République Algérienne Démocratique et Populaire
Ministère de L'Enseignement Supérieur et de la Recherche Scientifique
Université Colonel Akli Mohand Oulhadj- Bouira
Faculté des Sciences et des Sciences de Appliqueés

Mémoire de Master
Département de Mathématiques
Spécialité : Recherche Opérationnelle

Thème : Estimation à Noyau de la Densité de Probabilité dans le Cas des Données de Composition

---

*Présentée par :**Nacef Mourad***
*Sous la direction de Monsieur : **Said Beddek***

Devant le jury composée de :

Mr. Hamid Karim	**Président**	**M.A.A**	**U.A.M.O Bouira**
Mr .Beddek Said	**Encadreur**	**M.A.A**	**U.A.M.O Bouira**
Mme. Boudene Khadidja	**Examinatrice**	**M.A.A**	**U.A.M.O Bouira**
Mr. Hamdouni Omar	**Examinateur**	**M.A.A**	**U.A.M.O Bouira**

Année Universitaire 2022-2023

Remerciements

Je commence par exprimer ma gratitude envers Dieu.

pour le courage, la patience et la volonté

qui ont guidé notre parcours.

Un grand merci à Mr. ***BEDDEK SAID*** *pour son encadrement,*

ses précieux conseils, et ses orientations.

Je tiens également à remercier chaleureusement les membres du jury

pour avoir accepté d'examiner et d'évaluer notre travail.

Je remercie tous les enseignants du département de mathématiques

pour leur contribution essentielle à notre formation.

Votre soutien et votre confiance ont été cruciaux pour notre succès.

Merci du fond du coeur.

Dedicaces

Du profond de mon coeur, je dédie ce travail à tous ceux qui me sont chers.

***À la mémoire de mon père, SAID**,*

Ce travail est dédié À Mon père (RAHIMAHO ALLAH),

Tu ne nous as rien refusé dans cette vie et tu nous a quittés bien trop tôt.

Que Dieu lui accorde Sa miséricorde Et le place au plus haut rang au paradis

***À ma chère mère Boudina Hadda**,*

aucune dédicace ne saurait exprimer adéquatement mon respect,

mon amour éternel et ma profonde considération pour les sacrificesque vous avez consentispour mon éducation et mon bien-être. Je vous remercie du fond du cœur pour tout le soutien et l'amour que vous m'avez donnés depuis mon enfance. Puisse Dieu vous accorder santé, bonheur et une longue vie.

À mes chers frères,

vous avez toujours été mon soutien inestimable , du plus aîné au plus jeune.

J'ai réussi dans mon parcours grâce à vos précieux conseils.

je n'oublierai jamais ces merveilleux moments passés ensemble.

Merci du fond du cœur.

À mes deux chères soeurs

Vous êtes l'une des bénédictions de Dieu.

Je vous ai toujours trouvés à mes côtés.

La vie sans vous est insipide.

À TOUTE MA FAMILLE

À mes chèrs amis

Votre amitié a illuminé ma vie de joie,

Cette réussite est aussi la vôtre, car vous avez été une source constante d'inspiration et de soutien. Je tiens à remercier particulièrement mes amis ***Bacha Mohamed*** *et* ***Hamma Mohamed***

pour leurs encouragements tout au long de mon parcours.

sincèrement :

Merci !

Table des matières

Table des figures

Introduction Générale

L'estimation de la densité de probabilité est un domaine essentiel de la statistique et de l'analyse des données, permettant de comprendre la distribution sous-jacente des données et d'extraire des informations cruciales. Cependant, lorsque les données présentent des caractéristiques particulières, telles que des contraintes de somme à un, comme c'est le cas des données de composition, l'approche classique de l'estimation de densité peut être inappropriée[**Aitchison and Lauder, 1985**]. Les données de composition se manifestent dans de nombreux domaines, tels que la géochimie, l'économétrie, la recherche biomédicale et bien d'autres, où elles sont souvent exprimées sous forme de contributions relatives plutôt qu'absolues.

Ce mémoire se penche sur le défi de l'estimation de densité de probabilité dans le contexte spécifique des données de composition. Les données de composition sont représentées comme des vecteurs dont les coordonnées expriment la contribution relative de chaque partie à l'ensemble, avec la somme des coordonnées étant constante[**Aitchison and Lauder, 1985**]. Pour aborder cette problématique, nous utilisons des outils avancés de la géométrie et de la statistique, en mettant particulièrement l'accent sur l'estimation à noyau.

Ce travail est structuré en quatre chapitres, chacun approfondissant un aspect spécifique de l'estimation à noyau pour les données de composition. Dans le premier chapitre, nous introduisons les données de composition, expliquant leur nature unique et leurs applications dans divers domaines. Nous présentons également les avancées conceptuelles et méthodologiques récentes qui ont permis de traiter efficacement ces données.

Le deuxième chapitre se concentre sur l'estimation de densité de probabilité multivariée à l'aide de la méthode du noyau. Nous expliquons en détail le concept fondamental des estimateurs à noyau, en mettant en évidence leur pertinence et leur simplicité d'utilisation par rapport à d'autres méthodes d'estimation de densité de probabilité .

Dans le troisième chapitre, nous explorons l'utilisation des noyaux dans l'espace du simplex (Sd) pour estimer les densités de données de composition. Nous examinons trois types de noyaux - Dirichlet, gaussiens multivariés et logistiques-normaux - et analysons leurs avantages et limites dans le contexte des données de composition.

Enfin, dans le quatrième chapitre, nous nous lançons dans la simulation de données de composition et analysons les performances des trois noyaux précédemment abordés. Nous utilisons une méthodologie de validation croisée pour optimiser les paramètres de lissage et évaluer la précision des estimations.

Ce mémoire offre ainsi une exploration approfondie de l'estimation à noyau de la densité de probabilité dans le cas des données de composition, en combinant des concepts théoriques avec des applications pratiques. Il vise à contribuer à la compréhension et l'application dans

divers domaines de la recherche et de l'analyse de données.

Chapitre 1

Données de composition

Introduction

Dans ce chapitre, nous allons introduire un cas particulier trés important de données dépendantes. Il s'agit de cas des données de compositions (En anglais on dit : Compositional data) dont les compositions sont définies comme des vecteurs dans lesquels les coordonnées représentent une contribution relative des différentes parties d'un ensemble. Par conséquent, leur somme est une constante $\boldsymbol{c}$ selon les unités de mesure. L'espace d'échantillonnage des données de composition [**Aitchison and Lauder, 1985**] est le simplex de dimension d (d-simplexe), noté par S^d , qui est défini comme suit :

$$S^d = \{(x_1, ..., x_d) : x_1 > 0, ..., x_d > 0; x_1 + ... + x_d = \boldsymbol{c}\}.$$

On retrouve en général ce genre de données dans différents domaines de la science appliquée et expérimentale. Notamment, dans la géochimiques, l'économétrie, la recherche biomédicale ou la recherche spatiale (voir **Aitchison [1982]** pour plus de détails). Ces données se présentent en général sous forme de grandeurs relatives plutôt qu'absolues. Ce qui rend les outils de la géométrie classique euclidienne inutilisables et inappropriés pour décrire les différentes interactions entres les compositions. Au cours de ces dernières années, d'énormes progrès basés sur des concepts alternatifs ont été réalisés. Ainsi, de nouveaux outils algébriques et géométriques plus adaptés à la structure spécifique du simplexe S^d et constituants une géométrie naturelle pour celui-ci ont été construits et ils sont maintenant largement utilisée et appliquée dans des études pratiques(par exemple **Pawlowsky-Glahn and Egozcue [2001]**).[**Chacón et al., 2011**]

1.1 Données de composition

Pour introduire ces données, nous nous référons à la description du [**Aitchison and Lauder, 1985**]. Ainsi, les données de composition, en général, font référence à un type particulier de données utilisées pour représenter des informations relatives aux parties ou aux composants d'un tout. Ces données se caractérisent par le fait que la somme des proportions de ces parties est constante ou égale à un. Les données de composition sont couramment rencontrées dans divers domaines, notamment la chimie, la géologie, la biologie, l'économie, et d'autres sciences naturelles et sociales.

Voici une définition plus détaillée :

Définition 1.1.1 *Les données de composition sont un ensemble de données dans lequel chaque observation est représentée par un vecteur de proportions, où chaque composant du vecteur représente la contribution relative d'une partie ou d'un composant spécifique à un tout. La caractéristique fondamentale des données de composition est que la somme des proportions de chaque composant dans chaque observation est constante, généralement égale à un.*

Par exemple, si vous étudiez la composition chimique d'un échantillon de sol, vous pourriez avoir des données de composition où chaque observation représente la proportion des différents éléments chimiques (comme l'azote, le phosphore, le potassium, etc.) dans le sol, et la somme de ces proportions est toujours égale à 100 [**Aitchison and Lauder, 1985**].

1.2 Géométrie dans un simplexe

1.2.1 Simplexe

Pour toute situation expérimentale ou observationnelle où nous enregistrons une D-composition de la pièce $(x_1, \ldots, x_D)$, nous avons le choix entre deux voies vers les espaces d'échantillonnage naturels. Si nous sommes concernés par des problèmes mathématiques comme la spécification d'une fonction de densité sur l'espace de l'échantillon, alors nous devons mettre l'accent sur la dimensionnalité de la composition et définir l'espace de l'échantillon en termes d'un sous-vecteur déterminant tel que $(x_1, \ldots, x_d)$ Cela nous am-ène à la définition suivante d'un simplex en termes de notation standard. [Cox et al., 1984]

Définition 1.2.1 *[Cox et al., 1984] Le simplexe de dimension d (d-simplex)est l'ensemble défini par ;*
$S^d = \{(x_1, \ldots, x_d) : x_1 > 0, \ldots, x_d > 0; x_1 + \ldots + x_d \leqslant 1\}$.

Puisque dans la plupart des applications le composant x_D est aussi important que n'impo-rte quel autre composant $(x_1, \ldots, x - d)$, nous préférerions naturellement, en l'absence de toute condition mathématique, une approche plus symétrique. Dans cette approche, nous pouvons adopter la définition suivante de [Cox et al., 1984] ;

Définition 1.2.2 *La frontière du simplex d-dimensionnel dans l'espace réel d-dimensionnel est l'ensemble défini par :*
$\varphi^d = \{(x_1, \ldots, x_d) : x_1 > 0, \ldots, x_d > 0; x_1 + \ldots + x_d = 1\}$.

Définition 1.2.3 *L'interieur de simplexe de dimension d (int d-simplex)est l'ensemble défini par ;*

$$I^d = \{(x_1, \ldots, x_d) : x_1 > 0, \ldots, x_d > 0; x_1 + \ldots + x_d < 1\}.$$

Remarque 1.2.1 *Par convention , un simplexe de dimension 0 est un point de l'espace.*
-Un simlexe du dimension 1 est le segment $[0, 1]$ *(Fig1.1).*
-un simplexe du dimension 2 est un triangle(Fig1.1).
-un simplexe tridimensionnelle est un tétraèdre (Fig1.1).

Remarque 1.2.2 *Un d-simplexe est la frontière du simplexe du dimension* $d + 1$*, en effet :*
En posant $d^{'} = d - 1$ *;*
On aura : $x_{d'} = 1 - x_1 - \ldots - x_d$ *;*
Par conséquent , on peut écrire :

$$S^d = \{(x_1, \ldots, x_d, x_{d'}) : x_1 > 0, \ldots, x_d, x_{d'} > 0, ; x_1 + \ldots + x_d + x_{d'} = 1\} ;$$

d'où ;

$$S^d \simeq \varphi^{d+1} ;$$

Proposition 1.2.1 *[Cox et al., 1984] Soit* $C^d = \cup_{k=1}^{d} S^k$ *,l'ensemmble de tout les simplexe de dimension* k *,* $k \leqslant d$.
L'inclusion est une relation d'ordre total dans C^d*,c'est à dire :*

$$\varphi^1 \subset \varphi^2 \subset \varphi^3 \subset \ldots \subset \varphi^d.$$

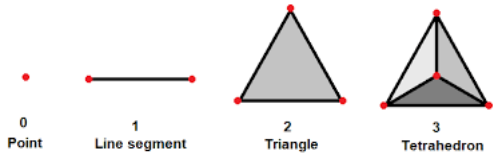

FIG. 1.1 – Représentation graphique des simplexes de dimension 0, 1, 2 et 3 de gauche à droite

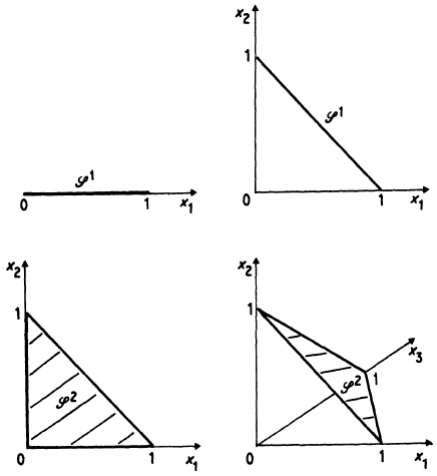

FIG. 1.2 – simplex

Définition 1.2.4 ***Diagrammes ternaires[Hardy, 2018] :***

Les diagrammes ternaires sont une méthode pratique pour représenter la variabilité des compositions en trois parties. Ils sont communément appelés diagrammes ternaires, triangles de référence ou barycentriques.

Ces diagrammes sont largement utilisés dans certaines disciplines, notamment dans les sciences géologiques. Le triangle représenté dans la figure 1.3 de [Hardy, 2018], avec les sommets A, B et C, est un triangle équilatéral ayant une altitude unitaire.

Pour tout point X situé à l'intérieur du triangle ABC, les perpendiculaires x_1, x_2, x_3 tirées de X vers les côtés opposés A, B et C satisfont aux conditions suivantes :

$$x_i \geqslant 0 \quad (i = 1, 2, 3) \quad x_1 + x_2 + x_3 = 0.$$

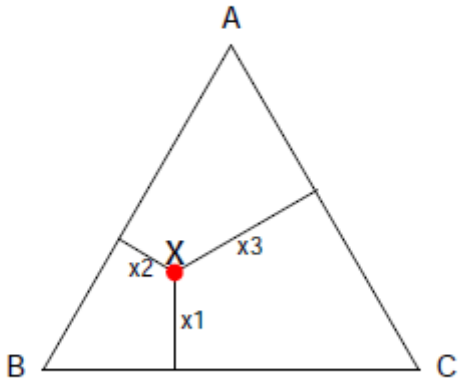

FIG. 1.3 – Exemple de diagramme ternaire. X est le barycentre des sommets A ;B ;C pondérés par les poids x1 ; x2 ; x3

1.2.2 Closure

d'après ([Chacón et al., 2011] et [Hardy, 2018]) La transformation qui permet de modifier la somme des composantes d'une composition (par exemple, passer d'une proportion à un pourcentage) est appelée *closure*. En notant k la somme des composantes souhaitée, la closure est définie comme suit :

$$C(k, x) = \left[\frac{k \cdot x_1}{\sum_{i=1}^{d} x_i}, \frac{k \cdot x_2}{\sum_{i=1}^{d} x_i}, ..., \frac{k \cdot x_d}{\sum_{i=1}^{d} x_i} \right]$$

Dans la suite du rapport, sauf indication contraire explicite, la notation $C(x)$ correspondra à $C(1, x)$. Deux vecteurs $x, y \in \mathbb{R}^{D+}$ tels que $x_i, y_i > 0, \forall i = 1, ..., d$ sont dits *compositionnellement équivalents* s'ils existent des valeurs $\lambda \in \mathbb{R}^+$ telles que $x = \lambda y$, ce qui équivaut à $C(x) = C(y)$.

1.2.3 Opérations dans le simplexe

Dans le simplexe, nous ne pouvons pas utiliser la géométrie euclidienne pour comparer des compositions. Remarquons cela en prenant en exemple les quatre 3-compositio-ns suivantes :
$x_1 = [0.1, 0.4, 0.5], x_2 = [0.2, 0.3, 0.5], x_3 = [0.4, 0.4, 0.2], x_4 = [0.5, 0.3, 0.2]$
La distance euclidienne entre les compositionsx_1 et x_2 est la même que celle entre x_3 et x_4, pourtant la proportion de la première composante a doublé dans le premier cas tandis qu'elle n'a augmenté que de 25 pour cent dans le second. De même, la multiplication et l'addition usuelles ne sont pas appropriées dans le simplexe : la multiplication d'une composition par un scalaire ne permet pas de rester dans le simplexe (elle change la valeur de k) et lorsque vous

additionnez deux compositions dans le simplexe, la valeur k de la composition résultante est doublée par rapport à chaque composition d'origine. Pour ces raisons, nous avons besoin de définir une nouvelle géométrie dans le simplexe . [Hardy, 2018]
Commençons d'abord par définir les opérations de bases qui sont définies par [Chacón et al., 2011] :
$\oplus$:La perturbation d'un élément $x \in S_d$ par un élément $y \in S_d$ est notée $x \oplus y$ et est définie comme suit : $C([x_1y_1, x_2y_2, ..., x_dy_d])$.
$\odot$:La puissance d'un élément $x \in S_d$ par une constante $\alpha \in \mathbb{R}$ est notée $\alpha \odot x$ et est définie comme suit : $C([x^{\alpha_1}, x^{\alpha_2}, ..., x^{\alpha_d}])$.

Les opérations ci-dessus confèrent au simplexe la structure d'un espace vectoriel. La première opération, notée $\oplus$, est la loi de composition interne appelée perturbation. La seconde opération, notée $\odot$, est la loi de composition externe appelée puissance. En plus de ces deux opérations, on introduit un produit scalaire ainsi que les concepts de distance et de norme associés à ce produit scalaire :

Le produit scalaire de deux compositions $x, y \in S_d$ est défini comme suit :

$$x, y \in s^d, \langle x, y \rangle_a = \frac{1}{2d} \sum_{i=1}^{d} \sum_{j=1}^{d} \ln \frac{x_i}{x_j} \ln \frac{y_i}{y_j} \ ,$$

La distance entre les compositions x et $y \in S_d$ est donnée par :

$$x, y \in s^d, d_a(x, y) = \sqrt{\frac{1}{2d} \sum_{i=1}^{d} \sum_{j=1}^{d} (\ln \frac{x_i}{x_j} \ln \frac{y_i}{y_j})^2} \ ,$$

La norme d'une composition $x \in S_d$ est définie comme :

$$\|x\|_a = \sqrt{\frac{1}{2d} \sum_{i=1}^{d} \sum_{j=1}^{d} (\ln \frac{x_i}{x_j})^2} \ ,$$

[Chacón et al., 2011] Cette approche permet de définir des objets géométriques tels que les lignes compositionnelles de direction x et passant par x_0 (analogues aux droites dans l'espace euclidien) : , avec $x, x_0 \in S_d$ et $\alpha \in \mathbb{R}$. Graphiquement, pour $d = 3$, cela ressemble à ceci. Les mesures statistiques telles que la variance doivent également être redéfinies dans ce simplexe pour tenir compte de sa géométrie. Pour un échantillon de taille n de D-compositions, la moyenne est définie comme suit :

$$\bar{g} = C([g_1, g_2, \ldots, g_d]),$$

où

$$\bar{g} = (\prod_{j=1}^{n} x_{ij})^{1/n} \quad i = 1, 2, \ldots, d \ ,$$

La matrice de covariance quant à elle est définie par :

$$T = \begin{pmatrix} t_{11} & t_{12} & \ldots & t_{1d} \\ \vdots & \vdots & \ddots & \vdots \\ t_{d1} & t_{d2} & \ldots & t_{dd} \end{pmatrix} ,$$

avec

$$t_{ij} = var(\ln \tfrac{x_i}{x_j})$$

ce qui permet de définir la variance totale comme suit :

$$\text{totvar}[x] = \frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} t_{ij}.$$

Ainsi, cette approche permet de prendre en compte la géométrie particulière du simplexe lors de la définition des concepts statistiques tels que la moyenne et la variance.

1.2.4 transformation des données de composition

L'estimation de la densité à noyau est un outil essentiel pour l'analyse statistique et est largement utilisée par les statisticiens, intégrée dans la plupart des logiciels statistiques. Bien que cette méthode ait été principalement développée pour les données réelles, à la fois univariées et multivariées [**Jones, 1995**] , elle trouve également d'autres applications. tandis que [**Bowman and Azzalini, 1997**] ont proposé une méthode de transformation pour traiter des données non standard.

[**Chacón et al., 2011**] L'espace du simplexe nécessite une redéfinition des opérations de base et a sa propre géométrie, ce qui le rend plus complexe à manipuler. Dans cette optique, nous examinerons ici des transformations permettant de ramener le simplexe à un espace euclidien en modifiant la géométrie d'Aitchison vers une approche plus familière. Il est important de noter que la valeur de k (somme des composantes) d'une composition n'a pas fondamentalement une grande importance. En effet, l'étude des données reste la même, qu'elles soient exprimées en pourcentages ou en proportions. Les rapports relatifs entre les composantes sont en revanche très informatifs. Par conséquent, les données de composition nécessitent un traitement particulier. [**Aitchison, 1982**] a introduit la méthodologie de travail avec les log-ratios, proposant notamment une stratégie basée sur des transformations pour traiter ces données. Les transformations du log-ratio additif (alr) et du log-ratio centré (clr) sont définies comme suit :

$$alr : \mathbb{R}^D \to \mathbb{R}^{D-1}, \quad alr(x) = \left[\ln\left(\frac{x_1}{x_D}\right), \ln\left(\frac{x_2}{x_D}\right), \ldots, \ln\left(\frac{x_{D-1}}{x_D}\right) \right] \tag{1.1}$$

La dernière composante d'une composition peut s'exprimer comme :

$$x_D = k - \sum_{i=1}^{D-1} x_i$$

[**Chacón et al., 2011**] , Les D composantes d'une D-composition sont intrinsèquement liées et existent réell-ement dans un espace de dimension $D-1$. L'approche log-ratio additif (alr) résout cette problématique en choisissant une composante comme composante de référence et en comparant les autres composantes à cette référence. La D-ième coordonnée $\ln\left(\frac{x_D}{x_D}\right)$ est exclue de la transformation alr car elle est identiquement nulle.

