# Estimation non paramétrique de densités conditionnelles : grande dimension, parcimonie et algorithmes gloutons.

Minh-Lien Jeanne Nguyen

NNT : 2019SACLS185

# THÈSE DE DOCTORAT

de

L'UNIVERSITÉ PARIS-SACLAY

École doctorale de mathématiques Hadamard (EDMH, ED 574)

*Établissement d'inscription :* Université Paris-Sud
*Laboratoire d'accueil :* Laboratoire de mathématiques d'Orsay, UMR 8628 CNRS

*Spécialité de doctorat :* Mathématiques appliquées

**Minh-Lien Jeanne NGUYEN**

# Estimation non paramétrique de densités conditionnelles : grande dimension, parcimonie et algorithmes gloutons.

*Date de soutenance :* 8 Juillet 2019

*Après avis des rapporteurs :*

| | |
|---|---|
| BÉATRICE LAURENT-BONNEAU | (INSA de Toulouse) |
| MARKUS REISS | (Humboldt-Universität zu Berlin) |

*Jury de soutenance :*

| | | |
|---|---|---|
| ARNAK DALALYAN | (ENSAE / CREST, GENES) | Examinateur |
| CLAIRE LACOUR | (Université Paris-Est Marne-la-Vallée) | Codirectrice de thèse |
| BÉATRICE LAURENT-BONNEAU | (INSA de Toulouse) | Rapporteur |
| OLIVIER LOPEZ | (Sorbonne Université) | Examinateur |
| PASCAL MASSART | (Université Paris-Sud) | Président du jury |
| MARKUS REISS | (Humboldt-Universität zu Berlin) | Rapporteur |
| VINCENT RIVOIRARD | (Université Paris-Dauphine) | Codirecteur de thèse |

# *Remerciements*

*Mes premiers remerciements s'adressent naturellement à mes deux directeurs de thèse et à leurs nombreuses qualités : leur profonde compréhension scientifique en m'offrant ce sujet de recherche qui m'a passionnée, leurs explications variées sur le monde académique qui ont accompagné mes premiers pas dans la recherche, leur exigence initiale de m'envoyer en conférences desquelles je tire mes plus beaux moments de thèse, leurs conseils avisés qui ont ponctué ces quatre années et dont je n'ai parfois compris l'importance que des mois après, leur générosité, leur rigueur et leur honnêteté dans leurs relectures innombrables, leur disponibilité, leur soutien et leur gentillesse dans mes moments de doute. Je n'imagine pas meilleurs directeurs de thèse.*

*En deuxième, je tiens à remercier les membres de mon jury qui ont gracieusement accepté de venir m'écouter pour me donner leur avis, avec une mention particulière pour mes deux rapporteurs qui ont vaillamment relu mon manuscrit en avant-première.*

*J'ai aussi une pensée pour chacun de mes professeurs de mathématiques, de la sixième au master, qui m'ont fait arpenter cette voie.*

*Enfin, je souhaite remercier tous ceux qui ont nourri ces quatre dernières années de moments inoubliables sans lesquels la vie serait bien vide :*

- *les foisonnantes discussions, questions et suggestions scientifiques qui ont enrichi mon travail tout en m'aidant à prendre du recul,*
- *les précieuses et généreuses aides et explications durant mes galères administratives et informatiques,*
- *les exposés extraordinaires des talentueux orateurs qui continuent de m'émerveiller sur la beauté des mathématiques et de la statistique,*
- *les festives rencontres "jeunes" ou moins jeunes qui contribuent au caractère chaleureux du monde académique,*
- *les joies de l'enseignement venant autant de mes collègues que de mes élèves (malheureusement accompagnées de leurs lots de copies),*
- *les repas quotidiens, moments de débats sur le monde actuel, d'idéologies grandiloquentes pour le monde de demain, ou d'anecdotes farfelues et désopilantes sur les petits riens de la vie,*
- *les bières partagées entre rires, taquineries et déboires doctoraux,*
- *les bouffées d'air et d'expiration de miasmes en chœur et en joie,*
- *les sacro-saintes pauses café-potins avant de repartir bosser,*
- *les bavardages et originales manières d'être entre co-bureaux,*

- *les longues résolutions d'énigmes extravagantes en salle de thé du* 430*,*
- *les échappées philosophiques, littéraires ou artistiques d'êtres exquisément éclairés.*

*Merci.*

*à mes proches et leur indéfectible soutien,*

# Table des matières

# Chapitre I

# Introduction

L'objet de cette thèse est l'estimation de densités conditionnelles. Plus spécifiquement, l'angle de recherche est de s'attaquer au « fléau de la grande dimension », terme qui englobe les multiples obstacles posés par l'estimation en grandes dimensions.

Cette introduction suit le plan suivant. La première partie expose quels obstacles sont à surmonter dans l'estimation de fonctions en grandes dimensions. En deuxième partie, je focalise mon travail sur l'estimation de densités conditionnelles, en en détaillant les motivations ainsi que l'état de l'art. En dernière partie, je présente mon travail de recherche en précisant mes contributions dans ce contexte.

## 1 Estimation en grandes dimensions.

Les progrès numériques en termes de stockage ont ouvert la voie vers l'estimation en grandes dimensions en mettant à la disposition des statisticiens des jeux de données de grandes tailles à la fois en quantité d'observations et en dimension, c'est-à-dire nombre de caractéristiques observées.

Le potentiel que représentent ces grands jeux de données pour améliorer nos estimations est cependant contrecarré par les multiples obstacles que pose l'estimation en grandes dimensions.

Une première explication intuitive est qu'un objet de grande dimension est difficile à appréhender. En particulier, l'humain, vivant dans un monde en 3D et projetant essentiellement toute image sur des surfaces en 2D, visualise difficilement les altérations produites par plus de trois caractéristiques variant en même temps.
De même, le paysage des statistiques est substantiellement modifié par le passage en grandes dimensions, mettant en lumière de nombreuses contraintes qui n'existent pas en petites dimensions. Entre autres, la grande dimension a des répercutions théoriques sur les vitesses de convergence d'estimateurs mais aussi en pratique, quand par exemple on s'intéresse aux coûts algorithmiques des méthodes.

Ce problème dans toutes ses ramifications est communément appelé ''fléau de la dimension''.

### 1.a Fléau théorique de la dimension - L'approche minimax pour l'estimation fonctionnelle

On s'intéresse à l'estimation non paramétrique de fonctions réelles. Ce cadre se positionne naturellement dans la très grande dimension, en visualisant les fonctions de $\mathbb{R}^d \to \mathbb{R}$ comme des

éléments de $\mathbb{R}^{\mathbb{R}^d}$.

Pour illustrer le fléau dans ce cadre, je me focalise sur l'approche minimax qui révèle explicitement le problème.

### 1.a.i L'approche minimax.

La théorie minimax repose sur une idée simple : la fonction à estimer étant inconnue, on cherche un estimateur qui doit bien estimer toutes les fonctions, ou tout du moins toute une classe de fonctions. Considérer alors la plus grande erreur commise sur cette classe permet d'assurer une vitesse de convergence minimale, pourvu que la fonction à estimer soit dans la classe de fonctions considérée.
Plus formellement : pour une classe de fonctions $\mathcal{F}$, $T_n$ un estimateur construit à partir d'un échantillon de taille $n$, $\ell$ une fonction de perte sur l'espace des fonctions, l'approche minimax considère le risque uniforme sur la classe

$$\sup_{f\in\mathcal{F}} \mathbb{E}[\ell(T_n, f)],$$

et définit une notion d'optimalité en minimisant cette quantité.

**Définition I.1.** *On appelle « vitesse minimax » sur $\mathcal{F}$ pour la perte $\ell$ l'infimum suivant :*

$$\varphi_n(\mathcal{F}) := \inf_{T_n} \sup_{f\in\mathcal{F}} \mathbb{E}[\ell(T_n, f)],$$

*où l'infimum est pris sur tous les estimateurs construits à partir d'un échantillon de taille $n$.*
*On qualifie alors de « minimax optimal » un estimateur $\hat{T}_n$ dont la vitesse de convergence est inférieure à cette vitesse à une constante près, i.e. : il existe une constante $c > 0$ telle que :*

$$\sup_{f\in\mathcal{F}} \mathbb{E}[\ell(\hat{T}_n, f)] \leq c\varphi_n(\mathcal{F}).$$

Cette vitesse optimale dépend naturellement du nombre $n$ d'observations, mais aussi de la régularité des fonctions dans $\mathcal{F}$ ainsi que de la dimension $d$ de l'espace de départ. Cette vitesse minimale a été étudiée dans différents modèles (régression, densité, bruit blanc, ...), pour différentes fonction de perte (ponctuelle, uniforme, distance $\mathrm{L}^2$, $\mathrm{L}^1$, ...), sur différentes classes de régularité (Hölder, Sobolev, Besov, ...). Pour une régularité $s$-höldérienne, la vitesse minimax optimale est généralement de l'ordre de $n^{-\frac{s}{2s+d}}$, voire $(\log(n)/n)^{\frac{s}{2s+d}}$. Ce facteur logarithmique supplémentaire est récurrent dans certains cas de figure, comme celui de la perte uniforme, i.e. quand $\ell(.^{(1)}, .^{(2)}) = \| .^{(1)} - .^{(2)} \|_\infty$ (voir par exemple [Khas' minskii 1979]).

Nous nous intéresserons essentiellement à l'ordre de grandeur de ces vitesses. Fixons la notation pour l'équivalence à constante près : pour toutes suites $(a_n)$ et $(b_n)$, on note $a_n \asymp b_n$ s'il existe deux constantes positives $c$ et $C$ telles que pour $n$ assez grand,

$$ca_n \leq b_n \leq Ca_n.$$

Pour illustrer notre propos, on va plutôt se concentrer sur l'estimation ponctuelle de densités, dont les résultats sont plus aboutis que pour la densité conditionnelle. On considère aussi des fonctions de régularité höldérienne, régularité naturelle pour les méthodes à noyau présentées dans la suite. Usuellement définie pour $s \in ]0, 1]$, étendons la définition pour $s \in \mathbb{R}_+^*$.

**Définition I.2** ($s$-Hölder). *Soient $s > 0$, $L > 0$ et $d \in \mathbb{N}_{>0}$. Une fonction $f : \mathbb{R}^d \to \mathbb{R}$ appartient à la « boule höldérienne de régularité $s$ et de rayon $L$ » sur $\mathcal{U} \subset \mathbb{R}^d$, notée $\mathcal{H}_d(s, L)$, si, en décomposant $s$ sous la forme $s = k + \{s\}$ avec $k := \lceil s - 1 \rceil$ le plus grand entier strictement inférieur à $s$ et $\{s\} \in ]0, 1]$ la partie fractionnaire de $s$,*

*(i) $f$ est de classe $C^k$, c'est-à-dire est $k$-fois différentiable sur $\mathcal{U}$ et de dérivées partielles continues,*

*(ii) les dérivées partielles d'ordre $k$ de $f$ sont $(\{s\}, L)$-höldérienne au sens canonique, c'est-à-dire : pour tout multi-indice $\alpha \in \mathbb{N}^d$ tel que $\sum_{j=1}^{d} \alpha_j = k$, pour tous $(x, y) \in \mathcal{U}^2$ :*

$$\left| \frac{\partial^k}{\partial_1^{\alpha_1} \dots \partial_d^{\alpha_d}} f(y) - \frac{\partial^k}{\partial_1^{\alpha_1} \dots \partial_d^{\alpha_d}} f(x) \right| \leq L \|y - x\|_1^{\{s\}}.$$

**Remarque 1.** *Les fonctions $f : \mathbb{R}^d \to \mathbb{R}$ de classe $C^k$ de dérivées partielles d'ordre $k$ bornées par $L$ (cf Chapitre II) appartiennent à la boule $\mathcal{H}_d(s, 2L)$ pour tout $s \in ]k - 1, k]$*

**Remarque 2.** *La définition des boules höldériennes dans le Chapitre III est légèrement moins restrictive car elle ne contraint que les dérivées partielles uni-directionnelles.*

Dans ce cadre, le résultat suivant a été obtenu :

**Théorème I.1.** *Pour l'estimation ponctuelle de densités de $\mathbb{R}^d$ dans $\mathbb{R}_+$, pour toute régularité $s > 0$,*

$$\varphi_n(\mathcal{H}_d(s, L)) \asymp n^{-\frac{s}{2s+d}}.$$

En particulier, la borne inférieure est prouvée par Farrell [1972] dans le cadre multivarié, la borne supérieure ayant déjà été atteinte dans [Epanechnikov 1969] par méthode à noyau pour le risque en norme $\|\cdot\|_2$.

**Adaptation à la régularité.** La question de l'adaptation, où l'on sous-entend « adaptation à la régularité », se pose à partir des années 90. La problématique est la suivante. On disposait certes de méthodes atteignant la vitesse minimax, mais elles n'étaient valables que pour une régularité à la fois. En particulier, elles pouvaient faire intervenir dans leurs implémentation cette régularité $s$. Par exemple, la méthode d'Epanechnikov [1969] pour estimer une fonction de classe $C^2$, utilise dans son estimateur un noyau d'ordre $2$, en symbiose avec la régularité, pour converger à vitesse $n^{-\frac{2}{4+d}}$. En particulier, cette méthode ne converge pas à vitesse optimale si $f$ est de classe $C^3$.

Comme a priori la régularité de la fonction à estimer est inconnue, on ne va plus supposer que $f$ appartient à la classe $\mathcal{F}_s$ de régularité $s$ connue, mais à une classe beaucoup plus grande $\bigcup_{s \in \mathcal{S}} \mathcal{F}_s$ (avec $\mathcal{S}$ une famille de régularités) pour s'intéresser à des méthodes qui convergeront d'autant plus vite que la fonction à estimer est régulière, en s'adaptant ainsi à chaque classe $\mathcal{F}_s$ qui pourra contenir $f$. En particulier, on s'intéresse aux vitesses $\{\Psi_n(\mathcal{F}_s)\}_{s \in \mathcal{S}}$ minimales telle que, à estimateur $\hat{T}_n$ fixé, pour tout $s \in \mathcal{S}$,

$$\lim_{n \to \infty} \Psi_n(\mathcal{F}_s)^{-1} \sup_{f \in \mathcal{F}_s} \mathbb{E}[\ell(\hat{T}_n, f)] < +\infty.$$

**Définition I.3.** *Un estimateur $\hat{T}_n$ est dit « adaptatif » sur $\{\mathcal{F}_s, s \in \mathcal{S}\}$ s'il existe une famille de vitesses positives $\Psi_n := \{\Psi_n(\mathcal{F}_s)\}_{s \in \mathcal{S}}$ telle que pour toute régularité $s \in \mathcal{S}$, il existe une constante $C$ telle que : pour $n$ assez grand,*

$$\sup_{f \in \mathcal{F}_s} \Psi_n(\mathcal{F}_s)^{-1} \mathbb{E}[\ell(\hat{T}_n, f)] \leq C,$$

*et qu'il existe une constante $c > 0$ telle que :*

$$\inf_{T_n} \sup_{f \in \mathcal{F}_s, s \in \mathcal{S}} \Psi_n(\mathcal{F}_s)^{-1} \mathbb{E}[\ell(T_n, f)] \geq c.$$

*Les vitesses $\Psi_n(\mathcal{F}_s), s \in \mathcal{S}$, sont alors appelées « vitesses minimax adaptatives optimales » sur $\{\mathcal{F}_s, s \in \mathcal{S}\}$.*

Une question qui s'est alors posée était : a-t-on systématiquement $\Psi_n(\mathcal{F}_s) \asymp \varphi_n(\mathcal{F}_s)$, quelle que soit l'adaptation sur $\mathcal{S}$ ? En particulier, ces vitesses sont identiques pour l'estimation de densités en risque $L_2$ :

$$\Psi_n(\mathcal{H}_d(s,L)) \asymp \varphi_n(\mathcal{H}_d(s,L)) \asymp n^{-\frac{s}{2s+d}}$$

(voir [Akakpo 2012] pour la borne supérieure adaptative).

Mais cette égalité est fausse en toute généralité : en particulier, pour l'estimation ponctuelle de densités multivariées, on peut comparer le résultat récent de Rebelles [2015] au Théorème I.1 :

**Théorème I.2.** *Pour l'estimation ponctuelle adaptative de densités sur $\{\mathcal{H}_d(s,L), s > 0\}$,*

$$\Psi_n(\mathcal{H}_d(s,L)) \asymp \left(\frac{\log n}{n}\right)^{\frac{s}{2s+d}}.$$

**Le fléau de la dimension.** Le fléau de la dimension s'exprime dans le fait qu'il n'existe pas d'estimateur convergent plus vite que $n^{\frac{-s}{2s+d}}$, vitesse qui est d'autant plus lente que la dimension est grande. Pire encore, si l'on s'intéresse aussi à la dépendance en $d$ (comme par exemple dans les modèles à dimension croissante), un facteur multiplicatif $d^d$ s'ajoute, ralentissant encore la convergence (cf [McDonald 2017]).

### 1.a.ii Parcimonie et réduction de dimension

Pour contrer ce caractère inéluctable du fléau, on va plutôt s'intéresser au sous-problème dans lequel il existe un sous-espace de dimension $r$ plus petite que $d$ qui contiendrait la majorité de l'information : cette restriction est appelée « hypothèse de parcimonie ».

Comme souvent, face à l'adversité, la littérature a été très prolixe : de nombreux travaux se sont intéressés à ce sous-problème. On pense par exemple à l'analyse en composante principale (PCA) qui propose de se restreindre aux directions de plus grande variance ; ou bien les modèles qui imposent une dimension fixée comme le modèle à indice unique ("single-index model" en anglais) qui suppose une dépendance linéaire entre les covariables pour se réduire à estimer une fonction univariée de cette combinaison linéaire ; ou encore les méthodes pénalisées telles que LASSO qui font payer l'ajout de variable dans le modèle par une pénalisation de régularisation.
Avec ces méthodes variées de nombreuses questions émergent :

- Quelle définition formelle donner à cette hypothèse de parcimonie ? En particulier, comment la relier à notre fonction cible ?
- À quel moment fait-on la restriction de dimension : avant, après ou en même temps que l'estimation de notre fonction ?
- Quelles sont les vitesses minimax optimales si l'on se restreint à des fonctions parcimonieuses ? Si notre référence par la suite sera la vitesse $n^{\frac{-s}{2s+r}}$, où l'on a simplement remplacé la dimension $d$ par la dimension « pertinente » $r$, cette question reste en fait ouverte (à ma connaissance).

- Suppose-t-on que l'on connaît la dimension pertinente $r$ ? Sinon, est-ce que l'adaptation à la structure de parcimonie se paie dans les vitesses minimax ? Autre question ouverte.

Dans la suite, pour l'estimation ponctuelle de $f$ au point $w$, l'hypothèse de parcimonie considérée est représentée par l'ensemble $\mathcal{R}$ des variables dont $f$ dépend sur un voisinage $\mathcal{U}$ de $w$.

**Définition I.4.** *On définit le sous-ensemble $\mathcal{R}$ de $\{1,\dots,d\}$ (dépendant de $\mathcal{U}$ et de $f$) tel que sur $\mathcal{U}$, $f$ est constante dans toutes les directions non indexées dans $\mathcal{R}$. On appelle « pertinentes » les composantes indexées dans $\mathcal{R}$, et on note $r$ le cardinal de $\mathcal{R}$.*

## 1.b Fléau numérique de la dimension - Complexité algorithmique

Les statistiques, en tant que mathématiques appliquées, confrontent la théorie à la réalité. En particulier, une méthode effective et efficace sera privilégiée même sans garantie théorique. On pense par exemple aux forêts aléatoires, qui ont été proposées en 2001 par Leo Breiman : si leurs garanties théoriques sont jusqu'à maintenant plutôt limitées en comparaisons avec d'autres méthodes, elles sont très appréciées par les praticiens car leurs résultats empiriques sont souvent les plus compétitifs sur le marché.

