Sur l'estimation non paramétrique de la fonction d'égalisation équipercentile. Application à la qualité de vie.

Kaouthar El Fassi

THÈSE DE DOCTORAT DE l'UNIVERSITÉ PIERRE ET MARIE CURIE

Spécialité : Mathématiques

Option : Statistique

présentée par :

Kaouthar EL FASSI

pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ PIERRE ET MARIE CURIE - PARIS VI

Sujet de la thèse :

Sur l'estimation non-paramétrique de la fonction d'"Égalisation" Équipercentile. Application à la qualité de vie.

Soutenue publiquement le 3 juin 2009 devant le jury composé de :

M. Anouar BENMALEK	Université Paris XI	*Examinateur*
M. Michel BRONIATOWSKI	Université Paris VI	*Examinateur*
M. Paul DEHEUVELS	Université Paris VI	*Président du jury*
M. Mounir MESBAH	Université Paris VI	*Directeur de thèse*
M. Geert MOLENBERGHS	Hasselt University - Belgium	*Rapporteur*
M. Mikhail NIKULIN	Université Bordeaux 2	*Rapporteur*

À mes parents,

à mon frère

et à mon mari,

Je dédie cette thèse.

Remerciements

Pour souscrire à la tradition, je vais me soumettre au difficile exercice de la rédaction des remerciements de thèse. Exprimer sa gratitude aux personnes qui vous ont entouré et soutenu au cours de ces années est tout naturel. Il ne s'agit aucunement d'une simple formalité, mais de sincères sentiments qui proviennent du fond du coeur. Cependant, la difficulté de cette tâche réside essentiellement dans le fait de n'oublier personne. Ainsi, je commence par remercier tous ceux dont le nom n'apparaît pas sur ces pages par omission et qui m'ont aidée d'une manière ou d'une autre dans mon parcours.

Je tiens en premier abord à exprimer mes remerciements à Monsieur Mounir Mesbah, Professeur à l'Université Pierre et Marie Curie, d'avoir accepté de diriger cette thèse. Je le remercie également pour la confiance qui m'a témoignée au cours de ces années. Les conférences et séminaires auxquels il m'a encouragée à participer, les rencontres que j'y ai faites constituent une richesse scientifique et personnelle dont je lui suis reconnaissante.

Je remercie Monsieur Paul Deheuvels, Directeur du Laboratoire de Statistique Théorique et Appliquée, qui m'a fait l'honneur de présider ce jury. Je le remercie aussi de m'avoir accueillie au sein de son laboratoire et de m'avoir permis de réaliser mes projets.

J'exprime mes vifs remerciements à Monsieur Geert Molenberghs, Professeur à l'Université de Hasselt en Belgique et à Monsieur Mikhail Nikulin, Professeur à l'université de Bordeaux 2, d'avoir accepté de rapporter sur ce travail. Je les remercie de l'intérêt qu'ils ont éprouvé pour cette recherche, de leurs pertinentes remarques et conseils qui se sont révélés très enrichissants. Aussi, je voudrais leur exprimer ma gratitude pour leur réactivité et la célérité de leur réponse. Je remercie également Messieurs Michel Broniatowski, Professeur à l'Université Paris VI et Anouar Benmalek, Maître de conférence à la Faculté de Pharmacie - Paris XI pour avoir bien voulu être membre de mon jury de thèse.

Pour ses précieux conseils et ses grandes disponibilité et générosité, je tiens à exprimer toute ma gratitude à Monsieur Belkacem Abdous, Professeur à l'Université Laval, Québec, Canada. Il m'a judicieusement guidée durant mes travaux de recherche et il a su orienter mes recherches aux bons moments et m'a permis de les conduire à terme. Qu'il trouve également ici l'expression de ma reconnaissance pour la qualité de sa collaboration, sa rigueur scientifique et sa compétence qui susciteront à jamais mon respect et mon admiration.

La réalisation de ce travail n'aurait pas été possible sans la contribution et le soutien de nombreuses personnes du laboratoire.
J'adresse mes remerciements, pour la sympathie qu'ils m'ont témoignée, aux professeurs et maîtres de conférence du L.S.T.A., Monsieur Gérard Biau, Monsieur Jérôme Dedecker, Monsieur Amor Keziou, Monsieur Djamal Louani, Monsieur Giovanni Peccati. Je remercie également Monsieur Philippe Saint-Pierre pour ses nombreux conseils et remarques, ainsi que pour toutes les discussions fructueuses que nous avons eues.

Je tiens aussi à remercier les secrétaires du laboratoire Madame Louise Lamart et Madame Anne Durande pour avoir toujours réglé efficacement les problèmes administratifs et pratiques avec le sourire et une bonne humeur communicative.

J'exprime ma profonde sympathie à mes collègues du laboratoire : Aboubacar, Amadou, Boris, Claire, David, Issam, Khalid, Lynda, Mamadou, Mohamed Ali, Mory, Nabil, Nour-eddine, Olivier B., Olivier F., Ousmane, Salim, Tarek, pour leur bonne humeur, leur écoute aussi bien scientifique qu'amicale ! Je leur souhaite une bonne continuation dans leurs travaux de recherche. Un merci particulier à François-Xavier pour son concours efficace et compétent, son soutien constant et sa perpétuelle bonne humeur. Je souhaite également remercier Lahcen pour son aide précieuse lors de mes débuts de thèse et d'être toujours disponible pour répondre à mes questions. J'ai également une pensée pour Anissa qui m'a offert son amitié cette dernière année. J'espère que notre amitié durera encore plus longtemps. Je ne pourrai oublier de remercier Boutayna et Véronique avec qui j'ai partagé des moments de galère et de joie,

au laboratoire ou ailleurs. Je leur souhaite de tout mon coeur une bonne continuation dans leur vie personnelle et professionnelle.

Mes remerciements vont à tous mes amis et en particulier à Asmaa pour son indéfectible amitié depuis toujours, pour sa confiance et son soutien fraternel malgré la distance qui nous séparait.

Que serais-je devenue sans ma famille et leur soutien ? merci à vous papa, maman et mon frère de m'avoir soutenue moralement et financièrement et encouragée de manière inconditionnelle en m'offrant les meilleures conditions afin d'en être là aujourd'hui. Malgré l'éloignement géographique vous avez su être toujours près de moi. Merci de vous être déplacés pour assister à ma soutenance.

Des remerciements tous particuliers à mes beaux-parents pour leur soutien et leurs prières dans les moments difficiles.

Mes remerciements les plus amoureux vont à Idriss, la personne avec qui je partage ma vie. Merci d'être entré dans ma vie, pas forcément dans la période la plus sereine, d'avoir accepté l'éloignement et le peu de disponibilité induits par cette thèse, d'avoir su écouter mes lamentations et de m'avoir apaisée dans les moments les plus difficiles.

Merci à tous.

Kaouthar EL FASSI

Table des matières

Introduction

Nous nous intéressons dans ce travail au problème de l'estimation non-paramétrique de la fonction d'égalisation équipercentile $G^{-1} \circ F$ entre deux scores observés X et Y, où F est la fonction de répartition de la variable aléatoire X et G^{-1} la fonction quantile de la variable aléatoire Y. Pour une variable d'intérêt, le but est d'établir l'équivalence entre des scores d'une échelle de mesure de référence et des scores obtenus à partir d'une autre échelle de mesure. Les scores peuvent être alors interchangeables. Les applications de l'égalisation des scores observés sont importantes dans un grand nombre de domaines et en particulier dans le domaine de la qualité de vie lié à la santé.

L'utilisation la plus courante de la fonction d'égalisation équipercentile ou, plus précisément, de la relation $\Delta(x) = G^{-1}(F(x)) - x$ est le changement de modèle $F(x) = G(x + \Delta(x))$ pour tout x. Selon Doksum (1974), Doksum et Sievers (1976), Switzer (1976) et Nair (1982), la fonction de changement $\Delta(x)$ est la seule fonction de x telle que $X + \Delta(X)$ et Y ont la même distribution lorsque F et G sont continues. Si les deux fonctions de répartition sont représentées sur le même graphique, alors $\Delta(x)$ n'est que la distance horizontale à partir de F en G en un point x. Dans l'analyse des données de la durée de vie, la fonction $\Delta(x)$ est une mesure utile pour la comparaison de deux traitements. Par exemple, si nous considérons que X est une réponse de contrôle et Y est une réponse au traitement, alors $\Delta(x)$ peut être interprétée comme l'effet du traitement ajouté au potentiel de la réponse du contrôle x. L'ensemble $\{x : \Delta(x) > 0\}$ détermine la part de la population pour qui le traitement est bénéfique (voir e.g., Doksum et Sievers (1976) et Switzer (1976)).

D'autre part, dans le domaine des fonctions de transformation, Bagdonavicius et Nikulin (2001, 2002) mettent en place des méthodes statistiques pour répondre à ce genre de problème, en supposant que les fonctions de répartition F et G appartiennent à une famille de lois paramétriques ou semi-paramétriques pour avoir les modèles de Cox, modèles de fragilité, etc. Ainsi, l'estimation de la fonctionnelle de transformation se fait selon le principe de Sedyakin (voir Sedyakin (1966)). Par ailleurs, nous signalons les travaux de Bol′šev (1959, 1963) qui propose une approche complètement différente pour approcher $G^{-1}(F(x))$. Cette approche repose sur la notion de transformations asymptotiquement normales et les approximations par polynômes orthogonaux.

L'estimateur naturel de la fonction d'égalisation équipercentile est donné par $G_m^{-1}(F_n(x))$ où G_m^{-1} et F_n sont les fonctions quantile et de répartition empiriques des fonctions G^{-1} et F respectivement. L'estimateur empirique de la fonction d'égalisation équipercentile est connu dans la littérature statistique sous le nom du processus Quantile-Quantile. Sa première utilisation remonte à Lorenz (1905). Plus tard, plusieurs auteurs ont étudié ses propriétés asymptotiques, voir e.g., Doksum (1974), Hollander et Korwar (1982), Aly (1986), Beirlant et Deheuvels (1990), Lu *et al.* (1994), Li *et al.* (1996).

Étant donné le caractère empirique de cet estimateur, la résolution de l'équation équipercentile peut à l'occasion présenter quelques difficultés. En effet, il pourrait être impossible de trouver une solution unique qui consiste à déterminer le score équivalent y d'un score x donné satisfaisant l'équation équipercentile. Traditionnellement, ce problème est contourné en lissant les fonctions de répartition empiriques par la méthode du noyau ou des splines. Le lissage s'effectue soit avant la résolution de l'équation équipercentile, soit après. Ces deux étapes sont appelées "pré-lissage" et "post-lissage". Pour de plus amples détails et références, nous renvoyons au livre de von Davier *et al.* (2003).

Les méthodes de lissage les plus utilisées sont les techniques de noyau, les splines cubiques et les méthodes log-linéaires. L'estimation par courbe non-paramétrique a connu des avancées

importantes ces dernières décennies. En particulier, plusieurs estimateurs lissés sont apparus dans la littérature. Les plus connus étant les estimateurs à noyau (voir Cheng et Peng (2002) et Cheng et Parzen (1997) pour une description du traitement unifié). L'estimation non-paramétrique de la fonction d'égalisation équipercentile par la méthode du noyau a été introduite par Holland et Thayer (1989) dans le but de rendre continues les fonctions de répartition des scores observés. Plus tard, von Davier *et al.* (2003) ont développé la méthode en considérant cinq étapes essentielles sans étudier leurs propriétés asymptotiques.

Dans ce travail, nous généralisons l'égalisation à noyau en introduisant l'ajustement par les polynômes locaux. Cette approche est connue pour ses avantages significatifs sur la méthode de lissage à noyau entre autres : elle réduit le biais, elle n'a pas d'effets sur le bord et fournit des estimateurs simultannés des dérivées. Elle a fait l'objet de nombreux travaux repris dans le livre de Fan et Gijbels (1996). Dans le cas de la fonction d'égalisation équipercentile, cette approche présente des aspects intéressants qui permet de généraliser les méthodes à noyau usuelles en corrigeant les effets de bords. La structure proposée contient beaucoup de problèmes classiques d'estimation fonctionnelle : densité de probabilité, régression, fonction risque *etc.* (*cf.* Abdous *et al.* (2003)).

L'objet de cette thèse est d'améliorer les étapes de lissage de la fonction d'égalisation équipercentile en bénéficiant de ces avantages. Nous obtenons ainsi de nouveaux estimateurs de la fonction d'égalisation équipercentile dont nous étudions les propriétés asymptotiques.

Organisation de la thèse

Dans le premier chapitre, nous commençons par rappeler les principales méthodes d'égalisation des scores communément utilisés et donner les propriétés souhaitables de l'égalisation des scores. Dans la littérature, diverses propriétés ont été proposées par plusieurs auteurs (voir e.g. Angoff (1971), Lord (1980), Petersen *et al.* (1989)). L'égalisation des scores étant

une procédure statistique basée sur l'analyse des données, nous présentons brièvement les principaux plans d'expérience rencontrés en pratique. Une liste plus détaillée, y compris la section pré-égalisation peut être trouvée dans Petersen *et al.* (1989),(Holland et Wightman (1982) et von Davier *et al.* (2003). Dans cette thèse, nous nous restreignons au cas du plan de deux groupes équivalents. Il est attrayant en raison de sa simplicité et consiste à administrer deux questionnaires différents en même temps.

Le deuxième chapitre se situe autour du lissage de la fonction d'égalisation équipercentile en utilisant l'approche de l'ajustement polynômial local. Nous donnons d'abord un estimateur pour une fonctionnelle quelconque et ensuite nous proposons d'étudier quatre nouveaux estimateurs de la fonction d'égalisation équipercentile. Ainsi nous précisons les éléments et outils qui permettent de formuler l'équation équipercentile dans le cas que nous considérons.

Dans le troisième chapitre, nous rappelons les principaux résultats des processus empirique et quantile. Nous donnons ensuite la convergence uniforme presque sûre de la quantité $G_m^{-1}(F_n(x))$ et son approximation par un pont brownien. Pour établir l'approximation par un pont brownien, nous reprenons les grandes lignes de la preuve donnée par Beirlant et Deheuvels (1990) sous les hypothèses d'égalité des distributions $F = G$ et d'uniformité sur $(0, 1)$ $(F(x) = x)$. Ces résultats sont utiles pour l'étude de l'estimateur $y_{n,m}^{[5]}$ dans les chapitres suivants.

Dans le chapitre 4, nous établissons la convergence uniforme presque sûre des estimateurs construits dans le chapitre du lissage. Ce résultat (Théorème 4.1.1) se base essentiellement sur les convergences classiques de la fonction de répartition empirique et la fonction quantile. Une version réduite de ce chapitre a fait l'objet d'un article accepté et peut également être trouvée dans El Fassi *et al.* (2009).

Le cinquième chapitre est consacré à l'étude de l'approximation par un pont brownien approprié des processus construits à partir des estimateurs polynômiaux locaux. Nous

présentons d'abord dans le lemme 5.2.1, l'approximation par un pont brownien approprié de l'estimateur polynômial local d'une fonctionnelle quelconque. Le résultat principal sur l'approximation des estimateurs de la fonction d'égalisation équipercentile est donné dans le théorème 5.2.1.

Dans le chapitre 6, nous évaluons la performance locale des estimateurs polynômiaux locaux de la fonction d'égalisation équipercentile en étudiant leurs erreurs ponctuelles en moyenne quadratique (MSE) à l'intérieur et aux bords du support. Nous déduisons par la suite la fenêtre asymptotiquement optimale minimisant le MSE.

Pour illustrer nos résultats, nous proposons quelques simulations sous R. Nous comparons les différents estimateurs en les appliquant sur un jeu de données réelles provenant d'une étude longitudinale multi-centrique de la cohorte ANRS C08.

L'objectif est de trouver, pour chaque score donné sur l'échelle de référence $SF12$ (SF-12® (2002)) son score équivalent sur l'échelle de mesure WhoQol (WHOQOL-HIV (2004)) adaptée aux patients atteints du VIH.

Ce travail se termine par un résumé des différents résultats obtenus et par la présentation de quelques perspectives de recherches futures.

Une partie des travaux présentés dans ce manuscrit a fait l'objet d'une note publiée dans les Comptes Rendus de l'Académie des Sciences sous la référence (El Fassi *et al.* (2009)) et de plusieurs communications orales lors des séminaires et conférences (El Fassi (2005), El Fassi (2006), El Fassi *et al.* (2008b), El Fassi (2008)) dont deux internationales. Par ailleurs, deux articles en collaboration avec le professeur Belkacem Abdous sont en phase finale de rédaction. Le premier porte sur l'approximation par un pont brownien des estimateurs de la fonction d'égalisation équipercentile et le second porte de l'erreur ponctuelle en moyenne quadratique.

Notations

Nous adopterons les notations ci-dessous dans les différents chapitres de ce manuscrit.

Abréviations et Symboles

$(\Omega, \mathcal{A}, \mathcal{P})$	L'espace probabiliste où les variables aléatoires sont considérées.
$\mathcal{P}$	La mesure de probabilité attachée à l'espace probabiliste $(\Omega, \mathcal{A})$.
$\mathbb{E}(X)$	L'espérance mathématique de la variable aléatoire X.
$Var(X)$	La variance de la variable aléatoire X.
$Cov(X, Y)$	La covariance entre les variables aléatoires X et Y.
$:=$	La définition d'une quantité.
$\overset{\mathcal{D}}{=}$	L'égalité en distribution.
$\overset{\text{p.s.}}{=}$	L'égalité en presque sûre.
$\xrightarrow{\mathcal{P}}$	La convergence en probabilité.
$\xrightarrow{\text{p.s.}}$	La convergence presque sûre.
$\mathbf{1}_A$	La fonction indicatrice qui vaut 1 sur l'ensemble A et 0 ailleurs.
$M^\top$	La transposée de la matrice M.
$a \wedge b$	Le minimum de a et b.
$a \vee b$	Le maximum de a et b.
$a_n = \mathcal{O}(b_n)$, $n \to \infty$	$\limsup_{n\to\infty} \lvert a_n/b_n \rvert < \infty$ avec a_n et b_n deux suites réelles.
$a_n = o(b_n)$, $n \to \infty$	$\lim_{n\to\infty} a_n/b_n = 0$ avec a_n et b_n deux suites réelles.

$a_n \overset{\text{p.s.}}{=} \mathcal{O}(b_n)$, $n \to \infty$	$a_n/b_n \xrightarrow{\text{p.s.}} C$ où C est une constante avec a_n et b_n deux suites aléatoires.
$a_n \overset{\text{p.s.}}{=} o(b_n)$, $n \to \infty$	$a_n/b_n \xrightarrow{\text{p.s.}} 0$ avec a_n et b_n deux suites aléatoires.
$a_n = \mathcal{O}_P(b_n)$, $n \to \infty$	$a_n/b_n \xrightarrow{\mathcal{P}} C$ où C est une constante avec a_n et b_n deux suites aléatoires.
$a_n = o_P(b_n)$, $n \to \infty$	$a_n/b_n \xrightarrow{\mathcal{P}} 0$ avec a_n et b_n deux suites aléatoires.

Chapitre 1

Généralités sur l'égalisation des scores

Résumé. Dans ce chapitre, nous présentons brièvement les différents plans d'expérience et méthodes d'égalisation de scores. L'égalisation des scores est une procédure statistique basée sur l'analyse de données. Il existe plusieurs plans de collectes de données. Dans ce travail, nous nous sommes limités au plan d'expérience de deux groupes équivalents.

1.1 Définition et propriétés de la fonction d'égalisation

Nous nous restreignons au cas, où chacune de ces formes multiples du questionnaire est "unidimensionnelle", et est censée mesurer la même dimension de la qualité de vie, par exemple, la mobilité, la sociabilité ou la dimension mentale de la qualité de vie. Les deux formes du questionnaire, sont donc censées mesurer, avec des questions (items) différentes, le même concept (mobilité, sociabilité, mental), que l'on peut considérer comme trait latent inobservé directement. Le terme aptitude, utilisé dans le domaine des sciences de l'éducation, caractérise donc la position de l'individu sur ce trait latent. La série de questions (items) servant à mesurer, à l'aide des réponses, cette aptitude permet de calculer un score par individu, qui est toujours un nombre entier ou réel, estimation de cette aptitude individuelle (mesure de cette aptitude). Le questionnaire est, dans la littérature psychométrique appelé

test, qu'il ne faut pas confondre avec le test d'hypothèses statistique.

Donc, avant d'égaliser des scores d'individus sur les formes multiples du même test, il est nécessaire d'établir une équivalence efficace entre les scores des tests. C'est ce qu'on appelle le problème de l'égalisation.

Dans le domaine de la qualité de vie l'intérêt principal des praticiens est la possibilité de "traduire" les scores d'une échelle de mesure à une autre échelle de référence. Par exemple, passer d'une température exprimée en degrés Celsius à une température exprimée en degrés Fahrenheit. Souvent, le but essentiel des cliniciens est d'utiliser ces scores égalisés pour comparer des groupes de patients.

Dans la littérature, diverses propriétés de l'égalisation ont été proposées par plusieurs auteurs (voir par exemple Angoff (1971), Lord (1980), Petersen *et al.* (1989)). Quatre de ces propriétés sont considérées essentielles :

- ✓ Même aptitude : les formes alternatives de test doivent mesurer la même caractéristique ;
- ✓ Equité : l'indifférence à être évalué par l'un des deux tests/questionnaires.
- ✓ L'invariance de population, i.e., l'égalisation est choisie indépendamment de l'échantillon ou du groupe d'individus ;
- ✓ La symétrie : l'égalisation des scores est réversible, i.e., l'égalisation des scores x aux scores y est l'inverse de l'égalisation des scores y aux scores x, autrement, lien$(x \rightarrow y)$= inverse(lien$(y \rightarrow x)$).

La définition explicite de l'équité est donnée par (Lord (1980) chapitre 13). Le concept de l'équité joue un rôle essentiel dans le problème de l'égalisation des scores des tests. Il est nécessaire qu'elle soit vérifiée avant que les scores puissent être égalisés. Si la condition de l'équité est observée après l'égalisation ou transformation des scores sur les formes des tests X et Y, les deux formes de test sont strictement parallèles dans le sens de la théorie classique de test. La vérification de la condition d'équité selon la définition de Lord (1980) est à peine faisable dans la pratique pour la simple raison que les fiabilités diffèrent souvent. Les individus ayant une faible aptitude ont un avantage avec une fiabilité relativement faible, Tandis que

les individus ayant une aptitude assez élevée ont un avantage avec une mesure précise de leur aptitude, autrement dit, avec un test relativement fiable. Ainsi, il est clair qu'il est important de rendre les tests comparables en terme de fiabilité.

Une autre caractéristique importante de l'égalisation est la symétrie. Autrement dit, si un score x sur l'échelle de mesure A est égalisé à un score y sur l'échelle de mesure B, alors le score y sur l'échelle de mesure B, à l'aide de la même méthode, sera égalisé au score x sur l'échelle de mesure A. Nous pouvons nous demander ce qui est remarquable à ce sujet. Les relations statistiques importantes ne sont pas toutes symétriques. En particulier, la prévision en statistique n'est pas symétrique.

Plusieurs méthodes peuvent être utilisées pour égaliser les scores. Elles sont classées par catégorie (voir par exemple Angoff (1987))

- ✓ Égalisation équipercentile ;
- ✓ Égalisation linéaire ;
- ✓ Égalisation par les méthodes de régression ;
- ✓ Égalisation par les méthodes des réponses aux items (IRT).

1.2 Plans d'expérience pour l'égalisation des scores

En pratique, la manière dont les données sont collectées, joue un rôle important dans l'égalisation. Les plans d'expérience de la collecte de données pour l'égalisation sont détaillés dans von Davier *et al.* (2003) et Kolen et Brennan (2004). Dans ce travail, nous nous limiterons au plan d'expérience de deux groupes équivalents.

1.2.1 Plan d'expérience d'un seul groupe

Ce plan d'expérience consiste à administrer deux tests ou traitements à un seul groupe ou échantillon d'individus. Un des avantages de ce plan d'expérience est que chaque individu répond aux items des deux tests. Toutefois, ce plan est rarement utilisé car l'administration

des tests nécessite beaucoup de temps, outre l'effet de la fatigue qui est toujours présent et pourrait jouer un rôle important quand les individus répondent aux items du deuxième test.

Par conséquent, une idée pertinente serait d'administrer les tests dans un ordre différent. L'échantillon d'individus est partagé en deux sous-échantillons et l'on administre les tests aux deux sous-échantillons dans un ordre équilibré. Techniquement, nous retombons sur le plan d'expérience équilibré des groupes aléatoires.

1.2.2 Plan d'expérience des groupes équivalents

Ce plan est aussi appelé plan des groupes aléatoires. Il est attrayant en raison de sa simplicité. Il exige que tous les tests soient disponibles et administrés en même temps. Ce qui réduit relativement le temps d'administration des tests.

Les deux groupes/échantillons fournissent des données qui peuvent être utilisées pour estimer directement les probabilités des scores. D'ailleurs, il n'y a aucune hypothèse additionnelle à considérer outre les hypothèses classiques,

(1) Les tests sont aléatoirement administrés aux deux groupes d'individus.

(2) Les deux groupes/échantillons sont indépendamment et aléatoirement tirés de la même population.

Si les conditions de l'administration de test sont les mêmes pour les deux tests, l'échantillonage en spirales peut être utilisé pour créer les groupes/échantillons (voir Kolen et Brennan (2004)). Le plan des groupes équivalents exige toujours les grandes tailles de l'échantillon afin de minimiser les différences aléatoires entre les deux groupes.

1.2.3 Plan d'expérience d'équilibrage

Dans le plan d'équilibrage, les tests sont administrés aux deux groupes/échantillons d'individus aléatoires et indépendants, dans différents ordres. Le premier groupe prend d'abord le premier test et ensuite le second, comme dans le plan d'expérience d'un seul groupe. L'autre groupe prend le second test d'abord et ensuite le premier. L'intérêt d'équilibrer l'ordre de

test est de s'assurer que tous les effets d'ordre sont pris en compte de façon équitable dans les scores obtenus pour les deux tests. Si quelques effets d'ordre telle que la fatigue sont efficacement contrôlés par l'équilibrage, alors l'avantage primaire en utilisant le plan d'un seul groupe avec équilibrage est qu'il a typiquement de plus petites conditions sur la taille des échantillons, contrainte principale dans le plan de groupes équivalents.

Dans la pratique, le plan d'un seul groupe avec équilibrage pourrait être utilisé au lieu du plan des groupes équivalents quand

(1) l'administration des deux tests est opérationnellement possible,

(2) on ne s'attend pas à ce que des effets d'ordre se produisent,

(3) il est difficile d'avoir un nombre suffisant d'individus dans une étude d'égalisation utilisant le plan des groupes équivalents.

1.2.4 Plan d'expérience de groupes non-équivalents avec des items communs

Dans ce plan, deux échantillons d'individus sont tirés de deux populations différentes. On administre le premier test au groupe tiré de la première population, et le second test au groupe tiré de la deuxième population et les deux groupes prennent un ensemble commun d'items. Ce plan est utilisé quand on ne peut pas administrer plus d'un test en une date donnée pour des raisons de sécurité ou autres soucis pratiques.

Bien que diverses méthodes d'égalisation aient été proposées pour ce plan d'expérience, aucune méthode ne fournit des ajustements complètement appropriés quand les groupes d'individus sont très différents.

1.3 Méthodes d'égalisation

L'égalisation est utilisé pour transformer des scores de tests sur deux échelles de mesure différentes en se basant sur diverses procédures statistiques. Il existe plusieurs techniques

et méthodologies qui peuvent être utilisées pour l'égalisation. D'une manière générale, il y a deux catégories de méthodes : les méthodes classiques de la théorie d'égalisation et les méthodes de la théorie de réponse aux items (IRT).

Dans cette section, nous présentons brièvement les méthodes classiques d'égalisation les plus utilisés dans la littérature : égalisation des moyennes, égalisation linéaire, égalisation équipercentile et la méthode de l'égalisation à noyau.

1.3.1 Égalisation des moyennes

Nous définissons deux variables aléatoires X et Y représentant deux scores de tests. Soient x et y les réalisations de X et Y respectivement. En outre, nous désignons respectivement par $\mu(X)$ et $\mu(Y)$ les moyennes des variables X et Y. Dans l'égalisation des moyennes, les différences entre les scores x et y et leurs moyennes $\mu(X)$ et $\mu(Y)$ sont égales, autrement dit,

$$x - \mu(X) = y - \mu(Y). \tag{1.1}$$

En résolvant pour y, nous obtenons $m_Y(x)$ la fonction d'égalisation des moyennes,

$$m_Y(x) = y = x - \mu(X) + \mu(Y). \tag{1.2}$$

Dans cette équation, la fonction $m_Y(x)$ fournit les scores x transformés à des scores y en utilisant l'égalisation des moyennes. Ainsi, l'égalisation des moyennes consiste à additionner une constante éventuellement négative (différence entre les deux moyennes) à tous les scores x pour trouver les scores égalisés y.

1.3.2 Égalisation linéaire

Dans l'égalisation linéaire, les scores qui ont une distance égale de leurs moyennes dans des unités d'écart type sont égales. Nous définissons respectivement les écarts type des variables X et Y par $\sigma(X)$ et $\sigma(Y)$. La conversion linéaire est définie en égalisant les scores normalisés

tels que

$$\frac{x-\mu(X)}{\sigma(X)} = \frac{y-\mu(Y)}{\sigma(Y)}. \tag{1.3}$$

Remarquons que si les écarts-type sont égaux, l'équation (1.3) pourrait être simplifiée pour retomber sur l'équation de l'égalisation des moyennes (1.2). Ainsi, les deux méthodes d'égalisation des moyennes et l'égalisation linéaire donneront le même résultat. En résolvant pour y, l'équation (1.3) devient

$$Lin_Y(x) = y(x) = \sigma(Y)\left[\frac{x-\mu(X)}{\sigma(X)}\right] + \mu(Y), \tag{1.4}$$

où $Lin_Y(x)$ est l'équation de la transformation linéaire des scores x en scores y. Une expression alternative est

$$Lin_Y(x) = y(x) = \frac{\sigma(Y)}{\sigma(X)}x + \left[\mu(Y) - \frac{\sigma(Y)}{\sigma(X)}\mu(X)\right] \tag{1.5}$$

qui a la forme d'une fonction de la régression simple de Y en X. Cette dernière n'est pas considérée comme une méthode d'égalisation des scores car elle ne vérifie pas la propriété de la symétrie. Dans la section suivante, nous expliquons la différence entre l'égalisation linéaire des scores et la régression simple de Y en X.

1.3.3 Propriétés de l'égalisation des moyennes et de l'égalisation linéaire

En utilisant l'équation de l'égalisation des moyennes (1.2), l'espérance et l'écart type du score converti $m_Y(X)$ sont donnés par

$$\mathbf{E}\left(m_Y(X)\right) = \mathbf{E}\big(X - \mu(X) + \mu(Y)\big) = \mu(Y) \tag{1.6}$$

et

$$\sigma\left(m_Y(X)\right) = \sigma(X). \tag{1.7}$$

D'autre part, en utilisant l'équation de l'égalisation linéaire (1.5), l'espérance et l'écart type du score transformé $Lin_Y(X)$ sont donnés par

$$\mathbf{E}\left(Lin_Y(X)\right) = \mathbf{E}\left[\frac{\sigma(Y)}{\sigma(X)}x + \mu(Y) - \frac{\sigma(Y)}{\sigma(X)}\mu(X)\right] = \mu(Y) \tag{1.8}$$

et

$$\sigma\left(Lin_Y(X)\right) = \sigma\left(\frac{\sigma(Y)}{\sigma(X)}X\right) = \sigma(Y). \tag{1.9}$$

Par conséquent, l'espérance et l'écart type de X transformés sur l'échelle de Y sont respectivement égaux à l'espérance et à l'écart type de Y. Si en effet, l'égalisation des scores x en scores y est réalisé, nous pouvons déduire l'égalisation inverse qui consiste à transformer les scores y en scores x. Ces transformations sont notées respectivement par $m_X(y)$ et $Lin_X(y)$ et sont symétriques par définition.

L'équation de l'égalisation linéaire (1.5) a la forme d'une une équation de régression linéaire. La différence est que, pour la régression linéaire, le terme $\sigma(Y)/\sigma(X)$ est multiplié par la corrélation entre X et Y. Cependant, nous ne pouvons qualifier une équation de régression linéaire comme une fonction d'égalisation linéaire car la régression de X en Y est différente de la régression de Y en X, à moins que le coefficient de corrélation soit égal à 1. C'est d'ailleurs pour cette raison que les équations de régression ne peuvent pas, en général, être utilisées comme des fonctions d'égalisation. En effet, la régression de Y en X est différente de la régression de X en Y. Tandis que l'égalisation linéaire peut être utilisée pour transformer des scores x en scores y, ou le cas inverse, pour transformer des scores y en scores x.

En résumé, dans l'égalisation des moyennes, la transformation des scores est obtenue en égalisant les différences entre les scores et leurs moyennes, alors que dans l'égalisation linéaire, nous égalisons les scores x et y normalisés. Dans l'égalisation des moyennes, les scores x sont ajustés par une quantité constante qui est égale à la différence entre les moyennes de X et de Y. La moyenne de X transformée sur l'échelle de Y est égale à la moyenne de Y ; alors que dans l'égalisation linéaire, la moyenne et l'écart type de X transformés sur l'échelle de Y sont respectivement égaux à la moyenne et l'écart type de Y. En général, l'égalisation des moyennes est moins compliquée que l'égalisation linéaire, mais cette dernière fournit des résultats plus significatifs par rapport à l'égalisation des moyennes.

1.3.4 Égalisation équipercentile

Nous nous tournons maintenant vers la plus importante des méthodes de l'égalisation des scores observés, l'égalisation équipercentile. La définition de l'égalisation équipercentile a été développée par Braun et Holland (1982).

À cet effet, nous fixons d'abord quelques notations. Soient X et Y deux variables aléatoires représentant deux scores à deux tests ou deux versions d'un même test ou encore les résultats de deux instruments de mesure de la qualité de vie. Supposons que ces scores appartiennent aux intervalles $[x_1, x_N]$ et $[y_1, y_M]$ respectivement. Les deux entiers N et M sont fixés et connus. Nous désignerons par $F(\cdot)$ et $G(\cdot)$ les fonctions de répartition associées à X et Y respectivement définies par

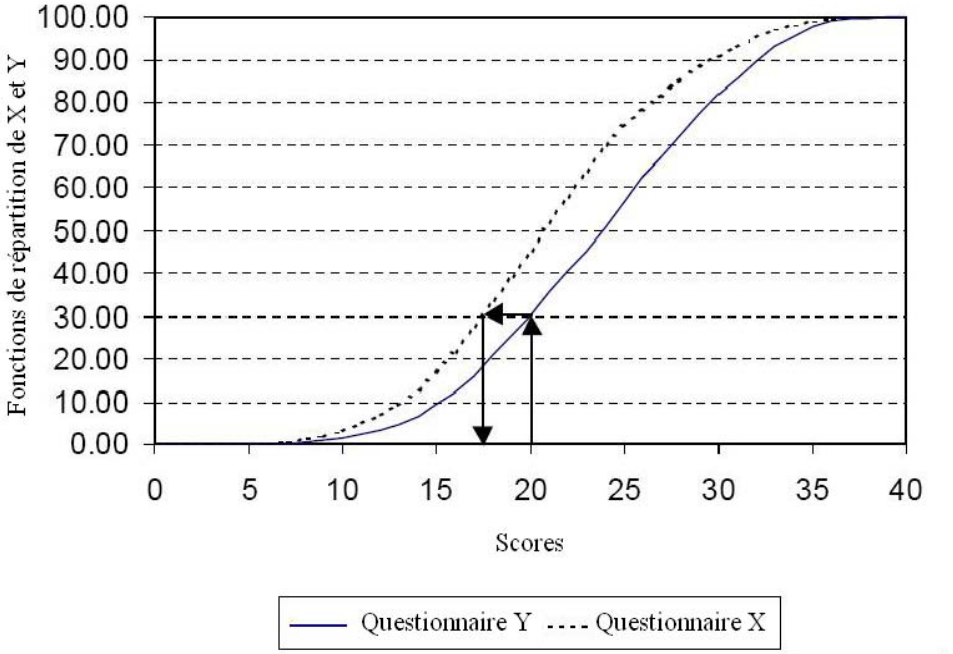
$$F(x) = P(X \leq x), \quad \text{et} \quad G(y) = P(Y \leq y).$$

Nous dirons que deux scores x et y sont équivalents si et seulement si

$$F(x) = G(y)^1 \tag{1.10}$$

[1]Dans un contexte d'application différent (fiabilité), on retrouve ce qui est communément appelé fonction de transformation (Bagdonavicius et Nikulin (2007)). "Denote by $\lambda_0(t) = \lambda_{x_0}$ a baseline rate function. It corresponds to the failure rate of a given population of items when hypothetically all items are in the same ideal, normal, usual conditions, given by a stress $x_0(\cdot) \in E$, where E is a set of all possible or admissible stresses. We shall write also S_0 instead S_{x_0} . We denote by $f_{x(\cdot)}(t)$ the time under $x_0(\cdot)$ equivalent to t under the stress $x(\cdot)$ in the sense that the probability that an item used under the stress $x(\cdot)$ would survive till the moment t is equal to the probability that an item used under the normal stress $x_0(\cdot)$ would survive till the moment $f_{x(\cdot)}(t)$: $S_{x(\cdot)}(t) = \mathbf{P}\{T_{x(\cdot)} \geq t\} = \mathbf{P}\{T_{x_0(\cdot)} \geq f_{x(\cdot)}(t)\} = S_{x_0(\cdot)}\left(f_{x(\cdot)}(t)\right)$, where $T_{x(\cdot)}$ denote the failure time under a stress $x(\cdot) \in E$. It means that for any t, $t > 0$, $f_{x(\cdot)}(t) = S^{-1}_{x_0(\cdot)}\left[S_{x(\cdot)}(t)\right], x(\cdot) \in E$. The functional $f_{x(\cdot)}(t)$: $E \times [0, \infty) \to [0, \infty)$ is called the transfer functional. For any $x(\cdot) \in E$ denote $\psi_{x(\cdot)}$ the inverse function of $f_{x(\cdot)}$. The last equality implies that $T_{x(\cdot)} \cong \psi_{x(\cdot)}(T_{x_0})$, i.e. the distributions of $T_{x(\cdot)}$ and $\psi_{x(\cdot)}(T_{x_0})$ coincide. The function $\psi_{x(\cdot)} = \psi_{x(\cdot)}(t)$ is called time transformation function, (TTF). "

La méthode de l'égalisation équipercentile est représentée sur la figure ci-dessous,



Par ailleurs, si les variables X et Y sont discrètes, alors pour un score x, il peut être impossible de trouver un score y satisfaisant *l'équation équipercentile* (1.10). Pour éviter ce problème, une solution commune consiste à approcher F et G par des fonctions de répartition continues par interpolation linéaire par exemple et résoudre par la suite l'équation équipercentile correspondante (voir par exemple Holland et Thayer (1989)). De ce fait, si nous définissons correctement la fonction inverse G^{-1} de G, la fonction d'égalisation équipercentile de X en Y peut être définie par

$$y(x) = Equi_Y(x) = G^{-1}(F(x)) \tag{1.11}$$

où la fonction inverse généralisée G^{-1} est définie par

$$G^{-1}(p) := \inf\{u : G(u) \geq p\} \quad \text{pour} \quad 0 < p < 1, \tag{1.12}$$
$$G^{-1}(0) = G^{-1}(0^+) = \lim_{p \downarrow 0} G^{-1}(p) \text{ et } G^{-1}(1) = G^{-1}(1^-) = \lim_{p \uparrow 1} G^{-1}(p).$$

Les propriétés de la fonction G et de son inverse G^{-1} sont détaillées dans Serfling (1980), par exemple

(i) $G^{-1}(G(z)) \leq z$, pour $-\infty < z < \infty$, avec égalité si G est continue et strictement croissante.

