaire dans ce cas d'utiliser notre méthode d'estimation semi-paramétrique pour les données de dénombrement de la Table 3.4 au lieu des modèles paramétriques binomial et bêta-binomial (généralisé) ; voir, par exemple, Rodríguez-Avi *et al.* (2007).

x	0	1	2	3	4	5	6	7
Tr.$a_0 = 1$	4.41	4.76	−3.47	−10.18	−17.87	−8.64	−7.57	−6.74
Tr.$a_0 = 2$	4.40	4.75	−2.99	−10.03	−17.85	−8.66	−7.59	−6.75
Binomial	4.86	7.60	2.03	−13.97	−18.70	−16.47	−15.06	−12.50
Poisson	3.44	6.41	6.16	−0.15	−11.91	−23.12	−24.79	−16.78
Bin. nég.	2.66	5.80	6.79	2.98	−5.41	−13.98	−16.20	−11.06

TAB. 3.11 – Valeurs de $Z(x)$ associées aux résultats dans Table 3.7 (1/3)

Quant aux données de dénombrement émanant du mélange de Poisson de la Table 3.7 sur l'ensemble $\{0, 1, \ldots, 20\}$ de $\mathbb{N}$, nous n'avons que 4.16% des valeurs de $Z(x)$ dans la bande ± 1.96 lorsque l'on utilise les noyaux discrets standards et 8.33% des $Z(x)$ avec les noyaux associés discrets triangulaires $a_0 \in \{1, 2\}$ (voir Tables 3.11, 3.12 et 3.13 ainsi que Figure 3.4). Le modèle paramétrique de Poisson n'est pas adapté du fait que ces données proviennent d'un mélange de distributions de Poisson. Il peut être plus approprié d'envisager une méthode d'estimation paramétrique avec un mélange de modèles (Böhning, 2000) ou un paramètre appartenant à $\mathbb{R}^d$ (par exemple, $d = 2$ dans le cas de nos données). Toutefois, même si le départ paramétrique n'est pas bien choisi, la partie non-paramétrique apporte une correction. Ici, il serait bien de retenir la méthode d'estimation purement non-paramétrique (3.1).

x	8	9	10	11	12	13	14	15	16	17
Tr.$a_0 = 1$	-4.84	-0.51	1.80	2.48	3.13	2.77	2.98	2.69	2.74	3.14
Tr.$a_0 = 2$	-4.85	-0.52	1.80	2.48	3.12	2.77	2.98	2.69	2.74	3.14
Binomial	-5.45	0.43	3.23	5.18	5.17	5.54	4.78	4.02	3.23	2.43
Poisson	-6.10	2.69	8.10	10.29	10.1	8.72	6.83	5.03	3.63	2.73
Bin. nég	-3.37	2.98	6.72	8.05	7.73	6.54	5.08	3.75	2.72	2.09

TAB. 3.12 – Suite de Table 3.11 (2/3)

x	18	19	20	21	22	23	$Z(x) \in [\pm 1.96]$
Tr.$a_0 = 1$	3.54	7.01	12.16	21.86	-Inf	-Inf	**8.33%**
Tr.$a_0 = 2$	3.69	7.01	12.16	21.88	41.20	-Inf	**8.33%**
Binomial	3.19	4.51	7.09	12.50	23.30	44.78	**4.16%**
Poisson	2.39	2.71	3.93	6.63	12.06	22.83	**4.16%**
Bin. nég	1.90	2.22	3.30	5.60	10.22	19.34	**4.16%**

TAB. 3.13 – Suite et fin de Table 3.11 (3/3)

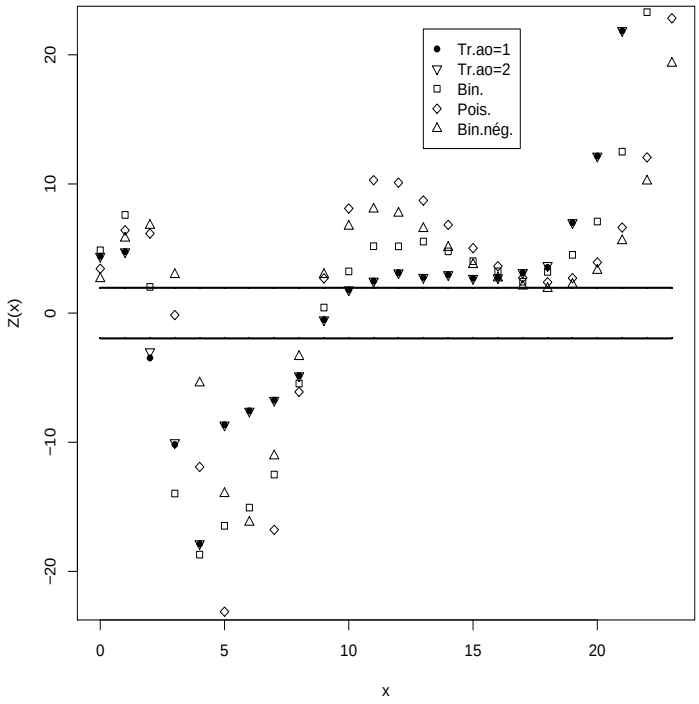

FIG. 3.4 – Valeurs de $Z(x)$ associées aux résultats dans Tables 3.11, 3.12 et 3.13

3.8 Conclusion

Dans ce chapitre, nous avons introduit une procédure d'estimation simple et très performante selon les cas pour des distributions de données de dénombrement. Il est possible d'envisager un départ paramétrique autre que les lois de Poisson et binomial considérées dans ce travail. Les estimateurs semi-paramétriques proposés ici en (3.17) et (3.28) surpassent en performance ceux non-paramétriques et paramétriques dans le cas d'échantillons de tailles finies lorsque la loi à estimer se situe dans un voisinage de la loi de départ (ici, Poisson ou binomiale). De plus, ils offrent plus de liberté dans le choix des noyaux discrets ainsi que des fenêtres de lissages. Ceci permet d'obtenir de meilleurs ajustements qu'avec le modèle paramétrique classique. Soulignons aussi le grand intérêt des modèles de diagnostique dans le choix de la méthode d'estimation appropriée. De plus, les résultats obtenus par les estimateurs semi-paramétriques proposés s'interprétent natutrellement par les distributions pondérées de Poisson ou binomiale ; ce qui n'est pas le cas de l'estimation non-paramétrique pure (3.1). Enfin, une extension directe de ce travail est de considérer une régression non-paramétrique sur les données de dénombrement en utilisant des noyaux discrets.

Chapitre 4

Régression non-paramétrique

4.1 Introduction

Considérons un échantillon formé d'une suite de n couples (x_i, y_i), $i = 1, \ldots, n$, à valeurs dans $\mathbb{N} \times \mathbb{R}$. Sans aucune spécification de distribution sur les y_i et en absence d'une forme de relation évidente entre y_i et x_i, le modèle classique de régression non-paramétrique de y_i sur x_i s'impose à travers :

$$y_i = m(x_i) + e_i, \tag{4.1}$$

où y_i est une réalisation de la variable aléatoire réelle (v.a.r.) Y_i à expliquer, x_i est une réalisation d'une variable explicative de dénombrement X_i, e_i correspond aux résidus non observables ϵ_i tels que $\mathbb{E}(\epsilon_i) = 0$ et var$(\epsilon_i) = \sigma^2$, et $m : \mathbb{N} \mapsto \mathbb{R}$ est la fonction discrète inconnue de régression. En considérant que les X_i sont des variables aléatoires, la fonction m peut être exprimée comme $m(x_i) = \mathbb{E}(Y_i|X_i = x_i)$. Dans ce dernier chapitre, nous nous intéressons au lissage discret ou à l'estimation non-paramétrique par noyaux associés discrets de cette fonction de régression discrète m en prenant en compte sa structure discrète.

Mis à part les travaux sur l'estimateur naïf, il existe quelques travaux sur le lissage non-paramétrique de variables ou fonctions discrètes depuis l'article des pionniers Aitchison & Aitken (1976) ; cependant, les noyaux discrets qu'ils proposent conviennent essentiellement aux données catégorielles. Une approche pour lisser la fonction de régression discrète m en (4.1) est de considérer le régresseur discret comme un régresseur continu, puis de lui appliquer l'une des nombreuses méthodes d'estimation non-paramétrique de fonctions de régression continues. Voir, par exemple, Chen (2000b), Collomb (1981), Gasser & Müller (1979), Michels (1992) pour des données continues et Cardot *et al.* (1999) pour des données fonctionnelles. Mais malheureusement, la structure propre de comptage du régresseur n'est pas prise en compte. Récemment, dans le cas particulier de la régression non-paramétrique binomiale où la variable réponse suit une loi binomiale $\mathcal{B}\{N_i, m(x_i)\}$ pour chaque covariable x_i, Okumura &

Naito (2004, 2006a) ont dû transformer la variable discrète x avant d'utiliser l'estimateur à noyau continu symétrique de Nadaraya (1964) et Watson (1964) pour une fonction de régression continue. Okumura & Naito (2006b) étendent aussi cette méthode à la régression multinomiale.

Dans ce travail, nous mettons encore en oeuvre la méthode des noyaux associés discrets pour estimer la fonction discrète de régression m en (4.1) sur $\mathbb{N}$ sans aucune transformation sur les variables discrètes. Nous adaptons l'estimateur continu de Nadaraya-Watson au cas discret. Cet estimateur pondéré est l'un des plus anciens et aussi des plus simples à mettre en oeuvre. A l'aide de simulations et de deux exemples de jeux de données, nous essayons d'illustrer la nécessité de cette approche et sa capacité à apporter de meilleures explications des données de la vie réelle. Cependant, notons que les deux exemples présentés correspondent à des séries temporelles et qu'ils ne sont utilisés qu'à titre d'illustrations.

Le premier jeu de données que nous utilisons (voir Table 4.1) correspond aux observations du chiffre de vente d'un produit pendant les 25 premiers jours qui suivent son introduction sur le marché. Les 160 observations (x_i, y_i), $i = 1, 2, \ldots, 25$, représentent le jour x_i de ventes et la moyenne de ventes correpondante $y_i \in \{y_{Ai}, y_{Bi}, \ldots, y_{Hi}\}$. Le nombre de centres de ventes dans chacune des villes $(A, B, \ldots, H)$ n'est pas disponible, sauf pour la ville H où il n'y a qu'un seul centre de vente. Il faut noter de plus que le jour $x_1 = 0$ représente le premier jour de lancement du produit sur le marché et son résultat de vente dépend essentiellement de la campagne publicitaire qui a précédé ; $x_2, x_3, \ldots$ correspondent aux autres jours pour lesquels les ventes du produit résultent aussi de la publicité informelle faite par les clients. Pendant les cinq premiers jours $x_1, x_2, \ldots, x_5$, les observations ont une allure croissante qui traduit l'intérêt de la clientèle pour ce nouveau produit. Au-delà du cinquième jour, le nombre moyen de ventes se met à décroitre ; ceci peut s'expliquer par le fait que la clientèle a suffisamment acheté le produit au cours des 5 premiers jours et qu'elle est satisfaite. Une autre explication est que le produit n'a pas atteint le succés escompté. Après qu'une quantité maximale vendue ait été atteinte, les ventes s'effondrent lentemement. Les chiffres de vente dans certains centres commerciaux ne sont pas disponibles au-delà du 13ème jour x_{14}, ce qui explique les données manquantes dans la Table 4.1. L'objectif de ce type d'étude est de déterminer une ligne d'action pour la campagne publicitaire qui est en général couteuse.

Le second jeu de données (voir Table 4.2) portent sur l'étude de la moyenne journalière de graisse (kg/jour) fournit par le lait d'une vache pendant 35 semaines (McCulloch, 2001 ; pages 40-45). La quantité de graisse contenue dans le lait augmente pendant les quatorze premières semaines avant de commencer à diminuer. Cela correspondrait à un cycle de production de lait qui dépendrait des périodes d'allaitement des veaux. Figure 4.1 y présente deux réalisations de modèles linéaires généralisés (McCullagh & Nelder, 1989 ; Kokonendji *et al.*, 2007a). En fait, le premier est un modèle normal pour la variable réponse log-transformée Y_i avec une fonction de lien identité et le second représente un modèle normal pour la variable réponse Y_i avec

x_i	y_{Ai}	y_{Bi}	y_{Ci}	y_{Di}	y_{Ei}	y_{Fi}	y_{Gi}	y_{Hi}
0	7.2	9.7	5.0	9.0	12.0	7.0	9.2	7
1	16.4	15.9	*5.0	16.1	16.2	15.4	16.2	16
2	21.4	18.9	22.2	19.7	17.2	20.7	19.9	23
3	22.0	19.7	24.2	20.5	18.1	22.4	21.1	20
4	20.4	19.2	23.0	19.8	17.6	21.7	20.5	23
5	18.2	18.1	20.4	18.4	16.8	19.8	*8.8	24
6	16.1	16.6	17.7	16.6	15.7	17.6	*6.5	12
7	14.2	15.0	15.2	14.8	14.5	15.3	*4.0	13
8	12.6	13.5	13.0	13.0	13.3	13.3	*1.8	9
9	11.1	12.0	11.1	11.5	12.1	11.5	9.9	9
10	9.9	10.6	9.5	10.0	10.9	9.9	8.5	8
11	8.7	9.4	8.2	8.8	9.8	8.6	7.5	10
12	7.8	8.3	7.1	7.7	8.7	7.4	6.8	8
13	6.9	7.3	6.2	6.7	7.8	6.5	6.3	7
14	–	6.4	5.4	5.9	6.9	5.6	5.9	2
15	5.4	5.6	4.7	–	6.1	4.9	5.6	*12
16	4.8	–	4.1	4.6	5.4	–	–	3
17	4.3	–	3.6	–	–	3.8	–	5
18	–	3.8	–	–	–	3.3	–	4
19	–	–	–	3.2	3.7	–	4.3	2
20	–	2.9	–	2.8	3.2	–	–	2
21	–	–	–	–	2.9	2.3	3.5	5
22	–	–	–	–	2.5	–	3.2	5
23	–	–	2.1	1.8	2.2	–	–	2
24	–	–	1.9	–	–	–	2.5	–

TAB. 4.1 – Données du chiffre de vente (les observations manquantes sont notées par – et celles notées avec * peuvent être considérées comme des valeurs particulières)

x_i	1	2	3	4	5	6	7	8	9	10	11	12
y_i	0.31	0.39	0.50	0.58	0.59	0.64	0.68	0.66	0.67	0.70	0.72	0.68
x_i	13	14	15	16	17	18	19	20	21	22	23	24
y_i	0.65	0.64	0.57	0.48	0.46	0.45	0.31	0.33	0.36	0.30	0.26	0.34
x_i	25	26	27	28	29	30	31	32	33	34	35	
y_i	0.29	0.31	0.29	0.20	0.15	0.18	0.11	0.07	0.06	0.01	0.01	

TAB. 4.2 – Moyenne journalière de graisse (kg/jour) dans le lait produit par une vache sur 35 semaines (McCulloch, 2001)

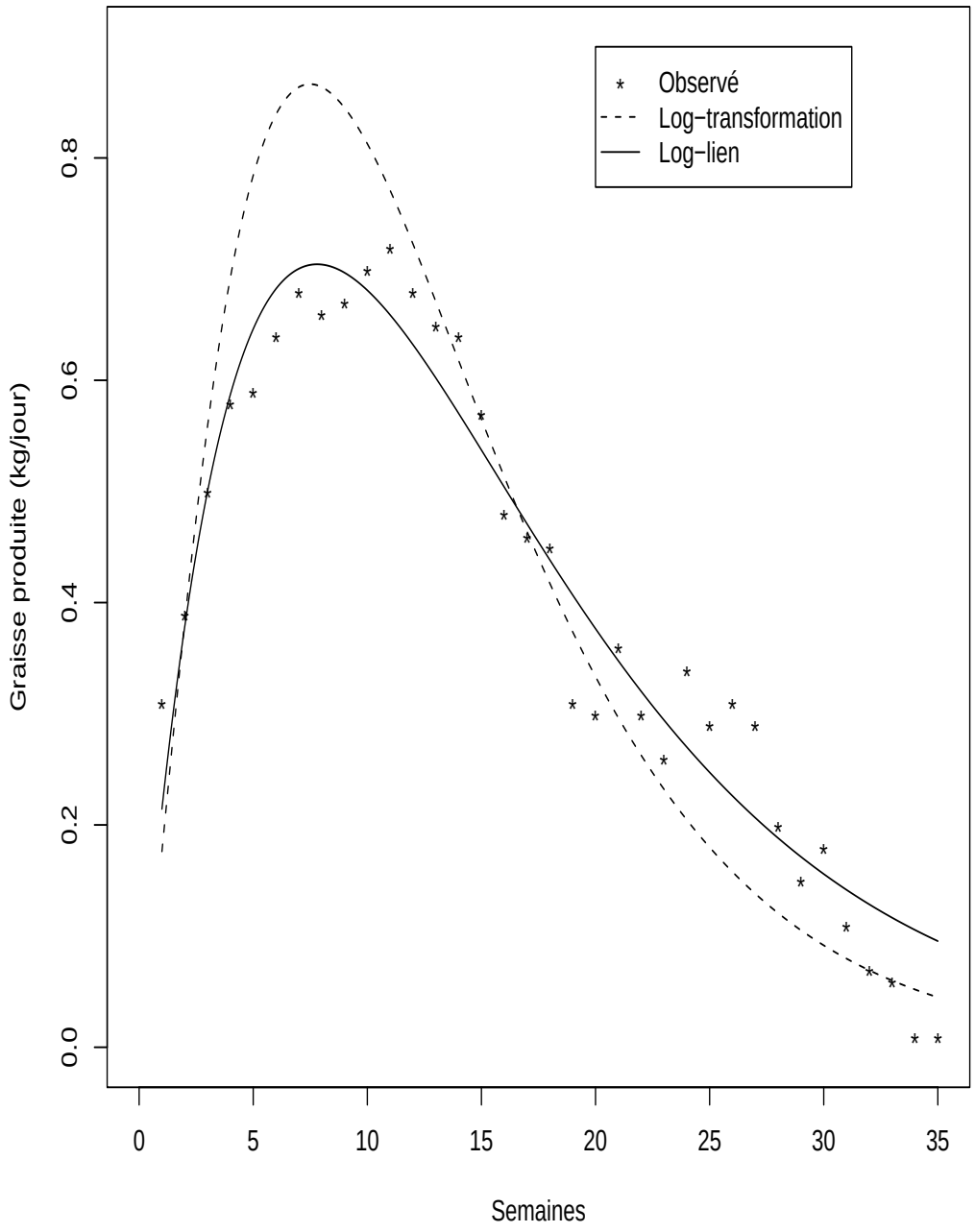

FIG. 4.1 – Deux modèles linéaires généralisés des données de la Table 4.2, avec le meilleur $R^2 = 0.7679$ (McCulloch, 2001)

une fonction de lien logarithmique. Les deux modèles ont le même prédicteur linéaire $\beta_0 + \beta_1 x_i + \beta_2 \log x_i$, où x_i représente la semaine. Nous observons que les deux modèles ne donnent pas un ajustement convenable même si il y a une amélioration du modèle log-lien ; en particulier, ils ne détectent pas le plateau associé aux observations $x = 19, 20, \ldots, 27$. Dans ce travail, nous comparerons ces modèles avec les résultats obtenus par notre nouvelle approche de (4.1).

Dans la Section 4.2, nous utilisons la méthode des noyaux associés discrets [Chapitre 3, Section 3.2] pour construire l'estimateur de m de (4.1) sur le modèle de celui de Nadaraya-Watson. Nous donnons quelques propriétés fondamentales de cet estimateur et nous adaptons la procédure de validation croisée pour un choix optimal du paramètre de lissage discret. Une étude par simulation ainsi que deux applications sont faites dans la Section 4.3. De plus, une comparaison est réalisée entre les différents noyaux discrets et aussi avec le noyau continu d'Epanechnikov (1969). Nous concluons dans la Section 4.4 par quelques remarques et des perspectives d'extension.

4.2 Version discrète de l'estimateur de Nadaraya-Watson

On considère le modèle non-paramétrique de régression discrète défini en (4.1). L'analogue discret de l'estimateur de Nadaraya (1964) et Watson (1964) pour la fonction discrète inconnue m de (4.1) est défini par

$$\widehat{m}_n(x) = \sum_{i=1}^{n} \omega_x(X_i) Y_i, \quad x \in \mathbb{N},$$

où

$$\omega_x(X_i) = \frac{K_{x,h}(X_i)}{\sum_{i=1}^{n} K_{x,h}(X_i)} = \omega_{x,h}(X_i) \tag{4.2}$$

représente la fonction poids telle que $\sum_{i=1}^{n} \omega_{x,h}(X_i) = 1$ en convenant que $0/0 = 0$, et $K_{x,h}(\cdot)$ est un noyau (associé) discret qui est défini en Section 3.2 du Chapitre 3. La fenêtre $h \equiv h(n, K)$ a pour rôle de déterminer le lissage discret de l'estimation. Pour un noyau discret approprié, de très petites largeurs de fenêtres h reproduisent presque les données tandis que de très grandes largeurs de h rapportent une estimation constante pour la fonction discrète de régression.

Pour l'étude du biais et de la variance de l'estimateur $\widehat{m}_n(x) \equiv \widehat{m}_n(x; h)$ de m, il est commode d'écire $\widehat{m}_n(x)$ comme le rapport

$$\widehat{m}_n(x) = \frac{N_n(x; h)}{D_n(x; h)},$$

avec

$$D_n(x; h) = \frac{1}{n} \sum_{i=1}^{n} K_{x,h}(X_i) = \widetilde{f}_n(x) \text{ et } N_n(x; h) = \frac{1}{n} \sum_{i=1}^{n} Y_i K_{x,h}(X_i).$$

Dans la section 3.2 du Chapitre 3, l'espérance et la variance de $D_n(x;h)$ sont approchées par :

$$\mathbb{E}\{D_n(x;h)\} = f\{\mathbb{E}(\mathcal{K}_{x,h})\} + \frac{\text{var}(\mathcal{K}_{x,h})}{2} f^{(2)}(x) + o(h)$$

et

$$\text{var}\{D_n(x;h)\} = \frac{1}{n} f(x) \{1 - f(x)\} \{\Pr(\mathcal{K}_{x,h} = x)\}^2 + R_n(x;h), \quad (4.3)$$

où le reste $R_n(x;h)$ est donné en (3.9). En ce qui concerne le numérateur $N_n(x;h)$, sachant que $m(x) = \mathbb{E}(Y_i|X_i = x)$, on exprime successivement :

$$\begin{aligned}
\mathbb{E}\{N_n(x;h)\} &= \mathbb{E}\{Y_1 K_{x,h}(X_1)\} \\
&= \sum_{z \in \aleph_x} m(z) f(z) \Pr(\mathcal{K}_{x,h} = z) \\
&= \mathbb{E}\{(mf)(\mathcal{K}_{x,h})\} \\
&= (mf)\{\mathbb{E}(\mathcal{K}_{x,h})\} + \frac{\text{var}(\mathcal{K}_{x,h})}{2} (mf)^{(2)}(x) + o(h),
\end{aligned}$$

avec $(mf)^{(2)} = m^{(2)} f + 2m^{(1)} f^{(1)} + m f^{(2)}$. Puis, à l'aide de l'espérance conditionnelle de Y^2 sachant X, la variance du numérateur $N_n(x;h)$ s'exprime comme suit :

$$\begin{aligned}
\text{var}\{N_n(x;h)\} &= \frac{1}{n} \mathbb{E}\{Y_1^2 K_{x,h}^2(X_1)\} - \frac{1}{n} \mathbb{E}^2\{Y_1 K_{x,h}(X_1)\} \\
&= \frac{1}{n} \sum_{y \in \aleph_x} \mathbb{E}(Y_1^2|X_1 = y) f(y) \{\Pr(\mathcal{K}_{x,h} = y)\}^2 \\
&\quad - \frac{1}{n} \left\{ \sum_{z \in \aleph_x} \mathbb{E}(Y_1|X_1 = z) f(z) \Pr(\mathcal{K}_{x,h} = z) \right\}^2 \\
&= \frac{1}{n} \mathbb{E}(Y_1^2|X_1 = x) f(x) \{\Pr(\mathcal{K}_{x,h} = x)\}^2 \\
&\quad - \frac{1}{n} \{\mathbb{E}(Y_1|X_1 = x) f(x) \Pr(\mathcal{K}_{x,h} = x)\}^2 + r_n(x;h) \\
&= \frac{1}{n} \left[\left\{ \mathbb{E}(Y_1^2|X_1 = x) - f(x) \mathbb{E}^2(Y_1|X_1 = x) \right\} f(x) \{\Pr(\mathcal{K}_{x,h} = x)\}^2 \right] \\
&\quad + r_n(x;h), \quad (4.4)
\end{aligned}$$

avec

$$\begin{aligned}
r_n(x;h) &= \frac{1}{n} \sum_{y \in \aleph_x \setminus \{x\}} \mathbb{E}(Y_1^2|X_1 = y) f(y) \{\Pr(\mathcal{K}_{x,h} = y)\}^2 \\
&\quad + \frac{1}{n} \{\mathbb{E}(Y_1|X_1 = x) f(x) \Pr(\mathcal{K}_{x,h} = x)\}^2 \\
&\quad - \frac{1}{n} \left\{ \sum_{z \in \aleph_x \setminus \{x\}} \mathbb{E}(Y_1|X_1 = z) f(z) \Pr(\mathcal{K}_{x,h} = z) \right\}^2 . \quad (4.5)
\end{aligned}$$

Nous déduisons des variances de $D_n(x;h)$ et de $N_n(x;h)$ que nous avons une convergence en moyenne d'ordre 2 de $D_n(x;h)$ et de $N_n(x;h)$ vers leurs espérances respectives. En effet, quand $n \to +\infty$ et $h = h(n) \to 0$, on a

$$\mathbb{E}[D_n(x;h) - \mathbb{E}\{D_n(x;h)\}]^2 = \text{var}\{D_n(x;h)\} \to 0 \tag{4.6}$$

et

$$\mathbb{E}[N_n(x;h) - \mathbb{E}\{N_n(x;h)\}]^2 = \text{var}\{N_n(x;h)\} \to 0. \tag{4.7}$$

Posons

$$A = \frac{N_n(x;h) - \mathbb{E}\{N_n(x;h)\}}{\mathbb{E}\{N_n(x;h)\}} \quad \text{et} \quad B = \frac{D_n(x;h) - \mathbb{E}\{D_n(x;h)\}}{\mathbb{E}\{D_n(x;h)\}},$$

avec $B \to 0$ quand n grand. Pour le calcul de l'espérance et de la variance de l'estimateur $\widehat{m}_n(x)$ (voir, par exemple, Bosq & Lecoutre, 1987), nous écrivons que

$$\begin{aligned}
\frac{N_n(x;h)}{D_n(x;h)} &= \frac{\mathbb{E}\{N_n(x;h)\}}{\mathbb{E}\{D_n(x;h)\}} \times \frac{1+A}{1+B} \\
&\doteq \frac{\mathbb{E}\{N_n(x;h)\}}{\mathbb{E}\{D_n(x;h)\}}(1+A)(1-B) \quad \text{p.s.} \\
&= \frac{\mathbb{E}\{N_n(x;h)\}}{\mathbb{E}\{D_n(x;h)\}}(1+A-B-AB),
\end{aligned} \tag{4.8}$$

où $\doteq$ indique un équivalent asymptotique et «p.s.» signifie «presque surement». En passant à l'espérance, nous obtenons

$$\begin{aligned}
\mathbb{E}\left\{\frac{N_n(x;h)}{D_n(x;h)}\right\} &\doteq \frac{\mathbb{E}\{N_n(x;h)\}}{\mathbb{E}\{D_n(x;h)\}}\left[1 - \frac{\text{cov}\{N_n(x;h), D_n(x;h)\}}{\mathbb{E}\{N_n(x;h)\}\mathbb{E}\{D_n(x;h)\}}\right] \\
&= \frac{\mathbb{E}\{N_n(x;h)\}}{\mathbb{E}\{D_n(x;h)\}}\left[2 - \frac{\mathbb{E}\{N_n(x;h)D_n(x;h)\}}{\mathbb{E}\{N_n(x;h)\}\mathbb{E}\{D_n(x;h)\}}\right].
\end{aligned}$$

D'abord, dans cette dernière expression, on exprime

$$\begin{aligned}
\mathbb{E}\{N_n(x;h)D_n(x;h)\} &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}\{Y_i K_{x,h}(X_i)K_{x,h}(X_j)\} \\
&= \frac{1}{n^2}[n\mathbb{E}\{Y_1 K_{x,h}^2(X_1)\} + n(n-1)\mathbb{E}\{Y_1 K_{x,h}(X_1)K_{x,h}(X_2)\}] \\
&= \frac{1}{n}\mathbb{E}\{Y_1 K_{x,h}^2(X_1)\} + \frac{n-1}{n}\mathbb{E}\{Y_1 K_{x,h}(X_1)\}\mathbb{E}\{K_{x,h}(X_2)\}.
\end{aligned}$$

Ensuite, en écrivant

$$\mathbb{E}\{N_n(x;h)\} = \mathbb{E}\{Y_1 K_{x,h}(X_1)\} \text{ et } \mathbb{E}\{D_n(x;h)\} = \mathbb{E}\{K_{x,h}(X_2)\},$$

nous avons

$$\lim_{n\to+\infty}\left[2-\frac{\mathbb{E}\{N_n(x;h)D_n(x;h)\}}{\mathbb{E}\{N_n(x;h)\}\mathbb{E}\{D_n(x;h)\}}\right] = 1+\lim_{n\to+\infty}\frac{1}{n}\left[\frac{\mathbb{E}\{Y_1K_{x,h}^2(X_1)\}}{\mathbb{E}\{Y_1K_{x,h}(X_1)\}\mathbb{E}\{K_{x,h}(X_2)\}}-1\right] = 1, \quad (4.9)$$

avec $\mathbb{E}\{Y_1K_{x,h}^2(X_1)\}[\mathbb{E}\{Y_1K_{x,h}(X_1)\}\mathbb{E}\{K_{x,h}(X_2)\}]^{-1} < +\infty$. Finalement, ceci conduit à

$$\mathbb{E}\left\{\frac{N_n(x;h)}{D_n(x;h)}\right\} \doteq \frac{\mathbb{E}\{N_n(x;h)\}}{\mathbb{E}\{D_n(x;h)\}},$$

et à l'expression du biais tel que :

$$\begin{aligned}\text{biais}\{\widehat{m}_n(x)\} &= \mathbb{E}\{\widehat{m}_n(x)\}-m(x)\\ &\doteq \frac{(mf)\{\mathbb{E}(\mathcal{K}_{x,h})\}+\{\text{var}(\mathcal{K}_{x,h})/2\}(mf)^{(2)}(x)}{f\{\mathbb{E}(\mathcal{K}_{x,h})\}+\{\text{var}(\mathcal{K}_{x,h})/2\}f^{(2)}(x)}-m(x). \quad (4.10)\end{aligned}$$

Nous vérifions immédiatement que $\text{biais}\{\widehat{m}_n(x)\} \to 0$ quand $h \to 0$ sous les hypothèses suivantes du noyau associé discret : $\mathbb{E}(\mathcal{K}_{x,h}) \sim x$ et $\text{var}(\mathcal{K}_{x,h}) \to 0$ quand $h \to 0$ [Equations (3.3) et (3.4)].