Pour donner un ordre de grandeur des jeux de données massives disponibles, on peut considérer par exemple le nombre de tweets par jour, environs 500 millions, sur un réseau de plus de 300 millions de personnes interagissant, à multiplier par le nombre de jours durant lequel on prolonge l'expérience ; ou encore l'ADN humain qui contient environs $3,4$ milliards de paires de bases à multiplier par le nombre d'individus observés. Juste lire ce type de données est coûteux en temps et lancer un algorithme pour les analyser implique de les parcourir plusieurs fois. Le problème se complique encore quand notre société hyper-connectée demande d'être informée minute par minute, c'est-à-dire de ré-actualiser les réponses en prenant en compte les données collectées dans la minute précédente.
Créer des algorithmes efficaces en temps est donc crucial pour gérer ce type de données.

### 1.b.i Complexité algorithmique

Un critère pour déterminer l'efficacité d'un programme est la complexité algorithmique. Ici, on ne se soucie que de la complexité en temps, à différentier de la complexité en espace (sous-entendu, de stockage). Dans un soucis de référentiel objectif et en particulier pour que ne soient pas pris en compte les différences de performances des machines individuelles (qualité du processeur, langage de programmation utilisé, ...), cette complexité en temps se calcule le plus généralement en nombre d'opérations de base (par exemple, une addition, une multiplication, un test d'inégalité) en fonction de de la taille $n$ des données mises en entrée de l'algorithme et dans le pire des cas, c'est-à-dire pour l'entrée de taille $n$ qui prend le plus de temps.
Seul l'ordre de grandeur de la complexité nous intéresse, c'est-à-dire quand $n$ est grand, car pour les programmes courts, la réponse est en fait "instantanée" selon un ressenti humain.

Évidemment, plus cet ordre de grandeur est petit, plus l'algorithme est efficace. Spécifions quelques niveaux de coût :

- ◇ *complexité linéaire* : $\mathcal{O}(n)$. Typiquement, lire simplement les données. C'est donc le coût minimum si l'on veut utiliser toutes les données.

◇ *complexité quasi-linéaire* : $\mathcal{O}(\log(n)^b n)$, avec $b > 0$. Comme son nom le suggère, on ne perd pas beaucoup de temps par rapport à la complexité linaire, le facteur logarithmique étant négligeable par rapport à $n$, surtout quand $b$ n'est pas trop grand. Quand $b = 1$ en particulier, on parle de *complexité linéarithmique*.

◇ *complexité quadratique* : $\mathcal{O}(n^2)$. Typiquement, lire une matrice $n \times n$.

◇ *complexité cubique* : $\mathcal{O}(n^3)$. Typiquement, inverser une matrice $n \times n$.

#### 1.b.ii Fléau numérique de la dimension

Pour revenir au fléau de la dimension, on va distinguer la dimension du nombre d'observations pour discerner l'influence spécifique de la dimension sur les temps d'exécution : ainsi, on notera plutôt $n$ le nombre d'observations, $d$ leur dimension pour une taille totale des données $d \times n$.

Le fléau numérique de la dimension se produit quand, typiquement, la complexité croît exponentiellement vite avec la dimension. Cela peut engendrer des situations où l'on ne peut plus se permettre d'attendre qu'une procédure se termine, car cela prendrait des mois, des années, voire plus que l'âge de l'univers, d'autant plus que les différents pans du fléau de la dimension intéragissent en empirant le problème. En effet, comme expliqué en Section 1.a, plus la dimension est grande, plus la vitesse de convergence minimax des estimateurs est lente. Ainsi, pour garantir un même niveau d'erreur, une grande dimension demandera plus d'observations, rallongeant d'autant plus le temps d'exécution.

On va donc s'intéresser à des méthodes plus rapides dont l'enjeu sera de ne pas trop détériorer la qualité de l'estimation.

#### 1.b.iii Algorithmes gloutons et procédure itératives

Les méthodes qualitativement performantes en petites dimensions nécessitent d'optimiser un certain critère pour, par exemple, réaliser un bon compromis biais-variance. On pense par exemple à la validation croisée (voir [Arlot and Celisse 2010] et les références à l'intérieur) ou aux méthodes de Lepski [Lepskii 1991 ; Goldenshluger and Lepski 2011a]. En grandes dimensions, cette optimisation peut être très coûteuse en temps, d'autant plus si la grille à parcourir dépend de la dimension des données.

Plusieurs procédés ont été considérés pour réduire le temps d'exécution des méthodes. En voici quelques uns :

- les algorithmes gloutons. Il s'agit d'algorithmes itératifs dont la stratégie est de se déplacer astucieusement dans la grille d'optimisation pour ne pas avoir à l'explorer toute entière. Un exemple classique est l'algorithme de Newton-Raphson pour chercher le minimum d'une fonction. La plupart des ces algorithmes dépendent d'une hypothèse de structure ; dans notre exemple, la convexité de la fonction.
- la parallélisation, qui consiste à faire tourner simultanément sur plusieurs clusters de calculs différentes parties du programme quand elles ne dépendent pas mutuellement de leurs résultats respectifs.
- les algorithmes de flux d'information. Dans un contexte où les données arrivent de manière continue, leur propriété importante pour prendre en compte l'ajout de données est de ne plus utiliser les premières données directement mais seulement l'estimation qui a été faite

sur ces premières données, de sorte à ne pas relancer la procédure sur des jeux de données toujours plus grands.

# 2 Estimation de densités conditionnelles

Mon travail de thèse s'est concentré sur l'estimation de densités conditionnelles.

Après avoir posé le modèle, j'explique pourquoi ce problème est intéressant. Enfin, je décris l'état de l'art pour l'estimation non paramétrique de densités conditionnelles en dimensions modérément grandes.

## 2.a Modèle

Posons le modèle considéré. On a collecté $n$ observations $W_i$, $i = 1 : n$, indépendantes et identiquement distribuées (abrégé *i.i.d.* dans la suite) d'un vecteur aléatoire $d$-dimensionnel. Parmi les $d$ variables mesurées, seulement certaines sont vraiment d'intérêt, les autres n'étant que des observations auxiliaires. Quitte à réordonner les composantes, on peut écrire chaque observation $W_i$, $i = 1 : n$, sous la forme d'un couple $(X_i, Y_i)$ de vecteurs aléatoires, $Y_i$ contenant les $d_2$ variables d'intérêt et $X_i$ les $d_1$ variables auxiliaires (de telle sorte que $d_1 + d_2 = d$).
On s'intéresse alors à la loi conditionnelle de $Y_1$ sachant $X_1$ (l'échantillon étant *i.i.d.*). On suppose que les variables aléatoires sont absolument continues par rapport à la mesure de Lebesgue, et on note $f_W$ la densité jointe de $W_1 = (X_1, Y_1)$, $\mathrm{f}_X$ la densité marginale de $X_1$ et $f$ la densité conditionnelle de $Y_1$ sachant $X_1$ que l'on définit ponctuellement : au point $w = (x, y) \in \mathbb{R}^{d_1+d_2}$, si $\mathrm{f}_X(x) > 0$[1], on définit la densité conditionnelle de $Y$ sachant $X = x$ par le ratio

$$f(x,y) := \frac{f_W(x,y)}{f_X(x)}.$$

## 2.b Motivations

L'estimation de densités conditionnelles répond à deux problématiques fondamentales en statistiques : retrouver la loi sous-jacente à un jeu de données et décrire les relations entre les différentes variables. De ce point de vue, l'estimation de densités conditionnelles est un problème plus riche que deux problèmes qui ont été bien plus intensivement étudiés :

- ⋆ l'estimation de densités, qui est englobée naturellement par l'estimation de densités conditionnelles en ne considérant aucune variable comme auxiliaire, et
- ⋆ le problème de régression. La densité conditionnelle contient de fait plus d'information que la fonction de régression, qui est simplement l'espérance conditionnelle, puisqu'à partir de la densité conditionnelle, on peut obtenir la fonction de régression, mais que l'inverse est faux.

En comparaison aux deux problèmes cités ci-dessus, la littérature est nettement plus maigre pour traiter le problème d'estimation de densités conditionnelles, alors qu'il y a une forte demande dans de nombreux domaines d'application tels que l'économie [Hall et al. 2004], l'astronomie [Izbicki

1. Les densités n'étant définie qu'à ensemble négligeable près, pour donner un sens à cette contrainte, on supposera nos densités continues au voisinage de $w$, et de fait, tous nos résultats dépendront d'hypothèses de régularité. Noter qu'en toute généralité, il n'y aucun espoir d'estimer correctement une fonction évaluée à un point de discontinuité.
Enfin, noter que dans le cas où $\mathrm{f}_X(x) = 0$, conditionner par $X = x$, évènement non réalisable, n'a pas de sens.

and Lee 2016], la médicine [Takeuchi et al. 2009], l'actuariat [Efromovich 2010b], la météorologie [Jeon and Taylor 2012].

Une des raison est que l'estimation de densités conditionnelles se heurte plus rapidement au fléau de la dimension, d'où la difficulté d'obtenir des méthodes performantes. Les travaux prenant en compte les difficultés qui en découlent sont plutôt rares, surtout quand on cherche des modèles où à la fois $X$ et $Y$ sont multivariés.

Dans la section suivante, on détaille l'état de l'art en estimation non paramétrique de densités conditionnelles multivariées.

## 2.c État de l'art

Différents types de méthodes non paramétriques ont été examinées pour estimer des densités conditionnelles : les estimateurs à noyau [Rosenblatt 1969 ; Hyndman et al. 1996 ; Bertin et al. 2016] et les différentes discussions pour sélectionner la fenêtre [Bashtannyk and Hyndman 2001 ; Fan and Yim 2004 ; Hall et al. 2004], les estimateurs par polynômes locaux [Fan et al. 1996 ; Hyndman and Yao 2002], les estimateurs par projection [Efromovich 1999; 2007], les estimateurs constants par morceaux [Györfi and Kohler 2007 ; Sart 2017], les estimateurs par copule [Faugeras 2009] entre autres. Cependant la plupart des travaux pré-cités ne prennent pas en compte le fléau de la dimension. Ils se placent pour la plupart dans un cadre où soit $X$, soit $Y$ est univarié, voire les deux, et leurs coûts d'exécution peuvent s'avérer très onéreux : en particulier, aucune de ces méthodes ne propose d'exemple ou d'application de dimension $d > 3$.

Si l'on s'intéresse plus spécifiquement aux méthodes à noyau, elles obtiennent de bons résultats théoriques, pour peu que la fenêtre soit attentivement sélectionnée : en particulier Hall et al. [2004] et Bertin et al. [2016] convergent à vitesse minimax optimale (et adaptative pour le second article) en sélectionnant la fenêtre respectivement par validation croisée et par la méthodologie de Goldenshluger and Lepski [2011a]. De plus, ces deux méthodes parviennent à contourner le fléau de la dimension théorique en détectant d'éventuelles composantes pertinentes, améliorant ainsi la vitesse de convergence dans les cas parcimonieux. Cependant ces deux sélections sont particulièrement intensives algorithmiquement puisqu'elles demande de calculer l'estimateur à noyau sur une grille exhaustive de fenêtres multivariées, pour en sélectionner le meilleur (pour un critère bien choisi). Comme le cardinal de la grille croît exponentiellement avec la dimension (par exemple de taille $(\log n)^d$ quand on prend $\log n$ valeurs par direction), ces deux méthodes sont fortement soumises au fléau de la dimension numérique.

À ma connaissance, seulement deux méthodes à noyau attaquent frontalement ce fléau numérique. Holmes et al. [2010] proposent une version approchée de la validation croisée pour diminuer significativement le nombre d'appels au noyau, grâce la construction d'un arbre dual. Mais les résultats théoriques en pâtissent lourdement puisque seule la consistance de la procédure est obtenue, sans vitesse de convergence.

Une autre approche est d'effectuer en amont une étape de réduction de dimension pour ensuite estimer une densité conditionnelle de plus petite dimension. C'est ce que proposent Fan et al. [2009] dans un modèle où $Y$ est scalaire : ils approximent la densité conditionnelle par une version approchée où la dépendance de $Y$ en $X$ est restreinte à une combinaison linéaire des composantes de $X$. Une fois la meilleure combinaison linéaire trouvée, il ne leur reste ainsi plus qu'une densité conditionnelle bivariée à estimer. Ils prouvent une convergence à vitesse $n^{-\frac{1}{3}}$, qui n'est cependant ni adaptative puisqu'ils supposent une régularité $C^3$, ni celle minimax optimale $n^{-\frac{3}{8}}$ pour les fonctions bivariées de régularité $3$. De plus, leur hypothèse de parcimonie où $Y$ ne dépend de $X$ qu'à travers une combinaison linéaire de $X$ est particulièrement forte. En dehors de ce

cadre, cette hypothèse peut induire une perte d'information significative. On peut noter qu'il ne s'agit plus d'une méthode pleinement non paramétrique mais plutôt d'une approximation semi-paramétrique puisque la combinaison linéaire est estimée paramétriquement et la densité conditionnelle bivariée non-paramétriquement.

Dans la même veine, on peut citer Otneim and Tjøstheim [2018] qui supposent une corrélation gaussienne sur les variables. Sous hypothèse de corrélation gaussienne qu'ils estiment paramétriquement, il ne leur reste plus qu'à estimer non-paramétriquement des fonctions univariés : les fonctions de répartition et les densités marginales de chaque covariable. Dans leurs exemples numériques, la méthode semble gérer des jeux de données de dimension $6$ et détecter implicitement d'éventuelles variables non pertinentes, mais ce n'est pas prouvé. Théoriquement ils prouvent la normalité asymptotique de leur méthode. Cependant, il est difficile de déterminer la vitesse de convergence réelle de leur méthode car elle dépend d'un paramètre qui est contraint par plusieurs de leurs hypothèses et dont la calibration n'est pas discutée dans leurs simulations.

Les méthodes par projections peuvent aussi être compétitives si l'on restreint $Y$ à être scalaire. Efromovich [2010a] obtient de bons résultats théoriques : une inégalité oracle impliquant la vitesse adaptative minimax optimale, et la détection d'une dimension intrinsèque plus petite en cas de parcimonie grâce à une décomposition en série spécifiquement choisie. Néanmoins, cette décomposition se paie sur le temps d'exécution : la complexité de l'algorithme n'est plus quasi-linéaire en $n$ (au sens où l'exposant en $n$ est strictement supérieur à $1$). La méthode ne peut donc s'appliquer que pour de petits échantillons.
Plus récemment, les méthodes par projection de Izbicki and Lee [2016; 2017] ouvrent la voie du côté numérique du fléau. Le premier papier propose une méthode particulièrement rapide, pouvant ainsi gérer des $X_i$ de très grande dimension (plus de $1000$ composantes) et la vitesse de convergence tient compte d'une éventuelle dimension intrinsèque plus petite. La seconde méthode traduit les coefficients de leur estimateur par projection comme des fonction de régression, permettant d'y déployer tous les résultats de régression connus (données mixtes ou sur des variétés, détection de variables pertinentes, procédures rapides). Cependant, ces deux méthodes pèchent sur deux critères importants : leur meilleures vitesses ne sont pas minimax optimales et reposent sur des paramètres particulièrement non adaptatifs, car ils dépendent à la fois de la régularité et du degré de parcimonie de la densité conditionnelle à estimer.

# 3 Outils préliminaires : estimateur à noyau adapté à la densité conditionnelle

## 3.a Rappels sur les estimateurs à noyau

### 3.a.i Vers l'estimateur à noyau

L'estimateur à noyau de densités a été inventé par Rosenblatt [1956] en partant de l'idée de "dériver" la fonction de répartition empirique. Soit $F_n : \mathbb{R} \to [0,1]$ la fonction empirique d'un échantillon univarié $\{V_i\}_{i=1}^n$ *i.i.d.*, définie par

$$F_n : v \mapsto \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{V_i \leq v}.$$

Cette fonction n'est pas dérivable aux points d'observations. Regardons sa variation sur un petit intervalle : au point $v \in \mathbb{R}$, pour un écart $h > 0$,

$$\begin{aligned} F_n(v+h) - F_n(v-h) &= \frac{1}{n}\sum_{i=1}^{n}(\mathbb{1}_{V_i \leq v+h} - \mathbb{1}_{V_i \leq v-h}) \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{v-h<V_i\leq v+h} \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{]-1,1]}(\tfrac{V_i-v}{h}) \end{aligned}$$

Pour $h$ petit, on considère alors l'estimateur de la densité $\frac{F_n(v+h)-Fn(v-h)}{2h}$ que l'on ré-écrit sous la forme

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K(\tfrac{V_i-v}{h}),$$

avec $K \equiv \frac{1}{2}\mathbb{1}_{]-1,1]}$. Cet estimateur est en fait un histogramme à fenêtre glissante, que l'on va étendre en estimateur à noyau en prenant d'autres fonctions $K$, en particulier des fonctions régulières de manière à proposer un estimateur lisse de la densité.
Dans le but de conserver une densité, la fonction $K$ est choisie d'intégrale $1$ (et éventuellement positive) : une telle fonction est appelé « noyau », d'où la dénomination « estimateur à noyau ».

On vient ainsi de définir un estimateur à noyau de fenêtre $h > 0$ et de noyau $K$ de la densité de $V_1$ :

$$\hat{f}_{V,h} : v \mapsto \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K(\tfrac{V_i-v}{h}).$$

Concrètement, cet estimateur applique $K$, notre fonction de masse $1$, sur chaque observation $V_i$, puis en prend la moyenne. Cette construction est satisfaisante intuitivement car pour un noyau qui contient majoritairement sa masse autour de $0$, plus la densité sera grande localement, plus ce voisinage contiendra d'observations et plus les noyaux autour de ces observations seront de grande valeur, et plus l'estimateur sera grand localement (cf Figure I.1 au voisinage de $-1$).

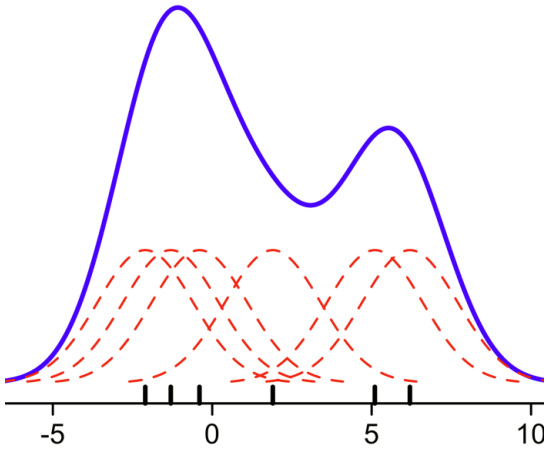



FIGURE I.1 – Construction d'un estimateur à noyau pour $n = 6$ observations. Sur chaque observation (représentée par un tiret noir sur l'axe des abscisses), on applique le noyau gaussien (courbes discontinues rouges) pondéré par $1/n$, puis on en prends la somme (courbe bleue).

### 3.a.ii Estimateurs à noyau multivariés

Soit $\{W_i\}_{i=1}^n$ un échantillon multivarié à valeurs dans $\mathbb{R}^d$. Oublions un instant la fenêtre. Pour construire un estimateur multivarié, on va simplement rendre $K$ multivarié. Soit $\mathbb{K} : \mathbb{R}^d \to \mathbb{R}$ un noyau, c'est-à-dire intégrable et vérifiant $\int_{\mathbb{R}^d} \mathbb{K} = 1$. On définit alors un estimateur de la densité de $W_1$ :

$$\hat{f}_W : w \mapsto \frac{1}{n} \sum_{i=1}^{n} \mathbb{K}(W_i - w).$$

Ajoutons notre paramètre sous forme matricielle. Soit $H$ une matrice $d \times d$ symétrique et définie positive, et on définit l'estimateur à noyau multivarié par

$$\hat{f}_{W,H} : w \mapsto \frac{1}{n} \sum_{i=1}^{n} \mathbb{K}(H^{-1}(W_i - w))/\det(H).$$

Noter l'analogie entre $H$ et la racine carrée de la matrice de covariance pour un noyau gaussien standard.