(ii) $G(G^{-1}(p)) \geq p$, pour $0 < p < 1$, avec égalité si G est continue.

(iii) $G(z) \geq p$, si et seulement si $z \geq G^{-1}(p)$.

Donc, pour chaque x fixe, la valeur $Equi_Y(x)$ est la plus petite valeur de y telle que

$$P(Y < y) \leq F(x) \leq P(Y \leq y).$$

Si G est strictement croissante, donc $Equi_Y(x)$ est unique.

Si, par exemple, on peut rendre les fonctions de répartition continues donc la fonction égalisation devient

$$y(x) = Equi_Y(x) = G^{-1}(F(x)) \tag{1.13}$$

où x et y sont des scores continus.
Cette transformation est en fait une transformation des scores x en scores y. Son inverse consiste à transformer les scores y en scores x. Il est bien connu que l'égalisation équipercentile vérifie les propriétés définies précédemment.

Lorsque les scores sont des variables discrètes, la définition de l'égalisation équipercentile est plus compliquée.Holland et Thayer (1989) ont présenté une justification statistique pour l'utilisation de la fonction quantile dans l'égalisation des scores. Dans leur approche, ils utilisent une étape appelée "Processus de continuisation". Si l'on suppose que X est une variable aléatoire discrète et U une variable aléatoire uniformément distribuée sur $\left[-\frac{1}{2}; \frac{1}{2}\right]$, et si l'on définit une nouvelle variable aléatoire $X^* = X + U$, on retombe sur une nouvelle variable aléatoire continue. La fonction de répartition de cette variable correspond à la fonction de rang de percentile. L'inverse de cette fonction de répartition existe et est la fonction quantile. Holland et Thayer (1989) ont généralisé leur approche pour incorporer les processus de continuisation qui sont basés sur des distributions autres que l'uniforme. Cette approche a été développée par von Davier *et al.* (2003).

1.3.5 Égalisation à noyau

L'égalisation à noyau est une approche basée sur l'estimation non-paramétrique des fonctions de répartition (Tapia et Thompson (1978) ; Silverman (1986)). Elle a été introduite par Holland et Thayer (1989) dans le but de rendre continues les fonctions de répartition des scores observés. von Davier *et al.* (2003) ont présenté un résumé de la procédure de l'égalisation à noyau sous forme de cinq étapes :

Étape 1 : Pré-lissage.

Cette étape consiste fondamentalement à estimer les probabilités des réalisations x_i et y_j des scores X et Y respectivement en utilisant les modèle log-linéaires polynômiaux qui permettent d'analyser des tables de contingence multidimensionnelles construites à partir de plusieurs variables catégorielles et d'étudier précisément les liaisons entre ces variables (Morineau *et al.* (1996), Holland et Thayer (2000)). Ces modèles sont utilisés afin d'estimer des probabilités en adaptant des fonctions polynomiales au logarithme de la densité de l'échantillon. Par exemple, pour les probabilités de X, les fonctions polynomiales sont de la forme suivante (Kolen et Brennan (2004)),

$$\log(nf(x)) = \beta_0 + \beta_1 x + \ldots + \beta_p x^p,$$

où n étant la taille de l'échantillon de X, $f(x)$ est la fonction densité de X, β_i avec $i = 0, 1, \ldots, p$ sont les coefficients du modèle et p est le degré du polynôme. Le même modèle polynomial peut être utilisé pour estimer les probabilités de Y (pas nécessairement du même degré). Les paramètres β_i peuvent être estimés par la méthode du maximum de vraisemblance. Holland et Thayer (1987) ont présenté une description détaillée de ces modèles et ont décrit des algorithmes pour la méthode de l'estimation par le maximum de vraisemblance. Le degré p du polynôme contrôle le degré du lissage.

Étape 2 : Estimation des probabilités des scores.

Dans cette étape, les probabilités de scores estimées dans la première étape sont transformées pour obtenir les probabilités des scores sur une population commune. La trans-

formation des distributions estimées est faite en utilisant ce que l'on appelle "fonction du plan d'expérience". Différentes fonctions du plan d'expérience sont utilisées pour les divers plans de collecte de données. Pour le plan des groupes équivalents, les probabilités estimées sont exactement identiques que les probabilités lissées de l'étape 1. En d'autres termes, dans le plan des groupes équivalents, les probabilités des scores dans la population sont estimées en pré-lissant les probabilités des scores dans l'échantillon.

Étape 3 : Continuisation.

Cette étape traite l'égalisation équipercentile des scores discrets. Traditionnellement, ce problème est résolu en utilisant l'interpolation linéaire. von Davier *et al.* (2003) ont proposé d'utiliser la méthode à noyau gaussien pour adapter une fonction de répartition continue à la fonction de répartition discrète des scores. Pour motiver donc la continuisation des fonctions de répartition F et G, on considère une variable aléatoire discrète X de moyenne μ_X et de variance σ_X^2, une variable aléatoire U indépendante de X et suivant une loi normale $\mathcal{N}(0,1)$. On constate qu'il est possible d'approcher la variable X par $X+h_XU$ où h_X est une constante positive assez petite. Cependant, puisque U est de loi continue, il est intuitivement évident que $X + h_XU$ soit continûment distribuée. La nouvelle variable aléatoire $X + h_XU$ a la même moyenne que X. En revanche, les deux variables n'ont pas la même variance pour $h_X > 0$,

$$Var(X + h_XU) = \sigma_X^2 + h_X^2 > 0.$$

Il est facile de transformer linéairement $X + h_XU$ de telle sorte que les deux variables aient les mêmes moyennes et variances pour tout h_X. Cette transformation est donnée par (von Davier *et al.* (2003)),

$$X(h_X) = a_X(X + h_XU) + (1 - a_X)\mu_X, \tag{1.14}$$

où

$$a_X = \sqrt{\frac{\sigma_X^2}{\sigma_X^2 + h_X^2}}. \tag{1.15}$$

De façon analogue, on définit la variable $Y(h_Y)$,

$$Y(h_Y) = a_Y(Y + h_Y V) + (1 - a_Y)\mu_Y, \tag{1.16}$$

où

$$a_Y = \sqrt{\frac{\sigma_Y^2}{\sigma_Y^2 + h_Y^2}}. \tag{1.17}$$

et V est indépendante de Y suivant une loi normale $\mathcal{N}(0,1)$. Le théorème ci-dessous résume le comportement de $X(h_X)$ quand $h_X \to 0$ et $h_X \to \infty$ (voir von Davier *et al.* (2003)).

Théorème 1.3.1. *Sous les notations ci-dessus, on a,*

(a) $\lim_{h_X \to 0} a_X = 1$,

(b) $\lim_{h_X \to \infty} a_X = 0$,

(c) $\lim_{h_X \to \infty} h_X a_X = \sigma_X$,

(d) $\lim_{h_X \to 0} X(h_X) = X$,

(e) $\lim_{h_X \to \infty} X(h_X) = \sigma_X U + \mu_X$.

Quand $h_X > 0$, $X(h_X)$ a une fonction de répartition continue définie par

$$\widehat{F}_{h_X}(x) = \mathbf{P}(X(h_X) \leq x). \tag{1.18}$$

Le théorème suivant prouve que la fonction de répartition de $X(h_X)$ est une version continuisée de F .

Théorème 1.3.2. *Si $X(h_X)$ est définie par (1.14) et $\widehat{F}_{h_X}(x)$ est la fonction de répartition définie par (1.18), alors,*

$$\widehat{F}_{h_X}(x) = \sum_{i=1}^{n} p_i \Phi(R_{iX}(x)), \tag{1.19}$$

où p_i est obtenue dans la première étape du pré-lissage, $\Phi(\cdot)$ est la fonction de répartition de la loi gaussienne et

$$R_{iX}(x) = \frac{x - a_X x_i - (1 - a_X)\mu_X}{a_X h_X}. \tag{1.20}$$

Preuve du théorème 1.3.2.

$$\begin{aligned}
\widehat{F}_{h_X}(x) &= \mathbf{P}\left(X(h_X) \leq x\right) \\
&= \mathbf{P}\left(a_X(X + h_X U) + (1 - a_X)\mu_X \leq x\right) \\
&= \mathbf{P}\left(a_X h_X U \leq x - a_X X - (1 - a_X)\mu_X\right) \\
&= \sum_{i=1}^{n} p_i \mathbf{P}\left(a_X h_X U \leq x - a_X x_i - (1 - a_X)\mu_X\right) \\
&= \sum_{i=1}^{n} p_i \mathbf{P}\left(U \leq \frac{x - a_X x_i - (1 - a_X)\mu_X}{a_X h_X}\right) \\
&= \sum_{i=1}^{n} p_i \Phi\left(\frac{x - a_X x_i - (1 - a_X)\mu_X}{a_X h_X}\right)
\end{aligned}$$

Ce qui achève la démonstration. □

Il existe diverses méthodes pour le choix du paramètre de la fenêtre h_X (voir Silverman (1986) et Wand et Jones (1995)). Si h_X est grand (i.e. $h_X > 10\sigma_X$), $\widehat{F}_{h_X}(x)$ est une approximation "normale" de F. Dans le cas où h_X et h_Y sont tous les deux grands, le résultat de la fonction d'égalisation équipercentile est celui de la fonction d'égalisation linéaire. Ce cas est détaillé dans l'étape d'égalisation.

Étape 4 : Égalisation.

Après avoir obtenu $\widehat{F}_{h_X}(x)$ et $\widehat{G}_{h_Y}(y)$, les estimateurs des deux fonctions de répartition respectivement de Xet Y, le processus de l'égalisation est relativement direct. Il suffit de calculer les fonctions inverses de $\widehat{F}_{h_X}(x)$ et $\widehat{G}_{h_Y}(y)$ notées respectivement $\widehat{F}^{-1}_{h_X}(x)$ et $\widehat{G}^{-1}_{h_Y}(y)$. En utilisant la formule (1.13) qui décrit la fonction générale de l'égalisation équipercentile, on obtient,

$$\widehat{y}(x) = \widehat{G}^{-1}_{h_Y}(\widehat{F}_{h_X}(x)).$$

De façon analogue, la fonction de l'égalisation à noyau pour transformer Y en X est donnée par

$$\widehat{x}(y) = \widehat{F}^{-1}_{h_X}(\widehat{G}_{h_Y}(y)).$$

Le théorème suivant démontre que, quand h_X et h_Y sont grandes, la fonction d'égalisation

équipercentile à noyau peut être approchée par la fonction d'égalisation linéaire.

Théorème 1.3.3. *Si $y(x)$ est défini par (1.13), alors*

$$\lim_{h_X,h_Y\to\infty} y(x) = \mu_Y + \frac{\sigma_Y}{\sigma_X}(x-\mu_X) = Lin_Y(x).$$

Preuve. D'après le théorème 1.3.1, quand h_X et h_Y tendent vers l'infini, on a

$$\lim_{h_X\to\infty} \widehat{F}_{h_X}(x) = \Phi\left(\frac{x-\mu_X}{\sigma_X}\right),$$
$$\lim_{h_Y\to\infty} \widehat{G}_{h_Y}(y) = \Phi\left(\frac{y-\mu_Y}{\sigma_Y}\right).$$

Par conséquent

$$\lim_{h_X\to\infty} \widehat{G}^{-1}_{h_Y}(u) = \mu_Y + \sigma_Y\Phi^{-1}(u),$$

où Φ^{-1} est l'inverse de Φ fonction de répartition de la loi normale standard. Donc,

$$\begin{aligned}\lim_{\substack{h_X\to\infty\\ h_Y\to\infty}} \widehat{F}_{h_X}(x) &= \Phi\left(\frac{x-\mu_X}{\sigma_X}\right),\\ &= \Phi\left(\frac{y-\mu_Y}{\sigma_Y}\right).\end{aligned}$$

□

Étape 5 : Calcul de l'erreur standard de l'égalisation.

Dans la procédure de l'égalisation, il y a deux sources d'erreur principales ; erreur aléatoire et erreur systématique. La première résulte du fait qu'on utilise des échantillons aléatoirement tirés d'une population donnée, pour estimer des paramètres tels que la moyenne, la variance, fonction quantile (Kolen (1988)). Si la population entière était utilisée dans l'étude de l'égalisation, alors aucune erreur aléatoire ne serait présente. L'erreur aléatoire d'égalisation peut être réduite en utilisant des échantillons de tailles assez grandes et en choisissant un plan d'expérience d'égalisation approprié (Kolen (1988), Kolen et Brennan (2004)). Ils ont énuméré quatre types d'erreurs systématiques produites lors de l'égalisation des scores. D'abord, les erreurs systématiques se produisent quand des facteurs ou des hypothèses statistiques ne sont pas respectées, par exemple,

quand l'égalisation linéaire est utilisée pour estimer une relation d'égalisation entre des scores qui n'est pas linéaire à la base. Deuxièmement, les erreurs systématiques se produisent quand la méthode d'estimation de la fonction d'égalisation présente un biais. Le troisième type d'erreur systématique peut être produit suite à une mauvaise administration de questionnaires aux échantillons d'individus. Finalement, les erreurs systématiques peuvent être dues à l'administration des questionnaires à des échantillons non représentatifs de la population.

Des erreurs d'égalisation aléatoires sont mesurées par "l'erreur standard de l'égalisation". Cette dernière est égale à l'écart-type de l'estimateur de la fonction d'égalisation (Kolen et Brennan (2004)). Dans le plan des groupes équivalents, deux échantillons d'individus sont aléatoirement tirés de la même population et prennent chacun un questionnaire différent. Alors, l'égalisation est obtenue en utilisant les données des deux échantillons (Kolen et Brennan (2004)).

Pour une estimation de la fonction d'égalisation donnée, l'erreur aléatoire de l'égalisation pour un score donné x_i est donnée par,

$$REE = \hat{y}(x_i) - E(\hat{y}(x_i)),$$

où $\hat{y}(\cdot)$ est un estimateur de la fonction d'égalisation de x_i. L'erreur standard de l'égalisation est donnée par

$$SEE(\hat{y}(x_i)) = \sqrt{Var\,(\hat{y}(x_i))}.$$

von Davier *et al.* (2003) ont proposé une formule générale pour le calcul de l'erreur standard de l'égalisation (SEE) basée sur la δ-méthode. Cette formule générale peut être appliquée à tout type de données.

Le biais de la fonction d'égalisation mesure des erreurs systématiques de l'égalisation. Il est défini comme la différence entre la vraie valeur de la fonction d'égalisation $y(x_i)$ et

l'espérance de son estimateur $\hat{y}(x_i)$,

$$Biais(\hat{y}(x_i)) = E(\hat{y}(x_i)) - y(x_i).$$

En termes de quantités carrées, on a

$$MSE(\hat{y}(x_i)) = Var(\hat{y}(x_i)) + [Biais(\hat{y}(x_i))]^2 .$$

Les résultats de von Davier *et al.* (2003) sont présentés dans un contexte appliqué, sans étude des propriétés asymptotiques. Dans les chapitres suivants, nous nous sommes intéressés aux divers modes de convergence de plusieurs estimateurs obtenus par la technique de lissage de l'ajustement polynomial local. Ces estimateurs sont plus généraux que l'estimateur à noyau. Dans le chapitre suivant, nous présentons le lissage de la fonction d'égalisation équipercentile. La convergence uniforme presque sûre est étudiée dans le chapitre 4 et l'approximation par un pont brownien dans le chapitre 5. Le chapitre 6 est consacré à l'étude du risque quadratique.

Chapitre 2

Lissage de la fonction d'égalisation équipercentile par des polynômes locaux

Résumé. Dans ce chapitre, nous donnons, en premier abord, la procédure de l'estimation d'une fonctionnelle par l'approche de l'ajustement par des polynômes locaux. Nous construisons, par la suite, divers scénarios d'estimation de la fonction d'égalisation équipercentile. Cette approche revient à résoudre un problème des moindres carrés pondérés.

2.1 Lissage des fonctions d'égalisation équipercentile

L'estimation basée sur les fréquences pour des distributions de probabilité discrètes est un procédé basique en statistiques et est assez bien compris. Par exemple, il est bien connu que l'estimateur de fréquence $\widehat{p}_j$ d'une probabilité p_j est le meilleur estimateur asymptotiquement normal. Toutefois ces résultats, usuellement consistants, tombent en défaut lorsque l'échantillon utilisé a un nombre épars d'observations sur un nombre fini de catégories. Typiquement, ceci apparaît quand un nombre substantiel de catégories n'est pas observé. Une

approche possible pour améliorer l'estimation dans une telle catégorie est d'emprunter l'information à une catégorie proche. Ceci se fait en utilisant des techniques de lissage (voir Burman (1987), Hall et Titterington (1987), Dong et Simonoff (1995), Simonoff (1995) et Aerts *et al.* (1997). Il est bien établi que les techniques de lissage tendent à introduire une réduction du biais et de la variance. Elles réalisent une amélioration des estimateurs basés sur les fréquences pourvu que les tailles des échantillons soient suffisamment grandes avec une légère hypothèse de régularité sur la distribution des échantillons.

En effet, étant donné le comportement empirique de l'équation équipercentile, la valeur $G_m^{-1}(F_n(x))$ peut ne pas être unique en raison de la discontinuité. Pour pallier ce problème, différents procédés de lissage ont été adaptés au problème de l'égalisation équipercentile. Il y a essentiellement deux catégories de tels procédés de lissage : le prélissage et le postlissage. Le prélissage consiste à lisser les distributions de fréquences avant de les utiliser pour déterminer la fonction équipercentile. Le but de cette étape est essentiellement de supprimer les irrégularités du score tout en préservant leurs formes et de contourner le problème de la non existence d'une solution à l'équation équipercentile. L'étape de postlissage consiste à lisser la fonction de l'égalisation équipercentile obtenue à partir des distributions de score observées (non lissées). Une fois encore, cette étape permet de régulariser la courbe associée à l'égalisation équipercentile en préservant sa forme et sa position. Pour une comparaison de ces deux procédés voir Butler et Hanson (1997). On se réfère aussi à von Davier *et al.* (2003) et Kolen et Brennan (2004) pour une structure unifiée de l'égalisation pouvant être appliquée à diverses données et méthodes d'égalisation.

Les méthodes de lissage les plus utilisées sont les techniques de noyaux, les splines et les méthodes log-linéaires polynomiales. L'estimation non paramétrique a connu des avancées importantes ces dernières dizaines d'années. En particulier, plusieurs estimateurs lissés de la fonction quantile sont apparus dans la littŕature. Les plus connus étant les estimateurs de la fonction quantile à noyau (voir Falk (1983), Zelterman (1990), Cheng et Parzen (1997) et Abdous *et al.* (2003) pour une description détaillée). Notre souhait dans ce travail est

d'améliorer les étapes de lissage dans l'égalisation en bénéficiant de ces avancées. De façon plus précise, nous allons adopter l'approche de l'ajustement polynomial local. Cette approche est connue pour ses avantages significatifs sur la méthode de lissage basée sur le noyau entre autres, elle réduit le biais, elle n'a pas d'effets sur le bord et fournit des estimateurs simultannés des dérivées.

Le développement de l'estimation polynomiale locale a fait l'objet de nombreux travaux tant au point de vue théorique qu'appliqué. Cette approche a été introduite par Stone (1977, 1980, 1982) et Cleveland (1979). Elle a, jusqu'à présent, surtout été utilisée pour l'estimation de l'espérance conditionnelle et a mené aux techniques dites de "régression polynomiale locale". Ces méthodes sont des versions sophistiquées de la régression polynomiale classique, où l'on applique localement cette dernière. Plus précisément, pour un point x, la fonction de régression est estimée par un polynôme de degré fixé calculé à partir des données "proches" de x.

Historiquement, les premières applications de l'estimation polynomiale locale remonte à Stone (1977). Lejeune (1985) a étudié le comportement de l'erreur moyenne quadratique. Il a montré une équivalence entre la régression polynomiale locale et la régression par noyaux optimaux, qui toutes les deux éliminent le biais jusqu'à un ordre donné et sont de variance minimale. Fan (1992, 1993) et Fan et Gijbels (1996) ont montré que la régression polynomiale locale présente plusieurs avantages sur les autres méthodes de régression non-paramétriques : ils ont remarqué sa capacité à contourner le principal écueil des méthodes à noyau classiques en corrigeant automatiquement les effets de bords tout en conservant les propriétés d'optimalité théoriques. Ruppert et Wand (1994) ont montré que son extension au cas multivarié est aisée.

En résumé, les vingt dernières années ont été marquées par un développement remarquable de l'estimation polynomiale locale, tant au point de vue théorique qu'appliqué .

L'estimation polynomiale locale de la fonction de régression a connu un grand succès. Ces techniques ont été appliquées pour l'estimation de la fonction de répartition et fonction de

densité. Cheng *et al.* (1997) ont obtenu des estimateurs locaux linéaires de la densité. Lejeune et Sarda (1992) ont minimisé une norme L^2 localement pondérée par une fonction poids, entre l'estimateur de la fonction de répartition F_n et un polynôme afin d'obtenir l'estimateur de la fonction de répartition, et ensuite ont utilisé ses dérivées comme estimateur de la densité. Cette approche a également été utilisée par Cheng et Peng (2002)dans le but d'estimer la fonction de répartition ainsi que la fonction quantile. Ils retrouvent les mêmes propriétés avantageuses que présente la régression polynomiale locale et obtiennent l'erreur en moyenne quadratique de ces estimateurs.

Ayant fait ses preuves sur le plan théorique, l'estimation polynômiale locale est maintenant utilisée dans de nombreux domaines où l'on applique la statistique comme en témoignent, entre autres, les travaux de Fan et Gijbels (1996), Fan *et al.* (1998) et Yu et Jones (1998).

Dans le cadre de cette thèse, nous proposons un traitement unifié de l'estimation des fonctions de répartition et quantile F et G^{-1}. La structure proposée contient beaucoup de problèmes classiques d'estimation fonctionnelle : densité de probabilité, régression, fonctions de risque *etc.* (*cf.* Abdous *et al.* (2003)).

2.2 Ajustement polynomial local d'une fonction arbitraire

Nous considérons une fonction de distribution inconnue et arbitraire H, nous désignons par $\Phi(x, H)$ une fonctionnelle indéxée par $x \in (a, b)$ avec $-\infty < a < b < \infty$ deux constantes connues et finies. Nous supposons que nous disposons d'un échantillon $Z_1, \ldots, Z_n$ issu de la distribution H et que $\Phi_n(x)$ est un estimateur de $\Phi(x, H)$. Dans la plupart des cas, cet estimateur est obtenu en remplaçant H dans $\Phi(x, H)$ par son estimateur empirique H_n. Par exemple, dans le contexte d'estimation de la fonction quantile, nous avons

$$\Phi(u, H) := H^{-1}(u) \text{ et } \Phi_n(u) = H_n^{-1}(u) \text{ avec } x \in (0, 1).$$

L'idée sous jacente à l'estimation polynomiale locale est en fait un simple développement de Taylor. Considérons x fixé et que la fonction qui à z associe $\Phi(z, H)$ est suffisamment régulière au voisinage de x. Cette hypothèse, bien que difficile à vérifier en pratique, est essentielle pour valider théoriquement la construction de l'estimateur localement polynomial. Alors, par une approximation de Taylor, lorsque z est situé dans un voisinage du point x,

$$\begin{aligned}\Phi(z,H) &\approx \Phi(x,H) + \Phi'(x,H)(z-x) + \frac{\Phi''(x,H)}{2} + \ldots + \frac{\Phi^{(r)}(x,H)}{r!}(z-x)^r \\ &\approx \sum_{k=0}^{r} \frac{\Phi^{(k)}(x,H)}{k!}(z-x)^k := \sum_{k=0}^{r} a_k (z-x)^k, \end{aligned} \tag{2.1}$$

où $\Phi^{(k)}(x, H)$ est la $k^{\text{ème}}$ dérivée de $\Phi(z, H)$ par rapport à z, évaluée en x. Ce développement permet d'approcher localement la fonction $\Phi(z, H)$ par un polynôme de degré r. Plus formellement, cette idée se traduit par la résolution d'un problème de moindres carrés pondérés. Une version discrète du critère de minimisation pour l'estimation classique de la fonctionnelle $\Phi(x, H)$ est donnée par

$$\min_a \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{Z_i - x}{h}\right) \left\{ \Phi_n(Z_i) - \sum_{k=0}^{r} a_k (Z_i - x)^k \right\}^2, \tag{2.2}$$

où a est le vecteur de dimension $(r+1)$ dont les composantes sont les a_k pour $k = 0, \ldots, r$ et r est l'ordre du polynôme. Dans cette expression, $K(\cdot)$ est une densité de probabilité arbitraire, appelée le noyau d'ajustement qui est supposée pour le moment positive et d'intégrale finie. Elle est utilisée pour diriger la minimisation de (2.2) en mettant un poids plus important aux données Z_i proches de x. La quantité $h = h(n) > 0$ désigne une suite de paramètres de lissage et est appelée la fenêtre d'ajustement. Elle détermine la largeur du voisinage considéré autour de x. Dans notre travail, nous adopterons une autre formulation, version continue, de ce problème fournie par,

$$\min_a \int_a^b \frac{1}{h} K\left(\frac{z - x}{h}\right) \left\{ \Phi_n(z) - \sum_{k=0}^{r} a_k (z - x)^k \right\}^2 \mathrm{d}z. \tag{2.3}$$

Nous désignons par $\mathbf{a} = (\hat{a}_0(x), \hat{a}_1(x), \ldots, \hat{a}_r(x))^\top \in \mathbb{R}^{r+1}$le vecteur qui minimise l'expression (2.3). D'après l'égalité en (2.1), la dérivée $k^{\text{ème}}$ $\Phi^{(k)}(x, H)$ peut être donc estimer

par $\widehat{a_k}(x) \times k!$, pour $k = 0, 1, \ldots, r$. Par conséquent, les coefficients $(\hat{a}_0(x), \hat{a}_1(x), \ldots, \hat{a}_r(x))$ représentent les estimateurs des paramètres $(\Phi(x; H), \Phi^{(1)}(x; H), \ldots, \Phi^{(r)}(x; H))$. La solution $\mathbf{a} = (\hat{a}_0(x), \hat{a}_1(x), \ldots, \hat{a}_r(x))^\top$ du critère (2.3), est donnée par,

$$\begin{aligned}\mathbf{a} &= \mathbb{H}^{-1}\mathbb{M}^{-1}\tilde{\Phi}(x,h) = (\mathbb{M}\mathbb{H})^{-1}\tilde{\Phi}(x,h) \\ &= \begin{pmatrix} \mu_0 & h\mu_1 & \cdots & h^r\mu_r \\ \mu_1 & h\mu_2 & \cdots & h^r\mu_{r+1} \\ \vdots & \vdots & \ddots & \vdots \\ \mu_r & h\mu_{r+1} & \cdots & h^r\mu_{2r} \end{pmatrix}^{-1} \begin{pmatrix} \tilde{\Phi}_0(x,h) \\ \tilde{\Phi}_1(x,h) \\ \vdots \\ \tilde{\Phi}_r(x,h) \end{pmatrix},\end{aligned}$$

où $\mathbb{H}^{-1}$ est l'inverse de la matrice diagonale $\mathrm{diag}(1, h, \ldots, h^r)$ de dimension $(r+1) \times (r+1)$, $\mathbb{M}^{-1}$ est l'inverse de la matrice des moments de dimension $(r+1) \times (r+1)$ avec $M_{ij} = \mu_{i+j-2}$ pour $i, j = 1, \ldots, r+1$ avec

$$\mu_\ell := \mu_\ell(x) = \frac{1}{h}\int_a^b \left(\frac{z-x}{h}\right)^\ell K\left(\frac{z-x}{h}\right) dz = \int_{(a-x)/h}^{(b-x)/h} y^\ell K(y)\mathrm{d}y < \infty, \text{ pour } \ell = 0, \ldots, 2r \tag{2.4}$$

et $\widetilde{\Phi}(x,h) = [\widetilde{\Phi}_0(x,h), \ldots, \widetilde{\Phi}_r(x,h)]^\top$ où

$$\widetilde{\Phi}_i(x,h) = \frac{1}{h}\int_a^b \left(\frac{z-x}{h}\right)^i K\left(\frac{z-x}{h}\right)\Phi_n(z)\mathrm{d}z, \quad i = 0, \ldots, r.$$

Alors, nous obtenons l'estimateur polynomial local de la fonction $\Phi(\cdot)$ qui est donné par,

$$\Phi_{nk}(x) = \int_a^b \frac{1}{h}K_k\left(\frac{z-x}{h}\right)\Phi_n(z)\mathrm{d}z \tag{2.5}$$

où

$$K_k(y) = e_k^\top \mathbb{H}^{-1}\mathbb{M}^{-1}(1, y, \ldots, y^r)^\top K(y)$$

avec $e_k = (0, \ldots, 0, 1, 0, \ldots, 0)$, 1 est à la $(k+1)^{\text{ème}}$ position avec $k = 0, 1, \ldots, r$. Pour plus d'exemples particuliers de fonctions Φ, voir Abdous *et al.* (2003).

Dans ce travail, nous nous limiterons aux ajustements constant, linéaire et quadratique i.e. $r = 0, 1, 2$. En effet, si nous fixons $r = 0, 1, 2$ dans le critère (2.3) et utilisons la valeur optimale obtenue de $a_0(x)$ comme estimateur de $\Phi(x)$, alors nous obtenons les estimateurs constant, linéaire et quadratique suivants,

$$\Phi_{nr}(x) = \int_a^b \frac{1}{h} K_r\left(\frac{z-x}{h}\right) \Phi_n(z)\mathrm{d}z = \int_{(a-x)/h}^{(b-x)/h} K_r(y)\Phi_n(x+hy)\mathrm{d}y, \quad r = 0, 1, 2 \qquad (2.6)$$

où

$$K_r(y) := K_r(y, x) = \begin{cases} \dfrac{1}{\mu_0} K(y), & \text{pour } r = 0\,; \\ \dfrac{\mu_2 - \mu_1 y}{\mu_0\mu_2 - \mu_1^2} K(y), & \text{pour } r = 1\,; \\ \dfrac{(\mu_2\mu_4 - \mu_3^2) - (\mu_1\mu_4 - \mu_2\mu_3)y + (\mu_1\mu_3 - \mu_2^2)y^2}{(\mu_2\mu_4 - \mu_3^2)\mu_0 - (\mu_1\mu_4 - \mu_2\mu_3)\mu_1 + (\mu_1\mu_3 - \mu_2^2)\mu_2} K(y), & \text{pour } r = 2. \end{cases} \qquad (2.7)$$

avec μ_ℓ est défini par (2.4).

Le noyau K_r dépend du point d'intérêt x à travers les moments μ_ℓ. Cette dépendance s'estompe lorsque le point u appartient à l'intérieur de (a, b) et que le noyau initial K est de support $[-1, 1]$. Il prend des formes différentes selon que le point u est proche de a, est à l'intérieur de (a, b) ou est proche de b. En effet, le noyau K_r s'adapte automatiquement aux régions de bord. Il prend différentes formes qui dépendent de la région à laquelle le point d'intérêt u appartient. En l'occurrence, nous décomposons (a, b), le support Φ, en trois régions : La région gauche $B_L := \{z : a < z \leq a + h\}$, la région intérieure $I := \{z : a + h < z < b - h\}$ et la région droite $B_R := \{z : b - h \leq z < b\}$.

Dans la suite, nous présentons quelques lemmes qui seront utilisés dans les chapitres ci-après. Le premier résume les propriétés principales du noyau K_r qui vont nus être utiles dans la suite de ce travail.

Lemme 2.2.1. *Soit K une densité de probabilité symétrique sur $[-1,1]$ ayant des moments d'ordre $2r$ finis où r est un entier dans $\{0,1,2\}$. Fixons u dans (a,b) et supposons que la fenêtre h satisfait $(b-a)>2h$. Alors, le noyau associé $K_r(\cdot) := K_r(\cdot,u)$ donné par (2.7) est une fonction bornée qui dépend de la région à laquelle u appartient, i.e.*

$$K_r(z,u) = \mathbb{I}_{B_L}(u)K_r^L(z,u) + \mathbb{I}_I(u)K_r^I(z) + \mathbb{I}_{B_R}(u)K_r^R(z,u)$$

où

(i) Si $x \in B_L$, le noyau de la région gauche $K_r^L(\cdot,u)$ est défini sur $[-\alpha,1]$ avec α étant fixé dans $[0,1]$ tel que $u = a+\alpha h$. Ses moments d'ordre ℓ satisfont

$$\nu_{r,\ell}^L = \int_{-\alpha}^{1} z^\ell K_r^L(z,u)\mathrm{d}z = \begin{cases} 1, & \text{pour } \ell = 0\,; \\ 0, & \text{pour } \ell = 1,\ldots,r\,; \\ \neq 0, & \text{pour } \ell = r+1. \end{cases}$$

(ii) Si $x \in I = (a+h, b-h)$, le noyau de la région intérieure $K_r^I(\cdot)$ est à support $[-1,1]$ et peut être réécrit

$$K_0^I(y) = K_1^I(y) = K(y), \qquad \text{et } K_2^I(y) = \frac{\mu_4 - \mu_2 y^2}{\mu_4 - \mu_2^2}K(y).$$

Ses moments coincident avec ceux du noyau K pour $r = 0,1$, alors que pour $r = 2$, nous avons

$$\nu_{r,\ell}^I = \int_{-1}^{1} z^\ell K_r^I(z)\mathrm{d}z = \begin{cases} 1, & \text{pour } \ell = 0\,; \\ 0, & \text{pour } \ell = 1,\ldots,2r-1\,; \\ \neq 0, & \text{pour } \ell = 2r. \end{cases}$$

(iii) Si $x \in B_R$, le noyau de la région droite $K_r^R(\cdot,u)$ est défini sur $[-1,\alpha]$ avec α étant fixé dans $[0,1]$ tel que $u = b-\alpha h$. Ses moments d'ordre ℓ satisfont

$$\nu_{r,\ell}^R = \int_{-1}^{\alpha} z^\ell K_r^R(z,u)\mathrm{d}z = \begin{cases} 1, & \text{pour } \ell = 0\,; \\ 0, & \text{pour } \ell = 1,\ldots,r\,; \\ \neq 0, & \text{pour } \ell = r+1. \end{cases}$$

Preuve. Selon (2.7), les noyaux K_r dépendent du point d'intérêt u via les moments μ_ℓ. Mais, quelque soit la valeur de u dans (a, b), ces noyaux sont bornés. En effet, il est aisé de voir que le numérateur dans (2.7) est borné, alors que le dénominateur peut être réécrit comme le déterminant de la matrice $\mathbb{M}$, $(r+1) \times (r+1)$

$$\mathbb{M} = \begin{pmatrix} \mu_0 & \mu_1 & \dots & \mu_r \\ \mu_1 & \mu_2 & \dots & \mu_{r+1} \\ \vdots & \vdots & \ddots & \vdots \\ \mu_r & \mu_{r+1} & \dots & \mu_{2r} \end{pmatrix}. \tag{2.8}$$

Puisque K est une densité de probabilité de moments finis, l'inégalité $|\det \mathbb{M}| \geq C > 0$ est vérifiée pour tout $r \geq 0$, voir Freud (1971), (Chapitre 11). Les affirmations restantes sont directes. □

En résumé, pour expliciter la dépendance entre le noyau K_r et la région où le point d'intérêt u est localisé, nous adopterons les notations suivantes

$$\Phi_{nr}(u) = \int_a^b \frac{1}{h} K_r \left(\frac{z-u}{h} \right) \Phi_n(z) \mathrm{d}z \tag{2.9}$$

$$= \begin{cases} \displaystyle\int_{-\alpha}^{1} K_r^L(z) \Phi_n(u+hz) \mathrm{d}z, & \text{si } u = a + \alpha h \text{ avec } \alpha \in (0,1]\,; \\ \displaystyle\int_{-1}^{1} K_r^I(z) \Phi_n(u+hz) \mathrm{d}z, & \text{si } u \in I\,; \\ \displaystyle\int_{-1}^{\alpha} K_r^R(z) \Phi_n(u+hz) \mathrm{d}z, & \text{si } u = b - \alpha h \text{ avec } \alpha \in (0,1]. \end{cases}$$

Le lemme suivant est une adaptation du lemme de Bochner (voir e.g. Parzen (1962)).

Lemme 2.2.2. *Supposons que $K(\cdot)$ est une densité de probabilité symétrique et bornée à support $[-1, 1]$. On considère $\Phi(y)$ une fonction intégrable sur l'intervalle (a, b). nous définissons*

$$\Phi_h(x) = \int_a^b \frac{1}{h} K_r \left(\frac{z-x}{h} \right) \Phi(z) \mathrm{d}z,$$

avec K_r étant donné par (2.7). Alors, en tout point de continuité x de $\Phi(\cdot)$, nous avons

$$\lim_{h \to 0} \Phi_h(x) = \Phi(x).$$

Preuve. D'abord, nous considérons le cas où x appartient à la région intérieure I, donc pour tout $\delta > 0$, nous avons

$$\begin{aligned}
|\Phi_h(x) - \Phi(x)| &= \left| \int_a^b \frac{1}{h} K_r\left(\frac{z-x}{h}\right) \Phi(z)\mathrm{d}z - \Phi(x) \right| \\
&\leq \int_{-1}^{1} |K_r^I(u)|\, |\Phi(x+hu) - \Phi(x)|\, \mathrm{d}u \\
&\leq \sup_{|hu|\leq\delta} |\Phi(x+hu) - \Phi(x)| \int_{-1}^{1} |K_r^I(u)|\mathrm{d}u + \\
&\quad \int_{|hu|>\delta} |uK_r^I(u)| \frac{|\Phi(x+hu)|}{|u|}\mathrm{d}u + \quad |\Phi(x)| \int_{|hu|>\delta} |K_r^I(u)|\mathrm{d}u \\
&\leq \sup_{|z|\leq\delta} |\Phi(x+z) - \Phi(x)| \int_{-1}^{1} |K_r^I(u)|\mathrm{d}u + \\
&\quad \frac{1}{\delta} \sup_{|u|>\delta/h} |uK_r^I(u)| \int_a^b |\Phi(v)|\mathrm{d}v + |\Phi(x)| \int_{|u|>\delta/h} |K_r^I(u)|\mathrm{d}u
\end{aligned}$$

Or, Φ est continue en x et le noyau K_r^I est borné et a son support sur $[-1, 1]$. Pour conclure, il suffit que les hypothèses $h \to 0$ et $\delta \to 0$ soient vérifiées.