Pour déterminer la variance de $\widehat{m}_n(x)$, à partir de (4.8) nous obtenons

$$\begin{aligned}\text{var}\left\{\frac{N_n(x;h)}{D_n(x;h)}\right\} &\doteq \frac{[\mathbb{E}\{N_n(x;h)\}]^2}{[\mathbb{E}\{D_n(x;h)\}]^2}\text{var}(1+A-B-AB)\\ &\doteq \frac{[\mathbb{E}\{N_n(x;h)\}]^2}{[\mathbb{E}\{D_n(x;h)\}]^2}\{\text{var}(A)+\text{var}(B)-2\text{cov}(A,B)\},\end{aligned}$$

avec A et B définis après (4.7). Nous exprimons

$$\begin{aligned}\text{var}(A) &= \frac{\text{var}[N_n(x;h)-\mathbb{E}\{N_n(x;h)\}]}{\mathbb{E}^2\{N_n(x;h)\}}\\ &= \frac{\text{var}\{N_n(x;h)\}}{\mathbb{E}^2\{N_n(x;h)\}},\end{aligned}$$

et, par un calcul similaire, il vient que

$$\text{var}(B) = \frac{\text{var}\{D_n(x;h)\}}{\mathbb{E}^2\{D_n(x;h)\}}.$$

Ainsi, la variance de $\widehat{m}_n(x)$ est donnée par :

$$\begin{aligned}\text{var}\left\{\frac{N_n(x;h)}{D_n(x;h)}\right\} &\doteq \frac{\mathbb{E}^2\{N_n(x;h)\}}{\mathbb{E}^2\{D_n(x;h)\}}\\ &\quad\times\left[\frac{\text{var}\{N_n(x;h)\}}{\mathbb{E}^2\{N_n(x;h)\}}+\frac{\text{var}\{D_n(x;h)\}}{\mathbb{E}^2\{D_n(x;h)\}}-2\frac{\text{cov}\{N_n(x;h),D_n(x;h)\}}{\mathbb{E}\{N_n(x;h)\}\mathbb{E}\{D_n(x;h)\}}\right].\end{aligned}$$

Quand $n \to +\infty$, en utilisant (4.3) et (4.4) pour les variances de $D_n(x;h)$ et $N_n(x;h)$, respectivement, ainsi que (4.9) pour leur covaraince, nous obtenons l'approximation suivante de la variance de $\widehat{m}_n(x)$:

$$\begin{aligned}
\text{var}\{\widehat{m}_n(x)\} &= \text{var}\left\{\frac{N_n(x;h)}{D_n(x;h)}\right\} \\
&= \frac{\text{var}\{N_n(x;h)\}}{\mathbb{E}^2\{D_n(x;h)\}} + O\left(\frac{1}{n}\right) \\
&= \frac{1}{n} \times \frac{\{\mathbb{E}(Y_1^2|X_1=x) - f(x)\mathbb{E}^2(Y_1|X_1=x)\}\, f(x)}{[f\{\mathbb{E}(\mathcal{K}_{x,h})\}\{\text{var}(\mathcal{K}_{x,h})/2\}f^{(2)}(x)]^2}\, \{\Pr(\mathcal{K}_{x,h}=x)\}^2 \\
&\quad + \frac{r_n(x;h)}{\mathbb{E}^2\{D_n(x;h)\}} + O\left(\frac{1}{n}\right). \qquad (4.11)
\end{aligned}$$

Dans la sous-section suivante, nous simplifierons cette dernière expression selon les deux familles de noyaux discrets asymétriques et symétriques. Dans le cadre de la régression non-paramétrique continue, plusieurs auteurs (*e.g.* Michels, 1992 ; et leurs références) arrivent à des résultats simplifiés similaires pour les expressions asymptotiques de l'espérance et de la variance de l'estimateur à noyau de régression. De là, on peut étudier le risque global à travers l'erreur quadratique moyenne intégrée :

$$MISE(h) = \sum_{x\in\mathbb{N}} \text{var}\{\widehat{m}_n(x;h)\} + \sum_{x\in\mathbb{N}} \text{biais}^2\{\widehat{m}_n(x;h)\}. \qquad (4.12)$$

Pour comparer la qualité de la régression non-paramétrique entre les différents noyaux aussi bien discrets que continus, on utilise le coefficient de détermination R^2 bien connu lequel quantifie la proportion de variation des y_i expliquée par le régresseur x_i :

$$R^2 = \frac{\sum_{i=1}^n (\widehat{y}_i - \overline{y})^2}{\sum_{i=1}^n (y_i - \overline{y})^2}, \qquad (4.13)$$

où $\widehat{y}_i = \widehat{m}_n(x_i;h)$, $\overline{y} = n^{-1}(y_1 + \ldots + y_n)$ et

$$\frac{1}{n}\sum_{i=1}^n (y_i - \overline{y})^2 = \frac{1}{n}\sum_{i=1}^n (\widehat{y}_i - \overline{y})^2 + \frac{1}{n}\sum_{i=1}^n (y_i - \widehat{y}_i)^2.$$

4.2.1 Noyaux binomial et triangulaires discrets

Nous présentons dans cette sous-section les estimateurs de régression associés aux noyaux binomial et triangulaires discrets, et les expressions simplifiés de leurs biais en (4.10) et variance en (4.11).

EXEMPLE 4.2.1 (Binomial). L'estimateur non-paramétrique de régression associé au noyau binomial $B_{x,h}$ s'écrit explicitement par :

$$\begin{aligned}\widehat{m}_n^B(x) &= \frac{\sum_{i=1}^n Y_i B_{x,h}(X_i)}{\sum_{i=1}^n B_{x,h}(X_i)} \\ &= \frac{\sum_{i=1}^n Y_i \frac{(x+1)!}{X_i!(x+1-X_i)!}\left(\frac{x+h}{x+1}\right)^{X_i}\left(\frac{1-h}{x+1}\right)^{x+1-X_i}}{\sum_{i=1}^n \frac{(x+1)!}{X_i!(x+1-X_i)!}\left(\frac{x+h}{x+1}\right)^{X_i}\left(\frac{1-h}{x+1}\right)^{x+1-X_i}}.\end{aligned}$$

Nous rappellons ici que $\mathbb{E}(\mathcal{B}_{x,h}) = x + h$ and $\text{var}(\mathcal{B}_{x,h}) = V^B(x,h) + o(h)$ avec $V^B(x,h) = (x+h)/(x+1) - xh/(x+1)$. De là, l'espérance de $D_n^B(x;h) = \widetilde{f}_n(x)$ peut s'écrire comme

$$\mathbb{E}\{D_n^B(x;h)\} \doteq f(x) + hf^{(1)}(x) + \frac{V^B(x,h)}{2} f^{(2)}(x),$$

et pour le numérateur $N_n^B(x;h)$ nous avons

$$\mathbb{E}\{N_n^B(x;h)\} \doteq (mf)(x) + h(mf)^{(1)}(x) + \frac{V^B(x,h)}{2}(mf)^{(2)}(x),$$

avec $(mf)^{(1)} = m^{(1)}f + mf^{(1)}$ et $(mf)^{(2)} = m^{(2)}f + 2m^{(1)}f^{(1)} + mf^{(2)}$. De manière classique, considérons l'approximation $\mathbb{E}\{D_n^B(x;h)\} \doteq f(x)$, le biais en (4.10) de $\widehat{m}_n^B(x)$ devient alors

$$\text{biais}\{\widehat{m}_n^B(x)\} \doteq h\frac{(mf)^{(1)}(x)}{f(x)} + \frac{1}{2}\left(\frac{x+h-xh}{x+1}\right)\frac{(mf)^{(2)}(x)}{f(x)}$$

et la variance en (4.11) s'écrit ici

$$\begin{aligned}\text{var}\{\widehat{m}_n^B(x)\} &= \frac{\mathbb{E}(Y_1^2|X_1=x) - f(x)\mathbb{E}^2(Y_1|X_1=x)}{nf(x)}\left\{(1-h)\left(\frac{x+h}{x+1}\right)^x\right\}^2 \\ &\quad + \frac{r_n^B(x;h)}{\mathbb{E}^2\{D_n(x;h)\}} + O\left(\frac{1}{n}\right).\end{aligned}$$

Le biais ponctuel de l'estimateur non-paramétrique de régression associé au noyau binomial et, en général, aux noyaux discrets asymétriques dépend des différences finies d'ordre 1 et 2 de la fonction produit mf. Tandis que pour les noyaux discrets triangulaires ci-dessous, les biais dépendent uniquement de la différence finie du second ordre $(mf)^{(2)}$ de mf, comme c'est la cas pour l'estimateur à noyau associé discret d'une fonction de masse de probabilité [Section 3.2].

EXEMPLE 4.2.2 (Triangulaires discrets). Soit $a \in \mathbb{N}^*$, les estimateurs à noyaux discrets triangulaires $T_{a;x,h}$ associés pour la régression s'écrivent

$$\begin{aligned}\widehat{m}_n^{T_a}(x) &= \frac{\sum_{i=1}^n Y_i T_{a;x,h}(X_i)}{\sum_{i=1}^n T_{a;x,h}(X_i)} \\ &= \frac{\sum_{i=1}^n Y_i\{(a+1)^h - |X_i - x|^h\}}{\sum_{i=1}^n \{(a+1)^h - |X_i - x|^h\}}.\end{aligned}$$

En utilisant $\mathbb{E}(\mathcal{T}_{a;x,h}) = x$ et var$(\mathcal{T}_{a;x,h}) = V(a,h)$ définie dans l'Exemple 3.2.2 [Chapitre 3, Section 3.2], l'espérance de $D_n^{T_a}(x;h) = \widetilde{f}_n(x)$ s'obtient par

$$\mathbb{E}\{D_n^{T_a}(x;h)\} \doteq f(x) + \frac{V(a,h)}{2} f^{(2)}(x),$$

et nous exprimons

$$\mathbb{E}\{N_n^{T_a}(x;h)\} \doteq (mf)(x) + \frac{V(a,h)}{2}(mf)^{(2)}(x).$$

Puis, les expressions simplifiées du biais en (4.10) et de la variance en (4.11) de $\widehat{m}_n^{T_a}(x)$ obtenues en utilisant $\mathbb{E}\{D_n(x)\} \doteq f(x)$ sont, respectivement,

$$\text{biais}\{\widehat{m}_n^{T_a}(x)\} \doteq \frac{V(a,h)}{2}\frac{(mf)^{(2)}(x)}{f(x)}$$

et

$$\text{var}\{\widehat{m}_n^{T_a}(x)\} = \frac{\mathbb{E}(Y_1^2|X_1=x) - f(x)\mathbb{E}^2(Y_1|X_1=x)}{nf(x)} \left\{\frac{(a+1)^h}{P(a,h)}\right\}^2 + o\left(\frac{1}{n}\right).$$

Le biais ponctuel est proportionnel à la variance $\mathrm{V}(a,h)$ de $\mathcal{T}_{a;x,h}$ et aussi au rapport $(mf)^{(2)}(x)/f(x)$.

4.2.2 Risque asymptotique ponctuel

Nous formulons ici un résultat sur le risque ponctuel de l'estimateur $\widehat{m}_n$ de la fonction discrète de régression m qui est liée à la fonction de masse f du régresseur. Pour cela, nous développons une approximation plus fine de l'espérance.

Proposition 4.2.1 *Pour $x \in \mathbb{N}$, soient $m(x) = \mathbb{E}(Y|X=x)$ et $f(x) = \Pr(X=x)$ définies de $\mathbb{N} \mapsto \mathbb{R}$. Soit $\widehat{m}_n(x)$ l'estimateur de $m(x)$ à noyau associé discret $K_{x,h}$ sur $\aleph_x$. Alors, quand $n \to +\infty$ et $h = h(n) \to 0$, en tout point x où $f(x) \neq 0$ on a les développements asymptotiques*

$$\mathbb{E}\{\widehat{m}_n(x)\} - m(x) = \left\{m^{(2)}(x) + 2m^{(1)}(x)\left(\frac{f^{(1)}}{f}\right)(x)\right\}\frac{\text{var}(\mathcal{K}_{x,h})}{2} + O(1/n)^2 + o(h)$$

et

$$\text{var}\{\widehat{m}_n(x)\} = \frac{\mathbb{E}(Y_1^2|X_1=x) - f(x)\mathbb{E}^2(Y_1|X_1=x)}{nf(x)}\left\{\Pr(\mathcal{K}_{x,h}=x)\right\}^2 + o\left(\frac{1}{n}\right).$$

Par conséquent, l'erreur quadratique s'écrit

$$\begin{aligned} MSE(x) &= \left\{m^{(2)}(x) + 2m^{(1)}(x)\left(\frac{f^{(1)}}{f}\right)(x)\right\}^2 \frac{\text{var}^2(\mathcal{K}_{x,h})}{4} \\ &\quad + \frac{\mathbb{E}(Y_1^2|X_1=x) - f(x)\mathbb{E}^2(Y_1|X_1=x)}{nf(x)}\left\{\Pr(\mathcal{K}_{x,h}=x)\right\}^2 + o\left(h^2 + \frac{1}{n}\right). \end{aligned}$$

DEMONSTRATION : Suivant Bosq & Lecoutre (1987 ; pages 119-121) pour le cas continu avec ici $g = mf$, nous pouvons écrire

$$\begin{aligned}\widehat{m}_n(x) &= m(x) + \frac{N_n(x;h) - g(x)}{f(x)} - \frac{g(x)\{D_n(x;h) - f(x)\}}{f^2(x)} \\ &\quad - \frac{\{N_n(x;h) - g(x)\}\{D_n(x;h) - f(x)\}}{f^2(x)} \\ &\quad + \frac{N_n(x;h)}{f^3(x)}\{D_n(x;h) - f(x)\}^2\{1 + o(1)\} \text{ p.s.} \end{aligned} \tag{4.14}$$

avec $D_n(x;h) = n^{-1}\sum_{i=1}^n K_{x,h}(X_i)$ et $N_n(x;h) = n^{-1}\sum_{i=1}^n Y_i K_{x,h}(X_i)$. En passant à l'espérance, nous obtenons

$$\begin{aligned}\mathbb{E}\{\widehat{m}_n(x)\} - m(x) &= \frac{\mathbb{E}\{N_n(x;h) - g(x)\}}{f(x)} - \frac{g(x)\mathbb{E}\{D_n(x;h) - f(x)\}}{f^2(x)} \\ &\quad - \frac{\mathbb{E}[\{N_n(x;h) - g(x)\}\{D_n(x;h) - f(x)\}]}{f^2(x)} \\ &\quad + \frac{\mathbb{E}[N_n(x;h)\{D_n(x;h) - f(x)\}^2]}{f^3(x)}\{1 + o(1)\},\end{aligned}$$

où $o(1)$ ne dépend pas de n et tend vers 0 quand $h \to 0$. Tout d'abord, sous la condition suivante du noyau associé discret $\mathbb{E}(\mathcal{K}_{x,h}) \sim x$ quand $h \to 0$, les espérances de $D_n(x;h)$ et $N_n(x;h)$ s'expriment par

$$\mathbb{E}\{D_n(x;h)\} - f(x) = \frac{\text{var}(\mathcal{K}_{x,h})}{2} f^{(2)}(x) + o(h)$$

et

$$\mathbb{E}\{N_n(x;h)\} - (mf)(x) = \frac{\text{var}(\mathcal{K}_{x,h})}{2}(mf)^{(2)}(x) + o(h),$$

avec $(mf)^{(2)} = m^{(2)}f + 2m^{(1)}f^{(1)} + mf^{(2)}$.

Ensuite, nous avons

$$\begin{aligned}&\mathbb{E}[N_n(x;h)\{D_n(x;h) - f(x)\}^2] \\ =\ &\mathbb{E}\left([N_n(x;h) - \mathbb{E}\{N_n(x;h)\}][D_n(x;h) - \mathbb{E}\{D_n(x;h)\}]^2\right) + 2[\mathbb{E}\{D_n(x;h)\} - f(x)] \\ &\times\mathbb{E}\big([N_n(x;h) - \mathbb{E}\{N_n(x;h)\}][D_n(x;h) - \mathbb{E}\{D_n(x;h)\}]\big) \\ &+\mathbb{E}[\{D_n(x;h) - f(x)\}^2]\mathbb{E}\{N_n(x;h)\} \\ =\ &O(1/n)^2 + O(1/n) + \mathbb{E}\{D_n(x;h) - f(x)\}^2\mathbb{E}\{N_n(x;h)\}.\end{aligned}$$

En effet, en écrivant

$$\begin{aligned}\mathbb{E}[\{N_n(x;h) - \mathbb{E}\{N_n(x;h)\}][D_n(x;h) - \mathbb{E}\{D_n(x;h)\}] &= \mathbb{E}\{N_n(x;h)D_n(x;h)\} \\ &\quad -\mathbb{E}\{N_n(x;h)\}\mathbb{E}\{D_n(x;h)\},\end{aligned}$$

nous avons

$$\mathbb{E}\{N_n(x;h)D_n(x;h)\} = \frac{1}{n}\mathbb{E}\{Y_1K_{x,h}^2(X_1)\} + \frac{n-1}{n}\mathbb{E}\{Y_1K_{x,h}(X_1)\}\mathbb{E}\{K_{x,h}(X_2)\}$$

ainsi que $\mathbb{E}\{N_n(x;h)\} = \mathbb{E}\{Y_1K_{x,h}(X_1)\}$ et $\mathbb{E}\{D_n(x;h)\} = \mathbb{E}\{K_{x,h}(X_2)\}$. Ceci conduit à

$$\begin{aligned} & \mathbb{E}\big([N_n(x;h) - \mathbb{E}\{N_n(x;h)\}][D_n(x;h) - \mathbb{E}\{D_n(x;h)\}]\big) \\ = & \frac{1}{n}[\mathbb{E}\{Y_1K_{x,h}^2(X_1)\} - \mathbb{E}\{Y_1K_{x,h}(X_1)\}\mathbb{E}\{K_{x,h}(X_2)\}] \\ = & O(1/n). \end{aligned}$$

De manière similaire, il vient que $\mathbb{E}\big([N_n(x;h)-\mathbb{E}\{N_n(x;h)\}][D_n(x;h)-\mathbb{E}\{D_n(x;h)\}]^2\big) = O(1/n)^2$. D'où, nous obtenons

$$\begin{aligned} \mathbb{E}\{\widehat{m}_n(x)\} - m(x) = & \left\{\frac{(mf)^{(2)}(x)}{f(x)} - \frac{mf^{(2)}(x)}{f(x)}\right\}\frac{\text{var}(\mathcal{K}_{x,h})}{2} \\ & + \frac{\mathbb{E}\{D_n(x;h) - f(x)\}^2\mathbb{E}\{N_n(x;h)\}}{f^3(x)} \\ & - \frac{f(x)\mathbb{E}[\{N_n(x;h) - g(x)\}\{D_n(x;h) - f(x)\}]}{f^3(x)} + O(1/n)^2 + o(h). \end{aligned}$$

De là, nous exprimons le numérateur du second terme de l'expression précédente comme

$$\begin{aligned} & \mathbb{E}\{D_n(x;h) - f(x)\}^2\mathbb{E}\{N_n(x;h)\} - f(x)\mathbb{E}[\{N_n(x;h) - g(x)\}\{D_n(x;h) - f(x)\}] \\ = & \mathbb{E}\{N_n(x;h)\}\mathbb{E}\{D_n^2(x;h)\} - \mathbb{E}\{D_n(x;h)\}\mathbb{E}\{N_n(x;h)D_n(x;h)\} + o(h) \\ = & \mathbb{E}\{N_n(x;h)\}[\text{var}\{D_n(x;h)\} + \mathbb{E}^2\{D_n(x;h)\}] - \mathbb{E}\{D_n(x;h)\} \\ & \times[\text{cov}\{D_n(x;h), N_n(x;h)\} + \mathbb{E}\{D_n(x;h)\}\mathbb{E}\{N_n(x;h)\}] + o(h) \\ = & \mathbb{E}\{N_n(x;h)\}[O(1/n) + \mathbb{E}^2\{D_n(x;h)\}] - \mathbb{E}\{D_n(x;h)\} \\ & \times[O(1/n) + \mathbb{E}\{D_n(x;h)\}\mathbb{E}\{N_n(x;h)\}] + o(h). \end{aligned}$$

Finalement, nous aboutissons à

$$\mathbb{E}\{\widehat{m}_n(x)\}-m(x) = \left\{m^{(2)}(x) + 2m^{(1)}(x)\left(\frac{f^{(1)}}{f}\right)(x)\right\}\frac{\text{var}(\mathcal{K}_{x,h})}{2} + O(1/n)^2 + o(h). \tag{4.15}$$

Pour la variance de l'estimateur $\widehat{m}_n$, à partir de (4.14) nous obtenons :

$$\begin{aligned} \text{var}\{\widehat{m}_n(x)\} = & \frac{\text{var}\{N_n(x;h)\}}{f^2(x)} + \frac{g^2(x)}{f^4(x)}\text{var}\{D_n(x;h)\} \\ & - 2\frac{g(x)}{f^3(x)}\text{cov}\{N_n(x;h), D_n(x;h)\} + o\left(\frac{1}{n}\right). \end{aligned}$$

Dans cette dernière expression, nous exprimons les différents termes par

$$\begin{aligned}\text{var}\{N_n(x;h)\} &= \frac{1}{n}\left[\left\{\mathbb{E}(Y_1^2|X_1=x) - f(x)\mathbb{E}^2(Y_1|X_1=x)\right\} f(x)\left\{\Pr(\mathcal{K}_{x,h}=x)\right\}^2\right] \\ &\quad + R_n(x;h),\end{aligned}$$

puis $\text{var}\{D_n(x;h)\} = O(1/n)$ et $\text{cov}\{N_n(x;h), D_n(x;h)\} = O(1/n)$. Ceci amène à exprimer

$$\text{var}\{\widehat{m}_n(x)\} = \frac{\mathbb{E}(Y_1^2|X_1=x) - f(x)\mathbb{E}^2(Y_1|X_1=x)}{nf(x)}\left\{\Pr(\mathcal{K}_{x,h}=x)\right\}^2 + o\left(\frac{1}{n}\right). \tag{4.16}$$

Les expressions de l'espérance (4.15) et de la variance (4.16) de $\widehat{m}_n(x)$ permettent de déduire l'erreur quadratique énoncée en un point $x \in \mathbb{N}$. ■

4.2.3 Choix de fenêtre par validation croisée

Dans le contexte de la régression à noyau discret de données de dénombrement, nous étudions la sélection de fenêtre par la méthode de validation croisée (Hardle & Marron, 1985). Pratiquement, pour un noyau associé discret donné, la fenêtre optimale s'obtient par

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

avec

$$CV(h) = \frac{1}{n}\sum_{i=1}^{n}\{Y_i - \widehat{m}_{-i}(X_i;h)\}^2 M(X_i), \tag{4.17}$$

où $\widehat{m}_{-i}(X_i;h) = \sum_{j\neq i}^{n} Y_j K_{X_i,h}(X_j) / \sum_{j\neq i}^{n} K_{X_i,h}(X_j)$ est obtenu en otant l'observation X_i dans $\widehat{m}_n(X_i;h)$ et $0 \leq M(\cdot) \leq 1$ est un poids qui permet d'éviter les difficultés liées à la division par 0 ou à une faible vitesse de convergence due au biais de bordure.

En effet, pour trouver la fenêtre optimale, on a besoin de minimiser par rapport à h un estimateur de $MISE$ (4.12) qu'on développe comme suit :

$$\begin{aligned}MISE(h) &= \mathbb{E}\left\{\sum_{x\in\mathbb{N}} \widehat{m}_n^2(x;h) f(x) M(x)\right\} - 2\mathbb{E}\left\{\sum_{x\in\mathbb{N}} \widehat{m}_n(x;h) m(x) f(x) M(x)\right\} \\ &\quad + \mathbb{E}\left\{\sum_{x\in\mathbb{N}} m^2(x) f(x) M(x)\right\},\end{aligned}$$

où f est la fonction de masse inconnue de X. Comme le dernier terme $\mathbb{E}\left\{\sum_{x\in\mathbb{N}} m^2(x) f(x) M(x)\right\}$ ne dépend pas de h, il faut minimiser un estimateur de l'expression

$$\mathbb{E}\left\{\sum_{x\in\mathbb{N}} \widehat{m}_n^2(x;h) f(x) M(x)\right\} - 2\mathbb{E}\left\{\sum_{x\in\mathbb{N}} \widehat{m}_n(x;h) m(x) f(x) M(x)\right\}$$

qui dépend de la fonction discrète de régression inconnue m. Considérons

$$\widehat{H}_n \;=\; \frac{1}{n}\sum_{i=1}^{n} \widehat{m}^2_{-i}\,(X_i;h)\,M(X_i).$$

Puisque les v.a. $X_1,\ldots,X_n$ sont i.i.d., on a alors

$$\mathbb{E}(\widehat{H}_n) \;=\; \mathbb{E}\{\widehat{m}^2_{-1}(X_1;h)M(X_1)\}$$

et

$$\mathbb{E}\left\{\sum_{x\in\mathbb{N}} \widehat{m}^2_n(x;h)f(x)M(x)\right\} = \mathbb{E}\{\widehat{m}^2_n(X_1;h)M(X_1)\}.$$

Ainsi, il est facile de voir que $\widehat{H}_n$ est un estimateur de $\mathbb{E}\{\sum_{x\in\mathbb{N}} \widehat{m}^2_n(x;h)f(x)M(x)\}$ mais pas nécessairement sans biais parce qu'ils n'ont pas la même distribution. De manière similaire, le second terme est estimé par

$$\widehat{G}_n = \frac{1}{n}\sum_{i=1}^{n} \widehat{m}_{-i}\,(X_i;h)\,Y_iM(X_i).$$

En fait, on a d'une part

$$\mathbb{E}(\widehat{G}_n) \;=\; \mathbb{E}\{\widehat{m}_{-1}(X_1;h)Y_1M(X_1)\}$$

et, d'autre part, on exprime

$$\mathbb{E}\left\{\sum_{x\in\mathbb{N}} \widehat{m}_n(x;h)m(x)f(x)M(x)\right\} = \mathbb{E}\{\widehat{m}_n(X_1;h)Y_1M(X_1)\}.$$

Finalement, un estimateur de la $MISE(h)$ est donnée par

$$\widehat{H}_n - 2\widehat{G}_n = \frac{1}{n}\sum_{i=1}^{n} \widehat{m}^2_{-i}(X_i;h)M(X_i) - \frac{2}{n}\sum_{i=1}^{n} \widehat{m}_{-i}\,(X_i)\,Y_iM(X_i)$$

car, en rajoutant le terme $n^{-1}\sum_{i=1}^{n} {Y_i}^2M(X_i)$ qui ne dépend pas de h, on a établi l'estimateur asymptotiquement sans biais de (4.17).

Notons que, pour l'utilisation pratique de la fonction poids $M(\cdot)$ en (4.17), nous pouvons considérer $M(X_i) = \omega_{X_i,h}(X_i) \equiv M(X_i;h)$ défini en (4.2) et dépendant de h. Dans ce cas, la minimisation globale par rapport à h de l'estimateur de la $MISE(h)$ ainsi obtenu

$$\widehat{MISE}(h) = \widehat{H}_n - 2\widehat{G}_n + \frac{1}{n}\sum_{i=1}^{n} Y_i^2M(X_i;h)$$

procure des résultats équivalents pour cette régression à noyau discret sur des données de dénombrement avec la fenêtre de validation croisée modifiée $h_{mcv} = \arg\min_{h>0} \widehat{MISE}(h)$. Cependant, la vitesse de convergence est faible.

4.3 Illustrations

Dans cette section, nous mettons en oeuvre l'estimateur à noyau discret $\widehat{m}_n$ dans une étude par simulation et sur des données $(x_i, y_i) \in \mathbb{N}\times\mathbb{R}$, $i = 1, 2, \ldots, n$ des Tables 4.1 et 4.2. Pour analyser et effectuer un travail comparatif, on complète les Exemples 4.2.1 et 4.2.2 avec les deux noyaux discrets de Poisson et binomial négatif ainsi que le noyau optimal continu d'Epanechnikov (1969). De manière précise, le noyau de Poisson $P_{x,h}$ est tel que

$$P_{x,h}(z) = \frac{(x+h)^z e^{-(x+h)}}{z!}, \quad z \in \mathbb{N}\,;$$

le noyau binomial négatif $BN_{x,h}$ est défini par

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

et le noyau continu d'Epanechnikov (1969) est donné par :

$$K^E(z) = \frac{3}{4}(1-z^2), \;\; z \in [-1,1].$$

Dans les illustrations qui suivront, pour chaque noyau utilisé, la fenêtre optimale h_{cv} est obtenue générallement par la méthode de validation croisée avec le poids $M \equiv 1$. Puis, nous déterminons le coefficient de détermination R^2 correspondant. Finalement, les courbes de régression tracées dans les différentes figures sont obtenues par interpolation linéaire entre les points de régression. Dans les applications suivantes, on ne s'intéressera pas à l'étude des diagnostiques ou des résidus. Cependant, nous utiliserons parfois le terme de meilleur ajustement dans le sens de la $MISE$ si la fenêtre est choisie par la procédure de validation croisée.

4.3.1 Etude par simulation

Considérons la fonction discrète de régression

$$m(x) = \frac{2^x}{x!}, \; x \in \mathbb{N}.$$

À l'issue de 1000 répliques, nous calculons la moyenne des erreurs quadratiques intégrées optimales ainsi que leurs écart-types pour les estimateurs (voir Table 4.3). Pour chaque simulation, les fenêtres optimales de lissage discret sont obtenues par la procédure de validation croisée. De là, les erreurs quadratiques intégrées optimales sont calculées en utilisant les valeurs optimales des fenêtres de lissage discret. Les résultats de la Table 4.3 montrent que les noyaux associés discrets triangulaires avec de petits bras sont plus performants que les noyaux binomial et d'Epanechnikov, même

dans le cas d'échantillons de grande taille. Nous ne recommandons pas l'utilisation des noyaux de Poisson et binomial négatif qui ne sont pas sousdispersés et qui ne satisfont pas la condition (3.4) sur la variance des noyaux associés discrets [Section 3.2, Chapitre 3]. Finalement, notant que les moyennes des erreurs quadratiques intégrées optimales simulées des meilleurs estimateurs à noyaux discrets et continu sont voisines lorsque la taille de l'échantillon n augmente.

n	Triang.1	Triang.2	Binomial	*Epanech.*	Poisson	Bin. nég.
20	11.77 (8.2)	24.51 (19.9)	43.87 (49.6)	55.06 (107.2)	65.98 (51.2)	82.73 (122.5)
50	7.01 (8.4)	17.00 (14.1)	27.03 (26.5)	32.02 (31.6)	69.10 (36.1)	87.01 (43.6)
80	4.47 (4.5)	14.13 (8.7)	18.18 (13.6)	23.30 (19.8)	69.42 (27.6)	88.45 (42.9)
100	3.67 (2.9)	12.68 (7.9)	17.03 (11.4)	21.93 (16.5)	70.51 (25.7)	92.12 (37.3)
200	2.43 (1.4)	10.20 (4.0)	12.60 (5.6)	14.81 (15.0)	76.48 (21.2)	102.17 (23.4)
500	1.86 (0.7)	8.46 (2.0)	9.92 (2.5)	9.96 (10.0)	83.45 (17.3)	116.13 (22.4)
1000	1.60 (0.4)	8.09 (1.4)	9.19 (1.5)	8.27 (7.9)	89.69 (13.0)	126.78 (20.1)

TAB. 4.3 – Moyennes des erreurs quadratiques intégrées optimales simulées et leurs écart-types (entre parenthèses) pour les estimateurs à noyaux discrets et d'Epanechnikov. Les résultats présentés ont été multipliés par 10^3

4.3.2 Moyenne de quantité de graisse

Table 4.4 et Figure 4.2 présentent les résultats correspondant aux régressions non-paramétriques sur les données de Table 4.2. Le noyau associé discret triangulaire avec $a = 2$, $h_{cv} = 0.1$ et $R^2 = 99.140\%$ produit le meilleur résultat au sens de R^2 et de $MISE$. Il est suvi par le binomial ($R^2 = 97.179\%$), puis le continu d'Epanechnikov ($R^2 = 96.967\%$). Ces trois noyaux mettent en évidence le plateau associé aux observations $x_i = 19, 20, \ldots, 25$. À l'opposé, les noyaux de Poisson et binomial négatif, comme les modèles linéaires généralisés présentés en Introduction, ne détectent pas cette allure ; ils surestiment ou sousestiment la plupart des valeurs de y.

Pour ce jeu données pour lequel à un x_i est associé un unique y_i, la régression non-paramétrique utilisant le noyau discret triangulaire $a = 1$ procure de meilleurs résultats au sens de R^2 et de $MISE$. Par contre, la courbe de régression reproduit quasiment les données. Nous améliorons le lissage discret en prenant d'autres valeurs du bras a qui est libre et dépend du choix de l'utilisateur. Cependant, pour le noyau associé discret triangulaire $a = 2$ et en faisant varier le paramètre de lissage, la courbe de régression obtenue devient plus régulière ($h = 0.5$ et $R^2 = 97.321$). Ceci est nettement plus perceptible en augmentant le bras a, puis en prenant plusieurs valeurs de

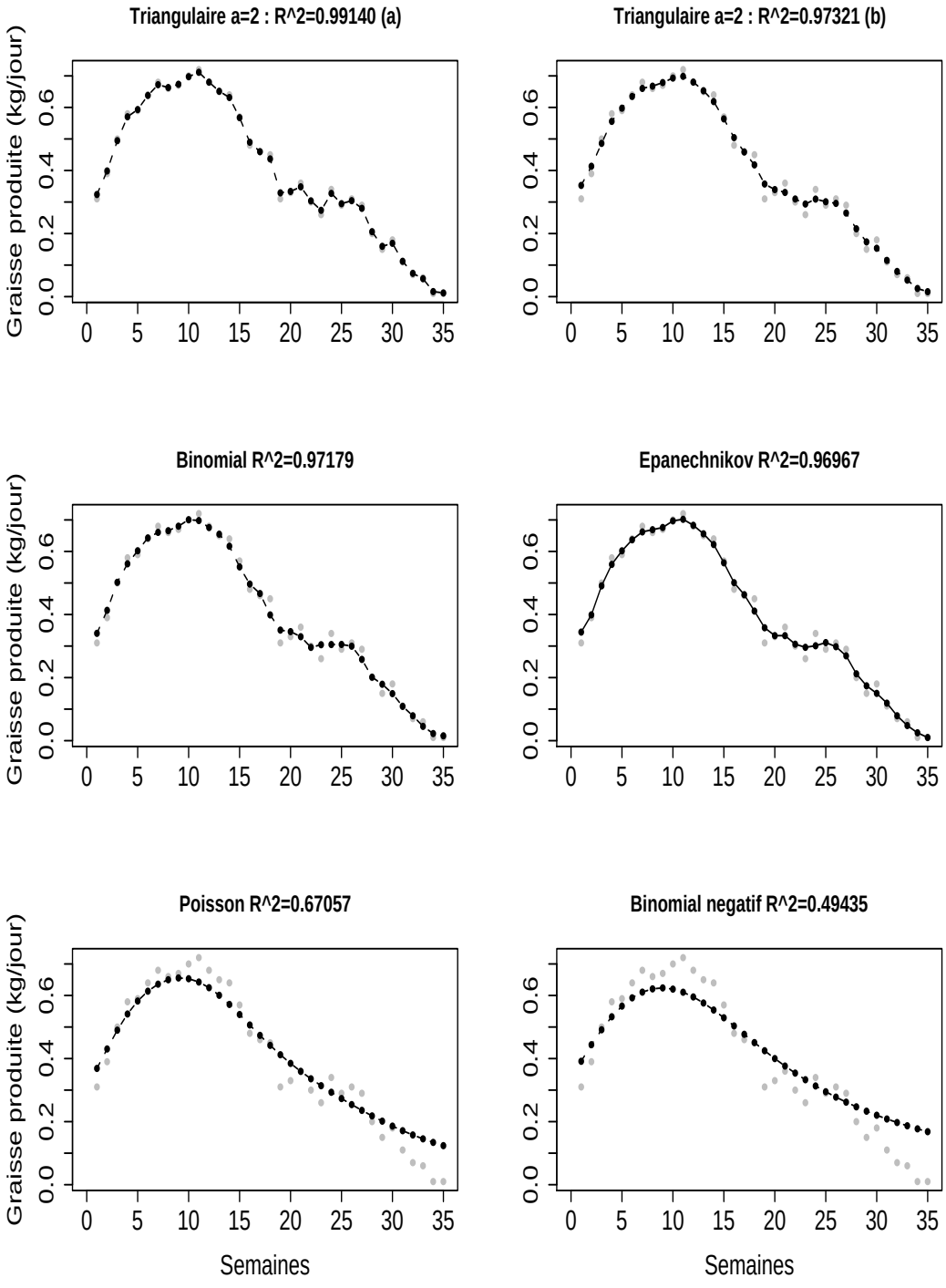

FIG. 4.2 – Régressions sur les données de graisse (Table 4.2) par les estimateurs à noyaux discrets et à noyau continu d'Epanechnikov

Noyau	Triang.2 (a)	Triang.2 (b)	Binomial	*Epanech.*	Poisson	Bin. nég.
h_{cv}	0.1	0.5*	0.101	4.0	0.151	0.224
R^2	99.140	97.321	97.179	96.967	67.057	49.435

TAB. 4.4 – Coefficient de détermination (en %) des régressions sur les données de graisse (Table 4.2) par les estimateurs à noyaux discrets et continu

h autres que h_{cv}. Ainsi, en utilisant le noyau associé discret triangulaire $a = 4$ et $h \in \{0.1, 0.3, 0.7, 1\}$, on obtient les résultats présentés en Table 4.5 et Figure 4.3. On voit alors que les courbes de régression obtenues sont plus lisses bien que le R^2 ne varie pas beaucoup. Si le paramètre de lissage optimale h_{cv} donne une qualité de régression optimale, l'usage d'autres valeurs h de la fenêtre permet d'obtenir une bonne régression à l'allure plus régulière. Le choix reste ouvert pour trouver un bras optimal a qui donne le meilleur R^2 associé à la fenêtre optimale.