Dans notre cas, on se restreindra aux noyaux multivariés sous forme de produit de $d$ noyaux univariés :

$$\mathbb{K} : w \mapsto \prod_{j=1}^{d} K(w_j),$$

et aux fenêtres vectorielles $h \in \mathbb{R}_{>0}^d$ :

$$H = \mathrm{diag}(h),$$

ce qui nous amène à l'estimateur suivant :

$$\hat{f}_{W,h} : w \mapsto \frac{1}{n} \sum_{i=1}^{n} K_h(W_i - w),$$

avec la notation $K_h : w \mapsto \prod_{j=1}^{d} K(w_j/h_j)/h_j$.

Naturellement se posent les questions du choix de la fenêtre et du noyau, ainsi que leurs influences. Noter qu'un changement de fenêtre peut en fait se traduire comme un changement de noyau : par exemple, choisir une fenêtre plus petite revient à prendre un noyau de masse plus concentrée autour de $0$. Ainsi, on peut réellement considérer la fenêtre comme un paramètre d'échelle et créer des classes d'équivalence $\{K_h : h \in \mathbb{R}_{>0}^d\}$.

### 3.a.iii Choix du noyau

Noter que tous les travaux ne se sont pas restreints à cette paramétrisation par la fenêtre (comme par exemple [Panaretos and Konis 2012]), mais la sélection du noyau en est alors bien plus complexe. A contrario, la paramétrisation par la fenêtre permet de transférer toute la dépendance en $n$ du noyau dans le choix de la fenêtre. Ainsi, à changement de fenêtre près, le choix du noyau est nettement moins restrictif.

Le critère qui nous importera dans la suite est son ordre.

**Définition I.5.** *Soit $p > 0$ un entier strictement positif. On dit qu'un noyau $K$ est « d'ordre $p$ » si pour tout entier $l < p$ ,*

$$\int t^l K(t) dt = 0$$

*et si*

$$\int |t^p K(t)|dt < +\infty$$

*On dit que $K$ est « exactement d'ordre $p$ » si $K$ est d'ordre $p$ et vérifie :*

$$\int t^p K(t)dt \neq 0.$$

Certains travaux se sont intéressés plus précisément au noyau optimal (à changement de fenêtre près) : par exemple [Epanechnikov 1969] en cherchant de manière déterministe le meilleur noyau pour la perte $L^2$, ou [Goldenshluger and Lepski 2011b] en choisissant empiriquement le meilleur noyau parmi une famille de noyaux. Cependant, les améliorations se sont avérées peu significatives restant de l'ordre de la constante dans les vitesses de convergence [Rosenblatt 1971] : pour converger à vitesse minimax optimale, il suffit en fait que l'ordre du noyau soit plus grand que la régularité la fonction estimée. Cette condition complique toutefois l'adaptation.

### 3.a.iv Sélection de la fenêtre : du sur- et sous-apprentissage au compromis biais-variance en passant par les inégalités de concentration

La vitesse de convergence d'un estimateur à noyau découle étroitement du choix de sa fenêtre. C'est une question difficile qui alimente la littérature depuis l'introduction des estimateurs à noyau [Rosenblatt 1956]. Les enjeux de cette sélection surviennent dans le simple contexte

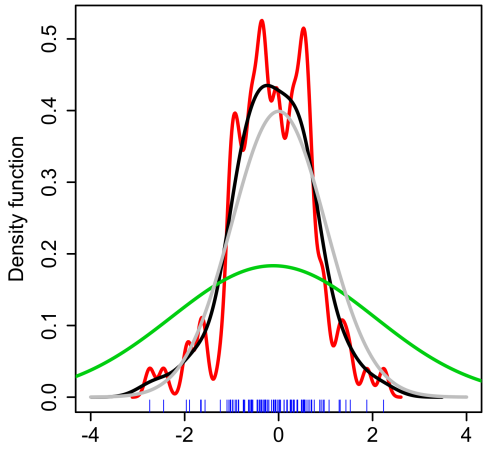



FIGURE I.2 – Trois estimateurs à noyau (en rouge, noir et vert) de fenêtres différentes (respectivement $0.05$, $0.337$ et $2$) pour estimer une densité gaussienne (en gris).

univarié, comme l'explicite la Figure I.2 : si la fenêtre $h$ est trop petit petite (cf courbe rouge), on considère, à tort, des amas aléatoires de données comme déterministes et inscrits dans la densité : on sur-apprend. Quand $h$ est trop grande (cf courbe verte), on met une masse trop étalée sur chaque observation, chargeant excessivement leurs voisinages éloignés et négligeant les modes de la densité : on sous-apprend. La sélection d'une "bonne" fenêtre (cf courbe noire) est donc sujette à un compromis entre sur- et sous-apprentissage, que l'on peut traduire en compromis biais-variance : l'estimateur rouge a une variance trop élevée, l'estimateur vert est trop biaisé.

Exprimons ces quantités en fonction de la fenêtre. Pour $w \in \mathbb{R}^d$, le risque quadratique ponctuel se décompose en biais et variance :

$$\begin{aligned}\mathbb{E}[(\hat{f}_{W,h}(w) - f_W(w))^2] &= \big(\mathbb{E}[\hat{f}_{W,h}(w)] - f_W(w)\big)^2 + \mathbb{E}[(\hat{f}_{W,h}(w) - \mathbb{E}[\hat{f}_{W,h}(w)])^2] \\ &=: \mathrm{Biais}(\hat{f}_{W,h}(w))^2 + \mathrm{Var}(\hat{f}_{W,h}(w)).\end{aligned}$$

On peut borner ces deux termes en fonction de $n$ et de la fenêtre $h$.

$$\mathrm{Var}(\hat{f}_{W,h}(w)) \leq \frac{C_1}{n\prod_{j=1}^{d} h_j}, \tag{I.1}$$

$C_1$ dépendant de $K$ et $\sup f_W$, et si $f_W$ est $(s,L)$-höldérienne et si l'ordre du noyau est supérieur à $s$, en utilisant un développement de Taylor :

$$|\mathrm{Biais}(\hat{f}_{W,h}(w))| \leq C_2 \sum_{j=1}^{d} h_j^s, \tag{I.2}$$

$C_2$ dépendant de $K$, $s$ et $L$. Le biais (ou plutôt sa majoration) est donc croissant avec les composantes de la fenêtre tandis que la variance est décroissante. Un compromis est donc nécessaire. La fenêtre minimisante est de l'ordre de $n^{-\frac{1}{2s+d}}$ et permet d'obtenir la vitesse minimax optimale. Cependant, cette fenêtre est fortement non adaptative en la régularité $s$. Noter tout de même que seul le biais dépend de la régularité, la borne de la variance ne dépendant des inconnues que dans la constante. La difficulté réside donc théoriquement dans le fait que le biais d'un estimateur soit difficile à évaluer (même si d'un point de vue plus pratique, les échantillons qui sont nécessairement de taille $n$ finie empêchent les considérations uniquement asymptotiques et posent des problèmes de calibration, ici de la constante devant $n^{-\frac{1}{2s+d}}$ dans la fenêtre à choisir).

Pour atteindre l'adaptation, la fenêtre ne doit plus être déterministe mais déduite des données. L'aléa induit par une fenêtre $\hat{h}$, fonction des observations, rend plus compliquée la majoration des termes de biais et de variance. Dans la suite, le contrôle du risque s'effectuera grâce à des inégalités de concentration qui exhiberont un comportement de grande probabilité sur lequel se baser et assureront que les autres cas restent négligeables.

**Inégalités de concentration.** La très connue « Loi des grands nombres » assure la convergence p.s. d'une moyenne empirique de variables *i.i.d.* vers leur espérance. Mais cela reste un résultats asymptotique. Confronté à un jeu de données fini, les inégalités de concentration permettent de déterminer un seuil relatif à la taille de l'échantillon donné de manière à ce qu'avec grande probabilité, la différence entre la moyenne empirique et l'espérance ne dépasse pas ce seuil. On considèrera en particulier l'inégalité de Bernstein :

**Lemma I.3** (Inégalité de Bernstein). *Soit $U_1,\dots,U_n$ des variables aléatoires indépendantes, absolument bornées par $c>0$, et de carré intégrable borné par $v>0$, i.e. : pour $i=1,\dots,n$,*

$$|U_i| \leq c \text{ p.s.}$$

$$\mathbb{E}[U_i^2] \leq v.$$

*Alors pour tout seuil $\lambda > 0$,*

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} U_i - \mathbb{E}[U_i]\right| \geq \lambda\right) \leq 2\exp\left(-\min\left(\frac{n\lambda^2}{4v}, \frac{3n\lambda}{4c}\right)\right).$$

Cette version de l'inégalité de Bernstein est une version légèrement plus faible que celle inscrite dans [Birgé and Massart 1998, p.366].

Les estimateurs à noyau étant des moyennes empiriques, on peut appliquer cette inégalité pour un seuil dépendant astucieusement de la fenêtre et de la taille d'échantillon $n$ pour contrôler les déviations de l'estimateur sur des évènements de grande probabilité.

## 3.b Estimation à noyau de densités conditionnelles

Pour estimer une densité conditionnelle, la littérature s'est majoritairement intéressée à un ratio d'estimateurs à noyau de la densité jointe $f_W$ et de la densité marginale $\mathrm{f}_X$ [Hall et al. 2004 ; Holmes et al. 2010] : typiquement, (en reprenant les notations en Section 2.a)

$$\frac{\hat{f}_{W,h}(w)}{\hat{\mathrm{f}}_{X,h'}(x)}.$$

Un défaut majeur de cet estimateur est qu'il n'estime pas directement la densité conditionnelle $f$, ce qui le rend non optimal sans hypothèse supplémentaire. En effet, même si les deux estimateurs au numérateur et au dénominateur sont minimax optimaux, leur vitesses de convergence dépendront respectivement de la régularité de la densité jointe et de celle de la densité marginale, qui peuvent être toutes deux inférieures à celle de la densité conditionnelle.
De même, une structure de parcimonie peut être propre à $f$. Typiquement, si la première composante des $X_i$ est indépendante des $Y_i$, $x_1$ est une direction non pertinente pour $f$, mais pertinente pour la densité jointe et la densité marginale.

Pour estimer plus directement la densité conditionnelle, l'idée de Bertin et al. [2016] est d'insérer à l'intérieur de la somme l'estimateur de la marginale en l'évaluant sur les observations $X_i$. Notons $\widetilde{\mathrm{f}}_X$ un estimateur de $\mathrm{f}_X$ (où l'on ne se restreint pas nécessairement à un estimateur à noyau), alors les estimateurs à noyau adaptés aux densités conditionnelles que je vais utiliser par la suite seront de la forme suivante :

$$\hat{f}_h : w \mapsto \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\widetilde{\mathrm{f}}_X(X_i)} K_h(W_i - w). \tag{I.3}$$

Noter tout de même que cet estimateur nécessite que l'estimateur de la marginale $\widetilde{\mathrm{f}}_X$ soit particulièrement rapide à calculer, puisque l'on doit l'appliquer en chaque observation. Des hypothèses sur la qualité d'estimation de $\widetilde{\mathrm{f}}_X$ seront aussi nécessaires.

### 3.b.i Sélection gloutonne de la fenêtre : l'algorithme RODEO

Pour sélectionner la fenêtre de l'estimateur (I.3), Bertin et al. [2016] utilisent la méthode de Goldenshluger and Lepski [2011a]. Bien que leurs résultats théoriques atteignent les vitesses minimax adaptatives optimales, la méthode est difficilement applicable pour des dimensions supérieures à $3$. En effet, elle nécessite de construire exhaustivement les estimateurs d'une grille nécessairement fine de fenêtres, pour les comparer deux à deux et sélectionner le meilleur selon un critère bien choisi. La grille croissant exponentiellement avec la dimension, cette méthode est fortement soumise au fléau de la dimension numérique.

On va donc s'intéresser à une sélection gloutonne de la fenêtre, au sens où l'on ne se permet plus d'explorer exhausitivement une grille complète des fenêtres potentiellement optimales. Plus spécifiquement, détaillons l'algorithme RODEO qui a été étudié en régression [Lafferty and Wasserman 2008] et en estimation de densités [Liu et al. 2007] : c'est une procédure itérative qui construit à partir de tests sur les données un chemin $1$-dimensionnel de fenêtres décroissantes traversant la grille $d$-dimensionnelle de fenêtres afin d'atteindre une fenêtre réalisant le compromis biais-variance.
Plus précisément, étant donnés les estimateurs à noyau $\{\hat{m}_h\}$ adaptés au problème considéré, on teste le long du parcours de l'algorithme jusqu'à quel niveau les composantes de la fenêtre doivent

continuer de décroître. La statistique de test clé est la dérivée partielle de l'estimateur $\hat{m}_h$ dans la direction $h_j$ : au point d'évaluation $w$,

$$Z_{hj} := \frac{\partial}{\partial h_j} \hat{m}_h(w), \tag{I.4}$$

que l'on compare à son écart-type (à facteur logarithmique près) pour décider si $h_j$ doit continuer à décroître. Le biais étant la quantité difficile à étalonner, $h_j Z_{hj}$ permet de quantifier le « biais associée à la direction $j$ », ce qui permet de réaliser le compromis biais-variance.

L'avantage de cette considération composante par composante permet de détecter leur pertinence ou non, et ainsi de s'adapter à une structure de parcimonie éventuelle [Liu et al. 2007 ; Lafferty and Wasserman 2008].

# 4 Contributions de cette thèse

Dans le but d'estimer des densités conditionnelles en dimensions modérément grandes, mes travaux de thèse se sont attachés à répondre aux objectifs suivants :

(i) proposer une méthode convergeant à vitesse minimax optimale (à un facteur logarithme près),

(ii) adaptative à la fois en la régularité et en une éventuelle structure de parcimonie : plus précisément, dans le cas d'une densité conditionnelle $s$-régulière et ne dépendant localement que de $r$ composantes, la vitesse de convergence attendue est $n^{-\frac{s}{2s+r}}$ (à facteur logarithmique près) ;

(iii) à l'aide d'une procédure gloutonne, ce qui exclut les procédures dont la complexité croît exponentiellement avec la dimension.

Mes contributions se séparent en trois parties, constituant les trois prochains chapitres (Chapitres II, III et IV), le dernier chapitre (Chapitre V) laissant place aux discussions et aux perspectives.

Dans le Chapitre II (principalement issu de mon premier article Nguyen [2018]), je reprends l'algorithme RODEO que j'adapte à la famille d'estimateurs de Bertin et al. [2016] (définie à l'équation I.3). Je nomme dans un premier temps cette méthode CDRODEO, puis « Direct CDRODEO » pour la distinguer de la procédure « RevDir CDRODEO » étudiée dans le Chapitre III. Une analyse de la concentration des statistiques $Z_{hj} := \frac{\partial}{\partial h_j} \hat{f}_h(w)$ permet de contrôler avec grande probabilité le chemin de fenêtres parcouru par l'algorithme : l'algorithme s'arrête en au plus $\mathcal{O}(\log n)$ itérations, ce qui lui assure une complexité linéarithmique $\mathcal{O}(d.n \log n)$, et sélectionne une fenêtre permettant d'obtenir à régularité $s$ fixée la vitesse minimax optimale de convergence $n^{\frac{-s}{2s+d}}$ (à un facteur logarithme près), ce qui valide les objectifs (i) et (iii).

De plus, s'il existe une structure de parcimonie sur la densité conditionnelle $f$ au sens indiqué dans la définition I.4, cette vitesse est adaptative en la dimension pertinente $r \in \{0, 1, \dots, d\} : n^{\frac{-s}{2s+r}}$, et la complexité diminue : $\mathcal{O}(d.n + r.n \log n)$ (voir la section 3 du Chapitre II pour plus de précisions, en particulier sur les hypothèses), ce qui valide partiellement l'objectif (ii).
À ma connaissance, seuls les algorithmes de type RODEO permettent une sélection de la fenêtre à la fois gloutonne et adaptative. Comparons-nous donc à [Liu et al. 2007], qui a adapté RODEO dans le cadre de l'estimation de densités (non conditionnelles). Premièrement, nous étendons notre résultats à toute régularité $s \in \mathbb{N}_{>0}$ alors que Liu et al. [2007] fixent leur régularité. D'ailleurs, comme le fait remarquer Comminges and Dalalyan [2012], ils semblent fixer leur régularité $s = 2$,

mais émettent des hypothèses sur les dérivées d'ordre $4$, ce qui rend leur résultat non minimax optimal. Par ailleurs, nous proposons une notion de « composantes pertinentes » plus naturelle et moins restrictive. Comme Liu et al. [2007] étudient le risque $\mathrm{L}_2$ de leur estimateur, leur définition de composante pertinente est nécessairement globale. De plus, elle dépend d'une densité de référence choisie par la méthode, sans rapport avec la densité estimée, ce qui complexifie l'interprétation de cette notion de parcimonie. Notre définition I.4 est une simple propriété locale de non-dépendance de $f$ envers certaines de ses composantes. Elle est ainsi plus facile à réaliser, ce qui atténue d'autant mieux le fléau de la dimension.

Dans le Chapitre III, nous améliorons nos résultats précédents sur deux points délicats : l'adaptation en la régularité et un problème de dépendance en l'initialisation qui détériore notre vitesse de convergence de quelques facteurs logarithmiques. Pour améliorer ce fait, notre nouvel algorithme, appelé « RevDir CDRodeo », construit toujours un chemin monotone pour chaque composante de la fenêtre mais avec l'alternative de croître (par une étape "Reverse") ou de décroître (par une étape Direct). Noter qu'un algorithme entièrement Reverse était proposé dans [Liu et al. 2007] mais uniquement dans leurs simulations et sans aucune garantie théorique.
Une étude plus fine des nouveaux chemins probables (toujours construits par des tests sur les statistiques $Z_{hj}$ que ce soit pour la croissance ou la décroissance de la fenêtre) nous permet d'atteindre la vitesse minimax optimale doublement adaptative, $((\log n)^{1+\varepsilon}/n)^{\frac{s}{2s+r}}$, avec $\varepsilon > 0$ aussi petit que l'on veut proche de $0$ (si l'on change un paramètre de notre procédure) et avec $r \in \{0, \dots, d\}$ et $s \in (1, p]$ inconnus ($p$ étant l'ordre du noyau de notre estimateur). Noter qu'il existe des noyaux de tout ordre. Noter aussi que l'on s'adapte aussi à des régularités non nécessairement entières en passant par les classes de régularité höldérienne (voir la définition I.2). La complexité de ce nouvel algorithme $\mathcal{O}(d.n \log n)$ reste linéarithmique, toutefois sans amélioration en cas de parcimonie. Noter enfin que le facteur logarithme que nous obtenons est au $\varepsilon$ près minimax adaptatif optimal, ce qui suggèrerait que l'adaptation en la structure de parcimonie ne semble ajouter aucun facteur logarithmique. Ceci reste subordonné à notre jeu d'hypothèses, en particulier une hypothèse de convexité du biais (voir l'hypothèse $\mathcal{M}$ en section 3.c du Chapitre III). Rappelons que notre procédure est gloutonne. Cette hypothèse est donc à rapprocher des hypothèses de convexité pour les descentes de gradient : en particulier, comme on n'explore pas exhaustivement toutes les fenêtres, on a besoin de s'assurer qu'un zéro de la dérivée (ici, nos tests sur les $Z_{hj}$) ne nous amène pas à un minimum local sous-optimal.

Dans le Chapitre IV, nous nous intéressons aux performances numériques de nos méthodes. Après des considérations sur le coût algorithmique, on s'intéresse à la calibration des paramètres ainsi qu'aux limites d'application des procédures CDRodeo, avant de comparer leurs performances.