La transformation alr permet de se débarrasser de la corrélation linéaire entre les composantes, mais elle dépend fortement du choix de la composante de référence. De plus, cette transformation ne conserve pas les distances, au sens où la distance euclidienne entre deux

compositions x et y n'est pas préservée par la transformation alr, c'est-à-dire $d(a(x,y)) \neq \|\text{alr}(x) - \text{alr}(y)\|_2$.

La transformation log-ratio centrée (clr) résout ces deux problèmes. Elle s'exprime comme suit :

$$clr : \mathbb{R}^D \to \mathbb{R}^D, \quad clr(x) = \left[\ln\left(\frac{x_1}{g(x)}\right), \ln\left(\frac{x_2}{g(x)}\right), \ldots, \ln\left(\frac{x_D}{g(x)}\right)\right] \tag{1.2}$$

avec $g(x)$ représentant le centre de la composition, défini comme :

$$g(x) = \left(\prod_{i=1}^{D} x_i\right)^{\frac{1}{D}}$$

[**Hardy, 2018**] , La moyenne géométrique de la composition x est donnée par $g(x) = (x_1 \ldots x_D)^{\frac{1}{D}}$. Essentiellement, la méthode de transformation proposée par [**Aitchison and Lauder, 1985**] pour l'estimation de la densité du noyau consiste à estimer la densité des données transformées en utilisant le log-ratio additif (alr), suivi d'une transformation vers le simplexe. Cependant, il est recommandé d'être prudent lors de l'application de la transformation alr, car elle est asymétrique dans ses composants et ne préserve pas les distances.

Pour surmonter ces problèmes, la transformation log-ratio isométrique (ilr) définie comme suit :

$$ilr : \mathbb{R}^D \to \mathbb{R}^{D-1}, \quad ilr(x) = y = [y_1, \ldots, y_{D-1}] \in \mathbb{R}^{D-1} \tag{1.3}$$

Où les coordonnées y_i sont données par :

$$y_i = \frac{1}{\sqrt{i(i+1)}} \ln\left(\frac{\prod_{i=1}^{j} x_j}{(x_{i+1})^i}\right)$$

La transformation ilr est isométrique, préservant ainsi les distances, et elle permet d'éviter les complications liées à la singularité de la matrice de covariance des données transformées. [**Chacón et al., 2011**]

1.2.5 Réduction de dimension

d'après [**Hardy, 2018**] L'Analyse en Composantes Principales (ACP) est une méthode couramment utilisée pour réduire la dimension d'un ensemble d'échantillons en synthétisant l'information qu'ils contiennent. Aitchison a proposé une redéfinition de l'ACP pour les données compositionnelles en les centrant et en les réduisant dans le simplexe avant de calculer les vecteurs propres de la matrice de covariance totale.

Cependant, dans cette approche, nous préférons exploiter le fait que la transformation ilr est un morphisme. Ainsi, nous définissons l'ACP pour les données compositionnelles comme l'ACP standard sur les données transformées en ilr. Ces deux définitions sont rigoureusement équivalentes et permettent d'obtenir les mêmes résultats d'analyse. En utilisant l'ACP sur les données ilr-transformées, nous pouvons bénéficier des propriétés de cette transformation et simplifier l'analyse des données compositionnelles.

1.3 Conclusion

En conclusion, les données de composition sont une classe de données importante, couramment utilisée dans divers domaines scientifiques et appliqués. Le simplexe, en tant qu'espace d'échantillonnage spécifique, nécessite des méthodes de traitement et d'a-nalyse adaptées, telles que les transformations alr, clr et ilr. Ces approches permettent de mieux comprendre les relations entre les compositions et ouvrent des perspectives intéressantes pour des analyses approfondies dans différents domaines de la recherche.

Chapitre 2

Estimation de la densité de probabilité multivariée par la méthode du noyau

Introduction

Dans cette section introductive, nous nous pencherons sur la problématique de l'estimation de densité de probabilité multivariée, en nous focalisant sur une catégorie spécifiq-ue d'approches appelées estimateurs à noyau de convolution, ou tout simplement, estimateurs à noyau. Ces méthodes de calcul à base de noyaux jouent un rôle fondamental dans la détermination des densités de probabilité dans des contextes multidimensionnels, représentant ainsi une extension significative au sein de la vaste famille des estimateurs à noyau.

L'approche d'estimation de densité de probabilité à l'aide de la méthode du noyau est très répandue et se démarque comme l'approche privilégiée par rapport à d'autres méthodes telles que les histogrammes, les techniques basées sur les séries orthogonales ou encore les estimateurs spline. Les estimateurs à noyau se distinguent par leurs propriétés favorables, leur simplicité d'interprétation et leur facilité de mise en œuvre. Le concept fondamental sous-tendant cette approche consiste à évaluer la densité f en un point donné $x \in \mathbb{R}^d$ (avec $d \in \mathbb{N}^*$) en comptabilisant le nombre d'observations qui se situent dans une zone de voisinage définie autour de x dans $\mathbb{R}^d$.

L'origine de cette théorie remonte aux contributions pionnières de [**Akaike, 1954**] détaillé et enrichi par le travail du [**S.BEDDEK, 2011**]. Dans le cas univarié, si nous disposons d'un ensemble de données $X_1, \ldots, X_n$, extrait d'une variable aléatoire unidimensionnelle X avec f com-me densité de probabilité, alors l'estimateur à noyau $bfn(x)$ de la densité f en un point $x \in \mathbb{R}$ donné se présente sous la forme suivante :

$$\widehat{f}_n(x) = \frac{1}{nh} \sum_{j=1}^{n} K\left(\frac{x - x_j}{h}\right), \tag{2.1}$$

où le paramètre h est lié à l'ampleur de la fenêtre de lissage, K désigne une fonction noyau définie sur l'ensemble $\mathbb{R}$, et les termes x_j correspondent aux observations dans l'échantillon.

Au cours de ce chapitre, nous approfondirons notre compréhension des estimateurs à noyau, mettant l'accent particulièrement sur les estimateurs à noyau de convolution pour l'estimation de densités de probabilité dans des contextes multivariés. De plus, nous explorerons différentes approches pour la sélection de la matrice de lissage, notamment les

techniques de validation croisée, dans le but d'assurer des performances optimales de l'estimateur à noyau dans des espaces multidimensionnels.

2.1 Estimateur à noyau :

Considérons un vecteur aléatoire d-dimensionnel $X = (X_1, \ldots, X_d)$ avec une fonction de densité d-variée $f(X_1, \ldots, X_d)$. Nous prenons également en compte un échantillon aléatoire $X^{(1)}, \ldots, X^{(i)}, \ldots, X^{(n)}$ issu de X, ce qui signifie que $X^{(i)}$ est défini comme ; $(X_1^{(i)}, \ldots, X_d^{(i)})$ pour $i = 1, \ldots, n$.

Définition 2.1.1 *L'estimateur à noyau de la densité de probabilité d-dimensionnelle f s'exprime généralement sous la forme :*

$$\widehat{f}_n = \frac{1}{n} \sum_{j=1}^{n} K(x - X^{(j)}), \tag{2.2}$$

avec :

$$K_H(x) = |H|^{-\frac{1}{2}} K\left(H^{-\frac{1}{2}} x\right), \tag{2.3}$$

où :

- *H est une matrice carrée d'ordre d, symétrique et définie positive, désignée comme la matrice des paramètres de lissage ou la matrice des fenêtres (bandwidth matrix).*
- *$K(.)$, appelée fonction noyau d-dimensionnelle, est une application de $\mathbb{R}^d$ dans $\mathbb{R}$, bornée et satisfaisant : $\int_{\mathbb{R}^d} K(x)dx = 1$.*

2.2 Construction du noyau multidimensionnel :

Il existe deux méthodes pour construire le noyau d-dimensionnel K à partir d'un noyau univarié symétrique ω.

Définition 2.2.1 *Le noyau produit, également connu sous le nom de noyau K^p, est défini comme suit :*

$$K^p(x) = \prod_{i=1}^{d} \omega(x_i). \tag{2.4}$$

Définition 2.2.2 *On désigne par noyau sphérique (spherically or radially symmetric Kernel) le noyau K^s défini comme suit :*

$$K^s(x) = C_{\omega.d}\omega\left[(x^T x)^{\frac{1}{2}}\right] \tag{2.5}$$

où :

$$C_{\omega.d}^{-1} = \int \omega\left[(x^T x)^{\frac{1}{2}}\right] dx \tag{2.6}$$

Remarque 2.2.1 *Pour l'estimation de la densité de probabilité multivariée, le noyau gaussien standard est le plus couramment utilisé lorsque le support de la densité est* $\mathbb{R}^d$. *Dans ce scénario, l'estimateur présente des propriétés asymptotiques favorables (voir* ***[Silverman]*** *et* ***[Scott, 2015]****). Cependant, cet estimateur montre des limitations lorsque certaines variables sont bornées. Un problème de biais se manifeste aux limites, entraînant la divergence de l'estimateur. La question du biais aux limites a été étudiée en profondeur dans le cas univarié . Cette problématique devient encore plus complexe dans le contexte multivarié, où elle s'ajoute à la dimension du support. Une première solution à ce problème pour le cas multivarié a été proposée par* ***[Bouezmarni and Rombouts, 2010]****. Ils suggèrent d'utiliser l'estimateur suivant :*

$$\hat{f}(x_1, \ldots, x_d) = \frac{1}{n} \sum_{i=1}^{n} \prod_{l=1}^{d} K^l(h_l, X_l^{(i)}) \cdot K_l(x_l)$$

Le vecteur $\mathbf{h} = (h_1, \ldots, h_d)^T$ représente les paramètres de lissage, et K_l est un noyau appliqué à la variable x_l. Les auteurs ont examiné deux scénarios impliquant des variables à support borné :

1. Lorsque le support de la variable est non-négatif, trois noyaux différents pour K_l sont envisagés :

— Le noyau K_L (noyau linéaire local) est défini comme suit :

$$k_L(h,t)(x) = \frac{a_0(x,h) - a_1(x,h)y}{a_0(x,h)a_2(x,h) - a_1^2(x,h)} K(y),$$

où $y = \frac{x-t}{h}$, K est un noyau symétrique univarié à support compact $[-1, 1]$, et

$$a_l(x,h) = \int_{\frac{x}{h}-1}^{\frac{x}{h}} t^l K(y)\, dy.$$

— Le noyau gamma K_G est donné par l'expression :

$$KG(h,t)(x) = \frac{xh}{t} \exp\left[-\frac{h}{t}\right] \frac{1}{xh+1} \Gamma(xh+1) = \Gamma\left(\frac{1}{h}, xh+1\right). \tag{2.7}$$

— Le noyau gamma K_{NG} est défini par :

$$KNG(h,t)(x) = \exp\left[\frac{h}{t}\right] \rho(x)\Gamma(p(x))t\rho(x)^{-1} = \Gamma\left(\frac{1}{h}, \rho(x)\right), \tag{2.8}$$

où $p(x)$ est donné par :

$$\rho(x) = \begin{cases} \frac{x}{h} & \text{si } x \geq 2h \\ \frac{1}{4}\left(\frac{x}{h}\right)^2 + 1 & \text{si } x \in [0, 2h] \end{cases}$$

2. Si la variable est à support compact (nous considérons le support $[0, 1]$ pour simplifier), le noyau Beta est utilisé :

Le noyau $K(h,t)(x)$ est défini comme suit :

$$K(h,t)(x) = \beta\left(\frac{xh+1}{h}\right), \tag{2.9}$$

ou le noyau Beta modifié $kNB(h,t)(x)$ est utilisé :

$$k_{NB}(h,t)(x) = \begin{cases} \beta\left(\frac{x}{h}; \frac{1-x}{h}\right) & \text{si} \quad x \in [0, 2h] \\ \beta\left(\frac{x}{h}; \frac{1-x}{h}\right) & \text{si} \quad x \in [2h, 1-2h] \\ \beta\left(\frac{x}{h}; \rho(1-x)\right) & \text{si} \quad x \in [1-2h, 1] \end{cases}$$

où $\beta(\alpha, b)$ est la fonction densité de la loi Beta avec des paramètres α et β, h est le paramètre de lissage, et $\rho(x) = 2h^2 + 2.25 - \sqrt{4h^2 + 6h^2 + 2.25 - x^2 - \frac{x}{h}}$

2.3 Qualité et Performance de l'Estimateur

L'évaluation de l'estimateur à noyau, défini par l'équation (2.2), dépend de sa proximité avec la densité cible. Cette proximité est quantifiée à travers le MISE (Erreur Quadratique Moyenne Intégrée) ou l'AMISE (Erreur Quadratique Moyenne Intégrée Asymptotique). Pour introduire notre discussion, commençons par présenter le théorème suivant, qui généralise le théorème de Taylor à plusieurs dimensions :

Définition 2.3.1 *Soit g une fonction à d dimensions et $\boldsymbol{\alpha}_n = (\alpha_1^n, \ldots, \alpha_d^n)^T$, où $n \in \mathbb{N}$, une séquence de vecteurs de dimension d tels que chaque composante α_i^n tend vers 0 lorsque n tend vers l'infini. Soit $\mathbf{D}_g(\mathbf{x})$ le vecteur des dérivées partielles d'ordre 1 de g et $\boldsymbol{\chi}_g(\mathbf{x})$ la matrice hessienne de g, c'est-à-dire la matrice carrée d'ordre d où l'élément (i,j) est égal à $\frac{\partial^2 g(\mathbf{x})}{\partial x_i \partial x_j}$. Si chaque élément de $\boldsymbol{\chi} g(\mathbf{x})$ est continu dans un voisinage de $\mathbf{x}$, alors :*

$$g(\mathbf{x} + \boldsymbol{\alpha}_n) = g(\mathbf{x}) + \boldsymbol{\alpha}_n^T + \frac{1}{2}\boldsymbol{\alpha}_n^T \boldsymbol{\chi}_g(\mathbf{x}) \boldsymbol{\alpha}_n + o(\boldsymbol{\alpha}_n^T \boldsymbol{\alpha}_n)$$

Ceci représente le développement de Taylor d'ordre 2 pour la fonction g au point $\mathbf{x}$.

Pour les développements ultérieurs, nous devrons aussi poser les hypothèses supplémen-taires suivantes concernant f, H et K :

1. Chaque élément de la matrice hessienne de f, notée $\boldsymbol{\chi}_f(\mathbf{x})$, est borné, continu et a une intégrale carrée finie pour tout $\mathbf{x} \in \mathbb{R}^d$.

2. $\mathrm{H} = \mathrm{H}(n)$ est une suite de matrices de lissage où chaque élément tend vers zéro lorsque n tend vers l'infini, et où $\lim_{n\to\infty} n^{-1/2}|\mathrm{H}| = 0$.

3. K est un noyau à d variables satisfaisant les conditions suivantes :

$$\int \mathrm{K}(z)dz = 1, \quad \int z\mathrm{K}(z)dz = 0, \quad \int zz^T\mathrm{K}(z)dz = \mu_2(\mathrm{K}) \cdot I,$$

où $\mu_2(\mathrm{K}) = \int z_i^2 \mathrm{K}(z)dz$ est fini et indépendant de i.

Il est important de noter que la condition (3) est satisfaite par tous les noyaux sphériq-ues symétriques ainsi que par les noyaux produits construits à partir d'un noyau symétri-que univarié de variance finie. Nous aurons également besoin des définitions et résultats suivants :

Définition 2.3.2 *Considérons une matrice carrée $\mathbf{A}$. La trace de $\mathbf{A}$, notée $tr(\mathbf{A})$, correspond à la somme des éléments présents en diagonale dans $\mathbf{A}$. Une propriété intéressante à noter est que pour toute matrice carrée $\mathbf{B}$ de même ordre que $\mathbf{A}$, la relation suivante est vérifiée :*

$$tr(\mathbf{AB}) = tr(\mathbf{BA}) \tag{2.10}$$

. Cette propriété met en évidence une équivalence importante entre les traces de deux matrices lorsqu'elles sont multipliées dans différents ordres.

Définition 2.3.3 *[**Henderson and Searle, 1979**] ont proposé une approche intrigante. Considéro-ns une matrice carrée A d'ordre d. Le vecteur associé à A, noté $vecA$, est construit en empilant les colonnes de A les unes sous les autres de gauche à droite, créant ainsi un vecteur de dimension $(d^2 \times 1)$. Le demi-vecteur de A, noté $vechA$, est obtenu en prenant le vecteur $vecA$ et en éliminant les éléments situés au-dessus de la diagonale, aboutissant à un vecteur de dimension $(\frac{1}{2})d(d+1) \times 1$. Lorsque A est une matrice symétrique, $vechA$ contient les éléments distincts de A, ce qui implique que $vecA$ contient les éléments de $vechA$ avec certaines répétitions.*

résultats importants

1 Il est possible de définir une matrice unique D_d de dimensions $(d^2 \times \frac{1}{2}d(d+1))$, composée de zéros et de uns. Cette matrice a la propriété suivante : pour toute matrice symétrique A de taille $(d \times d)$, l'équation

$$D_d \text{vech}A = \text{vec}A \tag{2.11}$$

est vérifiée. On nomme cette matrice D_d la "matrice de duplication" d'ordre d.

2 Une formule importante et applicable à toutes les matrices carrées A d'ordre d est la suivante :

$$(D_d^T \text{vec}A) = \text{vech}(A + A^T - dgA) \tag{2.12}$$

3

$$\text{tr}(A^T B) = (\text{vec}^T A)(\text{vec}B) \tag{2.13}$$

4 En ce qui concerne les transformations linéaires de variables lors de l'intégration sur $\mathbb{R}^d$, le résultat est le suivant :

$$\int g(Ax)dx = |A| \int g(y)dy, \tag{2.14}$$

où A est une matrice carrée d'ordre d inversible.

2.3.1 Erreur quadratique moyenne intégrée asymptotique

Dans un contexte similaire à celui du cas univarié, il est possible d'obtenir une approximation asymptotique simplifiée de l'erreur quadratique moyenne intégrée (MISE) pour l'estimateur à noyau de la densité multivariée. En prenant en considération les hypothès-es (1), (2) et (3) concernant les fonctions f, H et le noyau K, on peut dériver la relation suivante :

$$AMIS[\widehat{f}(.;H)] = n^{-1}|H|^{-\frac{1}{2}}R(K) + \frac{1}{4}\mu_2(K)^2 \int \text{tr}^2[H\chi_f(x)]dx \tag{2.15}$$

Le développement du deuxième terme peut être accompli en exploitant les relations (2.10), (2.11) et (2.13), ce qui conduit à l'expression :

$$\int \text{tr}^2[H\chi_f(x)]dx = (\text{vech}^T H)\Psi_4(\text{vech}H)$$

où Ψ_4, tel que défini dans (2.12), représente la matrice carrée d'ordre $\frac{1}{2}d(d+1)$:

$$\Psi_4 = \int \text{vech}[2\chi_f(x) - dg\chi_f(x)] \times \text{vech}^T[2\chi_f(x) - dg\chi_f(x)]dx$$

Bien que la matrice Ψ_4 puisse sembler complexe au premier abord, une formule simple pour ses éléments peut être obtenue en réalisant une intégration par parties. Pour une fonction g en d dimensions et un vecteur $R = (r_1, \ldots, r_d)$ d'entiers non négatifs, nous utiliserons la notation suivante :

$$g^{(R)}(x) = \frac{\partial^{|x|}}{\partial_{x_1}^{r_1}, \ldots, \partial_{x_d}^{r_d}} g(x)$$

à condition, bien sûr, que cette dérivée soit définie. Nous noterons $|R|$ la somme des éléments de R, c'est-à-dire $|R| = \sum_{i=1}^{d} r_i$. Il est possible de démontrer que :

$$\int f^{(R)}(x) f^{(R')}(x) dx = (-1)^{|R|} \int f^{(R+R')}(x) f(x) dx \tag{2.16}$$

lorsque $|(R + R')|$ est pair, sinon le résultat est 0. Par conséquent, chaque composante de Ψ_4 peut être exprimée sous la forme suivante :

$$\psi_R = \int f^{(R)}(x) f(x) dx$$

sous la condition que $|R|$ soit pair.

Remarque 2.3.1 *Il est important de noter que contrairement au cas univarié, il n'existe générale-ment pas d'expressions explicites pour la matrice de lissage optimale H qui minimise l'AMISE (Erreur Quadratique Moyenne Intégrée), et cette quantité ne peut être obtenue que numériquement (voir [**Wand, 1992**]). Des expressions relativement simples pour l'AMISE sont possibles dans les situations où $H \in S$ et $H \in D$. Plus précisément, lorsque :*

$$H = diag(h_1^2, \ldots, h_d^2), \tag{2.17}$$

alors l'AMISE peut être exprimée comme suit :

$$AMIS[\widehat{f}(.; H)] = n^{-1} R(K) \left(\prod_{j=1}^{d} h_j \right)^{-1} + \frac{1}{4} \mu_2(k)^2 \left(h_1^2, \ldots, h_d^2 \right)^T \Psi_D \left(h_1^2, \ldots, h_d^2 \right) \tag{2.18}$$

Dans cette équation, D est une matrice carrée d'ordre d avec l'élément (i, j) égal à $\psi_{2e_i} + \psi_{2e_j}$, où e_i est le vecteur d-dimensionnel ayant 1 comme composante i et 0 ailleurs.