Triang.4	(a)	(b)	(c)	(d)
h	0.1	0.3	0.7	1
R^2	96.445	93.101	90.126	88.843

TAB. 4.5 – Suite et fin de Table 4.4 pour l'estimateur à noyau discret triangulaire $a = 4$

4.3.3 Données de vente

Table 4.6 et Figure 4.4 présentent les résultats de régressions non-paramétriques faites par des noyaux discrets (triangulaire avec $a \in \{1, 2\}$, binomial, Poisson et binomial négatif) et par le noyau continu d'Epanechninov pour le jeu de données de la Table 4.1.

Parmi les noyaux discrets, les noyaux triangulaires avec $a \in \{1, 2\}$ et binomial donnent les meilleurs résultats dans les sens de $MISE$ et de R^2 qui soit autour de 95%, 92% et 80%, respectivement. Les noyaux de Poisson et binomial négatif donnent des valeurs sousestimées pour $x < 8$ et surestimées pour $x > 9$. En otant 6 points considérés comme extrêmes, les valeurs du R^2 augmentent de façon générale, mais

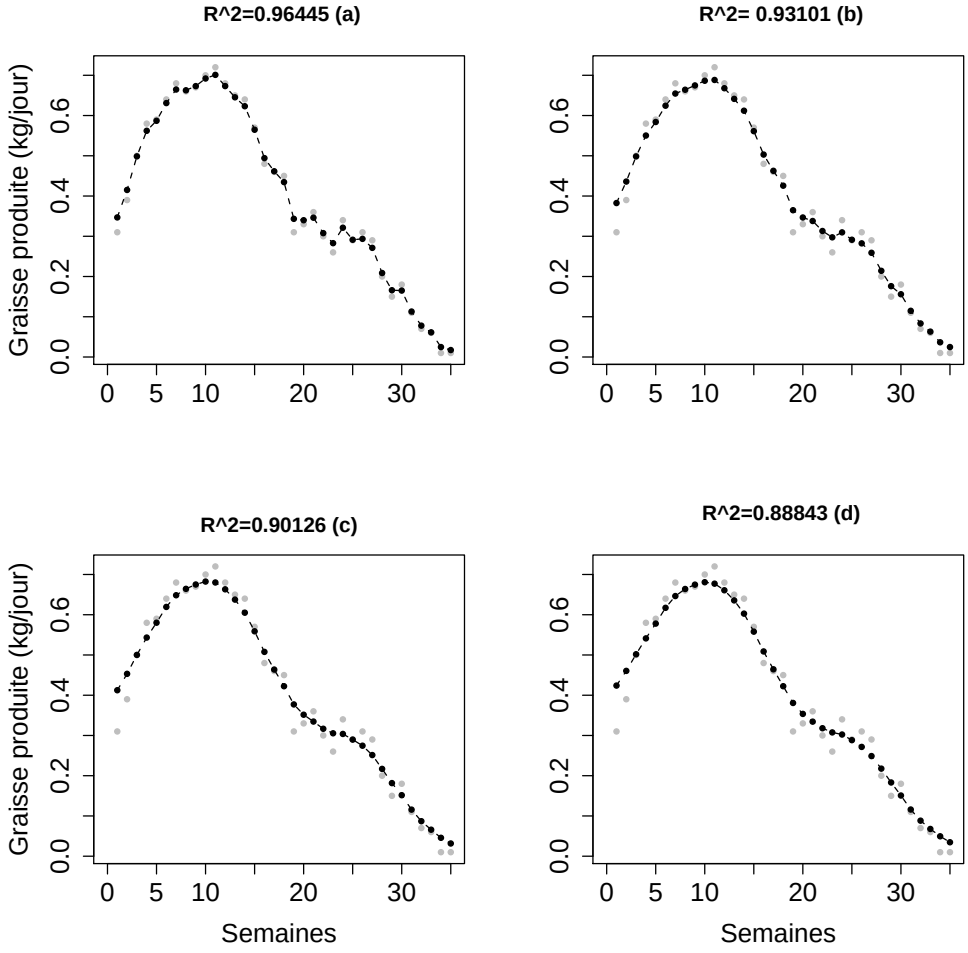

FIG. 4.3 – Suite et fin de Figure 4.2 pour les estimateurs à noyaux discrets triangulaires avec $a = 4$ et $h \in \{0.1(a), 0.3(b), 0.7(c), 1(d)\}$

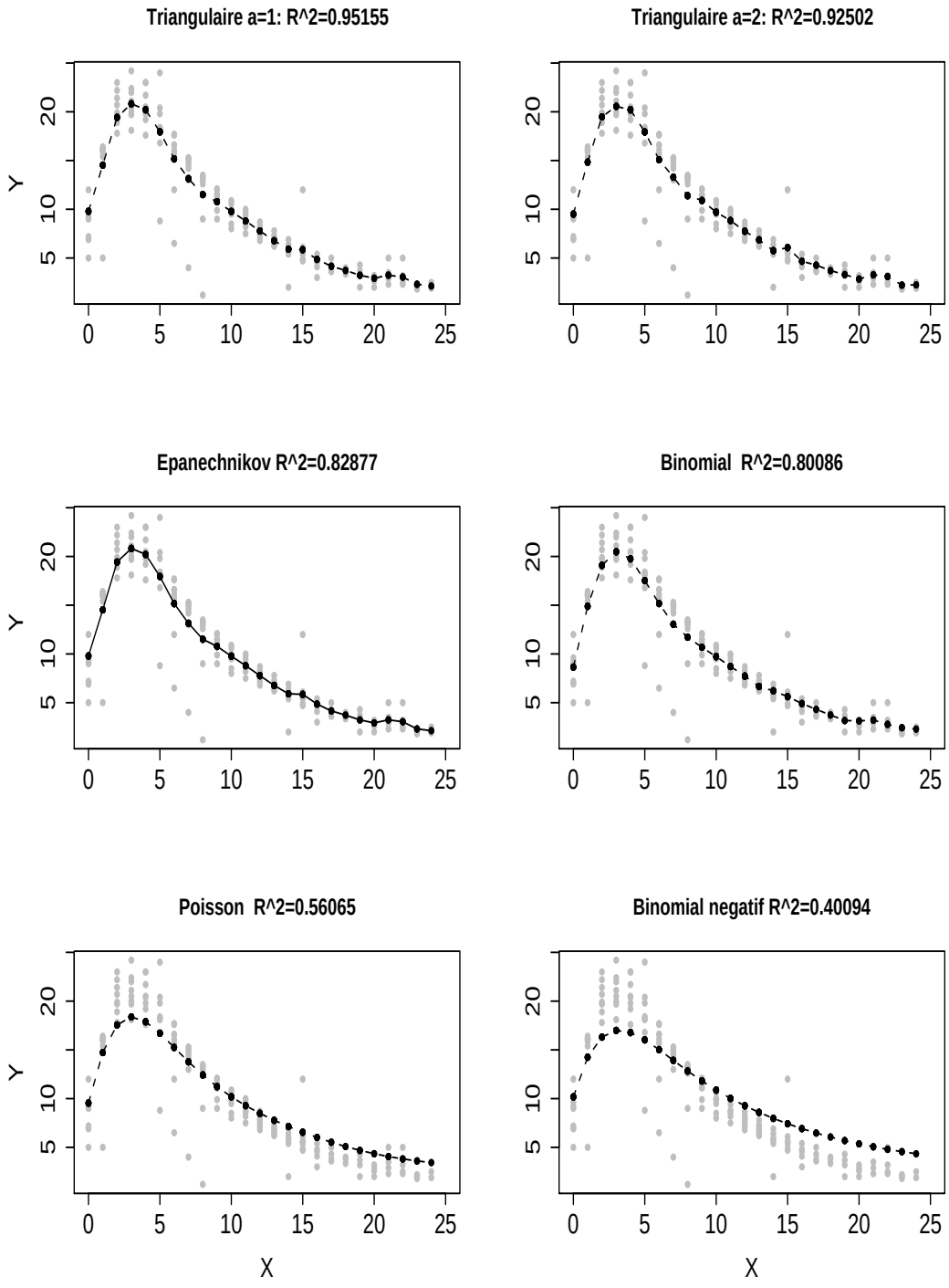

FIG. 4.4 – Régressions sur les données de vente Table 4.1 par les estimateurs à noyaux discrets et continu

Données	Noyau	Triang.1	Triang.2	*Epanech.*	Binomial	Poisson	Bin. nég.
Complètes	h_{cv}	0.558	0.132	2.427	0.064	0.206	0.327
	R^2	95.155	92.502	82.877	80.086	56.065	40.094
Incomplètes (sans *)	h_{cv}	0.170	0.052	2.121	0.078	0.209	0.327
	R^2	94.075	92.922	93.915	89.151	63.159	45.230

TAB. 4.6 – Coefficient de détermination (en %) des régressions sur les données de vente (Table 4.1) par les estimateurs à noyaux discrets et continu

l'ordre des performances selon les différents noyaux est le même.

La valeur du R^2 pour le noyau continu d'Epanechnikov est proche de celle des noyaux associés discrets triangulaires, mais ces derniers sont plus appropriés pour la structure de ce type de données.

4.4 Conclusion

Dans ce dernier chapitre, nous avons introduit un estimateur approprié de fonction discrète de régression et performant pour une variable explicative de dénombrement (4.1). Les noyaux (associés) discrets ont l'avantage d'être faciles à mettre en oeuvre et directement applicables sans aucune transformation préalable. Selon les noyaux (associés) discrets, particulièrement le binomial et le triangulaire avec un petit bras, les simulations réalisées et les deux jeux de données utilisés montrent l'utilité de la méthode proposée. Cette méthode est aussi compétitive, si ce n'est meilleure, que la régression par le noyau continu optimal. Bien qu'il n'y ait pas encore de noyau associé discret optimal, nous obtenons de bons ajustements. Ceci s'explique d'une part par la variance faible des noyaux associés discrets binomial et triangulaire avec de petit bras et, d'autre part, par le support fini de ces noyaux discrets. Une autre raison est suggérée par la bonne approximation entre les noyaux discrets et le noyau gaussien qui a l'avantage d'être le meilleur dans la régression non-paramétrique à noyau continu symétrique.

Des extensions évidentes et pratiques du modèle (4.1) sont envisageables pour plusieurs régresseurs. Soit un échantillon aléatoire $X_1, X_2, \ldots, X_n$ à valeurs dans $\mathbb{N}^d$ tel que $X_i = (X_{i1}, X_{i2}, \ldots, X_{id})^T$, $i = 1, 2, \ldots, n$, et Y une variable aléatoire réelle. L'estimateur de régression multidimensionnelle est défini à travers un produit de noyaux

(associés) discrets $K^j_{x_j,h_j}$ tels que

$$\widehat{m}_n(\underline{x}) = \frac{\sum_{i=1}^n Y_i \prod_{j=1}^p K^j_{x_j,h_j}(X_{ij})}{\sum_{i=1}^n \prod_{j=1}^p K^j_{x_j,h_j}(X_{ij})},$$

avec $\underline{x} = (x_1, \ldots, x_d)^T$. Les noyaux (associés) discrets $K^j_{x_j,h_j}$ peuvent être de même type ou de types différents. Une autre possibilité concerne des régresseurs à la fois discrets et continus ; ceci consiste à faire une régression non-paramétrique utilisant un produit des noyaux (associés) univariés discrets et continus. On peut améliorer aussi par une régression semi-paramétrique où le modèle de la variable endogène y_i sur le vecteur $p \times 1$ des régresseurs $x_i = (x_{1i}, \ldots, x_{pi})^T$ est formé d'une fonction paramétrique $g(x_i, \beta)$ et d'un facteur non-paramétrique $m(x_i)$ tels que

$$y_i = g(x_i, \beta)m(x_i) + e_i,$$

pour $i = 1, 2, \ldots, n$; voir, par exemple, Glad (1998). Enfin, on peut envisager une application dans l'estimation non-paramétrique d'une fonction discrète de hasard (voir Tutz & Pritscher, 1996).

Conclusion générale et perspectives

Ces travaux ont permis de mettre en place une procédure d'estimation non-paramétrique par la méthode des noyaux (associés) discrets pour des fonctions discrètes liées aux données de dénombrement.

Nous avons précisement défini la notion de noyau associé discret à partir d'une loi de probabilité discrète. Cette définition s'étend facilement aussi bien aux noyaux associés continus que discrets catégoriels (voir Annexe A). De là, nous avons construit une classe de noyaux discrets standards à partir des lois classiques de Poisson, binomiale et binomiale négative. Toutefois, certaines propriétés asymptotiques de ces noyaux discrets restent à améliorer, notamment celles des risques quadratiques intégrés des estimateurs correspondants qui ne convergent pas vers 0. Parmi les noyaux discrets standards, notre noyau binomial est celui dont l'estimateur associé donne les meilleurs ajustements notamment pour des échantillons de petites tailles. Ensuite, nous avons amélioré les estimateurs à noyaux discrets standards en introduisant des noyaux associés discrets triangulaires ayant de meilleures propriétés asymptotiques dont cette fois-ci la convergence des risques quadratiques intégrés des estimateurs correspondants vers 0. Pour des échantillons de grande taille, les estimateurs à noyaux discrets triangulaires sont plus performants que ceux à noyau binomial. Tandis que pour des échantillons de tailles petites ou modérées, notre estimateur à noyau binomial reste très compétitif. Par rapport aux estimateurs empirique et à noyau discret catégoriel d'Aitchison & Aitken (1976) dont les risques quadratiques intégrés convergent également vers 0, les estimateurs à noyaux discrets triangulaires sont aussi plus performants. Nous avons d'abord utilisé la méthode des noyaux associés discrets mise en place pour estimer la fonction de masse de probabilité f sur $\mathbb{N}$. Ensuite, nous avons appliqué cette méthode pour l'estimation non-paramétrique de la fonction discrète de poids ω de la loi de Poisson pondérée ainsi que de la loi binomiale pondérée. Enfin, une application a été présentée pour estimer la fonction discrète de régression m sur $\mathbb{N}$.

Cette méthode non-paramétrique utilisant les noyaux associés discrets est appropriée pour toute structure discrète ou discrétisée. Elle trouve son usage naturel et légitime sur les distributions de données de dénombrement par rapport aux méthodes analogues continues. De plus, elle est simple à mettre en oeuvre et plus performante que certaines méthodes utilisant des noyaux continus. Enfin, dans les cas d'échantillons de tailles petites et modérées, elle fournit des résultats probants et une alternative adéquate à l'estimateur empirique ou naïf.

Les travaux réalisés offrent plusieurs perspectives. En effet, des ordres de performance des noyaux associés discrets ont pu être établis. Ceci a été fait, d'une part, à l'intérieur de la classe des noyaux discrets standards et, d'autre part, dans celle des noyaux associés discrets triangulaires. Cependant, il reste à déterminer un noyau associé discret optimal parmi l'ensemble des noyaux discrets comme l'est celui d'Epanechnikov (1969) pour le cas continu et symétrique.

De plus, il faut envisager des généralisations au cas multivarié des différents estimateurs à noyau discret proposés ici. Pour des travaux dans le cas continu, on peut se référer à Abdous & Berlinet (1998) et leurs références. En effet, soit un échantillon de vecteurs aléatoires $X_1, X_2, \ldots, X_n$ de fonction de masse de probabilité f inconnue défini sur $\aleph = \mathbb{N}^d$. L'estimateur à noyau discret $\widetilde{f}_n$ de f admet une version multidimensionnelle qui se présente de manière générale par :

$$\widetilde{f}_n\left(\underline{x}\right) = \frac{1}{n}\sum_{i=1}^{n} K_{\underline{x},H}\left(X_i\right),$$

où $\underline{x} = (x_1, \ldots, x_d)^T$ est le vecteur cible, $X_i = (X_{i1}, X_{i2}, \ldots, X_{id})^T$, $i = 1, 2, \ldots, n$ et H est la matrice inversible de variance-covariance des fenêtres h de dimension $d \times d$ telle que

$$H = \begin{bmatrix} h_{11} & \ldots & h_{1d} \\ h_{21} & \ldots & h_{2d} \\ \ldots & h_{jj} & \ldots \\ h_{d1} & \ldots & h_{dd} \end{bmatrix}.$$

La fonction $K_{\underline{x},H}$ est le noyau associé discret sur $\aleph_{\underline{x}} \subseteq \mathbb{N}^d$. Plus simplement, nous pouvons présenter l'estimateur $\widetilde{f}_n$ en utilisant un produit de noyaux associés univariés discrets $K^j_{x_j,h_j}$, $j = 1, 2, \ldots, d$, tel que

$$\widetilde{f}_n\left(\underline{x}\right) = \frac{1}{n}\sum_{i=1}^{n}\left\{\prod_{j=1}^{d} K^j_{x_j,h_j}\left(X_{ij}\right)\right\}.$$

Cet estimateur multivarié à noyau discret est une version discrète de celui donné par Duong (2004) dans le cas continu. Selon les types de données dont on disposera, il sera nécessaire d'avoir recours à un noyau associé multivarié combinant les noyaux associés univariés discrets et continus.

Bibliographie

[1] Abdous, B. & Berlinet, A. (1998). Pointwise improvement of multivariate kernel density estimates. *Journal of Multivariate Analysis* **65**, 109–128.

[2] Agarwal, R.P. & Bohner, M. (1999). Basic calculus on time scales and some of its applications. *Results in Mathematics* **35**, 3–22.

[3] Aitchison, J. & Aitken, C.G.G. (1976). Multivariate binary discrimination by the kernel method. *Biometrika* **63**, 413–420.

[4] Alanko, T. & Lemmens, P.H. (1996). Response effects in consumption surveys : an application of the beta-binomial model to self-reported drinking frequencies. *Journal of Official Statistics* **12**, 253–273.

[5] Balakrishnan, N. & Kozubowski, T.J. (2008). A class of weighted Poisson processes. *Statistics & Probability Letters* **78**, 2346–2352.

[6] 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.

[7] Böhning, D. (2000). *Computer-assisted analysis of mixtures and applications. Meta-analysis, disease mapping and others*. Monographs on Statistics and Applied Probability 81. Chapman & Hall.

[8] Bosq, D. & Lecoutre, J.P. (1987). *Théorie de l'Estimation Fonctionnelle*. Economica, Paris.

[9] Bouvier, A., George, M. & Le Lionnais, F. (2005). *Dictionnaire de Mathématiques*. Presses Universitaires de France, Paris.

[10] Bowman, A. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika* **71**, 352–360.

[11] Cardot, H., Ferraty, F. & Sarda, P. (1999). Functional linear model. *Statistics & Probability Letters* **45**, 11–22.

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

[13] Chaubey Y.P., Sen A. & Sen K.P. (2007). A new smooth density estimator for non-negative random variables. Technical Report No. 01/07, Concordia University, Montréal.

[14] Chen, S.X. (1999). Beta kernels estimators for density functions. *Computational Statistics & Data Analysis* **31**, 131–145.

[15] Chen, S.X. (2000a). Gamma kernel estimators for density functions. *Annals of the Institute of Statistical Mathematics* **52**, 471–480.

[16] Chen, S.X. (2000b). Beta kernel smoothers for regression curves. *Statistica Sinica* **10**, 73–91.

[17] Collomb, G. (1981). Estimation Non-paramétrique de la Régression : Revue Bibliographique. *International Statistical Review* **49**, 75–93.

[18] Devroye, L. (1987). *A Course in Density Estimation*. Birkhäuser, Boston.

[19] Duong, T. (2004). *Bandwidth selectors for multivariate kernel density estimation*. Thèse pour le grade de Docteur d'Université de Western, Australie.

[20] Epanechnikov, V. (1969). Nonparametric estimates of a multivariate probability density. *Theory of Probability and its Applications* **14**, 153–158.

[21] Feller, W. (1966). *An Introduction to Probability Theory and Its Applications* (2nd ed.). John Wiley & Sons, New York.

[22] Ferraty, F. & Vieu, P. (2006). *Nonparametric Functional Data Analysis : Theory and Practice*. Springer, Berlin.

[23] Gasser, Th. & Müller, H.-G. (1979). Kernel estimation of regression functions. In : *Smoothing Techniques for Curve Estimation* (Edited by Th. Gasser and M. Rosenblatt). pp. 23-68. Springer, Heideberg.

[24] Glad, I.K. (1998). A note on unconditional properties of a parametrically guided Nadaraya-Watson estimator. *Statistics & Probability Letters* **37**, 101–108.

[25] 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. *Journal of the Royal Statistical Society Ser. A* **83**, 255–279.

[26] Gupta, R.C. & Ong, S.H. (2005). Analysis of long-tailed count data by Poisson mixtures. *Communications in Statistics - Theory and Methods* **34**, 557–573.

[27] Hall, P. (1987). On Kullback-Leibler loss and density estimation. *The Annals of Statistics* **15**, 1491–1519.

[28] Hardle, W. & Marron, J.S. (1985). Optimal bandwidth selection in nonparametric regression function estimation. *The Annals of Statistics* **13**, 1465–1481.

[29] Hille, E. (1948). *Functional Analysis and Semi-Groups*. American Statistical Society Colloquium, New York.

[30] Hjort, N.L. & Glad, I.K. (1995). Nonparametric density estimation with a parametric start. *The Annals of Statistics* **24**, 882–904.

[31] Izenman, A.J. (1991). Recent developments in nonparametric density estimation. *Journal of the American Statistical Association* **86**, 205–224.

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

[33] Jones, M.C. (1992). Estimating densities, quantiles, quantile densities and density quantiles. *Annals of the Institute of Statistical Mathematics* **44**, 721–727.

[34] Karlis, D. & Ntzoufras, L. (2003). Analysis of sport data by using bivariate Poisson models. *The Statistician* **52**, 381–393.

[35] Kokonendji, C.C., Demétrio, C.G.B. & Zocchi, S.S. (2007a). On Hinde-Demétrio regression models for oversdispersed count data. *Statistical Methodology* **4**, 277–291.

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

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

[38] Lejeune, M. & Sarda, P. (1992). Smooth estimation of distribution and density function. *Computational Statistics & Data Analysis* **14**, 457–471.

[39] Li, Q. & Racine, J.S. (2007). *Nonparametric Econometrics : Theory and Practice*. Princeton University Press, Princeton and Oxford.

[40] Mack, Y.P. & Müller, H.-G. (1989). Derivative estimation in nonparametric regression with random predictor variables. *Sankhyā* **A51**, 59–72.

[41] Marron, J.S. (1987). A comparison of cross-validation techniques in density estimation. *The Annals of Statistics* **15**, 152–162.

[42] Marron, J.S. & Padgett, W.J. (1987). Asymptotically optimal bandwidth selection for kernel density estimators from randomly right censored samples. *The Annals of Statistics* **15**, 1520–1535.

[43] Marsh, L. C. & Mukhopadhyay, K. (1999). Discrete Poisson kernel density estimation with an application to wildcat coal strikes. *Applied Economics Letters* **6**, 393–396.

[44] McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Models* (2nd ed). Chapman & Hall, London.

[45] McCulloch, C.E. (2001). *An Introduction to Generalized Linear Mixed Models*. 46a Reunião Anual da RBRAS - 9o SEAGRO, University of São Paulo - ESALQ, Piracicaba.

[46] Michels, P. (1992). Asymmetric kernel functions in non-parametric regression analysis and prediction. *The Statistician* **41**, 439–454.

[47] Mizère, D. (2006). *Contributions à la Modélisation et à l'Analyse Statistique des Données de Dénombrement*. Thèse pour le grade de Docteur d'Université de Pau et des Pays de l'Adour, soutenue le 26.01.2006, Pau.

[48] 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. *Revue de Statistique Appliquée* **54**, 61-84.

[49] Nadaraya, E.A. (1964). On estimating regression. *Theory of Probability and its Applications* **9**, 141–142.

[50] Nikoloulopoulos, A.K. & Karlis, D. (2008). On modeling count data : a comparison of some well-known discrete distributions. *Journal of Statistical Computation and Simulation* **78**, 437-457.

[51] Okumara, H. & Naito, K. (2004). Weighted kernel estimators in nonparametric binomial regression. *Journal of Nonparametric Statistics* **16**, 39–62.

[52] Okumara, H. & Naito, K. (2006a). Bandwidth selection for kernel binomial régression. *Journal of Nonparametric Statistics* **18**, 343–356.

[53] Okumara, H. & Naito, K. (2006b). Non-parametric kernel regression for multinomial data. *Journal of Multivariate Analysis* **97**, 2009–2022

[54] Ouyang, D., Li, Q. & Racine, J. (2006). Cross-validation and the estimation of probability distributions with categorical data. *Journal of Nonparametric Statistics* **18**, 69–100.

[55] Patil, P.N., Wells, M.T. & Marron, J.S. (1994). Some heuristics of kernel-based estimators of ratio functions. *Journal of Nonparametric Statistics* **4**, 203–209.

[56] Parzen, E. (1962). On estimation of a probability density function and mode. *Annals of Mathematical Statistics* **33**, 1065–1076.

[57] Puig, P. (2003). Characterizing additively closed discrete models by a property of their maximum likelihood estimators with application to generalized Hermite distributions. *Journal of the American Statistical Association* **96**, 687–692.

[58] Puig, P. & Valero, J. (2006). Count data distributions : some characterizations with applications. *Journal of the American Statistical Association* **101**, 332–340.

[59] Rodríguez-Avi, J., Conde-Sánchez, A., Sáez-Castillo, A.J. & Olmo-Jiménez, M.J. (2007). A generalization of the beta-binomial distribution. *Journal of the Royal Statistical Society Ser. C* **56**, 51–61.

[60] Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics* **27**, 832–837.

[61] Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics* **9**, 65–78.

[62] Scaillet, O. (2004). Density estimation using inverse and reciprocal inverse Gaussian kernels. *Journal of Nonparametric Statistics* **16**, 217–226.

[63] Schumaker, L.L. (1981). *Spline Functions : Basic Theory*. Wiley, New York.

[64] Scott, D.W. (1992). *Multivariate Density Estimation - Theory, Practice, and Visualization*. Wiley, New York.

[65] Shmueli, G., Minka, T.P., Kadane, J.P., Borle, S. & Boatwright, P. (2005). A useful distribution for fitting discrete data : revival of the Conway–Maxwell–Poisson distribution. *Journal of the Royal Statistical Society Ser. C* **54**, 127–142.

[66] Simonoff, J.S. (1996). *Smoothing Methods in Statistics*. Springer, New York.

[67] 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.

[68] Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, London.

[69] Stone, C. J. (1984). An asymptotically optimal window selection rule for kernel density estimates. *The Annals of Statistitics* **12**, 1285–1297.

[70] Tsybakov, A.B. (2004). *Introduction à l'Estimation Non-Paramétrique*. Springer, Paris.

[71] Tutz, G. & Pritscher, L. (1996). Nonparametric estimation of discrete hazard functions. *Lifetime Data Analysis* **2**, 291–308.

[72] Watson, G.S. (1964). Smooth regression analysis. *Sankhyā Ser. A* **26**, 359–372.

[73] White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50**, 1–26.

[74] Yang, Y. (2007). Consistency of cross validation for comparing regression procedures. *The Annals of Statistics* **6**, 2450–2473.

Annexe A

Noyaux associés continus et discret catégoriel

Sommaire :

A.1 Définition

D'après les études faites dans ce travail, nous pouvons unifier la définition d'un «noyau associé» à partir d'une loi de probabilité quelconque (continue ou discrète) afin d'effectuer le lissage d'une fonction inconnue f sur $\aleph \subseteq \mathbb{R}$.

Définition A.1 *Soit $x \in \aleph$ et $h > 0$. On appelle «noyau associé» $K_{x,h}(\cdot)$ toute densité ou fonction de masse de probabilité liée à une variable aléatoire $\mathcal{K}_{x,h}$, de support $\aleph_x$ contenant au moins x et indépendant de h, tels que*

1. $\bigcup_{x \in \aleph} \aleph_x \supseteq \aleph$,
2. $\lim_{h \to 0} \mathbb{E}(\mathcal{K}_{x,h}) = x$,
3. $\text{var}(\mathcal{K}_{x,h}) < +\infty$,
4. $\lim_{h \to 0} \text{var}(\mathcal{K}_{x,h}) = 0$.

Nous présentons quelques exemples des noyaux continus symétriques (*e.g.* normal) et asymétriques (*e.g.* gamma, gaussien inverse réciproque) ainsi que discret catégoriel (Aitchison & Aitken, 1976) comme des noyaux associés. En fait, nous vérifions que

toutes les conditions (1 à 4) de la définition précédente sont satisfaites. Nous illustrons graphiquement la forme des différents types de noyaux associés continus à travers des exemples ainsi que celle du noyau associé discret catégoriel. Mais commençons par quelques propriétés des estimateurs à noyaux continus.

A.2 Propriétés de l'estimateur à noyau continu

Dans cette section, nous étudions brièvement les estimateurs à noyaux associés continus symétriques et asymétriques. Nous donnons quelques unes de leurs propriétés.

Soit $X_1, X_2, \ldots, X_n$ un n-échantillon aléatoire indépendant et identiquement distribué (i.i.d.) de densité de probabilité inconnue f sur $\aleph \subseteq \mathbb{R}$. De manière générale, l'estimateur à noyau continu de f est défini par

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

où $h > 0$ est la fenêtre de lissage et $K_{x,h}$ la fonction noyau associé continu.

Le résultat suivant assure la convergence ponctuelle en moyenne de l'estimateur à noyau continu. Ce résultat est similaire à celui de la Propostion 1.2.4 du Chapitre 1 et a aussi été énoncé par Chaubey *et al.* (2007).