# Chapter II

# Greedy estimation of sparse conditional densities

This chapter is based on the submitted paper Nguyen [2018].

**Abstract :** • *In this chapter, we consider the problem of estimating a conditional density in moderately large dimensions. Much more informative than regression functions, conditional densities are of main interest in recent methods, particularly in the Bayesian framework (studying the posterior distribution, finding its modes...). Considering a recently studied family of kernel estimators, we select a pointwise multivariate bandwidth by revisiting the greedy algorithm* RODEO *(Regularisation Of Derivative Expectation Operator). The method addresses several issues: being greedy and computationally efficient by an iterative procedure, avoiding the curse of high dimensionality under some suitably defined sparsity conditions by early variable selection during the procedure, converging at a quasi-optimal minimax rate.* •

## Table of contents

# 1 Introduction

## 1.a Motivations

The problem of the conditional density estimation is defined as follows. We observe a $n$-sample of a couple $(X, Y)$, in which $Y$ is the vector of interest while $X$ gathers auxiliary variables. We denote $d$ the joint dimension. In particular we are interested in the inference of the $d$-dimensional conditional density $f$ of $Y$ conditionally to $X$.
There is a growing demand for methods of conditional density estimation in a wide spectrum of applications such as Economy[Hall et al. 2004], Cosmology [Izbicki and Lee 2016], Medicine [Takeuchi et al. 2009], Actuaries [Efromovich 2010b], Meteorology [Jeon and Taylor 2012] among others. It can be explained by the double role of the conditional density estimation: deriving the underlying distribution of a dataset and determining the impact of the vector $X$ of auxiliary variables on the vector of interest $Y$. In this aspect, the conditional density estimation is richer than both the unconditional density estimation and the regression problem. In particular, in the regression framework, only the conditional mean $\mathbb{E}[Y|X]$ is estimated instead of the full conditional density, which can be especially poorly informative in case of an asymmetric or multi-modal conditional density. Conversely, from the conditional density estimators, one can, *e.g.*, derive the conditional quantiles [Takeuchi et al. 2006] or give accurate predictive intervals [Fernández-Soto et al. 2002]. Furthermore, since the posterior distribution in the Bayesian framework is actually a conditional density, the present method also offers an alternative to the ABC methodology (for Approximate Bayesian Computation) [Beaumont et al. 2002 ; Marin et al. 2012 ; Biau et al. 2015] in the case of an intractable-yet-simulable model.
The challenging issue in conditional density estimation is to circumvent the "curse of dimensionality". The problem is twofold: theoretical and practical. In theory, it is stigmatized by the minimax approach, stating that in a $d$-dimensional space the best convergence rate for the pointwise risk over a $p$-regular class of functions is $\mathcal{O}(n^{-\frac{p}{2p+d}})$: in particular, the larger $d$, the slower the rate. In practice, the larger the dimension is, the larger the sample size is needed to control the estimation error. In order to maintain reasonable running times in moderately large dimensions, methods have to be designed especially greedy.
Furthermore, one interesting question is how to retrieve the eventual *relevant* components in case of sparsity structure on the conditional density $f$. For example, if we have at disposal plenty of auxiliary variables without any indication on their dependency with our vector of interest $Y$, the ideal procedure will take in input the whole dataset and still achieve a running time and a minimax rate as fast as if only the relevant components were given and considered for the estimation. More precisely, two goals are simultaneously addressed : converging at rate $\mathcal{O}(n^{-\frac{2p}{2p+r}})$ with $r$ the relevant dimension, *i.e.* the number of components that influence the conditional density $f$, and detect the irrelevant components at an early stage of the procedure in order to afterwards only work on the relevant data and thus speed up the running time.

## 1.b Existing methodologies

Several nonparametric methods have been proposed to estimate conditional densities: kernel density estimators [Rosenblatt 1969 ; Hyndman et al. 1996 ; Bertin et al. 2016] and various methodologies for the selection of the associated bandwidth [Bashtannyk and Hyndman 2001 ; Fan and Yim 2004 ; Hall et al. 2004]; local polynomial estimators [Fan et al. 1996 ; Hyndman and Yao 2002]; projection series estimators [Efromovich 1999; 2007]; piecewise constant estimator [Györfi and Kohler 2007 ; Sart 2017]; copula [Faugeras 2009]. But while most of the aforemen-

tioned works are only defined for bivariate data or at least when either $X$ or $Y$ is univariate, they are also computationally intractable as soon as $d > 3$.
It is in particular the case for the kernel density methodologies (Hall, Racine ,Li 2004, Bertin et al. 2016): they achieve the optimal minimax rate, and even the detection of the relevant components, thanks to an adequate choice of the bandwidth (for the two aforementioned methods by cross validation and Goldenshluger-Lepski methodology), but the computational cost of these bandwidth selections is prohibitive even for moderate sizes of $n$ and $d$. To the best of our knowledge, only two kernel density methods have been proposed to handle large datasets. [Holmes et al. 2010] propose a fast method of approximated cross-validation, based on a dual-tree speed-up, but they do not establish any rate of convergence and only show the consistency of their method. For scalar $Y$, [Fan et al. 2009] proposed to perform a prior step of dimension reduction on $X$ to bypass the curse of dimensionality, then they estimate the bivariate approximated conditional density by kernel estimators. But the proved convergence rate $n^{-\frac{1}{3}}$ is not the optimal minimax rate $n^{-\frac{3}{8}}$ for the estimation of a bivariate function of assumed regularity $3$. Moreover, the step of dimension reduction restricts the dependency of $X$ to a linear combination of its components, which may induce a significant loss of information.
Projection series methods for scalar $Y$ have also been proposed. [Efromovich 2010a] extends his previous work [Efromovich 2007] to a multivariate $X$. Theoretically the method achieves an oracle inequality, thus the optimal minimax rate. Moreover it performs an automatic dimension reduction on $X$ when there exists a smaller intrinsic dimension. However the computation cost is prohibitive when both $n$ and $d$ are large. More recently, Izbicki and Lee have proposed two methodologies using orthogonal series estimators [Izbicki and Lee 2016; 2017]. The first method is particularly fast and can handle very large $X$ (with more than $1000$ covariates). Moreover the convergence rate adapts to an eventual smaller unknown intrinsic dimension of the support of the conditional density. The second method originally proposes to convert successful high dimensional regression methods into the conditional density estimation, interpreting the coefficients of the orthogonal series estimator as regression functions, which allows to adapt to all kind of figures (mixed data, smaller intrinsic dimension, relevant variables) in function of the regression method. However both methods converge slower than the optimal minimax rate. Moreover their optimal tunings depend in fact on the unknown intrinsic dimension.
For multivariate $X$ and $Y$, [Otneim and Tjøstheim 2018] propose a new semiparametric method, called Locally Gaussian Density Estimator: they rewrite the conditional density as a product of a function depending on the marginal distribution functions (easily estimated since univariate, then plug-in), and a term which measures the dependency between the components, which is approximated by a centred Gaussian whose covariance is parametrically estimated. Numerically, the methodology seems robust to addition of covariates of $X$ independent of $Y$, but it is not proved. Moreover they only establish the asymptotic normality of their method.

## 1.c Our strategy and contributions

Our challenge is to handle large datasets, thus we assume at our disposal a sample of large size $n$ and of moderately large dimension. Then our work is motivated by the following three objectives:

(i) achieving the optimal minimax rate (up to a logarithm term);

(ii) being greedy, meaning that the procedure must have reasonable running times for large $n$ and moderately large dimensions, in particular when $d > 3$ ;

(iii) adapting to a potential sparsity structure of $f$. More precisely, in the case where $f$ locally depends only on a number $r$ of its $d$ components, $r$ can be seen as the local *relevant* dimension. Then the desired convergence rate has to adapt to the unknown relevant dimension $r$: under this sparsity assumption, the benchmark for the estimation of a $p$-regular function is to achieve a convergence rate of the order $\mathcal{O}(n^{-\frac{2p}{2p+r}})$, which is the optimal minimax rate if the relevant components were given by an oracle.

Our strategy is based on kernel density estimators. The considered family has been recently introduced and studied in [Bertin et al. 2016]. This family is especially designed for conditional densities and is better adapted for the objective (iii) than the intensively studied estimator built as the ratio of a kernel estimator of the joint density over one of the marginal density of $X$. For example, a relevant component for the joint density and the marginal density of $X$ may be irrelevant for the conditional density and it is the case if a component of $X$ is independent of $Y$. Note though that many more cases of irrelevance exist since we define the relevance as a local property.
The main issue with kernel density estimators is the selection of the bandwidth $h \in \mathbb{R}_+^d$, and in our case, we also want to complete the objective (ii), since the pre-existing methodologies of bandwidth selection do not satisfy this restriction and thus cannot handle large datasets. In the following, it is performed by a new algorithm we call CDRODEO, which is derived from the algorithm RODEO [Lafferty and Wasserman 2008 ; Liu et al. 2007], which has respectively been applied for the regression and the unconditional density estimation. The greediness of the algorithm allows us to address datasets of large sizes while keeping a reasonable running time (see Section 3.e for further details). We give a simulated example with a sample of size $n = 10^5$ and of dimension $d = 5$ in Section 4. Moreover, RODEO-type algorithms ensure an early detection of irrelevant component, and thus achieve the objective (iii) while improving the objective (ii).
From the theoretical point of view, if the regularity of $f$ is known, our method achieves an optimal minimax rate (up to a logarithmic factor), which is adaptive to the unknown sparsity of $f$. The last property is mostly due to the RODEO-type procedures. The improvement of our method in comparison to the paper [Liu et al. 2007] which estimates the *unconditional* density with RODEO is twofold. First, our result is extended to any regularity $p \in \mathbb{N}_{>0}$, whereas [Liu et al. 2007] fixed $p = 2$. Secondly, our notion of relevance is both less restrictive and more natural. In [Liu et al. 2007], they studied the $L_2$-risk of their estimator, therefore they have to consider a notion of global relevance, whereas we consider a pointwise approach, which allows us to define a local property of relevance, which can be applied to a broader class of functions. Moreover, their notion of relevance is not intrinsic to the unknown density, but in fact depends on a tuning of the method, a prior chosen *baseline density* which has no connexion with the density, which limits the interpretation.

## 1.d Overview

The chapter is organized as follows. We introduce the CDRODEO method in Section 2. The theoretical results are in Section 3, in which we specify the assumptions and the tunings of the procedure from which are derived the convergence rate and the complexity cost of the method. A numerical example is presented in Section 4. The proofs are in the last section.

## 2 CDRODEO method

Let $W_1, \dots, W_n$ be a sample of a couple $(X, Y)$ of multivariate random vectors: for $i = 1, \dots, n$,

$$W_i = (X_i, Y_i),$$

with $X_i$ valued in $\mathbb{R}^{d_1}$ and $Y_i$ in $\mathbb{R}^{d_2}$. We denote $d := d_1 + d_2$ the joint dimension.
We assume that the marginal distribution of $X$ and the conditional distribution of $Y$ given $X$ are absolutely continuous with respect to the Lebesgue measure, and we define $f : \mathbb{R}^d \to \mathbb{R}$ such as for any $x \in \mathbb{R}^{d_1}$, $f(x, \cdot)$ is the conditional density of $Y$ conditionally to $X = x$. We denote $\mathrm{f}_X$ the marginal density of $X$.
Our method estimates $f$ pointwisely : let us fix $w = (x, y) \in \mathbb{R}^d$ the point of interest.

**Kernel estimators.** Our method is based on kernel density estimators. More specifically, we consider the family proposed in [Bertin et al. 2016], which is especially designed for the conditional density estimation. Let $K : \mathbb{R} \to \mathbb{R}$ be a kernel function, *ie:* $\int_{\mathbb{R}} K(t)dt = 1$, then for any bandwidth $h \in (\mathbb{R}_+^*)^d$, the estimator of $f(w)$ is defined by:

$$\hat{f}_h(w) := \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\widetilde{\mathrm{f}}_X(X_i)} \prod_{j=1}^{d} h_j^{-1} K\left(\frac{w_j - W_{ij}}{h_j}\right), \tag{II.1}$$

where $\widetilde{\mathrm{f}}_X$ is an estimator of $\mathrm{f}_X$, built from another sample $\widetilde{X}$ of $X$. We denote by $n_X$ the sample size of $\widetilde{X}$. The choices of $K$ and $\widetilde{\mathrm{f}}_X$ are specified later (see section 3.b).

**Bandwidth selection.** In kernel density estimation, selecting the bandwidth is a critical choice which can be viewed as a bias-variance trade-off. In [Bertin et al. 2016], it is performed by the Goldenshluger-Lepski methodology (see [Goldenshluger and Lepski 2011a]) and requires an optimization over an exhaustive grid of couples $(h, h')$ of bandwidths, which leads to intractable running time when the dimension exceeds 3 (and large dataset).
That is why we focus in a method which excludes optimization over an exhaustive grid of bandwidths to rather propose a greedy algorithm derived from the algorithm RODEO. First introduced in the regression framework [Wasserman and Lafferty 2006 ; Lafferty and Wasserman 2008], a variation of RODEO was proposed in [Liu et al. 2007] for the density estimation. Our method we called CDRODEO (for Conditional Density RODEO) addresses the more general problem of conditional density estimation.
Like RODEO (which means Regularisation Of Derivative Expectation Operator), the CDRODEO algorithm generates an iterative path of decreasing bandwidths, based on tests on the partial derivatives of the estimator with respect to the components of the bandwidth. Note that the greediness of the procedure leans on the selection of this path of bandwidths, which enables us to address high dimensional problems of functional inference.
Let us be more precise: we take a kernel $K$ of class $\mathcal{C}^1$ and consider the statistics $Z_{hj}$ for $h \in (\mathbb{R}_+^*)^d$ and $j = 1 : d$, defined by:

$$Z_{hj} := \frac{\partial}{\partial h_j} \hat{f}_h(w).$$

$Z_{hj}$ is easily computable, since it can be expressed by:

$$Z_{hj} = \frac{-1}{n h_j^2} \sum_{i=1}^{n} \frac{1}{\widetilde{\mathrm{f}}_X(X_i)} J\left(\tfrac{w_j - W_{ij}}{h_j}\right) \prod_{k \neq j}^{d} h_k^{-1} K\left(\frac{w_k - W_{ik}}{h_k}\right), \tag{II.2}$$

where $J : \mathbb{R} \to \mathbb{R}$ is the function defined by:

$$t \mapsto K(t) + tK'(t). \tag{II.3}$$

**Algorithm 1** CDRODEO algorithm

1. *Input:* the point of interest $w$, the data $W$, $\beta \in (0,1)$ the bandwidth decreasing factor, $h_0 > 0$ the bandwidth initialization value, a parameter $a > 1$.
2. *Initialization:*
   (a) Initialize the bandwidth: *for* $j = 1 : d$, $h_j \leftarrow h_0$.
   (b) Activate all the variables: $\mathcal{A} \leftarrow \{1, \dots, d\}$.
3. *While* ($\mathcal{A} \neq \emptyset$) & ($\prod_{k=1}^{d} h_k \geq \frac{\log n}{n}$):
   *for all* $j \in \mathcal{A}$:
   (a) Update $Z_{hj}$ and $\lambda_{hj}$.
   (b) *If* $|Z_{hj}| \geq \lambda_{hj}$: update $h_j \leftarrow \beta h_j$.
   *else:* remove $j$ from $\mathcal{A}$.
4. *Output:* $h$ (and $\hat{f}_h(w)$).

The details of the CDRODEO procedure are described in **Algorithm 1** and can be summed up in one sentence: for a well-chosen threshold $\lambda_{hj}$ (specified in Section 3.c), the algorithm performs at each iteration the test $|Z_{hj}| > \lambda_{hj}$ to determine if the component $j$ of the current bandwidth must be shrunk or not. It can be interpreted by the following principle: the bandwidth of a kernel estimator quantifies within which distance of the point of interest $w$ and at which degree an observation $W_i$ helps in the estimation. Heuristically, the larger the variation of $f$ is, the smaller the bandwidth is required for an accurate estimation. The statistics $Z_{hj} = \frac{\partial}{\partial h_j}\hat{f}_h(w)$ are used as a proxy of $\frac{\partial}{\partial w_j}f(w)$ to quantify the variation of $f$ in the direction $w_j$. Note in particular that since the partial derivatives vanish for irrelevant components, this bandwidth selection leads to an implicit variable selection, and thus to avoid the curse of dimensionality under sparsity assumptions.

# 3 Theoretical results

This section gathers the theoretical results of our method.

## 3.a Assumptions

We consider $K$ a compactly supported kernel. For any bandwidth $h \in (\mathbb{R}_+^*)^d$, we define the neighbourhood $\mathcal{U}_h(u)$ of $u \in \mathbb{R}^{d'}$ (typically, $u = x$ or $w$, and $d' = d_1$ or $d$) as follows:

$$\mathcal{U}_h(u) := \left\{ u' \in \mathbb{R}^{d'} : \forall j = 1 : d', u'_j = u_j - h_j z_j, \text{with } z \in (\text{supp}(K))^{d'} \right\}.$$

Then we denote the CDRODEO initial bandwidth $h^{(0)} = \left(\frac{1}{\log n}, \dots, \frac{1}{\log n}\right)$ and for short, $\mathcal{U}_n(u) := \mathcal{U}_{h^{(0)}}(u)$.
We also introduce the notation $\|\cdot\|_{\infty,\mathcal{U}}$ for the supremum norm over a set $\mathcal{U}$.

The following first assumption ensures a certain amount of observations in the neighbourhood of our point of interest $w$.

**Assumption II.1** ($\mathrm{f}_X$ bounded away of 0). *We assume* $\delta := \inf_{u \in \mathcal{U}_n(x)} \mathrm{f}_X(u) > 0$.

Note that if the neighbourhood $\mathcal{U}_n(x)$ does not contain any observation $X_i$, the estimation of the conditional distribution of $Y$ given the event $X = x$ is obviously intractable.

The second assumption specifies the notions of "sparse function" and "relevant component", under which the curse of high dimensionality can be avoided.

**Assumption II.2** (Sparsity condition). *There exists a subset* $\mathcal{R} \in \{1, \dots, d\}$ *such that for any fixed* $\{z_j\}_{j \in \mathcal{R}}$, *the function* $\{z_k\}_{k \in \mathcal{R}^c} \mapsto f(z_1, \dots, z_d)$ *is constant on* $\mathcal{U}_n(w)$.

In other words, if we denote $r$ the cardinal of $\mathcal{R}$, Assumption II.2 means that $f$ locally depends on only $r$ of its $d$ variables. We call *relevant* any component in $\mathcal{R}$. The notion of relevant component depends on the point where $f$ is estimated. For example, a component $w_j$ which behaves as $\mathbb{1}_{[0,1]}(w_j)$ in the conditional density is only relevant in the neighbourhood of 0 and 1. Note that this local property addresses a broader class of functions, which extends the application field of Theorem II.2 and improves the convergence rate of the method.

Finally, the conditional density is required to be regular enough.

**Assumption II.3** (Regularity of $f$). *There exists a known integer* $p$ *such that* $f$ *is of class* $\mathcal{C}^p$ *on* $\mathcal{U}_n(w)$ *and such that* $\partial_j^p f(w) \neq 0$ *for all* $j \in \mathcal{R}$.

## 3.b Conditions on the estimator of $\mathrm{f}_X$

Given the definition of the estimator (II.1), we need an estimator $\widetilde{\mathrm{f}}_X$ of $\mathrm{f}_X$.

**If $\mathrm{f}_X$ is known.** We take $\widetilde{\mathrm{f}}_X \equiv \mathrm{f}_X$. This case is not completely obvious. In particular, it tackles the case of *un*conditional density estimation, if we set by convention $d_1 = 0$ and $\mathrm{f}_X \equiv 1$.