Si le point d'intérêt x appartient à la région de bord gauche B_L, i.e. $x = a + \alpha h$ avec $\alpha \in (0, 1]$, alors la preuve suit les mêmes étapes où le noyau K_r^I et son support $[-1, 1]$ sont respectivement remplacés par K_r^L et $[-\alpha, 1]$. Des arguments similaires sont utilisés quant à la région de bord droite. □

Le lemme suivant exhibe la vitesse de convergence de $\Phi_h(x)$ vers $\Phi(x)$.

Lemme 2.2.3. *Supposons que $K(\cdot)$ est une densité de probabilité symétrique et bornée à support $[-1, 1]$. Soit K_r le noyau associé obtenu de (2.7) avec $r = 0, 1$ ou 2. Supposons que $\Phi(\cdot)$ est une fonction intégrable sur (a, b) et qu'elle admet des dérivées bornées jusqu'à l'ordre 4. Alors,*

(i) Si x appartient à la région de bord gauche B_L i.e. $x = a + \alpha h$, pour $\alpha \in [0, 1]$, alors

$$\Phi_h(x) - \Phi(x) = h^{r+1} \nu_{r,r+1}^L \frac{\Phi^{(r+1)}(a)}{(r+1)!} + o(h^{r+1}).$$

où $\nu_{r,k}^L = \int_{-\alpha}^{1} u^k K_r^L(u)\mathrm{d}u$, $k \geq 0$.

(ii) Si x appartient à la région intérieure I, i.e. $a+h<x<b-h$, alors

$$\Phi_h(x)-\Phi(x)=\begin{cases} h^2\mu_2\Phi^{(2)}(x)/2!+o(h^2), & si\ r=0,1\ ; \\ h^4\nu^I_{2,4}\Phi^{(4)}(x)/4!+o(h^4), & si\ r=2\ ; \end{cases}$$

où $\mu_2=\int_{-1}^1 u^2K(u)\mathrm{d}u$ *et* $\nu^I_{r,k}=\int_{-1}^1 u^kK^I_r(u)\mathrm{d}u$, $k\geq 0$.

(iii) Si x appartient à la région de bord droite B_R i.e. $x=b-\alpha h$, pour $\alpha\in[0,1]$, alors

$$\Phi_h(x)-\Phi(x)=h^{r+1}\nu^R_{r,r+1}\frac{\Phi^{(r+1)}(b)}{(r+1)!}+o(h^{r+1}).$$

où $\nu^R_{r,k}=\int_{-1}^{\alpha} u^kK^R_r(u)\mathrm{d}u$, $k\geq 0$.

Preuve. La preuve de ces expressions asymptotiques est directe. Elle se fonde sur le thèoprème de Taylor et sur les propriétés des noyaux K^L_r, K^I_r et K^R_r cités dans le lemme 2.2.1. Notons que

$$\begin{aligned}
\Phi_h(x)-\Phi(x) &= \int_a^b \frac{1}{h}K_r\left(\frac{z-x}{h}\right)[\Phi(z)-\Phi(x)]\,\mathrm{d}z \\
&= \mathbb{I}(x=a+\alpha h)\int_{-\alpha}^1 K^L_r(u)\,[\Phi(a+(u+\alpha)h)-\Phi(a+\alpha h)]\,\mathrm{d}u \\
&\quad +\mathbb{I}(a+h<x<b-h)\int_{-1}^1 K^I_r(u)\,[\Phi(x+hu)-\Phi(x)]\,\mathrm{d}u \\
&\quad +\mathbb{I}(x=b-\alpha h)\int_{-1}^{\alpha} K^R_r(u)\,[\Phi(b+(u-\alpha)h)-\Phi(b-\alpha h)]\,\mathrm{d}u \\
&= R_1(x)+R_2(x)+R_3(x).
\end{aligned}$$

Un développement de Taylor au voisinage de a permet d'écrire, quand $x\in B_L$, i.e. $x=a+\alpha h$

$$\begin{aligned}
R_1(a+\alpha h) &= \int_{-\alpha}^1 K^L_r(u)\,[\Phi(a+(u+\alpha)h)-\Phi(a+\alpha h)]\,\mathrm{d}u \\
&= \begin{cases} h\nu^L_{0,1}\Phi'(a)+o(h), & \text{si } r=0\,; \\ h^2\nu^L_{1,2}\Phi^{(2)}(a)/2!+o(h^2), & \text{si } r=1\,; \\ h^3\nu^L_{2,3}\Phi^{(3)}(a)/3!+o(h^3), & \text{si } r=2\,; \end{cases}
\end{aligned}$$

avec $\nu^L_{r,k}=\int_{-\alpha}^1 u^kK^L_r(u)\mathrm{d}u$, $k\geq 0$.

Si x appartient à I, i.e. $a+h<x<b-h$, alors

$$\begin{aligned} R_2(x) &= \int_{-1}^{1} K_r^I(u)\left[\Phi(x+uh)-\Phi(x)\right]\mathrm{d}u \\ &= \begin{cases} h^2\nu_{0,2}^I\Phi^{(2)}(x)/2! + o(h^2), & \text{si } r=0\,; \\ h^2\nu_{1,2}^I\Phi^{(2)}(x)/2! + o(h^2), & \text{si } r=1\,; \\ h^4\nu_{2,4}^I\Phi^{(4)}(x)/4! + o(h^4), & \text{si } r=2\,; \end{cases} \end{aligned}$$

avec $\nu_{r,k}^I = \int_{-1}^{1} u^k K_r^I(u)\mathrm{d}u$, $k \geq 0$.

Finalement, pour $x \in B_R$, i.e. $x = x_N - \alpha h$, nous avons

$$\begin{aligned} R_3(b-\alpha h) &= \int_{-1}^{\alpha} K_r^R(u)\left[\Phi(b+(u-\alpha)h)-\Phi(b-\alpha h)\right]\mathrm{d}u \\ &= \begin{cases} h\nu_{0,1}^R\Phi^{'}(b) + o(h), & \text{si } r=0\,; \\ h^2\nu_{1,2}^R\Phi^{(2)}(b)/2! + o(h^2), & \text{si } r=1\,; \\ h^3\nu_{2,3}^R\Phi^{(3)}(b)/3! + o(h^3), & \text{si } r=2\,; \end{cases} \end{aligned}$$

avec $\nu_{r,k}^R = \int_{-1}^{\alpha} u^k K_r^R(u)\mathrm{d}u$, $k \geq 0$. □

Dans la prochaine section, nous appliquerons ce cadre général d'ajustement polynomial local à notre problème d'égalisation équipercentile.

2.3 Ajustement polynomial local de la fonction égalisation équipercentile

L'adaptation de cette technique d'estimation à notre problème d'égalisation équipercentile est immédiate. En effet, nous avons plusieurs scénarios possibles. En reprenant l'estimateur empirique de la fonction équipercentile $y_{m,n}(x) = G_m^{-1}(F_n(x))$, on voit qu'il est possible de lisser $F_n(\cdot)$ uniquement, ou bien lisser $G_m^{-1}(\cdot)$ uniquement, ou bien lisser simultanément et séparément $F_n(\cdot)$ et $G_m^{-1}(\cdot)$ ou encore lisser l'estimateur $G_m^{-1}\circ F_n(\cdot)$. Ces divers scénarios nous conduisent à considérer les 5 estimateurs suivants :

i) Nous utilisons l'estimateur empirique et définissons

$$y_{m,n}^{[1]}(x) = G_m^{-1}(F_n(x)), \quad x \in [x_1, x_N].$$

ii) Nous lissons F_n uniquement et définissons

$$y_{m,n}^{[2]}(x) = G_m^{-1}(\widetilde{F_n}(x)), \quad x \in [x_1, x_N].$$

Ici, la fonction quantile empirique G_m^{-1} reste inchangée, tandis que F_n est remplacée par $\widetilde{F_n}(x)$, son estimateur d'ajustement polynomial local obtenu de (2.9) après avoir remplacé Φ_n par F_n et posé $a = x_1$ et $b = x_N$, i.e.

$$\widetilde{F_n}(x) = \int_{x_1}^{x_N} \frac{1}{h} K_r\left(\frac{z-x}{h}\right) F_n(z)\mathrm{d}z.$$

iii) Nous lissons G_m^{-1} uniquement et définissons

$$y_{m,n}^{[3]}(x) = \widetilde{G_m^{-1}}(F_n(x)) = \int_0^1 \frac{1}{h} K_r\left(\frac{z-F_n(x)}{h}\right) G_m^{-1}(z)\mathrm{d}z, \quad x \in [x_1, x_N].$$

Seule la fonction quantile G_m^{-1} est lissée. Elle est remplacée par $\widetilde{G_m}^{-1}$, obtenue de (2.6) après avoir remplacé Φ_n par G_m^{-1} et posé $a = 0$ et $b = 1$.

iv) Nous lissons séparément et simultanément F_n et G_m^{-1} et définissons

$$\begin{aligned} y_{m,n}^{[4]}(x) = \widetilde{G_m^{-1}}(\widetilde{F_n}(x)) &= \int_0^1 \frac{1}{k} L_r\left(\frac{z-\widetilde{F_n}(x)}{k}\right) G_m^{-1}(z)\mathrm{d}z, \quad \text{avec} \\ \widetilde{F_n}(x) &= \int_{x_1}^{x_N} \frac{1}{h} K_r\left(\frac{z-x}{h}\right) F_n(z)\mathrm{d}z. \end{aligned}$$

Ici, nous avons lissé séparément G_m^{-1} et F_n en prenant deux fenêtres différentes k et h et deux noyaux K_r et L_r.

v) Nous lissons en une seule fois le processus Q-Q, $G_m^{-1} \circ F_n(\cdot)$ et définissons

$$y_{m,n}^{[5]}(x) = \widetilde{G_m^{-1} \circ F_n}(x)) = \int_{x_1}^{x_N} \frac{1}{h} K_r\left(\frac{z-x}{h}\right) G_m^{-1} \circ F_n(z)\mathrm{d}z, \quad x \in [x_1, x_N].$$

En effet, la quantité $[G_m^{-1}(F_n(x)) - x]$ notée aussi $G_m^{-1} \circ F_n(x) - x$, est connue dans la littérature et se nomme généralement processus Quantile-Quantile ou encore Q-Q plot et sa première utilisation remonte à (1905) par Lorenz. Ses propriétés asymptotiques ont été étudiées par plusieurs auteurs par exemple, Doksum (1974), Doksum et Sievers (1976), Aly (1986), Beirlant et Deheuvels (1990). Les principales propriétés asymptotiques sont récapitulées dans le chapitre suivant. Tous ces travaux utilisent les techniques du processus Q-Q afin de comparer deux distributions F et G. Alors que dans notre cas, ces techniques sont utilisées dans le but d'estimer l'"équivalent" y (dans la distribution G) pour un x (dans la distribution F).

Afin de pallier le problème de la discontinuité que présente l'estimateur empirique de la fonction d'égalisation équipercentile, nous avons fait appel à la technique de lissage par l'ajustement polynomial local. L'objet des chapitres suivants est d'étudier le comportement asymptotique de ces estimateurs. Dans le contexte de la qualité de vie, x représente le score mesuré par un questionnaire donné et y serait le score "équivalent" obtenu par un autre questionnaire. Souvent, les épidémiologistes et cliniciens disposent d'un instrument X et aimeraient traduire (ou interpréter) le score mesuré par cet instrument en un score équivalent qui aurait été obtenu par un autre instrument de référence Y.

Chapitre 3

Théorie asymptotique du processus Q-Q Égalisation équipercentile

Résumé. Nous rappelons dans ce chapitre les principaux résultats des processus empirique, quantile et Quantile-Quantile. En effet, l'estimation de la fonction de répartition et la fonction quantile a suscité beaucoup d'attention dans la littérature (voir par exemple Csörgő (1983), Shorack et Wellner (1986)). Notons que, puisque nous sommes principalement intéressés par les problèmes liés à la qualité de vie, nous supposons que les supports de F et G sont des intervalles bornés $S(F)$ et $S(G)$ respectivement.

3.1 Propriétés asymptotiques du processus empirique, processus quantile

Soit $\{Z_i : i \geq 1\}$ une suite de variables aléatoires indépendantes et identiquement distribuées, de fonction de répartition H continue à support $S(H)$. La fonction de répartition

empirique associée à la suite $(Z_1, \ldots, Z_n)$ est définie, pour chaque entier $n \geq 1$, par

$$H_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{Z_i \leq x\}}, \quad \text{pour } x \in S(H). \tag{3.1}$$

Ainsi, nous définissons le processus empirique $\{\beta_n(x); x \in S(H)\}$ par

$$\{\beta_n(x); x \in S(H)\} = \{\sqrt{n}(H_n(x) - H(x)); x \in S(H)\}. \tag{3.2}$$

Il est bien connu que pour tout échantillon aléatoire $Z_1, \ldots, Z_n$ de fonction de répartition continue H à support $S(H)$, les variables aléatoires $U_i = H(Z_i)$, $(i = 1, 2, \ldots, n)$ forment un échantillon aléatoire de distribution uniforme sur $[0, 1]$ (voir Csörgő (1983), chapitre 1). Nous définissons ainsi le processus uniforme empirique correspondant $\{\alpha_n(u); 0 \leq u \leq 1\}$ par

$$\{\alpha_n(u); 0 \leq u \leq 1\} = \{\sqrt{n}(E_n(u) - u), 0 \leq u \leq 1\}, \quad n \geq 1, \tag{3.3}$$

où $E_n(u) = H_n(H^{-1}(u)) = \sum_{i=1}^{n} \mathbb{I}(U_i \leq u)/n$. Alors, les deux processus $\beta_n(\cdot)$ et $\alpha_n(\cdot)$ satisfont

$$\{\beta_n(H^{-1}(u)); 0 \leq u \leq 1\} \stackrel{\mathcal{D}}{=} \{\alpha_n(u); 0 \leq u \leq 1\}, \quad n \geq 1. \tag{3.4}$$

En outre, nous définissons le processus quantile uniforme $\{u_n(u); 0 \leq u \leq 1\}$ par

$$\{u_n(u); 0 \leq u \leq 1\} = \{\sqrt{n}(E_n^{-1}(u) - u), 0 \leq u \leq 1\}, \quad n \geq 1,$$

et le processus quantile général $\{q_n(u); 0 \leq u \leq 1\}$ par

$$\{q_n(u); 0 \leq u \leq 1\} = \{\sqrt{n}(H_n^{-1}(u) - H^{-1}(u)), 0 \leq u \leq 1\}, \quad n \geq 1.$$

Lemme 3.1.1. *Soit H une fonction de répartition continue à support fini $S(H)$ sur $\mathbb{R}$. Soit $Z_1, \ldots, Z_n$ un échantillon de fonction de répartition H. D'après la loi forte des grands nombres, nous avons, $\forall\ x \in S(H)$,*

$$H_n(x) \xrightarrow{p.s.} H(x), \quad \textit{quand } n \to \infty. \tag{3.5}$$

Si $H^{-1}(u)$ est continue en u, alors

$$H_n^{-1}(u) \xrightarrow{p.s.} H^{-1}(u), \tag{3.6}$$

où p.s. signifie presque sûrement.

Corollaire 3.1.1. Cantelli et Glivenko (1933) ont obtenu, séparément, que cette convergence est uniforme sur $S(H)$,

$$\sup_{x\in S(H)} |H_n(x) - H(x)| \xrightarrow{\text{p.s.}} 0, \text{ quand } n \to \infty. \tag{3.7}$$

En outre, si H est deux fois dérivable telle que $\inf_{0\leq u\leq 1} H'(H^{-1}(u)) > 0$ et $\sup_{0\leq u\leq 1} H''(H^{-1}(u)) < \infty$. Alors,

$$\sup_{0\leq u\leq 1} |H_n^{-1}(u) - H^{-1}(u)| \xrightarrow{\text{p.s.}} 0, \text{ quand } n \to \infty. \tag{3.8}$$

Remarque 3.1.1. Il est commode de se ramener au cas du processus empirique uniforme $\alpha_n(u)$ défini par (3.3). Nous avons

$$\sup_{0\leq u\leq 1} |E_n(u) - u| = \sup_{0\leq u\leq 1} |E_n^{-1}(u) - u|, \tag{3.9}$$

et nous obtenons

$$\sup_{0\leq u\leq 1} |E_n^{-1}(u) - u| \xrightarrow{\text{p.s.}} 0. \tag{3.10}$$

Preuve. Voir le livre de Csörgő (1983) p. 4-5.

Définition 3.1.1. Un processus de Wiener noté $\{W(t), t \geq 0\}$ est un processus gaussien, centré $(E(W(t)) = 0)$, à trajectoires continues et de fonction de covariance

$$\mathbb{E}\left(W(t)W(s)\right) = (t \wedge s)\sigma^2, \forall\ t \geq 0, \forall\ s \geq 0.$$

Lorsque $\sigma^2 = 1$, le processus de Wiener est dit *standard*.

Définition 3.1.2. Un pont brownien noté $\{B(t) : 0 \leq t \leq 1\}$ est un processus gaussien, centré $(E(B(t)) = 0)$, à trajectoires continues, et de fonction de covariance

$$\mathbb{E}\left(B(t)B(s)\right) = (t \wedge s) - ts, \quad \text{pour } 0 \leq t \leq 1 \text{ et } 0 \leq s \leq 1.$$

Si $\{W(t),\ 0 \leq t < \infty\}$ est un processus de Wiener, alors

$$B(t) = W(t) - tW(1), \quad 0 \leq t \leq 1, \tag{3.11}$$

est un pont brownien. Le pont brownien est identique à la restriction d'un processus de Wiener standard à $[0,1]$, conditionné par $\{W(1)=0\}$.

Nous savons que le processus empirique $\beta_n(x)=\sqrt{n}\,[H_n(x)-H(x)]$ converge en loi vers $B_n(H(x))$ (voir Billingsly (1968)), où $\{B_n(t):0\leq t\leq 1\}$ désigne une suite de ponts browniens. La vitesse de convergence de $\|\beta_n - B_n(H)\|$ qui peut être établie par une construction optimale de B_n est une question importante dans le domaine des statistiques et des probabilités. Plusieurs auteurs se sont intéressés à ce sujet (voir par exemple les livres de Csörgő et Révész (1981) et de Shorack et Wellner (1986)). La meilleure estimation obtenue est due à Komlós *et al.* (1975). De leurs méthodes sont issus beaucoup de travaux (voir par exemple les articles de Mason et van Zwet (1987), Massart (1989)). Nous avons à ce sujet les théorèmes suivants :

Théorème 3.1.1. *Soit $Z_1,\ldots,Z_n$ un échantillon de fonction de répartition continue H à support fini $S(H)=(a,b)$ avec*

$$a=\sup\{x:H(x)=0\},\quad b=\inf\{x:H(x)=1.\}$$

On suppose que les conditions suivantes sont vérifiées :

(i) H est deux fois dérivable ;

(ii) $H'(x)=h(x)>0$, sur (a,b) et $0<h(a):=h(a+)<\infty$, $0<h(b):=h(b-)<\infty$;

(iii) pour $\gamma>0$, nous avons

$$\sup_{a<x<b} H(x)(1-H(x))\frac{|h'(x)|}{h^2(x)}=\sup_{0<y<1} y(1-y)\frac{|h'(H^{-1}(y))|}{h^2(H^{-1}(y))}\leq\gamma.$$

Alors, il existe une suite de ponts browniens $\{B_n(t);0\leq t\leq 1\}$ telle que

$$\sup_{t\in S(H)}|\beta_n(t)-B_n(H(t))| \;=\; \mathcal{O}(n^{-1/2}\log n),\ p.s. \tag{3.12}$$

où le processus $\beta_n(\cdot)$ est défini par

$$\{\beta_n(t);t\in S(H)\}=\{\sqrt{n}\,[H_n(t)-H(u)]\,;t\in S(H)\}. \tag{3.13}$$

Preuve. Voir l'article de Komlós *et al.* (1975).

Théorème 3.1.2. *Soit $Z_1, \ldots, Z_n$ un échantillon de fonction de répartition continue H à support fini $S(H) = (a, b)$. On suppose vérifiées les conditions (i) et (ii) du théorème (3.1.1). Alors, il existe une suite de ponts browniens $\{\widetilde{B}_n(t); 0 \leq t \leq 1\}$ telle que*

$$\sup_{\delta_n \leq u \leq 1-\delta_n} |\rho_n(u) - \widetilde{B}_n(u)| \overset{p.s.}{=} \mathcal{O}(n^{-1/2} \log n), \tag{3.14}$$

où le processus $\rho_n(\cdot)$ est donné par,

$$\{\rho_n(u); 0 \leq u \leq 1\} = \{\sqrt{n} h(H^{-1}(u)) \left[H_n^{-1}(u) - H^{-1}(u)\right]; 0 \leq u \leq 1\}. \tag{3.15}$$

où $\delta_n = 25 n^{-1} \log\log n$. Si en plus des conditions (i) et (ii), on suppose que

(iv)

$$A := \overline{\lim_{x \downarrow a}}\, h(x) < \infty, \quad B := \overline{\lim_{x \uparrow b}}\, h(x) < \infty;$$

(v) Soit

(v,α) $\min(A, B) > 0$;

(v,β) Si $A = 0$ (resp. $B = 0$) alors h ne décroît pas (resp. ne croît pas) sur un intervalle à droit de a (resp. à gauche de b) ;

Alors, si (v, α) est obtenu,

$$\sup_{0 \leq u \leq 1} |\rho_n(u) - \widetilde{B}_n(u)| \overset{p.s.}{=} \mathcal{O}(n^{-1/2} \log n), \tag{3.16}$$

et si (v, β) est vérifié,

$$\sup_{0 < u < 1} |\rho_n(u) - \widetilde{B}_n(u)| \overset{p.s.}{=} \begin{cases} \mathcal{O}(n^{-1/2} \log n), & \text{si } \gamma < 2\,; \\ \mathcal{O}(n^{-1/2} (\log\log n)^{\gamma} (\log n)^{(1+\varepsilon)(\gamma-1)}), & \text{si } \gamma \geq 2, \end{cases} \tag{3.17}$$

où γ est défini dans le théorème (3.1.1) et $\varepsilon > 0$ est arbitraire.

Preuve. Voir l'article de Csörgő et Révész (1978).

3.2 Propriétés asymptotiques du processus Quantile-Quantile

Soient $X = X_1, \ldots, X_n$ (respectivement (resp.) $Y = Y_1, \ldots, Y_m$) des variables aléatoires, indépendantes et identiquement distribuées de fonction de répartition $F(x) = P(X \leq x)$ (resp. $G(x) = P(X \leq x)$) à support fini $S(F) = [x_1, x_N]$ (resp. $S(G) = [y_1, y_M]$). Soit $F^{-1}(\cdot)$ (resp. $G^{-1}(\cdot)$) la fonction *inverse*, ou *de quantiles*, de $F(\cdot)$ (resp. $G(\cdot)$) définie, pour $0 \leq s \leq 1$, par

$$F^{-1}(s) := \inf\{x : F(x) \geq s\} \quad \text{pour} \quad 0 < s < 1, \tag{3.18}$$
$$F^{-1}(0) = F^{-1}(0^+) = \lim_{s \downarrow 0} F^{-1}(s) \text{ et } F^{-1}(1) = F^{-1}(1^-) = \lim_{s \uparrow 1} F^{-1}(s).$$

(resp. $$G^{-1}(s) := \inf\{x : G(x) \geq s\} \quad \text{pour} \quad 0 < s < 1, \tag{3.19}$$
$$G^{-1}(0) = G^{-1}(0^+) = \lim_{s \downarrow 0} G^{-1}(s) \text{ et } G^{-1}(1) = G^{-1}(1^-) = \lim_{s \uparrow 1} G^{-1}(s).)$$

On suppose que les dérivées des fonctions $F(\cdot)$ et $G(\cdot)$ existent et sont respectivement notées par $f(\cdot)$ et $g(\cdot)$. Désignons par

$$F_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{]-\infty,t]}(X_i), \text{ pour } t \in S(F), \tag{3.20}$$

la fonction de répartition empirique de l'échantillon $X_1, \ldots, X_n$, et désignons par $X_{1,n} \leq \ldots \leq X_{n,n}$, les statistiques d'ordre correspondantes. La fonction de quantile empirique associée à $F_n(\cdot)$ est alors donnée par

$$F_n^{-1}(p) = \begin{cases} X_{k,n} & \text{si } \frac{k-1}{n} < p \leq \frac{k}{n}, \\ X_{1,n} & \text{si } p = 0. \end{cases} \tag{3.21}$$

Définissons, de manière analogue, la fonction de répartition empirique et la fonction de quantile empirique correspondant à l'échantillon $Y = Y_1, \ldots, Y_m$ par

$$G_m(y) := \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}_{]-\infty,y]}\big(Y_i\big), \text{ pour } y \in S(G), \tag{3.22}$$

et

$$G_m^{-1}(s) = \begin{cases} Y_{j,m} & \text{si } \frac{j-1}{m} < s \leq \frac{j}{m}, \\ Y_{1,m} & \text{si } s = 0, \end{cases} \tag{3.23}$$

où on désigne par $Y_{1,m} \leq \ldots \leq Y_{m,m}$, les statistiques d'ordre correspondant à $Y_1, \ldots, Y_m$. Par conséquent, le processus Q-Q empirique, $G_m^{-1}(F_n(x))$ est un estimateur naturel de $G^{-1}(F(x))$. Ainsi, nous définissons le processus Quantile-Quantile par

$$\Delta_{n,m}(x) := \sqrt{\frac{nm}{n+m}} g(G^{-1}(F(x))) \left[G_m^{-1}(F_n(x)) - G^{-1}(F(x))\right], \quad x \in S(F), \tag{3.24}$$

où $g(\cdot)$ est la dérivée de $G(\cdot)$.

Nombreux sont les articles sur les propriétés asymptotiques du processus Q-Q (voir par exemple, Doksum (1974), Aly (1986), Beirlant et Deheuvels (1990)). Ce sont des analogues des théorèmes de Glivenko-Cantelli pour les processus Q-Q. Nous avons à ce sujet les théorèmes suivants adaptés au cas des supports finis $S(F)$ et $S(G)$:

Théorème 3.2.1. *Soient X et Y deux variables aléatoires de fonctions de répartition F et G et de supports $S(F) = [x_1, x_N]$ et $S(G) = [y_1, y_M]$ respectivement. Supposons que G soit continûment dérivable et telle que $G' = g$, $\inf_{0 \leq u \leq 1} g(G^{-1}(u)) > 0$ et $\sup_{0 \leq u \leq 1} |g'(G^{-1}(u))| < \infty$. Alors,*

$$\sup_{x \in S(F)} \left|G_m^{-1}(F_n(x)) - G^{-1}(F(x))\right| \xrightarrow{p.s.} 0. \tag{3.25}$$

Nous utilisons le fait (Komlós *et al.* (1975) et Csörgő et Révész (1978)) que nous pouvons définir quatre suites de ponts Browniens $B_{F,n}$, $B_{G,m}$, $\widetilde{B}_{F,n}$ et $\widetilde{B}_{G,m}$ pour $n \geq 1$, $m \geq 1$, sur le même espace de probabilité telles que

(i) $\left\{B_{F,n}, \widetilde{B}_{F,n} : n \geq 1\right\}$ et $\left\{B_{G,m}, \widetilde{B}_{G,m} : m \geq 1\right\}$ sont indépendantes ;

(ii) quand $n \to \infty$ et $m \to \infty$ on a presque sûrement,

$$\begin{aligned} \sup_{S(F)} \left|\sqrt{n}\left[F_n(x) - F(x)\right] - B_{F,n}(F(x))\right| &= \mathcal{O}\left(\frac{\log n}{\sqrt{n}}\right), \\ \sup_{S(G)} \left|\sqrt{m}\left[G_m(x) - G(x)\right] - B_{G,m}(G(x))\right| &= \mathcal{O}\left(\frac{\log m}{\sqrt{m}}\right), \end{aligned} \tag{3.26}$$

et

$$\sup_{0\leq t\leq 1}\left|\sqrt{n}f\left(F^{-1}(t)\right)\left[F_n^{-1}(t)-F^{-1}(t)\right]-\widetilde{B}_{F,n}(t)\right|=\mathcal{O}\left(\frac{\log n}{\sqrt{n}}\right), \tag{3.27}$$
$$\sup_{0\leq t\leq 1}\left|\sqrt{m}g\left(G^{-1}(t)\right)\left[G_m^{-1}(t)-G^{-1}(t)\right]-\widetilde{B}_{G,m}(t)\right|=\mathcal{O}\left(\frac{\log m}{\sqrt{m}}\right).$$

Dans le théorème suivant, nous avons adapté le résultat obtenu par Beirlant et Deheuvels (1990) au cas des fonctions de répartition quelconques :

Théorème 3.2.2. *Supposons que G est deux fois dérivable sur $S(G)$ et que $\forall\ \ \gamma>0$ nous avons $\sup_{y\in[0,1]} y(1-y)|g'(G^{-1}(y))|/g^2(G^{-1}(y))\leq\gamma$. Notons $G'(t)=g(t)>0$. Il existe une suite de ponts browniens doublement indexée $\{B_{m;n}(x): m\geq 1, n\geq 1, x\in S(F)\}$ telle que, lorsque $n\to\infty$ et $m\to\infty$*

$$\begin{aligned}\sup_{x\in S(F)}\left|\Delta_{n,m}(x)-B_{n,m}(x)\right| &= \mathcal{O}_P\left(\sqrt{\frac{m}{n+m}}\,\frac{\log n}{\sqrt{n}}\right)+\mathcal{O}_P\left(\sqrt{\frac{n}{n+m}}\,\frac{\log m}{\sqrt{m}}\right)\\ &\quad+\mathcal{O}_P\left(\sqrt{\frac{n}{n+m}}\,\frac{(\log n)^3}{n}\right)^{1/4}\end{aligned} \tag{3.28}$$

où

$$B_{m,n}(x):=\sqrt{\frac{nm}{n+m}}\left(m^{-1/2}\widetilde{B}_{G,m}(F(x))+n^{-1/2}B_{F,n}(F(x))\right). \tag{3.29}$$

3.3 Preuves

3.3.1 Preuve du théorème 3.2.1

Preuve. La convergence uniforme presque sûre du processus Q-Q a été étudiée par plusieurs auteurs (par exemple Doksum (1974, 1976), Aly (1986), Aly, Csörgő and Horvàth (1987)). Nous avons

$$\begin{aligned}\left|G_m^{-1}(F_n(x))-G^{-1}(F(x))\right| &= \left|G_m^{-1}\circ F_n(x)-G^{-1}\circ F(x)\right|\\ &\leq \left|G_m^{-1}\circ F_n(x)-G^{-1}\circ F_n(x)\right|+\left|G^{-1}\circ F_n(x)-G^{-1}\circ F(x)\right|.\end{aligned}$$

La convergence uniforme presque sûre de $G_m^{-1}(\cdot)$ est obtenue par (3.8) en remplaçant H par G. Par conséquent,

$$\sup_{x_1\leq x\leq x_N}\left|G_m^{-1}\circ F_n(x)-G^{-1}\circ F_n(x)\right|\leq \sup_{u\in[0,1]}\left|G_m^{-1}(u)-G^{-1}(u)\right|\xrightarrow{\text{p.s.}}0$$

La convergence uniforme presque sûre de $\left|G^{-1}\circ F_n(x)-G^{-1}\circ F(x)\right|$ est déduite en utilisant la continuité uniforme de G^{-1} et le résultat classique de la convergence uniforme presque sûre de $F_n(x)$, $\sup_{x\in S(F)}|F_n(x)-F(x)|\xrightarrow{\text{p.s.}}0$. □

3.3.2 Preuve du théorème 3.2.2

Preuve. Nous reprenons les étapes de la preuve donnée par Beirlant et Deheuvels. Notons que

$$\begin{aligned}\Delta_{n,m}(x)=&\sqrt{\frac{n}{n+m}}\frac{g\left(G^{-1}(F(x))\right)}{g\left(G^{-1}(F_n(x))\right)}\widetilde{B}_{G,m}(F_n(x))\\&+\sqrt{\frac{m}{n+m}}\frac{g\left(G^{-1}(F(x))\right)}{g\left(G^{-1}(\theta_{x,n})\right)}B_{F,n}(F(x))+R_1(x)\end{aligned}$$

avec

$$\begin{aligned}R_1(x):=&\sqrt{\frac{m}{n+m}}\frac{g\left(G^{-1}(F(x))\right)}{g\left(G^{-1}(\theta_{x,n})\right)}\left\{\sqrt{n}\left[F_n(x))-F(x)\right]-B_{F,n}(F(x))\right\}\\&+\sqrt{\frac{n}{n+m}}\frac{g\left(G^{-1}(F(x))\right)}{g\left(G^{-1}(F_n(x))\right)}\\&\times\left\{\sqrt{m}g\left(G^{-1}(F_n(x))\right)\left[G_m^{-1}(F_n(x))-G^{-1}(F_n(x))\right]-\widetilde{B}_{G,m}(F_n(x))\right\}\end{aligned}$$

En posant

$$B_{n,m}(x):=\sqrt{\frac{nm}{n+m}}\left(m^{-1/2}\widetilde{B}_{G,m}(F(x))+n^{-1/2}B_{F,n}(F(x))\right),\quad x\in S(F)$$

La quantité $\Delta_{n,m}(x)$ peut être réécrite sous la forme

$$\Delta_{n,m}(x)=B_{m,n}(x)+R_1(x)+R_2(x)$$

avec

$$\begin{aligned}
R_2(x) &= \sqrt{\frac{n}{n+m}}\left\{\frac{g\left(G^{-1}(F(x))\right)}{g\left(G^{-1}(F_n(x))\right)}\widetilde{B}_{G,m}(F_n(x)) - \widetilde{B}_{G,m}(F(x))\right\} \\
&\quad + \sqrt{\frac{m}{n+m}}B_{F,n}(F(x))\left\{\frac{g\left(G^{-1}(F(x)\right)}{g\left(G^{-1}(\theta_{x,n})\right)} - 1\right\} \\
&= \sqrt{\frac{n}{n+m}}\frac{g\left(G^{-1}(F(x))\right)}{g\left(G^{-1}(F_n(x))\right)}\left\{\widetilde{B}_{G,m}(F_n(x)) - \widetilde{B}_{G,m}(F(x))\right\} \\
&\quad + \sqrt{\frac{n}{n+m}}\widetilde{B}_{G,m}(F(x))\left\{\frac{g\left(G^{-1}(F(x))\right)}{g\left(G^{-1}(F_n(x))\right)} - 1\right\} \\
&\quad + \sqrt{\frac{m}{n+m}}B_{F,n}(F(x))\left\{\frac{g\left(G^{-1}(F(x)\right)}{g\left(G^{-1}(\theta_{x,n})\right)} - 1\right\} \\
&:= R_{21}(x) + R_{22}(x) + R_{23}(x)
\end{aligned} \tag{3.30}$$

Puisque les conditions du Lemme 3 de Beirlant et Deheuvels (1990) entraînent que $0 < g(x) < \infty$ pour tout $x \in S(G)$, on a pour $m \wedge n \to \infty$

$$\sup_{x\in S(F)}|R_1(x)| \overset{\text{p.s.}}{=} \sqrt{\frac{m}{n+m}}\mathcal{O}\left(\frac{\log n}{\sqrt{n}}\right) + \sqrt{\frac{n}{n+m}}\mathcal{O}\left(\frac{\log m}{\sqrt{m}}\right) \tag{3.31}$$

Quant à R_2, le développement de Taylor permet d'écrire

$$\frac{g\left(G^{-1}(u)\right)}{g\left(G^{-1}(v_n)\right)} - 1 = \frac{u - v_n}{g\left(G^{-1}(v_n)\right)}\frac{g'\left(G^{-1}(\xi)\right)}{g\left(G^{-1}(\xi)\right)}$$

où $u, v_n \in [0,1]$ et $u \wedge v_n < \xi < u \vee v_n$. Ainsi, en prenant $u = F(x)$ et $v_n = F_n(x)$ et en supposant que

$$\inf_{0\le u\le 1} g\left(G^{-1}(u)\right) > 0, \quad \sup_{0\le u\le 1}|g'\left(G^{-1}(u)\right)| < \infty$$

on obtient

$$\left|\frac{g\left(G^{-1}(F(x))\right)}{g\left(G^{-1}(F_n(x))\right)} - 1\right| = \mathcal{O}(|F_n(x) - F(x)|), \text{ p.s.}$$

De façon similaire, si on prend $u = F(x)$ et $v_n = \theta_{x,n}$, on aura

$$\left|\frac{g\left(G^{-1}(F(x))\right)}{g\left(G^{-1}(\theta_{x,n})\right)} - 1\right| = \mathcal{O}(|\theta_{x,n} - F(x)|) = \mathcal{O}(|F_n(x) - F(x)|), \text{ p.s.}$$

Par conséquent, on a l'égalité p.s. suivante

$$\begin{aligned}
&R_{22}(x) + R_{23}(x) = \\
&\quad \mathcal{O}(|F_n(x) - F(x)|)\left\{\sqrt{\frac{n}{n+m}}\sup_{u\in(0,1)}|\widetilde{B}_{G,m}(u)| + \sqrt{\frac{m}{n+m}}\sup_{x\in S(F)}|B_{F,n}(F(x))|\right\}
\end{aligned} \tag{3.32}$$

Par ailleurs, il est bien connu que

$$\sup_{x\in S(F)} |F_n(x) - F(x)| \overset{\text{p.s.}}{=} \mathcal{O}\left(\left(\frac{\log\log n}{n}\right)^{1/2}\right) \tag{3.33}$$

et que, pour toute suite de ponts browniens $\{B_n(t), 0 \leq t \leq 1\}$,

$$P\left(\sup_{0\leq t\leq 1} |B_n(t)| > x\right) \leq 2\exp^{-2x^2}, \quad x \geq 0$$

(voir, e.g. Csörgő et Révész (1981)). Une application du lemme de Borel-Cantelli et un choix approprié de x permettent d'écrire

$$\sup_{0\leq t\leq 1} |B_n(t)| \overset{\text{p.s.}}{=} \mathcal{O}\left((\log n)^{1/2}\right) \tag{3.34}$$

En injectant (3.33) et (3.34) dans (3.32), on obtient

$$R_{22}(x) + R_{23}(x) \overset{\text{p.s.}}{=} \sqrt{\frac{nm}{n+m}}\sqrt{\frac{\log\log n}{n}}\mathcal{O}\left(\sqrt{\frac{\log m}{m}} + \sqrt{\frac{\log n}{n}}\right). \tag{3.35}$$

Dans la suite et sans perte de généralité, nous supposerons que les ponts browniens $B_{F,n}$ $B_{G,m}$, $\widetilde{B}_{F,n}$ et $\widetilde{B}_{G,m}$ sont des restrictions à $[0,1]$ de suites de ponts browniens étendus définis sur la droite réelle et que ces ponts étendus sont eux-même définis à partir de suites de processus de Wiener étendus sur $\mathbb{R}$.