Proposition A.1 *Soit $f : \aleph \to \mathbb{R}$ une fonction continue et bornée et soit $x \in \aleph$ fixé. Si $\widetilde{f}_n(x)$ est un estimateur de $f(x)$ à noyau associé continu $K_{x,h}$ sur $\aleph_x$ alors*

$$\mathbb{E}\{\widetilde{f}_n(x)\} = \int_{\aleph\cap\aleph_x} f(t)K_{x,h}(t)dt \to f(x) \quad \textit{quand} \quad h = h(n) \to 0 \quad \textit{si} \quad n \to +\infty.$$

DÉMONSTRATION : Puisque $f(y)K_{x,h}(y) = 0$ pour tout $y \notin \aleph \cap \aleph_x$, on suppose $\aleph_x \subseteq \aleph$ pour tout $x \in \aleph$. Pour tout $\delta > 0$, on notera $\aleph_{x,\delta} = \{t \in \aleph_x : |t - x| < \delta\}$ et $\overline{\aleph}_{x,\delta} = \{t \in \aleph_x : |t - x| \geq \delta\}$ son complémentaire. Nous pouvons écrire $f(x) = f(x)\int_{\aleph\cap\aleph_x} K_{x,h}(t)dt$. De là, nous exprimons :

$$\begin{aligned}\left|\mathbb{E}\left\{\widetilde{f}_n(x)\right\} - f(x)\right| &= \left|\int_{\aleph\cap\aleph_x} \{f(t) - f(x)\} K_{x,h}(t)dt\right| \\ &\leq \int_{t\in\aleph_{x,\delta}} |f(t) - f(x)|K_{x,h}(t)dt + \int_{t\in\overline{\aleph}_{x,\delta}} |f(t) - f(x)|K_{x,h}(t)dt,\end{aligned}$$

où $\delta > 0$ est une constante arbitraire. Dans le but de majorer le premier terme de l'inégalité précédente, nous utilisons la continuité de la fonction f :

$$\forall \epsilon > 0, \exists\, \delta > 0, \forall t : |t - x| < \delta \Rightarrow |f(t) - f(x)| < \epsilon.$$

Ainsi, pour tout $\epsilon > 0$, il existe un $\delta > 0$ pour lequel

$$\int_{t \in \aleph_{x,\delta}} |f(t) - f(x)| K_{x,h}(t) dt \quad \leq \quad \epsilon. \tag{A.1}$$

Puis, comme f est bornée, il existe $M > 0$ tel que pour tout $t \in \aleph$, $|f(t)| < M$. Alors, en utilisant l'inégalité de Tchebychev-Markov, nous obtenons successivement :

$$\begin{aligned} \int_{t \in \overline{\aleph}_{x,\delta}} |f(t) - f(x)| K_{x,h}(t) dt &\leq 2M \int_{|t-x|>\delta} K_{x,h}(t) dt \\ &\leq \frac{2M}{\delta^2} \mathbb{E}\{(\mathcal{K}_{x,h} - x)^2\} \\ &\leq \frac{2M}{\delta^2} \left[\text{var}(\mathcal{K}_{x,h}) + \{\mathbb{E}(\mathcal{K}_{x,h}) - x\}^2 \right]. \end{aligned} \tag{A.2}$$

Sous les conditions 3 et 5 de Définition A.1 du noyau associé continu, les inégalités (A.1) et (A.2) conduisent au résultat final. ■

A.2.1 Cas continu symétrique

Soit $X_1, X_2, \ldots, X_n$ un n-échantillon aléatoire i.i.d. de densité de probabilité inconnue f sur l'ensemble des réels $\mathbb{R} = \aleph$. Les travaux de Rosenblatt (1956) puis de Parzen (1962) ont permis de définir l'estimateur à noyau continu de f par

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

avec

$$K_{x,h}(\cdot) = \frac{1}{h} K\left(\frac{x - \cdot}{h} \right).$$

Le paramètre $h > 0$ est la fenêtre de lissage, la fonction continue noyau K est telle que $K(t) \geq 0$, $\int_{\mathbb{R}} K(t) dt = 1$ et $K(x) = K(-x)$, et la fonction $K_{x,h}$ est donc le noyau associé. Notons que la fonction continue noyau K vérifie généralement les conditions suivantes :

(i) $\int_{\mathbb{R}} tK(t) dt = 0$,

(ii) $\int_{\mathbb{R}} K^2(t) dt < +\infty$,

(iii) $\int_{\mathbb{R}} t^2 K(t) dt < +\infty$.

Ces conditions sont nécessaires pour les calculs ultérieurs des propriétés de biais et de variance de l'estimateur $\widetilde{f}_n(x)$ sous de bonnes hypothèses (de régularité) de f.

Dans toute la suite de cette section, nous supposons que les dérivées d'ordre 1 et 2 de f existent et admettent une intégrale finie sur le support $\mathbb{R}$. Nous rappelons d'abord

que l'estimateur à noyau continu symétrique est une densité de probabilité. En effet, en posant le changement de variable $t = (x - X_1)/h$ et donc $dx = hdt$, nous obtenons :

$$\begin{aligned}\int_{\mathbb{R}} \widetilde{f}_n(x)dx &= \int_{\mathbb{R}} \frac{1}{h} K\left(\frac{x - X_1}{h}\right) dx \\ &= \int_{\mathbb{R}} K(t)dt \\ &= 1.\end{aligned}$$

Ensuite, nous analysons les risques quadratiques ponctuel MSE et global $MISE$ de l'estimateur $\widetilde{f}_n$. Pour cela, la décomposition

$$\begin{aligned}MISE(n, h, K, f) &= \int_{\mathbb{R}} MSE(x)dx \\ &= \int_{\mathbb{R}} \text{var}\left\{\widetilde{f}_{n,h,K}(x)\right\} dx + \int_{\mathbb{R}} \text{biais}^2\left\{\widetilde{f}_{n,h,K}(x)\right\} dx\end{aligned}$$

nous conduit à examiner le biais et la variance ponctuels de $\widetilde{f}_n$.

Pour le biais ponctuel, en posant $-t = (x - x_1)/h$ puis en utilisant le développement limité de Taylor $f(x + ht) \doteq f(x) + htf'(x) + \{(ht)^2/2\}f''(x)$ et l'hypothèse (i), nous exprimons :

$$\begin{aligned}\text{biais}\{\widetilde{f}_n(x)\} &= \int_{\mathbb{R}} \frac{1}{h} K\left(\frac{x - x_1}{h}\right) f(x_1)\, dx_1 - f(x) \\ &= \int_{\mathbb{R}} K(-t)\, f(x + ht)\, dt - f(x) \\ &\doteq f(x) \int_{\mathbb{R}} K(t)dt + hf'(x) \int_{\mathbb{R}} tK(t)dt + \frac{1}{2}h^2 f''(x) \int_{\mathbb{R}} t^2 K(t)dt - f(x) \\ &= \frac{1}{2}h^2 f''(x) \int_{\mathbb{R}} t^2 K(t)\, dt,\end{aligned}$$

où $\doteq$ indique un équivalent asymptotique. L'hypothèse (iii) permet alors d'obtenir un biais qui soit fini.

En ce qui concerne la variance ponctuelle, nous posons de nouveau $-t = (x - x_1)/h$ et sous les hypothèses (ii) et $f(x) < f_{max}(x) < +\infty$, pour n grand, nous obtenons successivement :

$$\begin{aligned}\text{var}\left\{\widetilde{f}_n(x)\right\} &= \frac{1}{n} \int_{\mathbb{R}} \frac{1}{h^2} K^2\left(\frac{x - x_1}{h}\right) f(x_1)\, dx_1 - \frac{1}{n}\left\{\int_{\mathbb{R}} \frac{1}{h} K\left(\frac{x - x_1}{h}\right) f(x_1)\, dx_1\right\}^2 \\ &= \frac{1}{nh} \int_{\mathbb{R}} K^2(t) f(ht + x)\, dt - \frac{1}{n}\left[\text{biais}\left\{\widehat{f}_{n,h,K}(x)\right\} + f(x)\right]^2 \\ &= \frac{1}{nh} \int_{\mathbb{R}} K^2(t) f(ht + x)\, dt - \frac{1}{n}\left\{O\left(h^2\right) + f(x)\right\}^2 \\ &\doteq \frac{1}{nh} f(x) \int_{\mathbb{R}} K^2(t)dt.\end{aligned}$$

Ainsi, si de plus $nh \to +\infty$, on obtient var $\left\{\widetilde{f}_{n,h,K}(x)\right\} \to 0$.

Finalement, pour un point x fixé, l'approximation de l'erreur quadratique moyenne en un point x fixé s'ecrit :

$$AMSE\,(x) = \frac{1}{nh} f(x) \int_{\mathbb{R}} K^2(t)dt + \left\{\frac{1}{2}h^2 f''(x) \int_{\mathbb{R}} t^2 K\,(t)\,dt\right\}^2;$$

puis, celle du risque quadratique intégré est :

$$\begin{aligned} AMISE(n,h,K,f) &= \int_{\mathbb{R}} AMSE(x)dx \\ &= \frac{1}{nh}\int_{\mathbb{R}} K^2(t)dt + \frac{1}{4}h^4 \left\{\int_{\mathbb{R}} t^2 K(t)dt\right\}^2 \int_{\mathbb{R}} \{f''(x)\}^2 dx. \end{aligned}$$

a. Choix de noyau

Soient deux noyaux continus symétriques fixés K_1 et K_2. Le critère utilisé pour mesurer l'efficacité relative de ces deux noyaux est le suivant :

$$eff(K_1, K_2) = \frac{AMISE(K_1)}{AMISE(K_2)}.$$

Ainsi, l'efficacité d'un noyau continu symétrique K par rapport au noyau optimal d'Epanechnikov (1969) est donnée par

$$eff(K) = \frac{3}{5\sqrt{5}} \frac{1}{\sqrt{\int_{\mathbb{R}} t^2 K(t)dt \int_{\mathbb{R}} K^2(t)dt}} \leq 1.$$

Nous récapitulons dans la Table A.1 quelques exemples de noyaux continus symétriques et leurs valeurs d'efficacité correspondantes, lesquelles sont toutes très proches de 1 et rendent le choix d'un noyau continu symétrique moins important.

b. Choix de fenêtres

Le choix de la fenêtre de lissage h est primordial dans la procédure d'estimation à noyau continu symétrique. Pour un tel noyau fixé K et un échantillon de taille n, l'expression de la fenêtre optimale se détermine par une minimisation directe du $MISE$ par rapport à h. La résolution de

$$\frac{\partial}{\partial h} AMISE(h) = 0$$

conduit à

$$h_{id} = \frac{1}{n^{1/5}} \left[\frac{\int_{\mathbb{R}} K^2(t)dt}{\{\int_{\mathbb{R}} t^2 K(t)dt\}^2 \int_{\mathbb{R}} f''(x)dx}\right]^{1/5}.$$

Noyau	Densité	Support	Efficacité
Epanechnikov	$(3/4)(1-u^2)\mathbf{1}_{[-1,1]}$	$[-1,1]$	1.000
Biweight	$(15/16)(1-u^2)^2\mathbf{1}_{[-1,1]}$	$[-1,1]$	0.994
Triangulaire	$(1-\lvert u\rvert)\mathbf{1}_{[-1,1]}$	$[-1,1]$	0.986
Gaussien	$(1/\sqrt{2\pi})\exp(-u^2/2)$	$\mathbb{R}$	0.951
Uniforme	$(1/2)\mathbf{1}_{[-1,1]}(u)$	$[-1,1]$	0.930

TAB. A.1 – Exemple de noyaux continus symétriques

En particulier, pour le noyau d'Epanechnikov, nous avons

$$h_{id}(K_{Epanechn.}) = \left[\frac{15}{n\int_{\mathbb{R}}\{f''(x)\}^2dx}\right]^{1/5}.$$

Toutefois, la fenêtre idéale h_{id} dépend de la dérivée seconde de la densité inconnue f. Alors, plusieurs méthodes sont utilisés pour déterminer une valeur optimale de h. Nous pouvons citer la procédure de «Plug-in» dont l'idée de base est d'estimer la quantité inconnue $\int_{\mathbb{R}}\{f''(x)\}^2dx$ en supposant que f appartient à une famille de distributions paramétriques telle qu'une loi normale centrée et de variance σ^2. Nous pouvons appliquer aussi les méthodes de validation croisée par les moindres carrées et par le maximum de vraisemblance. Dans le cas discret, nous avons présenté ces deux dernières procédures dans la section 1.5 du Chapitre 1.

A.2.2 Cas continu asymétrique

Soit $X_1, X_2, \ldots, X_n$ un n-échantillon aléatoire i.i.d. de densité de probabilité inconnue f sur l'ensemble $\aleph \subseteq \mathbb{R}$. L'estimateur à noyau continu (symétrique) présenté dans la section précédente a été conçu à la base pour des densités à supports continus et non-bornés. Chen (1999, 2000a) et plus tard Scaillet (2004) introduisent des noyaux continus asymétriques. Ces derniers sont appropriés pour l'estimation d'une densité continue à support compact ou borné d'un seul côté tel que $\aleph = [a,b]$ avec $a \in \mathbb{R}$ et $b \in \bar{\mathbb{R}}$, par exemple $[0,1]$ ou $[0,+\infty[$. Pour $x \in \aleph$, les estimateurs à noyau associé

continu asymétrique sont aussi de forme

$$\widetilde{f}_n(x) = \frac{1}{n}\sum_{i=1}^{n} K_{x,h}\left(X_i\right) = \widetilde{f}_{n,h,K}(x),$$

où $h > 0$ est le paramètre de lissage et $K_{x,h}$ est le noyau associé continu. Mais il n'est pas possible de dégager une forme générale de $K_{x,h}$ comme dans le cas symétrique où l'on a $K_{x,h}(\cdot) = (1/h)K\{(x-\cdot)/h\}$. La cible x et le paramètre h entrent dans la définition intrinséque du noyau associé $K_{x,h}$. Ce dernier se définit en exprimant les paramètres du type de noyau K choisi en fonction du couple (x, h).

Nous examinons quelques propriétés de l'estimateur à noyau continu asymétrique. Elles sont les analogues continues des propriétés étudiées pour l'estimateur à noyau associé discret en Section 1.2 du Chapitre 1.

Soulignons déjà que la fonction $x \mapsto \widetilde{f}_n(x)$ n'est pas nécessairement une densité de probabilité. Toutefois, cette vérification qui est simple dans le cas continu symétrique n'est pas aisée dans le cas asymétrique continu ; voir la Section B.1 de l'Annexe B pour des noyaux asymétriques discrets. Il est alors nécessaire de normaliser par la constante $C = \int_{\aleph} \widetilde{f}_n(x)dx = C(h, K; X_1, \ldots, X_n)$ qui est positive et finie. Dans la suite, nous considérons que cet l'estimateur à noyau $\widetilde{f}_n$ est normalisé.

Nous exprimons tout d'abord la relation suivante qui est fondamentale pour les calculs à venir :

$$\begin{aligned}
\mathbb{E}\left\{\widetilde{f}_n(x)\right\} &= \int_{\aleph\cap\aleph_x} K_{x,h}(t)f(t)dt \\
&= \int_{\aleph\cap\aleph_x} f(t)K_{x,h}(t)dt \\
&= \mathbb{E}\left\{f(\mathcal{K}_{x,h})\right\}.
\end{aligned}$$

En réalisant un développement limité de Taylor de $f(\mathcal{K}_{x,h})$ au point moyen $\mathbb{E}(\mathcal{K}_{x,h}) = m_{x,h}$ puis en prenant l'espérance $\mathbb{E}\left\{f(\mathcal{K}_{x,h})\right\}$, nous obtenons

$$f(\mathcal{K}_{x,h}) \doteq f(m_{x,h}) + (\mathcal{K}_{x,h} - m_{x,h})f'(x) + \frac{1}{2}(\mathcal{K}_{x,h} - m_{x,h})^2 f''(x),$$

suivi de

$$\begin{aligned}
\mathbb{E}\left\{f(\mathcal{K}_{x,h})\right\} &\doteq f(m_{x,h}) + \frac{1}{2}\mathbb{E}\left\{(\mathcal{K}_{x,h} - m_{x,h})^2\right\} f''(x), \\
&= f\left\{\mathbb{E}(\mathcal{K}_{x,h})\right\} + \frac{1}{2}\text{var}(\mathcal{K}_{x,h})f''(x).
\end{aligned}$$

De là, les conditions 2 et 4 de la Définition A.1 du noyau associé permettent de retrouver le résultat de convergence ponctuelle en moyenne de $\widetilde{f}_n$ donné en Proposition A.1.

Nous utilisons les approximations précédentes du développement limité de Taylor pour obtenir le biais ponctuel par :

$$\begin{aligned} \text{biais}\left\{\widetilde{f}_n(x)\right\} &= \mathbb{E}\left\{\widetilde{f}_n(x)\right\} - f(x) \\ &\doteq f\left\{\mathbb{E}(\mathcal{K}_{x,h})\right\} - f(x) + \frac{1}{2}\text{var}(\mathcal{K}_{x,h}) f''(x). \end{aligned}$$

A travers la Proposition A.1, le biais tend vers 0 quand $h = h(n)$ tend vers 0 et n tend vers $+\infty$.

Pour la variance ponctuelle de l'estimateur, nous obtenons :

$$\begin{aligned} \text{var}\{\widetilde{f}_n(x)\} &= \frac{1}{n}\int_{\aleph\cap\aleph_x} K_{x,h}^2(t) f(t) dt - \frac{1}{n}\left\{\int_{\aleph\cap\aleph_x} K_{x,h}(t) f(t) dt\right\}^2 \\ &= \frac{1}{n}\int_{\aleph\cap\aleph_x} K_{x,h}^2(t) f(t) dt - \frac{1}{n}\left[\text{biais}\left\{\widetilde{f}_n(x)\right\} + f\,(x)\right]^2, \end{aligned}$$

laquelle $\text{var}\{\widetilde{f}_n(x)\}$ tend vers 0 quand $n^{-1}\int_{\aleph\cap\aleph_x} K_{x,h}^2(t) f(t) dt \to 0$ si $n \to +\infty$ et $h = h(n) \to 0$. Donc, nous pouvions déduire

$$\begin{aligned} MISE &= \int_{\aleph} \mathbb{E}\left\{\widetilde{f}_n(x) - f(x)\right\} dx, \\ &= \int_{\aleph} \text{var}\left\{\widetilde{f}_n(x)\right\} dx + \int_{\aleph} \text{biais}^2\left\{\widetilde{f}_n(x)\right\} dx \end{aligned}$$

et sa convergence vers 0 quand $n^{-1}\int_{\aleph}\left\{\int_{\aleph\cap\aleph_x} K_{x,h}^2(t) f(t) dt\right\} dx \to 0$ si $n \to +\infty$ et $h = h(n) \to 0$.

A.3 Exemples de noyaux associés continus symétriques

Dans la Table A.1, nous avons présenté des exemples de noyaux continus symétriques et de moyenne nulle. Le noyau associé à la loi gaussienne fait l'objet ici d'une étude particulière.

En effet, la loi gaussienne $\mathcal{N}(\mu, \sigma)$ de moyenne $\mu \in \mathbb{R}$ et d'écart type $\sigma > 0$ est symétrique et définie sur $\mathbb{R} = \aleph$ telle que sa densité s'écrit :

$$g_{N(\mu,\sigma)}(t) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(t-\mu)^2}{2\sigma^2}\right\}, \quad t \in \mathbb{R}.$$

On définit le noyau associé gaussien $K_{N(x,h)}$ sur $\mathbb{R} = \aleph_x$ par :

$$K_{N(x,h)}(t) = \frac{1}{h\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{t-x}{h}\right)^2\right\}, \quad t \in \mathbb{R}.$$

Tout d'abord, il est facile de vérifier que les conditions de Définition A.1 d'un noyau associé sont évidemment satisfaites. En effet, pour la condition 1, nous avons $\aleph = \aleph_x = \mathbb{R}$ alors la réunion sur x des $\aleph_x$ est $\mathbb{R}$. Puis, les conditions 2 à 4 sont satisfaites car le noyau associé gaussien $K_{N(x,h)}$ a pour moyenne $\mathbb{E}(\mathcal{K}_{N(x,h)}) = x$ et a une variance $\text{var}(\mathcal{K}_{N(x,h)}) = h^2 < +\infty$ ainsi $\text{var}(\mathcal{K}_{N(x,h)}) = h^2 \to 0$ quand $h \to 0$.

Ensuite, nous observons l'allure de la forme du noyau associé gaussien en faisant varier la cible x et le paramètre de lissage h. Pour h fixé, la Figure A.1 montre que le noyau associé gaussien est de forme identique en différents points x ; il réalise une simple translation d'un point à l'autre. Dans la Figure A.2, en fixant la cible x et en faisant varier la fenêtre, le noyau associé gaussien est toujours symétrique autour de x mais c'est sa variance qui change. D'une manière plus générale, les noyaux associés continus symétriques gardent la même forme quelque soit le point x où ils sont calculés et c'est leur variance qui change en fonction de la valeur de h.

Enfin, nous vérifions ici que l'estimateur à noyau associé gaussien est encore une densité de probabilité. En effet, comme la loi gaussienne est symétrique on peut permutter t et x. Ainsi, nous obtenons rapidement le résultat :

$$\begin{aligned}\int_{\mathbb{R}} \widetilde{f}_n(x)dx &= \int_{\mathbb{R}} \frac{1}{h\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{t-x}{h}\right)^2\right\} dx \\ &= 1.\end{aligned}$$

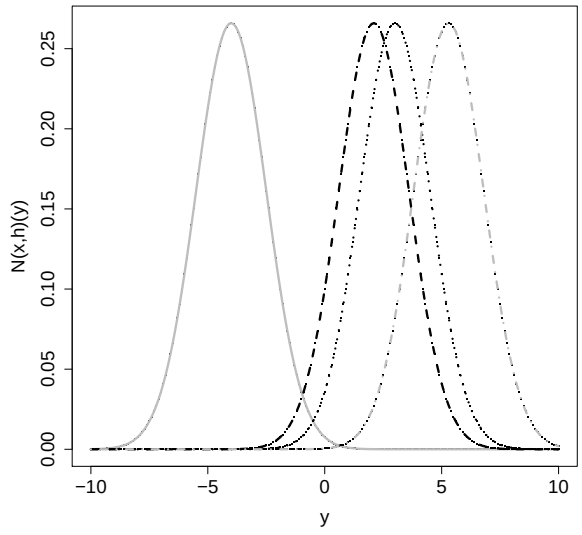

FIG. A.1 – Noyau associé gaussien pour $h = 1.5$ et $x \in \{-4, 2.1, 3, 5.3\}$

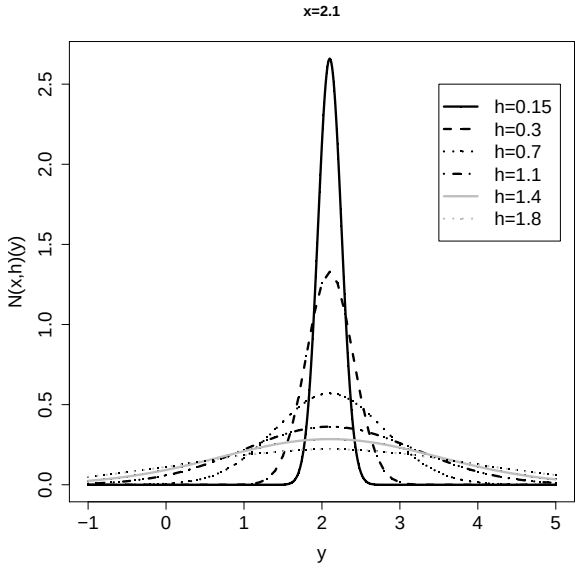

FIG. A.2 – Noyau associé gaussien en la cible $y = x = 2.1$ et selon des valeurs de h

A.4 Exemples de noyaux associés continus asymétriques

Dans cette section, nous présentons des noyaux associés continus asymétriques liées aux densités de probabilité gamma, bêta, gaussienne-inverse (IG) et gaussienne-inverse-réciproque (RIG) (voir Table A.2). En particulier, nous illustrons les noyaux associés asymétriques gamma et gaussien-inverse-réciproque. Précisons que dans les expressions des densités données dans la Table A.2 nous avons :

$$\Gamma(a) = \int_0^{+\infty} e^{-t} t^{a-1} dt \text{ et } B(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1} dt.$$

Chen (2000a) est le premier a utilisé la loi gamma pour construire des noyaux associés continus asymétriques. Pour $a = xh^{-1} + 1$ et $b = h$, on définit un noyau associé gamma par :

$$K_{G(xh^{-1}+1;h)}(t) = \frac{t^{xh^{-1}} e^{-th^{-1}}}{\Gamma(xh^{-1}+1) h^{xh^{-1}+1}}, \quad t \in \aleph_x = \mathbb{R}_+$$

de loi gamma de support $\aleph_x = \mathbb{R}_+$, d'espérance $\mathbb{E}(\mathcal{K}_{G(xh^{-1}+1;h)}) = x+h$ et de variance $\text{var}(\mathcal{K}_{G(xh^{-1}+1;h)}) = xh + h^2$. Ce noyau associé gamma $K_{G(xh^{-1}+1;h)}$ satisfait bien les conditions d'un noyau associé continu. En effet, la condition 1 est vérifiée car les supports sont tels que $\aleph = \aleph_x = \mathbb{R}_+$ et ils ne dépendent pas de x. Puis, pour la condition 2, nous avons $\mathbb{E}(\mathcal{K}_{G(xh^{-1}+1;h)}) \to x$ quand $h \to 0$. Enfin, pour les conditions

3 et 4, nous obtenons bien que var$(\mathcal{K}_{G(xh^{-1}+1;h)}) = xh + h^2$ est finie et tend vers 0 quand h tend vers 0.

Dans les Figures A.3 et A.4, nous présentons des allures de la courbe du noyau associé gamma. Nous fixons la fenêtre h et faisons varier la cible x, puis pour une cible fixée nous prenons des valeurs différentes du paramètre de lissage. Contrairement au cas continu symétrique, le changement de forme du noyau associé continu asymétrique est bien mis en évidence en fonction de la cible x puis de la fenêtre de lissage h.

Scaillet (2004) introduit le noyau associé gaussien-inverse-réciproque $K_{RIG((x-h)^{-1};h^{-1})}$ défini sur $\aleph_x =]0, +\infty[$ tels que :

$$K_{RIG((x-h)^{-1};h^{-1})}(t) = \frac{1}{\sqrt{2\pi ht}} \exp\left\{-\frac{x-h}{2h}\left(\frac{t}{x-h} - 2 + \frac{x-h}{t}\right)\right\}, \quad t > 0.$$

En utilisant les propriétés de l'espérance et de la variance de la loi à laquelle il est associé, nous vérifions aisément que $K_{RIG((x-h)^{-1};h^{-1})}$ est un noyau associé (Table A.3). Les Figures A.5 et A.6 montrent le comportement du noyau associé gaussien-inverse-réciproque quand pour h fixé la cible x varie, puis quand la fenêtre h varie à x fixé, respectivement. Observons que le noyau associé $K_{RIG((x-h)^{-1};h^{-1})}$ change de forme lorsque x varie de manière à ce que le mode soit toujours autour de la cible. Mais, lorsque h augmente la précision autour de la cible x diminue (Figure A.6).

Noyau	Support	Densité	Espérance	Variance
Gamma(a,b)	$[0, +\infty[$	$\{\Gamma(a)\}^{-1} b^{-a} t^{a-1} e^{-t/b}$	ab	ab^2
Bêta(a,b)	[0, 1]	$\{B(a,b)\}^{-1} t^{a-1}(1-t)^{b-1}$	$a(a+b)^{-1}$	$ab\left\{(a+b)^2(a+b+1)\right\}^{-1}$
IG(a,b)	$]0, +\infty[$	$\frac{\sqrt{b}}{\sqrt{2\pi t^3}} \exp\left\{-\frac{b}{2a}(\frac{t}{a} - 2 + \frac{a}{t})\right\}$	a	$a^3 b^{-1}$
RIG(a,b)	$]0, +\infty[$	$\frac{\sqrt{b}}{\sqrt{2\pi t}} \exp\{-\frac{b}{2a}(at - 2 + \frac{1}{at})\}$	$a^{-1} + b^{-1}$	$(ab)^{-1} + 2b^{-2}$

TAB. A.2 – Tableau récapitulatif des lois de probabilités continues asymétriques

A.5 Exemple de noyau associé discret catégoriel

Dans cette section, nous présentons un noyau associé discret adapté pour des données catégorielles qu'on peut supposer défini sur un ensemble discret fini $\aleph \subset \mathbb{R}$. Les

Type de noyau K	Noyau associé continu $K_{x,h}$	$\mathbb{E}(\mathcal{K}_{x,h})$	var$(\mathcal{K}_{x,h})$
Gamma(a,b)	$a = xh^{-1} + 1$ et $b = h$	$x + h$	$xh + h^2$
Bêta(a,b)	$a = xh^{-1} + 1$ et $b = (1-x)h^{-1} + 1$	$(x + h)(2h + 1)^{-1}$	$\frac{x(1-x)h+h^2+h^3}{(1+2h)^2(1+3h)}$
IG(a,b)	$a = x$ et $b = h^{-1}$	x	x^3h
RIG(a,b)	$a = (x-h)^{-1}$ et $b = h^{-1}$	x	$xh + h^2$

TAB. A.3 – Tableau récapitulatif des noyaux associés continus asymétriques

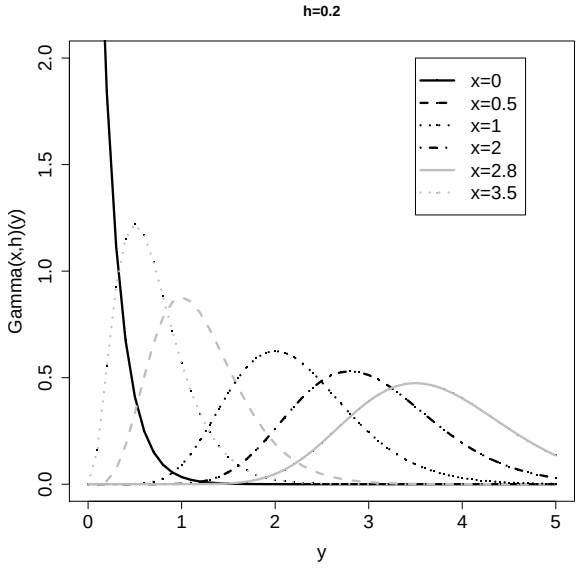

FIG. A.3 – Noyau associé gamma pour $h = 0.2$ et selon des valeurs de x

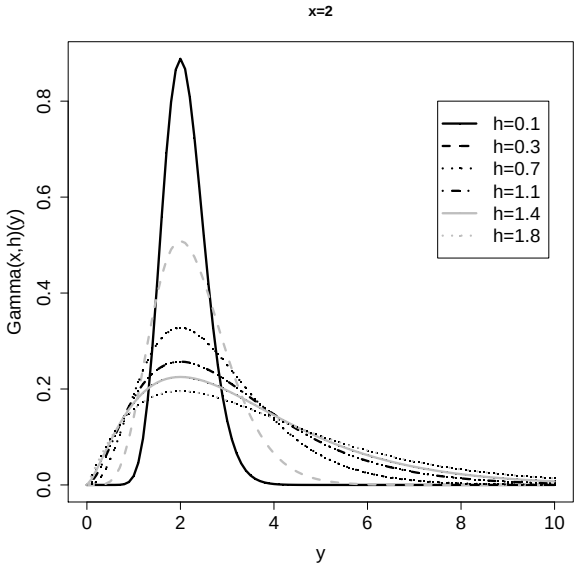

FIG. A.4 – Noyau associé gamma en la cible $y = x = 2$ et selon des valeurs de h

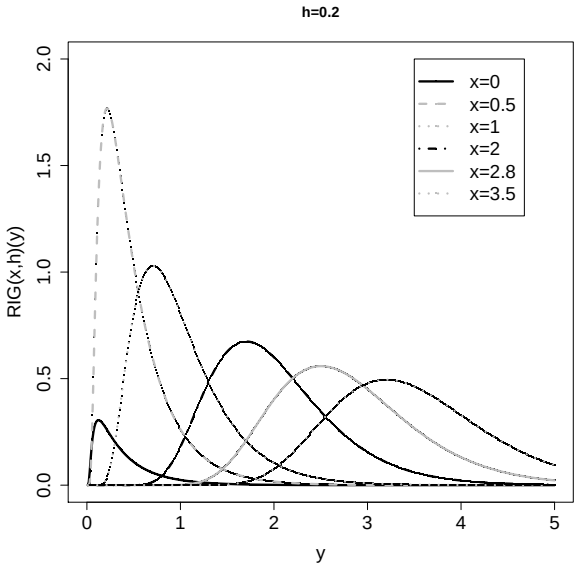

FIG. A.5 – Noyau associé gaussien-inverse-réciproque pour $h = 0.2$ et selon des valeurs de x (voir aussi Scaillet, 2004)

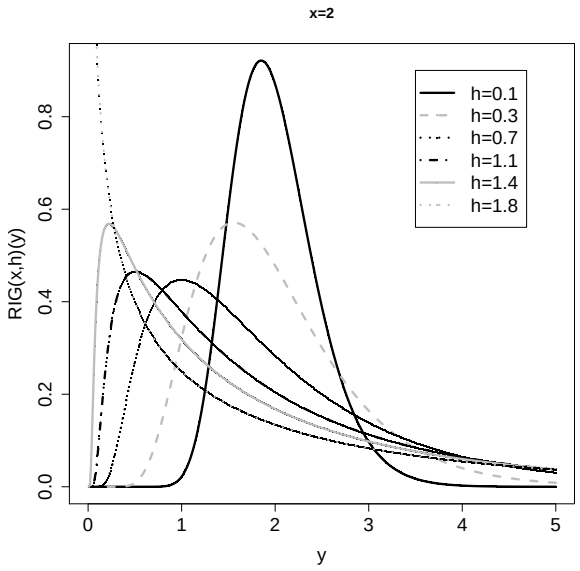

FIG. A.6 – Noyau associé gaussien-inverse-réciproque en la cible $y = x = 2$ et selon des valeurs de h (voir aussi Scaillet, 2004)

premiers travaux pour les estimateurs correspondants remontent à Aitchison & Aitken (1976). Pour $c \in \mathbb{N} \setminus \{0,1\}$, $c_0 \in \{0,1,\ldots,c-1\}$ et $\lambda \in]0,1]$, on notera $\mathcal{A}(c;c_0,\lambda)$ la loi d'Aitchison-Aitken*.

Définition A.2 *Soit $c \in \mathbb{N} \setminus \{0,1\}$, $c_0 \in \{0,1,\ldots,c-1\}$ et $\lambda \in]0,1]$. Une variable aléatoire discrète Y est de loi $\mathcal{A}(c;c_0,\lambda)$ de paramètres c le cardinal du support, c_0 le point de référence et λ la probabilité de succés, si pour tout y dans son support $\aleph_c = \{0,1,\ldots,c-1\}$ sa probabilité individuelle s'écrit :*

$$\Pr(Y=y) = (1-\lambda)1_{y=c_0} + \frac{\lambda}{c-1}1_{y\neq c_0}.$$

*Voir aussi le rapport de stage de Imen Ben Khalifa «Estimations non-paramétriques par noyaux associés et données de panel en marketing», réalisé de Février à Mai 2008 à l'Université de Pau et des Pays de l'Adour, sous la direction de M. Célestin C. Kokonendji.