**If $\mathrm{f}_X$ is unknown.** We need an estimator $\widetilde{\mathrm{f}}_X$ which satisfies the following two conditions:

(i) a positive lower bound: $\widetilde{\delta}_X := \inf_{u \in \mathcal{U}_n(x)} \widetilde{\mathrm{f}}_X(u) > n^{-\chi}$, for some $\chi > 0$,

(ii) a concentration inequality in local sup norm: there exists a constant $M_X > 0$ such that:

$$\mathbb{P}\left( \sup_{u \in \mathcal{U}_n(x)} \left| \frac{\mathrm{f}_X(u) - \widetilde{\mathrm{f}}_X(u)}{\widetilde{\mathrm{f}}_X(u)} \right| > M_X \frac{(\log n)^{\frac{d}{2}}}{n^{\frac{1}{2}}} \right) \leq \exp(-(\log n)^{\frac{5}{4}}).$$

The first condition on $\widetilde{\mathrm{f}}_X$ is used to control its inverse, and the second one ensures it is sufficiently accurate. Note that these conditions are feasible, as it is proved in the following proposition. Furthermore, the provided estimator of $\mathrm{f}_X$ (see the proof in Section 5.c.iii) is easily implementable and does not need any optimisation.

**Proposition II.1.** *Given a sample* $\widetilde{X}$ *with same distribution as* $X$ *and of size* $n_X = n^c$ *with* $c > 1$, *if* $\mathrm{f}_X$ *is of class* $\mathcal{C}^{p'}$ *with* $p' \geq \frac{d_1}{2(c-1)}$, *there exists an estimator* $\widetilde{\mathrm{f}}_X$ *which satisfies (i) and (ii).*

## 3.c CDRODEO parameters choice.

**Kernel $K$.** We choose the kernel function $K : \mathbb{R} \to \mathbb{R}$ of class $\mathcal{C}^1$, with compact support and of order $p$, *i.e.*: for $\ell = 1, \dots, p-1$, $\int_{\mathbb{R}} t^\ell K(t)dt = 0$, and $\int_{\mathbb{R}} t^p K(t)dt \neq 0$.
Note that considering a compactly supported kernel is fundamental for the local approach. In particular, it relaxes the assumptions by restricting them to a neighbourhood of $w$.
Taking a kernel of order $p$ is usual for the control of the bias of the estimator.

**Parameter $\beta$.** Let $\beta \in (0,1)$ be the decreasing factor of the bandwidth. The larger $\beta$, the more accurate the procedure, but the longer the computational time. From the theoretical point of view, it remains of little importance, as it only affects the constant terms. In practice, we set it close to 1.

**Bandwidth initialization.** We recall that we set $h_0 := \frac{1}{\log n}$ (and the initial bandwidth as $\left(\frac{1}{\log n}, \dots, \frac{1}{\log n}\right)$).

**Threshold $\lambda_{h,j}$.** For any bandwidth $h \in (\mathbb{R}_+^*)^d$ and for $j = 1 : d$, we set the threshold as follows:

$$\lambda_{hj} := \mathsf{C}_\lambda \sqrt{\frac{(\log n)^a}{n h_j^2 \prod_{k=1}^d h_k}}, \tag{II.4}$$

with $\mathsf{C}_\lambda := 4\|J\|_2 \|K\|_2^{d-1}$ (where $J$ is defined in (II.3)) and $a > 1$. The expression is obtained by using concentration inequalities on $Z_{hj}$. For the proof, the parameter $a$ has to be tuned such that:

$$(\log n)^{a-1} > \frac{\|f\|_{\infty,\, \mathcal{U}_n(w)}}{\delta}, \tag{II.5}$$

which is satisfied for $n$ large enough. The influence of this parameter is discussed in the next section, once the theoretical results are stated.

Hereafter, unless otherwise specified, the parameters are chosen as described in this section.

## 3.d Mains results

Let us denote $\hat{h}$ the bandwidth selected by CDRODEO. In Theorem II.2, we introduce a set $\mathcal{H}_{\mathsf{hp}}$ of bandwidths which contains $\hat{h}$ with high probability, which leads to an upper bound of the pointwise estimation error with high probability. In Corollary II.3, we deduce the convergence rate of CDRODEO from Theorem II.2.
More precisely, in Theorem II.2, we determine lower and upper bounds (with high probability) for the stopping iteration of each bandwidth component. We set:

$$\tau_n := \frac{1}{(2p+r)\log\frac{1}{\beta}} \log\left(\frac{n}{\mathsf{C}_\tau (\log n)^{2p+d+a}}\right), \tag{II.6}$$

and

$$T_n := \tau_n + \frac{\log\left(\mathsf{C}_T^{-1}\right)}{(2p+1)\log\frac{1}{\beta}}, \tag{II.7}$$

where

$$\mathsf{C}_\tau := \left( \frac{4(p-1)!\mathsf{C}_\lambda}{\left(\min_{j\in\mathcal{R}} \partial_j^p f(w)\right) \int_{\mathbb{R}} t^p K(t)dt} \right)^2, \mathsf{C}_T := \left( \frac{\min_{j\in\mathcal{R}} |\partial_j^p f(w)|}{24 \max_{j\in\mathcal{R}} |\partial_j^p f(w)|} \right)^2.$$

Then we define the set of bandwidths $\mathcal{H}_{\mathsf{hp}}$ by:

$$\mathcal{H}_{\mathsf{hp}} := \left\{ h \in \mathbb{R}_+^d : h_j = \frac{\beta^{\theta_j}}{\log n}, \text{ with } \theta_j \in \{\lfloor \tau_n \rfloor + 1, \dots, \lfloor T_n \rfloor\} \text{ if } j \in \mathcal{R}, \text{ else } \theta_j = 0 \right\}.$$

**Theorem II.2.** *Assume that $\widetilde{\mathrm{f}}_X$ satisfies (i) and (ii) of section 3.b and Assumptions II.1 to II.3 are satisfied. Then, the bandwidth $\hat{h}$ selected by CDRODEO belongs to $\mathcal{H}_{hp}$ with high probability. More precisely, for any $q > 0$ and for $n$ large enough:*

$$\mathbb{P}\left(\hat{h} \in \mathcal{H}_{hp}\right) \geq 1 - n^{-q}. \tag{II.8}$$

*Moreover, with probability larger than $1 - 2n^{-q}$, the CDRODEO estimator $\hat{f}_{\hat{h}}(w)$ verifies:*

$$\left|\hat{f}_{\hat{h}}(w) - f(w)\right| \leq C(\log n)^{\frac{p}{2p+r}(d-r+a)} n^{-\frac{p}{2p+r}} \tag{II.9}$$

*with*

$$C := 2r\mathsf{C}_\tau^{\frac{p}{2p+r}} \int_{t\in\mathbb{R}} \left|\frac{t^p}{p!} K(t)\right| dt \times \max_{k\in\mathcal{R}} \|\partial_k^p f\|_{\infty,\, \mathcal{U}_n(w)} + 4\|K\|_2^d \|f\|_{\infty,\, \mathcal{U}_n(w)}^{\frac{1}{2}} \delta^{-\frac{1}{2}} \mathsf{C}_T^{\frac{-r}{2(2p+1)}} \mathsf{C}_\tau^{\frac{-r}{2(2p+r)}}.$$

**Corollary II.3.** *Under the assumptions of Theorem II.2, for any $q \geq 1$:*

$$\left(\mathbb{E}\left[\left|\hat{f}_{\hat{h}}(w) - f(w)\right|^q\right]\right)^{1/q} \leq C(\log n)^{\frac{p}{2p+r}(d-r+a)} n^{-\frac{p}{2p+r}} + o\left(n^{-1}\right).$$

Corollary II.3 presents a generalization of the previous works on RODEO [Lafferty and Wasserman 2008] and [Liu et al. 2007] whose results are restricted to the regularity $p = 2$ and to simpler problems, namely regression and density estimation.

We compare the convergence rate of CDRODEO with the optimal minimax rate. In particular, our benchmark is the pointwise minimax rate, which is of order $\mathcal{O}\left(n^{-\frac{p}{2p+d}}\right)$, for the problem of $p$-regular $d$-dimensional density estimation, obtained by [Donoho and Low 1992].

Without sparsity structure ($r = d$), CDRODEO achieves the optimal minimax rate, up to a logarithmic factor. The exponent of this factor depends on the parameter $a$. For the proofs, we need $a > 1$ in order to satisfy (II.5), but if an upper bound (or a pre-estimator) of $\frac{\|f\|_{\infty,\, \mathcal{U}_n(w)}}{\delta}$ were known, we could obtain the similar result with $a = 1$ and a modified constant term. Note that the logarithmic factor is a small price to pay for a computationally-tractable procedure for high-dimensional functional inference, in particular see section 3.e for the computational gain of our procedure.

Under sparsity assumptions, we avoid the curse of high dimensionality and our procedure achieves the desired rate $n^{-\frac{p}{2p+r}}$ (up to a logarithmic term), which is optimal if the relevant components were known. Note that some additional logarithmic factors could be unavoidable due to the unknown sparsity structure, which needs to be estimated. Identifying the exact order of the logarithm term in the optimal minimax rate for the sparse case remains an open challenging question.

### 3.e Complexity

We now discuss the complexity of CDRODEO without taking into account the pre-computation cost of $\widetilde{\mathrm{f}}_X$ at the points $X_i$, $i = 1 : n$ (used for computing the $Z_{hj}$), but a fast procedure for $\widetilde{\mathrm{f}}_X$ is required, to avoid losing CDRODEO computational advantages.
For CDRODEO, the main cost lies in the computation of the $Z_{hj}$'s along the path of bandwidths. The condition $\prod_{k=1}^{d} h_k \geq \frac{\log n}{n}$ restricts to at most $\log_{\beta^{-1}} n$ updates of the bandwidth across all components, leading to a worst-case complexity of order $\mathcal{O}(d.n\log n)$.
But as shown in Theorem II.2, with high probability, $\hat{h} \in \mathcal{H}_{\mathsf{hp}}$, in which only the relevant components are active after the first iteration. In first iteration, the $Z_{h^{(0)}j}$'s computation costs $\mathcal{O}(d.n)$ operations, while the product kernel enables us to compute the $Z_{hj}$'s in following iteration with only $\mathcal{O}(r.n)$ operations, which leads to the complexity $\mathcal{O}(d.n + r.n\log n)$.
In order to grasp the advantage of CDRODEO greediness, we compare its complexity with optimization over an exhaustive bandwidth grid with $\log n$ values for each component of the bandwidth (which is often the case in others methods: Cross validation, Lepski methods...): for each bandwidth of $(\log n)^d$-sized grid, the computation of a statistic from the $d.n$-sized dataset needs at least $\mathcal{O}(d.n)$ operation, which leads to a complexity of order $\mathcal{O}(d.n(\log_{\beta^{-1}} n)^d)$. Using the parameters used in the simulated example in section 4 ($n = 2.10^5$, $d = 5$, $r = 3$, $\beta = 0.95$), the ratio of complexities is $\frac{d.n(\log n)^d}{r.n\log n} \approx 5.10^9$, and even without sparsity structure: $\frac{d.n(\log n)^d}{d.n\log n} \approx 3.10^9$. It means that CDRODEO run is a billion times faster on this data set.

## 4 Simulations

In this section, we test the practical performances of our method. In particular, we study CDRODEO on a 5-dimensional example. The major purpose of this section is to assess the numerical performances of our procedure. Let us describe the example. We set $d_1 = 4$ and $d_2 = 1$ and simulate an i.i.d sample $\{(X_i, Y_i)\}_{i=1}^n$ with the following distribution: for any $i = 1, \ldots, n$:

- the first component $X_{i1}$ of $X_i$ follows a uniform distribution on $[-1, 1]$,
- the other components $X_{ij}$, $j = 2 : 4$, are independent standard normal and are independent of $X_{i1}$,
- $Y_i$ is independent of $X_{i1}$, $X_{i3}$ and $X_{i4}$ and the conditional distribution of $Y_i$ given $X_{i2}$ is exponential with survival parameter $X_{i2}^2$.

The estimated conditional density function is then defined by:

$$f : (x, y) \mapsto \mathbb{1}_{[-1,1]}(x_1)\frac{1}{x_2^2}e^{-\frac{y}{x_2^2}}.$$

This example enables us to test several criteria: sparsity detection, behaviour when fonctions are not continuous, bimodality estimation, robustness when $\mathrm{f}_X$ takes small values.
In the following simulations, if not stated explicitly otherwise, RODEO is run with sample size $n = 200,000$, product Gaussian kernel, initial bandwidth value $h_0 = 0.4$, bandwidth decreasing factor $\beta = 0.95$ and parameter $a = 1.1$ and $\widetilde{\mathrm{f}}_X \equiv \mathrm{f}_X$.
Figure II.1 illustrates CDRODEO bandwidth selection. In which, the boxplots of each selected bandwidth component are built from 200 runs of CDRODEO at the point $w = (0, 1, 0, 0, 1)$. This figure reflects the specificity of CDRODEO to capture the relevance degree of each component, and

one could compare it with variable selection (as done in [Lafferty and Wasserman 2008]). The components $x_3$ and $x_4$ are irrelevant and for this point of interest, the components $x_2$ and $y$ are clearly relevant while the component $x_1$ is barely relevant as $f$ is constant in the direction $x_1$ in near neighbourhood of $x_1 = 0$. As expected, the irrelevant $h_3$ and $h_4$ are mostly deactivated at the first iteration, while the relevant $h_2$ and $h_5$ are systematically shrunk. The relevance degree of $x_1$ is also well detected as the values of $h_1$ are smaller than $h_0$, but significantly larger than $h_2$ and $h_5$.

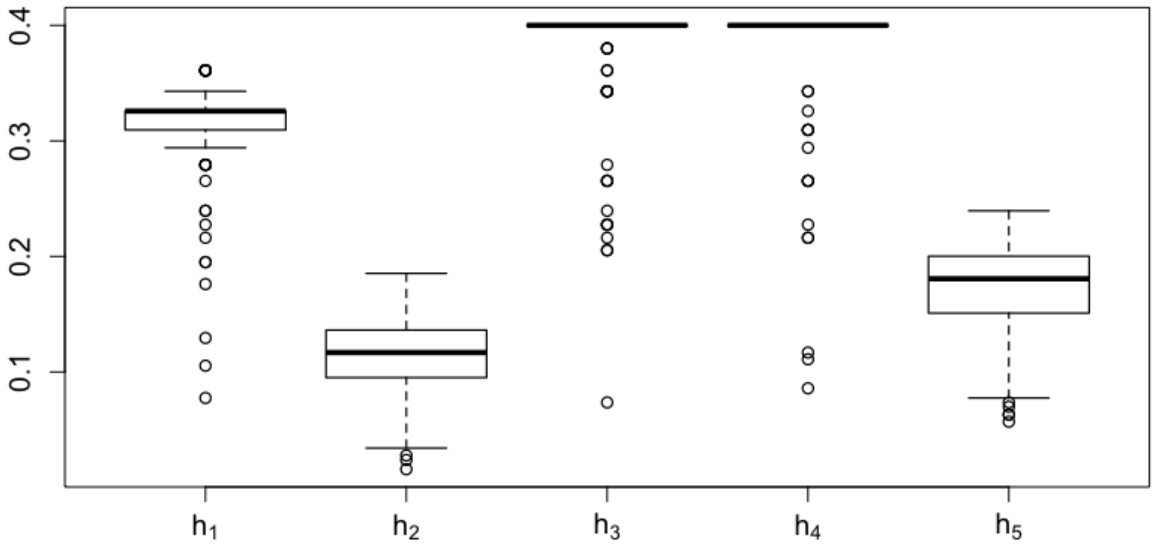



Figure II.1 – Boxplots of each component of $200$ CDRODEO selected bandwidths at the point $w = (0,1,0,0,1)$.

Figure II.2 gives CDRODEO estimation of $f$ from one $n$-sample. The function $f$ is well estimated. In particular, irrevance, jumps and bi-modality are features which are well detected by our method. As expected, main estimation errors are made on points of discontinuity for $x_1$ and $y$ or at the boundaries for $x_2$, $x_3$ and $x_4$. Note that the $\mathrm{f}_X$ values are particularly small at the boundaries of the plots in function of $x$, leading to lack of observations for the estimation. Note however that null value for $\mathrm{f}_X$ does not deteriorate the estimation (cf top left plot), since the estimate of $f$ vanishes automatically when there is no observation near the point of interest.

**Running time.** The simulations are implemented in R on a Macintosh laptop with a $3,1$ GHz Intel Core i7 processor. In the Figure Figure II.1, the $200$ runs of CDRODEO take $2952.735$ seconds (around $50$ minutes), or $14.8$ seconds per run.

# 5 Proofs

We first give the outlines of the proofs in Section 5.a. To facilitate the lecture of the proof, we have divided the proofs of the main results (Proposition II.1, Theorem II.2 and Corollary II.3) into intermediate results which are stated in Section 5.b and proved in Section 5.d. The proof of the main results are in Section 5.c.

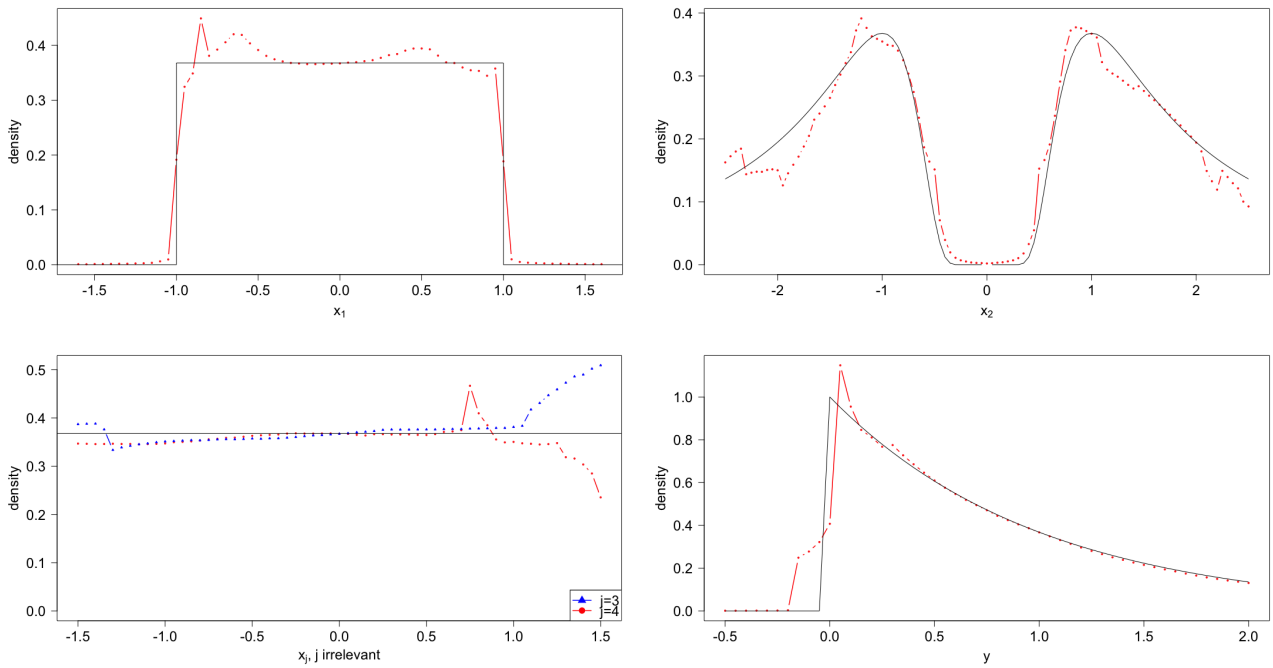



Figure II.2 – CDRODEO estimator (red or blue dashed lines) VS the true density (black solid line) in function of each component, the others component being fixed following $(x, y) = (0, 1, 0, 0, 1)$.