De plus, lorsque $H = h^2 I$, l'expression devient :

$$AMIS[\widehat{f}(.; H)] = n^{-1} h^{-1} R(K) + \frac{1}{4} h^4 \mu_2(K)^2 \int [\nabla^2 f(x)]^2 dx \tag{2.19}$$

Pour ce cas particulier, la matrice de lissage optimale qui minimise l'AMISE est donnée par :

$$h_{AMIS} = \left[\frac{d R(K)}{n \mu_2(K)^2 \int [\nabla^2 f(x)]^2 dx} \right]^{\frac{1}{(d+4)}} \tag{2.20}$$

Le minimum correspondant de l'AMISE peut alors être calculé comme suit :

$$\inf_{h>0} AMIS[\widehat{f}(.; H)] = \frac{d+4}{4d} \left[\left(\mu_2(K)^{2d} [dR(k)]^4 \left[\int [\nabla^2 f(x)]^2 dx \right]^d n^{-4} \right) \right]^{\frac{1}{d+4}} \tag{2.21}$$

Il convient de noter que selon cette dernière expression, la vitesse de convergence de $\inf_{h>0} AMISE[\widehat{f}(.;H)]$ est de l'ordre de $n^{-\frac{4}{d+4}}$, un taux qui diminue à mesure que la dimension de l'espace augmente. Cette diminution de la convergence, principalement due à la complexité de la dimension spatiale, peut rendre l'implémentation pratique des estimateurs à noyau de densité inappropriée dans des dimensions plus élevées. Néanmoins, cette méthode reste un outil extrêmement utile pour l'analyse de données dans des dimensions modérées de l'espace (voir [**Scott, 1991**]).

2.4 Choix de la matrice de lissage par les Mméthodes Cross validation

La performance de l'estimateur de densité à noyau dépend essentiellement du choix du paramètre de lissage. Dans le contexte univarié, ce choix équivaut à sélectionner un paramètre scalaire positif h qui régit le niveau de lissage. Cependant, en dimension multiple, le paramètre de lissage devient une matrice symétrique et définie positive H. Cette matrice influence à la fois le degré et la direction du lissage, ce qui rend sa sélection plus complexe. Jusqu'à présent, la majorité des efforts de recherche ont été concentrés sur la détermination automatique du paramètre de lissage optimal en univarié. De ce fait, de nombreux travaux ont été réalisés dans la littérature sur ce sujet. Pour une revue générale, les travaux de ([**Jones et al., 1996**]) sont consultables.

2.4.1 Matrice de lissage optimale

Pour évaluer les performances de l'estimateur $\widehat{f}(x,H)$, nous adoptons la mesure d'erreur quadratique moyenne intégrée (MISE) définie comme suit :

$$MISE\widehat{f}(.,H) = E\int[\widehat{f}(x,H) - f(x)]^2 dx$$

Notre objectif consiste à choisir la matrice de lissage qui minimise le MISE, c'est-à-dire à trouver :

$$H_{MIS} = argmin_{H\in F} MISE\widehat{f}(.,H)$$

Ici, F représente l'ensemble des matrices carrées symétriques et définies positives d'ordre d. Dans leur étude, [**Jones, 1995**] ont établi que sous certaines conditions, l'approximation suivante est valable :

$$MISE\widehat{f}(.,H) = AMISE\widehat{f}(.,H) + o(n^{-1}|H|^{-\frac{1}{2}} + tr^2 H) \tag{2.22}$$

Où :

$$AMISE\widehat{f}(.,H) = n^{-1}|H|^{-\frac{1}{2}}R(K) + \frac{1}{4}\mu_2(K)^2(\text{vech}^T H)\Psi_4(\text{vech}H) \tag{2.23}$$

La matrice carrée Ψ_4 d'ordre $\frac{1}{2}d(d+1)$ est définie par :

$$\Psi_4 = \text{vech}\{2\chi_f(x) - dg\chi_f(x)\} \times \text{vech}^T\{2\chi_f(x) - dg\chi_f(x)\}dx$$

Ces deux dernières expressions fournissent une estimation pratique du MISE par le AMISE. Ainsi, une analyse asymptotique devient plus favorable. Cela implique que nous visons à estimer :

$$H_{AMISE} = argmin_{H\in F} AMISE\widehat{f}(.;H)$$

Plutôt que d'adopter l'approche précédente :

$$H_{MISE} = argmin_{H\in F} MISE\widehat{f}(.;H)$$

Il est important de rappeler que les éléments de la matrice Ψ_4 peuvent être exprimés de la manière suivante :

$$\psi_r = \int f^{(r)}(x) f(x) dx$$

Ici, $r = (r_1, \ldots, r_d)$ est un vecteur d'ordre d dont les composantes sont des entiers non-négatifs, et $|r| = \sum_{i=1}^{d} r_i$. Par conséquent, l'AMISE est une fonctionnelle de la densité inconnue f, à travers les éléments de Ψ_4. Dans cette optique, l'utilisation d'estimateurs pilotes pour les fonctionnelles ψ_r devient impérative. Ces estimateurs peuvent ensuite être réutilisés pour obtenir un estimateur $\widehat{AMISE}$ de l'AMISE, qui peut être soumis à une minimisation numérique afin d'obtenir la matrice de lissage optimale H_{opt}. Il convient de noter que cette démarche est simplifiée lorsque la matrice H est diagonale (voir [Wand et al., 1994]).

Estimation fonctionnelle pilote

Considérons les éléments suivants :

$$\begin{aligned}\psi_r &= \int_{\mathbb{R}^d} f^{(r)}(x) f(x) dx \\ &= \mathbb{E} f^{(r)}(X)\end{aligned}$$

où X représente une variable aléatoire d-dimensionnelle avec une densité f. Ainsi, l'estimateur naturel de ψ_r émerge en calculant la moyenne empirique de $\widehat{f}^{(r)}(X)$, exprimée comme :

$$\begin{aligned}\widehat{\psi_r}(G) &= n^{-1} \sum_{i=1}^{n} \widehat{f}^{(r)}(X^{(i)}, G) \\ &= n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} K_G^{(r)}(X^{(i)} - X^{(j)}) \qquad (2.24)\end{aligned}$$

Dans ces équations, la matrice G représente la matrice de lissage pilote qui diffère de la matrice de lissage initiale H. par définition ,Il convient de noter que la matrice pilote G est symétrique et définie positive .

AMSE- Matrice pilote de lissage

Considérons $G = g^2 I$ avec $g > 0$. Soit $|r| = j$. Sous les hypothèses que $K^{(r)}$ est intégrable au carré, que $g = g_n \to \infty$ quand $n \to \infty$, et que $n^{-1}g^{-d-2j} \to 0$ quand $n \to \infty$, nous obtenons :

$$AMSE\widehat{\psi_r}(g) = 2n^{-2}g^{-d-2j}\psi_0 R(K^{(r)}) + \left[n^{-1}g^{-d-j}K^{(r)}(0) + \frac{1}{2}g^2\mu_2(k)\sum_{i=1}^{d}\psi_r + 2e_i\right]^2$$

Nous cherchons à déterminer :

$$g_r, AMSE = \arg\min_{g>0} AMSE\widehat{\psi_r}(g)$$

Il convient de souligner que la formulation de l'AMSE a été établie par [**Wand et al., 1994**]. Pour la majorité des noyaux, y compris le noyau gaussien, lorsque tous les éléments de r sont pairs, une relation de signes opposés se manifeste entre $K^{(r)}(0)$ et ψ_{r+2e_i} pour tous les $i = 1, \ldots, d$. Dans ce contexte, la valeur de $g_{r,\text{AMSE}}$ est déterminée de manière à annuler le terme de biais :

$$g_r, AMSE = \left[\frac{-2K^{(r)}(0)}{\mu_2(K)\sum_{i=1}^{d}\psi_{r+2e_i}n}\right]^{\frac{1}{d+j+2}}$$

Si au moins l'un des éléments de r est impair, alors $K^{(r)}(0) = 0$. Le minimum de l'AMSE est atteint lorsque g est égal à :

$$g_{r,\text{AMSE}} = \left[\frac{2\psi_0(2|r|+d)R(K^{(r)})}{\mu_2(K)^2(\sum_{i=1}^{d}\psi_{r+2e_i}n)}\right]^{\frac{1}{d+2j+4}}$$

Les valeurs de $g_{r,\text{AMSE}}$ dépendent des fonctionnelles d'ordre supérieur ψ_{r+2e_i}, qui font partie de l'ensemble Ψ_6. Il est possible d'estimer chaque élément de Ψ_6, c'est-à-dire chaque fonctionnelle ψ_{r+2e_i} avec $|r| = 4$, en utilisant un estimateur à noyau différent. Cependant, il devient évident que la matrice de lissage optimale dépend des éléments de Ψ_8. Cela nous ramène au même dilemme, nous poussant à estimer Ψ_8 à l'aide de la méthode du noyau, puis $\Psi_{10}, \Psi_{12}, \ldots$. Cela conduit à une séquence d'estimations à noyau en cascade, ce qui n'est pas souhaitable. Une alternative est donc nécessaire pour estimer Ψ_6.

Une idée prometteuse consiste à fixer un nombre maximal m d'estimations à noyau à effectuer, ce qui permet d'estimer successivement $\Psi_6, \Psi_8, \ldots, \Psi_{4+2m}$. Ensuite, les éléments de Ψ_{4+2m} sont estimés à l'aide d'une approximation normale de référence, définie comme suit :

$$\widehat{\psi_r}^{NR} = (-1)^{|r|}\Phi_{2s}^{(r)}(0), \tag{2.25}$$

pour $|r| = 4 + 2m$.

Il est important de noter que $\Phi_\Sigma(x)$ représente la densité normale multivariée de moyenne nulle et de matrice de variance-covariance Σ, évaluée en x. Dans ce contexte, Σ correspond à la matrice de variance-covariance de l'échantillon S.

Lemme 2.4.1 *[Duong and Hazelton, 2003]*
Si une unique matrice de lissage pilote G et le noyau gaussien sont utilisés pour estimer tous les éléments de Ψ_4, alors $\widehat{\Psi}_4$ est garantie d'être définie positive.

SAMSE- Matrice pilote de lissage

Afin d'assurer que la matrice Ψ_4 soit définie positive, [Duong and Hazelton, 2003] ont introduit un nouveau critère d'erreur appelé SAMSE (Somme Asymptotique des Erreurs Quadratiques Moyennes). Pour un ordre de dérivation j, le SAMSE est défini comme suit :

$$SAMSE_j(G) = \sum_{r;|r|=j} AMSE\widehat{\psi}_r(G)$$

Dans la continuité des arguments précédents, nous adoptons la forme $G = g^2 I$, où $g > 0$. [Duong and Hazelton, 2003] ont ainsi proposé de minimiser le SAMSE pour obtenir :

$$g_j, SAMSE = \arg\min_{g>0} SAMSE_j(g)$$

L'expression du SAMSE pour un ordre de dérivation j est donnée par :

$$\begin{aligned} SAMSE_j(G) &= \sum_{r;|r|=j} AMSE\widehat{\psi}_r(G) \\ &= 2n^{-2}g^{-2j-d}A_0 + 2n^{-2}g^{-2j-2d}A_1 + n^{-1}g^{-j-d+2}A_2 + \frac{1}{4}g^4 A_3 \end{aligned}$$

où les constantes A_0, A_1, A_2 et A_3 sont indépendantes de n et définies par :
$A_0 = \sum_{r;|r|=j} R(K^{(r)})$
$A_1 = \sum_{r;|r|=j} K^{(r)}(0)^2$
$A_2 = \mu_2(K)^2 \sum_{r;|r|=j} K^{(r)}(0)(\sum_{i=1}^{d} \psi_{r+2e_i})$
$A_3 = \mu_2(K)^2 \sum_{r;|r|=j} (\sum_{i=1}^{d} \psi_{r+2e_i})^2$
Les valeurs de A_0, A_1 et A_3 sont positives par construction, tandis que A_2 est négatif en raison de la relation de signes opposés entre $K^{(r)}(0)$ et ψ_{r+2e_i} lorsque r est pair, et de $K^{(r)}(0) = 0$ lorsque r est impair.
L'expression $SAMSE_j(G)$ peut être simplifiée. En effet, le premier terme$o(n^{-2}g^{-2j-d})$, est dominé par le deuxième terme, $o(n^{-2}g^{-2j-2d})$. Ainsi, en négligeant le premier terme (qui correspond à la variance asymptotique), nous obtenons :

$$SAMSE_j(G) = 2n^{-2}g^{-2j-2d}A_1 + (n^{-1}g^{-j-d+2}A_2) + \frac{1}{4}g^4 A_3 \tag{2.26}$$

Nous avons ainsi considéré uniquement la contribution du biais au carré. En dérivant par rapport à g, nous trouvons :

$$\frac{\partial}{\partial_g} SAMSE_j(G) = g^3[-(2j+2d)n^{-2}g^{-2j-2d-4}A_1 - (j+d-2)n^{-1}g^{-1-d-2}A_2 + A_3].$$

En annulant cette expression qui est une équation du second degré en $[n-1g-j-d-2]$, sa résolution donne :

$$g_j, SAMSE = \left[\frac{(4j+4d)A_2}{((-j-d+2)A_2 + \sqrt{(-j-d+2)^2A_2^2 + (8j+8d)A_1A_3})n}\right]^{\frac{1}{j+d+2}} \tag{2.27}$$

Ainsi, la matrice pilote de lissage SAMSE d'ordre j est définie. Cette méthode d'estimation de Ψ_4 utilise la même matrice pilote de lissage $G = g_{j,SAMSE}I$ pour tous les éléments de Ψ_4. Par conséquent, selon le lemme (2.4.1), Ψ_4 est définie positive. Un autre avantage majeur de la méthode SAMSE est son économie par rapport à la méthode AMSE.

présentation des méthode validation crossé

Au cours de cette partie , nous explorerons diverses approches relatives à la validation croisée. Notre attention se portera sur trois méthodologies principales : la validation croisée non biaisée (UCV), la validation croisée biaisée (BCV) ainsi que la validation croisée lissée (SCV). Par ailleurs, nous aurons l'occasion d'examiner les déclinaisons multivariées de ces trois procédés. Dans le but d'évaluer le niveau de convergence propre à chacune de ces méthodes, il conviendra au préalable de définir certains concepts fondamentaux de. Cette mise en contexte nous permettra ensuite d'approfondir l'étude de chacune de ces approches.

Définition 2.4.1 *Considérons deux suites de variables aléatoires réelles, notées* $[A_n]_{n\in\mathbb{N}}$ *;* $[B_n]_{n\in\mathbb{N}}$.

1. *On dit que* $A_n = O_p(B_n)$ *lorsque, pour tout* $\xi > 0$, *on a* $\lim_{n\to\infty} P\left(\left|\frac{A_n}{B_n}\right| > \xi\right) = 0$.
2. *On dit que* $A_n = O_p(B_n)$ *si, pour tout* $\xi > 0$, *il existe des valeurs* λ *et* M *tels que* $P\left(\left|\frac{A_n}{B_n}\right| > \lambda\right) < \xi$ *pour tout* $n > M$.

Définition 2.4.2 *Lorsque l'estimateur* $\widehat{H}$ *tend à converger vers* H_{AMISE} *avec un taux relatif de* $n^{-\alpha}$ *,*$(\alpha > 0)$*, la relation suivante est vérifiée :*

$$vech(\widehat{H} - H_{AMIS})O_p(j_{d'}n^{-\alpha})vech(H_{AMIS})$$

où :

— H_{AMIS} *désigne la matrice de lissage qui minimise l'estimateur* $AMIS$.

— $\widehat{H}$ *représente l'estimateur de* H_{AMIS}.

— $j_{d'}$ *est une matrice carrée d'ordre* $d' = \frac{1}{2}d(d+1)$, *avec des éléments égaux à 1.*

Il est essentiel de noter que cette relation exprime la convergence de $\widehat{H}$ *vers* H_{AMISE} *avec un taux relatif de décroissance* $n^{-\alpha}$, *où* α *est un paramètre positif.*

Définition 2.4.3 ***Dfinition 2.4.3 :***

Considérons deux séquences de matrices $[A_n]_{n\in\mathbb{N}}$ *et* $[B_n]_{n\in\mathbb{N}}$, *où chaque matrice* A_n *et* B_n *a les mêmes dimensions pour tout* $n \in \mathbb{N}$. *Les concepts suivants sont définis :*

1. *On dit que* $A_n = O_p(B_n)$ *si, pour chaque élément* $[A_n]_{ij}$ *de la matrice* A_n *et chaque élément* $[B_n]_{ij}$ *de la matrice* B_n, $[A_n]_{ij} = [B_n]_{ij}$.
2. *On dit que* $A_n = o_p(B_n)$ *si, pour chaque élément* $[A_n]_{ij}$ *de la matrice* A_n *et chaque élément* $[B_n]_{ij}$ *de la matrice* B_n, $[A_n]_{ij} = [B_n]_{ij}$.

Ces définitions illustrent les relations entre les éléments des matrices A_n *et* B_n, *auxquelles les notations* O_p *et* o_p *font référence.*

Lemme 2.4.2 *([**Duong and Hazelton, 2005**])*
Considérons les scénarios suivants :

— (A_1) *: Tous les éléments de la fonction* $\chi_f(x)$ *sont bornés, continus et intégrables au carré.*

— (A_2) *: Tous les éléments de* H *tendent vers zéro et* $n^{-1}|H|^{-\frac{1}{2}} \longrightarrow 0$ *lorsque* $n \longrightarrow \infty$.

— (A_3) *: Le noyau* K *est sphérique.*

Maintenant, définissons $\widehat{H} = argmin_{H\in F}\widehat{AMISE}(H)$ *comme l'estimateur de* H*. Nous évaluons son erreur quadratique moyenne comme suit :*

$$MSE(vech\widehat{H}) = E[vech(\widehat{H} - H_{AMISE})vech^T(\widehat{H} - H_{AMISE})].$$

Ainsi, nous pouvons exprimer cette erreur quadratique moyenne comme :

$$MSE(vech\widehat{H}) = [I_{d'} + o(j_{d'})]AMSE(vech\widehat{H}),$$

Où :

$$AMSE(vech\widehat{H}) = AV_{ar}(vech\widehat{H}) + [ABiais((vech\widehat{H}))][ABiais((vech\widehat{H}))]^T,$$

En outre :

$$ABiais((vech\widehat{H})) = [D_H^2 AMISE(H_{AMISE})^{-1}E[D_h(\widehat{AMISE} - AMISE)(H_{AMISE})]],$$

Et :

$$AV_{ar}(vech\widehat{H}) = Var[D_H^2 AMISE(H_{AMISE})]^{-1}Var[D_H(\widehat{AMISE} - AMISE)(H_{AMISE})][D_H^2 AMISE(H_{AMISE})]^{-1},$$

Où DH *représente l'opérateur différentiel par rapport à* $vechH$ *et* D_H^2 *est l'opérateur Hessien correspondant.*

Dans la littérature, ce lemme est généralement identifié sous le nom du lemme de $AMSE$.

Remarquons que lorsque l'équation suivante est vérifiée :

$$MSE(vech\widehat{H}) = O_p(j_{d'}n^{-2B})(vechH_{AMISE})(vech^T H_{AMISE}),$$

alors l'estimateur $\widehat{H}$ manifeste un taux de convergence équivalent à n^{-B}.