Traitons maintenant le terme $R_{21}(x)$. En effet, posons, $\beta_n(x) = \sqrt{n}(F_n(x) - F(x))$ pour $x \in S(F)$ et reprenons l'expression de R_{21} fournie par (3.30). Nous avons

$$\begin{aligned} R_{21}(x) &= \sqrt{\frac{n}{n+m}}\frac{g\left(G^{-1}(F(x))\right)}{g\left(G^{-1}(F_n(x))\right)}\left\{\widetilde{B}_{G,m}(F_n(x)) - \widetilde{B}_{G,m}(F(x))\right\} \\ &= \sqrt{\frac{n}{n+m}}\frac{g\left(G^{-1}(F(x))\right)}{g\left(G^{-1}(F_n(x))\right)}\left\{\widetilde{B}_{G,m}\left(F(x) + n^{-1/2}\beta_n(x)\right) - \widetilde{B}_{G,m}(F(x))\right\} \\ &:= R_{21,1}(x) + R_{21,2}(x) + R_{21,3}(x) \end{aligned}$$

avec

$$\begin{aligned}
R_{21,1}(x) &= \sqrt{\frac{n}{n+m}}\frac{g\left(G^{-1}(F(x))\right)}{g\left(G^{-1}(F_n(x))\right)}\Big\{W_m\left(F(x)+n^{-1/2}\beta_n(x)\right) \\
&\qquad\qquad - W_m\left(F(x)+n^{-1/2}B_{F,n}(F(x))\right)\Big\} \\
R_{21,2}(x) &= \sqrt{\frac{n}{n+m}}\frac{g\left(G^{-1}(F(x))\right)}{g\left(G^{-1}(F_n(x))\right)}\Big\{W_m\left(F(x)+n^{-1/2}B_{F,n}(F(x))\right) - W_m(F(x))\Big\} \\
R_{21,3}(x) &= -\sqrt{\frac{1}{n+m}}\frac{g\left(G^{-1}(F(x))\right)}{g\left(G^{-1}(F_n(x))\right)}\,\beta_n(x)W_m(1)
\end{aligned}$$

Notons qu'une application de (3.26) nous assure l'existence d'une constante $C > 0$ telle que, pour tout n assez grand,

$$\sup_{x\in S(F)} |\beta_n(x) - B_{F,n}(F(x))| \leq C\log n/\sqrt{n}, \quad \text{p.s.}$$

Par ailleurs, selon le Lemme 2.1 de Beirlant & Deheuvels (1990),

$$\lim_{h\downarrow 0}\sup_{a<u<b;|v-u|\leq h} |W(u)-W(v)|/\{2h\log(1/h)\}^{1/2} = 1, \quad \text{p.s.}$$

où $-\infty < a < b < \infty$ et W est un processus de Wiener standard étendu sur $\mathbb{R}$. Donc, en combinant ces deux résultats avec le fait que les suites $\{W_m(u), B_{F,n}(u), u \in \mathbb{R}\}$ sont égales en distribution à $\{W(u), B(u), u \in \mathbb{R}\}$ avec W un processus de Wiener étendu et B un pont brownien étendu et indépendant de W, on aboutit à

$$\sup_{x\in S(F)} \left|W\left(F(x)+n^{-1/2}\beta_n(x)\right) - W\left(F(x)+n^{-1/2}B(F(x))\right)\right| = \mathcal{O}\left(\log n/\sqrt{n}\right), \quad \text{p.s.}$$

Finalement, en utilisant le fait que $0 < g(x) < \infty$ pour tout x, on en déduit que

$$R_{21,1}(x) = \mathcal{O}_P\left(\log n/\sqrt{n+m}\right) \tag{3.36}$$

Quant au terme $R_{21,2}(x)$, notons que l'hypothèse $0 < g(x) < \infty$ permet d'exhiber une constante $C > 0$ telle que

$$\sup_{x\in S(F)} |R_{21,2}(x)| \leq C\sqrt{\frac{n}{n+m}}\sup_{x\in S(F)} \left|W_m\left(F(x)+n^{-1/2}B_{F,n}(F(x))\right) - W_m(F(x))\right|. \tag{3.37}$$

Par ailleurs, selon le Lemme 2.2 de Beirlant & Deheuvels (1990), l'égalité suivante est valide avec probabilité 1 pour toute fonction ψ continue sur $[0,1]$,

$$\lim_{T\to\infty} T^{1/4}(\log T)^{-1/2} \sup_{0<t<1} |W(t+T^{-1/2}\psi(t)) - W(t)| = \sup_{0<t<1} |\psi(t)|^{1/2}.$$

Ainsi, un raisonnement analogue à celui utilisé pour étudier $R_{21,1}$ permet d'écrire

$$\lim_{n\to\infty} \left(n^{1/4}(\log n)^{-1/2} \sup_{0<t<1} \left| W\left(t+n^{-1/2}B(t)\right) - W(t)\right| - \sup_{0<t<1} |B(t)|^{1/2} \right) = 0 \quad \text{p.s..}$$

Ce qui entraine que

$$\sup_{x\in S(F)} \left| W_m\left(F(x)+n^{-1/2}B_{F,n}(F(x))\right) - W_m(F(x))\right| = \\ n^{-1/4}(\log n)^{1/2} \left\{ o_P(1) + \sup_{0<t<1} |B_{F,n}(F(x))|^{1/2} \right\}.$$

Pour conclure, il suffit d'utiliser (3.37) et (3.34) pour obtenir

$$\sup_{x\in S(F)} |R_{21,2}(x)| = \mathcal{O}_P\left(\sqrt{\frac{n}{n+m}} n^{-1/4}(\log n)^{3/4} \right). \tag{3.38}$$

Le terme restant $R_{21,3}(x)$ est traité de façon similaire. En effet, il est aisé de vérifier que

$$\sum_m P\left(W_m(1) \geq 2\sqrt{\log m} \right) < \infty.$$

Donc, une application du lemme de Borel-Cantelli permet d'écrire $W_m(1) \overset{\text{p.s.}}{=} (\log m)^{1/2}$. Ce résultat combiné avec la loi du logarithme itéré de Chung (1949) appliquée à $\beta_n(x)$, i.e.

$$\sup_{x\in S(F)} |\beta_n(x)| \overset{\text{p.s.}}{=} \mathcal{O}(\log\log n)^{1/2}$$

impliquent que

$$\begin{aligned} \sup_{x\in S(F)} |R_{21,3}(x)| &\overset{\text{p.s.}}{=} \sqrt{\frac{1}{n+m}} \mathcal{O}\left(\sup_{x\in S(F)} |\beta_n(x)||W_m(1)| \right) \\ &\overset{\text{p.s.}}{=} \mathcal{O}\left(\sqrt{\frac{1}{n+m}} (\log\log n)^{1/2} (\log m)^{1/2} \right). \end{aligned} \tag{3.39}$$

En résumé, en regroupant les approximations (3.31), (3.35), (3.36), (3.38) et (3.39), nous obtenons le résultat. □

Chapitre 4

Convergence uniforme presque sûre des estimateurs de la fonction d'égalisation équipercentile

Ce chapitre est publié en version réduite sous la référence El Fassi et al. *(2009) : Ajustement polynomial local de la fonction d'égalisation équipercentile : Convergence uniforme presque sûre. C.R. Math. Acad. Sci. Paris, Ser. I 347 pages 195-200 (2009).*

Résumé. Soient X et Y deux variables aléatoires de fonctions de répartition F et G respectivement. La résolution de l'équation équipercentile permet d'exprimer l'équivalent équipercentile de x comme suit : $y(x) = G^{-1}(F(x))$, où G^{-1} désigne l'inverse de la fonction G. Dans ce chapitre, nous établissons la convergence uniforme presque sûre des estimateurs de la fonction d'égalisation équipercentile $G^{-1} \circ F$, obtenus par l'approche d'ajustement polynomial local.

4.1 Convergence uniforme presque sûre

4.1.1 Convergence uniforme presque sûre de l'estimateur local polynomial d'une fonctionnelle

Nous rappelons que l'estimateur local polynomial Φ_{nr} d'une fonctionnelle Φ est donné par (2.9),

$$\Phi_{nr}(u) = \int_a^b \frac{1}{h} K_r\left(\frac{z-u}{h}\right) \Phi_n(z)\mathrm{d}z$$

$$= \begin{cases} \displaystyle\int_{-\alpha}^{1} K_r^L(z)\Phi_n(u+hz)\mathrm{d}z, & \text{si } u = a+\alpha h \text{ avec } \alpha \in (0,1]\,; \\ \displaystyle\int_{-1}^{1} K_r^I(z)\Phi_n(u+hz)\mathrm{d}z, & \text{si } u \in I\,; \\ \displaystyle\int_{-1}^{\alpha} K_r^R(z)\Phi_n(u+hz)\mathrm{d}z, & \text{si } u = b-\alpha h \text{ avec } \alpha \in (0,1]. \end{cases}$$

Dans le lemme suivant, nous prouvons que des convergences similaires que celles de Glivenko-Cantelli (voir e.g. Csörgő (1983)) pour les estimateurs lissés de F_n et G_m^{-1}. Nous énoncerons le résultat en termes des fonctions génériques Φ et Φ_n.

Lemme 4.1.1. *Soit Φ une fonction continue sur $[a,b]$. Supposons que Φ est deux fois dérivable satisfaisant $\inf_{a\le u\le b}\Phi'(u) > 0$ et $\sup_{a\le u\le b}|\Phi''(u)| < \infty$. supposons que pour tout $x \in [a,b]$, l'estimateur empirique Φ_n satisfait*

$$\sup_{x\in[a,b]} |\Phi_n(x) - \Phi(x)| \xrightarrow{p.s.} 0, \quad n \to \infty. \tag{4.1}$$

Alors,

$$\sup_{x\in[a,b]} |\Phi_{nr}(x) - \Phi(x)| \xrightarrow{p.s.} 0, \quad n \to \infty. \tag{4.2}$$

Remarque 4.1.1. En prenant $\Phi(x) = F(x)$, $\Phi_n(x) = F_n(x)$ et $\Phi_{nr}(x) = \widetilde{F}_n(x)$, pour tout $x \in [x_1, x_N]$, nous obtenons

$$\sup_{x\in S(F)} \left|\widetilde{F}_n(x) - F(x)\right| \xrightarrow{\text{p.s.}} 0. \tag{4.3}$$

De même, dans le cas où $\Phi(x) = G^{-1}(p)$, $\Phi_n(p) = G_m^{-1}(p)$ et $\Phi_{nr}(p) = \widetilde{G^{-1}}_m(p)$, pour tout $p \in [0,1]$, nous avons

$$\sup_{p\in[0,1]} \left| \widetilde{G_m}^{-1}(p) - G^{-1}(p) \right| \xrightarrow{\text{p.s.}} 0. \tag{4.4}$$

4.1.2 Convergence uniforme presque sûre des estimateurs de la fonction d'égalisation équipercentile

Le théorème suivant présente les conditions suffisantes pour la convergence uniforme presque sûre de l'estimateur $y_{n,m}^{[i]}(x)$ de la fonction d'égalisation équipercentile, avec $y_{n,m}^{[i]}(x) = \varphi_m\left(\psi_n(x)\right)$, $i = 1, \ldots, 5$ où $\varphi_m(\cdot)$ et $\psi_n(\cdot)$ représentent respectivement les estimateurs empiriques ou lissés de la fonction quantile et la fonction de répartition.

Théorème 4.1.1. *Soient X et Y deux variables aléatoires de fonctions de répartition respectivement F et G. Désignons leur supports par $S(F) = [x_1, x_N]$ et $S(G) = [y_1, y_M]$. Supposons que G deux fois dérivable et satisfait*

$$\inf_{0\le u\le 1} g(G^{-1}(u)) > 0 \quad et \quad \sup_{0\le u\le 1} |g'(G^{-1}(u))| < \infty,$$

avec $G' = g$. Alors,

$$\sup_{x\in S(F)} \left| y_{m,n}^{[i]}(x) - G^{-1}(F(x)) \right| \xrightarrow{p.s.} 0, \quad i = 1, \ldots, 5, \quad quand \quad n \wedge m \to \infty.$$

4.2 Preuves

Preuve du Lemme 4.1.1. Notons d'abord que

$$\begin{aligned} |\Phi_{nr}(x) - \Phi(x)| \le & \sup_{z\in[a,b]} |\Phi_n(z) - \Phi(z)| \int_{(a-x)/h}^{(b-x)/h} |K_r(z)| \mathrm{d}z \\ & + \left| \int_a^b \frac{1}{h} K_r\left(\frac{x-z}{h}\right) \left[\Phi(z) - \Phi(x)\right] \mathrm{d}z \right|. \end{aligned}$$

Observons que la quantité $|\int_{(a-x)/h}^{(b-x)/h} |K_r(z)|\mathrm{d}z$ est bornée car

$$\int_{(a-x)/h}^{(b-x)/h} |K_r(z)|\mathrm{d}z = \mathbb{I}(x = a + \alpha h) \int_{-\alpha}^{1} |K_r^L(z)|\mathrm{d}z + \\ \mathbb{I}(a + h < x < b - h) \int_{-1}^{1} |K_r^I(z)|\mathrm{d}z + \mathbb{I}(x = b - \alpha h) \int_{-1}^{\alpha} |K_r^R(z)|\mathrm{d}z,$$

avec $0 \leq \alpha \leq 1$. Ainsi pour conclure, il suffit d'utiliser le lemme 2.2.2 et la convergence uniforme presque sûre $\sup_{z\in[a,b]} |\Phi_n(z) - \Phi(z)|$. □

Preuve du Théorème 4.1.1. Les quatre premiers estimateurs $y_{n,m}^{[i]}(x)$, $i = 1, \ldots, 4$ pourraient être combinés en une seule forme $y_{m,n}^{[i]}(x) = \varphi_m(\psi_n(x))$ où φ_m joue le rôle de G_m^{-1} ou de $\widetilde{G_m^{-1}}$, tandis que ψ_n représente soit F_n soit $\widetilde{F_n}$. Dans ces cas de figures, nous pouvons écrire pour $x \in S(F)$,

$$\begin{aligned} \left|y_{m,n}^{[i]}(x) - G^{-1}(F(x))\right| &= \left|\varphi_m(\psi_n(x)) - G^{-1}(F(x))\right| \\ &\leq \left|\varphi_m(\psi_n(x)) - G^{-1}(\psi_n(x))\right| + \left|G^{-1}(\psi_n(x)) - G^{-1}(F(x))\right| \\ &:= \Delta_1 + \Delta_2. \end{aligned}$$

Le premier terme Δ_1 est tel que

$$\Delta_1 \leq \sup_{x\in S(F)} \left|\varphi_m(\psi_n(x)) - G^{-1}(\psi_n(x))\right| \leq \sup_{u\in[0,1]} \left|\varphi_m(u) - G^{-1}(u)\right|.$$

Ainsi, sa convergence uniforme presque sûre dépend de la convergence uniforme presque sûre de $\varphi_m(\cdot)$ obtenue dans le lemme 4.1.1 et/ou du résultat classique

$$\sup_{u\in[0,1]} \left|G_m^{-1}(u) - G^{-1}(u)\right| \xrightarrow{\text{p.s.}} 0, \quad m \to \infty$$

qui reste valable en supposant les hypothèses indiquées ci-dessus sur G, (voir e.g. le corollaire 1.4.1 dans Csörgő (1983)). Par ailleurs, en utilisant le lemme 4.1.1 si $\psi_n(\cdot) = \widetilde{F_n}$ ou le théorème de Glivenko-Cantelli si $\psi_n(\cdot) = F_n$, nous notons que

$$\sup_{x\in S(F)} \left|\psi_n(x) - F(x)\right| \xrightarrow{\text{p.s.}} 0.$$

Cette convergence uniforme presque sûre combinée avec la continuité uniforme de G^{-1} sur $[0,1]$ entraîne que le second terme Δ_2 satisfait

$$\Delta_2 \leq \sup_{x\in S(F)} \left|G^{-1}(\psi_n(x)) - G^{-1}(F(x))\right| \xrightarrow{\text{p.s.}} 0, \quad n \to \infty$$

Ce qui permet de conclure quant à la preuve du théorème (4.1.1) pour les quatre estimateurs $y_{m,n}^{[i]}(x)$, $i = 1, \ldots, 4$. Il reste à traiter le cas de $y_{m,n}^{[5]}(x)$. En effet, nous avons

$$\begin{aligned}
\left|y_{m,n}^{[5]} - G^{-1}\circ F(x)\right| &\leq \left|\int_{x_1}^{x_N} \frac{1}{h} K_r\left(\frac{z-x}{h}\right)\left[G_m^{-1}\circ F_n(z) - G^{-1}\circ F(z)\right]\mathrm{d}z\right| \\
&\quad + \left|\int_{x_1}^{x_N} \frac{1}{h} K_r\left(\frac{z-x}{h}\right)\left[G^{-1}\circ F(z) - G^{-1}\circ F(x)\right]\mathrm{d}z\right| \\
&\leq \sup_{x_1\leq z\leq x_N}\left|G_m^{-1}\circ F_n(z) - G^{-1}\circ F(z)\right| \int_{(x_1-x)/h}^{(x_N-x)/h} |K_r(z)|\mathrm{d}z \\
&\quad + \left|\int_{x_1}^{x_N} \frac{1}{h} K_r\left(\frac{z-x}{h}\right) G^{-1}\circ F(z)\mathrm{d}z - G^{-1}\circ F(x)\right|
\end{aligned}$$

La conclusion suit de la convergence uniforme presque sûre du processus Q-Q, $G_m^{-1}\circ F_n(\cdot)$ et du lemme 2.2.2.

□

Chapitre 5

Approximation par un pont brownien des estimateurs de la fonction d'égalisation équipercentile

Résumé. Ce chapitre est consacré à l'approximation par un pont brownien des processus construits par les divers estimateurs polynômiaux locaux de la fonction d'égalisation équipercentile. Nous obtenons nos résultats en faisant usage d'approximations pondérées du processus empirique et quantile par des ponts browniens appropriés.

5.1 Introduction et Notations

Soient deux échantillons de variables aléatoires $X_1, \ldots, X_n$ et $Y_1, \ldots, Y_m$ de fonctions de répartition F et G respectivement. Nous considérons deux suites de ponts browniens $\{B_{F,n}(t), 0 \leq t \leq 1\}$ et $\{\widetilde{B}_{G,m}(s), 0 \leq s \leq 1\}$, définies sur le même espace de probabilité que les suites originales des variables aléatoires (voir e.g.Komlós *et al.* (1975) et Csörgő et Révész (1978)), telles que

(i) $\left\{B_{F,n} : n \geq 1\right\}$ et $\left\{\widetilde{B}_{G,m} : m \geq 1\right\}$ sont indépendantes ;

(ii) quand $n \to \infty$ et $m \to \infty$ nous avons presque sûrement,

$$\sup_{S(F)} \left|\sqrt{n}\left[F_n(x) - F(x)\right] - B_{F,n}(x)\right| = \mathcal{O}\left(\frac{\log n}{\sqrt{n}}\right) \tag{5.1}$$

et

$$\sup_{0\leq t\leq 1} \left|\sqrt{m}g\left(G^{-1}(t)\right)\left[G_m^{-1}(t) - G^{-1}(t)\right] - \widetilde{B}_{G,m}(t)\right| = \mathcal{O}\left(\frac{\log m}{\sqrt{m}}\right), \tag{5.2}$$

où g est la dérivée de la fonction G.

5.2 Approximation par un pont brownien

Dans un premier temps, nous établissons la vitesse uniforme de l'approximation de l'estimateur général polynomial local d'une fonctionnelle $\Phi_{nr}(\cdot)$ par un pont brownien.

5.2.1 Approximation par un pont brownien de l'estimateur polynomial local d'une fonctionnelle

Nous rappelons que l'estimateur polynomial local d'une fonctionnelle $\Phi_{nr}(\cdot)$ est donné par (2.9), i.e.,

$$\Phi_{nr}(x) = \int_a^b \frac{1}{h} K_r\left(\frac{z-x}{h}\right)\Phi_n(z)\mathrm{d}z$$

$$= \begin{cases} \displaystyle\int_{-\alpha}^{1} K_r^L(z)\Phi_n(x+hz)\mathrm{d}z, & \text{si } x = a + \alpha h \text{ avec } \alpha \in (0,1]\,; \\ \displaystyle\int_{-1}^{1} K_r^I(z)\Phi_n(x+hz)\mathrm{d}z, & \text{si } x \in (a+h, b-h)\,; \\ \displaystyle\int_{-1}^{\alpha} K_r^R(z)\Phi_n(x+hz)\mathrm{d}z, & \text{si } x = b - \alpha h \text{ avec } \alpha \in (0,1], \end{cases}$$

où Φ_n étant l'estimateur empirique de Φ, le noyau K_r est défini par (2.7), (a, b) est le support de Φ et h est le paramètre de la fenêtre.

Lemme 5.2.1. *Soit $\omega(\cdot)$ une fonction arbitraire définie sur $[a,b]$ telle que $0 < M_1 \leq |\omega(x)| \leq M_2 < \infty$ pour tout $x \in [a,b]$ admettant des dérivées jusqu'à l'ordre 4. Soit $\gamma(\cdot)$ une fonction ayant une dérivée bornée et satisfaisant,*

$$0 \leq \gamma(x) \leq 1, \text{ pour tout } x \in [a,b].$$

Supposons qu'il existe un espace de probabilité sur lequel nous définissons une suite de ponts browniens $\{B_n^*(y), 0 \leq y \leq 1\}$ telles que

$$\sup_{x\in[a,b]} \left|\sqrt{n}\omega(x)\left[\Phi_n(x) - \Phi(x)\right] - B_n^*(\gamma(x))\right| \overset{p.s.}{=} \mathcal{O}\left(\frac{\log n}{\sqrt{n}}\right). \tag{5.3}$$

Alors nous avons

$$\sup_{x\in[a,b]} \left|\sqrt{n}\omega(x)\left[\Phi_{nr}(x) - \Phi(x)\right] - B_n^*(\gamma(x))\right| = \mathcal{O}_P\left(\frac{\log n}{\sqrt{n}} + \left(h\log\frac{1}{h}\right)^{1/2} + \sqrt{n}h^{r+1}\right). \tag{5.4}$$

Remarque 5.2.1. Une application du lemme 5.2.1 permet de voir qu'en utilisant (5.1) et en posant $\omega(x) = 1$, $\gamma(x) = F(x)$, $\Phi(x) = F(x)$, $\Phi_n(x) = F_n(x)$, $\Phi_{nr}(x) = \widetilde{F}_n(x)$ et $B_n^*(F(x)) = B_{F,n}(F(x))$ pour $x \in [a,b] = [x_1, x_N]$, nous obtenons

$$\sup_{x\in S(F)} \left|\sqrt{n}\left[\widetilde{F}_n(x) - F(x)\right] - B_{F,n}(F(x))\right| = \mathcal{O}_P\left(\frac{\log n}{\sqrt{n}} + \left(h\log\frac{1}{h}\right)^{1/2} + \sqrt{n}h^{r+1}\right). \tag{5.5}$$

De façon similaire, en raison de (5.2), si nous remplaçons n par m et nous posons $\omega(p) = g(G^{-1}(p))$, $\gamma(p) = p$, $\Phi(p) = G^{-1}(p)$, $\Phi_m(x) = G_m^{-1}(p)$, $\Phi_{mr}(p) = \widetilde{G_m^{-1}}(p)$ et $B_m^*(p) = \widetilde{B}_{G,m}(p)$ pour $p \in [a,b] = [0,1]$, nous avons

$$\sup_{p\in[0,1]} \left|\sqrt{m}g(G^{-1}(p))[\widetilde{G_m^{-1}}(p) - G^{-1}(p)] - \widetilde{B}_{G,m}(p)\right| = \mathcal{O}_P\left(\frac{\log m}{\sqrt{m}} + \left(k\log\frac{1}{k}\right)^{1/2} + \sqrt{m}k^{r+1}\right), \tag{5.6}$$

où k est le paramètre de lissage de $\widetilde{G_m^{-1}}$.

5.2.2 Approximation par un pont brownien des estimateurs polynômiaux locaux de la fonction d'égalisation équipercentile

Nous considérons les vitesses uniformes des approximations de $\left\{y_{m,n}^{[i]}(x) - G^{-1}(F(x))\right\}$ par un pont Brownien sur $[0,1]$, qui est un processus gaussien à valeurs réelles de moyenne nulle et de fonction de covariance $s(1-t)$, avec $0 \leq s \leq t \leq 1$. Nos résultats sont fondés sur les approximations classiques des processus empirique $\left\{F_n(\cdot) - F(\cdot)\right\}$ et quantile $\left\{G_m^{-1}(\cdot) - G^{-1}(\cdot)\right\}$ par des ponts browniens appropriés (voir les équations (5.1) et (5.2)). Dans la suite, nous supposons que la quantité $\dfrac{n}{n+m}$ converge vers une constante dans $(0,1)$ quand $n+m \to +\infty$. Nous définissons les processus construits à partir des estimateurs polynômiaux locaux de la fonction d'égalisation équipercentile par

$$\Delta_{m,n}^{[i]}(x) = \sqrt{\frac{nm}{n+m}} g\left(G^{-1}(F(x))\right)\left[y_{m,n}^{[i]} - G^{-1}(F(x))\right], \quad i = 1, \ldots, 5. \tag{5.7}$$

Le but principal est d'étudier les approximations des processus $\Delta_{m,n}^{[i]}(x)$ par un pont brownien $B_{m,n}(F(x))$. Nous étudierons les quantités suivantes :

$$\sup_{x \in S(F)} \left|\Delta_{m,n}^{[i]}(x) - B_{m,n}(F(x))\right|,$$

avec

$$B_{m,n}(F(x)) = \sqrt{\frac{nm}{n+m}}\left(m^{-1/2}\widetilde{B}_{G,m}(F(x)) + n^{-1/2}B_{F,n}(F(x))\right), \quad x \in S(F) = [x_1, x_N].$$

Le comportement asymptotique du processus Quantile-Quantile a été étudié par plusieurs auteurs (voir, e.g., Doksum (1974), Hollander et Korwar (1982), Aly (1986), Beirlant et Deheuvels (1990), Lu *et al.* (1994)).

Dans le théorème suivant, nous prolongeons ces approximations aux estimateurs $y_{m,n}^{[i]}(x)$, $i = 1, \ldots, 5$. Les quatre premiers estimateurs $y_{n,m}^{[i]}(x)$, $i = 1, \ldots, 4$ pourraient être combinés en une seule forme $y_{m,n}^{[i]}(x) = \varphi_m\left(\psi_n(x)\right)$ où φ_m joue le rôle de G_m^{-1} ou de $\widetilde{G_m^{-1}}$, tandis que ψ_n représente soit F_n soit $\widetilde{F_n}$.

Théorème 5.2.1. *Supposons que $K(\cdot)$ est une densité de probabilité symétrique et bornée à support $[-1,1]$. Soit K_r le noyau associé obtenu par (2.7) avec $r = 0, 1$ ou 2. Supposons que la foncion $G^{-1}(F(\cdot))$ admet des dérivées bornées jusqu'à l'ordre $2r$ et que*

$$\inf_{0\leq u\leq 1} g\left(G^{-1}(u)\right) > 0, \quad \sup_{0\leq u\leq 1} |g'\left(G^{-1}(u)\right)| < \infty$$

où $G' = g$. Supposons également que les deux fenêtres $h = h(n)$ et $k = k(m)$ tendent vers 0 quand $n, m \to \infty$. Alors, il existe deux suites de ponts browniens indépendantes $B_{F,n}$ et $\widetilde{B}_{G,m}$ définies sur le même espace de probabilité associé aux variables aléatoires $X_1, \ldots, X_n$ et $Y_1, \ldots, Y_m$ telles que

$$\begin{aligned}\sup_{x\in S(F)} \left|\Delta^{[i]}_{m,n}(x) - B_{m,n}(F(x))\right| = \\ \sqrt{\frac{nm}{n+m}}\left\{\mathcal{O}_P\left(c_m\right) + \mathcal{O}_P(d_n) + \mathcal{O}_P\left(\frac{(\log n)^{3/4}}{n^{1/4}m^{1/2}}\right) + \mathcal{O}_P\left(\frac{1}{\sqrt{m}}\sqrt{d_n \log\frac{1}{d_n}}\right)\right. \\ \left. + \mathcal{O}_P\left(\left[2\sqrt{\frac{\log m}{m}} + \sqrt{\frac{\log n}{n}}\right]\left[d_n + \sqrt{\frac{\log n}{n}}\right]\right)\right\}\end{aligned} \tag{5.8}$$

pour $i = 1, \ldots, 4$ avec

$$c_m = \begin{cases} \log m/m, & si\ \varphi_m = G_m^{-1} \\ \log m/m + m^{-1/2}\left(k\log\dfrac{1}{k}\right)^{1/2} + k^{r+1}, & si\ \varphi_m = \widetilde{G_m^{-1}} \end{cases}$$

et

$$d_n = \begin{cases} \log n/n, & si\ \psi_n = F_n \\ \log n/n + n^{-1/2}\left(h\log\dfrac{1}{h}\right)^{1/2} + h^{r+1}, & si\ \psi_n = \widetilde{F}_n. \end{cases}$$

Quant à $\Delta^{[5]}_{n,m}$, nous avons

$$\begin{aligned}\sup_{x\in S(F)} \left|\Delta^{[5]}_{m,n}(x) - B_{m,n}(F(x))\right| = \\ \sqrt{\frac{nm}{n+m}}\mathcal{O}_P\left(h^{r+1} + (n^{-1/2} + m^{-1/2})\left(h\log\frac{1}{h}\right)^{1/2} + \frac{\log n}{n} + \frac{\log m}{m} + \frac{(\log n)^{3/4}}{n^{1/4}m^{1/2}}\right)\end{aligned} \tag{5.9}$$

Remarque 5.2.2. Une application du théorème 5.2.1 permet de voir que,

(i) en prenant $\varphi_m = G_m^{-1}$ et $\psi_n = F_n$, nous obtenons l'approximation par un pont brownien du processus Quantile-Quantile, i.e.,

$$\sup_{x\in S(F)} \left|\Delta_{m,n}^{[1]}(x) - B_{m,n}(F(x))\right| = \sqrt{\frac{nm}{n+m}}\mathcal{O}_P\left(\frac{\log n}{n} + \frac{\log m}{m} + \frac{(\log n)^{3/4}}{n^{1/4}m^{1/2}}\right);$$

(ii) en remplaçant φ_m par G_m^{-1} et ψ_n par $\widetilde{F_n}$, nous obtenons,

$$\sup_{x\in S(F)} \left|\Delta_{m,n}^{[2]}(x) - B_{m,n}(F(x))\right| =$$
$$\sqrt{\frac{nm}{n+m}}\mathcal{O}_P\left(h^{r+1} + n^{-1/2}\left(h\log\frac{1}{h}\right)^{1/2} + \frac{\log n}{n} + \frac{\log m}{m} + \frac{(\log n)^{3/4}}{n^{1/4}m^{1/2}}\right);$$

(iii) lorsque $\varphi_m = \widetilde{G_m^{-1}}$ et $\psi_n = F_n$, nous obtenons l'approximation suivante

$$\sup_{x\in S(F)} \left|\Delta_{m,n}^{[3]}(x) - B_{m,n}(F(x))\right| =$$
$$\sqrt{\frac{nm}{n+m}}\mathcal{O}_P\left(k^{r+1} + m^{-1/2}\left(k\log\frac{1}{k}\right)^{1/2} + \frac{\log n}{n} + \frac{\log m}{m} + \frac{(\log n)^{3/4}}{n^{1/4}m^{1/2}}\right);$$

(iv) en posant $\varphi_m = \widetilde{G_m^{-1}}$ et $\psi_n = \widetilde{F_n}$, nous avons

$$\sup_{x\in S(F)} \left|\Delta_{m,n}^{[4]}(x) - B_{m,n}(F(x))\right| =$$
$$\sqrt{\frac{nm}{n+m}}\mathcal{O}_P\left(h^{r+1} + k^{r+1} + n^{-1/2}\left(h\log\frac{1}{h}\right)^{1/2} + m^{-1/2}\left(k\log\frac{1}{k}\right)^{1/2} + \frac{\log n}{n} + \frac{\log m}{m} + \frac{(\log n)^{3/4}}{n^{1/4}m^{1/2}}\right);$$

(v) et finalement, si l'estimateur $y_{n,m}^{[5]}(x)$ est remplacé par $\widetilde{G_m^{-1}(F_n(x))}$, nous avons

$$\sup_{x\in S(F)} \left|\Delta_{m,n}^{[5]}(x) - B_{m,n}(F(x))\right| =$$
$$\sqrt{\frac{nm}{n+m}}\mathcal{O}_P\left(h^{r+1} + (n^{-1/2} + m^{-1/2})\left(h\log\frac{1}{h}\right)^{1/2} + \frac{\log n}{n} + \frac{\log m}{m} + \frac{(\log n)^{3/4}}{n^{1/4}m^{1/2}}\right).$$

5.3 Preuves

Preuve du lemme 5.2.1. Pour tout $x \in [a,b]$, nous posons

$$\rho_n(x) = \sqrt{n}\omega(x)\left[\Phi_n(x) - \Phi(x)\right].$$

Nous avons

$$\begin{aligned}
&\sqrt{n}\omega(x)\left[\Phi_{nr}(x) - \Phi(x)\right] - B_n^*(\gamma(x)) \\
&\qquad = \omega(x)\int_a^b \frac{1}{h}K_r\left(\frac{y-x}{h}\right)\frac{1}{\omega(y)}\left[\rho_n(y) - B_n^*(\gamma(y))\right]\mathrm{d}y \\
&\qquad\quad +\omega(x)\int_a^b \frac{1}{h}K_r\left(\frac{y-x}{h}\right)\frac{1}{\omega(y)}\left[B_n^*(\gamma(y)) - B_n^*(\gamma(x))\right]\mathrm{d}y \\
&\qquad\quad +\int_a^b \frac{1}{h}K_r\left(\frac{y-x}{h}\right)B_n^*(\gamma(x))\left[\frac{\omega(x)}{\omega(y)} - 1\right]\mathrm{d}y \\
&\qquad\quad +\sqrt{n}\omega(x)\int_a^b \frac{1}{h}K_r\left(\frac{y-x}{h}\right)\left[\Phi(y) - \Phi(x)\right]\mathrm{d}y \\
&\qquad = R_1(x) + R_2(x) + R_3(x) + R_4(x).
\end{aligned}$$

Puisque $0 < M_1 \leq \omega(x) \leq M_2 < \infty$ et $\displaystyle\int_{(a-x)/h}^{(b-x)/h} |K_r(z)|\,\mathrm{d}z < \infty$ pour tout $x \in [a,b]$, en utilisant (5.3), nous avons

$$\begin{aligned}
\sup_{x\in[a,b]}|R_1(x)| &\leq \sup_{x\in[a,b]}\int_{(a-x)/h}^{(b-x)/h}\left|K_r(z)\,\frac{\omega(x)}{\omega(x+hz)}\left[\rho_n(x+hz) - B_n^*(\gamma(x+hz))\right]\right|\mathrm{d}z \\
&\leq C\sup_{x\in[a,b]}|\rho_n(x) - B_n^*(\gamma(x))| \overset{\text{p.s.}}{=} \mathcal{O}\left(\frac{\log n}{\sqrt{n}}\right) \qquad (5.10)
\end{aligned}$$

avec $C > 0$ une constante fixée et finie. La quantité $R_2(x)$ peut être réécrite comme suit

$$\begin{aligned} R_2(x) &= \int_{(a-x)/h}^{(b-x)/h} K_r(u)\frac{\omega(x)}{\omega(x+hy)}\left[B_n^*(\gamma(x+hu)) - B_n^*(\gamma(x))\right]\mathrm{d}u \\ &= \mathbb{I}(x = a+\alpha h)\int_{-\alpha}^{1} K_r^L(u)\frac{\omega(x)}{\omega(x+hu)}\left[B_n^*(\gamma(x+hu)) - B_n^*(\gamma(x))\right]\mathrm{d}u \\ &\quad \mathbb{I}(a+h < x < b-h)\int_{-1}^{1} K_r^I(u)\frac{\omega(x)}{\omega(x+hu)}\left[B_n^*(\gamma(x+hu)) - B_n^*(\gamma(x))\right]\mathrm{d}u \\ &\quad +\mathbb{I}(x = b-\alpha h)\int_{-1}^{\alpha} K_r^R(u)\frac{\omega(x)}{\omega(x+hu)}\left[B_n^*(\gamma(x+hu)) - B_n^*(\gamma(x))\right]\mathrm{d}u. \\ &\leq C\left[\sup_{(u,x)\in\Omega} + \sup_{(u,x)\in\Omega^c}\right]\left|B_n^*(\gamma(x+hu)) - B_n^*(\gamma(x))\right| \end{aligned}$$

où $C > 0$ est une constante finie et l'ensemble Ω est donné par

$$\Omega = \{(u,x) \in [-1,1]\times[a,b] : |\gamma(x+hu) - \gamma(x)| \leq \epsilon\}$$

avec $\epsilon > 0$ étant un paramètre arbitraire. D'autre part, puisque γ est dérivable sur $[a,b]$, nous pouvons écrire pour tout $(u,x) \in [-1,1]\times[a,b]$

$$\gamma(x+hu) = \gamma(x) + hu\gamma'(\xi), \quad \text{où } x\bigwedge(x+hu) \leq \xi \leq x\bigvee(x+hu)$$

Par conséquent, si nous prenons $\epsilon := \epsilon_h = h\sup_{x\in[a,b]}|\gamma'(x)|$, alors $\Omega^c = \emptyset$ et

$$|R_2(x)| \leq C\sup_{\substack{u,v\in[0,1]\\|u-v|\leq\epsilon_h}}\left|B_n^*(u) - B_n^*(v)\right|.$$

Désignons par B un pont brownien étendu sur $\mathbb{R}$ dont la restriction sur $[0,1]$ coincide avec B_n^*. Une application du Théorème 1.5.2 dans Csörgő (1983) permet d'avoir pour un h suffisamment petit

$$\sup_{x\in[a,b]}|R_2(x)| \leq C\sup_{|x-y|\leq\epsilon_h}|B(x) - B(y)| = \mathcal{O}_P\left(\epsilon_h\log 1/\epsilon_h\right)^{1/2} = \mathcal{O}_P\left(h\log\frac{1}{h}\right)^{1/2}. \quad (5.11)$$

Réécrivons maintenant $R_3(x)$ comme suit

$$R_3(x) = B_n^*(\gamma(x))\omega(x)\int_a^b \frac{1}{h}K_r\left(\frac{y-p}{h}\right)\left[\frac{1}{\omega(y)} - \frac{1}{\omega(x)}\right]\mathrm{d}y,$$

et remplaçons la fonction $\Phi(\cdot)$ dans le lemme 2.2.3 par la fonction $1/\omega(\cdot)$ pour aboutir à

$$\sup_{x\in[a,b]} |R_3(x)| \;\leq\; Ch^{r+1} \sup_{0\leq t\leq 1} |B_n^*(t)| \overset{\text{p.s.}}{=} \mathcal{O}\left((\log n)^{1/2} h^{r+1}\right) \tag{5.12}$$

où $C > 0$ est une constante fixée et finie. Afin d'obtenir la dernière égalité, rappelons que pour toute suite de ponts browniens $\{B_n^*(t), 0 \leq t \leq 1\}$, nous avons

$$P\left(\sup_{0\leq t\leq 1} |B_n^*(t)| > \varepsilon\right) \leq 2\exp\left(-2\varepsilon^2\right), \quad \varepsilon \geq 0$$

(voir, e.g. Csörgő et Révész (1981)). Alors, en utilisant le lemme de Borel-Cantelli et en posant $\varepsilon = (\log n)^{1/2}$, on obtient

$$\sup_{0\leq t\leq 1} |B_n^*(t)| \overset{\text{p.s.}}{=} \mathcal{O}\left((\log n)^{1/2}\right). \tag{5.13}$$

Finalement, quant à la quantité $R_4(x)$, notons que

$$\begin{aligned} R_4(x) &= \sqrt{n}\omega(x) \int_a^b \frac{1}{h} K_r\left(\frac{y-x}{h}\right) [\Phi(y) - \Phi(x)]\, \mathrm{d}y \\ &= \sqrt{n}\omega(x) \Bigg\{ \mathbb{I}(x = a + \alpha h) \int_{-\alpha}^{1} K_r^L(u) \left[\Phi(x+hu) - \Phi(x)\right] \mathrm{d}u \\ &\qquad + \mathbb{I}(a + h < x < b - h) \int_{-1}^{1} K_r^I(u) \left[\Phi(x+hu) - \Phi(x)\right] \mathrm{d}u \\ &\qquad + \mathbb{I}(x = b - \alpha h) \int_{-1}^{\alpha} K_r^R(u) \left[\Phi(x+hu) - \Phi(x)\right] \mathrm{d}u \Bigg\}. \end{aligned}$$

Une application du lemme 2.2.3 entraîne que

$$\sup_{x\in[a,b]} |R_4(x)| = \mathcal{O}\left(\sqrt{n} h^{r+1}\right), \quad r = 0, 1, 2. \tag{5.14}$$

Pour conclure, notons que $(\log n)^{1/2} h^{r+1} = \mathcal{O}\left(\sqrt{n} h^{r+1}\right)$ et en regroupant les expressions (5.10)–(5.14), nous obtenons le résultat (5.4). □

Preuve du théorème 5.2.1. De manière similaire que la preuve de la convergence uniforme presque sûre des $y_{n,m}^{[i]}$, pour $i = 1, \ldots, 5$, la preuve de l'approximation par un pont brownien est coupé en deux parties également. Nous fournissons dans un premier temps une preuve

unifiée pour les quatre premiers estimateurs $y_{m,n}^{[i]}(x)$, $i = 1, \ldots, 4$ et ensuite nous établissons la preuve de l'estimateur $y_{m,n}^{[5]}(x)$.