## 5.a Outlines of the proofs

We first prove Proposition II.1 by constructing an estimator of $\mathrm{f}_X$ with the wanted properties. In this proof, we use some usual properties of a kernel density estimator (control of the bias, concentration inequality), which are gathered in Lemma II.4.

Theorem II.2 states two results: the bandwidth selection (II.8) and the estimation error of the procedure (II.9). For the proof of the bandwidth selection (II.8), Proposition II.8 mades explicit the highly probable behaviour of CDRODEO along a run, and thus the final selected bandwidth. In particular, the proof leans on an analysis of $Z_{hj}$, which is made in two steps. We first consider $\bar{Z}_{hj}$, a simpler version of $Z_{hj}$ in which we substitute the estimator of $\mathrm{f}_X$ by $\mathrm{f}_X$ itself, and we detail its behaviour in Lemma II.6. Then we control the difference $Z_{hj} - \bar{Z}_{hj}$ (see in 1. of Lemma II.7) to ensure $\bar{Z}_{hj}$ behaves like $Z_{hj}$.

To control the estimation error of the procedure (II.9), we similarly analyse $\hat{f}_h(w)$ in two parts: in Lemma II.5, we describe the behaviour of $\bar{f}_h(w)$, the simpler version of $\hat{f}_h(w)$ in which we substitute the estimator of $\mathrm{f}_X$ by $\mathrm{f}_X$ itself, and in 2. of Lemma II.7, we bound the difference $f_h - \bar{f}_h(w)$. Then the bandwidth selection (II.8) leads to the upper bound with high probability of the estimation error of $\hat{f}_h(w)$ (II.9).

Finally, we obtain the expected error of $\hat{f}_h(w)$ stated in Corollary II.3 by controlling the error on the residual event.

## 5.b Intermediate results

For any bandwidth $h_X \in \mathbb{R}_+^*$, we define the kernel density estimator $\tilde{\mathrm{f}}_X^K$ by: for any $u \in \mathbb{R}^{d_1}$,

$$\tilde{\mathrm{f}}_X^K(u) := \frac{1}{n_X . h_X^{d_1}} \sum_{i=1}^{n_X} \prod_{j=1}^{d_1} K_X \left( \frac{u_j - \widetilde{X}_{ij}}{h_X} \right), \tag{II.10}$$

where $K_X : \mathbb{R} \to \mathbb{R}$ a kernel which is compactly supported, of class $\mathcal{C}^1$, of order $p_X \geq \frac{d_1}{2(c-1)}$, where we recall that $c > 1$ is defined by $n_X = n^c$.
We also introduce the neighbourhood

$$\mathcal{U}_n'(x) := \{u' = u - h_X z : u \in \mathcal{U}_n(x), z \in \text{supp}(K_X)\}. \tag{II.11}$$

**Lemma II.4** ($\tilde{\mathrm{f}}_X^K$ behaviour). *We assume $\mathrm{f}_X$ is $\mathcal{C}^{p'}$ on $\mathcal{U}_n'(x)$ with $p' \leq p_X$, then for any bandwidth $h_X \in \mathbb{R}_+^*$,*

1. *if we denote $C_{bias_X} := \frac{\|K_X\|_1^{d_1-1}\|\cdot^{p'} K_X(\cdot)\|_1}{p'!} d_1 \max_{k=1:d_1} \|\partial_k^{p'} \mathrm{f}_X\|_{\infty,\,\mathcal{U}_n'(x)}$, then*

$$\left\|\mathbb{E}\left[\tilde{\mathrm{f}}_X^K\right] - \mathrm{f}_X\right\|_{\infty,\,\mathcal{U}_n(x)} \leq C_{bias_X} h_X^{p'}.$$

2. *If the condition*

$$Cond_X(h_X): \quad h_X^{d_1} \geq \frac{4\|K_X\|_\infty^{2d_1}}{9\|K_X\|_2^{2d_1}\|\mathrm{f}_X\|_{\infty,\,\mathcal{U}_n'(x)}} \frac{(\log n)^{\frac{3}{2}}}{n_X}$$

*is satisfied, then for $\lambda_X := \sqrt{\frac{4\|K_X\|_2^{2d_1}\|\mathrm{f}_X\|_{\infty,\,\mathcal{U}_n'(x)}}{h_X^{d_1} n_X}(\log n)^{\frac{3}{2}}}$ and for any $u \in \mathcal{U}_n(x)$:*

$$\mathbb{P}\left(\left|\tilde{\mathrm{f}}_X^K(u) - \mathbb{E}\left[\tilde{\mathrm{f}}_X^K(u)\right]\right| > \lambda_X\right) \leq 2\exp\left(-(\log n)^{\frac{3}{2}}\right).$$

In the following, we denote $\widetilde{A}_n := \left\{ \sup_{u \in \mathcal{U}_n(x)} \left| \frac{\mathrm{f}_X(u) - \widetilde{\mathrm{f}}_X(u)}{\widetilde{\mathrm{f}}_X(u)} \right| \leq M_X \frac{(\log n_X)^{3/4}}{n_X^{1/2}} \right\}$ the event where the ratio $\frac{\mathrm{f}_X}{\widetilde{\mathrm{f}}_X}$ is close enough to 1.

**Lemma II.5** ($\bar{f}_h(w)$ behaviour). *For any bandwidth $h \in (0, h_0]^d$, and any $i = 1 : n$, let us denote $\bar{f}_{hi}(w) := \frac{K_h(w-W_i)}{\mathrm{f}_X(X_i)}$. Then, if $K$ is chosen as in section 3.c, under Assumptions II.1 to II.3,*

1. *Let $C_{\bar{E}} := \|f\|_{\infty,\,\mathcal{U}_n(w)}\|K\|_1^d$. Then*

$$\left|\mathbb{E}\left[\bar{f}_{h1}(w)\right]\right| \leq \mathbb{E}\left[\left|\bar{f}_{h1}(w)\right|\right] \leq C_{\bar{E}}.$$

*Besides, if we denote $\overline{B}_h := \mathbb{E}\left[\bar{f}_h(w)\right] - f(w)$ the bias of $\bar{f}_h(w) := \frac{1}{n}\sum_{i=1}^n \bar{f}_{hi}(w)$, then:*

$$|\overline{B}_h| \leq C_{\overline{bias}} \sum_{k \in \mathcal{R}} h_k^p,$$

*with $C_{\overline{bias}} := \frac{2|\int_{t\in\mathbb{R}} t^p K(t)dt|}{p!} \max_{k\in\mathcal{R}} |\partial_k^p f(w)|$.*

2. *Let $\overline{\mathcal{B}}_h := \{|\bar{f}_h(w) - \mathbb{E}[\bar{f}_h(w)]| \leq \sigma_h\}$, where $\sigma_h := C_\sigma \sqrt{\frac{(\log n)^a}{n\prod_{k=1}^d h_k}}$ with $C_\sigma = \frac{2\|K\|_2^d \|f\|_{\infty,\,\mathcal{U}_n(w)}^{\frac{1}{2}}}{\delta^{\frac{1}{2}}}$. If*

*Cond(h): $\prod_{k=1}^d h_k \geq \frac{4^2\|K\|_\infty^{2d}}{9\delta^2 C_\sigma^2}\frac{(\log n)^a}{n}$ is satisfied, then:*

$$\mathbb{P}\left(\overline{\mathcal{B}}_h^c\right) \leq 2e^{-(\log n)^a}$$

3. *Let* $\mathcal{B}_{|\bar{f}|h} := \left\{ \left| \frac{1}{n} \sum_{i=1}^{n} |\bar{f}_{hi}(w)| - \mathbb{E}[|\bar{f}_{h1}(w)|] \right| \leq C_{\bar{E}} \right\}$. *Then*

$$\mathbb{P}\left(\mathcal{B}^c_{|\bar{f}|h}\right) \leq 2e^{-C_{\gamma|f|} n \Pi_{k=1}^{d} h_k},$$

*with* $C_{\gamma|f|} := \min\left( \frac{C_{\bar{E}}^2}{C_\sigma^2}; \frac{3\delta C_{\bar{E}}}{4\|K\|_\infty^d} \right)$.

**Lemma II.6** ($\bar{Z}_{hj}$ behaviour). *For any* $j \in \{1,\dots,d\}$ *and any bandwidth* $h \in (0,h_0]^d$, *we define* $\bar{Z}_{hij} := \frac{1}{f_X(X_i)} \frac{\partial}{\partial h_j} \left( \prod_{k=1}^{d} h_k^{-1} K\left( \frac{w_k - W_{ik}}{h_k} \right) \right)$, *and* $\bar{Z}_{hj} := \frac{1}{n}\sum_{i=1}^{n} \bar{Z}_{hij}$. *If* $K$ *is chosen as in Section 3.c,*

1. *Under Assumptions II.1 to II.3, for* $j \notin \mathcal{R}$:

$$\mathbb{E}\left[\bar{Z}_{hj}\right] = 0.$$

*whereas, for* $j \in \mathcal{R}$, *for* $n$ *large enough,*

$$\frac{1}{2} C_{E\bar{Z},j} h_j^{p-1} \leq \left| \mathbb{E}\left[\bar{Z}_{hj}\right] \right| \leq \frac{3}{2} C_{E\bar{Z},j} h_j^{p-1}, \tag{II.12}$$

*where* $C_{E\bar{Z},j} := \left| \frac{\int_{\mathbb{R}} t^p K(t) dt}{(p-1)!} \partial_j^p f(w) \right|$.
*Besides, let* $C_{E|\bar{Z}|} := \|f\|_{\infty,\,\mathcal{U}_n(w)} \|J\|_1 \|K\|_1^{d-1}$. *Then :*

$$\mathbb{E}\left[|\bar{Z}_{h1j}|\right] \leq C_{E|\bar{Z}|} h_j^{-1}. \tag{II.13}$$

2. *Let* $\mathcal{B}_{\bar{Z},hj} := \{ |\bar{Z}_{hj} - \mathbb{E}\left[\bar{Z}_{hj}\right]| \leq \frac{1}{2}\lambda_{hj} \}$. *Under Assumptions II.1 to II.3, if the bandwidth satisfies:*

*Cond*$_{\bar{Z}}(h)$: $\prod_{k=1}^{d} h_k \geq cond_{\bar{Z}} \frac{(\log n)^a}{n}$, *with* $cond_{\bar{Z}} := \frac{4\|J\|_\infty^2 \|K\|_\infty^{2(d-1)}}{3^2 \|f\|_{\infty,\,\mathcal{U}_n(w)} \|J\|_2^2 \|K\|_2^{2(d-1)}}$,

*then:* $\mathbb{P}\left(\mathcal{B}^c_{\bar{Z},hj}\right) \leq 2e^{-\gamma_{Z,n}}$, *with* $\gamma_{Z,n} := \frac{\delta}{\|f\|_{\infty,\,\mathcal{U}_n(w)}} (\log n)^a$.

3. *Let* $\mathcal{B}_{|\bar{Z}|,hj} := \{ |\frac{1}{n}\sum_{i=1}^{n} |\bar{Z}_{hij}| - \mathbb{E}\left[|\bar{Z}_{h1j}|\right]| \leq C_{E|\bar{Z}|} h_j^{-1} \}$. *Then, under Assumptions II.1 to II.3:*

$$\mathbb{P}\left(\mathcal{B}^c_{|\bar{Z}|,hj}\right) \leq 2e^{-C_{\gamma|\bar{Z}|} n \Pi_{k=1}^{d} h_k},$$

*with* $C_{\gamma|\bar{Z}|} := \min\left( \frac{\delta C_{E|\bar{Z}|}^2}{4\|f\|_{\infty,\,\mathcal{U}_n(w)} \|J\|_2^2 \|K\|_2^{2(d-1)}}; \frac{3\delta C_{E|\bar{Z}|}}{4\|K\|_\infty^{d-1}\|J\|_\infty} \right)$.

**Lemma II.7.** *For any* $h \in (0,h_0]^d$ *and any component* $j = 1:d$, *we denote* $\Delta_{Z,hj} := Z_{hj} - \bar{Z}_{hj}$ *and* $\Delta_h := \hat{f}_h(w) - \bar{f}_h(w)$. *Under Assumptions II.1 to II.3, if the conditions on* $f_X$ *are satisfied (see section 3.b), then,*

1. *for* $C_{M\Delta Z} := \frac{2C_{E|\bar{Z}|}M_X}{C_\lambda}$:

$$\mathbb{1}_{\mathcal{B}_{|\bar{Z}|,hj} \cap \widetilde{A}_n} |\Delta_{Z,hj}| \leq \frac{C_{M\Delta Z}}{(\log n)^{\frac{a}{2}}} \lambda_{hj}$$

2. *for* $C_{M\Delta} := \frac{2C_{\bar{E}}M_X}{C_\sigma}$:

$$\mathbb{1}_{\widetilde{A}_n \cap \mathcal{B}_{|\bar{f}|h}} |\Delta_h| \leq \frac{C_{M\Delta}}{(\log n)^{\frac{a}{2}}} \sigma_h.$$

We introduce the notation $h^{(t)}$, $t \in \mathbb{N}$, the state of the bandwidth at iteration $t$ if $\hat{h} = h$. In particular for a fixed $t \in \{0, \dots, \lfloor \tau_n \rfloor\}$, $h^{(t)}$ is identical for any $h \in \mathcal{H}_{\text{hp}}$. Then we consider the event:

$$\mathcal{E}_Z := \widetilde{A}_n \cap \bigcap_{j \notin \mathcal{R}} \left\{ \mathcal{B}_{\bar{Z}, h^{(0)} j} \cap \mathcal{B}_{|\bar{Z}|, h^{(0)} j} \right\} \cap \bigcap_{j \in \mathcal{R}} \left[ \bigcap_{h \in \mathcal{H}_{\text{hp}}} \left\{ \mathcal{B}_{\bar{Z}, hj} \cap \mathcal{B}_{|\bar{Z}|, hj} \right\} \cap \bigcap_{t=0}^{\lfloor \tau_n \rfloor} \left\{ \mathcal{B}_{\bar{Z}, h^{(t)} j} \cap \mathcal{B}_{|\bar{Z}|, h^{(t)} j} \right\} \right].$$

**Proposition II.8** (CDRodeo behaviour). *Under Assumptions II.1 to II.3, on $\mathcal{E}_Z$, $\hat{h} \in \mathcal{H}_{hp}$.*
*In other words, when $\mathcal{E}_Z$ happens:*

1. *non relevant components are deactivated during the iteration $0$;*
2. *at the end of the iteration $\lfloor \tau_n \rfloor$, the active components are exactly the relevant ones;*
3. *CDRodeo stops at last at the iteration $\lfloor T_n \rfloor$.*

*Moreover, for any $q > 0$:*

$$\mathbb{P}(\mathcal{E}_Z^c) = o(n^{-q}).$$

The following lemma gives a technical result to canonically obtain an upper bound of the bias of a kernel estimator. Let us denote $\cdot$ the multiplication terms by terms of two vectors.

**Lemma II.9.** *Let $u \in \mathbb{R}^{d'}$ and $h \in \left(\mathbb{R}_+^*\right)^{d'}$ a bandwidth. For $j = 1 : d'$, let $K : \mathbb{R} \to \mathbb{R}$ be a continuous function with compact support and with at least $p-1$ zero moments, i.e.: for $l = 1 : (p-1)$,*

$$\int_{\mathbb{R}} K(t) t^l dt = 0.$$

*Let $f_0 : \mathbb{R}^{d'} \to \mathbb{R}$ a function of class $\mathcal{C}^p$ on the neighborhood of $u$:*

$$\mathcal{U}_h(u) := \left\{ u' \in \mathbb{R}^{d'} : \forall j = 1 : d', u'_j = u_j - h_j z_j, \text{with } z_j \in \text{supp}(K) \right\}.$$

*Then:*

$$\int_{\mathbb{R}^{d'}} \left( \prod_{j=1}^{d'} h_j^{-1} K\left( \tfrac{u_j - u'_j}{h_j} \right) \right) f_0(u') du' - f_0(u) \int_{\mathbb{R}^{d'}} \left( \prod_{j=1}^{d'} K\left(z_j\right) \right) dz = \sum_{k=1}^{d'} (I_k + II_k), \tag{II.14}$$

*where*

$$I_k := \int_{z \in \mathbb{R}^{d'}} \left( \prod_{j=1}^{d'} K(z_j) \right) \rho_k dz,$$

*with the notations $\rho_k := \rho_k(z,h,u) = (-h_k z_k)^p \int\limits_{0 \le t_p \le \dots \le t_1 \le 1} \left( \partial_k^p f_0(\bar{z}_{k-1} - t_p h_k z_k e_k) - \partial_k^p f_0(\bar{z}_{k-1}) \right) dt_{1:p}$,*

*and $\bar{z}_{k-1} := u - \sum_{j=1}^{k-1} h_j z_j e_j$ (where $\{e_j\}_{j=1}^{d'}$ is the canonical basis of $\mathbb{R}^{d'}$), and*

$$II_k := (-h_k)^p \int_{t \in \mathbb{R}} \frac{t^p}{p!} K(t) dt \int_{z_{-k} \in \mathbb{R}^{d'-1}} \partial_k^p f_0(\bar{z}_{k-1}) \left( \prod_{j \neq k} K(z_j) \right) dz_{-k}.$$

Finally, we recall (without proof) the classical Bernstein's Inequality and Taylor's theorem with integral remainder.

**Lemma II.10** (Bernstein's inequality). *Let $U_1, \dots, U_n$ be independent random variables almost surely uniformly bounded by a positive constant $c > 0$ and such that for $i = 1, \dots, n$, $\mathbb{E}[U_i^2] \leq v$. Then for any $\lambda > 0$,*

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} U_i - \mathbb{E}[U_i]\right| \geq \lambda\right) \leq 2\exp\left(-\min\left(\frac{n\lambda^2}{4v}, \frac{3n\lambda}{4c}\right)\right).$$

Note that this version is a simple consequence of Birgé and Massart (p.366 of [Birgé and Massart 1998]).