En utilisant le lemme (2.4.2), il devient possible de calculer le taux de convergence de $\widehat{H}$ vers H_{AMISE} en fonction de l'espérance et de la matrice de variance-covariance de $D_H(\widehat{AMISE} - AMISE)(H_{AMISE})$. Pour rappel :

$$\widehat{AMISE} = n^{-1}R(K)|H|^{-\frac{1}{4}} + \frac{1}{4}\mu_2(K)^2(vech^T H)\widehat{\Psi}_4(vechH),$$

d'où découle :

$$(\widehat{AMISE} - AMISE)(H) = \frac{1}{4} + \frac{1}{4}\mu_2(K)^4(vech^T H)(\widehat{\Psi}_4 - \Psi_4(vechH)[1 + o_p(1)].$$

Ainsi, on obtient :

$$E[d_H(\widehat{AMISE} - AMISE)(H)] = \frac{1}{2}\mu_2(K)^2[I_{d'} + o(j_{d'})](E\widehat{\Psi}_4 - \Psi_4)(vechH)$$

et la variance-covariance de $D_H(\widehat{AMISE} - AMISE)(H)$ se traduit par :

$$V_{ar}[D_H(\widehat{AMISE} - AMISE)(H)] = \frac{1}{4}\mu_2(K)^4[I_{d'} + o(j_{d'})]V_{ar}[\widehat{\Psi}_4(vechH)]$$

2.4.2 Validation croisée non-biaisée (UCV)

Présentation de la méthode

La version multivariée du critère de la méthode UCV constitue une généralisation simple de sa contrepartie univariée, préalablement proposée par [**Rudemo, 1982**] et [**Bowman, 1984**] :

$$UCV(H) = \int_{\mathbb{R}^d} \widehat{f}(x;H)^2 dx - \frac{2}{n}\sum_{i=1}^{n} \widehat{f}_{-i}(X_i;H),$$

où :

$$\widehat{f}_{-i}(x;H) = \frac{1}{n-1}\sum_{j=1;j\neq i}^{n} K_H(x-X_i).$$

Notre objectif est de choisir la matrice de lissage qui minimise la fonction UCV (H), c'est-à-dire :

$$\widehat{H}_{UCV} = \arg\min_{H\in F} UCV(H).$$

Lemme 2.4.3 *La méthode UCV vise à minimiser une mesure similaire à ISE(H), qui peut être décrite de la manière suivante :*

$$ISE(H) = \int_{\mathbb{R}^d} [\widehat{f}(x,H)]^2 dx,$$

où :

$$\begin{aligned} ISE(H) &= \int_{\mathbb{R}^d} \widehat{f}(x;H)^2 dx - 2\int_{\mathbb{R}^d} f(x)\widehat{f}(x;H)dx + \int_{\mathbb{R}^d} f(x)^2 dx, \\ &= R(\widehat{f}(x)) - 2\int_{\mathbb{R}^d} f(x)\widehat{f}(x;H)dx + R(f(x)). \end{aligned}$$

Comme $R(f(x))$ est indépendant de H, ce terme peut être négligé. Il nous reste à estimer $\int_{\mathbb{R}^d} f(x)\widehat{f}(x;H)dx$. Notons que :

$$\int_{\mathbb{R}^d} f(x)\widehat{f}(x;H)dx = E(\widehat{f}(X;H)).$$

Son estimateur naturel est alors :

$$\frac{1}{n}\sum_{i=1}^{n} \widehat{f}(X_i;H).$$

Afin de préserver l'indépendance des variables, nous utilisons l'estimateur Jackknife suivant :

$$\widehat{f}_{-i}(x;H) = (n-1)^{-1}\sum_{j=1;j\neq i}^{n} K_H(x-X_j).$$

L'estimation de $\int_{\mathbb{R}^d} f(x)\widehat{f}(x;H)dx$ *est donc effectuée comme suit :*

$$\frac{1}{n}\sum_{i=1}^{n}\widehat{f}_{-i}(X_i;H).$$

En substituant cette expression dans la formule de $ISE(H)$*, nous obtenons le critère* $UCV(H)$ *à minimiser. Ce critère peut être développé de la manière suivante :*

$$\begin{aligned} UCV(H) &= n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}(K_H * K_H)(X_i - X_j) - 2n^{-1}(n-1)\sum_{i=1}^{n}\sum_{i=1;j\neq 1}^{n}K_H(X_i - X_j), \\ &= n^{-1}R(K)|H|^{-\frac{1}{2}} + n^{-1}(n-1)\sum_{i=1}^{n}\sum_{i=1;j\neq 1}^{n}(K_H * K_H - 2K_H)[X_i - X_j]. \end{aligned} \quad (2.28)$$

Des investigations ont été menées par [**Sain et al., 1994**] concernant la version multivariée de la méthode de sélection UCV. Toutefois, leur attention s'est principalement portée sur l'utilisation du noyau produit, ce qui revient à adopter une paramétrisation diagonale avec un noyau sphérique K^s. Ces chercheurs ont également étudié la vitesse de convergence relative de la méthode UCV dans le contexte de la paramétrisation diagonale.

Par la suite, [**Duong and Hazelton, 2005**] ont étendu ces conclusions pour une matrice de lissage H symétrique et définie positive quelconque ($H \in F$).

2.4.3 Validation croisée biaisée (BCV)

Présentation de la méthode

L'approche basée sur la méthode de sélection de la validation croisée biaisée (BCV) vise à minimiser un estimateur de l'AMISE :

$$AMIS[\widehat{f}(.;H)] = n^{-1}|H|^{-\frac{1}{2}}R(K) + \frac{1}{4}\mu_2(K)^2(vech^T H)\Psi(vechH)$$

Il est nécessaire d'estimer Ψ_4. On pose $G = H$ et on utilise des estimateurs légèrement différents. Étant donné que l'AMISE est un estimateur biaisé du MISE, l'estimateur BCV est également biaisé pour le MISE (même s'il converge asymptotiquement vers l'estimateur non biaisé). C'est d'ailleurs ce biais qui confère à la méthode BCV son nom, car il est introduit pour réduire la variance.

Il existe deux versions de la méthode BCV en fonction de l'estimateur utilisé pour ψ_r, où $|r| = 4$. Pour plus de détails, on peut se référer à [**Sain et al., 1994**] et [**Jones and Kappenman, 1992**].

Ainsi, il est possible d'utiliser l'estimateur :

$$\check{\psi}_r(H) = n^{-2}\sum_{i=1}^{n}\sum_{i=1,j\neq 0}^{n}(K_H^r * K_H)(X_i - X_j),$$

ou encore, on peut considérer l'estimateur :

$$\tilde{\psi}_r(H) = n^{-1}\sum_{i=1}^{n}\widehat{f}_{-i}^{(r)}(X_i;H) = n^{-1}(n-1)\sum_{i=1,j\neq 0}^{n}K_H^r(X_i - X_j).$$

Les estimateurs $\check{\Psi}_4$ et $\tilde{\Psi}_4$ sont obtenus en remplaçant les fonctionnelles ψ_r par les estimateurs $\check{\psi}_r$ et $\tilde{\psi}_r$ respectivement. La fonction BCV1 correspond à la version de la BCV utilisant $\check{\Psi}_4$:

$$BCV1(H) = n^{-1}R(K)|H|^{-\frac{1}{2}} + \frac{1}{4}\mu_2(K)^2(vech^T H)\check{\Psi}_4(vechH) \tag{2.29}$$

De même, la fonction BCV2 correspond à la version de la BCV avec $\tilde{\Psi}_4$:

$$BCV2(H) = n^{-1}R(K)|H|^{-\frac{1}{2}} + \frac{1}{4}\mu_2(K)^2(vech^T H)\tilde{\Psi}_4(vechH) \tag{2.30}$$

La méthode de sélection BCV vise à trouver la matrice $\widehat{H}_{BCV}$ qui minimise la fonction BCV appropriée. En d'autres termes :

$$\widehat{H}_{BCV1} = \arg\min_{H\in F} BCV1 \text{ et } \widehat{H}_{BCV2} = \arg\min_{H\in F} BCV2$$

[**Sain et al., 1994**] ont entrepris des recherches sur la méthode de sélection BCV dans le cas d'une paramétrisation diagonale. Ils ont calculé le taux relatif de convergence de la BCV diagonale. La généralisation de cette méthode pour une matrice de lissage symétrique et définie positive quelconque ($H \in F$) a été proposée par [**Duong and Hazelton, 2005**].

Taux relatif de convergence de la méthode BCV

Les deux estimateurs $\check{\psi}_r$ et $\tilde{\psi}_r$ présentent une similitude notable. En employant un noyau gaussien, l'équation $\phi_H^{(r)} * \phi_H = (-1)^{|r|}\phi_{2H}^{(r)}$ émerge. Par conséquent, la seule distinction réside dans l'utilisation de $2H$ par $\check{\psi}_r$ et de H par $\tilde{\psi}_r$. Cette divergence n'exerce aucune influence sur le taux relatif de convergence, puisqu'elle n'affecte ni l'ordre de biais ni la variance asymptotique. En conséquence, il est seulement requis de calculer le taux de BCV2.

Lemme 2.4.4 *Supposons que les conditions (A1), (A2) et (A3) de lemme (2.4.3) de AMSE sont satisfaites. Alors :*

$$ABiiais(vech\widehat{H}_{BCV}) = O(J_d n^{-\frac{2}{d+4}})vechH_{AMISE}$$

et

$$AVar(vech\widehat{H}_{BCV}) = O(J_d n^{-\frac{2}{d+4}})(vechH_{AMISE})(vech^T H_{AMISE})$$

En amalgamant le lemme (2.4.2) d'AMSE avec le précédent lemme (2.4.4), émerge le théorème suivant qui détermine le rythme relatif de convergence de la méthode de sélection BCV. C'est-à-dire, le rythme auquel $\widehat{H}_{BCV}$ converge vers H_{AMISE}.

Théorème 2.4.1 *Sous les hypothèses du lemme (2.4.4), le taux relatif de convergence de $\widehat{H}_{BCV}$ vers H_{AMISE} est $n^{-\frac{\min(d,4)}{2d+8}}$. Ce taux de convergence s'avère identique à celui de la méthode de sélection UCV. Notamment, [**Sain et al., 1994**] ont fourni un taux de convergence relatif pour la méthode BCV diagonale, équivalent à $n^{-\frac{\min(d,4)}{2d+8}}$, mais ce résultat se révèle inexact pour $d > 4$.*

2.4.4 Validation croisée lissée (SCV)

Présentation de la méthode

La méthode de choix SCV (Smoothed Cross Validation) peut être envisagée comme une fusion entre les approches UCV et BCV. Son but est de lisser le critère UCV. La version multivariée du critère SCV étend simplement sa forme univariée proposée par [**Hall et al., 1992**]. Cette forme se présente sous la manière suivante :

$$SCV(H) = n^{-1}R(K)|H|^{-\frac{1}{2}} + n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}(K_H * K_H * L_G * L_G - 2K_H * L_G * L_G + L_G * L_G)(X_i - X_j),$$

où L_G représente le noyau pilote avec la matrice de lissage pilote G. Le premier terme est l'estimateur de la variance asymptotique intégrée et le second terme est l'estimateur de l'intégrale du carré du biais. La méthode de choix SCV vise à trouver la matrice $\widehat{H}_{SCV}$ qui minimise le critère SCV(H).

Il est à noter que lorsque $G = O$ (matrice nulle), nous obtenons $SCV(H) = UCV(H)$. L_0 correspond à la fonction delta de Dirac.

Si $K = L = \phi$, alors le critère SCV prend une forme plus simple :

$$SCV = n^{-1}|H|^{-\frac{1}{2}}(\pi)^{-\frac{d}{2}} + n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\phi_{2H+2G} - 2\phi_{H+2G} + \phi_{2G})(X_i - X_j). \tag{2.31}$$

Matrice de lissage pilote optimale

Comment sélectionner la matrice pilote de lissage optimale ? Plusieurs approches ont été proposées :

— [**Sain et al., 1994**] prennent la matrice pilote égale à la matrice de lissage finale.

— [**Jones et al., 1991**], dans le contexte univarié, suggèrent de choisir le paramètre pilote qui minimise l'erreur quadratique moyenne relative (RMSE - relative mean squared error). Pour un paramètre univarié $\widehat{h}$, cela s'exprime comme :

$$RMSE(\widehat{h}) = E\left[\left(\frac{\widehat{h} - h_{AMISE}}{h_{AMISE}}\right)^2\right].$$

— [**Duong and Hazelton, 2005**], quant à eux, proposent de minimiser l'AMSE. Il est important de noter que minimiser l'AMSE équivaut à minimiser le RMSE, car le dénominateur du RMSE ne dépend pas du paramètre de lissage.

Une généralisation de la version univariée de $MSE(\widehat{h}) = E(\widehat{h} - h_{AMISE})^2$ est donnée par :

$$trMSE(vech\widehat{H}; G) = E[vech^T(\widehat{H} - H_{AMISE})vech(\widehat{H} - H_{AMISE})].$$

Pour le calcul de g_0, les résultats intermédiaires suivants, établis par[**Duong and Hazelton, 2005**], seront nécessaires. Pour cela, l'extension d'ordre supérieur de l'AMISE, notée $AMISE'$, est définie comme :

$$AMISE' = AMISE(H) + \frac{1}{8}\int_{\mathbb{R}^d} tr(H\chi_f(x)tr(H^2\chi_f^2(x))dx.$$

On notera $\widehat{AMISE'}$ pour l'estimateur de $AMISE'$.

Lemme 2.4.5 *Supposons que les conditions* $(A_1), (A_2)$ *et* (A_3) *énoncées dans le lemme de AMSE (2.4.2) sont vérifiées. Si l'on considère* $\widehat{H}$ *comme étant le minimiseur de* $\arg\min_{H\in F} AMISE'$, *alors on peut énoncer la relation suivante :*

La valeur quadratique moyenne de l'estimateur $\operatorname{vech}\widehat{H}$ *est donnée par*

$$MSE(\operatorname{vech}\widehat{H}) = [I_{d'} + o(j_{d'})]AMISE'(\operatorname{vech}\widehat{H}).$$

Il convient de noter que l'expression de l'extension d'ordre supérieur asymptotique de l'erreur quadratique moyenne est la suivante :

$$AMSE'(\operatorname{vech}\widehat{H}) = AVar'(\operatorname{vech}\widehat{H}) + [ABiais'(\operatorname{vech}\widehat{H})][ABiais'(\operatorname{vech}\widehat{H})]^T.$$

Lemme 2.4.6 *Supposons que les conditions du lemme (2.4.5) concernant* $AMSE'$ *soient satisfaites, en tenant également compte des hypothèses suivantes :*

(S1) *La fonction* f *possède des dérivées partielles jusqu'à l'ordre huit qui sont bornées et continues.*

(S2) *Chaque composante de* $\theta_6 = \int_{\mathbb{R}^d} \chi_f^3(x)f(x)dx$ *est finie.*

(S3) *La suite des paramètres de lissage pilotes* $g = g_n$ *satisfait* $g^{-2}H \longrightarrow 0$ *lorsque* $n \longrightarrow \infty$.

(S4) *Les noyaux* K *et* L *sont des noyaux gaussiens.*

Alors, nous avons :

$$ABiais'(\operatorname{vech}\widehat{H}_{SCV}; g) = n^{-\frac{2}{d+4}}g^2C_{\mu_1} + n^{-\frac{2}{d+4}}n^{-1}g^{-d-4}C_{\mu_2} + o(j_{d'}(g^4 + n^{-1}g^{-d-6}))(vechH_{AMISE}).$$

où

— $C_{\mu_1} = \frac{1}{2}n^{-\frac{2}{d+4}}D_d^T\operatorname{vec}(\theta_6 H_{AMISE})$

— $C_{\mu_2} = \frac{1}{8}(4\pi)^{-\frac{d}{2}}n^{\frac{2}{d+4}}[2D_d^T\operatorname{vec}H_{AMISE} + (\operatorname{tr}H_{AMISE}D_d^T\operatorname{vec}I_d)]$.

Lemme 2.4.7 *Supposons que les conditions du lemme (2.4.5) concernant* $AMSE'$ *soient satisfaites, et sous les conditions du lemme (2.4.6), on a :*

$$AVar'(\operatorname{vech}\widehat{H}_{SCV}; g) = o(j_{d'}(n^{-2}g^{-d-8} + n^{-1}(\operatorname{vech}H_{AMISE})(\operatorname{vech}^T H_{AMISE})))$$

Les trois lemmes (2.4.5), (2.4.6), (2.4.7) nous conduisent au théorème suivant :

Théorème 2.4.2 *Sous les conditions du lemme (2.4.5) et (2.4.6), le paramètre pilote qui minimise la trace de* $AMSE'(\operatorname{vech}\widehat{H}_{SCV}; g)$ *pour* $d > 0$ *est donné par :*

$$g_0 = \left[\frac{2(d+4)C_{\mu_2}^TC_{\mu_2}}{-(d+2)C_{\mu_2}^TC_{\mu_1} + \sqrt{C_{\mu_0}}\,n}\right]^{\frac{1}{d+6}}$$

où

$$C_{\mu_0} = (d+2)^2(C_{\mu_2}^TC_{\mu_1})^2 + 8(d+4)(C_{\mu_1}^TC_{\mu_1})(C_{\mu_2}^TC_{\mu_2})$$

Taux relatif de convergence de la méthode SCV

Le taux relatif de convergence de la méthode de sélection SCV découle directement du théorème (2.4.2) et du lemme (2.4.4) concernant $AMSE'$. Il est à noter que lorsque

$$trMSE(\text{vech}\,\widehat{H}) = o(n^{-2\alpha}\|\text{vech}\,H_{AMISE}\|^2)$$

alors la convergence relative de $\widehat{H}$ vers H_{AMISE} s'effectue à un taux de $n^{-\alpha}$.

Théorème 2.4.3 *Sous les conditions de lemme (2.4.6) et (2.4.7), pour $d > 1$ le taux relatif de convergence de $\widehat{H}_{SCV}$ vers H_{AMISE} est $n^{-\frac{2}{d+6}}$.*

Algorithmes de sélection des méthodes cross validation

Nous allons exposer les algorithmes des trois méthodes, à savoir UCV, BCV et SCV, proposées par [**Duong and Hazelton, 2005**]. En comparaison avec les méthodes UCV et BCV, la méthode de sélection SCV se révèle notablement plus complexe. En effet, pour la SCV, il est nécessaire d'estimer une matrice de lissage pilote.

Algorithme pour la méthode de sélection UCV

1. Minimiser numériquement l'équation (2.23) de UCV(H).

Algorithme pour la méthode de sélection BCV

1. Minimiser numériquement
   (a) L'équation (2.24) de BCV 1(H), ou
   (b) L'équation (2.25) de BCV 2(H).

Algorithme de sélection m-étape de la méthode SCV

1. Fixer m (le nombre d'étapes à faire).
2. Pour $J_{max} = 2m+4$, donner l'approximation normale de référence $\widehat{\psi}^{NR}$ avec $|r| = J_{max}$. Puis injecter celle-ci dans la formule de $g_{J_{max}-2,SAMSE}$.
3. Pour $J = J_{max} - 2, J_{max} - 4, \ldots, 6$:
   (a) Calculer les estimateurs à noyau pour les fonctionnelles ψ_r, d'ordre $J = |r|$, en utilisant l'estimateur de $g_{J,SAMSE}$.
   (b) Remplacer les estimateurs $\widehat{\psi}_r$ dans l'équation (2.22) pour obtenir des estimateurs de $g_{J-2,SAMSE}$.
4. Utiliser $g_{6,SAMSE}$ pour avoir l'estimateur à noyau $\widehat{\theta}_6$ de θ_6.
5. Utiliser $g_{4,SAMSE}$ pour obtenir l'estimateur à noyau Ψ_4. Injecter cet estimateur dans l'équation (2.18) pour avoir $\widehat{AMISE}(H)$.
6. Minimiser numériquement $\widehat{AMISE}(H)$ pour obtenir la matrice de lissage SAMSE-$\widehat{H}_{SAMSE}$.
7. Utiliser $\widehat{H}_{SAMSE}$ et $\widehat{\theta}_6$ pour obtenir l'estimateur $\widehat{g}_0$ du théorème (2.4.2).
8. Remplacer $\widehat{g}_0$ dans l'équation (2.26) pour obtenir la formule de SCV(H).
9. Minimiser numériquement SCV(H).

Conclusion

En résumé, ce chapitre a exploré en profondeur l'estimation non paramétrique à noyaux de densité de probabilité multivariée . Nous avons abordé plusieurs points essentiels qui définissent cette approche.

Nous avons commencé par introduire les estimateurs à noyau de densité de probabilité, qui jouent un rôle fondamental dans la détermination des densités de probabilité dans des contextes multidimensionnels. Ces estimateurs sont appréciés pour leurs propriétés favorables, leur simplicité d'interprétation et leur facilité de mise en œuvre. L'idée sous-jacente de cette méthode est d'évaluer la densité en comptabilisant le nombre d'observations dans un voisinage autour d'un point donné.

Nous avons examiné en détail les estimateurs à noyau de convolution pour l'estimation de densités de probabilité dans des contextes multivariés. Nous avons également exploré les différentes approches pour la sélection de la matrice de lissage H, un paramètre crucial dans l'estimation non paramétrique à noyaux de densité multivariée.

Dans cette optique, nous avons présenté les méthodes de validation croisée (UCV), (SCV) et (BCV) pour sélectionner la matrice de lissage optimale. Ces méthodes visent à assurer des performances optimales de l'estimateur à noyau dans des espaces multidimensionnels en ajustant la matrice de lissage en conséquence.

En fin de compte, le choix judicieux de la matrice de lissage est essentiel pour obtenir des résultats de densité de probabilité précis dans des espaces de dimensions élevées. Les méthodes de validation croisée (UCV), (SCV) et (BCV) offrent des outils puissants pour déterminer ce paramètre de manière efficace.

En conclusion, ce chapitre a jeté les bases pour comprendre en profondeur l'estimation non paramétrique à noyaux de densité de probabilité multivariée . Les méthodes de validation croisée jouent un rôle crucial dans la sélection du paramètre de lissage optimal, garantissant ainsi des résultats précis et fiables dans des dimensions élevées.

Chapitre 3

Exploration des Noyaux dans l'Espace Simplexe (S_d)

Introduction

Ce chapitre explore l'utilisation des noyaux pour estimer les densités de données dans le contexte du simplex S_d, un espace englobant les compositions multivariées soumises à des contraintes de somme à un.

Nous abordons les bases du simplex S_d et ses propriétés uniques, jetant le socle nécessaire à la compréhension des noyaux appliqués. Trois types de noyaux retiennent notre attention : Dirichlet, gaussiens multivariés et logistiques-normaux. Chacun propose une approche distincte pour modéliser et estimer la distribution de données multivariées sur S_d.

La première partie plonge dans les noyaux de Dirichlet, essentiels à l'analyse statistique des compositions multivariées. Leur formulation, propriétés et utilisation pour estimer les densités de données sur S_d sont explorées. Puis, nous étudions les noyaux gaussiens multivariés, adaptés au simplex S_d et aux particularités de cet espace. Enfin, les noyaux logistiques-normaux sont abordés, présentant une approche novatrice pour explorer les distributions sur S_d via des transformations log-ratio.