Posons $y_{m,n}^{[i]}(x) = \varphi_m(\psi_n(x))$ avec $\varphi_m = G_m^{-1}$ ou $\widetilde{G_m^{-1}}$, et $\psi_n = F_n$ ou $\widetilde{F_n}$, pour $i = 1, \ldots, 4$. Alors, en utilisant les mêmes notations que celles du Théorème 5.2.1, nous avons, pour $i = 1, \ldots, 4$,

$$\begin{aligned}\Delta_{m,n}^{[i]}(x) - B_{m,n}(F(x)) &= \sqrt{\frac{nm}{n+m}}\Big\{g\left(G^{-1}(F(x))\right)\left[\varphi_m(\psi_n(x)) - G^{-1}(F(x))\right] \\ &\qquad -m^{-1/2}\widetilde{B}_{G,m}(F(x)) - n^{-1/2}B_{F,n}(F(x))\Big\} \\ &:= T_1 + T_2\end{aligned}$$

où

$$T_1 = \sqrt{\frac{nm}{n+m}}\Big\{g\left(G^{-1}(F(x))\right)\left[\varphi_m(\psi_n(x)) - G^{-1}(\psi_n(x))\right] - m^{-1/2}\widetilde{B}_{G,m}(F(x))\Big\}$$

et

$$T_2 = \sqrt{\frac{nm}{n+m}}\Big\{g\left(G^{-1}(F(x))\right)\left[G^{-1}(\psi_n(x)) - G^{-1}(F(x))\right] - n^{-1/2}B_{F,n}(F(x))\Big\}.$$

Commençons d'abord par le traitement de T_1 et posons $T_1 := T_{11} + T_{12}$ où

$$\begin{aligned}T_{11} &= \sqrt{\frac{n}{n+m}}\frac{g\left(G^{-1}(F(x))\right)}{g\left(G^{-1}(\psi_n(x))\right)} \\ &\qquad \times\Big\{\sqrt{m}g\left(G^{-1}(\psi_n(x))\right)\left[\varphi_m(\psi_n(x)) - G^{-1}(\psi_n(x))\right] - \widetilde{B}_{G,m}(\psi_n(x))\Big\}\end{aligned}$$

et

$$T_{12} = \sqrt{\frac{n}{n+m}}\Big\{\frac{g\left(G^{-1}(F(x))\right)}{g\left(G^{-1}(\psi_n(x))\right)}\widetilde{B}_{G,m}(\psi_n(x)) - \widetilde{B}_{G,m}(F(x))\Big\}.$$

Or $G^{-1}(\cdot)$ admet des dérivées bornées et $\inf_{0\le u\le 1} g\left(G^{-1}(u)\right) > 0$. Donc, il existe deux

constantes finies $M_1, M_2 > 0$ telles que

$$\begin{aligned} |T_{11}| &\leq \frac{M_2}{M_1}\sqrt{\frac{n}{n+m}}\sup_{p\in[0,1]}\left|g\left(G^{-1}(p)\right)\sqrt{m}\left[\varphi_m(p)-G^{-1}(p)\right]-\widetilde{B}_{G,m}(p)\right| \\ &= \sqrt{\frac{nm}{n+m}}\mathcal{O}_P(c_m) \end{aligned} \tag{5.15}$$

où c_m et d_n sont donnés selon (5.2) et (5.6) par

$$c_m = \begin{cases} \log m/m, & \text{if } \varphi_m = G_m^{-1} \\ \log m/m + m^{-1/2}\left(k\log\frac{1}{k}\right)^{1/2} + k^{r+1}, & \text{if } \varphi_m = \widetilde{G_m^{-1}}. \end{cases}$$

Considérons maintenant le second terme T_{12}. Nous avons la décomposition suivante

$$\begin{aligned} T_{12} &= \sqrt{\frac{n}{n+m}}\widetilde{B}_{G,m}\left(\psi_n(x)\right)\left\{\frac{g\left(G^{-1}(F(x))\right)}{g\left(G^{-1}(\psi_n(x))\right)}-1\right\} \\ &\qquad +\sqrt{\frac{n}{n+m}}\left\{\widetilde{B}_{G,m}(\psi_n(x))-\widetilde{B}_{G,m}(F(x))\right\} \\ &:= T_{12,1}+T_{12,2}. \end{aligned}$$

Pour tout $(u, v_n) \in [0,1]\times[0,1]$, par le développement de Taylor d'ordre 1, il existe $\xi \in]u\wedge v_n, u\vee v_n[$ tel que

$$\frac{g\left(G^{-1}(u)\right)}{g\left(G^{-1}(v_n)\right)}-1 = \frac{u-v_n}{g\left(G^{-1}(v_n)\right)}\frac{g'\left(G^{-1}(\xi)\right)}{g\left(G^{-1}(\xi)\right)}$$

Par conséquent, en posant $u = F(x)$, $v_n = \psi_n(x)$ et en utilisant les hypothèses

$$\inf_{0\leq u\leq 1} g\left(G^{-1}(u)\right) > 0 \quad \text{and} \quad \sup_{0\leq u\leq 1}|g'\left(G^{-1}(u)\right)| < \infty,$$

nous avons

$$\left|\frac{g\left(G^{-1}(F(x))\right)}{g\left(G^{-1}(\psi_n(x))\right)}-1\right| \overset{\text{p.s.}}{=} \mathcal{O}(|\psi_n(x)-F(x)|). \tag{5.16}$$

En raison de (5.1) et (5.5), nous pouvons toujours écrire, pour n suffisamment grand

$$\sup_{x\in S(F)} n^{-1/2}\left|\sqrt{n}\left[\psi_n(x)-F(x)\right]-B_{F,n}(F(x))\right| = \mathcal{O}_P(d_n), \tag{5.17}$$

où

$$d_n = \begin{cases} \log n/n, & \text{si } \psi_n = F_n \\ \log n/n + n^{-1/2}\left(h\log\frac{1}{h}\right)^{1/2} + h^{r+1}, & \text{si } \psi_n = \widetilde{F}_n. \end{cases}$$

Ce résultat implique que

$$\begin{aligned} \sup_{x\in S(F)} |\psi_n(x) - F(x)| &= \mathcal{O}_P\left(d_n + \frac{1}{\sqrt{n}} \sup_{t\in[0,1]} |B_{F,n}(t)|\right) \\ &= \mathcal{O}_P\left(d_n + \sqrt{\frac{\log n}{n}}\right). \end{aligned} \tag{5.18}$$

Ainsi, par définition de $T_{12,1}$, nous avons

$$\begin{aligned} \sup_{x\in S(F)} |T_{12,1}| &= \sqrt{\frac{n}{n+m}} \sup_{x\in S(F)} \left| \widetilde{B}_{G,m}\left(\psi_n(x)\right) \left\{ \frac{g\left(G^{-1}(F(x))\right)}{g\left(G^{-1}(\psi_n(x))\right)} - 1 \right\} \right| \\ &= \sqrt{\frac{nm}{n+m}} \mathcal{O}_P\left(\sqrt{\frac{\log m}{m}}\left(d_n + \sqrt{\frac{\log n}{n}}\right)\right). \end{aligned} \tag{5.19}$$

D'autre part, pour traiter les expressions des quantités restantes, sans perte de généralités, nous supposons qu'un processus standard de Wiener est une restriction sur $\mathbb{R}^+$ d'un processus standard de Winer étendu sur $\mathbb{R}$. Dans le même sens, nous supposerons que les ponts browniens $B_{F,n}$ et $\widetilde{B}_{G,m}$ (définis dans (5.1) and (5.2) respectivement) sont des restrictions sur $[0,1]$ des ponts browniens étendus sur $\mathbb{R}$ désignés par $B'_{F,n}$ et $\widetilde{B}'_{G,m}$ respectivement. Ainsi, pour tout pont brownien étendu $B'(\cdot)$, nous avons $B'(t) = W'(t) - tW'(1)$, $t \in \mathbb{R}$ où $W'(\cdot)$ est un processus de Wiener étendu.

Pour tout $x \in [x_1, x_N]$, nous décomposons la quantité $T_{12,2}$ comme suit

$$\begin{aligned} T_{12,2} &= \sqrt{\frac{n}{n+m}} \left\{ \widetilde{B}'_{G,m}(\psi_n(x)) - \widetilde{B}'_{G,m}(F(x)) \right\} \\ &= \sqrt{\frac{n}{n+m}} \left\{ \widetilde{B}'_{G,m}\left(F(x) + n^{-1/2}\beta_n(x)\right) - \widetilde{B}'_{G,m}(F(x)) \right\} \\ &:= T_{12,2}^{[1]} + T_{12,2}^{[2]} + T_{12,2}^{[3]} \end{aligned}$$

où $\beta_n(x) = \sqrt{n}(\psi_n(x) - F(x))$ pour $x \in S(F)$ et

$$\begin{aligned} T_{12,2}^{[1]} &= \sqrt{\frac{n}{n+m}}\Big\{W'_m\left(F(x) + n^{-1/2}\beta_n(x)\right) - W'_m\left(F(x) + n^{-1/2}B_{F,n}(F(x))\right)\Big\} \\ T_{12,2}^{[2]} &= \sqrt{\frac{n}{n+m}}\Big\{W'_m\left(F(x) + n^{-1/2}B_{F,n}(F(x))\right) - W'_m(F(x))\Big\} \\ T_{12,2}^{[3]} &= -\sqrt{\frac{1}{n+m}}\beta_n(x)W'_m(1) \end{aligned}$$

avec $W'_m(\cdot)$ étant une suite de processus standards étendus de Wiener. Pour traiter le premier terme $T_{12,2}^{[1]}$, nous utilisons (5.17) combiné avec le lemme 2.1 dans Beirlant et Deheuvels (1990) (voir aussi Taylor (1974)), i.e., tout processus standard de Wiener $W'(\cdot)$ satisfait

$$\lim_{\varepsilon\downarrow 0}\sup_{a<u<b;|v-u|\leq\varepsilon}|W'(u) - W'(v)|/\{2\varepsilon\log(1/\varepsilon)\}^{1/2} = 1, \quad \text{p.s.}$$

Pour tout $-\infty < a < b < \infty$, nous obtenons

$$\sup_{x\in S(F)}|T_{12,2}^{[1]}| = \sqrt{\frac{nm}{n+m}}\mathcal{O}_P\left(m^{-1/2}\left(d_n\log\frac{1}{d_n}\right)^{1/2}\right). \tag{5.20}$$

Quant à la deuxième quantité $T_{12,2}^{[2]}$, nous utilisons le lemme 2.2 de Beirlant et Deheuvels (1990). Il dit que tout processus standard de Wiener étendu $W'(\cdot)$ satisfait avec une probabilité 1

$$\lim_{T\to\infty}T^{1/4}(\log T)^{-1/2}\sup_{0<t<1}|W'(t + T^{-1/2}\gamma(t)) - W'(t)| = \sup_{0<t<1}|\gamma(t)|^{1/2}$$

pour toute fonction continue γ sur $[0,1]$. En particulier, puisque tout pont brownien étendu $B'(\cdot)$ est continu avec une probabilité 1 nous posons $\gamma(\cdot) = B'(\cdot)$ et nous obtenons

$$\lim_{n\to\infty}\left(n^{1/4}(\log n)^{-1/2}\sup_{0<t<1}\left|W'\left(t + n^{-1/2}B'(t)\right) - W'(t)\right| - \sup_{0<t<1}|B'(t)|^{1/2}\right) = 0 \quad \text{p.s.}$$

avec $B'(\cdot)$ et $W'(\cdot)$ étant indépendants.

Par conséquent, en utilisant le fait que $\{B_{F,n}(t), W'_m\left(t + n^{-1/2}B_{F,n}(t)\right) - W'_m(t), -\infty < t < \infty\}$ et $\{B'(t), W'\left(t + n^{-1/2}B'(t)\right) - W'(t), -\infty < t < \infty\}$ sont égaux en distribution, nous

aboutissons à

$$\sup_{x\in S(F)} \left|W'_m\left(F(x)+n^{-1/2}B_{F,n}(F(x))\right)-W'_m(F(x))\right| = \\ n^{-1/4}(\log n)^{1/2}\left\{o_P(1)+\sup_{x\in S(F)}|B_{F,n}(F(x))|^{1/2}\right\}.$$

Ce résultat combiné avec (5.13) entraîne que

$$\begin{aligned}\sup_{x\in S(F)}|T^{[2]}_{12,2}| &= \sqrt{\frac{n}{n+m}}\sup_{x\in S(F)}\left|W'_m\left(F(x)+n^{-1/2}B_{F,n}(F(x))\right)-W'_m(F(x))\right| \\ &= \sqrt{\frac{nm}{n+m}}\mathcal{O}_P\left(\frac{(\log n)^{3/4}}{n^{1/4}m^{1/2}}\right). \qquad (5.21)\end{aligned}$$

Ensuite, pour obtenir le comportement asymptotique de $T^{[3]}_{12,2}$, observons que $W_m(1)$ satisfait $W_m(1) \overset{\text{p.s.}}{=} O(\log m)^{1/2}$ car

$$\sum_m P\left(W_m(1)\geq 2\sqrt{\log m}\right)<\infty.$$

Nous utilisons (5.18) et obtenons

$$\begin{aligned}\sup_{x\in S(F)}|T^{[3]}_{12,2}| &= \sqrt{\frac{1}{n+m}}\sup_{x\in S(F)}|\beta_n(x)||W_m(1)| \\ &= \sqrt{\frac{nm}{n+m}}\mathcal{O}_P\left(\sqrt{\frac{\log m}{m}}\left(d_n+\sqrt{\frac{\log n}{n}}\right)\right). \qquad (5.22)\end{aligned}$$

Finalement, le comportement asymptotique du second terme T_2 pourrait être établi de manière similaire. Par le développement de Taylor d'ordre 1, nous avons

$$\begin{aligned}T_2 &= \sqrt{\frac{nm}{n+m}}\left\{g\left(G^{-1}(F(x))\right)\left[G^{-1}(\psi_n(x))-G^{-1}(F(x))\right]-n^{-1/2}B_{F,n}(F(x))\right\} \\ &= \sqrt{\frac{m}{n+m}}\left\{\frac{g\left(G^{-1}(F(x))\right)}{g\left(G^{-1}(\theta_{x,n})\right)}\sqrt{n}\left[\psi_n(x)-F(x)\right]-B_{F,n}(F(x))\right\} \\ &:= T_{21}+T_{22}\end{aligned}$$

où $0\leq\psi_n(x)\wedge F(x)<\theta_{x,n}<\psi_n(x)\vee F(x)\leq 1$ et

$$\begin{aligned}T_{21} &= \sqrt{\frac{m}{n+m}}\frac{g\left(G^{-1}(F(x))\right)}{g\left(G^{-1}(\theta_{x,n})\right)}\left\{\sqrt{n}\left[\psi_n(x)-F(x)\right]-B_{F,n}(F(x))\right\} \\ T_{22} &= \sqrt{\frac{m}{n+m}}\left\{\frac{g\left(G^{-1}(F(x))\right)}{g\left(G^{-1}(\theta_{x,n})\right)}-1\right\}B_{F,n}(F(x)).\end{aligned}$$

Puisque $0 < g(\cdot) < \infty$, il existe deux constantes positives M_1 and M_2 telles que

$$\begin{aligned} |T_{21}| &\leq \frac{M_2}{M_1}\sqrt{\frac{m}{n+m}} \sup_{x\in S(F)} \left|\sqrt{n}\left[\psi_n(x) - F(x)\right] - B_{F,n}(x)\right| \\ &= \sqrt{\frac{mn}{n+m}}\mathcal{O}_P(d_n). \end{aligned} \tag{5.23}$$

La dernière égalité provient de (5.17). Quant à T_{22}, similairement que (5.16), nous notons que

$$\left|\frac{g\left(G^{-1}(F(x))\right)}{g\left(G^{-1}(\theta_{x,n})\right)} - 1\right| = \mathcal{O}(|\theta_{x,n} - F(x)|) = \mathcal{O}(|\psi_n(x) - F(x)|), \text{p.s.}$$

Par conséquent,

$$\begin{aligned} \sup_{x\in S(F)} |T_{22}| &= \sqrt{\frac{m}{n+m}}\mathcal{O}_P\left((\log n)^{1/2}\right)\mathcal{O}_P(\sup_{x\in S(F)} |\psi_n(x) - F(x)|) \\ &= \sqrt{\frac{nm}{n+m}}\mathcal{O}_P\left(\sqrt{\frac{\log n}{n}}\left(d_n + \sqrt{\frac{\log n}{n}}\right)\right). \end{aligned} \tag{5.24}$$

Le résultat final est obtenu en regroupant les expressions fournies par (5.15), (5.19), (5.20), (5.21), (5.22), (5.23) et (5.24). Ce qui achève la preuve du théorème 5.2.1 pour les quatres premiers estimateurs.

Quant à l'approximation du dernier estimateur $y^{[5]}_{m,n}$, nous suivrons les mêmes étapes que celles utilisées dans la preuve du lemme 5.2.1. En effet, par définition de $y^{[5]}_{m,n}$ et $B_{m,n}$, nous avons

$$\begin{aligned} &\Delta^{[5]}_{m,n} - B_{m,n}(F(x)) = \\ &\quad g\left(G^{-1}(F(x))\right)\int_{x_1}^{x_N} \frac{1}{h}K_r\left(\frac{z-x}{h}\right)\frac{1}{g\left(G^{-1}(F(z))\right)}\left[\Delta^{[1]}_{m,n}(z) - B_{m,n}(F(z))\right]\mathrm{d}z \\ &\quad + g\left(G^{-1}(F(x))\right)\int_{x_1}^{x_N} \frac{1}{h}K_r\left(\frac{z-x}{h}\right)\frac{1}{g\left(G^{-1}(F(z))\right)}\left[B_{m,n}(F(z)) - B_{m,n}(F(x))\right]\mathrm{d}z \\ &\quad + \int_{x_1}^{x_N} \frac{1}{h}K_r\left(\frac{z-x}{h}\right)B_{m,n}(F(x))\left[\frac{g\left(G^{-1}(F(x))\right)}{g\left(G^{-1}(F(z))\right)} - 1\right]\mathrm{d}z \\ &\quad + \sqrt{\frac{nm}{n+m}}g\left(G^{-1}(F(x))\right)\int_{x_1}^{x_N} \frac{1}{h}K_r\left(\frac{z-x}{h}\right)\left[G^{-1}(F(z)) - G^{-1}(F(x))\right]\mathrm{d}z \\ &:= R_1(x) + R_2(x) + R_3(x) + R_4(x). \end{aligned}$$

Il est clair qu'il existe une constante $C > 0$ telle que

$$\sup_{x\in S(F)} |R_1(x)| \;\leq\; C \sup_{x\in S(F)} \left|\Delta^{[1]}_{m,n}(x) - B_{m,n}(F(x))\right|.$$

Pour trouver une vitesse appropriée de (5.25) , il suffit d'utiliser l'expression asymptotique donnée par (5.8) où c_m et d_n sont remplacés par $\log m/m$ et $\log n/n$ respectivement. Ensuite, un simple changement de variables permet de voir que le second terme $R_2(x)$ satisfait

$$|R_2(x)| \;\leq\; \int_{(x_1-x)/h}^{(x_N-x)/h} |K_r(u)| \frac{g\left(G^{-1}(F(x))\right)}{g\left(G^{-1}(F(x+hu))\right)} \left|B_{m,n}(F(x+hu)) - B_{m,n}(F(x))\right| \mathrm{d}u.$$

Donc, les mêmes arguments que ceux utilisées dans (5.11) mènent à

$$\sup_{x\in S(F)} |R_2(x)| = \sqrt{\frac{nm}{n+m}}\mathcal{O}_P\left((n^{-1/2}+m^{-1/2})\left(h\log\frac{1}{h}\right)^{1/2}\right). \tag{5.25}$$

Similairement, la quantité $R_3(x)$ peut être réécrite comme suit

$$R_3(x) = B_{m,n}(F(x)) g\left(G^{-1}(F(x))\right) \times \int_{(x_1-x)/h}^{(x_N-x)/h} K_r(u) \left\{\frac{1}{g\left[G^{-1}(F(x+hu))\right]} - \frac{1}{g\left(G^{-1}(F(x))\right)}\right\} \mathrm{d}u.$$

Ainsi, les arguments utilisés dans (5.12) peuvent être adaptés et nous obtenons

$$\sup_{x\in S(F)} |R_3(x)| = \sqrt{\frac{nm}{n+m}}\mathcal{O}_P\left(h^{r+1}\left(\frac{\log m}{m}\right)^{1/2} + h^{r+1}\left(\frac{\log n}{n}\right)^{1/2}\right). \tag{5.26}$$

Finalement, nous avons

$$\begin{aligned}\sup_{x\in S(F)} |R_4(x)| &= \sqrt{\frac{nm}{n+m}} g\left(G^{-1}(F(x))\right) \times \\ &\qquad \int_{(x_1-x)/h}^{(x_N-x)/h} K_r(u)\left[G^{-1}(F(x+hu)) - G^{-1}(F(x))\right] \mathrm{d}u \\ &= \sqrt{\frac{nm}{n+m}}\mathcal{O}\left(h^{r+1}\right). \end{aligned} \tag{5.27}$$

En regroupant l'expression (5.8) pour $\Delta_{n,m}^{[1]}$ combinée avec (5.25)–(5.27), nous aboutissons à

$$\begin{aligned}
&\sup_{x\in S(F)} \left|\Delta_{m,n}^{[5]}(x) - B_{m,n}(F(x))\right| \\
&\quad = \sqrt{\frac{nm}{n+m}}\mathcal{O}_P\left(h^{r+1} + (n^{-1/2}+m^{-1/2})\left(h\log\frac{1}{h}\right)^{1/2} + \frac{\log n}{n} + \frac{\log m}{m} + \frac{(\log n)^{3/4}}{n^{1/4}m^{1/2}}\right).
\end{aligned}$$

□

Chapitre 6

Performance asymptotique des estimateurs lissés de la fonction d'égalisation équipercentile

Résumé. Dans ce chapitre, nous évaluons la performance locale des estimateurs polynômiaux locaux de la fonction d'égalisation équipercentile. Un critère de choix de la fenêtre est obtenu en considérant une mesure de perte que l'on appelle l'erreur ponctuelle en moyenne quadratique (MSE : Mean Squared Error).

6.1 Introduction

Dans le but de simplifier les calculs du MSE, nous étudions l'erreur moyenne quadratique des estimateurs empirique $G_m^{-1}(F_n(x))$ et lissé $\widetilde{G_m^{-1}(F_n(x))}$ de la fonction d'égalisation équipercentile. Quant aux autres estimateurs définis dans le Chapitre 2 (Section 2.3), des arguments similaires peuvent être utilisés pour calculer leurs risques quadratiques. En effet, lorsque les supports $S(F)$ et $S(G)$ des fonctions F et G sont bornés, le comportement de l'estimateur $\widetilde{G_m^{-1}(F_n(x))}$ dépend de la région où l'estimation est effectuée ; région intérieure ou régions de bords. Pour cette raison, nous allons évaluer la performance des estimateurs

proposés en utilisant l'erreur moyenne quadratique ponctuelle

$$MSE(x) = E\left[\widetilde{G_m^{-1}(F_n}(x)) - G^{-1}(F(x))\right]^2.$$

L'erreur en moyenne quadratique est un critère très répandu dans la littérature pour évaluer la précision d'une valeur estimée en un point fixe x. La décomposition habituelle du MSE est donnée par,

$$MSE(x) = Var\left[\widetilde{G_m^{-1}(F_n}(x))\right] + Biais^2\left[\widetilde{G_m^{-1}(F_n}(x))\right]. \tag{6.1}$$

Dans ce contexte, une fenêtre optimale peut être celle qui minimise l'expression (6.1). Rappelons que l' estimateur polynomial local $\widetilde{G_m^{-1}(F_n}(x))$ de la fonction d'égalisation équipercentile obtenu à partir du critère de minimisation (2.3) est donné par,

$$\begin{aligned}\widetilde{G_m^{-1}(F_n}(x)) &= \int_a^b \frac{1}{h} K_r\left(\frac{z-x}{h}\right) G_m^{-1}(F_n(z))\mathrm{d}z \qquad (6.2)\\ &= \begin{cases} \displaystyle\int_{-\alpha}^{1} K_r^L(z,\alpha) G_m^{-1}(F_n(x+hz))\mathrm{d}z, & \text{if } x = a+\alpha h \text{ with } \alpha \in (0,1]\,; \\ \displaystyle\int_{-1}^{1} K_r^I(z) G_m^{-1}(F_n(x+hz))\mathrm{d}z, & \text{if } x \in (a+h, b-h)\,; \\ \displaystyle\int_{-1}^{\alpha} K_r^R(z,\alpha) G_m^{-1}(F_n(x+hz))\mathrm{d}z, & \text{if } x = b-\alpha h \text{ with } \alpha \in (0,1], \end{cases}\end{aligned}$$

où K_r est défini par (2.7) et le lemme (2.2.1), (a, b) est le support de $G^{-1}(F(\cdot))$ et h est la fenêtre.

Dans la prochaine section, nous donnons le résultat sur le risque quadratique des estimateurs empirique et lissé de la fonction d'égalisation équipercentile.

6.2 Erreur en moyenne quadratique

6.2.1 Estimateur empirique

Nous considérons l'erreur ponctuelle en moyenne quadratique définie par

$$MSE(x) = Var\left[G_m^{-1}(F_n(x))\right] + Biais^2\left[G_m^{-1}(F_n(x))\right].$$

Le lemme qui suit présente l'espérance, la variance et la covariance de l'estimateur empirique de la fonction d'égalisation équipercentile.

Lemme 6.2.1. *Soient F et G deux fonctions de répartition. Supposons la fonction inverse G^{-1} admet des dérivées secondes bornées. Alors, pour m et n assez grand et pour tout x et y fixés dans le support de $G^{-1} \circ F$, l'estimateur empirique $G_m^{-1}(F_n(x))$ est tel que,*

$$E\left[G_m^{-1}(F_n(x))\right] = G^{-1}(F(x)) + \mathcal{O}\left(\frac{1}{m} + \frac{1}{n}\right), \tag{6.3}$$

$$Var\left[G_m^{-1}(F_n(x))\right] = \frac{m+n}{nm} F(x)(1-F(x))\left(G^{-1(1)}(F(x))\right)^2 + \mathcal{O}\left(\frac{m+n}{nm}\right)^2 \tag{6.4}$$

et

$$\begin{aligned} &Cov\left[G_m^{-1}(F_n(x)), G_m^{-1}(F_n(y))\right] = \\ &\quad \frac{m+n}{nm}\left[F(x \wedge y) - F(x)F(y)\right] G^{-1(1)}(F(x)) G^{-1(1)}(F(y)) + \mathcal{O}\left(\frac{m+n}{nm}\right)^2. \end{aligned} \tag{6.5}$$

6.2.2 Estimateur lissé dans la région intérieure

Nous considérons l'erreur ponctuelle en moyenne quadratique définie dans notre contexte par l'expression (6.1). La proposition suivante donne l'ordre de la convergence ponctuelle du terme de biais et de la variance de l'estimateur $\widetilde{G_m^{-1}(F_n(x))}$:

Proposition 6.2.1. *Soit K une densité de probabilité symétrique sur $[-1,1]$ ayant des moments d'ordre $2r$ finis où r étant un entier fixé dans $\{0,1,2\}$. Supposons que la fonction $G^{-1} \circ F$ admet des dérivées continues et bornées jusqu'à l'ordre 4 sur (a,b), avec $-\infty \leq a < b \leq \infty$. Fixons x dans la région intérieure $I = (a+h, b-h)$. et supposons que la fenêtre h satisfait $(b-a) > 2h$. Alors,*

$$\begin{aligned} Biais\left[\widetilde{G_m^{-1}(F_n(x))}\right] &= h^s \nu_{r,s}^I \frac{\left(G^{-1} \circ F\right)^{(s)}(x)}{s!} + o(h^s) + \mathcal{O}\left(\frac{m+n}{nm}\right) \\ Var\left[\widetilde{G_m^{-1}(F_n(x))}\right] &= \frac{m+n}{nm}\left\{\Psi_1(x) - h\theta_r^I \Psi_2(x) + \mathcal{O}\left(h^2 + \frac{m+n}{nm}\right)\right\} \end{aligned}$$

où $s = 2$ *si* $r = 0$ *ou* 1 *et* $s = 4$ *si* $r = 2$ *et* $\nu_{r,k}^I = \int_{-1}^1 u^k K_r^I(u)\mathrm{d}u, \quad k \geq 0$ *et* $\theta_r^I = \int_{-1}^1 \int_{-1}^1 (z \vee y) K_r^I(y) K_r^I(z)\mathrm{d}y\mathrm{d}z$ *et*

$$\Psi_1(x) = F(x)(1 - F(x)) \left[G^{-1(1)}(F(x))\right]^2,$$
$$\Psi_2(x) = F^{(1)}(x) \left[G^{-1(1)}(F(x))\right]^2.$$

6.2.3 Estimateur lissé aux régions de bords

Tout au long de cette section, nous supposons que la fonction $G^{-1}(F(x))$ est à support (a, b) tel que $-\infty < a$ et $b < \infty$. Dans le but d'étudier les effets de bords à gauche de l'estimateur polynomial local d'égalisation équipercentile, nous supposons que $x \in B_L$. Nous pouvons étudier les effets de bords à droite de manière similaire en supposant que $x \in B_R$. La proposition suivante nous donne les expressions explicites du biais et de la variance de l'estimateur lissé.

Proposition 6.2.2. *Sous les mêmes hypothèses de la proposition 6.2.1 et*

(i) si $-\infty < a$ *et* x *appartient à la région de bord gauche* B_L*, i.e.* $x = a + \alpha h$*, pour* $\alpha \in [0, 1]$*, alors*

$$Biais\left[\widetilde{G_m^{-1}(F_n(x))}\right] = h^{r+1}\nu_{r,r+1}^L \frac{(G^{-1} \circ F)^{(r+1)}(a)}{(r+1)!} + o(h^{r+1}) + \mathcal{O}\left(\frac{m+n}{nm}\right)$$
$$Var\left[\widetilde{G_m^{-1}(F_n(x))}\right] = \frac{m+n}{nm}\left\{(\alpha - \theta_r^L)hF^{(1)}(a)\left[G^{-1(1)}(0)\right]^2 + \mathcal{O}\left(h^2 + \frac{m+n}{nm}\right)\right\}$$

où $\nu_{r,k}^L = \int_{-\alpha}^1 u^k K_r^L(u, \alpha)\mathrm{d}u, \quad k \geq 0$ *et* $\theta_r^L = \int_{-\alpha}^1 \int_{-\alpha}^1 (z \vee y) K_r^L(y, \alpha) K_r^L(z, \alpha)\mathrm{d}y\mathrm{d}z$*, avec* $z \vee y = \max(y, z)$ *;*

(ii) Si $b < \infty$ *et* x *appartient à la région de bord droite* B_R *i.e.* $x = b - \alpha h$*, pour* $\alpha \in [0, 1]$*, alors*

$$Biais\left[\widetilde{G_m^{-1}(F_n(x))}\right] = h^{r+1}\nu_{r,r+1}^R \frac{(G^{-1} \circ F)^{(r+1)}(b)}{(r+1)!} + o(h^{r+1}) + \mathcal{O}\left(\frac{m+n}{nm}\right)$$
$$Var\left[\widetilde{G_m^{-1}(F_n(x))}\right] = \frac{m+n}{nm}\left\{(\alpha + \eta_r^R)hF^{(1)}(b)\left[G^{-1(1)}(1)\right]^2 + \mathcal{O}\left(h^2 + \frac{m+n}{nm}\right)\right\}$$

où $\nu_{r,k}^R = \int_{-1}^{\alpha} u^k K_r^R(u,\alpha)\mathrm{d}u, \quad k \geq 0$ *et* $\eta_r^R = \int_{-1}^{\alpha}\int_{-1}^{\alpha}(z \wedge y)K_r^R(y,\alpha)K_r^R(z,\alpha)\mathrm{d}y\mathrm{d}z$*, avec* $z \wedge y = \min(y,z)$.

Les termes dominants dans les expressions précédentes seront utilisées pour définir ce que l'on appelle l'erreur moyenne quadratique asymptotique (AMSE) qui nous permettra à son tour d'obtenir la fenêtre asymptotiquement optimale associée.

6.3 Choix de la fenêtre

La fenêtre h contrôle la largeur du voisinage considéré autour du point x lors de l'ajustement polynomial de la fonction d'égalisation équipercentile et détermine de ce fait la complexité de l'estimateur puisqu'elle détermine le degré de lissage de celui-ci. Le choix d'une fenêtre trop petite reproduit presque intégralement les données et dans ce contexte l'estimateur "sous-lisse" et est donc très variable. A contrario, un choix d'une fenêtre très grande conduit à un sur-lissage, ce qui se traduit par une augmentation du biais de l'estimateur. On peut citer en exemple le cas de la régression linéaire locale où, si l'on choisit $h = +\infty$, la courbe de régression linéaire locale se confond avec la régression linéaire simple sur l'ensemble des points. Le but est donc de choisir une fenêtre qui équilibre asymptotiquement le biais et la variance de l'estimateur.

Une importante littérature a été consacrée au calcul de MSE ou plus généralement au choix du critère du paramètre de lissage dans le contexte d'ajustement polynomial local. Fan et Gijbels (1992, 1995a,b, 1996) ont proposé des critères de choix de la fenêtre dans le cadre de la régression. Dans notre contexte, nous obtenons une fenêtre d'ajustement en minimisant l'expression de l'AMSE déduite de la somme des deux expressions de la proposition 6.2.1.

Proposition 6.3.1. *Si* K *et* $G^{-1}(F(\cdot))$ *satisfont les hypothèses de la Proposition 6.2.1, alors*

$$AMSE(x) \simeq h^{2s}\left[\nu_{r,s}^I \frac{\left(G^{-1}\circ F\right)^{(s)}(x)}{s!}\right]^2 + \frac{m+n}{nm}\left\{\Psi_1(x) - h\theta_r^I\Psi_2(x)\right\}$$

et

$$h^I = \left[\frac{m+n}{nm}\right]^{\frac{1}{2s-1}} \left[\frac{\theta_r^I \Psi_2(x)}{2s\left[\nu_{r,s}^I \dfrac{\left(G^{-1}\circ F\right)^{(s)}(x)}{s!}\right]^2}\right]^{\frac{1}{2s-1}} \tag{6.6}$$

où $s = 2$ *si* $r = 0$ *ou 1 et* $s = 4$ *si* $r = 2$ *et*

$$\Psi_1(x) = F(x)(1-F(x))\left[G^{-1(1)}(F(x))\right]^2,$$
$$\Psi_2(x) = F^{(1)}(x)\left[G^{-1(1)}(F(x))\right]^2.$$

6.4 Discussion

Dans ce chapitre, nous avons établi la convergence ponctuelle en moyenne quadratique de l'estimateur polynomial local de la fonction de l'égalisation équipercentile. cette étude nous a permis de déduire le paramètre de la fenêtre optimale h donnée par (6.6), lorsque x appartient à la région intérieure I. Si $r = 0$, nous tombons sur le cas de l'estimateur à noyau et le noyau associé de la région gauche $K_0^L(z,\alpha) = K(z)/\mu_0$ pour z dans $(-\alpha, 1)$ est positif (même problème pour la région de bord droite). Par conséquent la quantité $(\alpha - \theta_r^L)$ est toujours positive. Nous ne pouvons déterminer l'expression explicite de la fenêtre. Lorsque $r = 1, 2$, nous n'avons pas de problèmes dans la région intérieure. Tandis qu'aux régions de bords, les quantités $(\alpha - \theta_r^L)$ et $(\alpha + \eta_r^R)$ ne sont pas toujours négatives. Donc, l'estimateur se comporte moins bien aux bords.