L'espérance mathématique de cette loi s'exprime par

$$\begin{aligned}\mathbb{E}(Y) &= \sum_{y\in\{0,1,\ldots,c-1\}} \left\{ y(1-\lambda)1_{y=c_0} + \frac{y\lambda}{c-1}1_{y\neq c_0} \right\} \\ &= c_0(1-\lambda) + \frac{\lambda}{c-1}\left\{ \left(\sum_{y=0}^{c-1} y \right) - c_0 \right\} \\ &= c_0\left(1-\lambda-\frac{\lambda}{c-1}\right) + \frac{\lambda c}{2}.\end{aligned}$$

Pour la variance de $Y \sim \mathcal{A}(c; c_0, \lambda)$, nous obtenons :

$$\begin{aligned}\text{var}(Y) &= c_0^2(1-\lambda) + \frac{\lambda}{c-1}\left\{ \left(\sum_{y=0}^{c-1} y^2 \right) - c_0^2 \right\} - \left\{ c_0\left(1-\lambda-\frac{\lambda}{c-1}\right) + \frac{\lambda c}{2} \right\}^2 \\ &= c_0^2\left(1-\lambda-\frac{\lambda}{c-1}\right) + \frac{\lambda c(2c-1)}{6} - c_0^2\left(1-\lambda-\frac{\lambda}{c-1}\right)^2 - \frac{\lambda^2 c^2}{4} \\ &\quad -c_0\lambda c\left(1-\lambda-\frac{\lambda}{c-1}\right) \\ &= c_0^2\frac{c^2\lambda(1-\lambda)-\lambda c}{(c-1)^2} - c_0\frac{c^2\lambda(1-\lambda)-\lambda c}{c-1} + \frac{\lambda c}{2}\left(\frac{2c-1}{3} - \frac{\lambda c}{2}\right).\end{aligned}$$

Des représentations graphiques des lois d'Aitchison-Aitken sont illustrées à travers les allures du noyau associé correspondant que nous donnons dans la suite ; voir Figures A.7 et A.8. Une loi d'Aitchison-Aitken n'est autre qu'une loi uniforme discrète en dehors du point de référence.

REMARQUES : Soit $Y \sim \mathcal{A}(c; c_0, \lambda)$.

(a) Pour $c = 2$, le support se réduit à l'ensemble $\{0, 1\} = \aleph_2$. Nous retrouvons la loi de Bernoulli de paramètre λ ou $1-\lambda$.
(b) Lorsque $c \to +\infty$, nous avons le support $\aleph_{+\infty} = \mathbb{N}$.
(c) Si $\lambda = 0$, la loi d'Aitchison-Aitken correspond à une loi de Dirac et qui est donc indépendante de c. Tandis que si $\lambda = 1$, nous obtenons $\Pr(Y = y) = \{1/(c-1)\}1_{y\neq c_0}$.

Pour définir le noyau associé d'Aitchison-Aitken afin d'estimer des fonctions de masse de probabilité f sur le même support $\aleph = \{0, 1, \ldots, c\}$, on considère la cible $x = c_0$ dans $\aleph$ et le paramètre de lissage $h = \lambda$ dans $]0, 1]$. Le noyau $A_{c;x,h}$ est associé à la v.a. $\mathcal{A}_{c;x,h}$ de loi $\mathcal{A}(c; x, h)$ et de support $\aleph_c = \{0, 1, \ldots, c-1\} = \aleph$. L'espérance mathématique de $\mathcal{A}_{c;x,h}$ est

$$\mathbb{E}(\mathcal{A}_{c;x,h}) = x\left(1-h-\frac{h}{c-1}\right) + \frac{hc}{2},$$

et sa variance s'obtient par

$$\text{var}(\mathcal{A}_{c;x,h}) = x^2\frac{hc^2(1-h)-hc}{(c-1)^2} - x\frac{hc^2(1-h)-hc}{c-1} + \frac{hc}{2}\left(\frac{2c-1}{3} - \frac{hc}{2}\right). \quad \text{(A.3)}$$

Nous vérifions immédiatement les quatres conditions d'un noyau associé discret. En effet, la première condition $\cup_{x\in\aleph}\aleph_c \supseteq \aleph$ est satisfaite. La condition 2 vient de ce que $\mathbb{E}(\mathcal{A}_{c;x,h}) \to x$ quand $h \to 0$. Les conditions 3 et 4 sur la variance découlent de l'expression de var$(\mathcal{A}_{c;x,h})$ ci-dessus.

Dans les Figures A.7 et A.8, nous observons le comportement de $A_{c;x,h}$ en faisant varier la cible x et le paramètre de lissage discret h. La valeur de la probabilité modale $\Pr(\mathcal{A}_{c;x,h} = x) = 1-h$ dépend de celle de h et sa position dépend de la cible x qui est ici le point de référence.

Soit $X_1, X_2, \ldots, X_n$ un n-échantillon aléatoire indépendant et identiquement distribué de fonction de masse catégorielle ordonnée inconnue f sur $\aleph = \{0, 1, \ldots, c-1\}$ où $c \in \mathbb{N} \setminus \{0, 1\}$. L'estimateur à noyau associé discret $A_{c;x,h}$ est

$$\begin{aligned}\widetilde{f}_n(x) &= \frac{1}{n}\sum_{i=1}^{n} A_{c;x,h}(X_i) \\ &= \frac{1}{n}\sum_{i=1}^{n}\left\{(1-h)1_{X_i=x} + \frac{h}{c-1}1_{X_i\neq x}\right\},\end{aligned}$$

pour $x \in \aleph = \{0, 1, \ldots, c-1\}$.

Nous montrons que $x \mapsto \widetilde{f}_n(x)$ est une fonction de masse de probabilité. En effet, nous avons successivement :

$$\begin{aligned}\sum_{x=0}^{c-1}\widetilde{f}_n(x) &= \sum_{x=0}^{c-1}\left\{(1-h)1_{X_1=x} + \frac{h}{c-1}1_{X_1\neq x}\right\} \\ &= (1-h) + \frac{h}{c-1}(c-1) \\ &= 1.\end{aligned}$$

Les propriétés de biais et de variance de l'estimateur, et par conséquent du risque quadratique intégré, s'obtiennent par des calculs similaires à ceux des noyaux discrets standards en Section 1.2 de Chapitre 1. Ainsi, nous exprimons le biais ponctuel exact par

$$\begin{aligned}\text{biais}\{\widetilde{f}_n(x)\} &= f(x)\{\Pr(\mathcal{A}_{c;x,h} = x) - 1\} + \sum_{y\in\aleph_x\setminus\{x\}} f(y)\Pr(\mathcal{A}_{c;x,h} = y) \\ &= -hf(x) + \frac{h}{c-1}\left\{\sum_{i=0}^{c-1} f(i) - f(x)\right\} \\ &= \frac{-hc}{c-1}f(x) + \frac{h}{c-1}\sum_{i=0}^{c-1} f(i).\end{aligned}$$

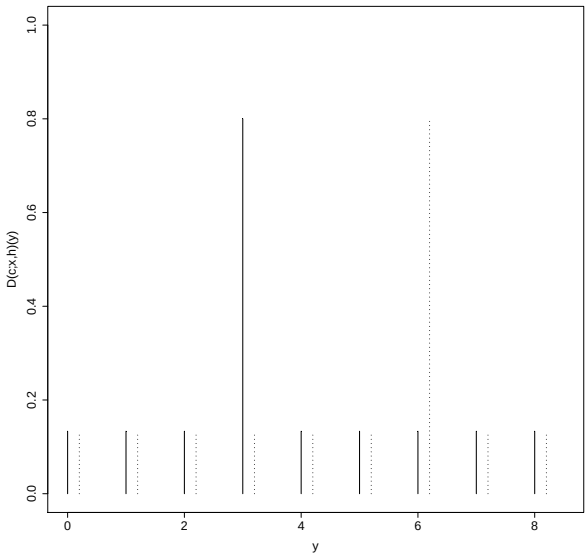

FIG. A.7 – Noyau associé d'Aitchison-Aitken pour h fixé et $y = x \in \{3, 6\}$

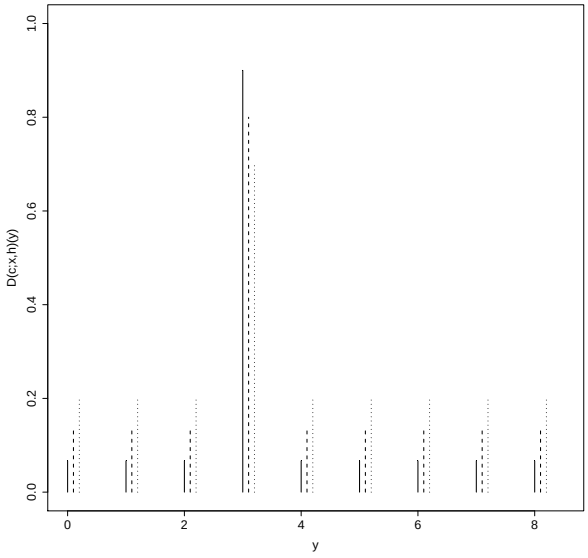

FIG. A.8 – Noyau associé d'Aitchison-Aitken pour x fixé et selon des valeurs de h

qui tend vers 0 quand $h \to 0$. Une possibilité pour réduire ce biais serait de modifier les paramètres pour avoir une loi discrète centrée en c_0 comme pour les lois triangulaires discrètes [Chapitre 2]. Pour la variance ponctuelle exacte, nous obtenons

$$\begin{aligned}
\text{var}\{\widetilde{f}_n(x)\} &= \frac{1}{n}\left[f(x)\left\{\Pr(\mathcal{A}_{c;x,h}=x)\right\}^2 + \sum_{y\in\aleph_x\setminus\{x\}} f(y)\left\{\Pr(\mathcal{A}_{c;x,h}=y)\right\}^2\right] \\
&\quad -\frac{1}{n}\left[f(x)\Pr(\mathcal{A}_{c;x,h}=x) + \sum_{y\in\aleph_x\setminus\{x\}} f(y)\Pr(\mathcal{A}_{c;x,h}=y)\right]^2 \\
&= \frac{1}{n}\left[f(x)(1-h)^2 + \frac{h^2}{(c-1)^2}\left\{\sum_{i=0}^{c-1} f(i) - f(x)\right\}\right] \\
&\quad -\frac{1}{n}\left[f(x)(1-h) + \frac{h}{c-1}\left\{\sum_{i=0}^{c-1} f(i) - f(x)\right\}\right]^2.
\end{aligned}$$

laquelle tend aussi vers celle de l'estimateur naïf $n^{-1}f(x)\left\{1-f(x)\right\}$ quand $h \to 0$. En utilisant le Théorème 1.2.6, le comportement du $MISE(n,h,A_c,f)$ est donné par

$$\begin{aligned}
MISE(n,h,A_c,f) &\doteq \frac{1}{n}(1-h)^2\sum_{x\in\aleph} f(x)\{1-f(x)\} \\
&\quad + \sum_{x\in\aleph}\left[f\{\mathbb{E}(\mathcal{A}_{c;x,h})\} - f(x) + \frac{1}{2}\text{var}(\mathcal{A}_{c;x,h})f^{(2)}(x)\right]^2.
\end{aligned}$$

Par conséquent, on a : $MISE(n,h,A_c,f) \to 0$ quand $n \to +\infty$ et $h = h(n) \to 0$; car $0 \leq \sum_{x\in\aleph} f(x)\left\{1-f(x)\right\} < 1$ et $\lim_{h\to 0}\left[f\{\mathbb{E}(\mathcal{A}_{c;x,h})\} - f(x) + (1/2)\text{var}(\mathcal{A}_{c;x,h})f^{(2)}(x)\right] = 0$. D'où, la convergence globale en moyenne quadratique des estimateurs à noyaux associés discrets d'Aitchison-Aitken.

Annexe B

Graphiques et tableaux supplémentaires

Sommaire :

B.1 Noyaux discrets standards

Dans cette partie, nous présentons des résultats liés aux noyaux discrets standards de la Section 1.2 du Chapitre 1 . En particulier, nous illustrons par quelques graphiques la forme variable du noyau binomial. De plus, nous montrons que les estimateurs à noyaux discrets standards ne sont pas des fonctions de masse de probabilité par des calculs formels réalisés sous le logiciel Maple.

Soit $x \in \mathbb{N}$ et $h \in]0,1]$. Le noyau binomial $B_{x,h}$ associé à la v.a. $\mathcal{B}_{x,h}$ de loi binomiale $\mathcal{B}\{x+1,(x+h)/(x+1)\}$ sur $\aleph_x = \{0,1,\ldots,x+1\}$ est défini par

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

Nous avons précisé à la section 1.2 du Chapitre 1 que le noyau $B_{x,h}$ est un noyau associé du premier ordre. En effet, il vérifie toutes les conditions d'un noyau associé discret exceptée la dernière. Dans ce qui suit, nous mettons en évidence le changement de forme du noyau binomial selon la cible x et la fenêtre h.

Dans la Figure B.1, nous fixons la fenêtre $h = 0.1$ et nous faisons varier la cible $x \in \{0,1,2,4,5,7\}$. La probabilité modale est située au niveau de la cible x et la forme

du noyau s'adapte de manière à ce que la moyenne soit autour de x. Dans les Figures B.2, B.3 et B.4, pour chaque cible $x \in \{0, 1, 7\}$ nous utilisons plusieurs fenêtres de lissage discret $h \in \{0.05, 0.1, 0.2, 0.4, 0.6, 0.7\}$. Pour des petites valeurs de h, nous observons que la valeur modale de la probabilité est en x, tandis que pour des plus grandes valeurs de h le mode se déplace en $x + 1$.

Soit $X_1, X_2, \ldots, X_n$ un n-échantillon aléatoire indépendant et identiquement distribué de fonction de masse de probabilité inconnue f sur $\aleph = \mathbb{N}$. Pour $x \in \mathbb{N}$ et $h \in]0, 1]$, l'estimateur à noyau binomial est donné par :

$$\widetilde{f}_n^B(x) = \frac{1}{n} \sum_{i=1}^{n} B_{x,h}(X_i).$$

Nous exprimons :

$$\begin{aligned} \sum_{x \in \mathbb{N}} \widetilde{f}_n^B(x) &= \sum_{x \in \mathbb{N}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{(x+1)!}{X_i!(x+1-X_i)!} \left(\frac{x+h}{x+1}\right)^{X_i} \left(\frac{1-h}{x+1}\right)^{x+1-X_i} \right\} \\ &= \sum_{x \in \mathbb{N}} \left\{ \frac{(x+1)!}{X_1!(x+1-X_1)!} \left(\frac{x+h}{x+1}\right)^{X_1} \left(\frac{1-h}{x+1}\right)^{x+1-X_1} \right\}. \end{aligned}$$

Numériquement, nous vérifions que cette somme ne fait pas toujours 1 et qu'ainsi cet estimateur n'est pas une fonction de masse de probabilité (Table B.1). Mais théoriquement, ce résultat n'est pas facile à prouver car il dépend de l'échantillon X_1 et de la fenêtre h.

De manière similaire, pour les estimateurs à noyau de Poisson (Exemple 1.2.1) et binomial négatif (Exemple 1.2.3), nous exprimons les sommes suivantes, respectivement :

$$\sum_{x \in \mathbb{N}} \widetilde{f}_n^P(x) = \sum_{x \in \mathbb{N}} \left\{ e^{-(x+h)} \frac{(x+h)^{X_1}}{X_1!} \right\}$$

et

$$\sum_{x \in \mathbb{N}} \widetilde{f}_n^{BN}(x) = \sum_{x \in \mathbb{N}} \left\{ \frac{(x+X_1)!}{x!X_1!} \left(\frac{x+h}{2x+1+h}\right)^{X_1} \left(\frac{x+1}{2x+1+h}\right)^{x+1} \right\},$$

avec $h > 0$. Les résultats numériques correspondants à ces deux dernières sommes sont présentés dans les Tables (B.2) et (B.3) en fonction de l'échantillon X_1 et de la fenêtre h. Notons que dans certains cas cette somme peut être très proche de 1.

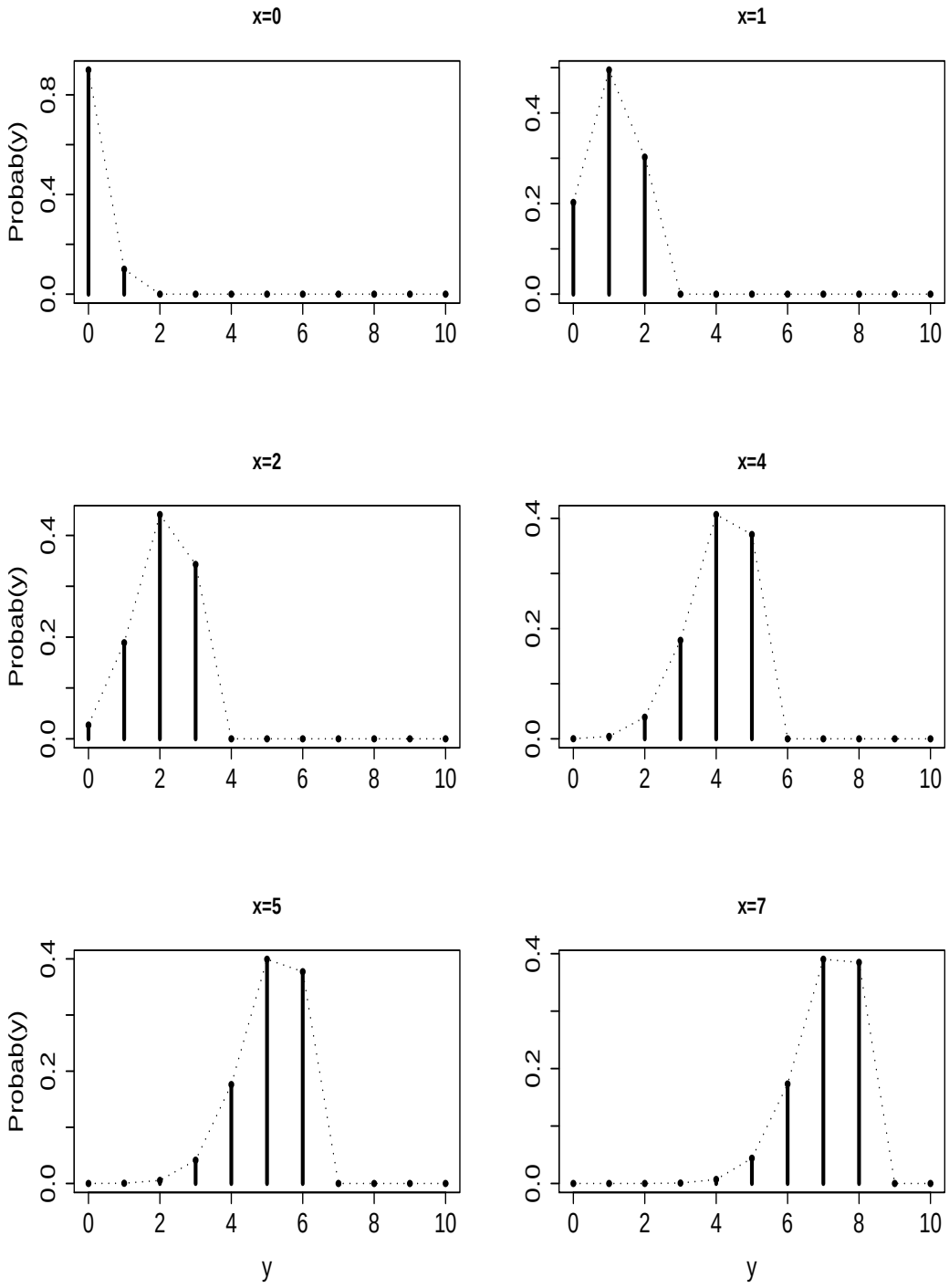

FIG. B.1 – Noyau binomial pour certaines valeurs de x et $h = 0.1$

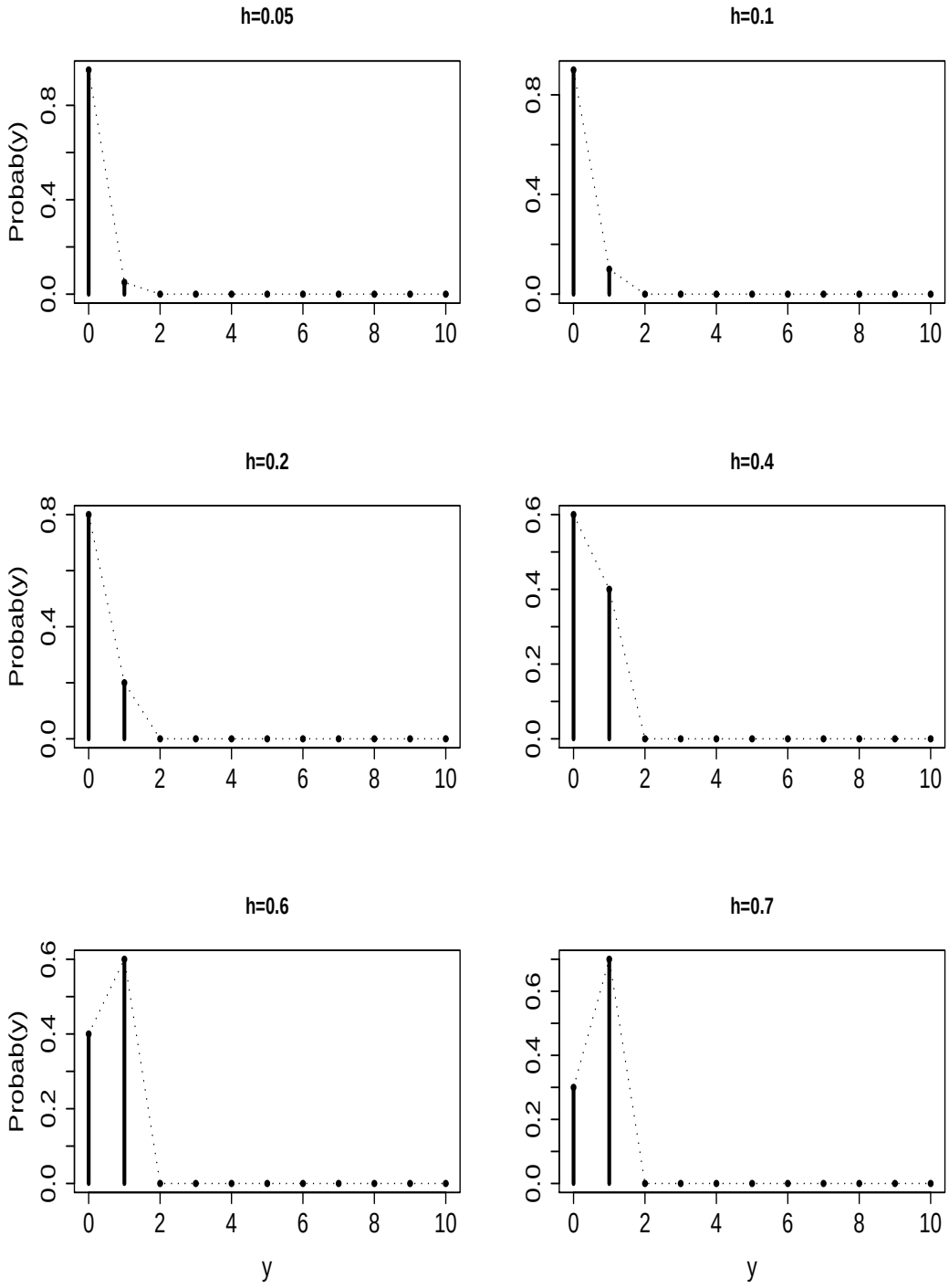

FIG. B.2 – Suite de Figure B.1 pour $x = 0$ et selon des valeurs de h

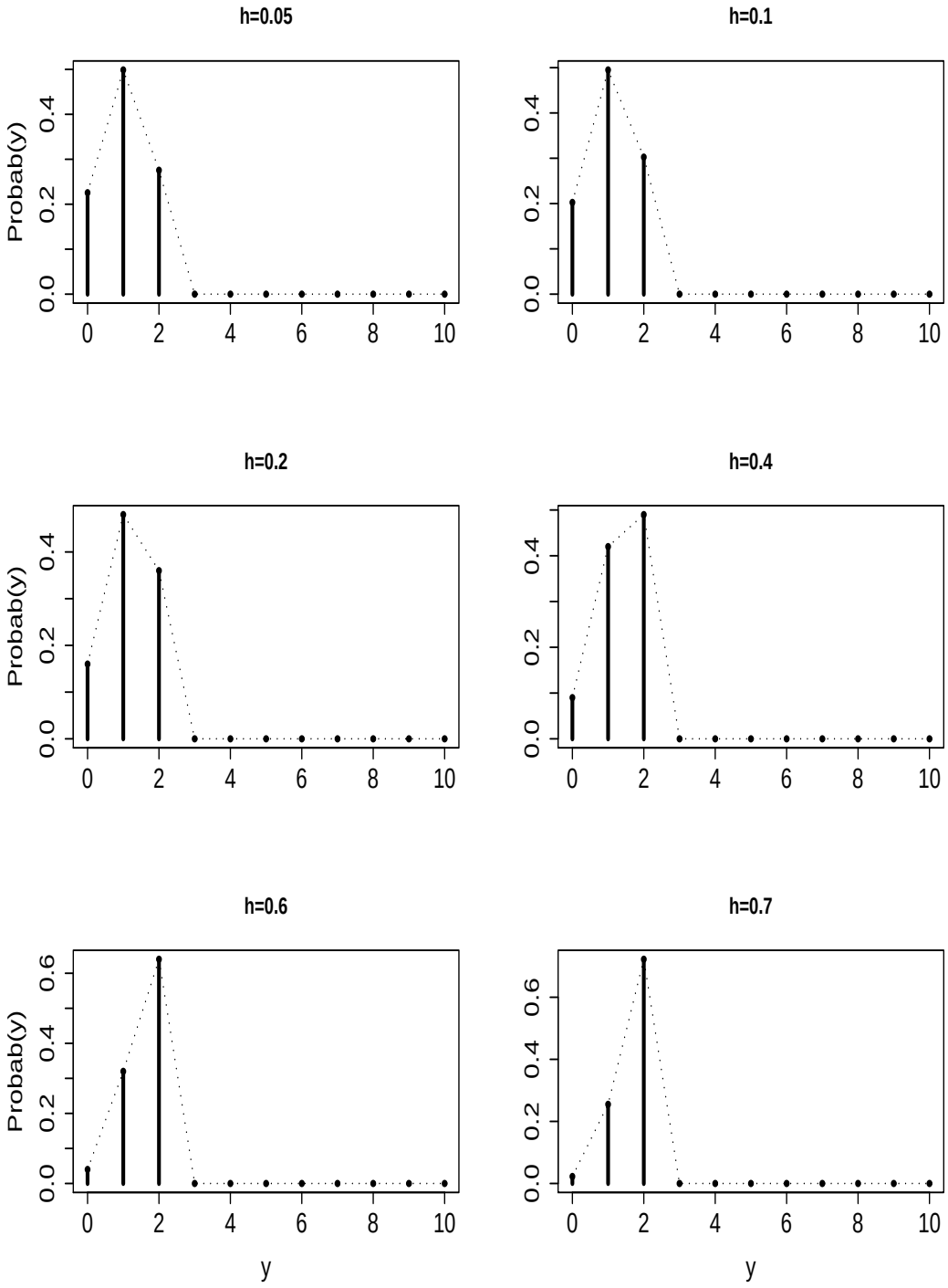

FIG. B.3 – Suite de Figure B.1 pour $x = 1$ et selon des valeurs de h

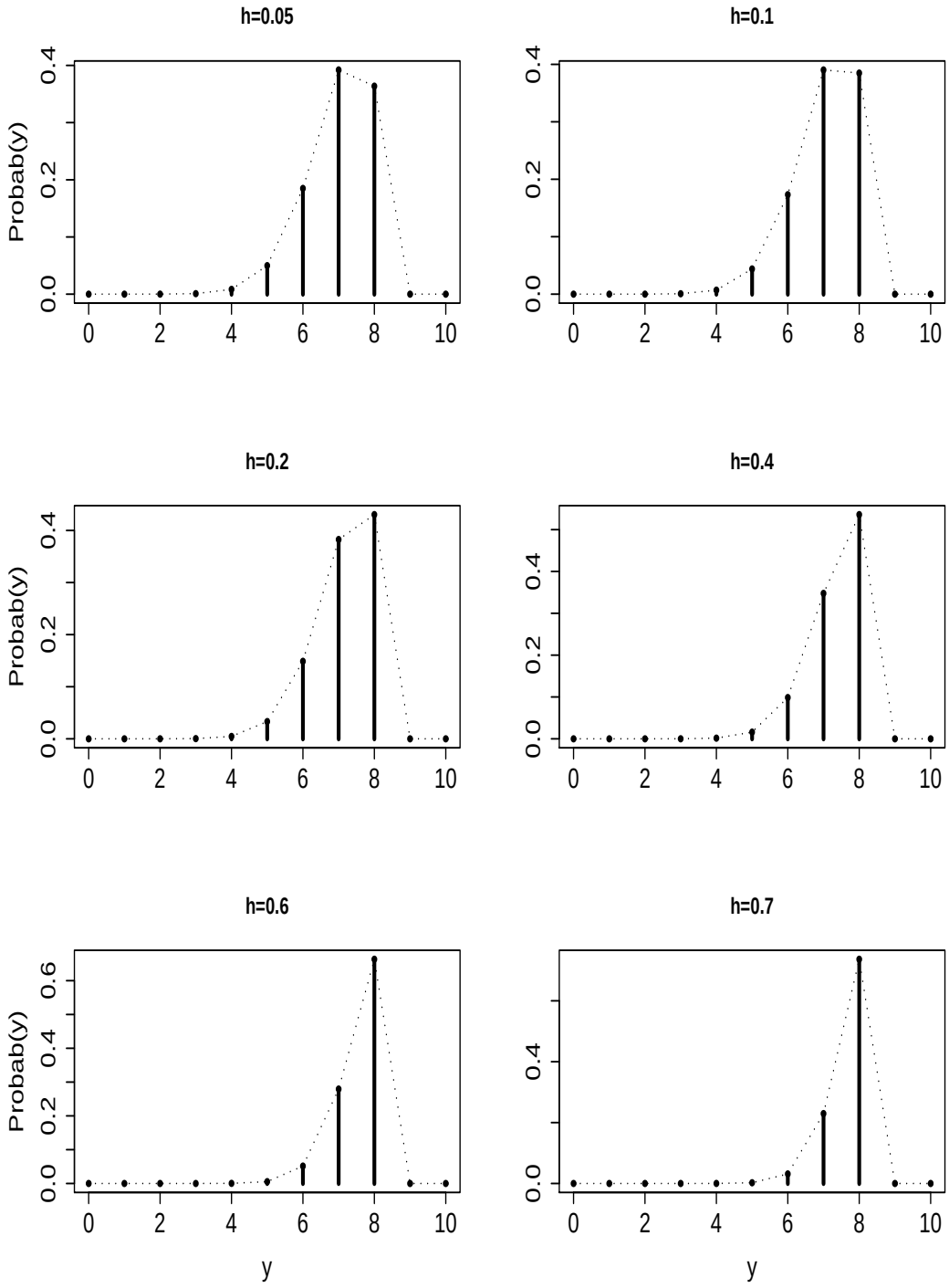

FIG. B.4 – Suite et fin de Figure B.1 pour $x = 7$ et selon des valeurs de h

$h = 0.1$				
X_1	0	1	5	10
$\sum_{x\in\mathbb{N}} \widetilde{f}_n(x)$	1.13226	0.82403	0.99689	0.99947
$h = 0.3$				
X_1	0	1	5	10
$\sum_{x\in\mathbb{N}} \widetilde{f}_n(x)$	0.83620	0.89968	0.99802	0.99967
$h = 0.7$				
X_1	0	1	5	10
$\sum_{x\in\mathbb{N}} \widetilde{f}_n(x)$	0.32353	0.98362	0.99960	0.99993

TAB. B.1 – Valeurs de la constante de normalisation pour le noyau binomial en fonction de l'échantillon et de la fenêtre

$h = 0.1$				
X_1	0	1	5	10
$\sum_{x\in\mathbb{N}} \widetilde{f}_n(x)$	1.43143	0.97260	1.000002	0.99999
$h = 0.5$				
X_1	0	1	5	10
$\sum_{x\in\mathbb{N}} \widetilde{f}_n(x)$	0.95952	1.03817	1.000017	1.000000001
$h = 1.1$				
X_1	0	1	5	10
$\sum_{x\in\mathbb{N}} \widetilde{f}_n(x)$	0.52659	0.88572	1.000000179	1.000000002

TAB. B.2 – Suite de Table B.1 pour le noyau de Poisson

$h = 0.1$				
X_1	0	1	5	10
$\sum_{x\in\mathbb{N}} \widetilde{f}_n(x)$	1.72947	1.080377	1.00798	1.00125
$h = 0.5$				
X_1	0	1	5	10
$\sum_{x\in\mathbb{N}} \widetilde{f}_n(x)$	1.31709	1.10593	1.00450	1.00055
$h = 1.1$				
X_1	0	1	5	10
$\sum_{x\in\mathbb{N}} \widetilde{f}_n(x)$	0.95203	0.97482	0.99822	0.99996

TAB. B.3 – Suite et fin de Table B.1 pour le noyau binomial négatif

B.2 Etude par simulation de MISE et AMISE

Dans cette partie, nous étudions les expressions exactes et approchées des risques quadratiques intégrés $MISE$ pour les estimateurs à noyaux discrets standards et l'estimateur naïf (voir Section 1.3) ainsi que pour les estimateurs à noyaux discrets triangulaires (voir Section 2.3). Pour cela, entre autres possibilités, nous considérons de manière classique une loi de Poisson de moyenne 2. À l'issue de 1000 répliques, nous calculons la moyenne des erreurs quadratiques intégrées optimales ISE ainsi que leurs écart-types pour les estimateurs (voir Table B.4). Pour chaque simulation, les fenêtres optimales de lissage discret sont obtenues par la procédure de validation croisée. De là, les erreurs quadratiques intégrées optimales sont calculées en utilisant les valeurs optimales des fenêtres de lissage discret. Les valeurs optimales des $MISE$ et des approximations $AMISE$ sont aussi présentées. Il apparait que les moyennes $\mathbb{E}(ISE)$ convergent vers les risques $MISE$ et $AMISE$ quand la taille n de l'échantillon augmente. Ceci confirme les expressions de $AMISE$ développées pour les estimateurs à noyaux discrets et l'estimateur fréquence.