Ce chapitre allie concepts théoriques et résultats empiriques, illustrant l'utilisation des noyaux dans l'estimation de densité sur S_d. Avantages et limites de chaque approche sont soulignés, mettant en lumière leurs applications potentielles dans divers domaines de recherche et d'analyse de données.

En résumé, ce chapitre propose un voyage captivant à travers les noyaux et les distributions multivariées sur le simplex S_d, offrant des outils essentiels pour analyser et interpréter des données complexes, tout en tenant compte des spécificités des compositions multivariées.

3.1 Noyaux de dirichlet

d'après [**Ouimet and Tolosana-Delgado, 2022**] Le simplex en dimension-d ainsi que son intérieur sont définis de la manière suivante : Dans le cas du simplex d-dimensionnel S_d, il est défini comme l'ensemble des points s appartenant à l'intervalle $[0,1]^d$, tout en satisfaisant la condition que la norme ℓ_1 de s ne dépasse pas 1 :

$$S_d = \{s \in [0,1]^d : \|s\|_1 \leq 1\}.$$

De manière similaire, l'intérieur du simplex en dimension-d, noté $Int(S_d)$, est défini com-me l'ensemble des points s dans l'intervalle ouvert $(0,1)^d$, où la norme ℓ_1 de s est strictement inférieure à 1 :

$$Int(S_d) = \{s \in (0,1)^d : \|s\|_1 < 1\}.$$

Dans ces expressions, la norme ℓ_1 de s, notée $\|s\|_1$, est obtenue en sommant les valeurs absolues de ses composantes individuelles :

$$\|s\|_1 = \sum_{i=1}^{d} |s_i|,$$

où d est un nombre naturel. Les paramètres $\alpha_1, \ldots, \alpha_d$ et $\beta > 0$ sont également définis dans ce contexte.

[**Ouimet and Tolosana-Delgado, 2022**] , La densité de la distribution de Dirichlet (α, β) est donnée par l'équation suivante :

$$K_{\alpha,\beta}(s) = \frac{\Gamma(\|\alpha\|_1 + \beta)}{\Gamma(\beta)\prod_{i=1}^{d}\Gamma(\alpha_i)} \cdot (1 - \|s\|_1)^{\beta-1} \prod_{i=1}^{d} s_i^{\alpha_i - 1}, \quad s \in S_d. \tag{3.1}$$

Cette équation décrit la densité de probabilité de la distribution de Dirichlet avec les paramètres (α, β), où s est un point appartenant au simplex S_d.

Pour un paramètre de lissage donné $b > 0$ et un ensemble d'observations échantillonnées $X_1, \ldots, X_n$ provenant d'une distribution F (dont la nature exacte est inconnue) avec une densité f définie sur l'espace S_d, nous introduisons l'estimateur du noyau de Dirichlet comme suit :

$$\widehat{f}_{n,b}(s) = \frac{1}{n}\sum_{i=1}^{n} K_{\frac{s}{b}+1,\,(1-\|s\|_1)b+1}(X_i), \quad s \in S_d. \tag{3.2}$$

Cet estimateur capture les propriétés de la distribution inconnue F en utilisant le paramètre de lissage b. Il calcule la moyenne pondérée des densités de Dirichlet ajustées aux points d'observation $X_1, \ldots, X_n$, où la forme de chaque noyau de Dirichlet est contrôlée par les caractéristiques de s et du paramètre de lissage[**Ouimet and Tolosana-Delgado, 2022**].

Exemples illustratifs de lissage

Deux exemples illustratifs de lissage de [**Ouimet and Tolosana-Delgado, 2022**]sont présentés dans la Figure 1.1 pour le cas où $d = 2$. Dans ces exemples, les densités cibles sont représentées par les deux sous-chiffres à gauche, tandis que les estimations correspondantes sont indiquées par les deux sous-chiffres à droite. Une observation importante est que la forme du noyau varie en fonction de la position s sur le simplex. Cette particularité diffère des estimateurs traditionnels, pour lesquels le noyau reste constant pour chaque point. Cette adaptation dynamique du noyau en fonction des variables permet aux estimateurs du noyau de Dirichlet (et plus généralement aux estimateurs asymétriques du noyau) de surmonter le problème de biais de convergence qui affecte les estimateurs du noyau classiques.

Notez que l'estimateur $\widehat{f}_{n,b}$ n'est pas une densité appropriée, car il ne s'intègre pas exactement à 1, mais il s'intègre asymptotiquement à 1. Pour le démontrer, nous introduisons

l'ensemble "Bulk" défini par :

$$\text{Bulk} = \left\{ x \in S_d : |x_i - r_i| \leq \frac{1}{\sqrt{1/b + d + 2}} \cdot \frac{b^{-1/6}}{2}, \quad \forall i \in [1, \ldots, d+1] \right\},$$

où $r_i = (s_i + b)/(1 + b(d+1))$, $x_{d+1} = 1 - \|x\|_1$, et $s_{d+1} = 1 - \|s\|_1$.

Ensuite, $X_1, \ldots, X_n \in$ Bulk avec une probabilité $1 - O(n \exp(b^{-1/3}/2))$ lorsque $n \longrightarrow \infty$, grâce à un argument simple basé sur l'union et la concentration. Ce résultat est valable tant que $b = o((\log n)^{-3})$, une hypothèse très faible étant donné que $b_{opt} \asymp n^{-2/(d+4)}$ selon [**Ouimet, 2022**].

Par conséquent, en utilisant l'approximation normale multivariée pour la densité de la distribution de Dirichlet ($\alpha = s + b, \beta = 1 - \|s\|_1 + b$) dérivée dans [**Ouimet, 2022**] , nous pouvons établir que, lorsque $n \longrightarrow \infty$,

$$\begin{aligned}
\int_{S_d} \widehat{f}_{n,b}(s) ds &= \frac{1}{n} \sum_{i=1}^{n} \int_{S_d} K_{\frac{s}{b}+1,(1-\|s\|_1)b+1}(X_i) ds \\
&= \frac{1}{n} \sum_{i=1}^{n} \int_{S_d} K_{\frac{s}{b}+1,(1-\|s\|_1)b+1}(X_i) \mathbf{1}_{\text{Bulk}}(X_i) ds + o - P(1) \\
&= \frac{1}{n} \sum_{i=1}^{n} \int_{\mathbb{R}^d} \frac{\exp\left(-\frac{1}{2} \delta_{x_i}^{-T} \sum_r^{-1} \delta_{X_i}\right)}{\sqrt{(2\pi)^d \frac{(1-\|r\|_1) \prod_{i=1}^d r_i}{(1/b+d+2)^d}}} ds + oP(1) \\
&= \frac{1}{n} \sum_{i=1}^{n} \int_{\mathbb{R}^d} \frac{\exp\left(1 - \frac{1}{2} z^t z\right)}{\sqrt{(2\pi)^d}} dz + o_{\mathbb{P}}(1) = 1 + o_{\mathbb{P}}(1).
\end{aligned}$$

En utilisant $\delta_{X_i} = \dfrac{X_i - r}{(1/b + d + 2)^{-1/2}}$ et $\sum_r = diag(r) - rr^T$, on peut élucider l'expression précédente en prenant en compte le changement de variables $z = \sum_r^{-1/2} \delta_{X_i}$. Cette considération est renforcée par le fait que la matrice $\sum_r$ est une matrice définie positive symétrique, ayant un déterminant égal à $(1 - \|r\|_1) \prod_{i=1}^d r_i$ par [**Tanabe and Sagae, 1992**].

Définition 3.1.1 *[**Ouimet and Tolosana-Delgado, 2022**] Les noyaux de Dirichlet se présentent sous forme de noyaux associés multivariés continus. Dans un contexte continu, les noyaux associés multivariés sont définis comme des fonctions de densité $K_{x,H}$ qui ont leur support sur un sous-ensemble de $[0, \infty)^d$, désigné par $\mathbb{T}_d^+$. Ces noyaux sont paramétrés par des points x situés sur le support et une matrice de lissage H qui est à la fois symétrique et définie positive. La matrice H peut prendre différentes formes : pleine, diagonale ou de type Scott (c'est-à-dire paramétrée par un unique paramètre h). La fonction $K_{x,H}$ doit obéir à la propriété fondamentale suivante : lorsque nous considérons un vecteur aléatoire $Z_{x,H}$ de dimension d distribué selon $K_{x,H}$, alors à mesure que $H \longrightarrow 0_{d\times d}^+$, les caractéristiques suivantes tendent vers :*

$$\mathbb{E}[Z_{x,H}] - x = a(x, H) \longrightarrow 0_d$$

$$Cov(Z_{x,H}) = B(x, H) \longrightarrow 0_{d\times d}^+$$

(Cette définition s'applique naturellement également au cas discret, et il est possible d'avoir un mélange de composantes discrètes et continues pour $Z_{x,H}$)[**Kokonendji and Somé, 2018**]. Conformément à cette définition, et avec des hypothèses de régularité relativement modestes sur la densité cible f, les propriétés asymptotiques du biais et de la variance au niveau ponctuel pour l'estimateur correspondant, noté $\tilde{f}_{n,H} = n^{-1}\sum_{i=1}^{n} K_{x,H}(X_i)$, ont été démontrées par [**Kokonendji and Somé, 2021**], pour un $x \in \mathbb{T}_d^+$ donné :

$$\mathbb{B}ias[\tilde{f}_{n,H}] = \nabla f(x)^T a(x,H) + \frac{1}{2}\mathrm{tr}[\chi_f(x)[B(x,H) + a(x,H)^T a(x,H)]] + o(\mathrm{tr}(B(x,H))) \quad (3.3)$$

$$\mathbb{V}ar[\tilde{f}_{n,H}] = n^{-1} f(x) \int_{\mathbb{T}_d^+} K_{x,H}^2(\mu) d\mu + o_x\left(\frac{n^{-1}}{(det(H))^{r_x}}\right), \quad (3.4)$$

(où x doit être à l'intérieur de $\mathbb{T}_d^+$) avec $r_x = \liminf_{n\longrightarrow\infty} \int_{\mathbb{T}_d^+} K_{x,H}^2(\mu) d\mu + (det(H))^p > 0$.

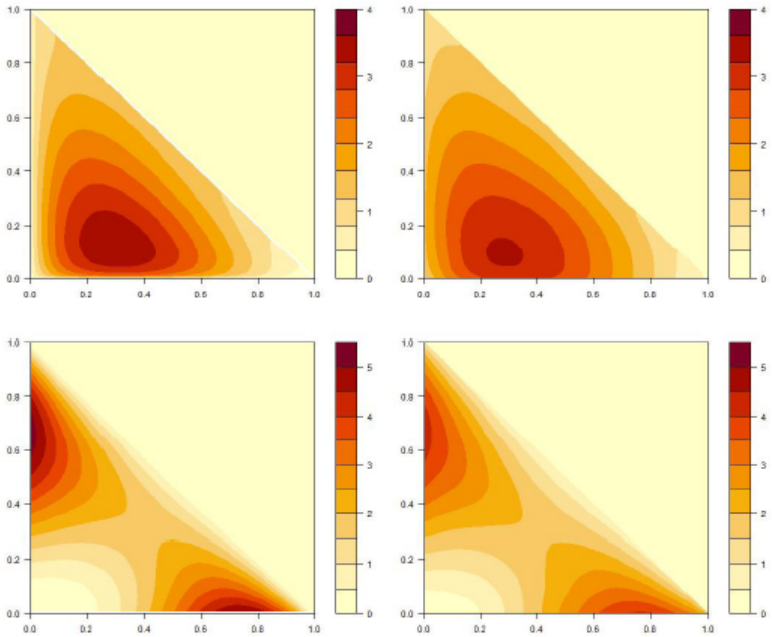

FIG. 3.1 – Exemples illustratifs de lissage pour $d = 2$.

Afin d'obtenir des expressions explicites pour le biais et la variance au niveau ponctuel de l'estimateur du noyau de Dirichlet, en utilisant les équations (1.3) et (1.4), nous pouvons estimer l'espérance et les covariances du vecteur aléatoire $\xi_s = (\xi_1, \ldots, \xi_d) \sim$ dirichlet $\left(\frac{s}{b} + 1, \frac{(1-\|s\|_1)}{b} + 1\right)$, pour $s \in S_d$. Pour tous $i, j \in [1, ..., d]$, des calculs simples conduisent aux expressions suivantes (voir [**Ouimet and Tolosana-Delgado, 2022**]), notées ($\bigstar$), :

$$\mathbb{E}[\xi_s] =^{(\bigstar)} \frac{(s_i/b) + 1}{(1/b) + d + 1} = \frac{s_i + b}{1 + b(d+1)} = s_i + b(1 - (d+1)s_i) + o(b^2). \quad (3.5)$$

$$\mathbb{C}ov =^{(\bigstar)} \frac{(s_i/b)+1)(((1/b)+d+1)\mathbb{I}_{[i=j]}-((s_j/b)+1))}{((1/b)+d+1)^2((1/b)+d+2)} = bs_i(\mathbb{I}_{[i=j]}-s_j)+O(b^2). \quad (3.6)$$

$$\mathbb{E}[(\xi_i-s_i)(\xi_j-s_j)] = \mathbb{C}ov(\xi_i,\xi_j)+(\mathbb{E}[\xi_i]-s_i)(\mathbb{E}[\xi_j]-s_j) = bs_i(\mathbb{I}_{[i=j]}-s_j)+O(b^2). \quad (3.7)$$

3.1.1 Principaux résultats

Chacun des résultats exposés dans cette section repose sur l'une des deux hypothèses suivantes :

1. La continuité Lipschitz de la densité f sur S_d. (1)
2. La double différentiabilité continue de la densité f sur S_d. (2)

Les indices illustrent comment les paramètres influencent les taux de convergence. Pour simplifier, nous utilisons la notation $[d] = (1,...,d)$ en plusieurs occasions. La largeur de bande $b = b(n)$ est toujours implicitement fonction du nombre d'observations. Nous notons l'espérance de $\widehat{f}_{n,b}(s)$ par :

$$\begin{aligned} f_b(s) &= \mathbb{E}[\widehat{f}_{n,b}(s)] \\ &= \mathbb{E}[K_{s/b+1,(1-\|s\|_1)/b+1}(X)] \\ &= \int_{S_d} f(x)K_{s/b+1,(1-\|s\|_1)/b+1}(x)\,dx \end{aligned}$$

Sinon, il est important de noter que si $\xi_s \sim \text{Dirichlet}(s/b+1,(1-\|s\|_1)/b+1)$, nous avons également la représentation suivante :

$$f_b(s) = \mathbb{E}[f(\xi_s)]$$

Les expressions asymptotiques des biais et variances ponctuels pour les estimateurs à noyau bêta ont été initialement déterminées par [**Chen, 1999**]. Le théorème ci-dessous étend cette analyse au contexte multidimensionnel.

Théorème 3.1.1 ***(biais et variance ponctuelle) :[Ouimet and Tolosana-Delgado, 2022]***

Supposons que la condition (2) soit vérifiée. À mesure que la taille de l'échantillon n tend vers l'infini, de manière uniforme pour s dans l'ensemble S_d, la relation suivante devient apparente :

Pour chaque b, nous observons que lorsque n tend vers l'infini :

$$\mathbb{B}iai[\widehat{f}_{n,b}(s)] = f_b(s)-f(s) = bg(x)+o(b)$$

Ici, g(s) peut être défini comme suit :

$$g(s) = \sum_{i\in[d]}(\mathbb{I}-(d+1)s_i)\frac{\partial}{\partial_{s_i}}f(s)+\frac{1}{2}\sum_{i,j\in[d]}s_i(1_{(i=j)}-s_j\frac{\partial^2}{\partial_{s_i}\partial_{s_j}}f(s))$$

En considérant le maintien de la condition (1), définissons des fonctions pour différents sous-ensembles d'indices $\jmath \subseteq [d]$:

$$\psi(s) = \psi_{\emptyset}(s) \text{ et } \psi_{\jmath}(s) = \left[(4\pi)^{d-|\jmath|} \cdot (1 - \|s\|_1) \prod_{i \in [d]/\jmath} s_i\right]^{-1/2}$$

De plus, pour tout $s \in int(S_d)$, tout sous-ensemble non vide $\jmath \subseteq [d]$, et tout $K \in (0, \infty)^d$, lorsque n tend vers l'infini, la variance se comporte comme suit :

$$Var(\widehat{f}_{n,b}(s)) = \begin{cases} n^{-1}b^{-d/2}\left(\psi(s)f(s) + O_s(b^{1/2})\right), \\ si \quad \frac{s_i}{b} \longrightarrow \infty \text{ pour tout } i \in [d] \text{ et} \\ \quad (1 - \|s\|_1)/b \longrightarrow \infty, \\ \\ n^{-1}b^{-(d+|\jmath|)/2}\left(\psi_{\jmath}(s)f(s)\prod_{i\in\jmath}\frac{\Gamma(2K_i+1)}{2^{2K_j+1}\Gamma^2(K_i+1)} + O_{K,S}(b^{1/2})\right), \\ si \quad \frac{s_i}{b} \longrightarrow K_i \text{ pour tout } i \in \jmath, \\ \quad \frac{s_i}{b} \longrightarrow \infty \text{ pour tout } i \in [d]/\jmath, \\ \quad \text{et } (1 - \|s\|_1)/b \longrightarrow \infty. \end{cases}$$

Ceci signifie que la variance ponctuelle se comporte comme $O_s(n^{-1}b^{-d/2})$ à l'intérieur du simplex, et cette valeur est multipliée par un facteur de $b^{-1/2}$ lorsque nous approchons de la frontière dans l'une des dimensions d. Lorsque nous sommes près d'un bord de dimension $d - |\jmath|$, alors la variance ponctuelle devient $O_s(n^{-1}b^{-(d+|\jmath|)/2})$.

Corollaire 3.1.1 ***Erreur quadratique moyenne [Ouimet and Tolosana-Delgado, 2022]***
Supposons que (2) soit vérifié. Lorsque $n \longrightarrow \infty$ et pour chaque $s \in int(S_d)$, on a :

$$\begin{aligned} MSE[\widehat{f}_{n,b}(s)] &= [\mathbb{E}|\widehat{f}_{n,b}(s) - f(s)|^2] \\ &= \mathbb{V}ar(\widehat{f}_{n,b}(s)) + (\mathbb{B}ias[\widehat{f}_{n,b}(s)])^2 \\ &= n^{-1}b^{-d/2}\psi(s)f(s) + b^2g^2(s) + O_s(n^{-1}b^{-d/2+1/2}) + o(b^2). \end{aligned}$$

En particulier, si $f(s), g(s) \neq 0$, le choix asymptotiquement optimal de b, en ce qui concerne le MSE, est donné par :

$$b_{opt}(s) = n^{-2/(d+4)}\left[\frac{d}{4}.\frac{\psi(s)f(s)}{g^2(s)}\right]^{-2/(d+4)},$$

avec

$$MSE[\widehat{f}_{n,b_{opt}}] = n^{-2/(d+4)}\left[\frac{1+\frac{d}{4}}{\left(\frac{d}{4}\right)^{\frac{d}{d+4}}}\right]\frac{(\psi(s)f(s))^{4/(d+4)}}{(g^2(s))^{-d/(d+4)}} + o_s(n^{-4/(d+4)}).$$

De manière plus générale, si $n^{-2/(d+4)}b \longrightarrow \lambda$ pour certains $\lambda > 0$ lorsque $n \longrightarrow \infty$, alors :

$$MSE[\widehat{f}_{n,b}(s)] = n^{-4/(d+4)}\left[\lambda^{-d/2}\psi(s)f(s) + \lambda^2 g^2(s)\right] + o_s(n^{-4/(d+4)}).$$

En intégrant le MSE et en montrant que la contribution provenant de points proches de la limite est négligeable, nous obtenons le résultat suivant.

Théorème 3.1.2 ***(Erreur Quadratique Intégrée Moyenne)[Ouimet and Tolosana-Delgado, 2022]***

Supposons que (2) soit vérifié. À mesure que $n \longrightarrow \infty$, nous avons :

$$\begin{aligned} MISE[\widehat{f}_{n,b}(s)] &= \int_{S_d} [\mathbb{E}|\widehat{f}_{n,b}(s) - f(s)|^2]\, ds \\ &= n^{-1}b^{-d/2}\int_{S_d}\psi(s)f(s)\,ds + b^2\int_{S_d} g^2(s)\,ds + o(n^{-1}b^{-d/2}) + o(b^2). \end{aligned}$$

En particulier, si $\int_{S_d} g^2(s)\,ds > 0$, le choix asymptotiquement optimal de b en termes de MISE est donné par :

$$b_{opt}(s) = n^{-2/(d+4)}\left[\frac{d}{4}\cdot\frac{\int_{S_d}\psi(s)f(s)\,ds}{\int_{S_d} g^2(s)\,ds}\right]^{-2/(d+4)},$$

avec :

$$MISE[\widehat{f}_{n,b_{opt}}] = n^{-4/(d+4)}\left[\frac{1+\frac{d}{4}}{\left(\frac{d}{4}\right)^{\frac{d}{d+4}}}\right]\frac{\left(\int_{S_d}\psi(s)f(s)\,ds\right)^{4/(d+4)}}{\left(\int_{S_d} g^2(s)\,ds\right)^{-d/(d+4)}} + o(n^{-4/(d+4)}).$$

De manière plus générale, si $n^{-2/(d+4)}b \longrightarrow \lambda$ pour certains $\lambda > 0$ lorsque $n \longrightarrow \infty$, alors :

$$MISE[\widehat{f}_{n,b}(s)] = n^{-4/(d+4)}[\lambda^{-d/2}\int_{S_d}\psi(s)f(s)\,ds + \lambda^2\int_{S_d} g^2(s)\,ds] + o(n^{-4/(d+4)})$$

.