6.5 Preuves

Avant d'aborder les résultats indiqués dans la section précédente, nous rappelons quelques résultats classiques sur les statistiques d'ordre. Pour tous entiers $n, m \geq 1$, soient $U_{(1)} \leq U_{(2)} \leq \cdots \leq U_n$ et $V_{(1)} \leq V_{(2)} \leq \cdots \leq V_m$ les statistiques d'ordre associées aux variables

aléatoires U et V i.i.d. de distribution uniforme sur $[0,1]$. Sans perte de généralité, nous supposons que $U = F(X)$ et $V = G(Y)$, alors il s'en suit que $Y_{(j)} = G^{-1}(V_{(j)})$ pour $j = 1, \ldots, m$. Par conséquent, sous des conditions de régularité sur la fonction G, nous pouvons écrire

$$Y_{(j)} = G^{-1}(V_{(j)}) = G^{-1}(p) + (V_{(j)} - p)G^{-1(1)}(p) + \frac{1}{2}(V_{(j)} - p)^2 G^{-1(2)}(\xi) \tag{6.7}$$

où $G^{-1(1)}$, $G^{-1(2)}$ sont respectivement les dérivées première et seconde de G^{-1} et ξ est tel que $V_{(j)} \wedge p < \xi < V_{(j)} \vee p$, avec $x \wedge y = \min(x, y)$ et $x \vee y = \max(x, y)$, $\forall\ x$ et y.
Soit $p_j = j/(m+1)$ pour $j = 1, \ldots, m$ et rappelons que les statistiques d'ordre $V_{(j)}$ satisfont

$$E\left[V_{(j)}\right] = p_j, \quad Var\left[V_{(j)}\right] = \frac{p_j\,(1-p_j)}{m+2} \quad \text{et} \quad Cov\left[V_{(j)}, V_{(k)}\right] = \frac{p_j\,(1-p_k)}{m+2}, \qquad \text{si } j < k$$

(voir, e.g., Arnold *et al.* (2008) ou Gibbons et Chakraborti (2003)). Alors, l'expression dans (6.7) entraîne que pour tous points p et q dans $(0,1)$, nous avons

$$\begin{aligned} E\left[Y_{(j)}\right] &= G^{-1}(p) + (p_j - p)G^{-1(1)}(p) + \mathcal{O}\left(E(V_{(j)} - p)^2\right), \\ Var\left[Y_{(j)}\right] &= \frac{p_j\,(1-p_j)}{m+2}\left(G^{-1(1)}(p)\right)^2 + \mathcal{O}\left(E(V_{(j)} - p)^2\right)^2 \end{aligned}$$

et

$$\begin{aligned} Cov\left[Y_{(j)}, Y_{(k)}\right] = \frac{p_j\,(1-p_k)}{m+2} & G^{-1(1)}(p)G^{-1(1)}(q) \\ & + \mathcal{O}\left(E(V_{(j)} - p)^2 E(V_{(k)} - q)^2\right), \qquad \text{si } j < k. \end{aligned}$$

Par ailleurs, étant donné que la fonction quantile peut être réécrite comme suit

$$G_m^{-1}(p) = \sum_{j=1}^{m} Y_{(j)} \mathbb{I}(j - 1 < mp \leq j), \quad \forall\ p \in (0,1),$$

donc à l'aide de l'expression de l'expression ci-dessus, il s'ensuit que

$$\begin{aligned} E\left[G_m^{-1}(p)\right] &= \sum_{j=1}^{m} \left\{G^{-1}(p) + (p_j - p)G^{-1(1)}(p)\right\} \mathbb{I}(j - 1 < mp \leq j) \\ &\quad + \mathcal{O}\left(\sum_{j=1}^{m} \left\{\frac{p_j\,(1-p_j)}{m+2} + (p_j - p)^2\right\} \mathbb{I}(j - 1 < mp \leq j)\right) \\ &= G^{-1}(p) + \mathcal{O}\left(\frac{1}{m}\right), \end{aligned} \tag{6.8}$$

$$\begin{aligned} Var\left[G_m^{-1}(p)\right] &= \frac{1}{m+2}\left(G^{-1(1)}(p)\right)^2 \sum_{j=1}^{m} p_j\left(1-p_j\right)\mathbb{I}(j-1<mp\leq j) \\ &\quad + \frac{1}{(m+2)(m+1)}\mathcal{O}(1) \\ &= \frac{1}{m+2}p(1-p)\left(G^{-1(1)}(p)\right)^2 + \mathcal{O}\left(\frac{1}{m^2}\right) \end{aligned} \tag{6.9}$$

et

$$\begin{aligned} &Cov\left[G_m^{-1}(p),G_m^{-1}(q)\right] \\ &\quad = \sum_{j=1}^{m}\sum_{k=1}^{m} Cov\left[Y_{(j)},Y_{(k)}\right]\mathbb{I}(j-1<mp\leq j)\mathbb{I}(k-1<mq\leq k) \\ &\quad = \frac{1}{m+2}\sum_{j,k=1}^{m}\left\{\frac{j\wedge k}{m+1}\left(1-\frac{j\vee k}{m+1}\right)G^{-1(1)}(p)G^{-1(1)}(q)\right. \\ &\qquad \left. + \mathcal{O}\left(|p_j-p|+|p_k-q|\right)\right\}\mathbb{I}(j-1<mp\leq j)\mathbb{I}(k-1<mq\leq k) \\ &\quad = \frac{1}{m+2}(p\wedge q-pq)G^{-1(1)}(p)G^{-1(1)}(q)+\mathcal{O}\left(\frac{1}{m^2}\right). \end{aligned} \tag{6.10}$$

Ces expressions nous serviront pour étudier le comportement de l'estimateur empirique $G_m^{-1}\circ F_n$.

Preuve du lemme 6.2.1. Nous désignons par E^* l'espérance conditionnelle par rapport aux variables aléatoires $X_1,\ldots,X_n$. Alors, en utilisant l'expression asymptotique fournie par (6.8), nous remarquons que,

$$\begin{aligned} E\left[G_m^{-1}(F_n(x))\right] &= E\left\{E^*\left[G_m^{-1}(F_n(x))\right]\right\} \\ &= E\left[G^{-1}(F_n(x))+\mathcal{O}\left(\frac{1}{m}\right)\right]. \end{aligned}$$

Alors, (6.3) sera issue de l'expression de Taylor de $G^{-1}(F_n(x))$ au voisinage de $F(x)$ et du fait que la dérivée seconde $G^{-1(2)}$ est bornée, i.e.,

$$\begin{aligned} &E\left[G^{-1}(F_n(x))\right] \\ &\quad = E\left[G^{-1}(F(x))+(F_n(x)-F(x))G^{-1(1)}(F(x))+\frac{1}{2}(F_n(x)-F(x))^2G^{-1(2)}(\xi)\right] \\ &\quad = G^{-1}(F(x))+Var(F_n(x))\mathcal{O}(1) \end{aligned}$$

où ξ est tel que $F(x) \wedge F_n(x) < \xi < F(x) \vee F_n(x)$.
D'autre part, pour calculer la variance de $G_m^{-1}(F_n(x))$, nous désignons par Var^* la variance conditionnelle par rapport aux variables aléatoires $X_1, \ldots, X_n$ et notons que

$$Var\left[G_m^{-1}(F_n(x))\right] = E\left\{Var^*\left[G_m^{-1}(F_n(x))\right]\right\} + Var\left\{E^*\left[G_m^{-1}(F_n(x))\right]\right\}.$$

Alors, en utilisant (6.9), nous obtenons

$$\begin{aligned}
&E\left\{Var^*\left[G_m^{-1}(F_n(x))\right]\right\} \\
&\quad = E\left\{\frac{1}{m+2}F_n(x)(1-F_n(x))\left(G^{-1(1)}(F_n(x))\right)^2\right\} + \mathcal{O}\left(\frac{1}{m^2}\right) \\
&\quad = \frac{1}{m+2}\frac{n-1}{n}F(x)(1-F(x))\left(G^{-1(1)}(F(x))\right)^2 \\
&\qquad + \frac{1}{m+2}E\left\{2F_n(x)(1-F_n(x))(F_n(x)-F(x))G^{-1(1)}(\xi)G^{-1(2)}(\xi)\right\} \\
&\qquad + \mathcal{O}\left(\frac{1}{m^2}\right) \\
&\quad = \frac{1}{m+2}F(x)(1-F(x))\left(G^{-1(1)}(F(x))\right)^2 + \mathcal{O}\left(\frac{1}{m^2}+\frac{1}{mn}\right). \qquad (6.11)
\end{aligned}$$

De manière similaire, (6.8) combiné avec l'expression de Taylor permet d'obtenir

$$\begin{aligned}
Var\left\{E^*\left[G_m^{-1}(F_n(x))\right]\right\} &= Var\left\{G^{-1}(F_n(x)) + \mathcal{O}\left(\frac{1}{m}\right)\right\} \\
&= Var\left[G^{-1}(F(x)) + (F_n(x)-F(x))G^{-1(1)}(F(x)) + \frac{1}{2}(F_n(x)-F(x))^2 G^{-1(2)}(\xi)\right] \\
&= \frac{1}{n}F(x)(1-F(x))\left(G^{-1(1)}(F(x))\right)^2 + \mathcal{O}\left(\frac{1}{n^2}\right). \qquad (6.12)
\end{aligned}$$

En combinant les deux approximations dans (6.11) et (6.12), nous obtenons (6.4).
Finalement, l'approximation de la covariance suit les mêmes étapes, i.e.,

$$\begin{aligned}
Cov\left[G_m^{-1}(F_n(x)), G_m^{-1}(F_n(y))\right] &= E\left\{Cov^*\left[G_m^{-1}(F_n(x)), G_m^{-1}(F_n(y))\right]\right\} \\
&\quad + Cov\left\{E^*\left[G_m^{-1}(F_n(x))\right], E^*\left[G_m^{-1}(F_n(y))\right]\right\},
\end{aligned}$$

avec Cov^* étant la covariance conditionnelle sachant $X_1, \ldots, X_n$.

L'expression (6.10) permet d'avoir

$$
\begin{aligned}
& E\left\{Cov^*\left[G_m^{-1}(F_n(x)), G_m^{-1}(F_n(y))\right]\right\} \qquad (6.13)\\
&= E\left\{\frac{1}{m+2}\left[F_n(x)\wedge F_n(y) - F_n(x)F_n(y)\right] G^{-1(1)}(F_n(x))G^{-1(1)}(F_n(y)) + \mathcal{O}\left(\frac{1}{m^2}\right)\right\}\\
&= \frac{1}{m+2}\left[F(x)\wedge F(y) - F(x)F(y)\right] G^{-1(1)}(F(x))G^{-1(1)}(F(y)) + \mathcal{O}\left(\frac{1}{m^2}+\frac{1}{mn}\right).
\end{aligned}
$$

Alors que (6.8) entraîne que

$$
\begin{aligned}
& Cov\left\{E^*\left[G_m^{-1}(F_n(x))\right], E^*\left[G_m^{-1}(F_n(y))\right]\right\} \qquad (6.14)\\
&= Cov\left\{G^{-1}(F_n(x)) + \mathcal{O}\left(\frac{1}{m}\right), G^{-1}(F_n(y)) + \mathcal{O}\left(\frac{1}{m}\right)\right\}\\
&= E\Bigg\{\left[(F_n(x)-F(x))G^{-1(1)}(F(x)) + \frac{1}{2}(F_n(x)-F(x))^2 G^{-1(2)}(\xi_1) + \mathcal{O}\left(\frac{1}{m}+\frac{1}{n}\right)\right]\\
&\qquad \times\left[(F_n(y)-F(y))G^{-1(1)}(F(y)) + \frac{1}{2}(F_n(y)-F(y))^2 G^{-1(2)}(\xi_2) + \mathcal{O}\left(\frac{1}{m}+\frac{1}{n}\right)\right]\Bigg\}\\
&= \frac{1}{n}\left[F(x\wedge y) - F(x)F(y)\right] G^{-1(1)}(F(x))G^{-1(1)}(F(y)) + \mathcal{O}\left(\frac{1}{n^2}+\frac{1}{mn}\right),
\end{aligned}
$$

où ξ_1 et ξ_2 satisfont $F(x)\wedge F_n(x) < \xi_1 < F(x)\vee F_n(x)$ et $F(y)\wedge F_n(y) < \xi_1 < F(y)\vee F_n(y)$ respectivement. Pour conclure la preuve de (6.5), il suffit de combiner les deux expressions dans (6.13) et (6.14).

Preuve des propositions 6.2.1 et 6.2.2. En premier abord, nous établissons les termes du biais dans les deux propositions. En effet, si le support de $G^{-1}\circ F$ est borné à gauche et le point d'intérêt appartient à la région de bord gauche, i.e., $x = a + \alpha h$, avec $\alpha \in [0,1]$,

alors, d'après (6.2) et (6.3),

$$\begin{aligned} E\left[\widetilde{G_m^{-1}(F_n}(x))\right] &= \int_{-\alpha}^{1} K_r^L(z,\alpha) E\left[G_m^{-1}(F_n(x+hz))\right] \mathrm{d}z \\ &= \int_{-\alpha}^{1} K_r^L(z,\alpha) \left\{ G^{-1}(F(x+hz)) + \mathcal{O}\left(\frac{1}{m}+\frac{1}{n}\right)\right\} \\ &= G^{-1}(F(x)) + h^{r+1}\frac{(G^{-1}\circ F)^{(r+1)}(x)}{(r+1)!}\int_{-\alpha}^{1} z^{r+1} K_r^L(z,\alpha)\mathrm{d}z \\ &\quad + o(h^{r+1}) + \mathcal{O}\left(\frac{1}{m}+\frac{1}{n}\right). \end{aligned}$$

La dernière égalité découle de l'expression de Taylor de $G^{-1}\circ F$, du théorème de convergence dominée et du (i) de la Proposition 2.2.1. La preuve du biais lorsque x appartient à l'intérieur et à la région de bord droite, suit les mêmes étapes. Quant à l'expression de la variance, nous utilisons (6.5) afin d'écrire

$$\begin{aligned} Var\left[\widetilde{G_m^{-1}(F_n}(x))\right] &= \int_{\Omega} K_r(y)K_r(z)Cov\left[G_m^{-1}(F(x+hy)), G_m^{-1}(F(x+hz))\right]\mathrm{d}y\mathrm{d}z \\ &= \frac{m+n}{nm}\int_{\Omega} K_r(y)K_r(z)\Theta(x+hy,x+hz)\mathrm{d}y\mathrm{d}z + \mathcal{O}\left(\frac{m+n}{nm}\right)^2, \end{aligned}$$

où

$$\{\Omega, K_r(\cdot)\} = \begin{cases} \{[-\alpha,1]\times[-\alpha,1], K_r^L(\cdot)\}, & \text{si } a > -\infty \text{ et } x\in[a,a+h)\,; \\ \{[-1,1]\times[-1,1], K_r^I(\cdot)\}, & \text{si } -\infty < a < b < \infty \text{ et } x\in[a+h,b-h]\,; \\ \{[-1,\alpha]\times[-1,\alpha], K_r^R(\cdot)\}, & \text{si } b<\infty \text{ et } x\in(b-h,b]\,; \end{cases}$$

et

$$\begin{aligned} &\Theta(x+hy,x+hz) \\ &\quad := \left[F(x+h(y\wedge z)) - F(x+hy)F(x+hz)\right] G^{-1(1)}(F(x+hy))G^{-1(1)}(F(x+hz)) \\ &\quad = \left\{F(x)(1-F(x)) + h[(z\wedge y) - (y+z)F(x)]F^{(1)}(x) + \mathcal{O}(h^2)\right\} \\ &\times \left\{\left[G^{-1(1)}(F(x))\right]^2 + h(y+z)G^{-1(1)}(F(x))G^{-1(2)}(F(x))F^{(1)}(x) + \mathcal{O}(h^2)\right\} \\ &\quad = \Psi_1(x) + h(z\wedge y)\Psi_2(x) - (y+z)h\Psi_3(x) + \mathcal{O}(h^2) \end{aligned}$$

avec

$$\Psi_1(x) = F(x)(1-F(x))\left[G^{-1(1)}(F(x))\right]^2$$
$$\Psi_2(x) = F^{(1)}(x)\left[G^{-1(1)}(F(x))\right]^2$$
$$\Psi_3(x) = F(x)F^{(1)}(x)G^{-1(1)}(F(x))\left\{G^{-1(1)}(F(x)) - (1-F(x))G^{-1(2)}(F(x))\right\}.$$

Alors, d'après les propriétés du noyau K_r citées dans la Proposition 2.2.1, nous avons

$$Var\left[\widetilde{G_m^{-1}(F_n(x))}\right] = \frac{m+n}{nm}\left\{\Psi_1(x) - h\left[\theta_r\Psi_2(x) + 2\nu_{r,1}\Psi_3(x)\right]\right\} + \mathcal{O}\left(\frac{m+n}{nm}\right) + \mathcal{O}\left(h^2\right)$$

avec $\theta_r = \int_\Omega (z \vee y)K_r(y)K_r(z)\mathrm{d}y\mathrm{d}z$ et $\nu_{r,1}$ représente $\nu_{r,1}^L$ ou $\nu_{r,1}^I$ ou $\nu_{r,1}^R$. Or selon la Proposition 2.2.1, les moments $\nu_{r,1}^I$ s'annulent pour $x \in I$, par conséquent l'expression dans (ii) se déduit facilement. Quant au cas des régions de bord, quand $x = a + \alpha h$ pour $\alpha \in (0,1)$, le développement de Taylor de $\Psi_i(a+\alpha h)$ et le fait que $F(a) = 0$ permettent de déduire le résultat dans (i) de la proposition 6.2.2. Finalement, afin d'obtenir (ii) de la proposition 6.2.2, nous utilisons les mêmes arguments et le fait que $F(b) = 1$. □

Notons que, si le support de $G^{-1} \circ F$ est $\mathbb{R}$, alors on n'aura pas à s'inquiéter aux effets de bords et le noyau $K_r(\cdot)$ coïncide avec $K_r^I(\cdot)$.

Chapitre 7

Simulations

Résumé. Dans ce chapitre, nous proposons une étude numérique, à l'aide du logiciel R (R Development Core Team (2008)), pour illustrer nos résultats précédents avec des comparaisons des performances des estimateurs empirique et polynômiaux locaux.

7.1 Présentation

Dans le cadre de l'égalisation équipercentile des scores observés, nous avons programmé sous R les estimateurs empirique et lissés obtenus par l'approche de l'ajustement polynomial local utilisée dans nos résultats. Nous décrivons ci-dessous les différentes étapes de cette implémentation en précisant l'ensemble des outils utilisés.

7.1.1 Estimateurs utilisés

Soient X et Y deux variables aléatoires de fonctions de répartition respectives F et G. On se donne $X_1, \ldots, X_n$ et $Y_1, \ldots, Y_m$ deux échantillons de variables aléatoires distribuées comme X et Y, de tailles $n > 1$ et $m > 1$, et à valeurs sur $S(F)$ et $S(G)$, supports de F et G respectivement. F_n et G_m désignent alors les fonctions de répartition empiriques basées sur les deux échantillons. Nous rappelons la forme des estimateurs de la fonction d'égalisation

équipercentile permettant d'estimer les scores y à partir des scores x introduits au chapitre 2 :

$$\begin{aligned}
y_{m,n}^{[1]}(x) &= G_m^{-1}(F_n(x)); \\
y_{m,n}^{[2]}(x) &= G_m^{-1}(\widetilde{F_n}(x)); \\
y_{m,n}^{[3]}(x) = \widetilde{G_m^{-1}}(F_n(x)) &= \int_0^1 \frac{1}{h} K_r\left(\frac{z - F_n(x)}{h}\right) G_m^{-1}(z)\mathrm{d}z; \\
y_{m,n}^{[4]}(x) = \widetilde{G_m^{-1}}(\widetilde{F_n}(x)) &= \int_0^1 \frac{1}{k} L_r\left(\frac{z - \widetilde{F_n}(x)}{k}\right) G_m^{-1}(z)\mathrm{d}z; \\
y_{m,n}^{[5]}(x) = \widetilde{G_m^{-1} \circ F_n}(x)) &= \int_{x_1}^{x_N} \frac{1}{h} K_r\left(\frac{z - x}{h}\right) G_m^{-1} \circ F_n(z)\mathrm{d}z,
\end{aligned}$$

avec

$$\widetilde{F_n}(x) = \int_{x_1}^{x_N} \frac{1}{h} K_r\left(\frac{z - x}{h}\right) F_n(z)\mathrm{d}z$$

et

$$\widetilde{G_m^{-1}}(p) = \int_0^1 \frac{1}{h} K_r\left(\frac{s - p}{h}\right) G_m^{-1}(s)\mathrm{d}s.$$

De manière analogue, les estimateurs de la fonction d'égalisation équipercentile permettant d'estimer le score x à partir de y sont donnés par :

$$\begin{aligned}
x_{m,n}^{[1]}(y) &= F_n^{-1}(G_m(y)); \\
x_{m,n}^{[2]}(y) &= F_n^{-1}(\widetilde{G_m}(y)); \\
x_{m,n}^{[3]}(y) &= \widetilde{F_n^{-1}}(G_m(y)); \\
x_{m,n}^{[4]}(y) &= \widetilde{F_n^{-1}}(\widetilde{G_m}(y)); \\
x_{m,n}^{[5]}(y) &= \widetilde{F_n^{-1} \circ G_m}(y)).
\end{aligned}$$

La symétrie est l'une propriété importante de l'égalisation des scores. von Davier *et al.* (2007) ont proposé une méthode générale pour construire des fonctions hybrides d'égalisation qui combinent des fonctions d'égalisation linéaires et non-linéaires tout en préservant la propriété de la symétrie (Dorans et Holland (2000)).

Dans nos simulations, nous utilisons des estimateurs polynomiaux locaux linéaires i.e. $r = 1$. La mise en œuvre de ces estimateurs repose clairement sur des choix ad hoc du noyau d'ordre supérieur K_r et des paramètres de lissage h.

7.1.2 Noyaux d'ordre supérieur

Rappelons l'expression du noyau K_r pour $r = 1$:

$$K_r(x) = \frac{\mu_2 - \mu_1 x}{\mu_0\mu_2 - \mu_1^2} K(x).$$

avec

$$\mu_\ell = \frac{1}{h}\int_a^b \left(\frac{z-u}{h}\right)^\ell K\left(\frac{z-u}{h}\right) \mathrm{d}z = \int_{(a-u)/h}^{(b-u)/h} x^\ell K(x)\mathrm{d}x < \infty, \text{ pour } \ell = 0, \ldots, 2r.$$

Nous utilisons ensuite le noyau d'Epanechnikov, i.e. $K(x) = \frac{3}{4}\left(1-x^2\right)\mathbb{1}\left(|x| \leq 1\right)$, qui est optimal pour le critère quadratique moyen (voir Tsybakov (2004), p.17). Pour ce choix, lorsque x appartient à la région intérieure (I) du support $S(F) = [a, b]$, i.e. $x \in I = [a+h, b-h]$ avec $(b-a)/2 \geq h$, nous avons,

$$K_1^I(x) = \frac{3}{4}\left(1-x^2\right)\mathbb{1}\left(|x| \leq 1\right).$$

Maintenant, si x appartient à la région de bord gauche B_L, i.e. $x = a + \alpha h$ avec $\alpha \in [0, 1]$, alors pour $x \in [-\alpha, 1]$, nous avons,

$$K_1^L(x) = -12\frac{(12\alpha^3 - 24\alpha^2 + 16\alpha - 8) + 15x(1-\alpha)^2}{(1+\alpha)^4(3\alpha^2 - 18\alpha + 19)}\left(1-x^2\right)\mathbb{1}\left(-\alpha \leq x \leq 1\right).$$

Par symétrie, si x appartient à la région de bord droite B_R, i.e. $x = b - \alpha h$, alors pour $x \in [-1, \alpha]$, nous avons

$$K_1^R(x) = -12\frac{(12\alpha^3 - 24\alpha^2 + 16\alpha - 8) - 15x(1-\alpha)^2}{(1+\alpha)^4(3\alpha^2 - 18\alpha + 19)}\left(1-x^2\right)\mathbb{1}\left(-1 \leq x \leq \alpha\right).$$

7.1.3 Paramètres de lissage

Concernant les choix optimaux de la fenêtre h, nous rappelons les expressions des paramètres de lissage des estimateurs polynomiaux locaux de la fonction de répartition et de la fonction quantile obtenus de la même manière que la fenêtre de l'estimateur de la fonction d'égalisation équipercentile. Donc, pour l'estimateur polynomial local de la fonction de répartition $\widetilde{F_n}(x)$:

$$h_{\widetilde{F_n}} = \left[\frac{n^{-1} F'(x) \theta_r^I}{\left(\mu_2^I\right)^2 \left(F^{(2)}(x)\right)^2} \right]^{1/3}. \tag{7.1}$$

Pour l'estimateur polynomial local de la fonction quantile $\widetilde{G_m^{-1}}$:

$$h_{\widetilde{G_m^{-1}}} = \left[\frac{m^{-1} [(G^{-1})'(p)]^2 \theta_r^I}{\left(\mu_2^I\right)^2 \left((G^{-1})^{(2)}(p)\right)^2} \right]^{1/3}. \tag{7.2}$$

Pour l'estimateur polynomial local $y_{n,m}^{[5]}(x)$ de la fonction d'égalisation équipercentile :

$$h_{\widetilde{G_m^{-1} \circ F_n}} = \left[\frac{m+n}{nm} \right]^{\frac{1}{3}} \left[\frac{\theta_r^I F'(x) \left[(G^{-1})'(F(x))\right]^2}{\left[\mu_2^I \left(G^{-1} \circ F\right)^{(2)}(x)\right]^2} \right]^{\frac{1}{3}}.$$

où $\theta_r^I = 2 \int_{-1}^{1} K_r(u) \mathbf{K_r}(u) \mathrm{d}u$ avec $\mathbf{K_r}(t) = \int_{-1}^{t} K_r(u) \mathrm{d}u$.
Pour les estimateurs $y_{n,m}^{[2]}(x) = G_m^{-1}(\widetilde{F_n}(x))$, $y_{n,m}^{[3]}(x) = \widetilde{G_m^{-1}}(F_n(x))$ et $y_{n,m}^{[4]}(x) = \widetilde{G_m^{-1}}(\widetilde{F_n}(x))$, nous utiliserons donc la fenêtre asymptotique optimale associée aux fonctions lissées $\widetilde{F_n}$ et $\widetilde{G_m^{-1}}$.

Clairement, l'utilisation de ces expressions théoriques des fenêtres optimales peut s'avérer délicate en pratique dans la mesure où elles dépendent des paramètres $F'(\cdot)$ et $(G^{-1})'(\cdot)$. Ces paramètres sont généralement inconnus ; et, lorsque les lois de distribution sont supposées connues, leurs expressions ne sont pas toujours explicites. Nous nous limiterons ainsi dans cette section à des lois pour lesquelles les quantités F' et $(G^{-1})'$ sont connues explicitement (loi exponentielle et loi de Weibull) ou pour laquelle F et G^{-1} peuvent être approchées numériquement (loi normale).

7.2 Calcul des estimateurs et leurs réciproques sur données simulées

Afin d'illustrer nos résultats théoriques, nous générons plusieurs échantillons pour les variables aléatoires X et Y ; et cherchons à établir une échelle de correspondances entre les valeurs X et Y en utilisant chacun des cinq estimateurs proposés. Dans ce qui suit, n et m désignent respectivement les tailles des échantillons de X et Y et $[a, b]$ désigne le support de F.

Pour pouvoir utiliser l'expression exacte de la fenêtre de l'estimateur $y^{[5]}_{n,m}(x)$, il est clair que les fonctions F et G doivent être distinctes afin de ne pas obtenir des quantités nulles dans le dénominateur.

Cas 1 : $X \sim \mathcal{N}(0,1)$, $Y \sim \mathcal{E}(1)$, n=250, m=300 et $[a,b] = [0,3]$.

Pour une valeur x de X donnée, nous pouvons déterminer son équivalent y ayant le même rang percentile que x via la graphe des fonctions de répartition lissées obtenues des deux échantillons X et Y. Dans l'exemple suivant, pour $x = 2$ nous déterminons facilement la valeur de y sur la figure (7.1) telle que $\widetilde{F_n}(x) = \widetilde{G_m}(y)$.

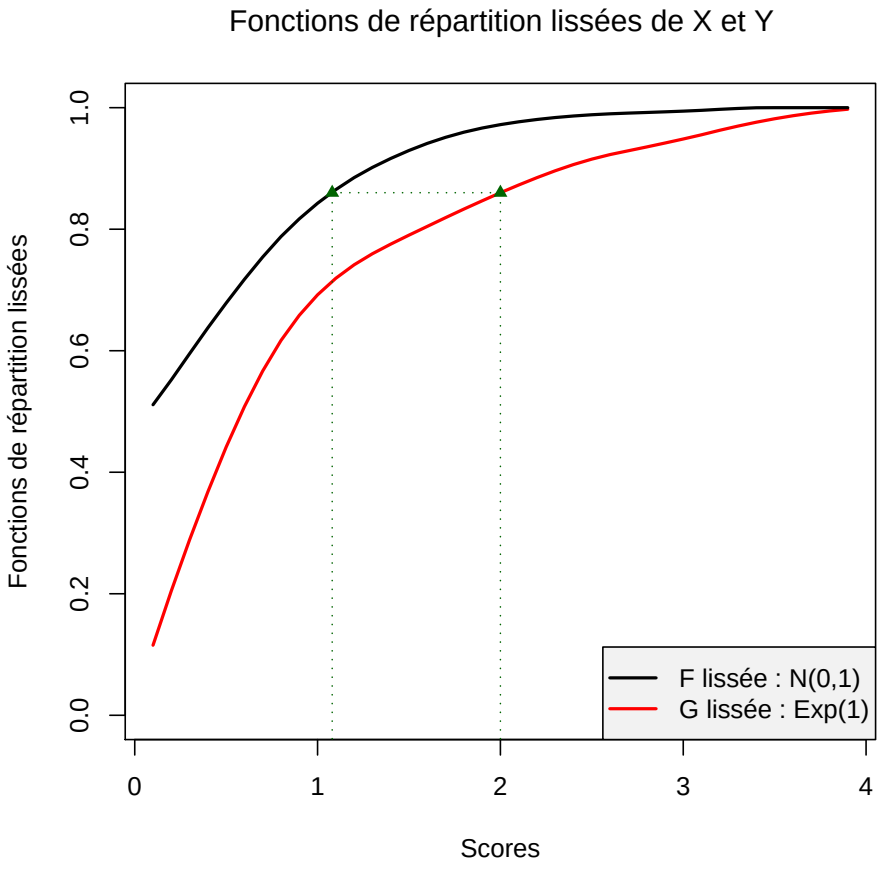


FIG. 7.1 – Fonctions de répartition de $\mathcal{N}(0,1)$ et $\mathcal{E}(1)$.

Dans les figures (7.2) et (7.3), les graphes en trait continu représentent les estimateurs de la fonction d'égalisation équipercentile de y en x, tandis que les graphes en trait-point

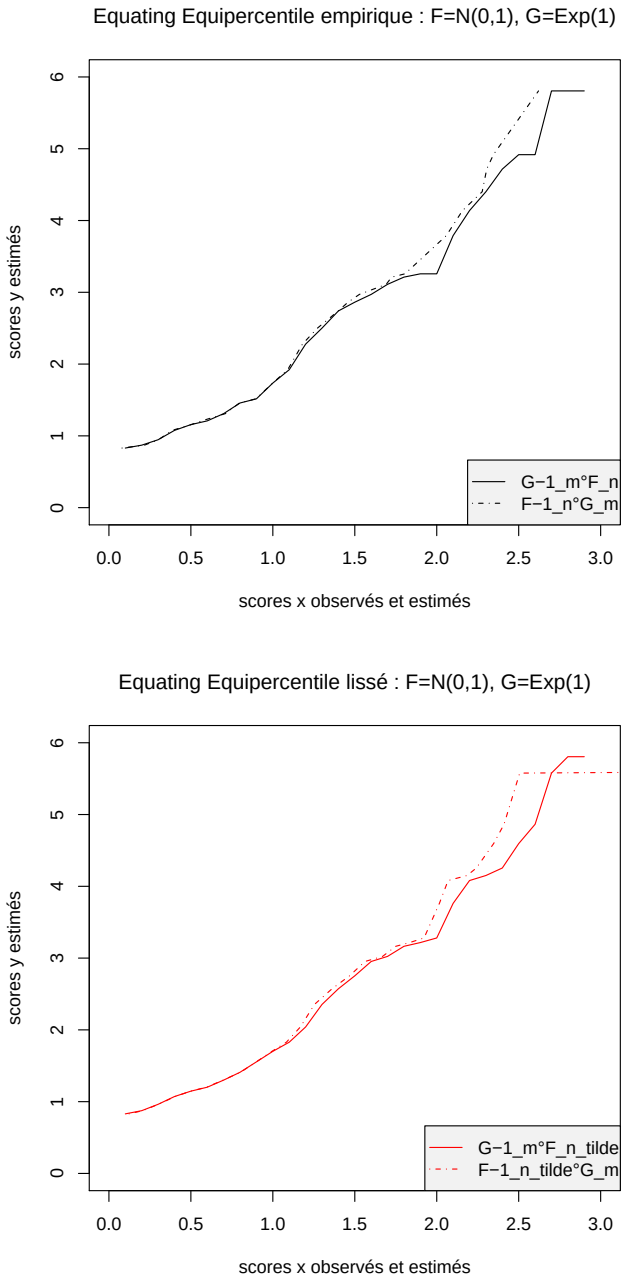


FIG. 7.2 – Estimateurs $y_{n,m}^{[1]}$ et $y_{n,m}^{[2]}$ de la fonction d'égalisation équipercentile de x en y avec leurs réciproques.

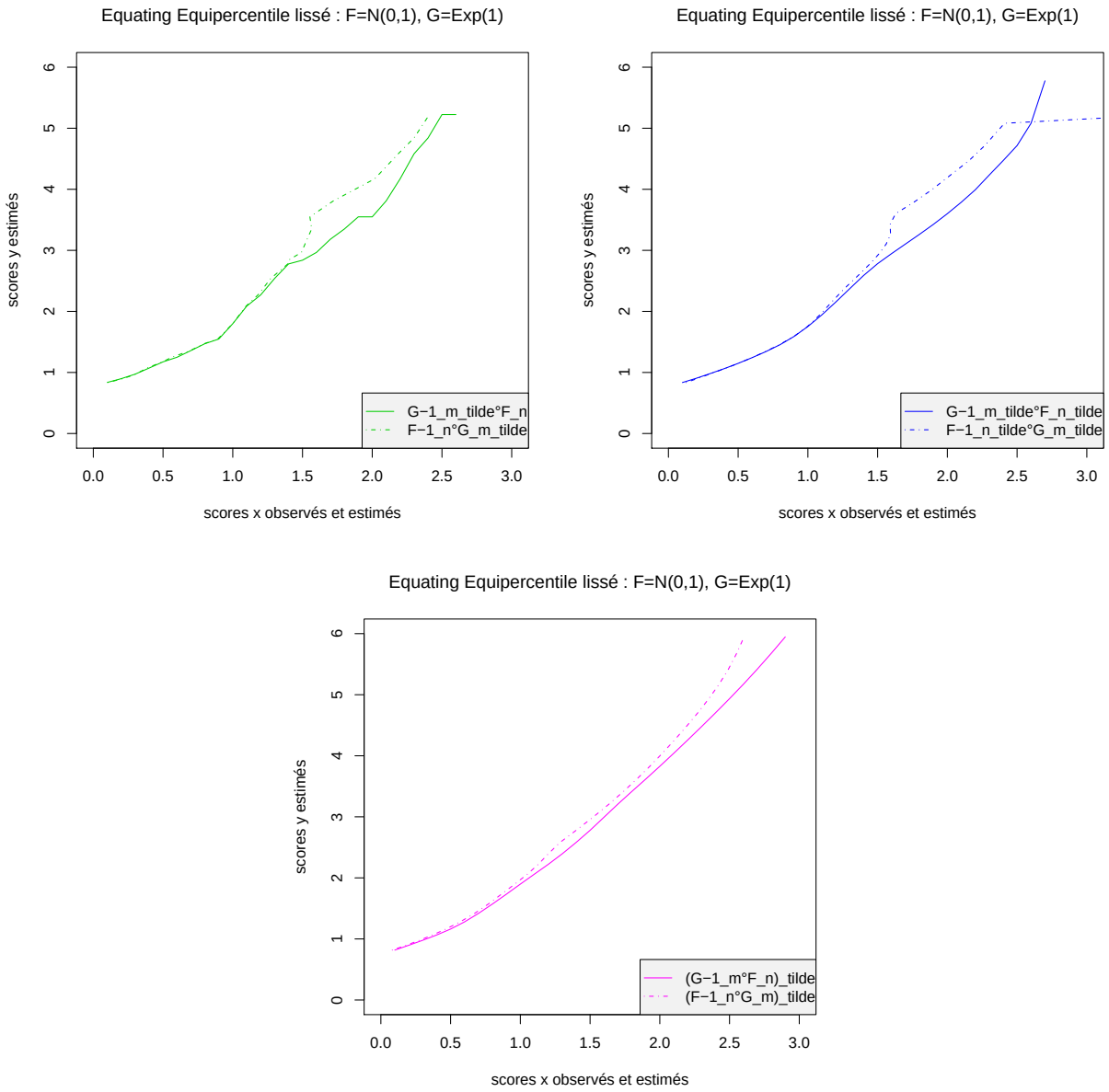


FIG. 7.3 – Estimateurs $y^{[3]}_{n,m}$, $y^{[4]}_{n,m}$ et $y^{[5]}_{n,m}$ de la fonction d'égalisation équipercentile de x en y avec leurs réciproques.

représentent les estimateurs de la fonction d'égalisation équipercentile de x en y.

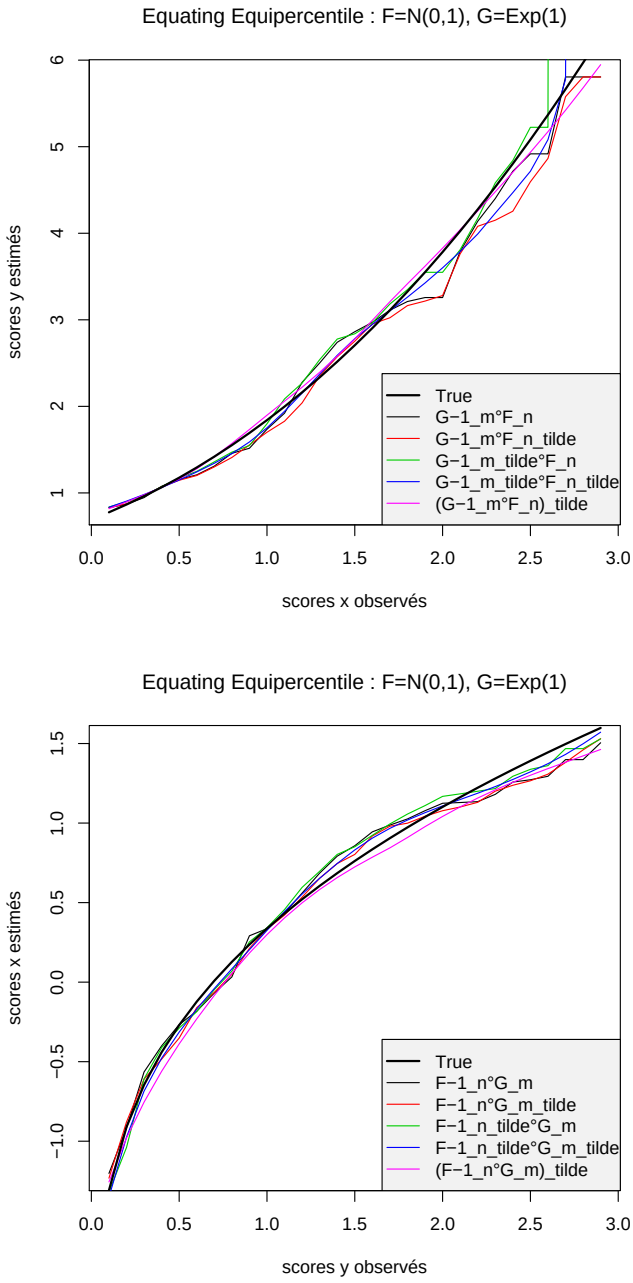


FIG. 7.4 – Estimateurs de la fonction d'égalisation équipercentile de x en y et leurs réciproques.