TAB. B.4 – Résultats simulés de $\mathbb{E}(ISE)$ et leurs écart-types (entre parenthèses) ainsi que de $MISE$ et $AMISE$ pour les estimateurs à noyaux discrets et l'estimateur fréquence. Les résultats présentés ont été multipliés par 10^3

(a)

Noyau	Triangulaire $a = 1$			Triangulaire $a = 2$		
n	$MISE$	$\mathbb{E}(ISE)$	$AMISE_a$	$MISE$	$\mathbb{E}(ISE)$	$AMISE_a$
20	**10.82**	**14.10** (15.94)	**11.21**	15.37	14.28 (13.14)	13.70
50	**5.47**	**6.81** (7.07)	**5.27**	9.89	8.63 (5.91)	8.85
80	**4.02**	4.80 (4.37)	**3.78**	5.60	6.34 (4.23)	6.65
100	**3.63**	4.14 (3.55)	**3.29**	4.95	5.42 (3.43)	6.03
200	**2.64**	**2.57** (1.83)	**2.25**	3.94	2.82 (1.98)	3.26
500	**1.22**	**1.28** (0.87)	**1.28**	1.57	1.46 (0.94)	1.83
700	**0.93**	**0.93** (0.62)	**0.95**	1.12	1.17 (0.71)	1.56

(b)

Noyau	Binomial			Poisson		
n	$MISE$	$\mathbb{E}(ISE)$	$AMISE^*$	$MISE$	$\mathbb{E}(ISE)$	$AMISE$
20	12.81	21.52 (23.36)	14.56	21.15	17.88 (9.42)	35.66
50	7.13	7.10 (7.87)	6.33	15.73	15.76 (5.76)	26.74
80	4.92	**4.46** (3.68)	4.61	14.68	15.02 (4.09)	25.77
100	4.60	**3.94** (3.41)	4.04	13.80	15.07 (3.68)	25.44
200	3.47	2.72 (2.01)	2.90	14.04	14.95 (2.50)	22.32
500	2.86	2.03 (1.12)	2.21	13.62	13.75 (1.59)	21.87
700	2.75	1.91 (0.92)	2.07	13.58	12.95 (1.07)	21.79

(c)

Noyau	Binomial négatif			Naif	
n	$MISE$	$\mathbb{E}(ISE)$	$AMISE^*$	$\mathbb{E}(ISE)$	$MISE$
20	27.83	27.60 (7.58)	79.72	36.48 (24.73)	39.65
50	26.76	27.73 (5.39)	76.33	15.84 (10.76)	15.86
80	26.34	27.63 (4.31)	75.48	10.03 (6.97)	9.91
100	24.85	27.78 (4.07)	75.20	8.13 (5.97)	7.93
200	25.80	28.26 (2.99)	74.63	3.89 (2.76)	3.96
500	25.00	26.76 (2.09)	74.29	1.57 (1.09)	1.59
700	24.94	25.96 (1.87)	74.23	1.14 (0.78)	1.13

B.3 Lissages discrets des données simulées

Dans cette section, nous faisons des ajustements de la distribution des données d'un mélange de Poisson $f = 0.4\mathcal{P}(0.5) + 0.6\mathcal{P}(10)$ pour plusieurs valeurs de la fenêtre h. Nous utilisons les noyaux discrets binomial, Poisson et binomial négatif. Dans les graphiques présentés, les bâtons en noir représentent les fréquences empiriques des

données simulées et les autres bâtons correspondent aux estimations discrètes obtenues.

h	$\sum_{x\in\mathbb{N}} \widetilde{f}_n(x)$	ISE^0
0.01	1.03747	0.00077
0.03	1.03101	0.00073
0.05	1.02459	0.00070
0.11*	**1.00555**	**0.00066**
0.2	0.97762	0.00081
0.3	0.94745	0.00130
0.5	0.88976	0.00346
0.75	0.82239	0.00898
0.9	0.78438	0.01422

TAB. B.5 – Qualité de lissages discrets par l'estimateur à noyau binomial de la distribution des données simulées du mélange de Poisson f avec $h_0^* = 0.11149$ et $n = 1000$

h	$\sum_{x\in\mathbb{N}} \widetilde{f}_n(x)$	ISE^0
0.01	1.12389	0.00739
0.05	1.11203	0.00731
0.1	1.09750	0.00725
0.19*	**1.07215**	**0.00723**
0.3	1.04256	0.00738
0.5	0.99261	0.00812
0.9	0.90669	0.01142
1	0.88790	0.01260
1.2	0.85329	0.02059

TAB. B.6 – Suite de Table B.5 pour le noyau de Poisson avec $h_0^* = 0.19099$

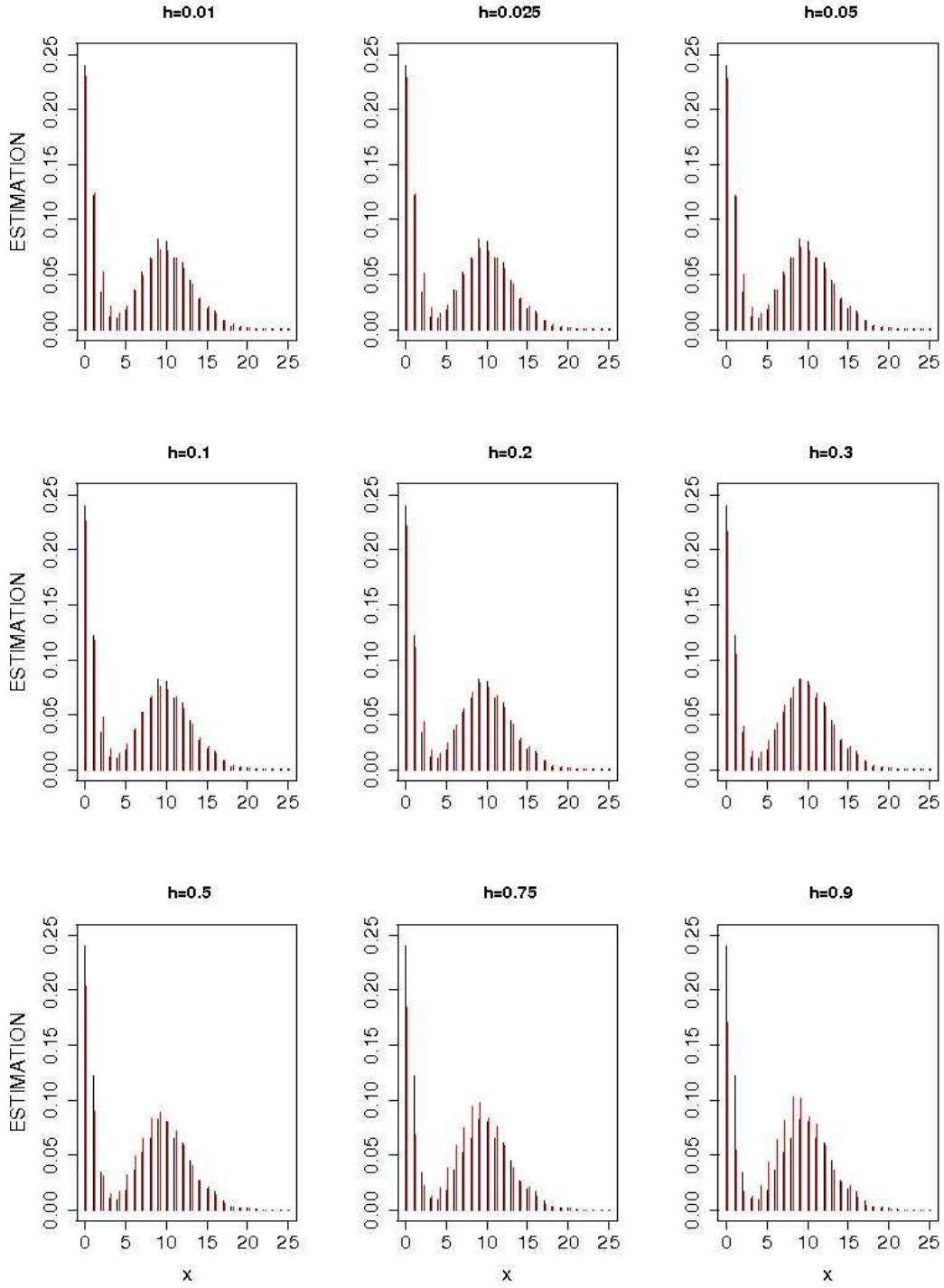

FIG. B.5 – Lissages par l'estimateur à noyau binomial selon différentes fenêtres de la distribution des données simulées du mélange de Poisson f avec $n = 1000$

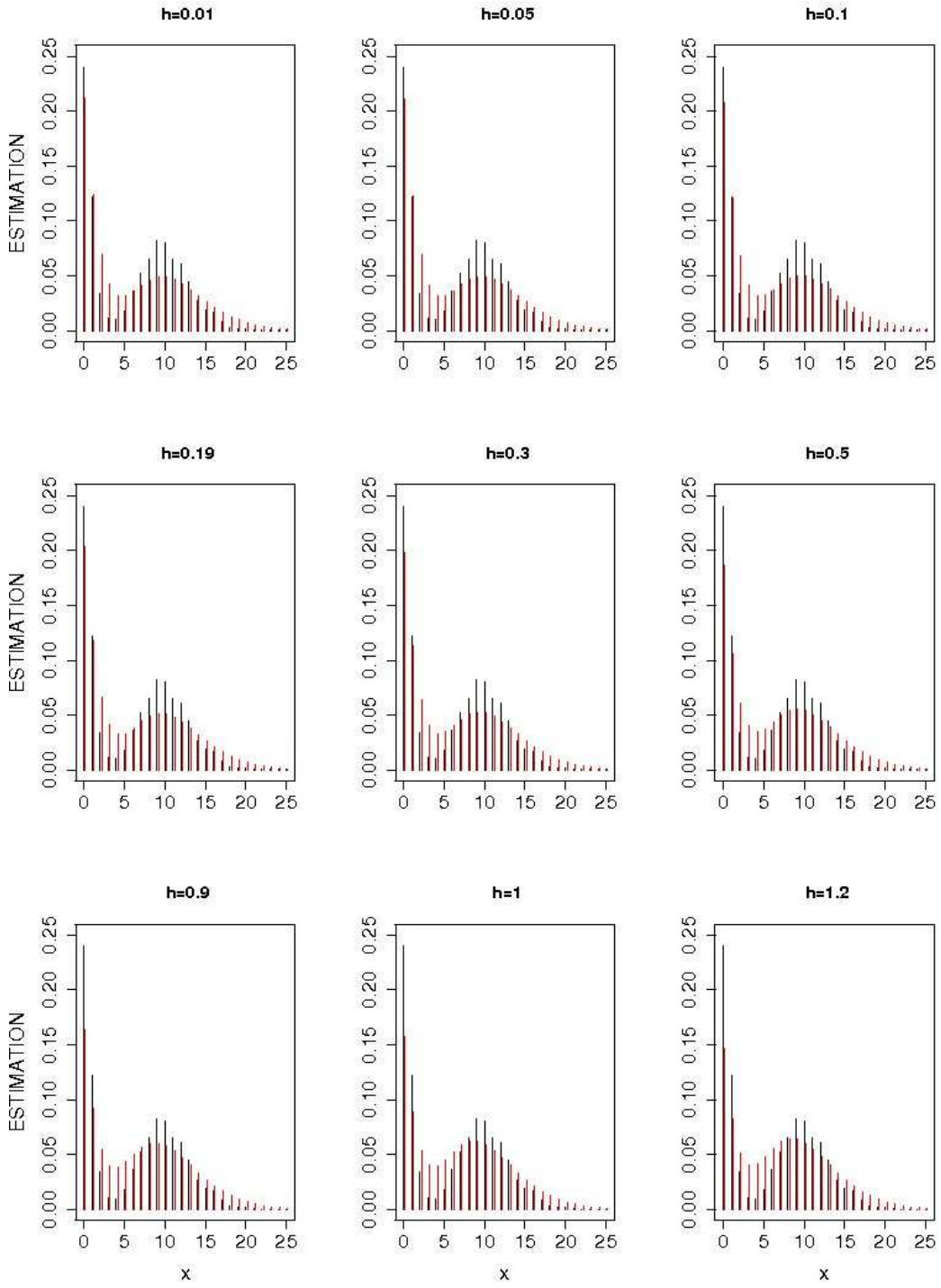

FIG. B.6 – Suite de Figure B.5 avec le noyau de Poisson

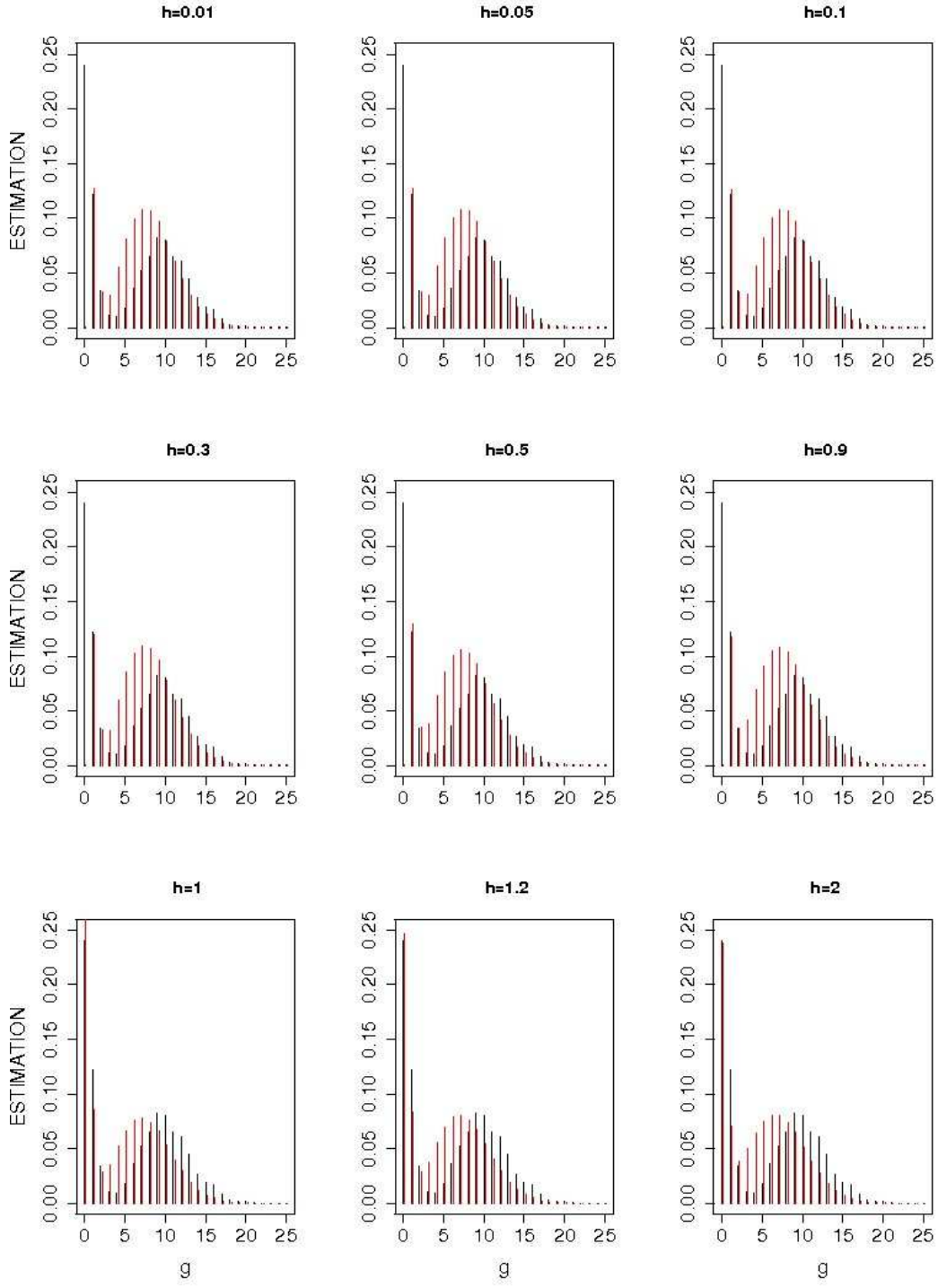

FIG. B.7 – Suite et fin de Figure B.5 avec le noyau binomial négatif

h	$\sum_{x\in\mathbb{N}} \widetilde{f}_n(x)$	ISE^0
0.01	0.28311	0.07398
0.1	0.28667	0.07442
0.29*	0.29330	0.07545
0.5	0.30993	0.07567
0.9	0.32048	0.07791
1	**0.44916**	**0.01301**
1.2	0.44265	0.01351
2	0.47015	0.01733
5	0.45852	0.03395

TAB. B.7 – Suite et fin de Table B.5 pour le noyau binomial négatif avec $h_0^* = 0.29156$

B.4 Lissages discrets des données de buts

Ici, nous présentons les ajustements de la distribution des données de buts en football de la Ligue 1 française (saison 2005-2006) avec $n = 380$. Nous illustrons aussi le choix de la fenêtre optimale en minimisant les courbes des fonctions de l'erreur quadratique intégrée $h \mapsto ISE^0(h)$ et du critère de validation croisée $h \mapsto CV(h)$. Les noyaux discrets utilisés sont binomial, Poisson et binomial négatif.

h	$\sum_{x\in\mathbb{N}} \widetilde{f}_n(x)$	ISE^0
0.01	0.97127	0.00343
0.03	0.96974	0.00338
0.05	**0.96818**	**0.00336**
0.1	0.96416	0.00344
0.2	0.95559	0.00418
0.3*	0.94635	0.00574
0.5	0.92599	0.01151
0.75	0.89738	0.02456
0.9	0.87875	0.03623

TAB. B.8 – Qualité de lissages discrets par le noyau binomial de la distribution de buts de Ligue 1 (saison 2005-2006) avec $h_0^* = 0.29939$ et $n = 380$

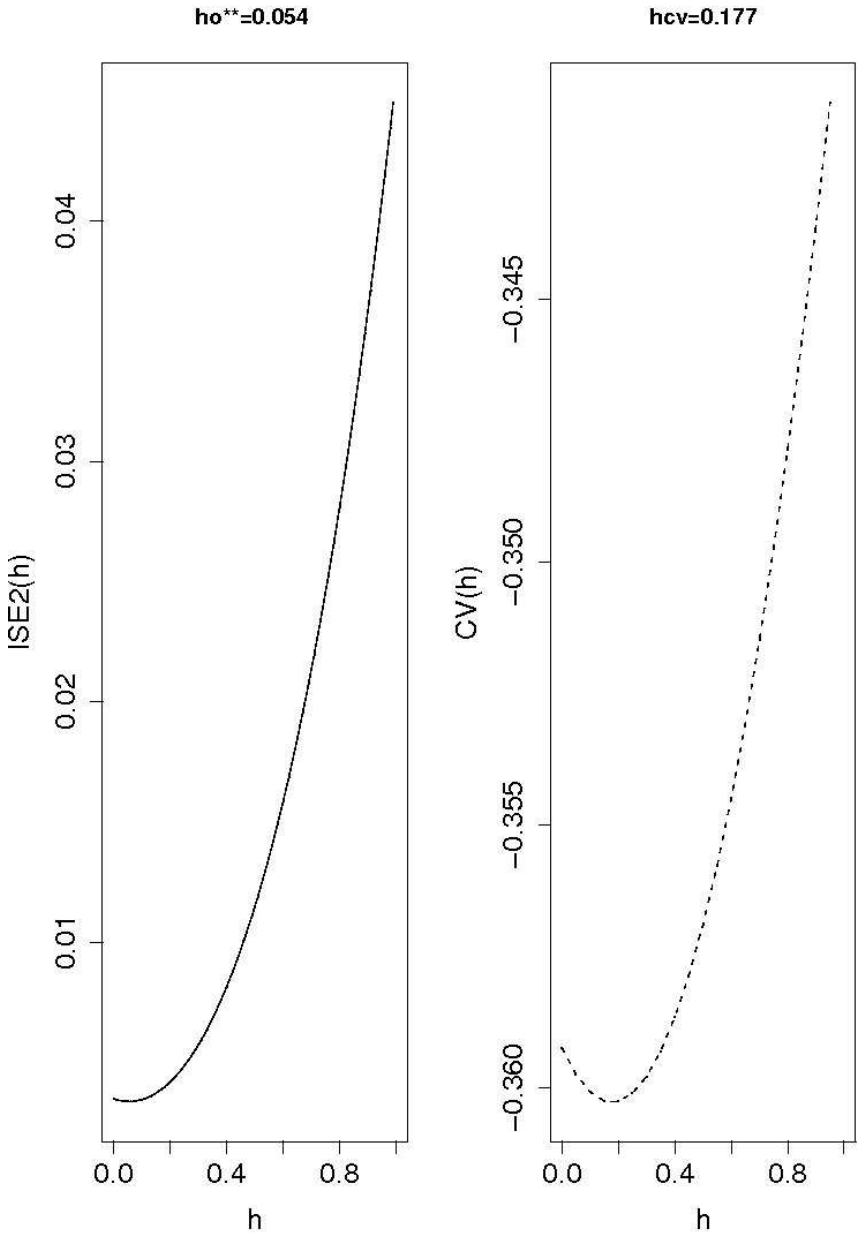

FIG. B.8 – Courbes des fonctions $h \mapsto ISE^0(h)$ (notée $ISE2(h)$ sur le graphe) et de validation croisée $h \mapsto CV(h)$ avec le noyau binomial de la distribution de buts de Ligue 1 (saison 2005-2006) avec $n = 380$

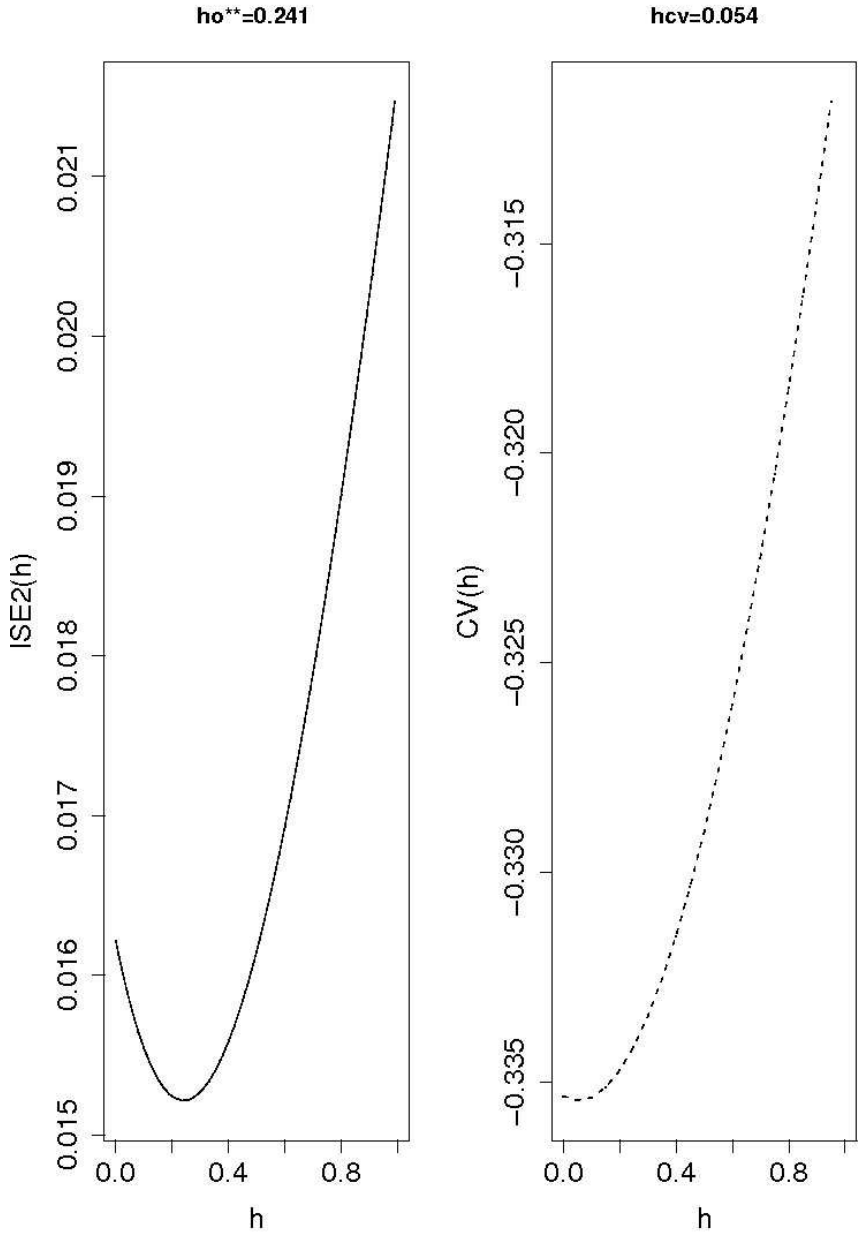

FIG. B.9 – Suite de Figure B.8 pour le noyau de Poisson

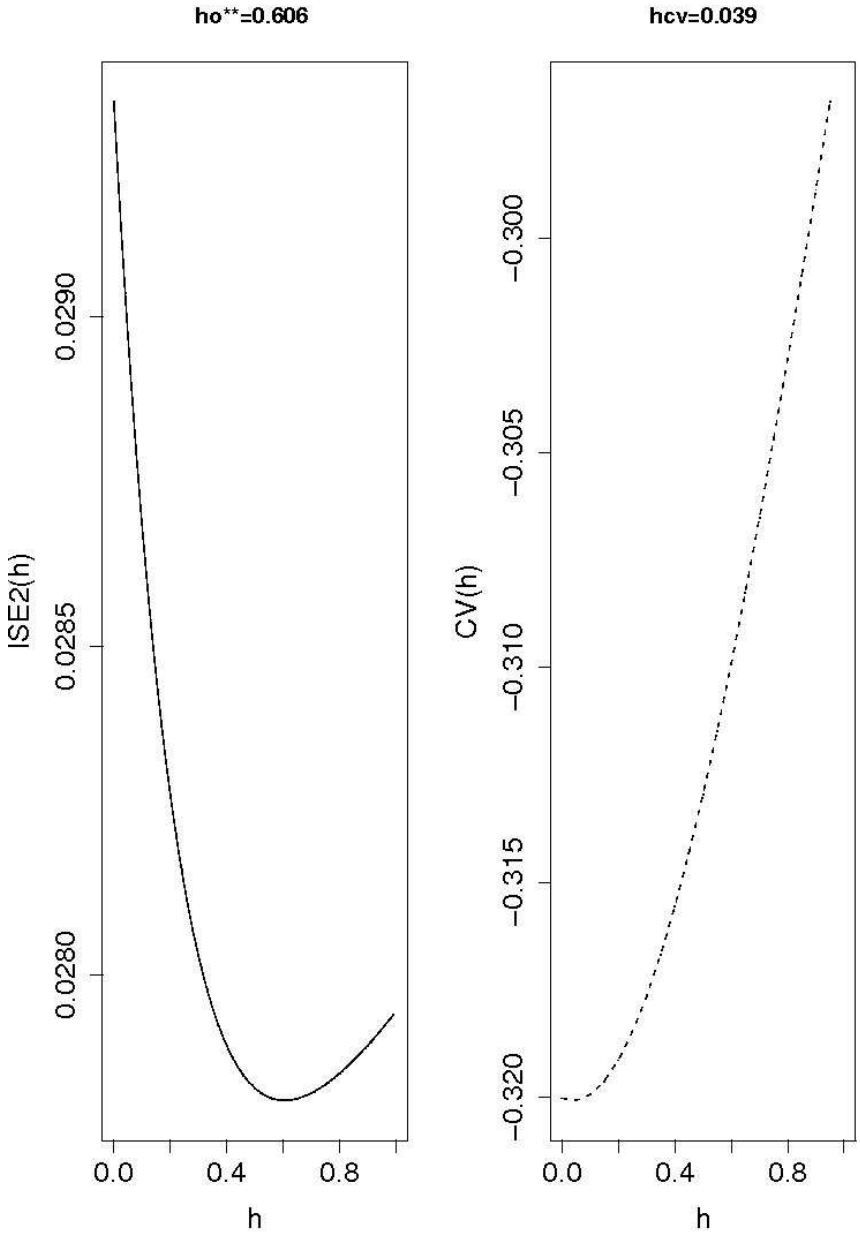

FIG. B.10 – Suite et fin de Figure B.8 pour le noyau Binomial négatif

h	$\sum_{x\in\mathbb{N}} \widetilde{f}_n(x)$	ISE^0
0.01	1.05434	0.01613
0.05	1.05115	0.01583
0.1	1.04689	0.01554
0.3	**1.02679**	**0.01527**
0.5	1.001951	0.01615
0.7*	0.97282	0.01789
0.9	0.93998	0.02025
1	0.92237	0.02160
1.2	0.88522	0.02456

TAB. B.9 – Suite de Table B.8 pour le noyau de Poisson et $h_0^* = 0.70105$

h	$\sum_{x\in\mathbb{N}} \widetilde{f}_n(x))$	ISE^0
0.01	0.24309	0.02998
0.1	0.24439	0.03013
0.3	0.24516	0.03044
0.5	0.29740	0.04645
1	0.37895	0.02076
1.2	**0.36842**	**0.02024**
1.6*	0.41461	0.03525
2	0.43552	0.05529
5	0.30718	0.08533

TAB. B.10 – Suite et fin de Table B.8 pour le noyau binomial négatif et $h_0^* = 1.59764$

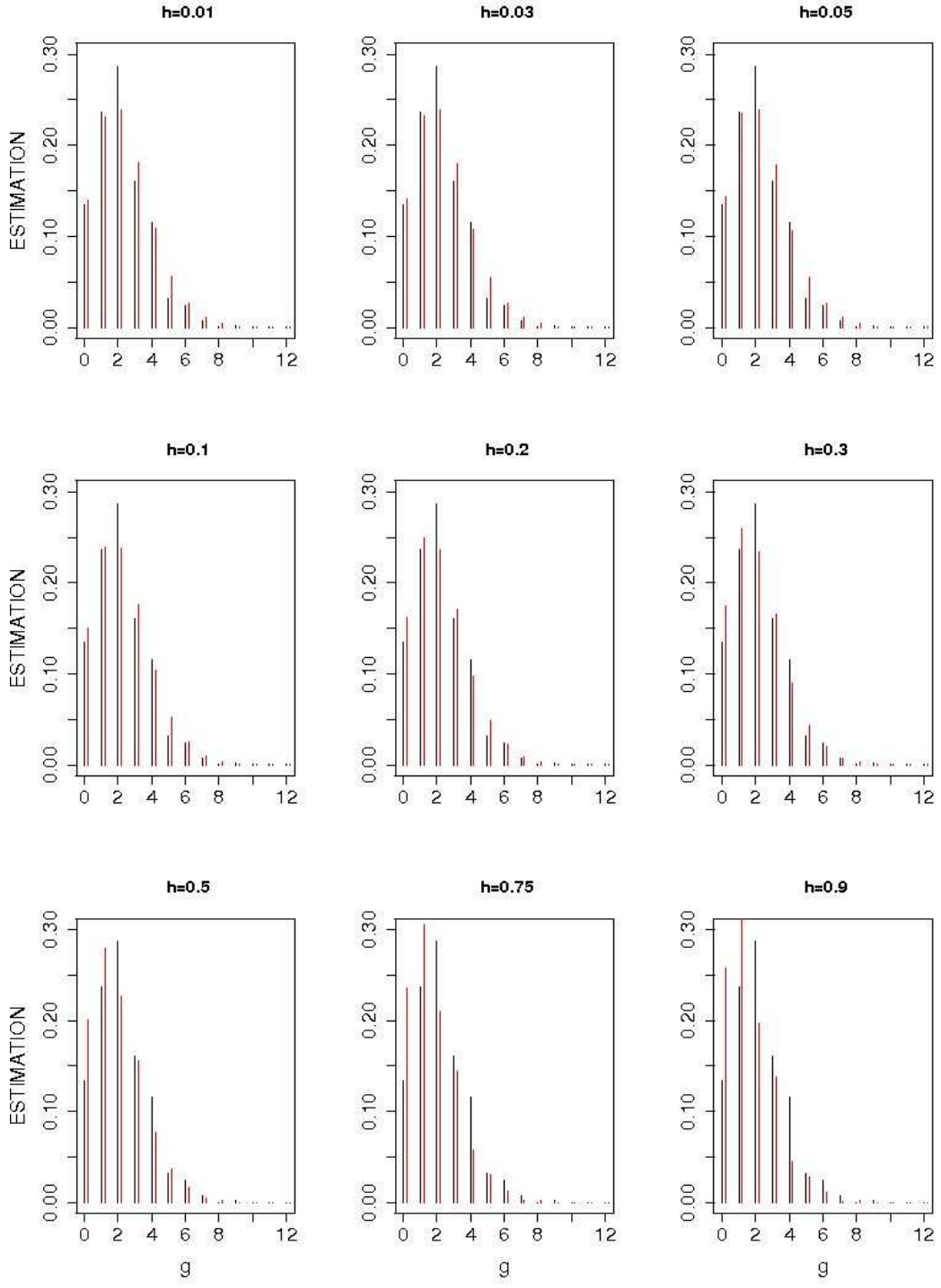

FIG. B.11 – Lissages discrets par le noyau binomial selon différentes fenêtres de la distribution de buts de Ligue 1 (saison 2005-2006) avec $n = 380$