3.2 noyau normale

Actuellement, parmi les choix privilégiés pour le noyau dans un estimateur de densité multivariée, la densité normale multivariée standard (connue sous le nom de noyau gaussien bivarié) jouit d'une grande popularité. Cependant, le paramètre de lissage h peut être étendu à une matrice symétrique et définie positive **H**, désignée comme matrice de lissage [**Duong and Hazelton, 2005**] ont introduit la distribution normale sur S^D en reliant la fonction de densité des coordonnées orthonormales. L'idée est simple : une composition aléatoire est considérée comme possédant une distribution normale sur S^D si les coordonnées orthonormales correspondantes, notées $\text{ilr}(x)$, suivent la densité normale standard dans l'espace réel $\mathbb{R}^{D-1}$, symbolisée par ϕ.

Dans cette section, nous proposons le noyau gaussien sur S^D, baptisé noyau iln, défini comme suit :

$$K_{iln}(x|X,H) = \phi_H(\text{ilr}(x) - \text{ilr}(X)), \quad x \in S^D$$

où :

— ilr(x) et ilr(X) représentent les coordonnées orthonormales respectives de x et X.

— $\phi_H(Y) = (2\pi)^{-(D-1)/2}|H|^{-1/2}\exp\left[-\frac{1}{2}Y^T H^{-1}Y\right]$, avec $Y \in \mathbb{R}^{D-1}$.

La matrice de lissage **H** est une matrice symétrique définie positive d'ordre $(D-1)\times(D-1)$. Ce paramètre de lissage général est connu sous le nom de matrice de lissage complète[**Chacón et al., 2011**].

3.2.1 Choix de la matrice de lissage

La Figure 3.2 de met en relief les contours des noyaux pour $D = 3$ à la fois dans l'espace des coordonnées et dans le simplexe S_3. Ces contours illustrent la configuration typique de cercles et d'ellipses caractéristique de la distribution normale multivariée dans l'espace des coordonnées. Concernant le processus de centrage, le noyau *iln* se conforme à l'équation suivante :

$$\mathbf{K}_{iln}(x|X,H) = \mathbf{K}_{iln}(x \ominus X|e,H)$$

Cela reflète une opération similaire à la soustraction vectorielle dans le domaine réel.

En ce qui concerne la structure de variance, l'utilisation du noyau *iln* permet une distribution plus étendue de la probabilité autour du point X. Contrairement à l'approche utilisant un unique paramètre de largeur de bande h, qui applique un lissage uniforme le long de toutes les directions de coordonnées, l'usage d'une matrice de largeur de bande complète H permet des niveaux de lissage distincts dans différentes directions, même en dehors des axes de coordonnées.

L'ensemble F englobe les matrices carrées d'ordre d, symétriques et définies positives. En général, une matrice H appartenant à F possède $\frac{1}{2}d(d+1)$ éléments indépendants, ce qui peut entraîner une multitude de paramètres de lissage à sélectionner, même pour des dimensions spatiales raisonnables. Cependant, en imposant des contraintes sur H, diverses simplifications peuvent être obtenues, entraînant l'examen de différents scénarios en fonction des restrictions appliquées à H au sein de l'ensemble F(voir et [**Chacón et al., 2011**].

Cas 1 : Une approche simple consiste à prendre H appartenant à l'ensemble S, où S est défini comme $S = \{H \in F \mid H = h^2 I, h > 0\}$. En d'autres termes, un même paramètre de lissage h est choisi pour chaque axe de coordonnées dans l'espace. Cela définit l'estimateur suivant :

$$\widehat{f}(x;h) = \frac{1}{nh^d}\sum_{i=1}^{n} K\left(\frac{x_1 - X_1^{(i)}}{h}, \ldots, \frac{x_d - X_d^{(i)}}{h}\right), \quad x = (x_1,\ldots,x_d)^T \in \mathbb{R}^d$$

Cas 2 : Une autre option est de considérer H appartenant à l'ensemble D, où D est défini comme $D = \{H \in F \mid H = \text{diag}(h_1^2,\ldots,h_d^2); h_i > 0 \text{ pour } i = 1,\ldots,d\}$. Cela conduit à la formulation de l'estimateur suivant :

$$\widehat{f}(x;h_1,\ldots,h_d) = \frac{1}{n\left(\sum_{l=1}^{d} h_l\right)^{-1}}\sum_{i=1}^{n} K\left(\frac{x_1 - X(i)_1}{h_1}, \ldots, \frac{x_d - X(i)_d}{h_d}\right), \quad x = (x_1,\ldots,x_d)^T \in \mathbb{R}^d$$

Cas 3 : Dans certaines situations, le lissage peut être nécessaire dans des directions différentes de celles définies par les axes de coordonnées. Dans ce cas, une matrice H appartenant à l'ensemble A est envisagée, où A est défini comme $A = F - D$. Cela permet de traiter des cas où le lissage doit être effectué le long d'axes distincts.

Remarque 3.2.1 *Il est essentiel de noter que les ensembles S, D et A respectent les relations d'inclusion suivantes : $S \subset D \subset F$ et $A \subset F$.*

Pour illustrer ces concepts, plongeons-nous dans le contexte bivarié :

La Figure 3.2 dépeint les contours-plotspour le noyau gaussien bivarié avec diverses paramétrisations, présentés de gauche à droite :

- $(a) : H \in S$
- $(b) : H \in D - S$
- $(c) : H \in A$

Tout d'abord, nous observons que dans chaque cas, les contours du noyau adoptent une forme elliptique. Cependant, lorsque $H \in S$, ces ellipses se transforment en cercles. En revanche, pour $H \in D - S$, les ellipses s'alignent sur les axes de coordonnées. Dans le scénario où $H \in A$, des ellipses à orientations variées se manifestent.

Comparativement, il existe un intérêt substantiel à évaluer les performances de l'estimateur en fonction des différentes paramétrisations. Cependant, il est évident qu'une certaine perte d'efficacité est inévitable.

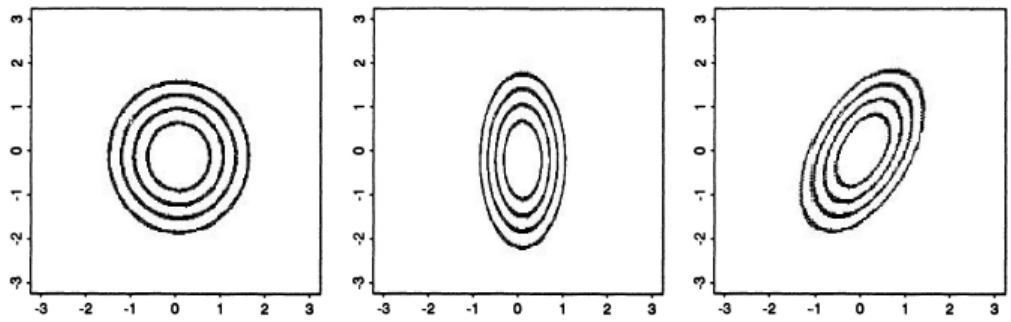

FIG. 3.2 – LesLes contours-plots du noyau gaussien

Remarque 3.2.2 *[**Aitchison and Lauder, 1985**] ont constaté que les paramètres usuels pour la largeur de lissage de la matrice H ne fonctionnent pas avec le noyau aln en raison de la transformation alr, ce qui rend les résultats sensibles aux permutations des composants. Ils ont recommandé de vérifier l'invariance par rapport aux permutations des composants et à la variabilité du diviseur dans la transformation alr. Pour résoudre ce problème, ils ont suggéré d'utiliser une matrice de lissage $H = hS$, proportionnelle à la matrice de covariance S de l'échantillon transformé. On souligne que cette méthode convient uniquement aux données normales multivariées et n'est pas adaptée à des densités plus générales dans l'espace réel. Ainsi, le noyau aln est approprié pour la classe aln, mais en utilisant leur approche avec le noyau iln, différentes formes de densités peuvent être estimées avec divers paramètres de matrice de lissage H [**Chacón et al., 2011**]* • .

Remarque 3.2.3 *Il est crucial de souligner l'invariance envers les variations de la base orthonormale pour le noyau iln. Les vecteurs ilr, utilisés dans notre travail, peuvent être interprétés comme des coordonnées en relation avec une base orthonormale établie sur le simplex. Le vecteur ilr mentionné dans l'équation (1.3), issu du premier chapitre et obtenu en utilisant une base orthonormale spécifique , aurait tout aussi bien pu être calculé avec une autre base orthonormale. Par conséquent, dans toute analyse de composition impliquant des vecteurs ilr, il est impératif de confirmer l'invariance lors des changements de bases orthonormales[**Chacón et al., 2011**].*

3.3 noyau logistique-normal

[**Aitchison, 1982**], Identifié des catégories de distributions pertinentes sur le simplexe S^d, dont la complexité est comparable à la classe de distributions normales d-dimensionnelles $N^d(\mu, \Sigma)$. Cela est rendu possible en induisant des distributions sur S^d depuis la classe normale multivariée $N^d(\mu, \Sigma)$ dans $\mathbb{R}^d$ par une transformation appropriée entre $\mathbb{R}^d$ et S^d.

Par exemple, une de ces transformations, appliquée à $Y \in \mathbb{R}^d$ pour obtenir $x \in S^d$, est la transformation logistique additive :

$$x_i = \frac{\exp(y_i)}{[\exp(y_1) + \ldots + \exp(y_d) + 1]} \quad (i = 1, \ldots, d) \tag{3.8}$$

avec sa transformation inverse, le log-ratio :

$$y_i = \log\left(\frac{x_i}{x_{d+1}}\right) \quad (i = 1, \ldots, d) \tag{3.9}$$

où :

$$x_{d+1} = 1 - x_1 - \ldots - x_d.$$

Ceci donne lieu à la classe logistique-normale additive $L^d(\mu, \Sigma)$, caractérisée par une densité typique :

$$L(x|\mu, \Sigma) = (x_1 \ldots x_{d+1})^{-1}\phi(y|\mu, \Sigma)$$

où $\phi(y|\mu, \Sigma)$ est la densité de la distribution $N^d(\mu, \Sigma)$ évaluée en y.

Supposons que l'ensemble de données de compositions C soit constitué de n compositions $(X_1, ..., X_n)$, et notons $Y_1, \ldots, Y_n$ les compositions log-ratio correspondantes définies par (3.7). Nous désignerons par x une composition typique dans S^d. Afin d'appliquer une méthode d'estimation de densité par noyau, nous devons d'abord définir un noyau $K(x|X, \lambda)$, une fonction de densité définie pour $x \in S^d$, centrée autour du point de données typique X, et permettant le choix approprié du facteur de lissage λ.

Ensuite, la fonction de densité ajustée à x est donnée par :

$$p(x|C, \lambda) = \frac{1}{n}\sum_{j=1}^{n} K(x|X_j, \lambda) \tag{3.10}$$

où nous déterminons la valeur de lissage à l'aide de la méthode de [**Habbema et al., 1974**] comme étant le λ qui maximise la pseudo-vraisemblance :

$$\prod_{i=1}^{n} \frac{1}{n-1}\sum_{j\neq i} K(X_i|X_j, \lambda) \tag{3.11}$$

Notre première étape consiste donc à examiner le choix approprié des noyaux, et à cette fin, nous nous tournons naturellement vers les classes bien connues de fonctions de densité sur le simplexe.

[**Aitchison, 1982**],Étant donné que les compositions log-ratio $Y_1, \ldots, Y_n$ sont des vecteurs dans $\mathbb{R}^d$, nous pouvons envisager l'estimation de densité du noyau dans S^d en suivant les étapes suivantes :

1. Dans une première étape, nous utilisons la méthode classique pour estimer la densité du noyau en Y dans $\mathbb{R}^d$, en utilisant les ensembles de données $Y_1, \ldots, Y_n$. Cela est accompli avec un noyau multivarié $\phi(y|Y, \lambda T)$, où T est une matrice de covariance appropriée.

2. Dans une deuxième étape, nous appliquons la transformation logistique de $y \in \mathbb{R}^d$ à $x \in S^d$ pour obtenir une estimation de la densité dans S^d.

Il peut être démontré que cette procédure est équivalente à l'utilisation d'un noyau sur S^d défini par :

$$K(x|X, \lambda) = (x_1 \ldots x_{d+1})^{-1} \phi(y|Y, \lambda T) \tag{3.12}$$

$$= L(x|Y, \lambda T) \tag{3.13}$$

Cependant, il est essentiel de prendre des précautions lors de cette approche. Le choix de la matrice de covariance T doit être fait judicieusement. À première vue, choisir $T = I_d$ pourrait sembler approprié, correspondant à l'utilisation d'un noyau de distribution normale sphérique. Une autre option courante pourrait être $T = \text{diag}(t_1^2, ..., t_d^2)$, où $t_1, ..., t_d$ représente l'écart type estimé de la c-ème composante des vecteurs de composition log-ratio $Y_1, ..., Y_n$. Cette approche implique l'utilisation d'un noyau de distribution normale multivariée avec des composantes indépendantes (Hermans et Habbema, 1976). Cependant, il est crucial de noter que cette méthode présente un inconvénient majeur, à savoir qu'elle n'est pas invariante aux permutations des composantes. En d'autres termes, elle n'est pas robuste face aux choix du diviseur dans la transformation log-ratio (3.7)[**Aitchison, 1982**].

3.4 Conclusion

En conclusion, ce chapitre ouvre la voie à une exploration approfondie des approches à noyaux pour la modélisation de densité dans l'espace simplex S_d. La compréhension de ces méthodes fournira aux chercheurs et aux analystes des outils précieux pour appréhender et interpréter des données multivariées complexes, tout en tenant compte des particularités inhérentes aux compositions sur S_d.

Chapitre 4

Simulation et Analyse de Données de Composition avec les Trois Noyaux

4.1 Introduction

Dans ce chapitre, nous plongeons dans l'univers stimulant de la simulation de données de composition, en nous concentrant sur trois échantillons distincts (A, B et C) de dimensions variées, ainsi que sur l'utilisation de trois types de noyaux pour estimé la densité de probabilité : le noyau de Dirichlet, le noyau gaussien multivarié et le noyau logistique normal. Notre objectif est d'évaluer les performances de ces noyaux en calculant le RMSE (Root Mean Squared Error) à l'aide de paramètres de lissage H optimisés par une méthode de validation croisée.

4.2 Le choix de langage de programmation

Dans ce mémoire, le langage de programmation R, version 4.3.1, a été employé pour mener à bien la simulation de divers systèmes distincts.

Le langage R se positionne en tant qu'environnement et langage de programmation spécifiquement orienté vers l'analyse statistique, la manipulation de données et la création de visualisations graphiques. Sa popularité est répandue parmi les chercheurs, les statisticiens et les praticiens en science des données, car il constitue un environnement propice pour la réalisation d'analyses sophistiquées, la formulation de modèles statistiques complexes et la conception de représentations visuelles expressives, toutes destinées à éclairer les processus décisionnels.

Le langage R se caractérise par son statut open-source, mettant en évidence son accessibilité universelle et encourageant son adaptation par la communauté. Son origine trouve sa source dans la création initiée par Ross Ihaka et Robert Gentleman à l'Université d'Auckland, en Nouvelle-Zélande, suivi par son développement collaboratif par une communauté internationale de développeurs.

4.3 Méthodologie

La méthodologie adoptée dans cette étude intègre l'utilisation de trois ensembles de données distincts, désignés sous les noms A, B et C, dont la création a été initiée grâce à l'utilisation de la fonction `set.seed()`. L'échantillon A, composé de 50 observations, se déploie dans un espace de deux dimensions. Pour l'échantillon B, rassemblant 100 observations, son espace s'étend à trois dimensions. Enfin, l'échantillon C, qui totalise 100 observations, opère dans un espace à quatre dimensions.

La démarche entreprise dans cette phase s'oriente autour de l'exploration de trois types de noyaux différents pour parvenir à l'estimation de la densité de probabilité multivariée. Le noyau de Dirichlet, intrinsèquement adapté aux données de composition, constitue la première pierre angulaire de cette démarche. En parallèle, le noyau gaussien multivarié et le noyau logistique normal viennent enrichir l'approche analytique, chacun apportant une perspective singulière à l'ensemble.

Une facette essentielle de cette méthodologie se manifeste à travers l'adoption de la validation croisée. Celle-ci vise à obtenir un paramètre de lissage H, spécifiquement calibré pour chaque type de noyau, dans le but précis de minimiser le RMSE (*Root Mean Squared Error*). Cette phase de validation croisée joue un rôle pivot pour garantir l'équité dans la comparaison des noyaux, en réduisant les erreurs de prédiction et en optimisant ainsi les résultats obtenus.

En somme, la méthodologie adoptée juxtapose la création de jeux de données diversifiés avec l'expérimentation de trois catégories distinctes de noyaux pour aboutir à l'estimation de la densité de probabilité multivariée. La validation croisée, en sélectionnant des paramètres de lissage H adaptés à chaque noyau, ambitionne la réduction du RMSE tout en garantissant des résultats solides et significatifs.

4.3.1 Algorithme de simulation

1. **Importation des Bibliothèques :** Importez les bibliothèques nécessaires pour la simulation, telles que `MASS`,`kde2d`, `caret`, `ks` et `compositions`.
2. **Génération des Données :** Générez les trois ensembles données de composition aléatoires en utilisant le package `DirichletReg`[Maier, 2014] , et la la fonction standard de programme **R** `set.seed()`. Choisissez le nombre d'observations (n) et la dimension (d) pour chaque ensemble de données.
3. **Définition du Grid de Paramètres :** Définissez un grid de valeurs pour le paramètre de lissage (H) que vous souhaitez explorer.
4. **Définition de la Fonction d'Estimation :** Créez une fonction d'estimation qui prend comme entrée le paramètre de lissage (H), les échantillons d'entraînement et de test, et estime la densité de probabilité à l'aide d'un noyau spécifique (par exemple, noyau de Dirichlet).
5. **Définition de la Fonction de Perte :** Créez une fonction de perte qui calcule *RMSE* de cette estimation , en prenant en compte les éventuelles valeurs manquantes.
6. **effectuez La Validation Croisée :** Pour chaque valeur de H dans le grid, effectuez la validation croisée en estimant la densité de probabilité et en calculant la performance (RMSE).
7. **Sélection du Meilleur Paramètre :** Identifiez le paramètre de lissage (H) qui minimise la performance (RMSE) obtenue grâce à la validation croisée.

8. **Calcul de la Performance Finale :** Utilisez le paramètre de lissage optimal pour estimer la densité de probabilité sur l'ensemble des données. Calculez la performance finale en termes de RMSE.
9. **Visualisation des Résultats :** Tracez les boîtes (boxplots) de densité de probabilité estimé à l'aide de la fonction `ggplot2` . Affichez le paramètre de lissage optimal sélectionné ainsi que la performance (RMSE).

4.3.2 Les Packages Utilisé pour Faciliter la Simulation

Dans cette simulation , nous explorons l'utilisation judicieuse de packages logiciels pour faciliter la mise en œuvre de notre simulation de densité de probabilité multivariée. Ces packages, soigneusement sélectionnés pour leurs fonctionnalités spécifiques, ont grandement contribué à simplifier le processus de simulation, tout en nous fournissant des outils puissants pour analyser et évaluer les résultats obtenus.

1. Package `DirichletReg`[**Maier, 2014**]
   Ce package `DirichletReg` Spécialement conçu pour l'analyse de données de composition, ce package a facilité la génération de données aléatoires . Son utilisation nous a permis de créer des ensembles de données avec différentes dimensions et tailles, formant ainsi la base de notre expérimentation.
2. Package `caret`[**Kuhn et al., 2008**]
   Ce package `caret` a été un compagnon précieux lors de l'implémentation de la validation croisée pour la sélection des paramètres de lissage optimaux.
3. Package `MASS`[**Ripley et al., 2013**]
   ce package `MASS` s'est révélé essentiel pour ses fonctions de manipulation et d'analyse statistique avancée
4. Package `ks`[**Werner et al., 2012**]
   ce package `ks` nous a fourni des outils pour estimer la densité de probabilité à partir de nos données.
5. Package `compositions`[**van den Boogaart et al., 2013**]
   Grâce à package `compositions`,nous avons bénéficié d'outils spécifiquement conçus pour le traitement des données de composition
6. Package `ggplot2` [**Della Vedova et al., 2019**]
   Le package `ggplot2` joue un rôle crucial dans la simulation en permettant de générer des estimations visuelles de la densité de probabilité à partir des données.

4.4 Résultats Obtenu

Dans cette section, nous présentons les résultats obtenus après avoir examiné les trois échantillons distincts (A, B et C) de dimensions variées en utilisant les trois types de noyaux : le noyau de Dirichlet, le noyau gaussien multivarié et le noyau logistique normal. Les performances de chaque combinaison échantillon-noyau ont été évaluées en termes de RMSE (Root Mean Squared Error), et les paramètres de lissage H optimaux pour chaque cas ont été sélectionnés à l'aide de la méthode de validation croisée.