**Lemma II.11** (Taylor's theorem). *Let $g : [0,1] \to \mathbb{R}$ be a function of class $\mathcal{C}^q$. Then we have:*

$$g(1) - g(0) = \sum_{l=1}^{q} \frac{g^{(l)}(0)}{l!} + \int_{t_1=0}^{1} \int_{t_2=0}^{t_1} \dots \int_{t_q=0}^{t_{q-1}} (g^{(q)}(t_q) - g^{(q)}(0)) dt_q dt_{q-1} \dots dt_1.$$

## 5.c Proofs of Theorem II.2, Corollary II.3 and Proposition II.1,

### 5.c.i Proof of Theorem II.2

We introduce $\mathcal{E}_f := \bigcap_{h \in \mathcal{H}_{\mathsf{hp}}} \left(\overline{\mathcal{B}}_h \cap \mathcal{B}_{|\bar{f}|h}\right)$ and denote $\mathcal{E} := \mathcal{E}_Z \cap \mathcal{E}_f$. On $\mathcal{E}$, $\hat{h}$ belongs to $\mathcal{H}_{\mathsf{hp}}$ (cf Proposition II.8). Thus:

$$\mathbb{1}_{\mathcal{E}}\left(\hat{f}_{\hat{h}}(w) - f(w)\right) = \mathbb{1}_{\mathcal{E}} \sum_{h \in \mathcal{H}_{\mathsf{hp}}} \mathbb{1}_{\hat{h}=h}\left(\hat{f}_h(w) - f(w)\right). \tag{II.15}$$

For any $h \in \mathcal{H}_{\mathsf{hp}}$, we denote $\Delta_h := \hat{f}_h(w) - \bar{f}_h(w)$ and $\overline{B}_h := \mathbb{E}\left[\bar{f}_h(w)\right] - f(w)$, and we decompose the loss as follows:

$$\left|\hat{f}_h(w) - f(w)\right| \leq |\Delta_h| + \left|\bar{f}_h(w) - \mathbb{E}\left[\bar{f}_h(w)\right]\right| + \left|\overline{B}_h\right|. \tag{II.16}$$

Using Lemma II.7, since $\mathcal{E} \subset \widetilde{A}_n \cap \mathcal{B}_{|\bar{f}|h}$:

$$\mathbb{1}_{\mathcal{E}} |\Delta_h| \leq \frac{\mathsf{C}_{\mathsf{M}\Delta}}{(\log n)^{\frac{a}{2}}} \sigma_h. \tag{II.17}$$

Moreover, by Lemma II.5, since $\mathcal{E} \subset \widetilde{A}_n \cap \overline{\mathcal{B}}_h$:

$$\begin{aligned}
\left|\bar{f}_h(w) - \mathbb{E}\left[\bar{f}_h(w)\right]\right| &\leq \sigma_h && \text{(II.18)}\\
&= \mathsf{C}_\sigma \sqrt{\frac{(\log n)^a}{n \prod_{k=1}^{d} h_k}} \\
&\leq \mathsf{C}_\sigma \sqrt{\frac{(\log n)^a}{n h_0^d \beta^{r(T_n - \tau_n) + r\tau_n}}} \\
&= \mathsf{C}_\sigma \mathsf{C}_T^{\frac{-r}{2(2p+1)}} \mathsf{C}_\tau^{\frac{-r}{2(2p+r)}} (\log n)^{\frac{p(a+d-r)}{2p+r}} n^{-\frac{p}{2p+r}}. && \text{(II.19)}
\end{aligned}$$

And, also:

$$|\overline{B}_h| \leq \mathsf{C}_{\overline{\mathsf{bias}}} \sum_{k \in \mathcal{R}} h_k^p \leq r \mathsf{C}_{\overline{\mathsf{bias}}} \beta^{p\tau_n} h_0^p = r \mathsf{C}_{\overline{\mathsf{bias}}} \mathsf{C}_\tau^{\frac{p}{2p+r}} (\log n)^{\frac{p(a+d-r)}{2p+r}} n^{-\frac{p}{2p+r}}. \tag{II.20}$$

To conclude,

$$\begin{aligned}
\mathbb{1}_{\mathcal{E}}|\hat{f}_{\hat{h}}(w)-f(w)| &\leq \mathbb{1}_{\mathcal{E}}\sum_{h\in\mathcal{H}_{\mathsf{hp}}}\mathbb{1}_{\hat{h}=h}\left|\hat{f}_h(w)-f(w)\right|, \text{ by (II.15)}\\
&\leq \mathbb{1}_{\mathcal{E}}\sum_{h\in\mathcal{H}_{\mathsf{hp}}}\mathbb{1}_{\hat{h}=h}\left(|\Delta_h|+\left|\bar{f}_h(w)-\mathbb{E}\left[\bar{f}_h(w)\right]\right|+\left|\overline{B}_h\right|\right), \text{ by (II.16)}\\
&\leq \mathbb{1}_{\mathcal{E}}\sum_{h\in\mathcal{H}_{\mathsf{hp}}}\mathbb{1}_{\hat{h}=h}\left[\left(1+\frac{\mathsf{C}_{\mathsf{M}\Delta}}{(\log n)^{\frac{a}{2}}}\right)\sigma_h+\left|\overline{B}_h\right|\right], \text{ by (II.17) and (II.18)}\\
&\leq \mathbb{1}_{\mathcal{E}}\sum_{h\in\mathcal{H}_{\mathsf{hp}}}\mathbb{1}_{\hat{h}=h}\mathsf{C}(\log n)^{\frac{p(a+d-r)}{2p+r}}n^{-\frac{p}{2p+r}}, \text{ by (II.19) and (II.20)}\\
&= \mathbb{1}_{\mathcal{E}}\mathsf{C}(\log n)^{\frac{p(a+d-r)}{2p+r}}n^{-\frac{p}{2p+r}},
\end{aligned}$$

with for $n$ large enough (*ie*: $\frac{\mathsf{C}_{\mathsf{M}\Delta}}{(\log n)^{\frac{a}{2}}}\leq 1$),

$$\mathsf{C} := r\mathsf{C}_{\overline{\mathsf{bias}}}\mathsf{C}_\tau^{\frac{p}{2p+r}}+2\mathsf{C}_\sigma\mathsf{C}_T^{\frac{-r}{2(2p+1)}}\mathsf{C}_\tau^{\frac{-r}{2(2p+r)}}.$$

It remains to give an upper bound on $\mathbb{P}(\mathcal{E}^c)$. For any $q>0$:

$$\begin{aligned}
\mathbb{P}(\mathcal{E}^c) &\leq \mathbb{P}(\mathcal{E}_Z^c)+\mathbb{P}\left(\mathcal{E}_f^c\right)\\
&\leq o(n^{-q})+\sum_{h\in\mathcal{H}_{\mathsf{hp}}}\left(\mathbb{P}\left(\overline{\mathcal{B}}_h^c\right)+\mathbb{P}\left(\mathcal{B}_{|\bar{f}|h}^c\right)\right), \text{ using Proposition II.8}\\
&\leq o(n^{-q})+\sum_{h\in\mathcal{H}_{\mathsf{hp}}}\left(2e^{-(\log n)^a}+2e^{-\mathsf{C}_{\gamma|f|}n\prod_{k=1}^d h_k}\right),
\end{aligned}$$

using Lemma II.5, since for any $h\in\mathcal{H}_{\mathsf{hp}}$, Cond($h$) is satisfied. Moreover:

$$n\prod_{k=1}^d h_k \geq n\beta^{rT_n}h_0^d = \mathsf{C}_T^{\frac{r}{2p+1}}\mathsf{C}_\tau^{\frac{r}{2p+r}}(\log n)^{\frac{ra-2p(d-r)}{(2p+r)}}n^{\frac{2p}{2p+r}} \geq \frac{(\log n)^a}{\mathsf{C}_{\gamma|f|}},$$

for $n$ large enough. Hence:

$$\mathbb{P}(\mathcal{E}^c)\leq o(n^{-q})+|\mathcal{H}_{\mathsf{hp}}|4e^{-(\log n)^a}=o(n^{-q}),$$

for $n$ large enough, since $|\mathcal{H}_{\mathsf{hp}}|=(\lceil T_n\rceil-\lfloor\tau_n\rfloor)^r=\left(\frac{1}{(2p+1)(\log(\frac{1}{\beta}))}\log(\frac{C_T}{C_\tau})+1\right)^r$ is finite.

### 5.c.ii Proof of Corollary II.3

We consider the event $\mathcal{E}=\left\{\left|\hat{f}_{\hat{h}}(w)-f(w)\right|\leq \mathsf{C}(\log n)^{\frac{p}{2p+r}(d-r+a)}n^{-\frac{p}{2p+r}}\right\}$ for which we proved in Theorem II.2:

$$\mathbb{P}(\mathcal{E}^c)=o(n^{-A}),$$

for any $A>0$. For short, we denote $R_h:=\left|\hat{f}_h(w)-f(w)\right|$ for any bandwidth $h\in(\mathbb{R}_+^*)^d$. Then we decompose $R_{\hat{h}}$ as follows:

$$R_{\hat{h}}=\mathbb{1}_{\mathcal{E}}R_{\hat{h}}+\mathbb{1}_{\mathcal{E}^c}R_{\hat{h}}.$$

By definition of $\mathcal{E}$, we immediately obtain:

$$\mathbb{1}_{\mathcal{E}} R_{\hat{h}} \leq \mathsf{C} (\log n)^{\frac{p}{2p+r}(d-r+a)} n^{-\frac{p}{2p+r}} \tag{II.21}$$

For the second term, we first bound $\hat{f}_{\hat{h}}(w)$ a.s. In CDRODEO procedure, the loop stops when the current bandwidth becomes too small: $\prod_{k=1}^{d} h_k < \frac{\log n}{n}$. So the final bandwidth $\hat{h}$ satisfies:

$$\prod_{k=1}^{d} \hat{h}_k \geq \frac{\beta^d \log n}{n}.$$

Since $\hat{f}_{\hat{h}}(w) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\widetilde{\mathrm{f}}_X(X_i)} \left( \prod_{k=1}^{d} \hat{h}_k^{-1} K\left( \frac{w_k - W_{ik}}{\hat{h}_k} \right) \right)$ and using the above lower bound and the lower bound of $\widetilde{\mathrm{f}}_X$, we obtain:

$$\left| \hat{f}_{\hat{h}}(w) \right| \leq \frac{\|K\|_\infty^d}{\beta^d \widetilde{\delta}_X} \frac{n}{\log n}.$$

Hence, using Condition (i):

$$R_{\hat{h}} \leq f(w) + \frac{\|K\|_\infty^d}{\beta^d n^{-\chi}} \frac{n}{(\log n)}$$

Therefore, for any $q > 0$, using $(a+b)^q \leq 2^{q-1}(a^q + b^q)$:

$$\begin{aligned} \mathbb{E}\left[ \mathbb{1}_{\mathcal{E}^c} (R_{\hat{h}})^q \right] &\leq P(\mathcal{E}^c) 2^{q-1} \left( f(w)^q + \left( \frac{\|K\|_\infty}{\beta^d} \right)^q \frac{n^{q(1+\chi)}}{(\log n)^q} \right) \\ &= o\left( n^{-A'} \right), \end{aligned} \tag{II.22}$$

for any $A' > 0$ (since $P(\mathcal{E}^c) = o(n^{-A' + q(1+\chi)})$).
We conclude by combining (II.21) and (II.22):

$$\left( \mathbb{E}\left[ \left| \hat{f}_{\hat{h}}(w) - f(w) \right|^q \right] \right)^{1/q} \leq \mathsf{C} (\log n)^{\frac{p}{2p+r}(d-r+a)} n^{-\frac{p}{2p+r}} + o\left( n^{-1} \right). \tag{II.23}$$

### 5.c.iii Proof of Proposition II.1

We construct $\widetilde{\mathrm{f}}_X$ in two steps: we first construct an estimator $\tilde{\mathrm{f}}_X^K$ which satisfies

$$\mathbb{P}\left( \|\mathrm{f}_X - \tilde{\mathrm{f}}_X^K\|_{\infty, \mathcal{U}_n(x)} > M_X \frac{(\log n)^{\frac{3}{4}}}{n^{\frac{1}{2}}} \right) \leq \exp(-(\log n)^{\frac{5}{4}}), \tag{II.24}$$

then we show that if we set $\widetilde{\mathrm{f}}_X \equiv \tilde{\mathrm{f}}_X^K \vee (\log n)^{-\frac{1}{4}}$, $\widetilde{\mathrm{f}}_X$ satisfies Conditions (i) and (ii) for $n$ large enough.

We take $\tilde{\mathrm{f}}_X^K$ as the kernel density estimator defined in Equation (II.10), with a kernel $K_X : \mathbb{R} \to \mathbb{R}$ that is compactly supported, of class $\mathcal{C}^1$, of order $p_X \geq \frac{d_1}{2(c-1)}$. and a bandwidth $h_X \in \mathbb{R}_+^*$ specified later. Let us control the bias $\left\| \mathbb{E}\left[ \tilde{\mathrm{f}}_X^K \right] - \mathrm{f}_X \right\|_{\infty, \mathcal{U}_n(x)}$. We define $p'_X = \min(p', p_X + 1)$. In particular, $\mathrm{f}_X$

is of class $\mathcal{C}^{p'_X}$ and $K_X$ has $p'_X - 1$ zero moments.
Therefore we can apply Lemma II.4 :

$$\left\|\mathbb{E}\left[\tilde{\mathrm{f}}_X^K\right] - \mathrm{f}_X\right\|_{\infty,\,\mathcal{U}_n(x)} \leq \mathsf{C}'_{\mathsf{bias}_X} h_X^{p'_X},$$

where $\mathsf{C}'_{\mathsf{bias}_X} := \frac{\|K_X\|_1^{d_1-1}\|.^{p'_X} K_X(\cdot)\|_1}{p'_X!} d_1 \max_{k=1:d_1} \|\partial_k^{p'_X} \mathrm{f}_X\|_{\infty,\,\mathcal{U}'_n(x)}$.
Therefore, since

$$\begin{aligned}\left\|\tilde{\mathrm{f}}_X^K - \mathrm{f}_X\right\|_{\infty,\,\mathcal{U}_n(x)} &\leq \left\|\tilde{\mathrm{f}}_X^K - \mathbb{E}\left[\tilde{\mathrm{f}}_X^K\right]\right\|_{\infty,\,\mathcal{U}_n(x)} + \left\|\mathbb{E}\left[\tilde{\mathrm{f}}_X^K\right] - \mathrm{f}_X\right\|_{\infty,\,\mathcal{U}_n(x)} \\ &\leq \left\|\tilde{\mathrm{f}}_X^K - \mathbb{E}\left[\tilde{\mathrm{f}}_X^K\right]\right\|_{\infty,\,\mathcal{U}_n(x)} + \mathsf{C}'_{\mathsf{bias}_X} h_X^{p'_X},\end{aligned}$$

we have for any threshold $\lambda$:

$$\mathbb{P}\left(\left\|\tilde{\mathrm{f}}_X^K - \mathrm{f}_X\right\|_{\infty,\,\mathcal{U}_n(x)} \geq \lambda\right) \leq \mathbb{P}\left(\left\|\tilde{\mathrm{f}}_X^K - \mathbb{E}\left[\tilde{\mathrm{f}}_X^K\right]\right\|_{\infty,\,\mathcal{U}_n(x)} \geq \lambda - \mathsf{C}'_{\mathsf{bias}_X} h_X^{p'_X}\right). \tag{II.25}$$

Therefore, we have reduced the problem to a local concentration inequality of $\tilde{\mathrm{f}}_X^K$ in sup norm. In order to move from a supremum on $\mathcal{U}_n(x)$ to a maximum on a finite set of elements of $\mathcal{U}_n(x)$, let us construct a $\varepsilon$-net of $\mathcal{U}_n(x)$. We denote $A > 0$ such that:

$$\mathsf{supp}(K_X) \cup \mathsf{supp}(K) \subset \left[-\tfrac{A}{2}, \tfrac{A}{2}\right].$$

We set $N(\varepsilon)$ the smallest integer such that $\varepsilon N(\varepsilon) \geq \frac{A}{\log n}$, *i.e.*:

$$N(\varepsilon) := \left\lceil \frac{A}{\varepsilon \log n} \right\rceil,$$

then we introduce the notation $u_{(l)} \in \mathcal{U}_n(x)$, for a multi-index $l \in (1 : N(\varepsilon))^{d_1}$ defined, such that the $j^{\text{th}}$ component of $u_{(l)}$ is:

$$u_{(l)j} := x_j - \frac{A}{2\log n} + (2l_j - 1)\frac{\varepsilon}{2}.$$

Then $\{u_{(l)} : l \in (1 : N(\varepsilon))^{d_1}\}$ is a $\varepsilon$-net of $\mathcal{U}_n(x)$, in the meaning that for any $u \in \mathcal{U}_n(x)$, there exists $l \in \{1, \ldots, N(\varepsilon)\}^{d_1}$ such that $\|u - u_{(l)}\|_\infty := \max_{k=1:d_1} |u_k - u_{(l)k}| \leq \varepsilon$.
Therefore to obtain the desired concentration inequality, we only need to obtain the concentration inequality for each point of $\{u_{(l)} : l \in (1 : N(\varepsilon))^{d_1}\}$ and to control the following supremum

$$\sup_{u \in\, \mathcal{U}_n(x)} \min_{l \in (1:N(\varepsilon))^{d_1}} \left|\tilde{\mathrm{f}}_X^K(u) - \mathbb{E}\left[\tilde{\mathrm{f}}_X^K(u)\right] - \tilde{\mathrm{f}}_X^K(u_{(l)}) + \mathbb{E}\left[\tilde{\mathrm{f}}_X^K(u_{(l)})\right]\right|.$$

For this purpose, we obtain (from Taylor's Inequality): for any $u, v \in \mathbb{R}^{d_1}$,

$$\left|\prod_{k=1}^{d_1} K_X(u_k) - \prod_{k=1}^{d_1} K_X(v_k)\right| \leq d_1 \|K'_X\|_\infty \|K_X\|_\infty^{d_1-1} \|u - v\|_\infty.$$

Therefore, for any $u, v \in \mathcal{U}_n(x)$:

$$\begin{aligned}\left|\tilde{\mathrm{f}}_X^K(u) - \tilde{\mathrm{f}}_X^K(v)\right| &\leq \frac{1}{n_X . h_X^{d_1}} \sum_{i=1}^{n_X} \left|\prod_{k=1}^{d_1} K_X\left(\tfrac{u_k - \widetilde{X}_{ik}}{h_X}\right) - \prod_{k=1}^{d_1} K_X\left(\tfrac{v_k - \widetilde{X}_{ik}}{h_X}\right)\right| \\ &\leq \frac{d_1}{h_X^{d_1+1}} \|K'_X\|_\infty \|K_X\|_\infty^{d_1-1} \|u - v\|_\infty.\end{aligned}$$

Since $\{u_{(l)} : l \in (1 : N(\varepsilon))^{d_1}\}$ is a $\varepsilon$-net of $\mathcal{U}_n(x)$:

$$\sup_{u \in \mathcal{U}_n(x)} \min_{l \in (1:N(\varepsilon))^{d_1}} \left|\tilde{\mathrm{f}}_X^K(u) - \tilde{\mathrm{f}}_X^K(u_{(l)})\right| \leq \frac{d_1}{h_X^{d_1+1}} \|K_X'\|_\infty \|K_X\|_\infty^{d_1-1} \varepsilon.$$

Thus:

$$\sup_{u \in \mathcal{U}_n(x)} \min_{l \in (1:N(\varepsilon))^{d_1}} \left|\mathbb{E}\left[\tilde{\mathrm{f}}_X^K(u)\right] - \mathbb{E}\left[\tilde{\mathrm{f}}_X^K(u_{(l)})\right]\right| \leq d_1 \|K_X'\|_\infty \|K_X\|_\infty^{d_1-1} \frac{\varepsilon}{h_X^{d_1+1}}.$$

And so:

$$\sup_{u \in \mathcal{U}_n(x)} \min_{l \in (1:N(\varepsilon))^{d_1}} \left|\tilde{\mathrm{f}}_X^K(u) - \mathbb{E}\left[\tilde{\mathrm{f}}_X^K(u)\right] - \tilde{\mathrm{f}}_X^K(u_{(l)}) + \mathbb{E}\left[\tilde{\mathrm{f}}_X^K(u_{(l)})\right]\right| \leq 2d_1 \|K_X'\|_\infty \|K_X\|_\infty^{d_1-1} \frac{\varepsilon}{h_X^{d_1+1}}.$$

We denote $\mathsf{C}_{\mathsf{diff}} := 2d_1 \|K_X'\|_\infty \|K_X\|_\infty^{d_1-1}$. Then:

$$\begin{aligned}
\left\|\tilde{\mathrm{f}}_X^K - \mathbb{E}\left[\tilde{\mathrm{f}}_X^K\right]\right\|_{\infty,\, \mathcal{U}_n(x)} &\leq \max_{l \in (1:N(\varepsilon))^{d_1}} \left|\tilde{\mathrm{f}}_X^K(u_{(l)}) - \mathbb{E}\left[\tilde{\mathrm{f}}_X^K(u_{(l)})\right]\right| \\
&\quad + \sup_{u \in \mathcal{U}_n(x)} \min_{l \in (1:N(\varepsilon))^{d_1}} \left|\tilde{\mathrm{f}}_X^K(u) - \mathbb{E}\left[\tilde{\mathrm{f}}_X^K(u)\right] - \tilde{\mathrm{f}}_X^K(u_{(l)}) + \mathbb{E}\left[\tilde{\mathrm{f}}_X^K(u_{(l)})\right]\right| \\
&\leq \max_{l \in (1:N(\varepsilon))^{d_1}} \left|\tilde{\mathrm{f}}_X^K(u_{(l)}) - \mathbb{E}\left[\tilde{\mathrm{f}}_X^K(u_{(l)})\right]\right| + \mathsf{C}_{\mathsf{diff}} \frac{\varepsilon}{h_X^{d_1+1}}.
\end{aligned}$$

Then the inequality (II.25) becomes: for any threshold $\lambda$,

$$\begin{aligned}
\mathbb{P}\left(\left\|\tilde{\mathrm{f}}_X^K - \mathrm{f}_X\right\|_{\infty,\, \mathcal{U}_n(x)} \geq \lambda\right) &\leq \mathbb{P}\left(\left\|\tilde{\mathrm{f}}_X^K - \mathbb{E}\left[\tilde{\mathrm{f}}_X^K\right]\right\|_{\infty,\, \mathcal{U}_n(x)} \geq \lambda - \mathsf{C}'_{\mathsf{bias}_X} h_X^{p'_X}\right) \\
&\leq \mathbb{P}\left(\max_{l \in (1:N(\varepsilon))^{d_1}} \left|\tilde{\mathrm{f}}_X^K(u_{(l)}) - \mathbb{E}\left[\tilde{\mathrm{f}}_X^K(u_{(l)})\right]\right| \geq \lambda - \mathsf{C}'_{\mathsf{bias}_X} h_X^{p'_X} - \mathsf{C}_{\mathsf{diff}} \frac{\varepsilon}{h_X^{d_1+1}}\right) \\
&\leq N(\varepsilon)^{d_1} \max_{l \in (1:N(\varepsilon))^{d_1}} \mathbb{P}\left(\left|\tilde{\mathrm{f}}_X^K(u_{(l)}) - \mathbb{E}\left[\tilde{\mathrm{f}}_X^K(u_{(l)})\right]\right| \geq \lambda - \mathsf{C}'_{\mathsf{bias}_X} h_X^{p'_X} - \mathsf{C}_{\mathsf{diff}} \frac{\varepsilon}{h_X^{d_1+1}}\right)
\end{aligned} \tag{II.26}$$

We want to apply 2. of Lemma II.4. Therefore we fix the following settings:

- $h_X := n_X^{-\frac{c-1}{c.d_1}}$
- $\lambda := 2\lambda_X$, where $\lambda_X$ is the threshold in 2. of Lemma II.4;
- $\varepsilon := h_X^{1+\frac{d_1}{2}} n_X^{-\frac{1}{2}}$.