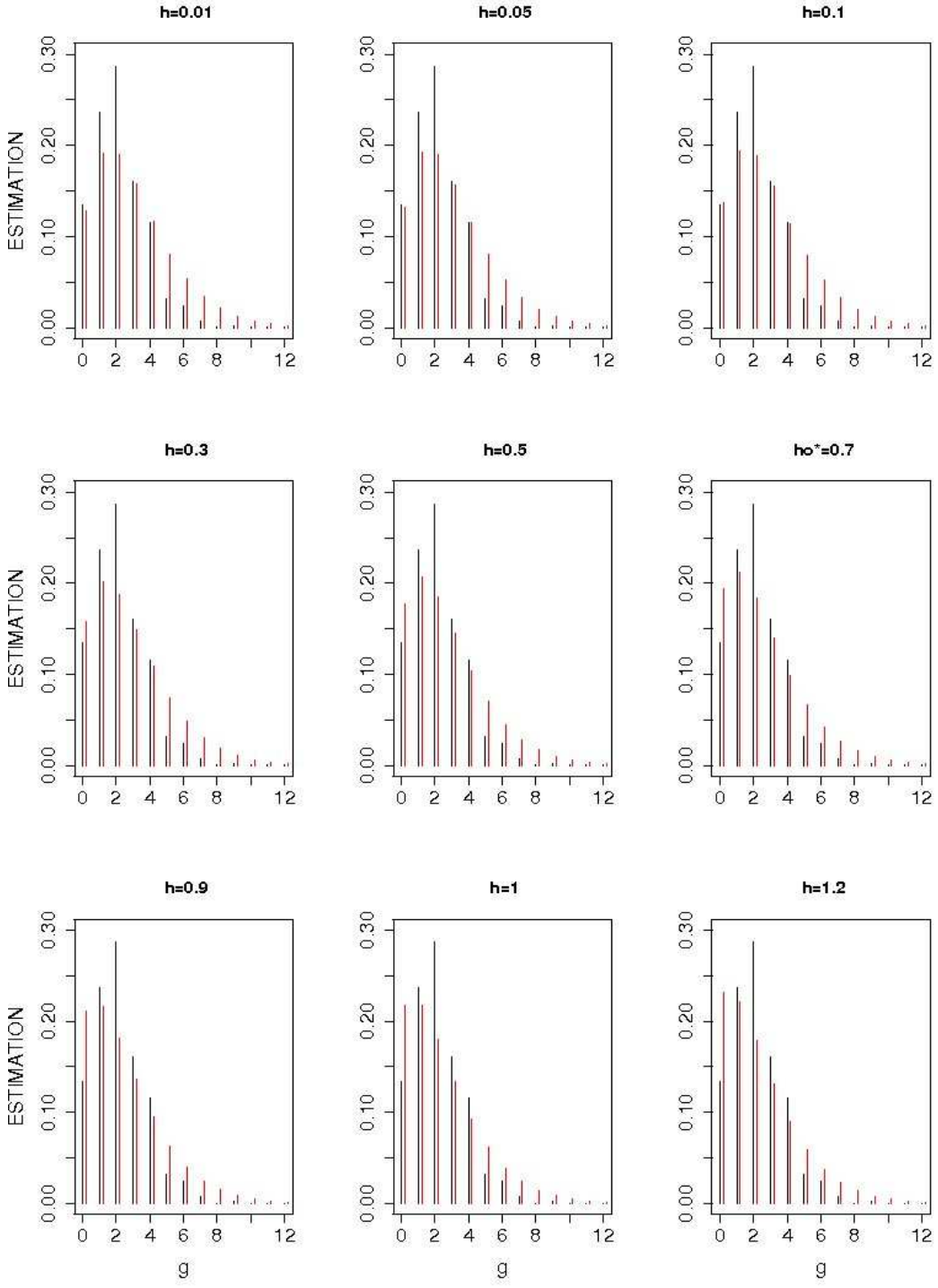

FIG. B.12 – Suite de Figure B.11 pour le noyau de Poisson

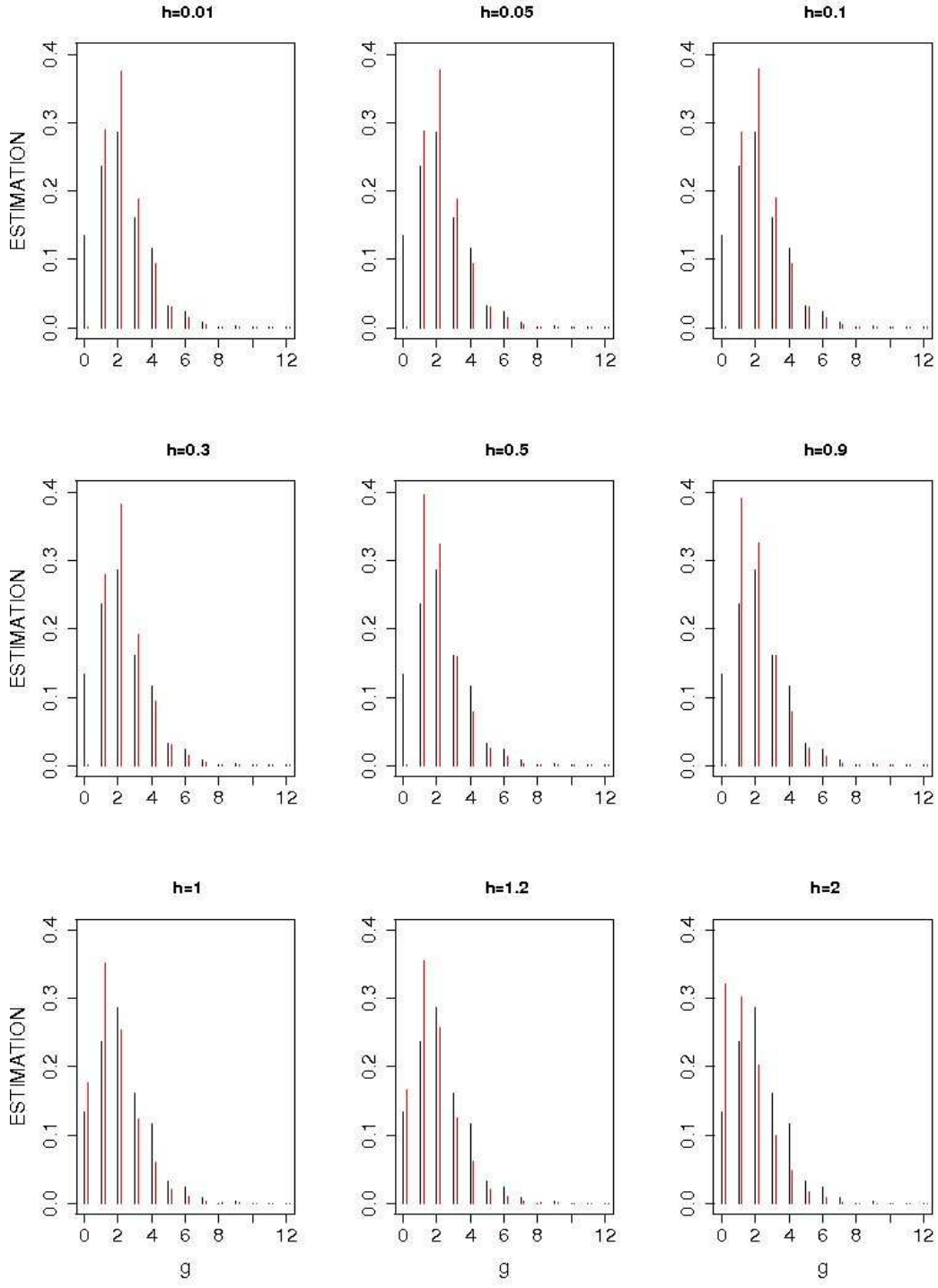

FIG. B.13 – Suite et fin de Figure B.11 pour le noyau binomial négatif

B.5 Autres lissages discrets

Cette partie complète les ajustements réalisés dans les deux sections précédentes sur les distributions des données du mélange de Poisson et des nombres de buts en football. Nous réalisons ici des estimations sur des échantillons de petites tailles $n \in \{30, 50\}$ des données simulées de distributions de Poisson de moyenne $\mu \in \{2, 5\}$ (Table B.11). Selon les échantillons, le choix de fenêtre est fait par les méthodes de validation croisée et d'excès de zéros. De plus, en faisant un nombre de simulations $n_{sim} = \{25, 50, 100, 200, 500, 1000\}$, nous déterminons la moyenne et la variance des fenêtres optimales au sens de ISE (Tables B.14 et B.15).

$n = 50$	$\mu = 2$										
	x_i	0	1	2	3	4	5				
	n_i	9	16	9	8	4	4				
	$\mu = 5$										
	x_i		1	2	3	4	5	6	7	8	9	10
	n_i		2	2	11	11	5	7	5	4	2	1
$n = 30$	$\mu = 2$										
	x_i	0	1	2	3	4					
	n_i	6	8	9	5	2					
	$\mu = 5$										
	x_i		1	2	3	4	5	6	7	8	9	10
	n_i		2	1	4	5	3	6	5	3	0	1

TAB. B.11 – Données simulées de Poisson $\mathcal{P}(\mu)$ avec $\mu \in \{2; 5\}$ et $n \in \{50; 30\}$

		Naïf	Binomial	Poisson	Bin. nég.
$n = 50$	$\mu = 2$				
	h_{cv}		0.164	0.292	0.467
	C		0.95178	1.02122	1.04910
	ISE	0.0153	**0.01218**	0.02016	0.02848
	$AMISE^*(f)$	0.02	**0.01205**	0.03420	0.10666
	h_0		0.260	0.612	1.285
	C		0.93906	0.96613	0.89742
	ISE	**0.0153**	0.01566	0.02513	0.03109
	$AMISE^*(f)$	0.02	**0.01196**	0.04912	0.22204
	$\mu = 5$				
	h_{cv}		0.001	0.197	0.376
			0.001	*0.169*	*0.280*
	C		0.98432	0.97084	0.93846
			0.98436	*0.97759*	*0.95161*
	ISE	0.01638	**0.00287**	0.00705	0.01438
		0.01643	***0.00291***	*0.00714*	*0.01468*
	$AMISE^*(f)$	0.02	**0.00853**	0.02205	0.07297
		0.02	***0.00867***	*0.02114*	*0.06666*

TAB. B.12 – Qualités de lissages discrets par le noyau de type Dirac et les noyaux discrets binomial, de Poisson et binomial négatif de la distribution f de Poisson $\mathcal{P}(\mu)$ avec $\mu \in \{2; 5\}$ et $n = 50$. ***En italique*** : avec le point $x = 0$

		Naïf	Binomial	Poisson	Bin. nég.
$n = 30$	$\mu = 2$				
	h_{cv}		0.626	0.466	0.576
	C		0.87316	0.99094	1.02753
	ISE	**0.00726**	0.03260	0.01785	0.02503
	$AMISE^*(f)$	0.03333	**0.01582**	0.04117	0.11153
	h_0		0.226	0.554	1.165
	C		0.94675	0.97421	0.91341
	ISE	**0.00726**	0.01090	0.01940	0.02784
	$AMISE^*(f)$	0.03333	**0.01833**	0.04420	0.19143
	$\mu = 5$				
	h_{cv}		0.001	0.334	0.490
			0.001	*0.141*	*0.289*
	C		0.97977	0.96162	0.92494
			0.97984	*0.97433*	*0.94470*
	ISE	0.01910	**0.00606**	0.00980	0.01790
		0.01914	***0.00611***	*0.01102*	*0.01900*
	$AMISE^*(f)$	0.03333	**0.01395**	0.02638	0.08014
		0.03333	***0.01418***	*0.02350*	*0.06891*

TAB. B.13 – Suite et fin de Table B.12 avec $\mu \in \{2; 5\}$ et $n = 30$

$\mu = 2$	$n = 30$			
	n_{sim}	25	50	100
	$\overline{\widehat{h}}_P\left(\text{var}(\widehat{h})\right)$	0.388 (0.099)	0.308 (0.064)	0.300 (0.068)
	$\overline{\widehat{h}}_B()$	0.216 (0.026)	0.226 (0.029)	0.201 (0.023)
	$\overline{\widehat{h}}_{BN}$ ()	0.576 (0.335)	0.722 (0.357)	0.788 (0.366)
	$n = 50$			
	n_{sim}	25	50	100
	$\overline{\widehat{h}}_P\left(\text{var}(\widehat{h})\right)$	0.300 (0.057)	0.344 (0.060)	0.304 (0.053)
	$\overline{\widehat{h}}_B$ ()	0.164 (0.011)	0.174 (0.018)	0.203 (0.016)
	$\overline{\widehat{h}}_{BN}$ ()	0.604 (0.330)	0.700 (0.305)	0.653 (0.331)
$\mu = 5$	$n = 30$			
	n_{sim}	25	50	100
	$\overline{\widehat{h}}_P\left(\text{var}(\widehat{h})\right)$	0.448 (0.138)	0.39 2(0.148)	0.403 (0.125)
	$\overline{\widehat{h}}_B$ ()	0.140 (0.007)	0.146 (0.011)	0.142 (0.010)
	$\overline{\widehat{h}}_{BN}$ ()	0.620 (0.297)	0.614 (0.272)	0.736 (0.310)
	$n = 50$			
	n_{sim}	25	50	100
	$\overline{\widehat{h}}_P\left(\text{var}(\widehat{h})\right)$	0.248 (0.055)	0.242 (0.066)	0.330 (0.089)
	$\overline{\widehat{h}}_B$ ()	0.120 (0.005)	0.138 (0.008)	0.174 (0.019)
	$\overline{\widehat{h}}_{BN}$ ()	0.544 (0.221)	0.628 (0.209)	0.525 (0.227)

TAB. B.14 – Valeurs de $\overline{\widehat{h}}$ et de var$(\widehat{h})$ pour un nombre $n_{sim} \in \{25; 50; 100\}$ d'échantillons de données de simulées d'un Poisson $\mathcal{P}(\mu)$ de moyennes $\mu \in \{2; 5\}$ et de tailles $n \in \{30; 50\}$

$\mu = 2$	$n = 30$			
	n_{sim}	200	500	1000
	$\widehat{\overline{h}}_P\left(\text{var}(\widehat{h})\right)$	0.378 (0.092)	0.354 (0.077)	0.341 (0.078)
	$\widehat{\overline{h}}_B$ ()	0.233 (0.032)	0.209 (0.023)	0.216 (0.028)
	$\widehat{\overline{h}}_{BN}$ ()	0.653 (0.320)	0.671 (0.332)	0.638 (0.336)
	$n = 50$			
	n_{sim}	200	500	1000
	$\widehat{\overline{h}}_P\left(\text{var}(\widehat{h})\right)$	0.297 (0.044)	0.283 (0.055)	0.301 (0.053)
	$\widehat{\overline{h}}_B$ ()	0.180 (0.019)	0.193 (0.020)	0.188 (0.019)
	$\widehat{\overline{h}}_{BN}$ ()	0.675 (0.315)	0.621 (0.316)	0.635 (0.311)
$\mu = 5$	$n = 30$			
	n_{sim}	200	500	1000
	$\widehat{\overline{h}}_P\left(\text{var}(\widehat{h})\right)$	0.398 (0.140)	0.387 (0.130)	0.379 (0.131)
	$\widehat{\overline{h}}_B$ ()	0.136 (0.007)	0.147 (0.012)	0.140 (0.010)
	$\widehat{\overline{h}}_{BN}$ ()	0.591 (0.270)	0.657 (0.275)	0.672 (0.291)
	$n = 50$			
	n_{sim}	200	500	1000
	$\widehat{\overline{h}}_P\left(\text{var}(\widehat{h})\right)$	0.348 (0.082)	0.352 (0.087)	0.316 (0.080)
	$\widehat{\overline{h}}_B$ ()	0.136 (0.007)	0.149 (0.011)	0.148 (0.011)
	$\widehat{\overline{h}}_{BN}$ ()	0.555 (0.225)	0.582 (0.219)	0.535 (0.211)

TAB. B.15 – Suite et fin de Table B.14 pour $n_{sim} \in \{200; 500; 1000\}$

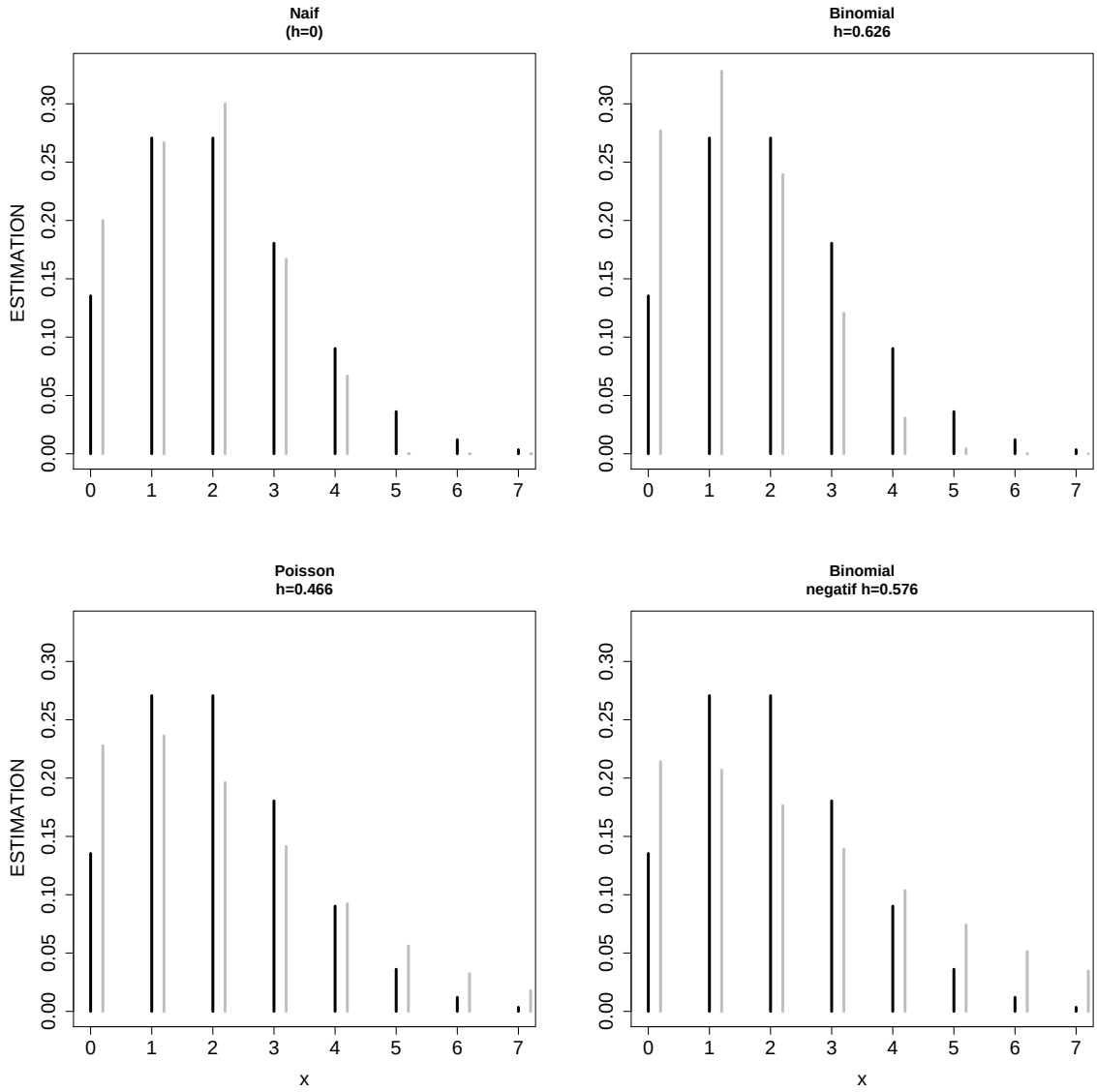

FIG. B.14 – Lissages discrets (bâtons gris) avec $h = h_{cv}$ par le noyau du type Dirac et les noyaux discrets binomial, de Poisson et binomial négatif de la distribution f de Poisson $\mathcal{P}(\mu)$ (bâtons noirs) avec $\mu = 2$ et $n = 30$

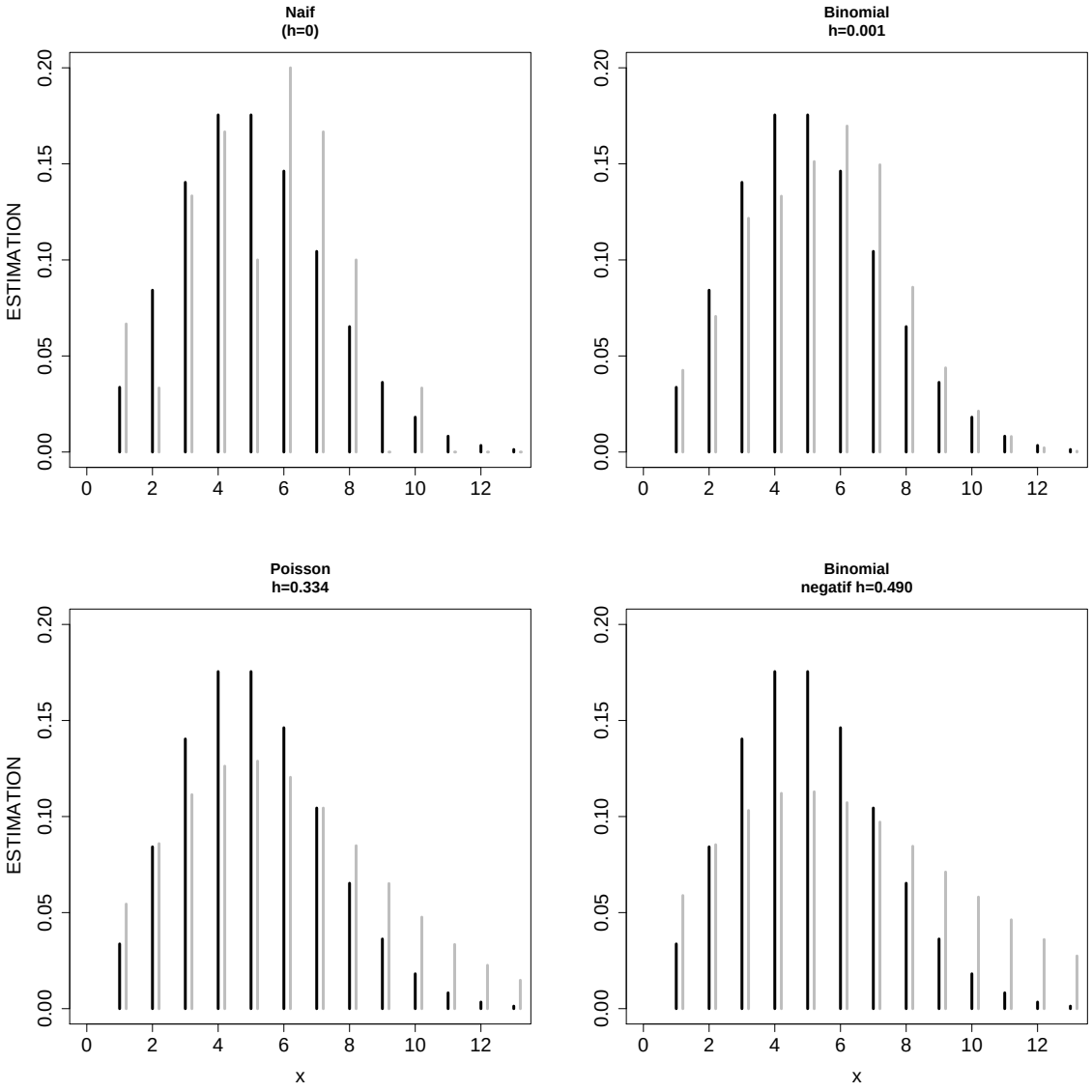

FIG. B.15 – Suite de Figure B.14 avec $\mu = 5$ et $n = 30$ (sans le point $x = 0$)

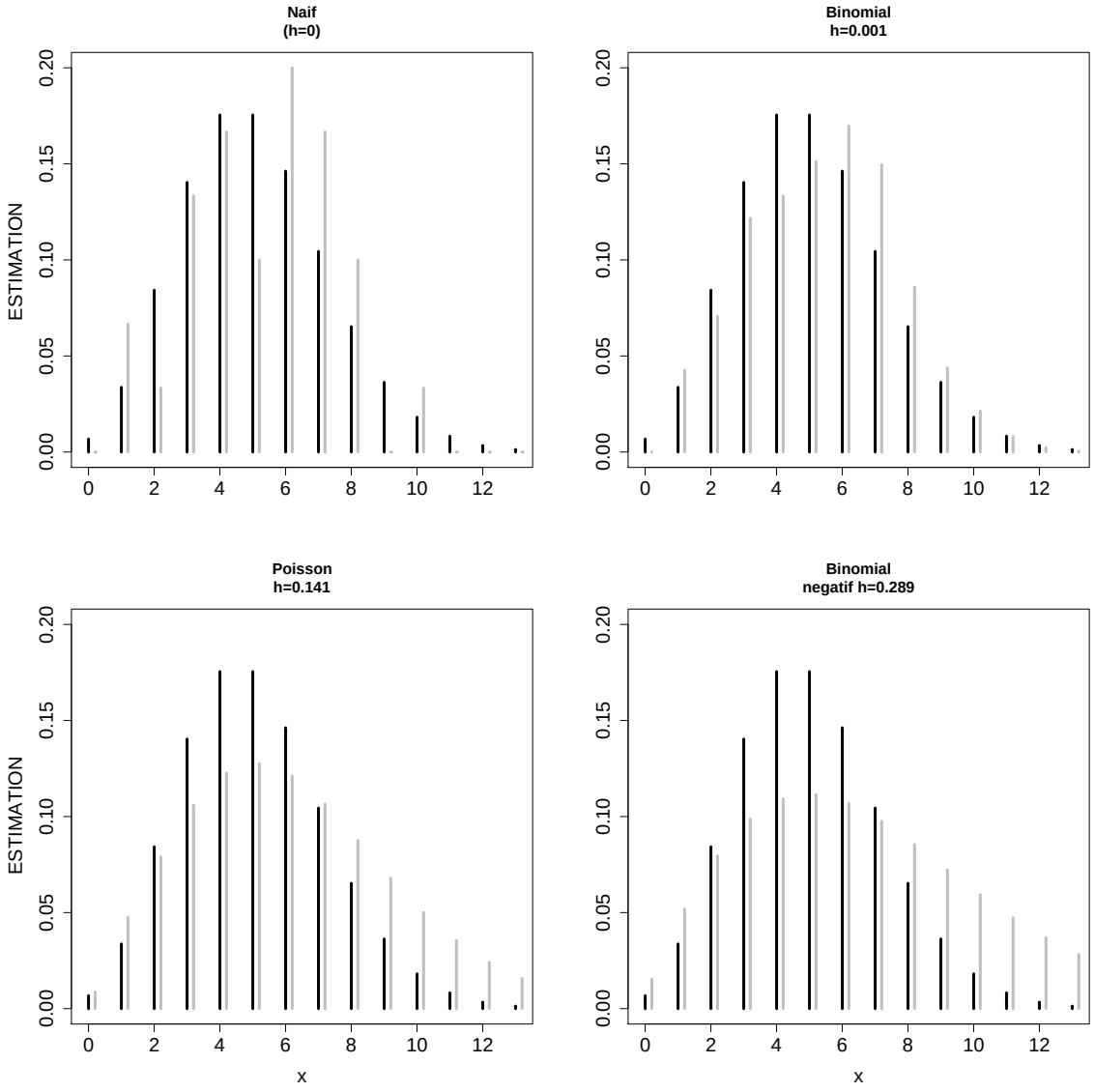

FIG. B.16 – Suite de Figure B.14 avec $\mu = 5$ et $n = 30$

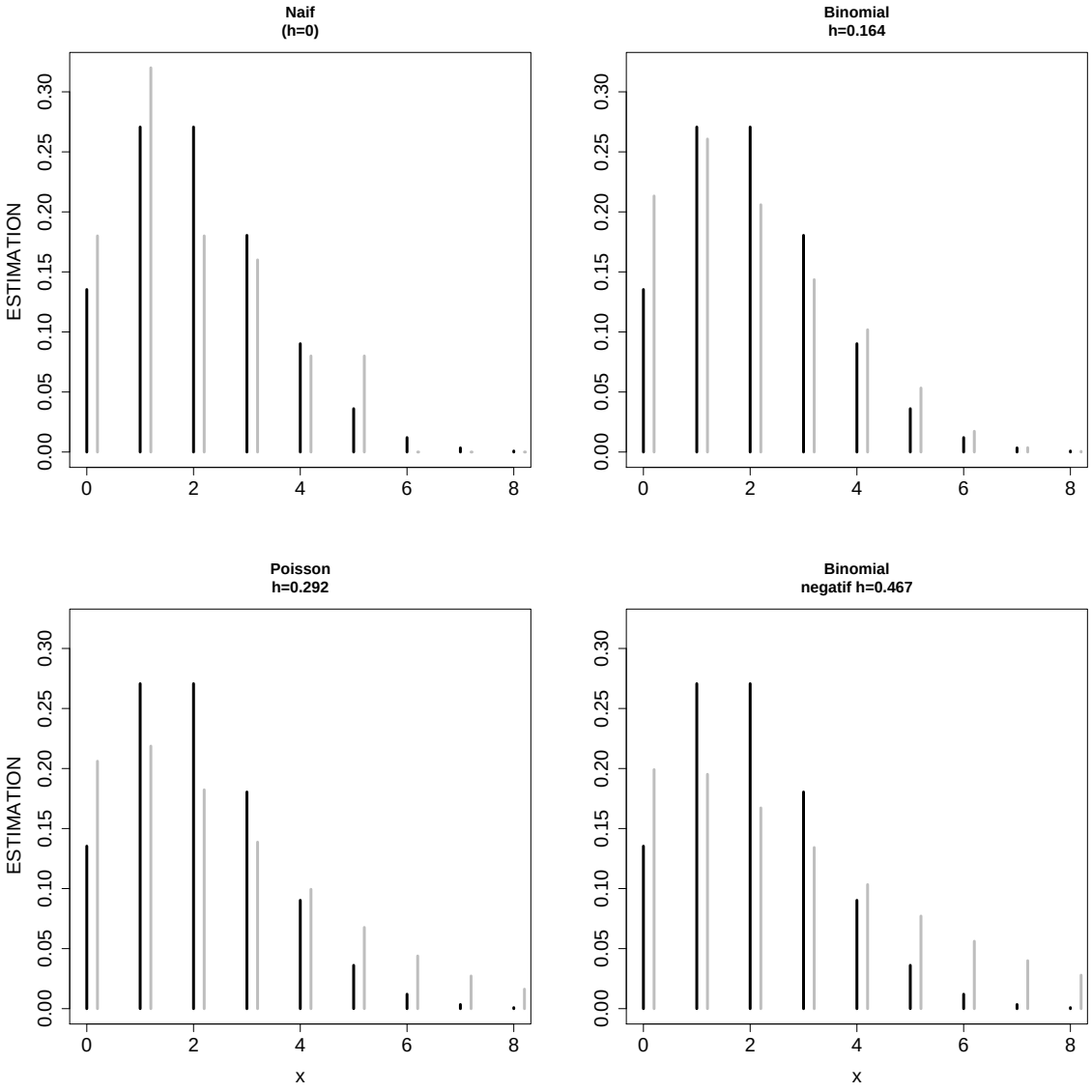

FIG. B.17 – Suite de Figure B.14 avec $\mu = 2$ et $n = 50$

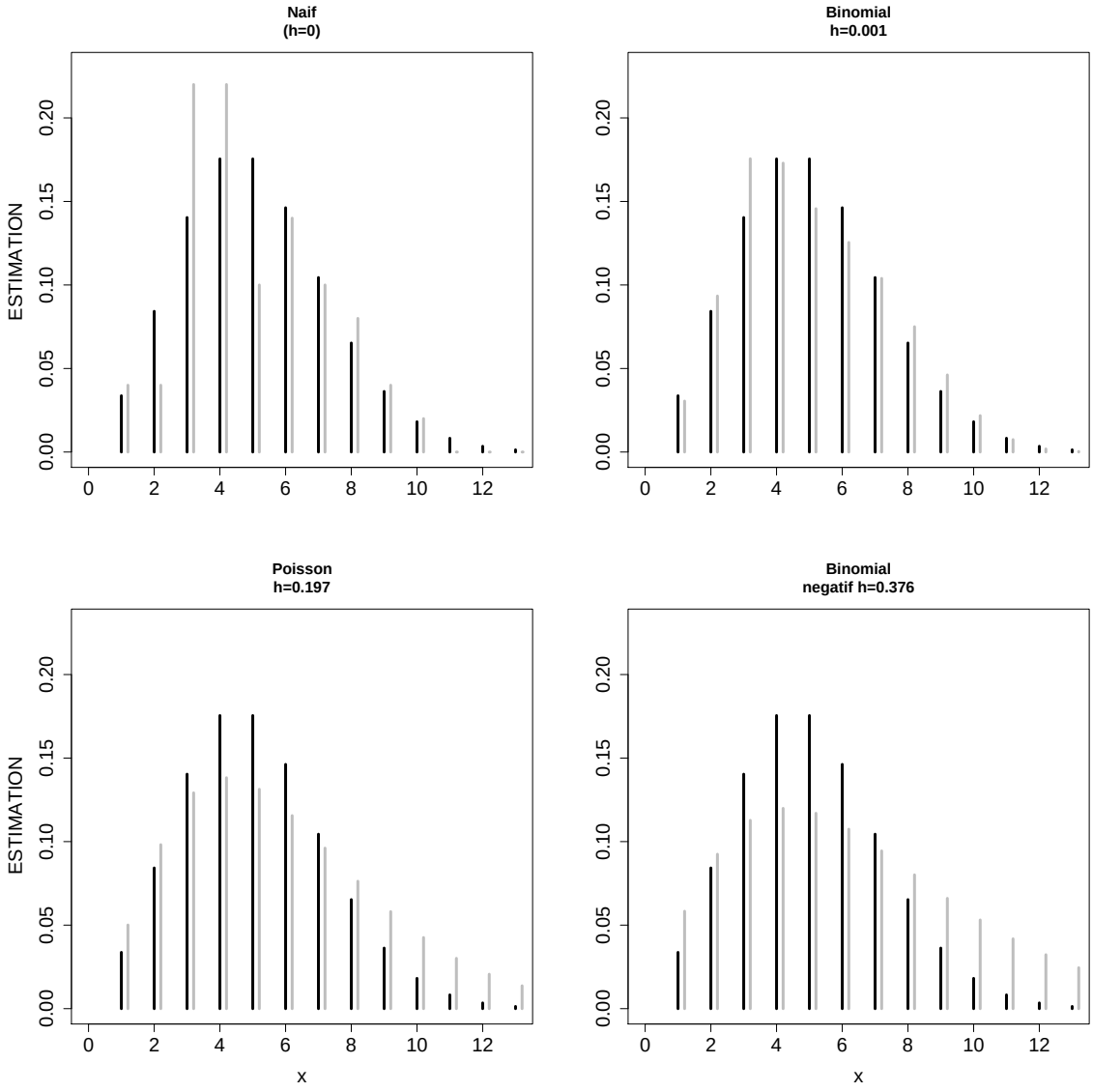

FIG. B.18 – Suite de Figure B.14 avec $\mu = 5$ et $n = 50$ (sans le point $x = 0$)