Cas 2 : $X \sim \mathcal{N}(0,1)$, $Y \sim Weibull(3,2)$, n=250, m=300 et $[a,b] = [0,3]$.

Pour déterminer le score équivalent de la valeur $x = 2$ par exemple, il suffit de lire sur la figure (7.5) la valeur correspondante telle que $\widetilde{F_n}(x) = \widetilde{G_m}(y)$.

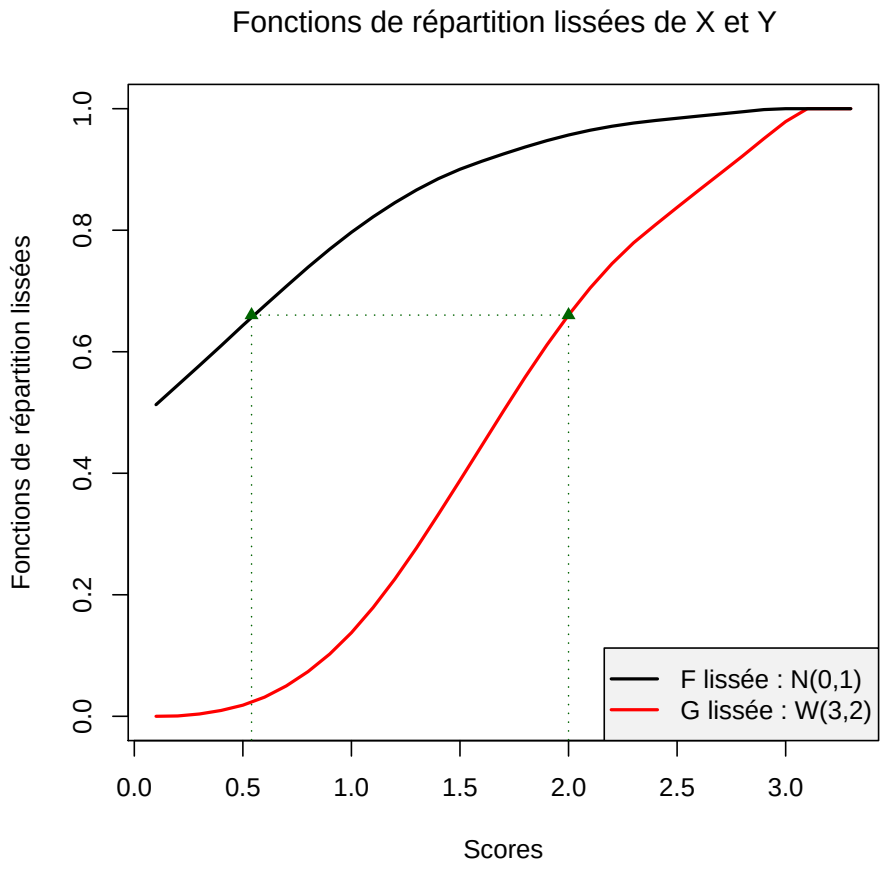


FIG. 7.5 – Fonctions de répartition de $\mathcal{N}(0,1)$ et $\mathcal{E}(1)$.

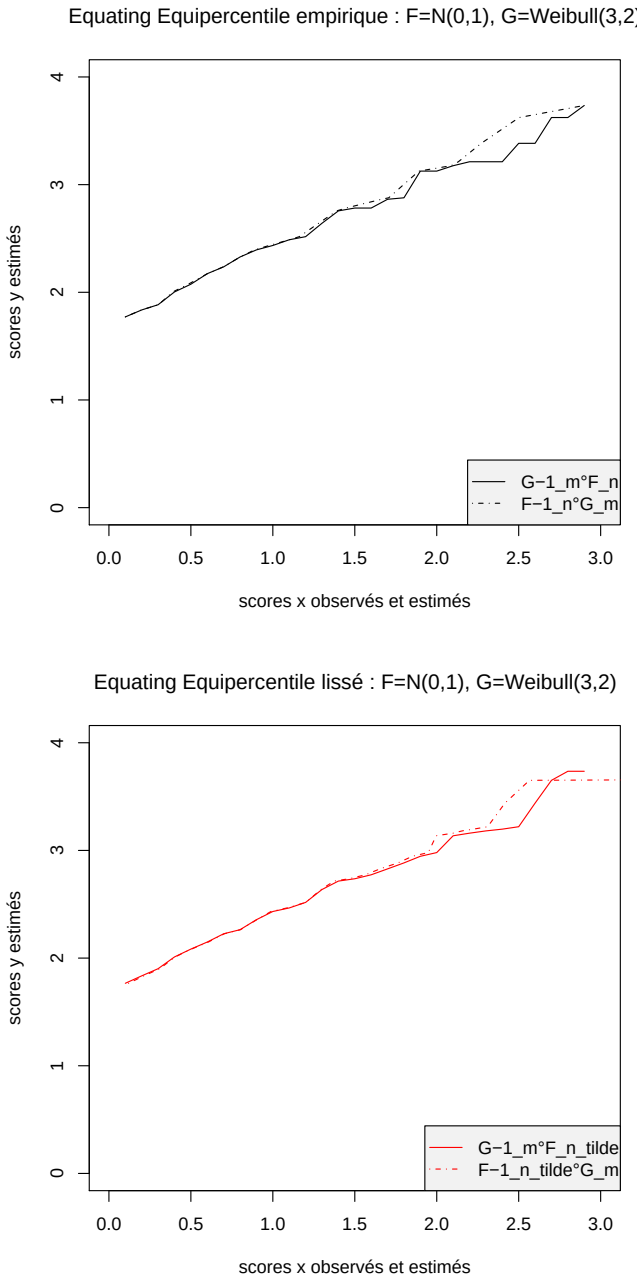


FIG. 7.6 – Comparaison des différents estimateurs de la fonction d'égalisation équipercentile de x en y avec leurs réciproques.

Dans les figures (7.6) et (7.7), les estimateurs de la fonction d'égalisation équipercentile

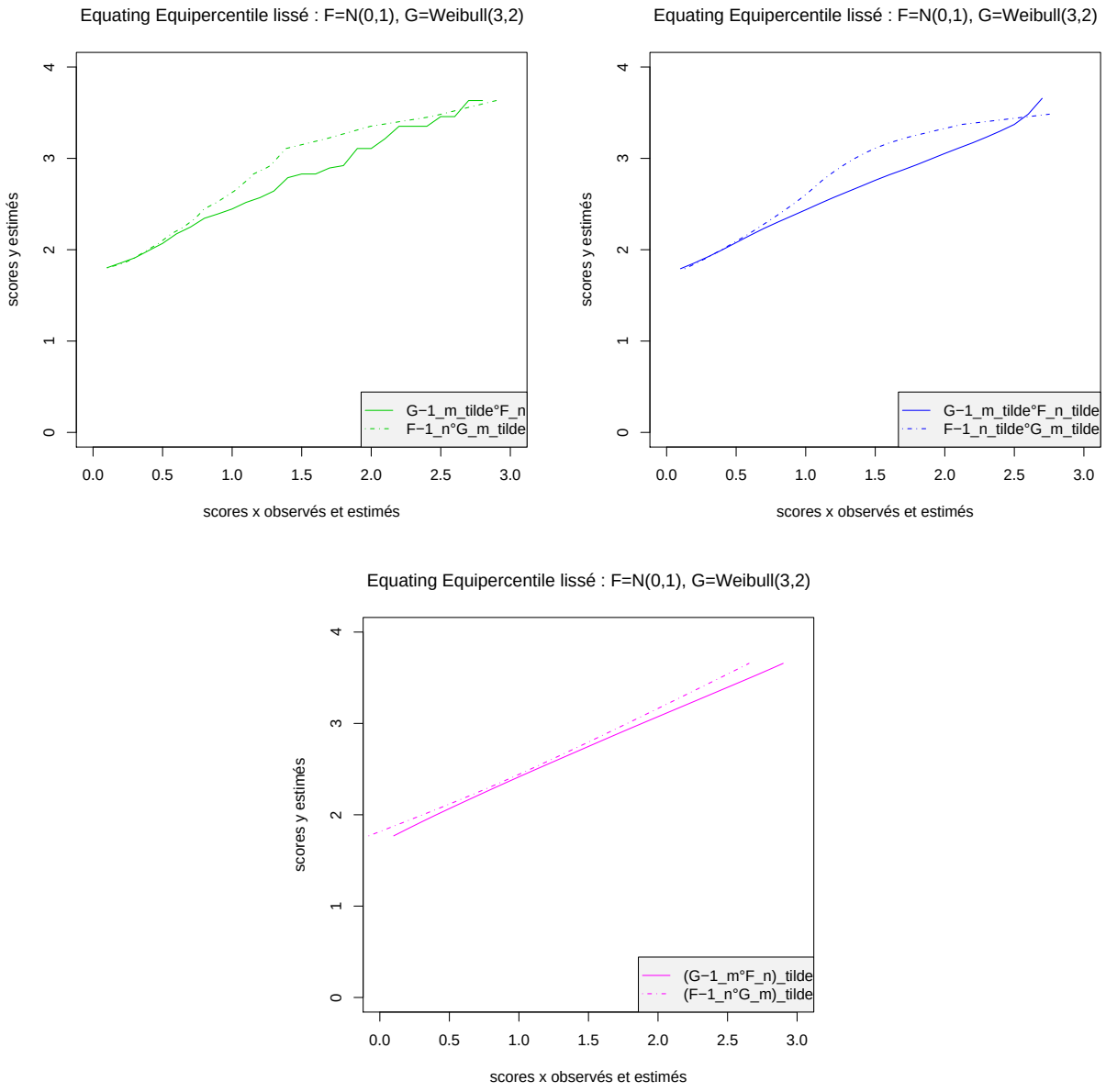


FIG. 7.7 – Comparaison des différents estimateurs de la fonction d'égalisation équipercentile de x en y avec leurs réciproques.

de y en x sont représentés en trait continu et leurs réciproques en trait-point.

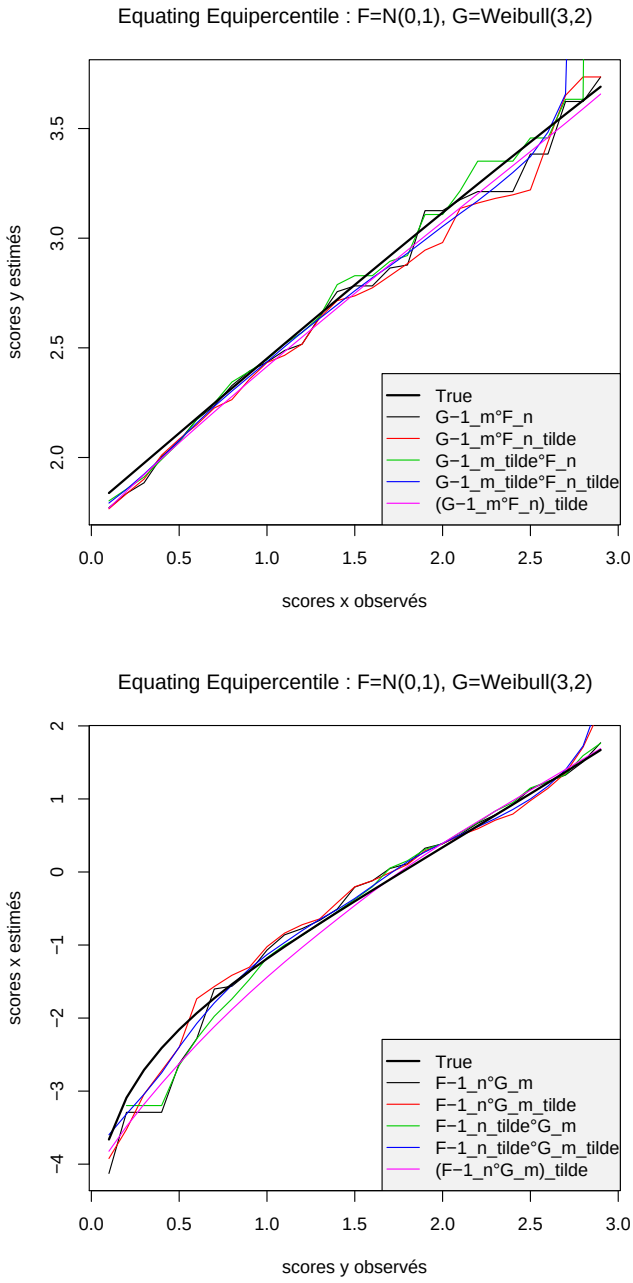


FIG. 7.8 – Estimateurs de la fonction d'égalisation équipercentile de x en y et leurs réciproques.

Cas 3 : $X \sim \mathcal{E}(1)$, $Y \sim Weibull(3,2)$, n=250, m=300 et $[a,b] = [0,3]$.

Il est facile de lire le score équivalent y de x via le graphe ci-dessous. Par exemple, pour $x = 2$ nous lisons la valeur correspondante telle que $\widetilde{F_n}(x) = \widetilde{G_m}(y)$.

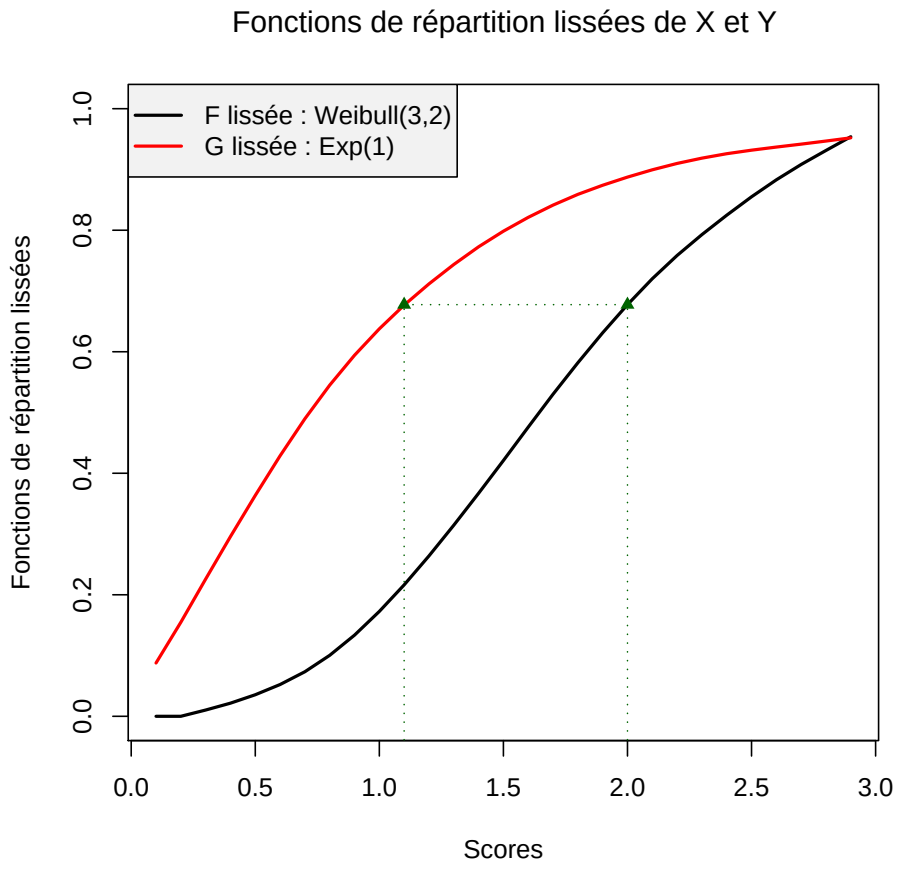


FIG. 7.9 – Fonctions de répartition de $Weibull(3,2)$ et $\mathcal{E}(1)$.

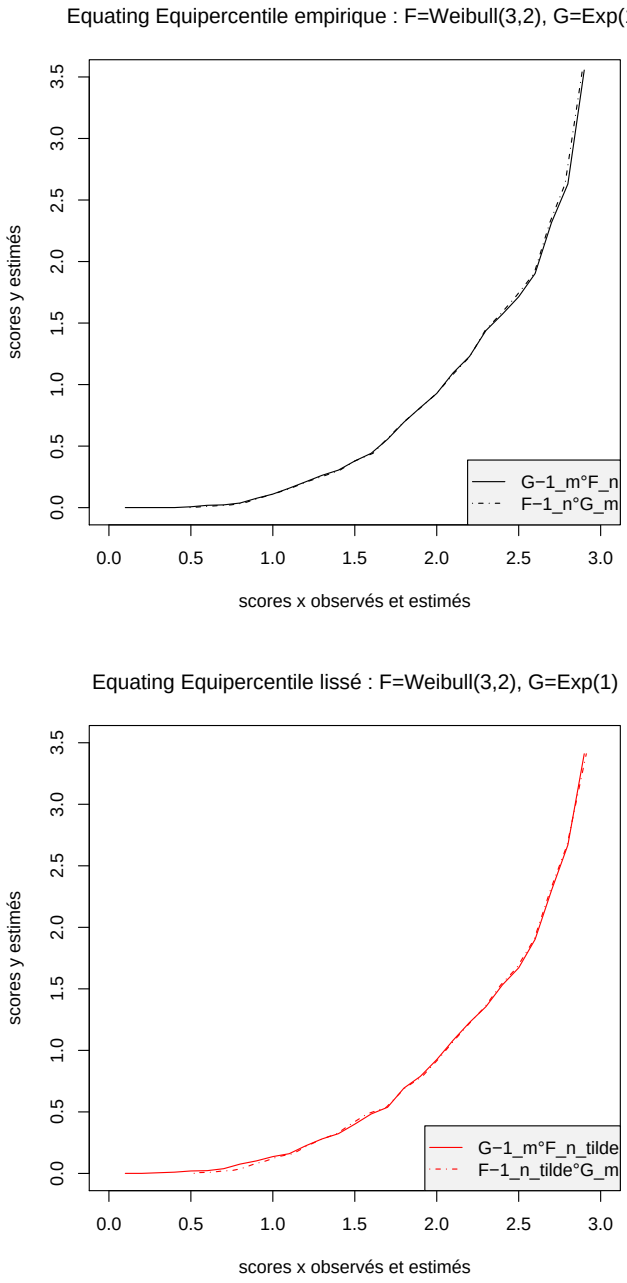


FIG. 7.10 – Comparaison des différents estimateurs de la fonction d'égalisation équipercentile de x en y avec leurs réciproques.

Dans les figures (7.10) et (7.11), les graphes en trait continu représentent les estimateurs

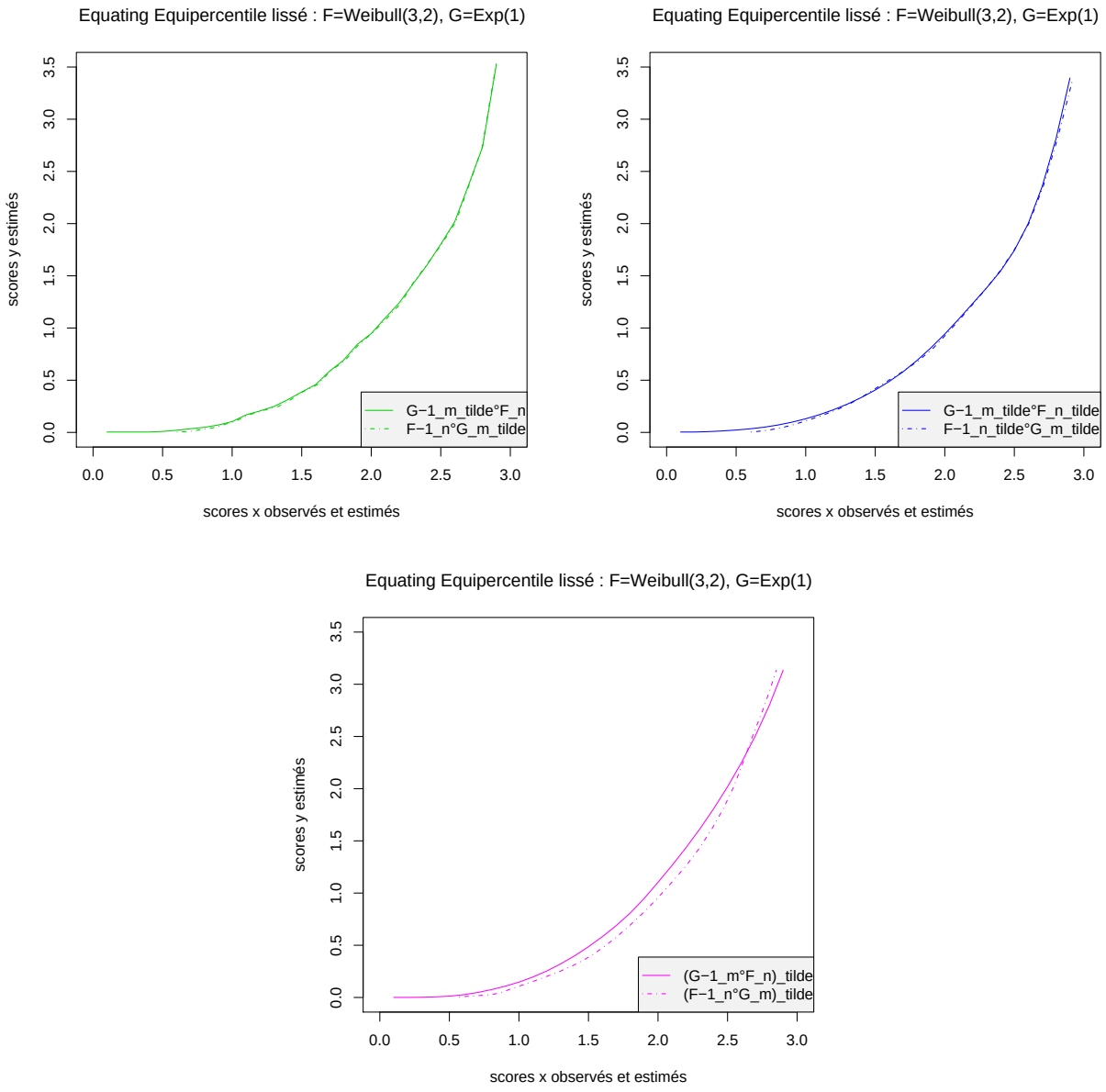


FIG. 7.11 – Comparaison des différents estimateurs de la fonction d'égalisation équipercentile de x en y avec leurs réciproques.

de la fonction d'égalisation équipercentile de y en x. Les réciproques de ces derniers sont représentés en trait-point.

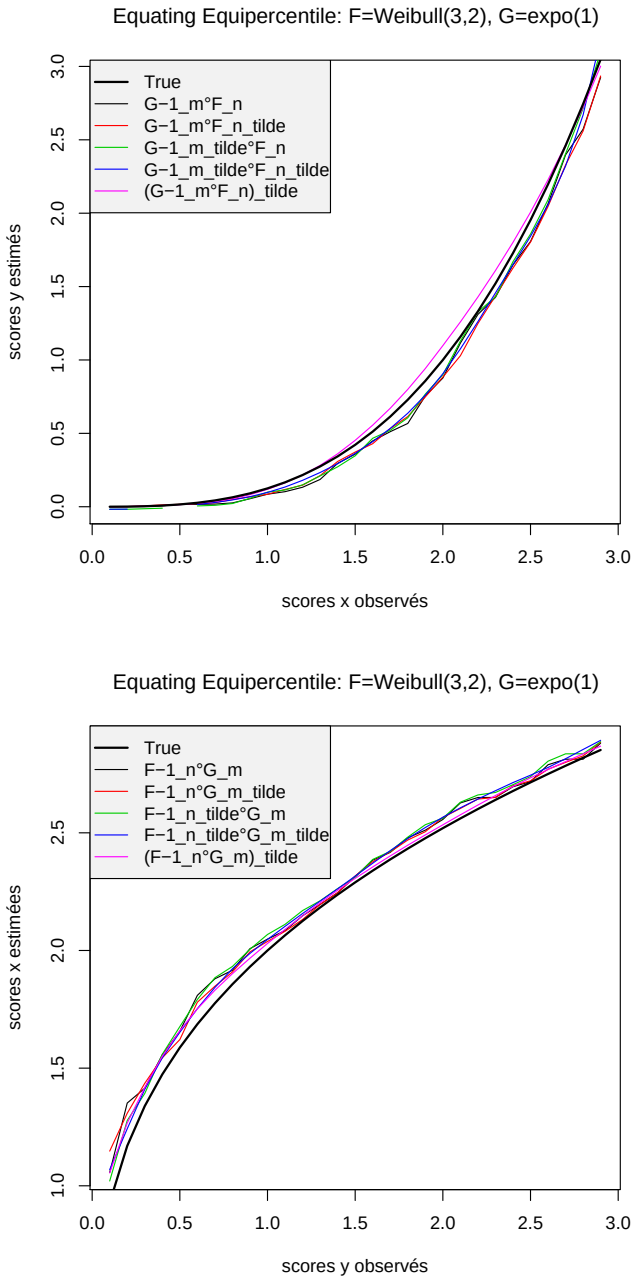


FIG. 7.12 – Estimateurs de la fonction d'égalisation équipercentile de x en y et leurs réciproques.

Les graphiques mettent en évidence les bonnes performances des cinq estimateurs utilisés. En particulier, dans le dernier cas nous obtenons de bonnes estimations sur les bords du support où typiquement l'estimation pose problème. Les estimateurs ont des performances assez proches mais l'estimateur $y^{[5]}_{n,m}(x) = \widetilde{G_m^{-1}(F_n(x))}$ semble l'emporter sur les autres estimateurs dans la plupart des cas quant à l'égalisation de x en y. Pour l'égalisation des scores y en x, c'est l'estimateur $x^{[5]}_{n,m}(y) = \widetilde{F_n^{-1}(G_m(y))}$ qui l'emporte.

Chapitre 8

Application aux scores de patients atteints par le VIH

Contexte d'étude

Nous disposons des scores de 233 patients atteints par le VIH sur deux échelles de mesure différentes (questionnaires). La première est une échelle de référence et s'appelle Short Form 12 (SF-12® (2002)) et la deuxième est une échelle spécifique pour l'étude des patients infectés par le VIH (WHOQOL-HIV (2004)). Ces données de scores sont produites et communiquées par l'Inserm de Marseille. Elles proviennent d'une étude longitudinale multi-centrique de la cohorte ANRS C08.

Le questionnaire SF-12 est une version abrégée du SF-36 (SF-36® (1992)) ne comportant que 12 questions sur les 36. Il permet de mesurer huit aspects de la qualité de vie : état de santé général et mental, fonctionnement physique et social, santé physique et émotionnelle, douleur et vitalité.

Durant le suivi des patients, deux questionnaires de qualité de vie sont donnés à remplir aux patients. Les réponses obtenues ont donné lieu ensuite au calcul de deux scores, un pour chaque questionnaire. Les auto-questionnaires ont changé en cours d'étude. En effet,

durant les cinq premières années, le questionnaire de qualité de vie est le SF-12, tandis que durant les quatre dernières année, le questionnaire est le WHOQOL-HIV. Ce remplacement de questionnaire s'explique par le fait que le SF-12, destiné à l'évaluation des services, des besoins et des traitements, adopte le point de vue associé des experts et des patients. Il s'agit donc plus de mesure de statut de santé ou de qualité de vie liée à la santé que de qualité de vie perçue par les patients eux-mêmes (voir e.g. Leplège et Coste (2001)). Ainsi l'échelle de qualité de vie générique SF-12 validée et largement utilisée chez les patients infectés par le VIH a fini par montrer ses limites. Elle ne permet pas de différencier précisément les différents états de qualité de vie physique et ne prend pas en compte la rapport à la sexualité, par exemple. L'échelle WHOQOL-HIV est issue d'une échelle générique de qualité de vie à laquelle il a été rajouté quelques questions spécifiques des problématiques de l'infection par le VIH. Ainsi, le WHOQOL-HIV est un outil spécifique de mesure de la qualité de vie des patients infectés par le VIH (WHOQOL-HIV (2004)) (voir la thèse de Boisson (2008) pour la description de la base de données où elle est décrite en détail).

Notre but donc est de construire une échelle de correspondances entre les deux scores au moyen de l'égalisation équipercentile à partir des cinq estimateurs étudiés.

Le but principal des cliniciens est de pouvoir "traduire" (égaliser) les scores agrégés de l'échelle WHOQOL-HIV en scores de l'échelle SF-12. Pour ce faire, nous construisons une échelle de correspondence entre les deux scores basés sur les estimateurs étudiés dans les chapitres précédents. Plus précisément, les scores de l'échelle SF-12 ont été estimés par la fonction d'égalisation équipercentile à partir des scores de l'échelle WHOQOL-HIV.

Dans la suite, X et Y désignent les vecteurs des scores issus des questionnaires "WHOQOL-HIV" et "SF-12" respectivement.

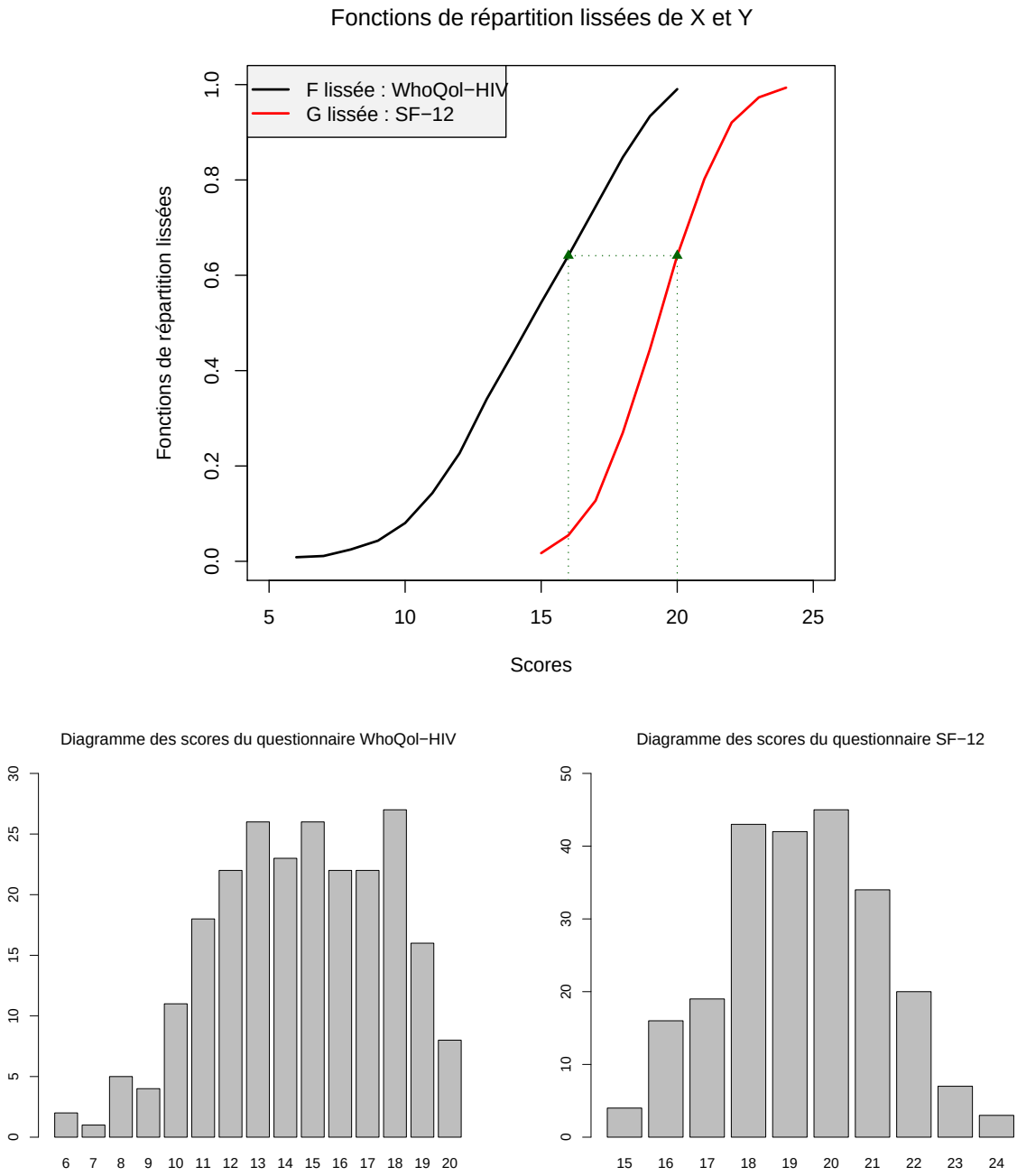


FIG. 8.1 – Distribution des scores des questionnaires “WHOQOL-HIV” et “SF-12”.

Pour chacune des 17 valeurs du vecteur X : 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19 et 20, on recherche une valeur correspondante de Y. Le problème posé est compliqué puisqu'il dépend essentiellement des choix optimaux du paramètre de lissage. L'utilisation des expressions théoriques n'est pas possible car F et G^{-1} sont inconnues. L'étude de méthodes alternatives reste un problème ouvert qui n'est pas abordé dans notre travail. Dans cette application, nous prenons le parti de n'utiliser qu'une seule valeur (arbitraire mais de l'ordre $n^{-1/3}$) des paramètres de lissage. On prendra donc $h = 233^{-1/3} \approx 0.16$.

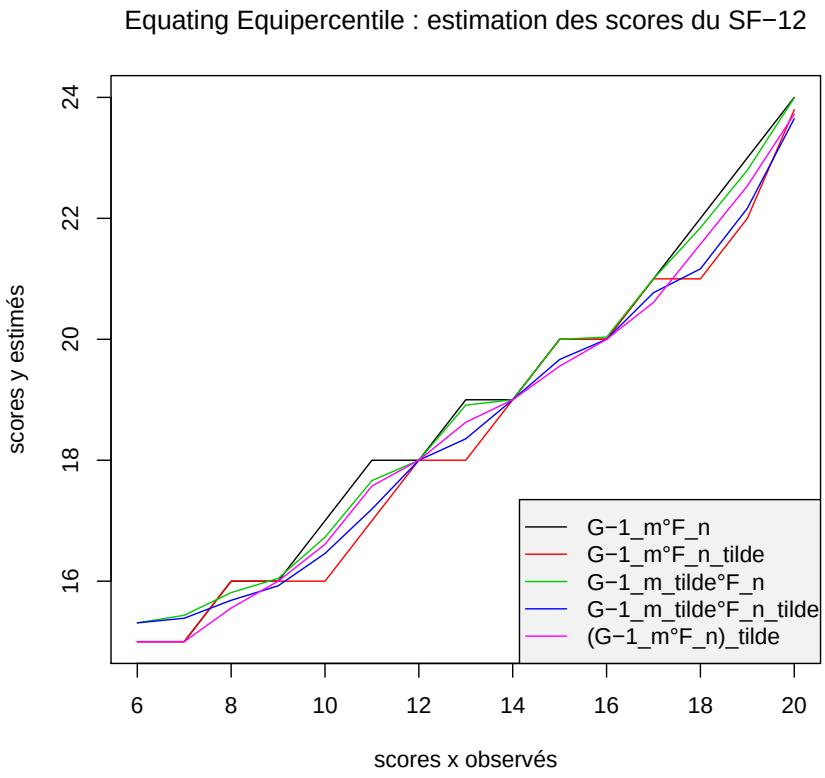


FIG. 8.2 – Estimation des scores du questionnaire "SF-12 à partir des scores du WhoQol".

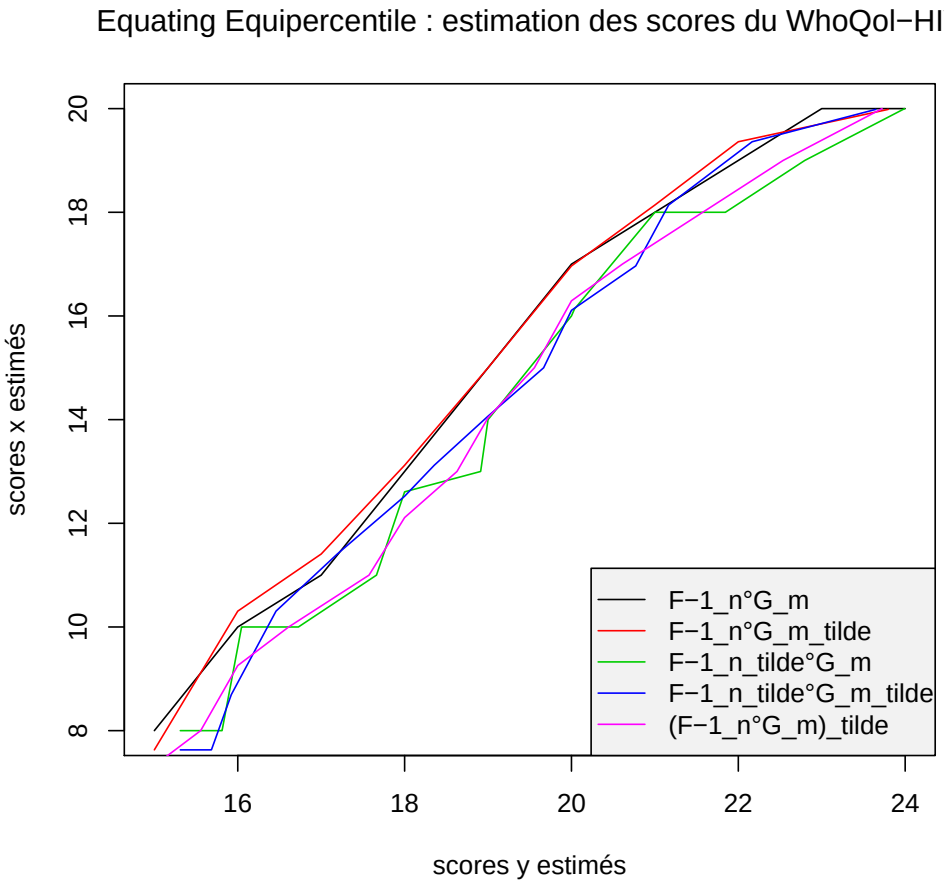


FIG. 8.3 – Estimation des scores du WhoQol-HIV à partir des scores estimés du “SF-12”.

Les figures (8.2) et (8.3) font apparaître des comportements comparables de nos estimateurs. Pour des choix mieux adaptés des valeurs de h, on pourrait espérer des résultats encore meilleurs.

Dans la suite, nous présentons les estimateurs permettant d'obtenir l'égalisation équipercentile des scores WhoQol-HIV en scores SF-12 et inversement.

Nous estimons les scores de SF-12 ($y_{n,m}^{[j]}$ pour $j = 1, \ldots, 5$) à partir des scores observés x_i de WhoQol-HIV. Ensuite, nous estimons les scores de WhoQol-HIV ($x_{n,m}^{[j]}$ pour $j = 1, \ldots, 5$) à partir des scores estimés de SF-12 ($y_{n,m}^{[j]}$). Le but est voir si l'on retombe sur les scores

x_i de WhoQol-HIV de départ afin de voir si les estimateurs vérifient bien la propriété de la symétrie.