Tab. 4.1 – Valeurs du RMSE pour chaque échantillon et chaque noyau

Échantillon	Noyau	RMSE
A (d=2, n=50)	Dirichlet	0.519
	Logistique Normale	0.221
	Gaussienne Multivariée	4.415e-51
B (d=3, n=100)	Dirichlet	0.535
	Logistique Normale	0.352
	Gaussienne Multivariée	1.899e-128
C (d=4, n=100)	Dirichlet	0.463
	Logistique Normale	0.296
	Gaussienne Multivariée	2.234e-160

Tab. 4.2 – Valeurs optimales de H pour chaque échantillon et chaque noyau

Échantillon	Noyau	H optimal
A (d=2, n=50)	Dirichlet	0.31
	Logistique Normale	0.01
	Gaussienne Multivariée	0.9999
B (d=3, n=100)	Dirichlet	0.28
	Logistique Normale	0.01
	Gaussienne Multivariée	$\begin{bmatrix} 1.0000005 & 0.9999995 \\ 0.9999995 & 1.0000005 \end{bmatrix}$
C (d=4, n=100)	Dirichlet	32
	Logistique Normale	0.01
	Gaussienne Multivariée	$\begin{bmatrix} 1.0000007 & 0.9999997 & 0.9999997 \\ 0.9999997 & 1.0000007 & 0.9999997 \\ 0.9999997 & 0.9999997 & 1.0000007 \end{bmatrix}$

Ci-après, vous découvrirez les graphiques (plots de contour) qui présentent l'estimation de la densité de probabilité pour l'échantillon **A** en employant trois noyaux distincts :

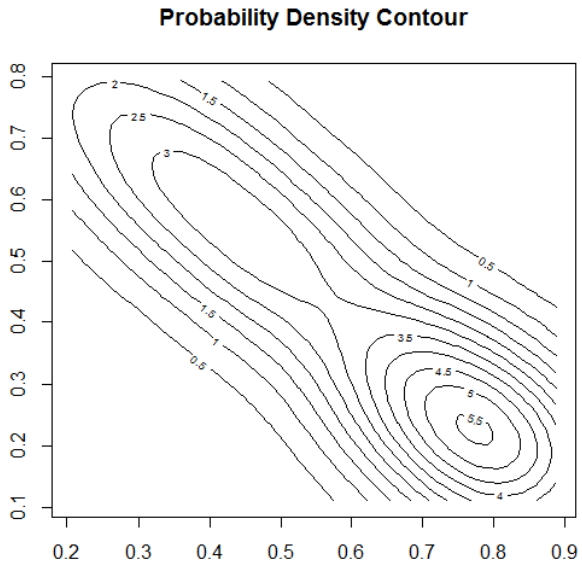

FIG. 4.1 – L'échantillon **A** avec le noyau de Dirichlet

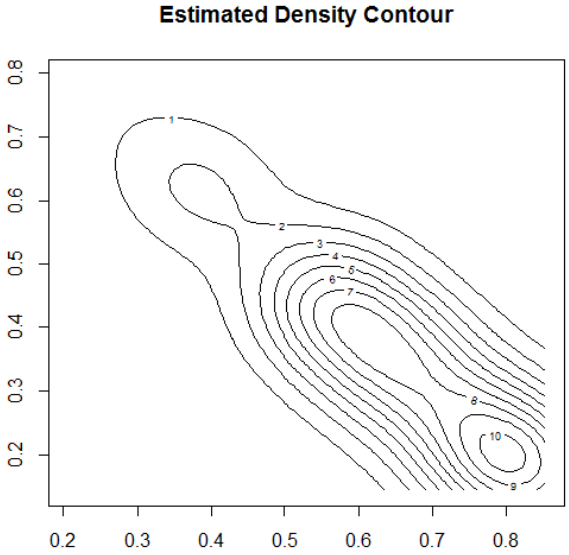

FIG. 4.2 – L'échantillon **A**avec le noyau logistique-normal

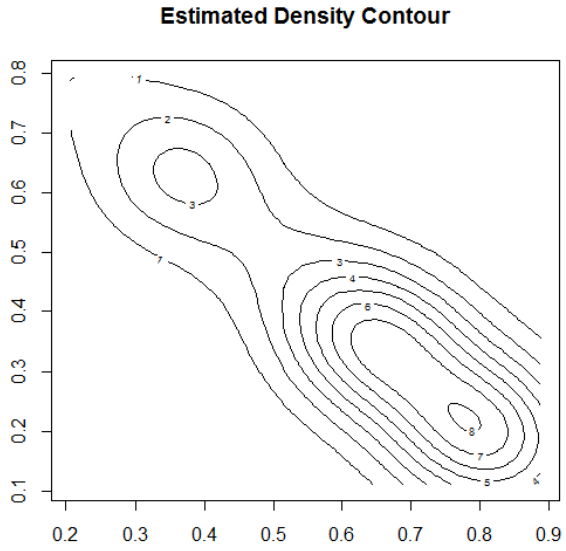

FIG. 4.3 – L'échantillon **A** avec le noyau normal

Voici les graphiques (contours) illustrant les estimations de densité de probabilité pour l'échantillon **B**, obtenues en appliquant trois noyaux distincts :

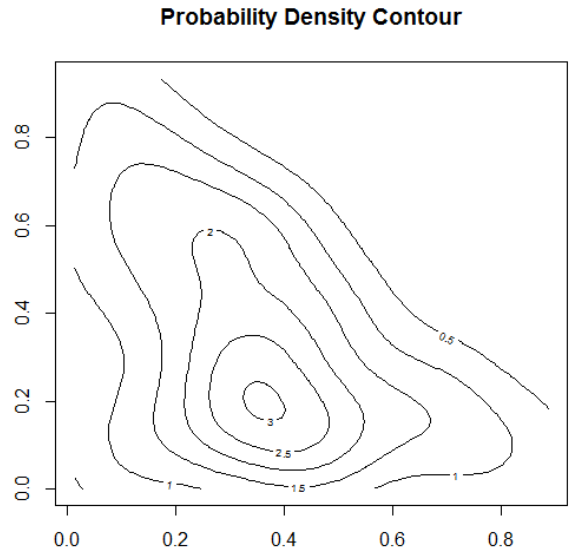

FIG. 4.4 – l'échantillon **B** avec le noyau de Dirichlet

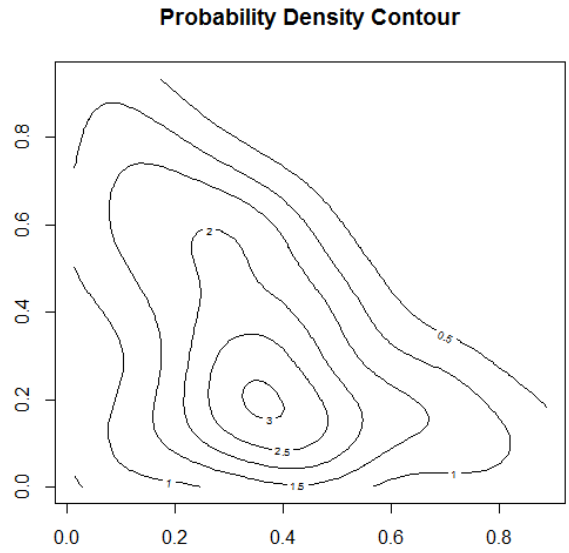

FIG. 4.5 – l'échantillon **B** avec le noyau logistique-normal

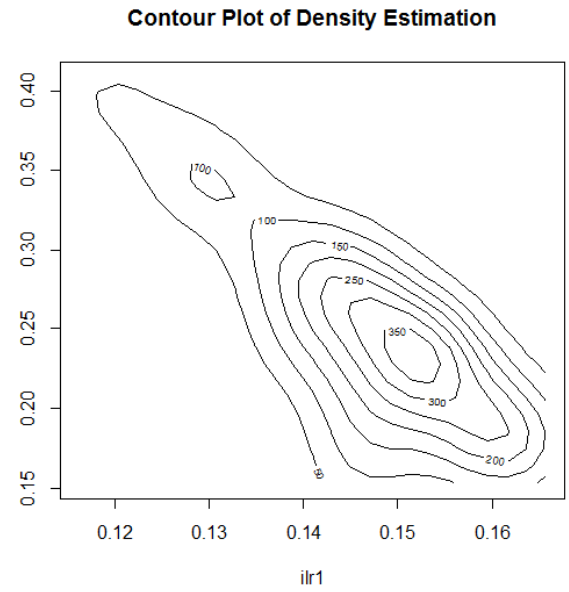

FIG. 4.6 – l'échantillon **B** avec le noyau normal

Les contours de l'estimation de densité de probabilité pour l'échantillon **C** sont finalement présentés en utilisant les trois noyaux différents :

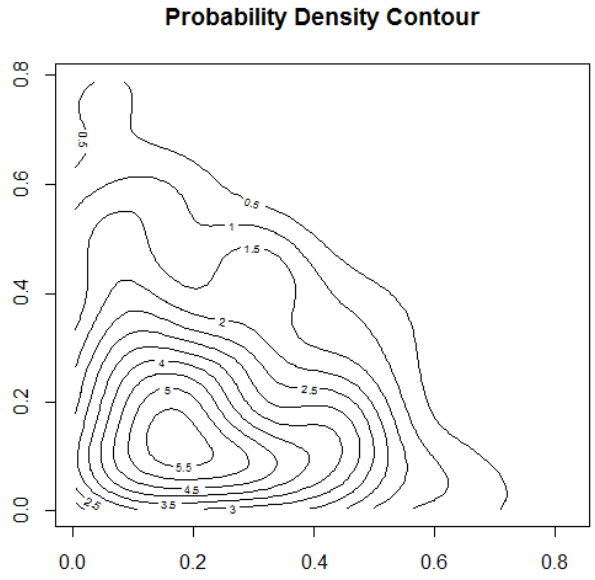

FIG. 4.7 – l'échantillon **C** avec le noyau de Dirichlet

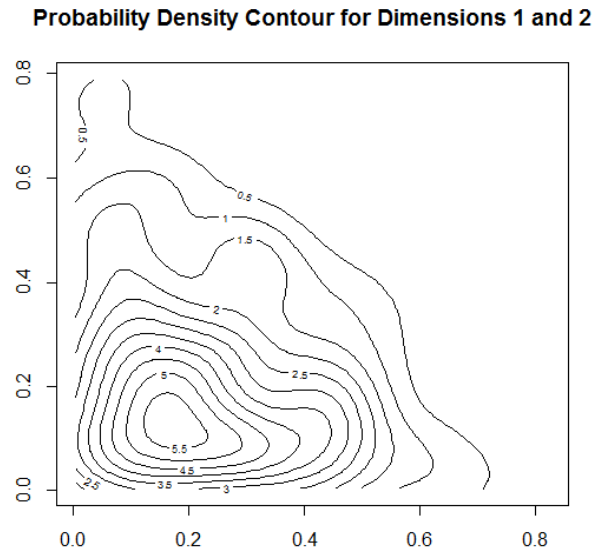

FIG. 4.8 – l'échantillon **C** avec le noyau logistique-normal

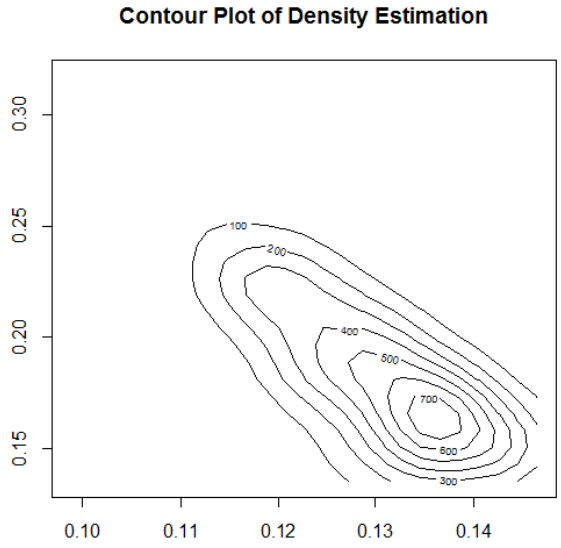

FIG. 4.9 – l'échantillon **C** avec le noyau normal

Ci-dessous, vous trouverez les graphiques en boîte (boxplots) illustrant les estimations de densité de probabilité estimé pour l'échantillon **A** en utilisant les trois différents noyaux :

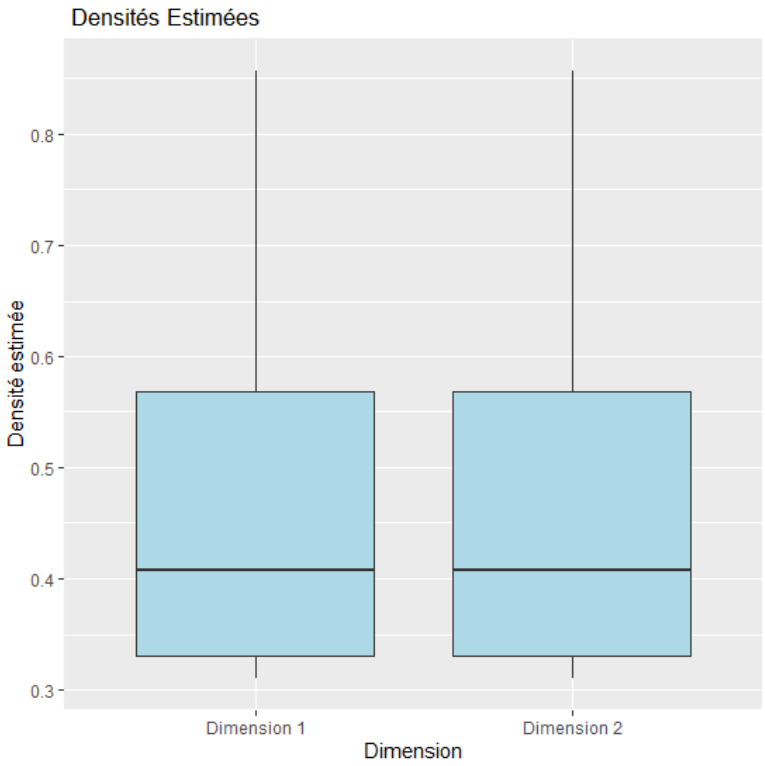

FIG. 4.10 – L'échantillon **A** avec le noyau de Dirichlet

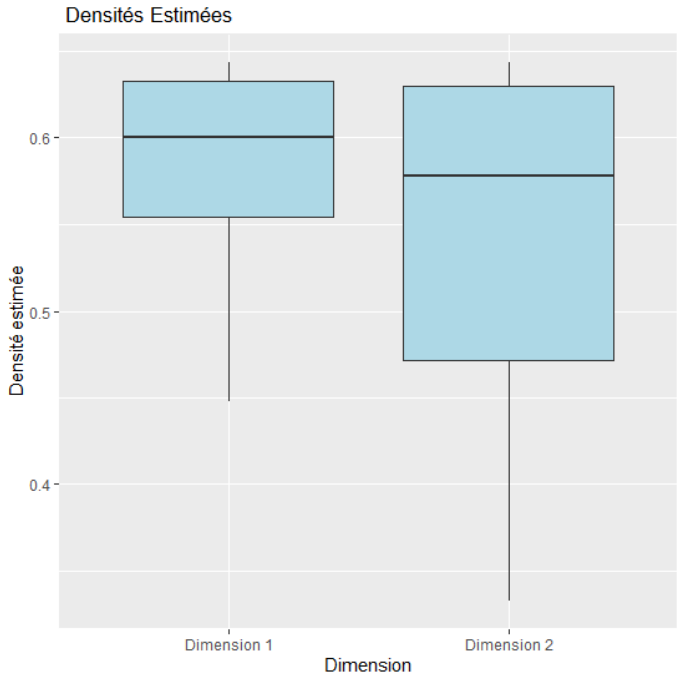

FIG. 4.11 – L'échantillon **A**avec le noyau logistique-normal

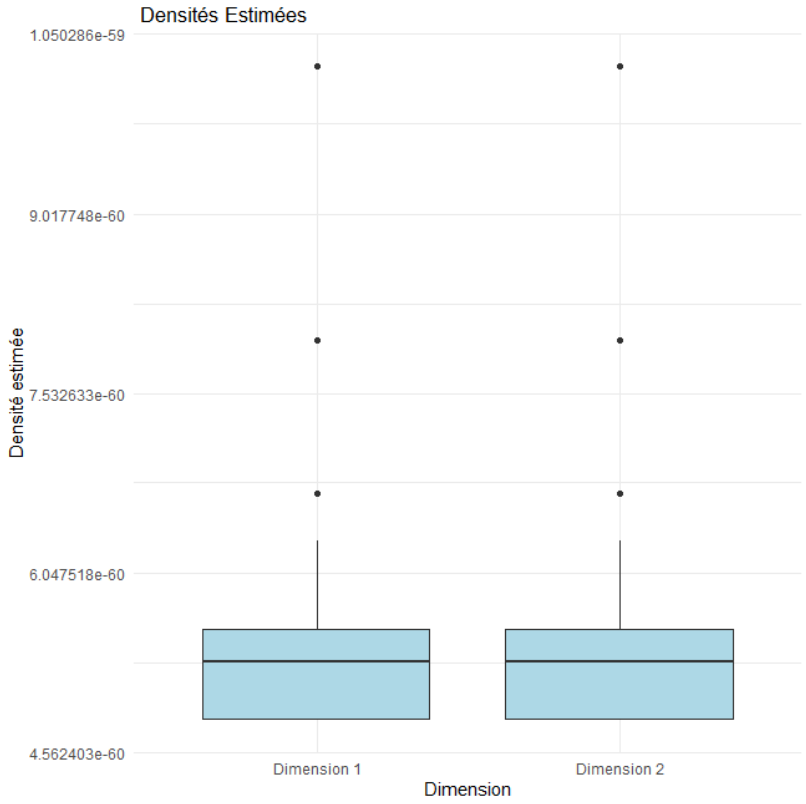

FIG. 4.12 – L'échantillon **A** avec le noyau normal

Voici les graphiques en boîte représentant les estimations de densité de probabilité pour l'échantillon **B**, réalisées en utilisant les trois noyaux différents :

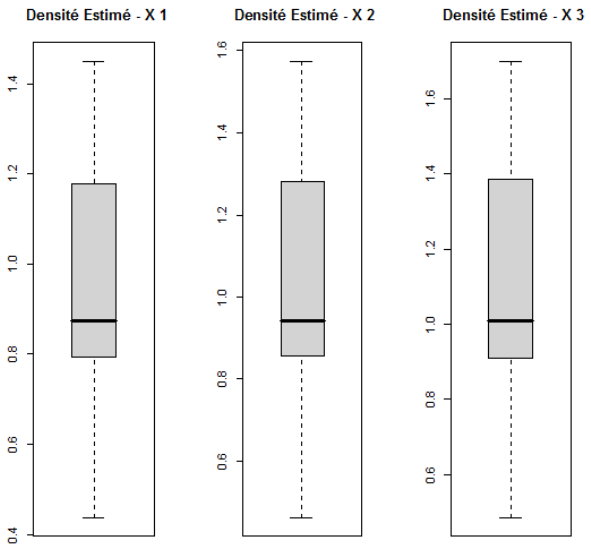

FIG. 4.13 – l'échantillon **B** avec le noyau de Dirichlet

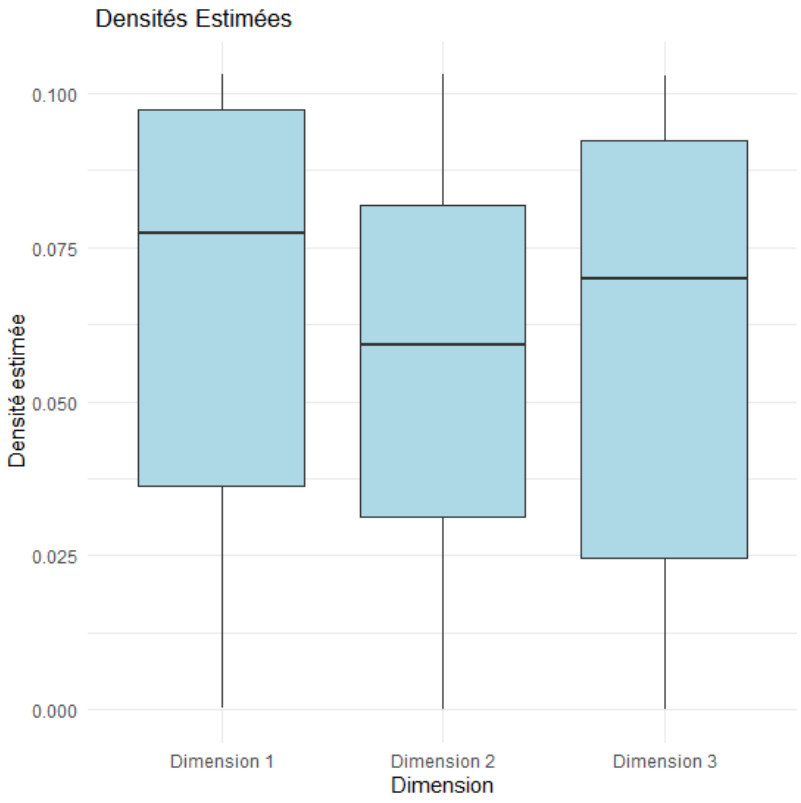

FIG. 4.14 – l'échantillon **B** avec le noyau logistique-normal

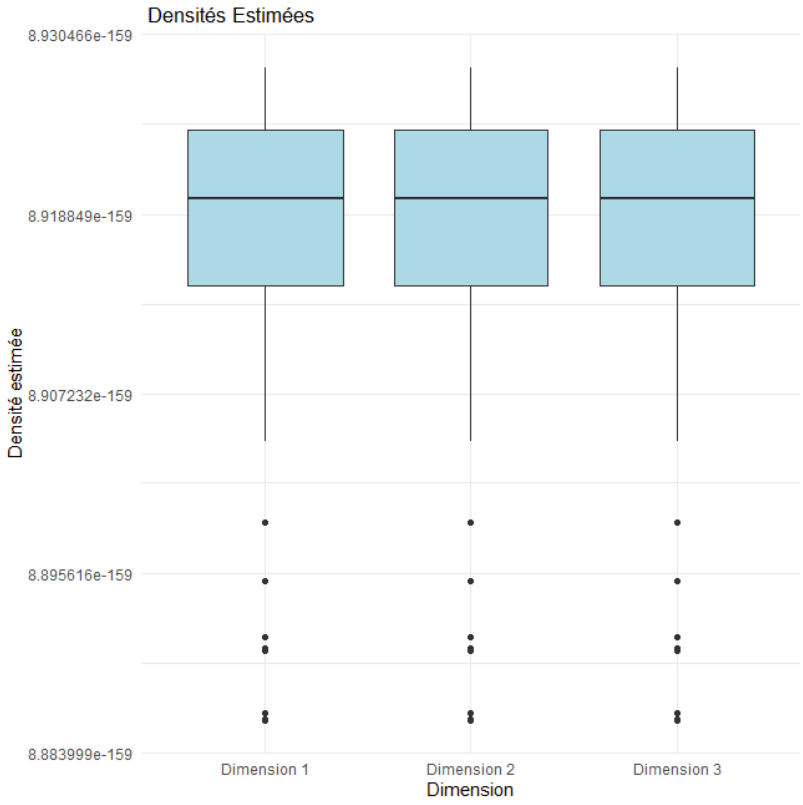

FIG. 4.15 – l'échantillon **B** avec le noyau normal

enfin les boxplote de la densité de probabilité estimé pour l'échantillon **C** avec les trois noyau