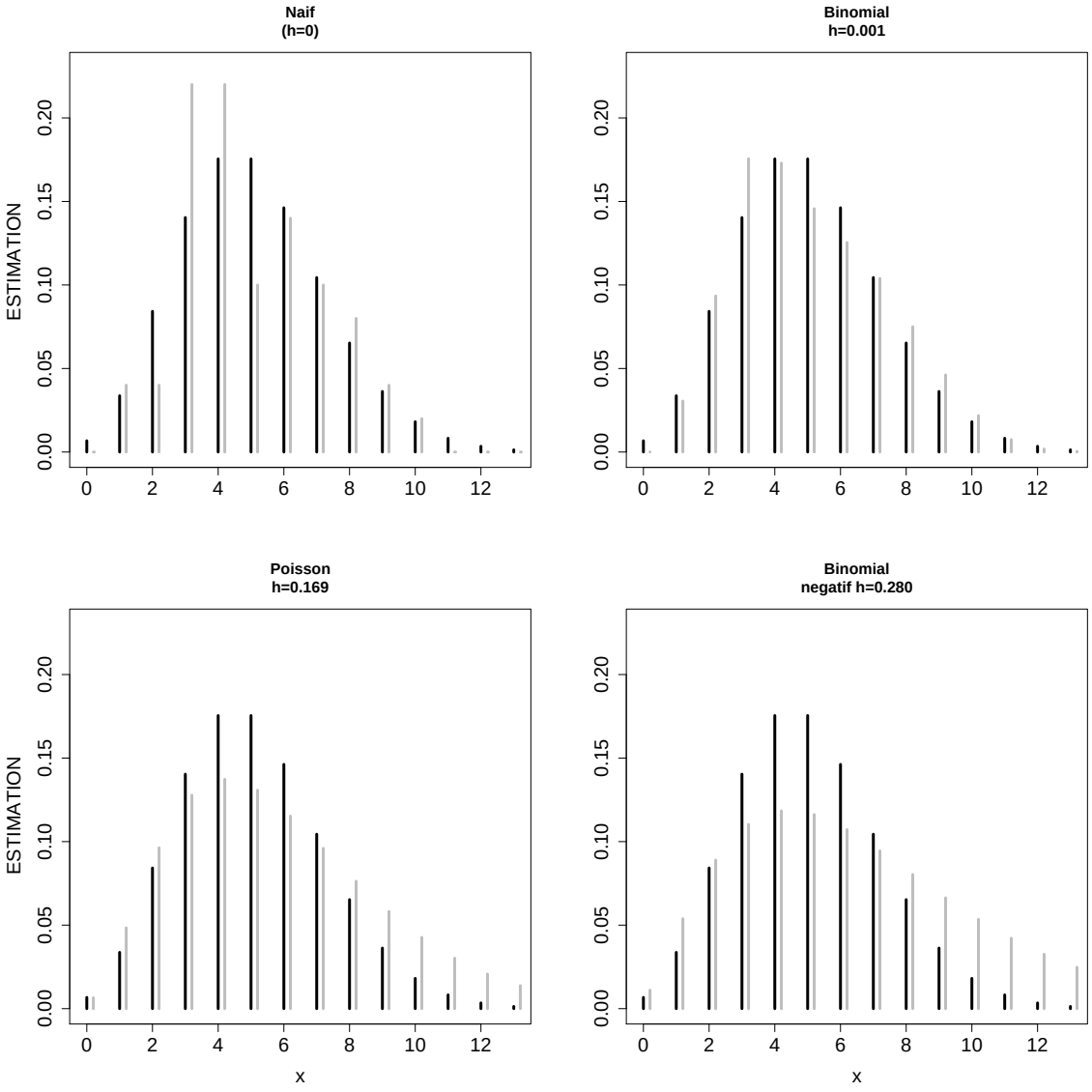

FIG. B.19 – Suite et fin de Figure B.14 avec $\mu = 5$ et $n = 50$

Annexe C

Programmes sous R

Sommaire :

C.1 Estimateurs à noyaux discrets standards

Nous présentons les programmes mis en place pour le choix de fenêtre par la méthode de validation croisée et pour les estimateurs à noyaux discrets standards sous le logiciel R. La méthode d'excès de zéros n'est pas évoquée dans toute cette partie car le calcul numérique qu'elle nécessite est facile.

C.1.1 Méthode de validation croisée par les moindres carrées

Le programme de validation croisée est donnée ici avec l'exemple du noyau binomial. Nous obtenons facilement les programmes concernant les noyaux discrets de Poisson et binomial négatif en les substituant au noyau binomial.

Description : Méthode de validation croisée avec le noyau binomial

Détails : La loi de probabilité binomiale de paramètres p et n se définit par

$$\Pr(z) = \text{choose}(n,z)*(p)\hat{}z*(1-p)\hat{}(n-z),$$

z = 0,1,..., n. Le noyau discret se construit avec p=(x+h)/(x+1) et n=x+1.

Arguments :
x : vecteur des points

h : paramètre de lissage discret
V : vecteur des observations de l'échantillon
N : effectifs des observations
n=sum(N) : nombre total d'observations = taille de l'échantillon

Usage :

```
Ycv= Y1 - Y2
plot(h,Ycv)
```

Code du 1er terme

Dans cette première partie, les observations X_i, $i = 1, 2, \ldots, n$ sont déclarées par le vecteur $V = (X_1, X_2, \ldots, X_n)$ et le vecteur N contient leurs effectifs tels que $N[1] + N[2] + \cdots + N[n] = n$.

Usage :

```
CV1bin=function(x,h,V,N,n)
CV1bin=edit(CV1bin,editor="nedit")
Y1=CV1bin(x,h,V,N,n)
```

```
function(x,V,N,n,h)
{   y=0
    S=rep(0,length(x))
    n=sum(N)
    for ( k in 1 : length(h))
    { m=rep(0,length(x)) ;
      for (i in 1 :length(x))          # boucle en i pour chaque point x[i]
      {for (j in 1 :length(N))         # boucle en j pour chaque observation V[j]
       {if(V[j]<=x[i]+1)          # Support {0, 1, ..., x + 1}
          { K= choose(x[i]+1,V[j])*((x[i]+h[k])/(x[i]+1))^(V[j])
              *((1-h[k])/(x[i]+1))^(x[i]+1-V[j])) # noyau binomial
            y=(N[j]/n)*K          # Estimation
          }
        m[i]=m[i]+y
      }
     }
   }
```

```
S[k]=sum(m^2)
     return(S)
   }
```

Code du 2ème terme

Dans cette deuxième partie, les observations X_i, $i = 1, 2, \ldots, n$ sont déclarées par le vecteur V tel que $V = (rep(X_1, N[1]), rep(X_2, N[2]), \ldots, rep(X_n, N[n]))$ avec $N[1] + N[2] + \cdots + N[n] = n$.

Usage :

```
CV2bin=function(x,h,N,n)
CV2bin=edit(CV1bin,editor="nedit")
Y2=CV2bin(x,h,N,n)
```

```
   function(x,V,N,n,h)
{  S2=rep(0,length(x))
     n=sum(N)
     for ( k in 1 : length(h))
     { S=rep(0,length(V))
      for (i in 1 :length(V))          # boucle en i pour chaque point x[i]
       {for (j in 1 :length(V))          # boucle en j pour chaque observation V[j]
        { if (i= !j)
         {if(V[j]<=V[i]+1)          # Support {0, 1, ..., x + 1}
          { K= choose(V[i]+1,V[j])*((V[i]+h[k])/(V[i]+1))^(V[j])
               *((1-h[k])/(V[i]+1))^(V[i]+1-V[j]) # noyau binomial
           S[i]=S[i]+(1/n)*(1/(n-1))*K
          }
         }
        }
       }
     S2[k]=2*sum(S)
     }
    return(S2)
   }
```

C.1.2 Estimateur à noyau Poisson

Description : Lissage discret d'une fonction de masse de probabilité par l'estimateur à noyau Poisson

Détails : La loi de probabilité discrète de Poisson se définit par

$$\Pr(z) = \text{lambda}\hat{}z \exp(\text{-lambda})/z !$$

pour z = 0, 1, 2, ... La moyenne et la variance sont égales à lambda. Le noyau discret se construit avec lambda=x+h.

Arguments :
x : vecteur des points
h : paramètre de lissage discret
V : vecteur des observations de l'échantillon
N : effectifs des observations
n=sum(N) : nombre total d'observations = taille de l'échantillon

Usage :

```
pois=function(x,h,V,N,n)
pois=edit(pois,editor="nedit")
Yp=pois(x,h,V,N,n)
```

Code de l'estimateur à noyau Poisson :

```
function(x,V,N,n,h)
{ y=0
   s=rep(0,length(x))
   n=sum(N)
   f0=c(N/n,rep(0,length(x)-length(N)))          # Estimateur fréquence
     for (i in 1 :length(x))          # boucle en i pour chaque point x[i]
     {for (j in 1 :length(N))          # boucle en j pour chaque observation V[j]
        { K=dpois(V[j],x[i]+h) # noyau Poisson
        y=(N[j]/n)*K          # Estimation
        }
      s[i]=s[i]+y
     }
   fn=s/sum(s)          # Estimations \widetilde{f}_n
   E=sum(s)          # Constante de normalisation C
   E[2]=sum((f0-fn)^2)          # ISE^0
```

```
 return(E)
}
```

C.1.3 Estimateur à noyau binomial

Description : Lissage discret d'une fonction de masse de probabilité par l'estimateur à noyau discret binomial

Détails : La loi de probabilité binomiale de paramètres p et n se définit par

$$\Pr(z) = \text{choose}(n,z)*(p)\hat{}z*(1-p)\hat{}(n-z),$$

z = 0,1,..., n. Le noyau discret se construit avec p=(x+h)/(x+1) et n=x+1.

Arguments :
x : vecteur des points
h : paramètre de lissage discret
V : vecteur des observations de l'échantillon
N : effectifs des observations
n=sum(N) : nombre total d'observations = taille de l'échantillon

Usage :

```
binom=function(x,h,V,N,n)
binom=edit(binom,editor="nedit")
Yb=binom(x,h,V,N,n)
```

Code de l'estimateur à noyau binomial :

```
function(x,V,N,n,h)
{   y=0
    s=rep(0,length(x))
    n=sum(N)
    f0=c(N/n,rep(0,length(x)-length(N)))          # Estimateur fréquence
      for (i in 1 :length(x))          # boucle en i pour chaque point x[i]
      {for (j in 1 :length(N))          # boucle en j pour chaque observation V[j]
       {if(V[j]<=x[i]+1)          # Support {0, 1, ..., x + 1}
          { K= choose(x[i]+1,V[j])*((x[i]+h)/(x[i]+1))^(V[j])
               *((1-h)/(x[i]+1))^(x[i]+1-V[j])) # noyau binomial
          y=(N[j]/n)*K          # Estimation
          }
        s[i]=s[i]+y
      }
```

```
    }
  fn=s/sum(s)          # Estimations f̃_n
  E=sum(s)          # Constante de normalisation C
  E[2]=sum((f0-fn)^2)          # ISE^0
  return(E)
}
```

C.1.4 Estimateur à noyau binomial négatif

Description : Lissage discret d'une fonction de masse de probabilité par l'estimateur à noyau binomial négatif

Détails : La loi de probabilité discrète binomiale négative se définit par

$$\Pr(z) = \text{Gamma(z+n)/}\big(\text{Gamma(n)z !}\big)\text{(1-p)\^{}z p\^{}n}$$

pour z = 0, 1, 2, ..., n > 0 et $0 < p \leq 1$. Le noyau discret se construit avec p=(x+1)/(2x+1+h) et n=x+1.

Arguments :
x : vecteur des points
h : paramètre de lissage discret
V : vecteur des observations de l'échantillon
N : effectifs des observations
n=sum(N) : nombre total d'observations = taille de l'échantillon

Usage :

```
binneg=function(x,h,V,N,n)
binneg=edit(binneg,editor="nedit")
Ybn=binneg(x,h,V,N,n)
```

Code de l'estimateur à noyau binomial négatif :

```
  function(x,V,N,n,h)
{  y=0
    s=rep(0,length(x))
    n=sum(N)
    f0=c(N/n,rep(0,length(x)-length(N)))          # Estimateur fréquence
      for (i in 1 :length(x))          # boucle en i pour chaque point x[i]
      {for (j in 1 :length(N))          # boucle en j pour chaque observation V[j]
        { K=gamma(V[j]+x[i]+1)/(gamma(x[i]+1)*gamma(V[j]+1))
```

```
                *((x[i]+h)/(2*x[i]+1+h))^(V[j])*((x[i]+1)/(2*x[i]+1+h))^((x[i]+1))
# noyau binomial négatif
              y=(N[j]/n)*K          # Estimation
             }
           s[i]=s[i]+y
         }
    fn=s/sum(s)          # Estimations f̃_n
    E=sum(s)          # Constante de normalisation C
    E[2]=sum((f0-fn)^2)          # ISE^0
    return(E)
  }
```

C.2 Noyaux associés discrets triangulaires

Pour les noyaux associés discrets triangulaires, nous présentons la mise en œuvre de la méthode de validation croisée et de l'estimateur à noyau discret.

C.2.1 Méthode de validation croisée par les moindres carrées

Description : Méthode de validation croisée avec les noyaux associés discrets triangulaires

Détails :

La loi de probabilité discrète triangulaire d'ordre h, de bras a et de centre x se définit par

$$\Pr(z) = \left((a+1)^h - \left(\text{abs}(z-x)\right)^h\right)/A,\ z = x \pm 1, x \pm 2, \ldots, x \pm a,$$

où A=(2*a+1)*(a+1)^h-2*sum(k^h), k=1,2,...,a, est la constante de normalisation.

Arguments :

x : vecteur des points
h : paramètre de lissage discret
a : bras (paramètre)
V : vecteur des observations de l'échantillon
N : effectifs des observations
n=sum(N) : nombre total d'observations = taille de l'échantillon

Usage :

```
Ycv= Y1 - Y2
```

```
plot(h,Ycv)
```

Code du 1er terme

Dans cette première partie, les observations X_i, $i = 1, 2, \ldots, n$ sont déclarées par le vecteur $V = (X_1, X_2, \ldots, X_n)$ et le vecteur N contient leurs effectifs tels que $N[1] + N[2] + \cdots + N[n] = n$.

Usage :

```
CV1trng=function(x,h,V,N,n,a)
CV1trng=edit(CV1trng,editor="nedit")
Y1=CV1trng(x,h,V,N,n,a)
```

```
    function(x,h,V,N,n,a)
{   y=0
      S=rep(0,length(x))
      n=sum(N)          # Taille de l'échantillon
      A=rep(0,length(h)) ; # Constante de normalisation de la ditribution triangulaire
      if (a==0)
       { u=0 ;
       }
      else
       { for (i in 1 :a)
       {
       u=u+(i^h)
       }
       }
       A=((2*a+1)*(a+1)^h)-2*u
      for ( k in 1 : length(h))
      { m=rep(0,length(x))
        for (i in 1 :length(x))          # boucle en i pour chaque point x[i]
        {for (j in 1 :length(N))          # boucle en j pour chaque observation V[j]
          {if (V[j]≥(x[i]-a) & V[j]≤(x[i]+a))        # Support {x ± 1, ..., x ± a}
           {K=((a+1)^h[k] - (abs(V[j]-x[i]))^h[k])/A[k] # Noyau associé discret
triangulaire
```

```
            y=(N[j]/n)*K          # Estimation
           }
         m[i]=m[i]+y
        }
       }
     }
S[k]=sum(m^2)
    return(S)
  }
```

Code du 2ème terme

Dans cette deuxième partie, les observations X_i, $i = 1, 2, \ldots, n$ sont déclarées par le vecteur V tel que $V = (rep(X_1, N[1]), rep(X_2, N[2]), \ldots, rep(X_n, N[n]))$ avec $N[1] + N[2] + \cdots + N[n] = n$.

Usage :

```
CV2trng=function(x,h,N,n,a)
CV2trng=edit(CV1trng,editor="nedit")
Y2=CV2trng(x,h,N,n,a)
```

```
  function(x,V,N,n,h)
{  S2=rep(0,length(x))
    n=sum(N)          # Taille de l'échantillon
    A=rep(0,length(h)) ; # Constante de normalisation de la ditribution triangulaire
    if (a==0)
     { u=0 ;
     }
    else
     { for (i in 1 :a)
     {
     u=u+(i^h)
     }
     }
     A=((2*a+1)*(a+1)^h)-2*u
    for ( k in 1 : length(h))
    { S=rep(0,length(V))
```

```
    for (i in 1 :length(V))          # boucle en i pour chaque point x[i]
     {for (j in 1 :length(V))        # boucle en j pour chaque observation V[j]
      { if (i= !j)
       {if (V[j]≥(x[i]-a) & V[j]≤(x[i]+a))      # Support {x ± 1, ..., x ± a}
        {K=((a+1)^h[k] - (abs(V[j]-x[i]))^h[k])/A[k]
        S[i]=S[i]+(1/n)*(1/(n-1))*K
        }
       }
      }
     }
    S2[k]=2*sum(S)
    }
  return(S2)
}
```

C.2.2 Estimateurs à noyaux associés discrets triangulaires

Description : Lissage discret d'une fonction de masse de probabilité par l'estimateur à noyau associé discret triangulaire

Détails :

La loi de probabilité discrète triangulaire d'ordre h, de bras a et de centre x se définit par

$$\Pr(z)= \left((a+1)^h - \left(\text{abs}(z-x)\right)^h\right)/A,\ z=x\pm1,x\pm2,\ldots,x\pm a,$$

où A=(2*a+1)*(a+1)^h-2*sum(k^h), k=1,2,...,a, est la constante de normalisation.

Arguments :

x : vecteur des points
h : paramètre de lissage discret
a : bras (paramètre)
V : vecteur des observations de l'échantillon
N : effectifs des observations
n=sum(N) : nombre total d'observations = taille de l'échantillon

Usage :

```
trng=function(x,h,V,N,n,a)
```

```
trng=edit(trng,editor="nedit")
Y=trng(x,h,V,N,n,a)
```

Code de l'estimateur à noyau associé discret triangulaire avec le bras a :

```
function(x,a,V,N,n,h)
{ y=0
  s=rep(0,length(x))
  n=sum(N)
  f0=c(N/n,rep(0,length(x)-length(N)))          # Estimateur fréquence
  u=0
  m=0
    for (k in 1 :a)
    { m=k^h
    u=u+m
    }
  A=(2*a+1)*(a+1)^h-2*u          # Constante de normalisation P(a,h)
    for (i in 1 :length(x))          # boucle en i pour chaque point x[i]
     {for (j in 1 :length(N))          # boucle en j pour chaque observation V[j
      {if (V[j]≥(x[i]-a) & V[j]≤(x[i]+a))        # Support {x ± 1, ..., x ± a}
       {K=((a+1)^h - (abs(V[j]-x[i]))^h)/A        # Noyau associé discret triangulaire
        y=(N[j]/n)*K          # Estimation
       }
      else{
      y=0
          }
     s[i]=s[i]+y
     }
    }
  fn=s/sum(s)          # Estimations f̃_n
  E=sum(s)          # Constante de normalisation C
  E[2]=sum((f0-fn)^2)          # ISE^0
  return(E)
```

```
    }
```

Code de l'estimateur à noyau associé discret triangulaire avec le bras modifié a_0=1 :

```
    function(x,a,V,N,n,h)
{ y=0
      s=rep(0,length(x))
      n=sum(N)
      f0=c(N/n,rep(0,length(x)-length(N)))          # Estimateur fréquence
      u=0
      for (i in 1 :length(x))          # boucle en i pour chaque point x[i]

       { if (i==1)
        {a=0
         }
         else
         {a=1
         }
         if (a==0)
         {
         u=0
         }
         else
         {
           for (k in 1 :a)
           { m=k^h
           u=u+m
           }
         }

      A=(2*a+1)*(a+1)^h-2*u          # Constante de normalisation P(a,h)
          {for (j in 1 :length(N))          # boucle en j pour chaque observation V[j]

            {if (V[j]≥(x[i]-a) & V[j]≤(x[i]+a))          # Support {x ± 1, . . . , x ± a}
             {K=((a+1)^h - (abs(V[j]-x[i]))^h)/A          # Noyau associé discret trian-
gulaire
```

```
            y=(N[j]/n)*K          # Estimation
           }
          else{
          y=0
              }
       s[i]=s[i]+y
        }
      }
   fn=s/sum(s)          # Estimations f̃_n
   E=sum(s)          # Constante de normalisation C
   E[2]=sum((f0-fn) ^2)          # ISE^0
   return(E)
}
```

C.3 Estimateur semi-paramétrique

Dans cette section, nous présentons les programmes relatifs au noyau binomial comme exemple. Ainsi, à partir des programmes des Sections C.1 et C.2, nous obtenons les programmes utilisant les noyaux de Poisson et binomial négatif ainsi que les noyaux associés discrets triangulaires.

C.3.1 Méthode de validation croisée par les moindres carrées

Description : Méthode de validation croisée en utilisant le noyau discret binomial

Arguments :
x : vecteur des points
h : paramètre de lissage discret
V : vecteur des observations de l'échantillon
N : effectifs des observations
n=sum(N) : nombre total d'observations = taille de l'échantillon

Usage :

```
Ycv= Y1 - Y2
plot(h,Ycv)
```

Code du 1er terme

Dans cette première partie, les observations X_i, $i = 1, 2, \ldots, n$ sont déclarées par le vecteur $V = (X_1, X_2, \ldots, X_n)$ et le vecteur N contient leurs effectifs tels que $N[1] + N[2] + \cdots + N[n] = n$.

Usage :

```
CV1bin=function(x,h,V,N,n)
CV1bin=edit(CV1bin,editor="nedit")
Y1=CV1bin(x,h,V,N,n)
```

```
function(x,V,N,n,h)
{ y=0
   S=rep(0,length(x))
   n=sum(N)
  mu=sum(V*N)/n
   p=mu/max(x)
   for ( k in 1 : length(h))
   { m=rep(0,length(x)) ;
     for (i in 1 :length(x))          # boucle en i pour chaque point x[i]
     {for (j in 1 :length(N))         # boucle en j pour chaque observation V[i]
      {for (l in 1 :length(N)) # boucle en l pour chaque observation V[l]
       {if ((V[j]<=x[i]+1) & (V[l]<=(x[i]+1)))          # Support {0, 1, ..., x + 1}
        { K1= choose(x[i]+1,V[j])*((x[i]+h)/(x[i]+1))^(V[j])
              *((1-h)/(x[i]+1))^(x[i]+1-V[j])) # noyau binomial
          K2= choose(x[i]+1,V[l])*((x[i]+h)/(x[i]+1))^(V[l])
              *((1-h)/(x[i]+1))^(x[i]+1-V[l])) # noyau binomial
          U=1/(dpois(V[l],mu)*dpois(V[j],mu))
          W=(dpois(x[i],mu))^2
          y= ((N[l]*N[j])/(n^2))*U*W*K1*K2          # Estimation
        }
       m[i]=m[i]+y
      }
     }
   }
```

```
 }

  S[k]=sum(m^2)
 return(S)
}
```

Code du 2ème terme

Dans cette deuxième partie, les observations X_i, $i = 1, 2, \ldots, n$ sont déclarées par le vecteur V tel que $V = (rep(X_1, N[1]), rep(X_2, N[2]), \ldots, rep(X_n, N[n]))$ avec $N[1] + N[2] + \cdots + N[n] = n$.

Usage :

```
CV2bin=function(x,h,N,n)
CV2bin=edit(CV1bin,editor="nedit")
Y2=CV2bin(x,h,N,n)
```

```
function(x,V,N,n,h)
{  S2=rep(0,length(x))
   n=sum(N)          # Taille de l'échantillon
   for ( k in 1 : length(h))
   { S=rep(0,length(V))
    for (i in 1 :length(V))          # boucle en i pour chaque point x[i]
     {mu1=(sum(V) - V[i])/(n-1)
     for (j in 1 :length(V))          # boucle en j pour chaque observation V[j]
      { mu2=(sum(V) - V[j])/(n-1)
      if (i= !j)
       {if(V[j]<=(V[i]+1))          # Support {0, 1, ..., x + 1}
        { K= choose(V[i]+1,V[j])*((V[i]+h[k])/(V[i]+1))^(V[j])
             *((1-h[k])/(V[i]+1))^(V[i]+1-V[j])) # noyau binomial
        U=(gamma(max(x)-V[j]+1))/(gamma(N-x[i]+1))*(p^(x[i]-V[j]))
             *((1-p)^(V[j]-x[i]))
        U=(dpois(V[i],mu1))/(dpois(V[j],mu2))
        S[i]=S[i]+(1/n)*(1/(n-1))*U*K
        }
      }
     }
```

```
        }
      S2[k]=2*sum(S)
      }
    return(S2)
  }
```

C.3.2 Estimateur à noyau binomial

Description : Estimation semi-paramétrique avec un noyau binomial

Arguments :

x : vecteur des points
h : paramètre de lissage discret
V : vecteur des observations de l'échantillon
N : effectifs des observations
n=sum(N) : nombre total d'observations = taille de l'échantillon

Usage :

```
binom=function(x,h,V,N,n)
binom=edit(binom,editor="nedit")
Yb=binom(x,h,V,N,n)
```

Code de l'estimateur à noyau binomial :

```
function(x,V,N,n,h)
{ y=0
    s=rep(0,length(x))
    n=sum(N)          # Taille de l'échantillon
    mu=sum(V*N)/n
    f0=c(N/n,rep(0,length(x)-length(N)))          # Estimateur fréquence
      for (i in 1 :length(x))          # boucle en i pour chaque point x
      {for (j in 1 :length(N))          # boucle en j pour chaque observation V
       {if(V[j]<=x[i]+1)          # Support {0, 1, ..., x + 1}
          { K= choose(x[i]+1,V[j])*((x[i]+h)/(x[i]+1))^(V[j])
              *((1-h)/(x[i]+1))^(x[i]+1-V[j]) # noyau binomial
          U=dpois(x[i],mu)/(dpois(V[j],mu))
          y=(N[j]/n)*K*U          # Estimation
          }
```

```
        s[i]=s[i]+y
       }
      }
  fn=s/sum(s)          # Estimations f̃_n
  E=sum(s)          # Constante de normalisation C
  E[2]=sum((f0-fn)^2)          # ISE^0
  return(E)
 }
```

C.4 Régression non-paramétrique

Comme dans la section précédente, nous présentons les programmes relatifs au noyau binomial. De là, on retrouve les programmes correspondants à l'utilisation des noyaux de Poisson, binomial négatif et triangulaires discrets pour l'estimation non-paramétrique de la fonction discrète de régression.

C.4.1 Méthode de validation croisée par les moindres carrées

Description : Méthode de validation croisée avec le noyau binomial

Arguments :
x : vecteur des points
h : paramètre de lissage discret
V : vecteur des observations de l'échantillon
N : effectifs des observations
n=sum(N) : nombre total d'observations = taille de l'échantillon

Usage :

```
CV1bin=function(x,h,V,N,n)
CV1bin=edit(CV1bin,editor="nedit")
Ycv=CV1bin(x,h,V,N,n)

 function(x,V,N,n,h)
{  A=0
    B=0
    n=sum(N)
    s=rep(0,length(x))
    S=rep(0,length(x))
    m=rep(0,length(V))
    w=rep(0,length(V))
```

```
 U=rep(0,length(h))
 n=sum(N)
 for ( k in 1 : length(h))
 {   S=rep(0,length(V))
s=rep(0,length(V))
m=rep(0,length(V))
w=rep(0,length(V))
    for (i in 1 :length(x))          # boucle en i pour chaque point x[i]
  {for (j in 1 :length(N))           # boucle en j pour chaque observation V[j]
    {if (i !=j)
     {if(V[j]<=V[i]+1)           # Support {0, 1, ..., x + 1}
        { K= choose(x[i]+1,V[j])*((V[i]+h[k])/(V[i]+1))^(V[j])
             *((1-h[k])/(V[i]+1))^(V[i]+1-V[j])) # noyau binomial
         A=K
          B=K*y[j]
          s[i]=s[i]+A ;
          S[i]=S[i]+B ;
         }
      }
    }
       m[i]=choose(V[i]+1,V[i])*((V[i]+h[k])/(V[i]+1))^V[i]
            *((1-h[k])/(V[i]+1))^(V[i]+1-V[i])
       w=m/sum(m)
        F=S/s
   }
 U[k]=(1/n)*sum(((y-F)^2)*w)
 }
return(U)
}
```

C.4.2 Estimateur à noyau binomial

Description : Lissage discret de la fonction discrète de régression par l'estimateur à noyau binomial

Arguments :
x : vecteur des points
h : paramètre de lissage discret
V : vecteur des observations de l'échantillon
N : effectifs des observations
n=sum(N) : nombre total d'observations = taille de l'échantillon

Usage :

```
regbinom=function(x,h,V,N,n)
regbinom=edit(regbinom,editor="nedit")
Ybin=regbinom(x,h,V,N,n)
```

Code de l'estimateur à noyau binomial :

```
function(x,V,N,n,h)
{   A=0
    B=0
    n=sum(N)
    s=rep(0,length(x))
    S=rep(0,length(x))
    m=rep(0,length(V))
    R=0
      for (i in 1 :length(x))          # boucle en i pour chaque point x[i]
      {for (j in 1 :length(N))          # boucle en j pour chaque observation V[j]
       {if(V[j]<=x[i]+1)          # Support {0, 1, ..., x + 1}
          { K= choose(x[i]+1,V[j])*((x[i]+h)/(x[i]+1))^(V[j])
               *((1-h)/(x[i]+1))^(x[i]+1-V[j])) # noyau binomial
         A=(N[j]/n)*K # Estimation
          B=(N[j]/n)*K*y[j]
          s[i]=s[i]+A
          S[i]=S[i]+B
         }
       }
     fn=s/sum(s) # vecteur des estimations normalisées
      E=sum(s) # Constante de normalisation C
      F=S/s # Estimation de la fonction de régression discrète
      m[i]=choose(V[i]+1,V[i])*((V[i]+h)/(V[i]+1))^V[i]*((1-h)/(V[i]+1))^(V[i]+1-V[i])
```

```
        moyY=sum(y)/n # moyenne des observations y
        w=m/sum(m)
        R=sum(((F-moyY)^2*w))/sum(((y-moyY)^2)*w) # Coefficient de détermination
R^2
       }
  return(F)
 }
```