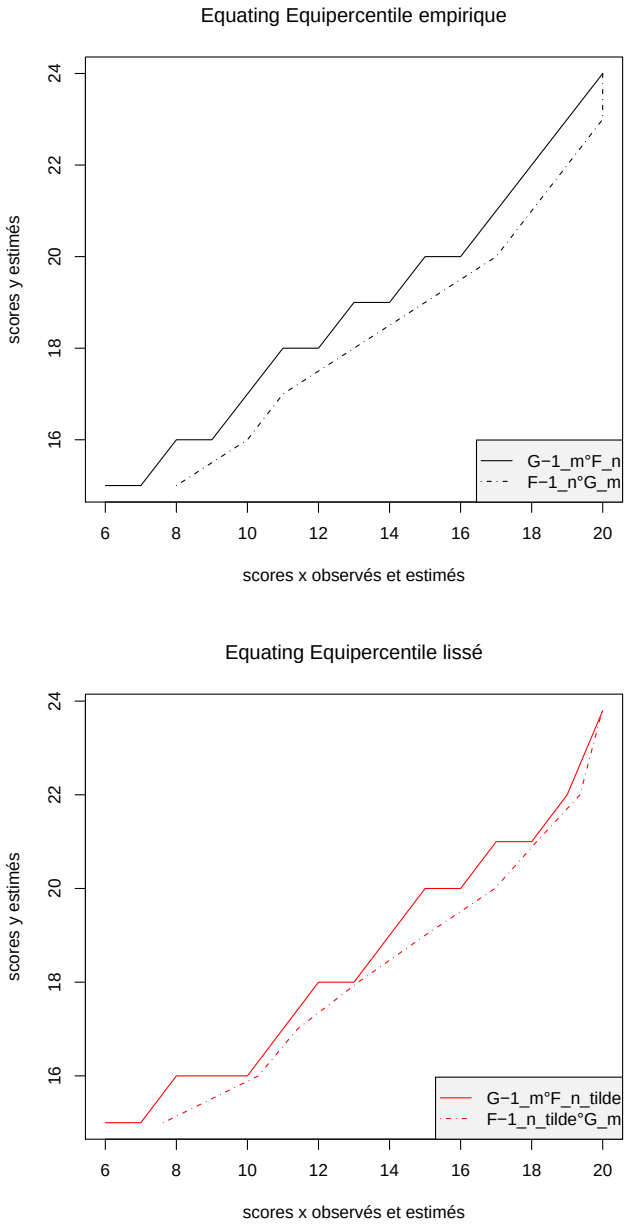


FIG. 8.4 – Estimateurs $y^{[1]}_{n,m}$ et $y^{[2]}_{n,m}$ de la fonction d'égalisation équipercentile des scores de WhoQol-HIV en scores SF-12 avec leurs réciproques.

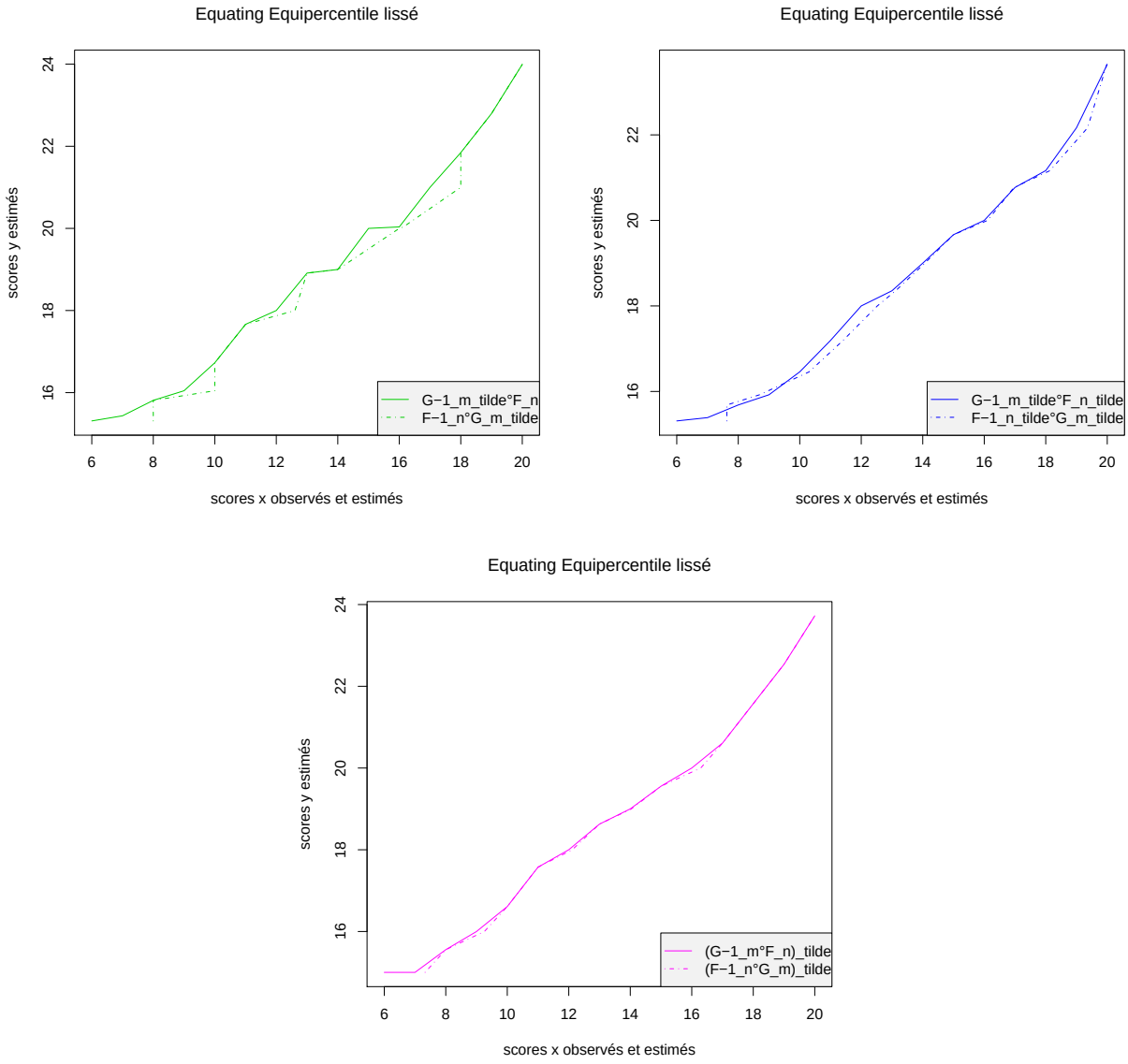


FIG. 8.5 – $y^{[3]}_{n,m}$, $y^{[4]}_{n,m}$ et $y^{[5]}_{n,m}$ de la fonction d'égalisation équipercentile des scores de WhoQol-HIV en scores SF-12 avec leurs réciproques.

En comparant les cinq derniers graphes (Figures (8.4) et (8.5)), nous constatons que l'estimateur $y^{[5]}_{n,m}$ donne les meilleures estimations tout en respectant le mieux la propriété de symétrie.

Échelle de correspondance des scores WhoQol-HIV en Scores SF-12

x_i	$y_{n,m}^{[1]}$	$y_{n,m}^{[2]}$	$y_{n,m}^{[3]}$	$y_{n,m}^{[4]}$	$y_{n,m}^{[5]}$
6	15	15.0	15.31	15.31	15.00
7	15	15.0	15.44	15.39	15.00
8	16	16.0	15.81	15.68	15.56
9	16	16.0	16.05	15.92	16.00
10	17	16.0	16.73	16.46	16.61
11	18	17.0	17.66	17.19	17.57
12	18	18.0	18.00	18.00	18.00
13	19	18.0	18.91	18.35	18.63
14	19	19.0	19.00	19.00	19.00
15	20	20.0	20.00	19.67	19.56
16	20	20.0	20.04	20.00	20.00
17	21	21.0	21.00	20.77	20.61
18	22	21.0	21.84	21.16	21.57
19	23	22.0	22.80	22.17	22.54
20	24	23.8	23.99	23.65	23.72

Comparaison des scores observés du WhoQol avec les scores estimés du WhoQol.

x_i	$x_{n,m}^{[1]}$	$x_{n,m}^{[2]}$	$x_{n,m}^{[3]}$	$x_{n,m}^{[4]}$	$x_{n,m}^{[5]}$
6	8.00	7.63	8.00	7.63	7.33
7	8.00	7.63	8.00	7.63	7.33
8	10.00	10.31	8.00	7.63	8.00
9	10.00	10.31	10.00	8.69	9.26
10	11.00	10.31	10.00	10.31	10.00
11	13.00	11.41	11.00	11.41	11.00
12	13.00	13.12	12.61	12.52	12.11
13	15.00	13.12	13.00	13.12	13.00
14	15.00	15.00	14.00	14.07	14.04
15	17.00	16.97	16.00	15.00	15.00
16	17.00	16.97	16.13	16.11	16.29
17	18.00	18.14	18.00	16.97	17.00
18	19.00	18.14	18.00	18.14	18.00
19	20.00	19.36	19.00	19.36	19.00
20	20.00	19.99	20.00	19.99	20.00

Les tableaux ci-dessus présentent les équivalents des scores x_i pour $i = 1, \ldots, 15$ où x_i représentent les scores du questionnaire WhoQol-HIV, $y_{n,m}^{[j]}$ pour $j = 1, \ldots, 5$ sont les scores estimés du SF-12 à partir des scores x_i et $x_{n,m}^{[j]}$ pour $j = 1, \ldots, 5$ sont les scores estimés du WhoQol-HIV à partir des scores estimés de SF-12 ($y_{n,m}^{[j]}$).

Conclusions et perspectives de recherche

Le point de départ de ce travail est une équation équipercentile posée par le statisticien dans le but d'estimer des scores y sur une échelle de mesure donnée à partir d'autres scores x d'une échelle différente ou inversement. Dans cette thèse, nous avons proposé d'étudier cinq estimateurs de la fonction d'égalisation équipercentile. Ces estimateurs ont été obtenus par l'approche de l'ajustement par des polynômes locaux.

Les principaux résultats que nous avons obtenus sont les suivants. Dans un premier temps, nous avons montré la convergence uniforme presque sûre de ces estimateurs. Ce résultat est très important pour la mise en œuvre des applications statistiques. Ensuite, nous avons établi l'approximation par un pont brownien approprié des processus construits à partir des estimateurs polynomiaux locaux de la fonction d'égalisation. Nous avons également calculé l'erreur en moyenne quadratique des estimateurs, critère de choix de la fenêtre qui s'insère dans le contexte d'estimation non-paramétrique permettant d'équilibrer asymptotiquement le biais et la variance de l'estimateur.

Dans la continuité de ce travail, il serait judicieux de faire une étude quant au choix de la fenêtre en essayant d'adapter certaines méthodes disponibles dans la littérature statistique

("plug-in", validation croisée, etc..) dans notre contexte de l'estimation non-paramétrique de la fonction d'égalisation équipercentile.

Dans ce travail, nous nous sommes limités au plan d'expérience des groupes équivalents. Il serait donc intéressant de généraliser ces résultats théoriques afin de pouvoir prendre en compte la dépendance entre les groupes d'individus.

Un axe de recherche qu'il serait intéressant d'explorer, est l'application de notre procédure non-paramétrique à des problèmes communément rencontrés dans l'analyse de survie et en fiabilité où les approches sont souvent paramétriques ou semi-paramétriques. En effet, l'analyse de survie occupe maintenant une place importante dans la modélisation statistique et trouve un intérêt dans de nombreux domaines d'applications.

Nous avons implémenté en R les estimateurs proposés afin d'obtenir pour chaque score donné sur une échelle de mesure son équivalent sur une autre échelle. Une étude par simulations montre que les estimateurs fonctionnent bien. Dans le but de peaufiner l'application sur les données des patients infectés par le VIH, nous envisageons d'étudier des critères de validation sur la précision de nos estimateurs.

Annexe

Programmes informatiques

Dans cette section, nous présentons les programmes informatiques nous permettant de calculer les estimateurs sur des données simulées. Il s'agit de trouver les scores y à partir de scores x et inversement. Le paramètre de lissage varie en fonction des lois simulées. Nous présentons ici le cas de la normale versus Weibull. Les autres cas de figure qui sont traités dans nos simulations s'obtiennent en adaptant des versions de la macro "*opt_bw*"pour les lois appropriées, en l'occurrence Weibull versus Weibull (avec des paramètres différents) et gaussienne versus exponentielle.

```
##############################################################################
############      Fonctions arithmétiques pour des polynômes      ############
##############################################################################
# Evaluation d'un polynôme en y dans [-1,1] #
    poly.eval <- function(p,y){
d <- length(p)-1
sapply(y,function(yy)sum(p*yy^(0:d)))
    }
# Intégration d'un polynôme #
    poly.integr<-function(p){
d <- length(p)-1
```

```
c(0,p/(1:(d+1)))
    }

# Multiplication d'un polynôme par un monôme #
# p est le polynôme et j l'ordre du monôme   #
    poly.multmon<-function(p,j){
c(rep(0,j),p)
    }
# Addition de deux poly #
    poly.add <- function(p1,p2){
d1 <- length(p1)-1
d2 <- length(p2)-1
d  <- max(d1,d2)
p1 <- c(p1,rep(0,d-d1))
p2 <- c(p2,rep(0,d-d2))
p1+p2
    }
# j-ième moment de K #
    mu <- function(j,K){
p  <- poly.multmon(K,j)
p  <- poly.integr(p)
poly.eval(p,1)-poly.eval(p,-1)
    }
# Multiplication de 2 polynômes p1 et p2 #
    mult.polynom <- function(p1,p2){
              n1 <- length(p1)
              if(n1==1){
```

```
                prod <- p1[1]*p2}
            else{
                prod <- p1[1]*p2
            for(i in 2:n1){
            temp <- p1[i]*poly.multmon(p2,i-1)
            prod <- poly.add(temp,prod)
              }
            }
    prod
    }
##############################################################################
#######################       END FUNCTIONS ANNEXES       ####################
##############################################################################
##############################################################################
# Fonction opt_bw                                                            #
# F: Weibull(lambda,kappa)                                                   #
# Q: N(0,1)                                                                  #
# lissage=1,2 et 3                                                           #
##############################################################################
opt_bw <- function(x,ech_x,ech_y,lissage,r){
lambda=3 # Paramètres de la Weibull #
kappa=2
bwF=NA
bwQ=NA
bwQF=NA
K   <- c(3/4,0,-3/4)
K01 <- K
```

```
n   <- length(ech_x)
m   <- length(ech_y)
Kr=K01
nuI <- mu(2*r,Kr)



# On calcule la fdr KKr du Noyau Kr à support [-1,1] #
temp <- poly.integr(Kr)
cste <- poly.eval(temp,-1)
KKr  <- poly.add(temp,-cste)

# Produit u*Kr*KKr et on intègre sur [-1,1] #
ftemp  <- mult.polynom(Kr,KKr)
sigmaI <- 2*mu(1,ftemp)

# Bandwidth of F (Normale) #
if(lissage==1){
der1_F <- dnorm(x,0,1)
der2_F <- DD(expression(pnorm(x,0,1)),"x",2)
der_x  <- der1_F
numerator   <- dnorm(x,0,1)*sigmaI
denominator <- n*((4*r)/(factorial(2*r))^2)*(nuI*der_x)^2
ratio       <- (numerator/denominator)
bwF         <- ratio^(1/3)
bwF
} # end case lissage==1 #
```

```
# Bandwidth of Q (Weibull) #
if(lissage==2){
if(x<0 | x>=1){return(NA)}
der1_Q <- DD(expression(lambda*(log(1/(1-x)))^(1/kappa)),"x",1)
der2_Q <- DD(expression(lambda*(log(1/(1-x)))^(1/kappa)),"x",2)
der_x  <- eval(der2_Q)
numerator   <- (eval(der1_Q)^2)*sigmaI
denominator <- m*((4*r)/(factorial(2*r))^2)*(nuI*der_x)^2
ratio       <- (numerator/denominator)
bwQ         <- ratio^(1/3)
bwQ
} # end case lissage==2 #

# Bandwidth of Q°F #
if(lissage==3){
der2_QF <- DD(expression(lambda*(log(1/(1-pnorm(x,0,1))))^(1/kappa)),"x",2)
der_x   <- eval(der2_QF)
numerator   <- dnorm(x,0,1)*((lambda/kappa)*(1/(1-pnorm(x,0,1)))
               *(log(1/(1-pnorm(x,0,1))))^(1/kappa-1))^2*sigmaI
denominator <- ((4*r)/(factorial(2*r))^2)*(nuI*der_x)^2
ratio       <- ((m+n)/(n*m))*(numerator/denominator)
bwQF        <- ratio^(1/3)
bwQF
} # end case lissage==3 #
return(list(bwF=bwF,bwQ=bwQ,bwQF=bwQF))
} # end function opt_bw #
##############################################################################
```

```
##############################################################################
# Fonction noyau                                                             #
# On définit l'expression des noyaux selon x, l'échantillon X et h           #
# La fonction noyau nécessite les fonctions poly.add, mu et poly.multmon     #
##############################################################################
noyau <- function(x,a,b,h){
alpha="null"
pos="center"
if (x<a | x>b) stop("x en dehors de l'intervalle [a,b]")
K   <- c(3/4,0,-3/4)
K01 <- K
if(a<=x & x<a+h){
    pos="left"
    alpha=(x-a)/h
    den1=((1+alpha)^4)*(3*(alpha^2)-18*alpha+19)
    num1_L=-16*c(12*(alpha^3)-24*(alpha^2)+16*alpha-8,15*(1-alpha)^2)
    polynom1_L=num1_L/den1
    K01=.75*poly.add(polynom1_L,-poly.multmon(polynom1_L,2))
} # end if à gauche #
if(b-h<x & x<=b){
    pos="right"
    alpha=(b-x)/h
    den1=((1+alpha)^4)*(3*(alpha^2)-18*alpha+19)
    num1_R=-16*c(12*(alpha^3)-24*(alpha^2)+16*alpha-8,-15*(1-alpha)^2)
    polynom1_R=num1_R/den1
    K01=.75*poly.add(polynom1_R,-poly.multmon(polynom1_R,2))
} # end if à droite #
```

```
return(list(K01=K01,alpha=alpha,position=pos))
} # END FUNCTION NOYAU #
##############################################################################
# Fonction a0                                                                #
# Estimation par polynômes locaux au point x                                 #
# La fonction a0 nécessite les fonctions noyau et poly.integr                #
# lissage==1 --> Fn en x lissé                                               #
# lissage==2 --> Qm en x lissé                                               #
# lissage==3 --> Qm°Fn en x lissé                                            #
##############################################################################
###############################     function a0     ##########################
a0 <- function(x,ech_x,ech_y,a,b,h,N,lissage){
    ech=ech_x
    if (lissage==2){
    a=0
    b=1
    ech=ech_y
    }
if(x<a | x>b) stop("x en dehors de l'intervalle [a,b]")
if(h>(b-a)/2){h <- (b-a)/2}
kern  <- noyau(x,a,b,h)
alpha <- kern$alpha
K     <- kern$K01
pos   <- kern$position
Z     <- a+(0:N)*(b-a)/N
KK    <- poly.integr(K)
S     <- sapply(x,function(x){
```

```
if(is.na(x)) return(NA)
# trouvons i et j tels que
# a <= Z[i] <= Z[j] < x+h <= Z[j+1]
indx <- 1:length(Z)

# Bornes d'encadrement pour la grille de Z suivant la position de x #
BL=x-h
BR=x+h
if(pos=="right") BR=x+alpha*h
if(pos=="left") BL=x-alpha*h
i <- min(indx[BL<Z])
j <- max(indx[Z<BR])
# si pas d'obs dans l'intervalle alors NA
if(is.na(i) | is.na(j)) return(NA)
# les Z dans l'intervalle plus deux bornes
Z <- c(BL,Z[i:j],BR)
# les valeurs dans l'intervalle
if(lissage==1){
Phi=ecdf(ech)
Phi <- Phi(Z)
}
if(lissage==2){
Phi=quantile(ech,probs=Z)
}
if(lissage==3){
Psi <- ecdf(ech)
Phi=quantile(ech_y,probs=Psi(Z))
```

```
}
S <- sum(diff(poly.eval(KK,(Z-x)/h))*Phi[-1])
if(lissage==1 && S<0){S=0} else if(lissage==1 && S>1){S=1}
S
}) # end sapply #
S
} # end function a0 #
##############################################################################
##############################################################################
# Fonction estim_y                                                           #
# La fonction estim_y nécessite la fonction a0                               #
# option==1 --> Qm°Fn sans lissage                                           #
# option==2 --> Qm°Fn_tilde                                                  #
# option==3 --> Qm_tilde°Fn                                                  #
# option==4 --> Qm_tilde°Fn_tilde                                            #
# option==5 --> Qm°Fn avec lissage                                           #
##############################################################################
############################   function estim_y   ############################

estim_y <- function(x,ech_x,ech_y,a,b,N,option){
if(x<a | x>b) stop("x en dehors de l'intervalle [a,b]")
if(option==1){
Psi <- ecdf(ech_x)
yhat=quantile(ech_y,probs=Psi(x))
} # endif option1 #

if(option==2){
```

```
bw <- opt_bw(x,ech_x,ech_y,1,1)
h=bw$bwF
yhat=quantile(ech_y,probs=a0(x,ech_x,ech_y,a,b,h,N,1))
} # endif option2 #

if(option==3){
Psi <- ecdf(ech_x)
Psi_x <- Psi(x)
if(Psi_x==0){Psi_x <- Psi_x+1E-4}
if(Psi_x==1){Psi_x <- Psi_x-1E-4}
bw <- opt_bw(Psi_x,ech_x,ech_y,2,1)
h=bw$bwQ
yhat=a0(Psi(x),ech_x,ech_y,a,b,h,N,2)
} # endif option3 #

if(option==4){
bw1 <- opt_bw(x,ech_x,ech_y,1,1)
h=bw1$bwF
Fn_tilde=a0(x,ech_x,ech_y,a,b,h,N,1)
if(Fn_tilde==1){Fn_tilde <- Fn_tilde-1E-4}
if(Fn_tilde==0){Fn_tilde <- Fn_tilde+1E-4}
bw2 <- opt_bw(Fn_tilde,ech_x,ech_y,2,1)
k=bw2$bwQ
yhat=a0(Fn_tilde,ech_x,ech_y,a,b,k,N,2)
} # endif option4 #

if(option==5){
```

```
bw <- opt_bw(x,ech_x,ech_y,3,1)
h=bw$bwQF
yhat=a0(x,ech_x,ech_y,a,b,h,N,3)
} # endif option5 #
yhat=as.numeric(yhat)
return(list(yhat=yhat))
} # end function estim_y #
################################################################################
##########################       END FUNCTIONS       ##########################
################################################################################
```

Bibliographie

Abdous, B., Berlinet, A., et Hengartner, N. (2003). A general theory for kernel estimation of smooth functionals of the distribution function and their derivatives. *Rev. Roumaine Math. Pures Appl.*, **48**(3), 217–232.

Aerts, M., Augustyns, I., et Janssen, P. (1997). Sparse consistency and smoothing for multinomial data. *Statist. Probab. Lett.*, **33**(1), 41–48.

Aly, E.-E. A. A. (1986). Strong approximations of the Q-Q process. *J. Multivariate Anal.*, **20**(1), 114–128.

Aly, E.-E. A. A., Csörgő, M., et Horváth, L. (1987). P-P plots, rank processes, and Chernoff-Savage theorems. In *New perspectives in theoretical and applied statistics (Bilbao, 1986)*, Wiley Ser. Probab. Math. Statist. Probab. Math. Statist., pages 135–156. Wiley, New York.

Angoff, W. H. (1971). *Scales, norms, and equivalent scores.* In R. L. Thorndike (Ed.), Educational measurement. American Council on Education., Washington, DC. A Wiley-Interscience Publication.

Angoff, W. H. (1987). Technical and practical issues in equating : A discussion of four papers. *Applied Psychological Measurement*, **11**, 291–300.

Arnold, B. C., Balakrishnan, N., et Nagaraja, H. N. (2008). *A First Course in Order Statistics (Classics in Applied Mathematics).* SIAM.

Bagdonavicius, V. et Nikulin, M. (2001). Mathematical models in the theory of accelerated experiments. *World Scientific, Cairo University, Egypt.*, **Mathematics and the 21st Century.**, 271–273.

Bagdonavicius, V. et Nikulin, M. (2002). *Accelerated life models : modeling and statistical analysis*. Editor : London , Chapman & Hall.

Bagdonavicius, V. et Nikulin, M. (2007). Statistical methods to analyze failures, wear and degradation with applications to accelerated testing and quality control : An engineering perspective. *Mathematical Methods in Reliability, 1-4 July 2007.*, pages 14–16.

Beirlant, J. et Deheuvels, P. (1990). On the approximation of P-P and Q-Q plot processes by Brownian bridges. *Statist. Probab. Lett.*, **9**(3), 241–251.

Boisson, V. (2008). *Test de type log-rank pour l'évolution longitudinale de la qualité de vie liée à la santé.*

Bol′šev, L. N. (1959). On transformations of random variables. *Theor. Probability Appl.*, **4**, 129–141.

Bol′šev, L. N. (1963). Asymptotic Pearson transformations. *Teor. Verojatnost. i Primenen.*, **8**, 129–155.

Braun, H. I. et Holland, P. (1982). *Observed-score test equating : A mathematical analysis of some ETS equating procedures.* In P. W. Holland and D.B Rubin (Eds.), Test equating (9–49)., New York : Academic.

Burman, P. (1987). Smoothing sparse contingency tables. *Sankhyā Ser. A*, **49**(1), 24–36.

Butler, O. D. et Hanson, B. A. (1997). *An examination of presmoothing and postsmoothing methods in equating a direct writing assessment.* Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, March).

Cantelli, F. P. (1933). Sulla determinazione empirica delle leggi di probabilita. *Giorn. Ist. Ital. Attuari.*, **4**, 421–424.

Cheng, C. et Parzen, E. (1997). Unified estimators of smooth quantile and quantile density functions. *J. Statist. Plann. Inference*, **59**(2), 291–307.

Cheng, M.-Y. et Peng, L. (2002). Regression modeling for nonparametric estimation of distribution and quantile functions. *Statist. Sinica*, **12**(4), 1043–1060.

Cheng, M.-Y., Fan, J., et Marron, J. S. (1997). On automatic boundary corrections. *Ann. Statist.*, **25**(4), 1691–1708.

Chung, K.-L. (1949). An estimate concerning the Kolmogoroff limit distribution. *Trans. Amer. Math. Soc.*, **67**, 36–50.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *J. Amer. Statist. Assoc.*, **74**(368), 829–836.

Csörgő, M. (1983). *Quantile Processes with Statistical Applications.* SIAM, Philadelphia.

Csörgő, M. et Horváth, L. (1993). *Weighted approximations in probability and statistics.* Wiley Series in Probability and Mathematical Statistics : Probability and Mathematical Statistics. John Wiley & Sons Ltd., Chichester. With a foreword by David Kendall.

Csörgő, M. et Révész, P. (1978). Strong approximations of the quantile process. *Ann. Statist.*, **6**(4), 882–894.

Csörgő, M. et Révész, P. (1981). *Strong approximations in probability and statistics.* Probability and Mathematical Statistics. Academic Press Inc. [Harcourt Brace Jovanovich Publishers], New York.

Csörgő, M., Csörgő, S., Horváth, L., et Mason, D. M. (1986). Weighted empirical and quantile processes. *Ann. Probab.*, **14**(1), 31–85.

Doksum, K. (1974). Empirical probability plots and statistical inference for nonlinear models in the two-sample case. *Ann. Statist.*, **2**, 267–277.

Doksum, K. A. et Sievers, G. L. (1976). Plotting with confidence : graphical comparisons of two populations. *Biometrika*, **63**(3), 421–434.

Dong, J. et Simonoff, J. S. (1995). A geometric combination estimator for d-dimensional ordinal sparse contingency tables. *Ann. Statist.*, **23**(4), 1143–1159.

Dorans, N. J. et Holland, P. W. (2000). Population invariance and the equatability of tests : basic theory and the linear case. *Journal of Educational Measurement*, **37**(2), 281–306.

El Fassi, K. (2005). Equating : Estimation d'un trait latent à partir d'un autre. *Groupe de travail Biostatistique et Variables Latentes. Mesure et Analyse Statistique de la Qualité et des Durées de Vie, Paris.*

El Fassi, K. (2006). Equating two quality of life scales using a multidimensional latent variable model. *International Conference on Statistical Latent Variable Models in Health Sciences, Perugia, Italie.*

El Fassi, K. (2008). Asymptotic theory of local polynomial estimate equipercentile equating function. *European seminar, Bordeaux, France.*

El Fassi, K., Abdous, B., et Mesbah, M. (2008a). Local polynomial fitting of the equipercentile equating function. *Preprint L.S.T.A.*

El Fassi, K., Abdous, B., et Mesbah, M. (2008b). Local polynomial smoothing kernel score equating methods for health related quality of life. *Third International Rasch Measurement Conference, Perth, Western Australia.*

El Fassi, K., Abdous, B., et Mesbah, M. (2009). Ajustement polynomial local de la fonction d'égalisation équipercentile : Convergence uniforme presque sûre. *C. R. Acad. Sci. Paris Sér. I Math.*, **7**(3-4), 195–200.

Fairbank, B. A. (1987). The use of presmoothing and postsmoothing to increase the precision of equating. *Applied Psuchological Measurement*, **11**, 245–262.

Falk, M. (1983). Relative efficiency and deficiency of kernel type estimators of smooth distribution functions. *Statist. Neerlandica*, **37**(2), 73–83.

Falk, M. (1985). Asymptotic normality of the kernel quantile estimator. *Ann. Statist.*, **13**(1), 428–433.

Fan, J. (1992). Design-adaptive nonparametric regression. *J. Amer. Statist. Assoc.*, **87**(420), 998–1004.

Fan, J. (1993). Local linear regression smoothers and their minimax efficiencies. *Ann. Statist.*, **21**(1), 196–216.

Fan, J. et Gijbels, I. (1992). Variable bandwidth and local linear regression smoothers. *Ann. Statist.*, **20**(4), 2008–2036.

Fan, J. et Gijbels, I. (1995a). Adaptative order polynomial fitting : bandwidth robustification and bias reduction. *Journal of Computational and Graphical Statistics*, **4**(3), 213–227.

Fan, J. et Gijbels, I. (1995b). Data-driven bandwidth selection in local polynomial fitting : variable bandwidth and spatial adaptation. *Journal of the Royal Statistical Society. Series B. Methodological*, **57**(2), 371–394.

Fan, J. et Gijbels, I. (1996). *Local polynomial modelling and its applications*, volume 66 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London.

Fan, J., Farmen, M., et Gijbels, I. (1998). Local maximum likelihood estimation and inference. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **60**(3), 591–608.

Freud, G. (1971). *Orthogonal polynomials*. Pergamon Press, Oxford.

Gibbons, J. D. et Chakraborti, S. (2003). *Nonparametric Statistical Inference, Fourth Edition (Statistics : a Series of Textbooks and Monogrphs)*. CRC.

Glivenko, V. (1933). Sulla determinazione empirica della leggi di probabilith. *Giorn. Ist. Ital. Attuari.*, **IV**, 92–99.

Hall, P. et Titterington, D. M. (1987). On smoothing sparse multinomial data. *Austral. J. Statist.*, **29**(1), 19–37.

Hanson, B. A., Zeng, L., et Colton, D. (1994). *A comparison of presmoothing and postsmoothing methods in equipercentile equating*, volume 94-4. ACT Research Report, Iowa City, IA : American College Testing.

Holland, P. W. et Thayer, D. T. (1987). *Note on the use of log-linear models for fitting discrete probability distributions*. Tech. Rep. No. TR- 87-79, Princeton, N J : Educational Testing Service.

Holland, P. W. et Thayer, D. T. (1989). *The kernel method of equating score distribution*. Tech. Rep. No. TR- 89-7, Princeton, N J : Educational Testing Service.

Holland, P. W. et Thayer, D. T. (2000). Univariate and bivariate loglinear models for discrete test score distributions. *Journal of Educational and Behavioral Statistics*, **25**(2), 133–183.

Holland, P. W. et Wightman, L. E. (1982). *Section pre-equating : A preliminary investigation. In Test Equating.* (edited by P. Holland and D. Rubin). New York : Academic Press.

Hollander, M. et Korwar, R. (1982). Nonparametric bayesian estimation of the horizontal distance between two populations. In *Nonparametrics Statistical Inference I, (eds : B.V. Gnedenko, M.L. Puri and I. Vincze)*, pages 409–416. North-Holland, New York.

Jones, M. C., Marron, J. S., et Sheather, S. J. (1996). A brief survey of bandwidth selection for density estimation. *J. Amer. Statist. Assoc.*, **91**(433), 401–407.

Kolen, M. (1988). Traditional equating methodology. *Educational Measurement : Issues and Practice*, **7**, 29–36.

Kolen, M. (1991). Smoothing methods for estimating test score distributions. *J. Educational Measurement*, **28**(3), 257–282.

Kolen, M. J. et Brennan, R. L. (2004). *Test equating, scaling, and linking : Methods and practices*. New York : Springer.

Komlós, J., Major, P., et Tusnády, G. (1975). An approximation of partial sums of independent RV's and the sample DF. I. *Z. Wahrsch. und Verw. Gebiete*, **32**, 111–131.

Lejeune, M. (1985). Estimation non-paramétrique par noyaux : régression polynomiale mobile. *Rev. Statist. Appl.*, **33**(3), 43–67.

Lejeune, M. et Sarda, P. (1992). Smooth estimators of distribution and density functions. *Comput. Statist. Data Anal.*, **14**(4), 457–471.

Leplège, A. et Coste, J. (2001). Mesure de la santé perceptuelle et de la qualité de vie : méthodes et applications. *Éditions Estem, Paris*.

Li, G., Tiwari, R. C., et Wells, M. T. (1996). Quantile comparison functions in two-sample problems, with application to comparisons of diagnostic markers. *J. Amer. Statist. Assoc.*, **91**(434), 689–698.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, New Jersey : Lawrence Erlbum Associates, Inc.

Lorenz, M. O. (1905). Methods on measuring the concentration of wealth. *.J. Amer. Statist. Assoc.*, **70**, 209–219.

Lu, H. H. S., Wells, M. T., et Tiwari, R. C. (1994). Inference for shift functions in the two-sample problem with right-censored data : with applications. *J. Amer. Statist. Assoc.*, **89**(427), 1017–1026.

Mason, D. M. et van Zwet, W. R. (1987). A note on the strong approximation to the renewal process. *Publ. Inst. Statist. Univ. Paris*, **32**(1-2), 81–91.

Massart, P. (1989). Strong approximation for multivariate empirical and related processes, via KMT constructions. *Ann. Probab.*, **17**(1), 266–291.

Morineau, A., Nakache, J., et Krzyzanowski, C. (1996). *Le modèle log-linéaire et ses applications.* CISIA CERESTA Editeur.

Mushkudiani, N. (2000). *Statistical applications of generalized quantiles.* Nonparametric tolerance regions and P-P plots, Dissertation, Technische Universiteit Eindhoven, Eindhoven, 2000.

Nadaraya, E. A. (1964). Some new estimators for distribution functions. *Theory Probab. Appl.*, **9**, 497–500.

Nair, V. N. (1982). Q-Q plots with confidence bands for comparing several populations. *Scand. J. Statist.*, **9**(4), 193–200.

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

Parzen, E. (1979). Nonparametric statistical data modeling. *J. Amer. Statist. Assoc.*, **74**(365), 105–131. With comments by John W. Tukey, Roy E. Welsch, William F. Eddy, D. V. Lindley, Michael E. Tarter and Edwin L. Crow, and a rejoinder by the author.

Petersen, N. S., Kolen, M., et Hoover, H. D. (1989). *Scaling, norming and equating.* In R. L. Linn (Ed.), Educational measurement (3ed., pp.221–262)., New York : American Council on Education/Macmillan.

R Development Core Team (2008). *R : A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Ruppert, D. et Wand, M. P. (1994). Multivariate locally weighted least squares regression. *Ann. Statist.*, **22**(3), 1346–1370.

Sedyakin, N. (1966). On one physical principle in reliability theory. *Engrg. Cybernetics,*, **3**, 80–87.

Serfling, R. J. (1980). *Approximation theorems of mathematical statistics*. John Wiley & Sons Inc., New York. Wiley Series in Probability and Mathematical Statistics.

SF-12® (2002). Health survey ©. *By Medical Outcomes Trust and Quality Metric Incorporated.*

SF-36® (1992). Health survey ©. *Medical Outcomes Study Short-Form General Health Survey.*

Shorack, G. R. et Wellner, J. A. (1986). *Empirical processes with applications to statistics*. Wiley Series in Probability and Mathematical Statistics : Probability and Mathematical Statistics. John Wiley & Sons Inc., New York.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman & Hall. Monographs on Statistics and Applied Probability.

Simonoff, J. S. (1995). Smoothing categorical data. *J. Statist. Plann. Inference*, **47**(1-2), 41–69. Statistical modelling (Leuven, 1993).

Stone, C. J. (1977). Consistent nonparametric regression. *Ann. Statist.*, **5**(4), 595–645. With discussion and a reply by the author.

Stone, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *Ann. Statist.*, **8**(6), 1348–1360.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.*, **10**(4), 1040–1053.

Switzer, P. (1976). Confidence procedures for two-sample problems. *Biometrika*, **63**(1), 13–25.

Tapia, R. A. et Thompson, J. R. (1978). *Nonparametric Probability Density Estimation.* Baltimore, MD : The Johns Hopkins University Press.

Taylor, S. J. (1974). Regularity of irregularities on a Brownian path. *Ann. Inst. Fourier (Grenoble)*, **24**(2), vii, 195–203. Colloque International sur les Processus Gaussiens et les Distributions Aléatoires (Colloque Internat. du CNRS, No. 222, Strasbourg, 1973).

Tsybakov, A. B. (2004). *Introduction à l'estimation non-paramétrique*, volume 41 of *Mathématiques & Applications (Berlin) [Mathematics & Applications]*. Springer-Verlag, Berlin.

von Davier, A. A., Holland, P. W., et Thayer, D. T. (2003). *The kernel method of test equating.* Statistics for Social Science and Public Policy. Springer-Verlag, New York.

von Davier, A. A., Fournier-Zajac, S., et Holland, P. W. (2007). Population invariance and the equatability of tests : basic theory and the linear case. *Report Number : RR-07-14.*

Wand, M. P. et Jones, M. C. (1995). *Kernel smoothing.*, volume 60 of *Monographs on Statistics and Applied Probability.* Chapman and Hall Ltd., London.

WHOQOL-HIV (2004). WHOQOL-HIV for quality of life assessment ammong people living with hiv and aids : Results from the filed test. *AIDS Care*, **16**(7), 882–889.

Yang, S.-S. (1985). A smooth nonparametric estimator of a quantile function. *J. Amer. Statist. Assoc.*, **80**(392), 1004–1011.

Yu, K. et Jones, M. C. (1998). Local linear quantile regression. *J. Amer. Statist. Assoc.*, **93**(441), 228–237.

Zelterman, D. (1990). Smooth nonparametric estimation of the quantile function. *J. Statist. Plann. Inference*, **26**(3), 339–352.

Zhang, S. et Karunamuni, R. J. (1998). On kernel density estimation near endpoints. *J. Statist. Plann. Inference*, **70**(2), 301–316.