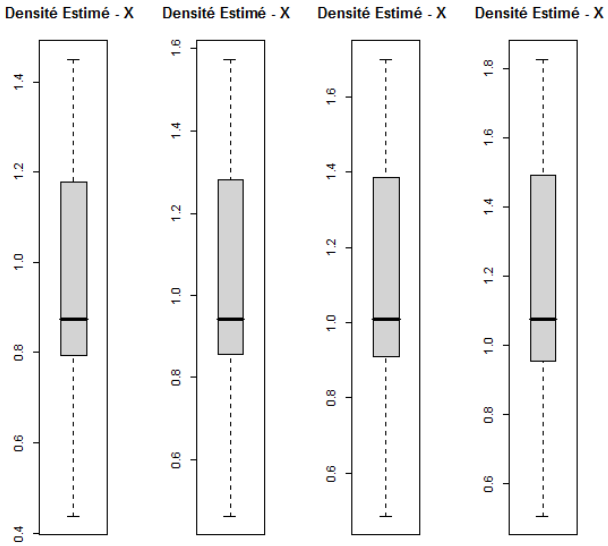

FIG. 4.16 – l'échantillon **C** avec le noyau de Dirichlet

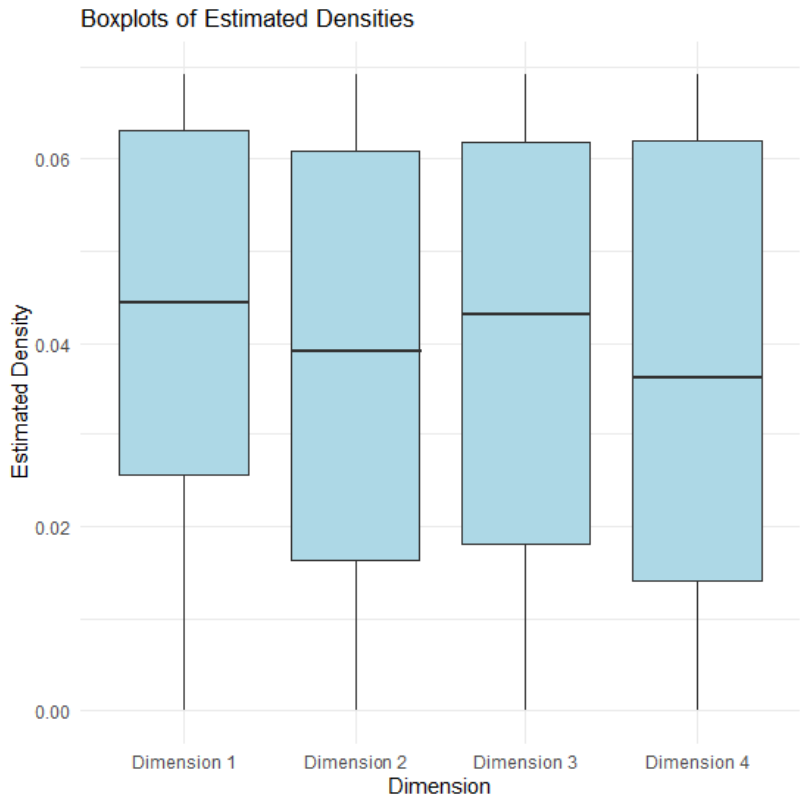

FIG. 4.17 – l'échantillon **C** avec le noyau logistique-normal

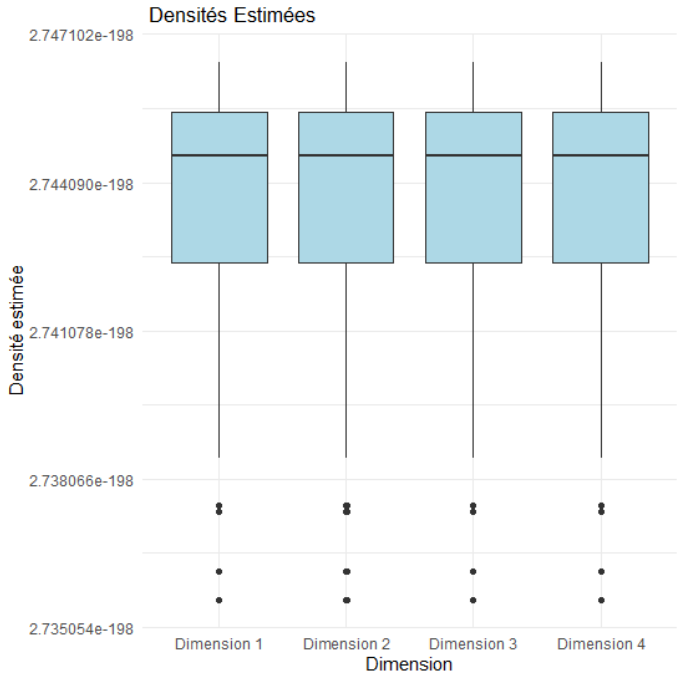

FIG. 4.18 – l'échantillon **C** avec le noyau normal

Ces résultats mettent en évidence des tendances intéressantes dans les performances des différentes méthodes de simulation ainsi que des noyaux utilisés. Les valeurs du RMSE permettent une évaluation de la précision des estimations de densité de probabilité, tandis que les paramètres de lissage (H) optimaux pour chaque combinaison révèlent le niveau d'ajustement nécessaire pour obtenir les meilleurs résultats.

4.5 Interprétation des Résultats

Cette section offre une synthèse des conclusions découlant de l'analyse des performances des différentes combinaisons échantillon-noyau dans le cadre de l'estimation de densité de probabilité multivariée. Trois échantillons distincts (A, B et C) de dimensions variables ont été soumis à l'examen, en utilisant trois types de noyaux distincts : le noyau de Dirichlet, le noyau gaussien multivarié et le noyau logistique normal. Ces résultats sont condensés dans les deux tableaux (tableau 4.1) et (tableau 4.2).

Le RMSE (Root Mean Squared Error) est une mesure d'erreur évaluant la divergence entre les valeurs prédites et les valeurs réelles. À mesure que le RMSE se rapproche de zéro, les prédictions deviennent plus précises. Les valeurs de RMSE obtenues pour chaque combinaison échantillon-noyau indiquent la différence entre les estimations de densité de probabilité et les données réelles.

Le paramètre de lissage H joue un rôle décisif dans l'ajustement de la méthode d'estimation de densité de probabilité. Dans le cas des noyaux de Dirichlet et de logistique normale, H se manifeste sous la forme d'un scalaire unique pour chaque combinaison échantillon-noyau. Toutefois, dans le contexte du noyau gaussien multivarié, H se présente sous la forme d'une matrice d'ordre d-1. Les valeurs optimales de H révèlent le degré de souplesse nécessaire pour atteindre les performances d'estimation optimales.

Les résultats révèlent des tendances intrigantes, mettant en évidence l'importance capitale du choix du noyau et du paramètre de lissage dans l'estimation de la densité de probabilité multivariée. Par exemple, dans l'échantillon A, le noyau gaussien multivarié se démarque en affichant un RMSE exceptionnellement faible, démontrant ainsi son adaptabilité remarquable à ces données. De plus, nos résultats révèlent que les valeurs optimales de H varient en fonction du noyau et de l'échantillon, soulignant la nécessité d'adapter l'ajustement selon le contexte. Notant que nos résultats confirment les résultats déja obtenus par d'autres chercheurs (voir **([Chacón and Duong, 2010])** conceranat la performance des différente configuration du paramètre du lissage. A savoir, définir le paramètre de lissage H comme une matrice pleinne donne de meilleurs résultats par rapport au cas où le paramètre de lissage est considéré comme un scalair, c'est à dire prendre $H = diag(h), h \in \mathbb{R}$.

De manière intéressante, une observation significative émerge de notre analyse des résultats présentés dans les tableaux 4.1 et 4.2, concernant le noyau logistique normal. Contrairement aux attentes basées sur les résultats antérieurs de [**Aitchison and Lauder, 1985**], notre étude indique que le noyau logistique normal surpasse le noyau de Dirichlet dans l'estimation de la densité de probabilité. Cette conclusion est étayée par des valeurs de RMSE plus faibles pour le noyau logistique normal dans tous les échantillons, suggérant une meilleure adéquation aux données observées. Ces découvertes mettent en lumière l'importance de considérer différentes approches de modélisation lors de l'estimation de la densité de probabilité, contribuant ainsi à éclairer le choix des méthodes les plus prometteuses pour une estimation précise dans divers contextes d'application.

En général, les boîtes (boxplots) indiquent que les distributions de probabilité estimées

pour chaque échantillon et méthode de noyau présentent des caractéristiques similaires. Les médianes décalées suggèrent une légère asymétrie dans les données. La variabilité des données varie d'une méthode à l'autre, mais elle semble modérée dans l'ensemble, sans valeurs aberrantes notables, à l'exception du cas du noyau normal pour l'échantillon C.

Ces observations laissent entendre que les différentes méthodes de noyau ont des performances similaires en termes de centralité et de variabilité pour chaque échantillon, bien que certaines puissent montrer une asymétrie légèrement différente. La présence de valeurs aberrantes dans le noyau normal de l'échantillon C pourrait indiquer une nécessité de traitement supplémentaire pour gérer ces observations extrêmes.

4.6 Conclusion

Ce chapitre a fourni un aperçu approfondi des performances des noyaux de Dirichlet, gaussien multivarié et logistique normal pour l'estimation de la densité de probabilité multivariée. Les résultats obtenus ont souligné l'importance du choix du noyau et du paramètre de lissage, guidant ainsi la sélection des méthodes les plus performantes pour l'estimation précise dans différentes situations d'application.

Conclusion Générale et Perspectives

Ce mémoire offre une exploration approfondie de l'estimation de la densité de probabilité dans le cas particulier des données de composition. Ces données costituent l'une des configurations de données dépendantes les plus répandues en sciences appliquées et expérimentales. Ceci est fait en combinant des concepts théoriques avec des applications pratiques. Les méthodes et approches discutées dans ces chapitres fournissent des outils précieux pour les chercheurs et les analystes travaillant avec des données multivariées complexes tout en tenant compte des particularités inhérentes aux compositions sur un simplex S^d. Ce travail vise à contribuer à la compréhension et à l'application de ces techniques dans divers domaines de la statistique en général et de l'analyse de données en particulier. Ce qui va ouvrir des perspectives intéréssantes pour la recherche appliquée et théorique notamment en explorant d'autres noyaux comme le noyau béta et gamma ainsi que d'autres critères d'erreur outre que le RMSE. D'autres approches pour le choix du paramètre de lissage pourrons aussi être explorées et comparées. Enfin, nous pouvons considérer d'autres jeux de données plus complexes issus de domaines variés et différents pour compléter et valider un peu plus nos résultats tant sur le plan théorique qu'expérimental.

Résumé

Ce mémoire offre une exploration approfondie et structurée de l'estimation à noyau de la densité de probabilité dans le cas des données de composition. À travers quatre chapitres, il combine des concepts théoriques avec des applications pratiques pour éclairer la compréhension et l'application dans divers domaines de la recherche et de l'analyse de données, mettant en évidence l'importance du choix judicieux du type de noyau en fonction du contexte des données de composition. Le premier chapitre introduit les données de composition et leurs applications variées, tandis que le deuxième se penche sur l'estimation de densité de probabilité multivariée à l'aide de la méthode du noyau. Dans le troisième chapitre, l'accent est mis sur l'utilisation de trois noyaux spécifiques - Dirichlet, gaussiens multivariés et logistiques-normaux - pour estimer les densités des données de composition, en analysant leurs avantages et limites. Enfin, le quatrième chapitre aborde la simulation de données de composition et évalue les performances des trois noyaux précédemment étudiés, en utilisant une méthodologie de validation croisée pour optimiser les paramètres de lissage et mesurer la précision des estimations. Ainsi, ce mémoire offre une contribution significative à la compréhension et à l'application de l'estimation à noyau dans le contexte spécifique des données de composition.

Mots Clés

Estimation de densité de probabilité , Données de composition , Simplexe (Sd) ,Transformations (alr, clr et ilr) , Estimation non paramétrique à noyaux , Estimateurs à noyau de convolution , Matrice de lissage , Méthodes de validation croisée , Noyaux de Dirichlet , Noyaux gaussiens multivariés , Noyaux logistiques normaux , Analyse de données multidimensionnelles.

Abstract

This thesis provides a comprehensive and structured exploration of kernel density estimation in the case of compositional data. Across four chapters, it combines theoretical concepts with practical applications to illuminate understanding and application across various fields of research and data analysis, highlighting the critical importance of judiciously selecting the kernel type based on the context of compositional data. The first chapter introduces compositional data and their diverse applications, while the second delves into multivariate probability density estimation using the kernel method. In the third chapter, the focus is on the use of three specific kernels - Dirichlet, multivariate Gaussian, and logistic-normal - to estimate densities of compositional data, analyzing their advantages and limitations. Finally, the fourth chapter addresses the simulation of compositional data and evaluates the performance of the three previously studied kernels, using cross-validation methodology to optimize smoothing parameters and assess estimation accuracy. Thus, this thesis makes a significant contribution to the understanding and application of kernel estimation in the specific context of compositional data.

Keywords

Probability density estimation, Composition data, Simplex (S^d), Transformations (alr, clr, and ilr), Non-parametric kernel estimation, Convolution kernel estimators, Smoothing matrix, Cross-validation methods, Dirichlet kernels, Multivariate Gaussian kernels, Normal logistic kernels, Multidimensional data analysis

Bibliographie

J Aitchison and IJ Lauder. Kernel density estimation for compositional data. *Journal of the Royal Statistical Society : Series C (Applied Statistics)*, 34(2) :129–137, 1985.

John Aitchison. The statistical analysis of compositional data. *Journal of the Royal Statistical Society : Series B (Methodological)*, 44(2) :139–160, 1982.

Hirotugu Akaike. An approximation to the density function. *Annals of the Institute of Statistical Mathematics*, 6(2) :127–132, 1954.

Taoufik Bouezmarni and Jeroen VK Rombouts. Nonparametric density estimation for multivariate bounded data. *Journal of statistical planning and inference*, 140(1) :139–152, 2010.

Adrian W Bowman. An alternative method of cross-validation for the smoothing of density estimates. *Biometrika*, 71(2) :353–360, 1984.

Adrian W Bowman and Adelchi Azzalini. *Applied smoothing techniques for data analysis : the kernel approach with S-Plus illustrations*, volume 18. OUP Oxford, 1997.

José E Chacón and Tarn Duong. Multivariate plug-in bandwidth selection with unconstrained pilot bandwidth matrices. *Test*, 19 :375–398, 2010.

José E Chacón, G Mateu-Figueras, and Josep-Antoni Martín-Fernández. Gaussian kernels for density estimation with compositional data. *Computers & Geosciences*, 37(5) :702–711, 2011.

Song Xi Chen. Beta kernel estimators for density functions. *Computational Statistics & Data Analysis*, 31(2) :131–145, 1999.

DR Cox, DV Hinkley, D Rubin, and BW Silverman. *Monographs on statistics and applied probability*. Springer, 1984.

Claire Della Vedova et al. Initiation au logiciel de statistiques r : réalisez vos premières visualisations avec le package ggplot2. *Bulletin de la Dialyse à Domicile*, 2(4) :229–238, 2019.

Tarn Duong and Martin Hazelton. Plug-in bandwidth matrices for bivariate kernel density estimation. *Journal of Nonparametric Statistics*, 15(1) :17–30, 2003.

Tarn Duong and Martin L Hazelton. Convergence rates for unconstrained bandwidth matrix selectors in multivariate kernel density estimation. *Journal of Multivariate Analysis*, 93 (2) :417–433, 2005.

JDF Habbema, J Hermans, and K Van den Broek. A stepwise discriminant analysis program using density estimation. 1974.

Peter Hall, JS Marron, and Byeong U Park. Smoothed cross-validation. *Probability theory and related fields*, 92(1) :1–20, 1992.

Clément Hardy. Exploration des méthodes d'analyse de données compositionnelles pour l'étude du microbiome. 2018.

Harold V Henderson and SR Searle. Vec and vech operators for matrices, with some uses in jacobians and multivariate statistics. *Canadian Journal of Statistics*, 7(1) :65–81, 1979.

M Chris Jones, James S Marron, and Simon J Sheather. A brief survey of bandwidth selection for density estimation. *Journal of the American statistical association*, 91(433) :401–407, 1996.

MC Jones and RF Kappenman. On a class of kernel density estimate bandwidth selectors. *Scandinavian Journal of Statistics*, pages 337–349, 1992.

MC Jones, James Stephen Marron, and Byeong U Park. A simple root n bandwidth selector. *The Annals of Statistics*, 19(4) :1919–1932, 1991.

Wand MP Jones. Kernel smoothing, chapman & hall, 1995.

Célestin C Kokonendji and Sobom M Somé. On multivariate associated kernels to estimate general density functions. *Journal of the Korean Statistical Society*, 47(1) :112–126, 2018.

Célestin C Kokonendji and Sobom M Somé. Bayesian bandwidths in semiparametric modelling for nonnegative orthant data with diagnostics. *Stats*, 4(1) :162–183, 2021.

Max Kuhn et al. Caret package. *Journal of statistical software*, 28(5) :1–26, 2008.

Marco Maier. Dirichletreg : Dirichlet regression for compositional data in r. 2014.

Frédéric Ouimet. A multivariate normal approximation for the dirichlet density and some applications. *Stat*, 11(1) :e410, 2022.

Frédéric Ouimet and Raimon Tolosana-Delgado. Asymptotic properties of dirichlet kernel density estimators. *Journal of Multivariate Analysis*, 187 :104832, 2022.

Vera Pawlowsky-Glahn and Juan José Egozcue. Geometric approach to statistical analysis on the simplex. *Stochastic Environmental Research and Risk Assessment*, 15 :384–398, 2001.

Brian Ripley, Bill Venables, Douglas M Bates, Kurt Hornik, Albrecht Gebhardt, David Firth, and Maintainer Brian Ripley. Package 'mass'. *Cran r*, 538 :113–120, 2013.

Mats Rudemo. Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics*, pages 65–78, 1982.

Stephan R Sain, Keith A Baggerly, and David W Scott. Cross-validation of multivariate densities. *Journal of the American Statistical Association*, 89(427) :807–817, 1994.

SAID S.BEDDEK. *Sur l'estimation non paramétrique de la densité de probabilité dans le cas multidimensionnelle*. PhD thesis, Université de Béjaia-Abderrahmane Mira, 2011.

David W Scott. Feasibility of multivariate density estimates. *Biometrika*, 78(1) :197–205, 1991.

David W Scott. *Multivariate density estimation : theory, practice, and visualization*. John Wiley & Sons, 2015.

BW Silverman. Density estimation for statistics and data analysis chapman & hall/crc, london (1986). *Search in*.

Kunio Tanabe and Masahiko Sagae. An exact cholesky decomposition and the generalized inverse of the variance–covariance matrix of the multinomial distribution, with applications. *Journal of the Royal Statistical Society : Series B (Methodological)*, 54(1) :211–219, 1992.

K Gerald van den Boogaart, Raimon Tolosana, Matevz Bren, and Maintainer K Gerald van den Boogaart. Package 'compositions'. *Compositional data analysis. Ver*, pages 1–40, 2013.

Matt P Wand, M Chris Jones, et al. Multivariate plug-in bandwidth selection. *Computational Statistics*, 9(2) :97–116, 1994.

MP Wand. Error analysis for general multtvariate kernel estimators. *Journal of Nonparametric Statistics*, 2(1) :1–15, 1992.

Hans-Joachim Werner, Peter J Knowles, Gerald Knizia, Frederick R Manby, and Martin Schütz. Molpro : a general-purpose quantum chemistry program package. *Wiley Interdisciplinary Reviews : Computational Molecular Science*, 2(2) :242–253, 2012.