For short, we denote $\mathsf{C}_{\lambda X} := 2\|K_X\|_2^{d_1} \|\mathrm{f}_X\|_{\infty,\, \mathcal{U}_n'(x)}^{\frac{1}{2}}$, so:

$$\lambda_X = \sqrt{\frac{4\|K_X\|_2^{2d_1} \|\mathrm{f}_X\|_{\infty,\, \mathcal{U}_n'(x)}}{h_X^{d_1} n_X} (\log n)^{\frac{3}{2}}} = \mathsf{C}_{\lambda X} (\log n)^{\frac{3}{4}} h_X^{-\frac{d_1}{2}} n_X^{-\frac{1}{2}} = \mathsf{C}_{\lambda X} (\log n)^{\frac{3}{4}} n_X^{-\frac{1}{2c}}.$$

In particular, since we take $p_X \geq \frac{d_1}{2(c-1)}$ and we assume $p' \geq \frac{d_1}{2(c-1)}$, then $p'_X = \min(p', p_X) \geq \frac{d_1}{2(c-1)}$. Hence we obtain for $n$ large enough:

$$\begin{aligned}
\mathsf{C}'_{\mathsf{bias}_X} h_X^{p'_X} &= \mathsf{C}'_{\mathsf{bias}_X} n_X^{-\frac{p'_X(c-1)}{c.d_1}} \\
&\leq \mathsf{C}'_{\mathsf{bias}_X} n_X^{-\frac{1}{2c}} \\
&\leq \frac{1}{2}\lambda_X = \frac{\mathsf{C}_{\lambda X}}{2} (\log n)^{\frac{3}{4}} n_X^{-\frac{1}{2c}}.
\end{aligned}$$

and also, since $c > 1$:

$$\begin{aligned}\mathsf{C}_{\mathsf{diff}}\frac{\varepsilon}{h_X^{d_1+1}} &= \mathsf{C}_{\mathsf{diff}} h_X^{-\frac{d_1}{2}} n_X^{-\frac{1}{2}} = \mathsf{C}_{\mathsf{diff}}\, n_X^{-\frac{1}{2c}} \\ &\leq \frac{1}{2}\lambda_X = \frac{\mathsf{C}_{\lambda X}}{2}(\log n)^{\frac{3}{4}} n_X^{-\frac{1}{2c}}.\end{aligned}$$

Thus:

$$\lambda - \mathsf{C}'_{\mathsf{bias}_X} h_X^{p'_X} - \mathsf{C}_{\mathsf{diff}}\frac{\varepsilon}{h_X^{d_1+1}} \geq \lambda_X,$$

and the inequality (II.26) becomes:

$$\mathbb{P}\left(\left\|\tilde{\mathrm{f}}_X^K - \mathrm{f}_X\right\|_{\infty,\,\mathcal{U}_n(x)} \geq \lambda\right) \leq N(\varepsilon)^{d_1} \max_{l\in(1:N(\varepsilon))^{d_1}} \mathbb{P}\left(\left|\tilde{\mathrm{f}}_X^K(u_{(l)}) - \mathbb{E}\left[\tilde{\mathrm{f}}_X^K(u_{(l)})\right]\right| \geq \lambda_X\right) \tag{II.27}$$

We verify that $\mathrm{Cond}_X(h_X)$ is satisfied for $n$ large enough:

$$\begin{aligned}h_X^{d_1} &= n_X^{-\frac{c-1}{c}} \\ &\geq \frac{4\|K_X\|_\infty^{2d_1}}{9\|K_X\|_2^{2d_1}\|\mathrm{f}_X\|_{\infty,\,\mathcal{U}'_n(x)}}\frac{(\log n)^{\frac{3}{2}}}{n_X}.\end{aligned}$$

Then we can apply 2. of Lemma II.4,

$$\mathbb{P}\left(\left|\tilde{\mathrm{f}}_X^K(u_{(l)}) - \mathbb{E}\left[\tilde{\mathrm{f}}_X^K(u_{(l)})\right]\right| > \lambda_X\right) \leq 2\exp\left(-(\log n)^{\frac{3}{2}}\right).$$

Thus the inequality (II.27) becomes:

$$\mathbb{P}\left(\left\|\tilde{\mathrm{f}}_X^K - \mathrm{f}_X\right\|_{\infty,\,\mathcal{U}_n(x)} \geq \lambda\right) \leq 2N(\varepsilon)^{d_1}\exp\left(-(\log n)^{\frac{3}{2}}\right). \tag{II.28}$$

Let us control $2N(\varepsilon)^{d_1}$:

$$\begin{aligned}2N(\varepsilon)^{d_1} &= 2\left\lceil\frac{A}{\varepsilon\log n}\right\rceil^{d_1} \\ &= 2\left\lceil\frac{A}{h_X^{1+\frac{d_1}{2}} n_X^{-\frac{1}{2}}\log n}\right\rceil^{d_1} \\ &= o\left(n_X^{d_1+1}\right).\end{aligned}$$

Then for $n$ large enough:

$$2N(\varepsilon)^{d_1}\exp\left(-(\log n)^{\frac{3}{2}}\right) \leq \exp\left(-(\log n)^{\frac{5}{4}}\right).$$

Therefore:

$$\mathbb{P}\left(\left\|\tilde{\mathrm{f}}_X^K - \mathrm{f}_X\right\|_{\infty,\,\mathcal{U}_n(x)} \geq \lambda\right) \leq \exp\left(-(\log n)^{\frac{5}{4}}\right).$$

Since $\lambda = 2\mathsf{C}_{\lambda X}(\log n)^{\frac{3}{4}} n^{-\frac{1}{2}}$, we have obtained the desired concentration inequality (II.24) with $M_X = 2\mathsf{C}_{\lambda X}$.

Now we consider $\widetilde{\mathrm{f}}_X \equiv \tilde{\mathrm{f}}_X^K \vee (\log n)^{-\frac{1}{4}}$. By construction, $\widetilde{\mathrm{f}}_X$ satisfies Condition (i). Let us show it also satisfies Condition (ii), *i.e.*:

$$\mathbb{P}\left(\sup_{u\in\,\mathcal{U}_n(x)}\left|\frac{\mathrm{f}_X(u)-\widetilde{\mathrm{f}}_X(u)}{\widetilde{\mathrm{f}}_X(u)}\right| > M_X\frac{(\log n)^{\frac{d}{2}}}{n^{\frac{1}{2}}}\right) \leq C_X\exp(-(\log n)^{\frac{5}{4}}).$$

We write:

$$\begin{aligned}\mathbb{P}\left(\sup_{u\in\,\mathcal{U}_n(x)}\left|\frac{\mathrm{f}_X(u)-\widetilde{\mathrm{f}}_X(u)}{\widetilde{\mathrm{f}}_X(u)}\right| > M_X\frac{(\log n)^{\frac{d}{2}}}{n^{\frac{1}{2}}}\right) &= \mathbb{P}\left(\exists u\in\,\mathcal{U}_n(x), \left|\mathrm{f}_X(u)-\widetilde{\mathrm{f}}_X(u)\right| > \widetilde{\mathrm{f}}_X(u)M_X\frac{(\log n)^{\frac{d}{2}}}{n^{\frac{1}{2}}}\right)\\ &\leq \mathbb{P}\left(\exists u\in\,\mathcal{U}_n(x), \left|\mathrm{f}_X(u)-\widetilde{\mathrm{f}}_X(u)\right| > (\log n)^{-\frac{1}{4}}M_X\frac{(\log n)^{\frac{d}{2}}}{n^{\frac{1}{2}}}\right)\\ &\leq \mathbb{P}\left(\left\|\mathrm{f}_X(u)-\widetilde{\mathrm{f}}_X(u)\right\|_{\infty,\,\mathcal{U}_n(x)} > M_X\frac{(\log n)^{\frac{d}{2}-\frac{1}{4}}}{n^{\frac{1}{2}}}\right).\end{aligned}$$

Since $d = d_1 + d_2 \geq 2$, $\frac{d}{2} - \frac{1}{4} \geq \frac{3}{4}$, we obtain from the previously proved concentration inequality (II.24):

$$\begin{aligned}\mathbb{P}\left(\sup_{u\in\,\mathcal{U}_n(x)}\left|\frac{\mathrm{f}_X(u)-\widetilde{\mathrm{f}}_X(u)}{\widetilde{\mathrm{f}}_X(u)}\right| > M_X\frac{(\log n)^{\frac{d}{2}}}{n^{\frac{1}{2}}}\right) &\leq \mathbb{P}\left(\left\|\tilde{\mathrm{f}}_X^K-\mathrm{f}_X\right\|_{\infty,\,\mathcal{U}_n(x)} \geq M_X\frac{(\log n)^{\frac{3}{4}}}{n^{\frac{1}{2}}}\right)\\ &\leq \exp\left(-(\log n)^{\frac{5}{4}}\right).\end{aligned}$$

## 5.d Proof of Proposition II.8 and the lemmas

### 5.d.i Proof of Proposition II.8

First, note that the final state of the bandwidth determines exactly at which iteration each component has been deactivated: for a fixed bandwidth $h \in (\mathbb{R}_+^*)^d$, if $\hat{h} = h$, we denote $\{\theta_k\}_{k=1}^d$ such as for $k = 1:d$, $h_k = h_0\beta^{\theta_k}$. In particular, $\theta_k$ is the iteration of deactivation of the component $k$.

We introduce the notation $h^{(t)}$, $t \in \mathbb{N}$, the state of the bandwidth at iteration $t$ if $\hat{h} = h$. It implies that $h^{(t)}$ is exactly defined by: $h_k^{(t)} = \beta^{\theta_k \wedge t} h_0$ for $k = 1:d$.

Notice that for a fixed $t \in \{0,\dots,\lfloor\tau_n\rfloor\}$, $h^{(t)}$ is identical for any $h \in \mathcal{H}_{\mathsf{hp}}$: by definition of $\mathcal{H}_{\mathsf{hp}}$, $h_j^{(t)} = h_0\beta^t$ if $j \in \mathcal{R}$, else $h_j^{(t)} = h_0$.

We recall the definition

$$\mathcal{E}_Z := \widetilde{A}_n \cap \bigcap_{j\notin\mathcal{R}}\left\{\mathcal{B}_{\bar{Z},h^{(0)}j}\cap\mathcal{B}_{|\bar{Z}|,h^{(0)}j}\right\}\cap\bigcap_{j\in\mathcal{R}}\left[\bigcap_{h\in\mathcal{H}_{\mathsf{hp}}}\left\{\mathcal{B}_{\bar{Z},hj}\cap\mathcal{B}_{|\bar{Z}|,hj}\right\}\cap\bigcap_{t=0}^{\lfloor\tau_n\rfloor}\left\{\mathcal{B}_{\bar{Z},h^{(t)}j}\cap\mathcal{B}_{|\bar{Z}|,h^{(t)}j}\right\}\right].$$

For any component $j$ and any bandwidth $h$, we decompose $Z_{hj}$ as follows:

$$\begin{aligned}\mathbb{1}_{\mathcal{E}_Z}Z_{hj} &= \mathbb{1}_{\mathcal{E}_Z}\bar{Z}_{hj} + \mathbb{1}_{\mathcal{E}_Z}\Delta_{Z,hj}\\ &= \mathbb{1}_{\mathcal{E}_Z}\mathbb{E}\left[\bar{Z}_{hj}\right] + \mathbb{1}_{\mathcal{E}_Z}(\bar{Z}_{hj} - \mathbb{E}\left[\bar{Z}_{hj}\right]) + \mathbb{1}_{\mathcal{E}_Z}\Delta_{Z,hj}.\end{aligned} \tag{II.29}$$

1. Let us fix $j \notin \mathcal{R}$ and $h = h^{(0)} = (h_0, \dots, h_0)$. Using 1. of Lemma II.6, $\mathbb{E}\left[\bar{Z}_{hj}\right] = 0$. Therefore:

$$\begin{aligned}\mathbb{1}_{\mathcal{E}_Z}|Z_{hj}| &\leq \mathbb{1}_{\mathcal{E}_Z}\left|\bar{Z}_{hj} - \mathbb{E}\left[\bar{Z}_{hj}\right]\right| + \mathbb{1}_{\mathcal{E}_Z}\left|\Delta_{Z,hj}\right| \\ &\leq \frac{1}{2}\lambda_{hj} + \mathbb{1}_{\mathcal{E}_Z}\left|\Delta_{Z,hj}\right|,\end{aligned}$$

using 2. of Lemma II.6, since $\mathcal{E}_Z \subset \mathcal{B}_{\bar{Z},hj}$. Now using 1. of Lemma II.7, since $\mathcal{E}_Z \subset \mathcal{B}_{|\bar{Z}|,hj} \cap \widetilde{A}_n$, we obtain:

$$\mathbb{1}_{\mathcal{E}_Z}|Z_{hj}| \leq \frac{1}{2}\lambda_{hj} + \frac{\mathsf{C}_{\mathsf{M}\Delta Z}}{(\log n)^{\frac{a}{2}}}\lambda_{hj}.$$

Then for $n$ large enough (*ie*: $(\log n)^{\frac{a}{2}} > 2\mathsf{C}_{\mathsf{M}\Delta Z}$), $\mathbb{1}_{\mathcal{E}_Z}|Z_{hj}| < \lambda_{hj}$. In other words, when $\mathcal{E}_Z$ happens, all irrelevant components deactivate at the iteration $0$.

2. Let us show that $\mathcal{E}_Z$ implies that the relevant components remain active until iteration $\lfloor \tau_n \rfloor + 1$.
It suffices to prove $|Z_{h^{(t)}j}| > \lambda_{h^{(t)}j}$, for any $j \in \mathcal{R}$ and any bandwidth $h^{(t)}, t = 0 : \lfloor \tau_n \rfloor$. (Indeed, by induction: $(h_0, \dots, h_0) = h^{(0)}$, and since the irrelevant components deactivate at the iteration 0, if the current bandwidth at the iteration $t$ is $h^{(t)}$ , then the fact that all the relevant components remain active for this bandwidth implies that the bandwidth at iteration $t+1$ is $h^{(t+1)}$).
Let us fix $j \in \mathcal{R}$, $t = 0 : \lfloor \tau_n \rfloor$ and we denote $h = h^{(t)}$. Using the decomposition (II.29), we obtain the following lower bound:

$$\mathbb{1}_{\mathcal{E}_Z}|Z_{hj}| \geq \mathbb{1}_{\mathcal{E}_Z}\left(\left|\mathbb{E}\left[\bar{Z}_{hj}\right]\right| - \left|\bar{Z}_{hj} - \mathbb{E}\left[\bar{Z}_{hj}\right]\right| - \left|\Delta_{Z,hj}\right|\right).$$

Then, combining:

- $\left|\mathbb{E}\left[\bar{Z}_{hj}\right]\right| \geq \frac{\mathsf{C}_{E\bar{Z},j}}{2} h_j^{p-1}$ (cf 1. of Lemma II.6),
- $\left|\bar{Z}_{hj} - \mathbb{E}\left[\bar{Z}_{hj}\right]\right| \leq \frac{1}{2}\lambda_{hj}$, since $\mathcal{E}_Z \subset \mathcal{B}_{\bar{Z},hj}$ (cf 2. of Lemma II.6),
- $\left|\Delta_{Z,hj}\right| \leq \frac{\mathsf{C}_{\mathsf{M}\Delta Z}}{(\log n)^{\frac{a}{2}}}\lambda_{hj}$, since $\mathcal{E}_Z \subset \mathcal{B}_{|\bar{Z}|,hj} \cap \widetilde{A}_n$ (cf 1. of Lemma II.7),

we obtain:

$$\mathbb{1}_{\mathcal{E}_Z}|Z_{hj}| \geq \mathbb{1}_{\mathcal{E}_Z}\left(\frac{\mathsf{C}_{E\bar{Z},j}}{2} h_j^{p-1} - \frac{1}{2}\lambda_{hj} - \frac{\mathsf{C}_{\mathsf{M}\Delta Z}}{(\log n)^{\frac{a}{2}}}\lambda_{hj}\right).$$

Now let us show: $\mathbb{1}_{\mathcal{E}_Z}|Z_{hj}| \geq \mathbb{1}_{\mathcal{E}_Z}\lambda_{hj}$.
First, if $n$ is large enough (ie $(\log n)^{\frac{a}{2}} \geq 2\mathsf{C}_{\mathsf{M}\Delta Z}$), then

$$\frac{\mathsf{C}_{\mathsf{M}\Delta Z}}{(\log n)^{\frac{a}{2}}}\lambda_{hj} \leq \frac{1}{2}\lambda_{hj}.$$

Then it suffices to prove:

$$\frac{\mathsf{C}_{E\bar{Z},j}}{2} h_j^{p-1} \geq 2\lambda_{hj},$$

*i.e.*:

$$h_j^{2p}\prod_{k=1}^{d} h_k \geq \frac{4^2\mathsf{C}_\lambda^2}{\mathsf{C}_{E\bar{Z},j}^2}\frac{(\log n)^a}{n}.$$

It is ensured for $t \leq \tau_n$, by definition of $\tau_n$ in (II.6):

$$h_j^{2p} \prod_{k=1}^{d} h_k = \frac{\beta^{t(2p+r)}}{(\log n)^{2p+d}} \geq \frac{\beta^{\tau_n(2p+r)}}{(\log n)^{2p+d}} = \frac{4^2 \mathsf{C}_\lambda^2}{\min_{k\in\mathcal{R}} \mathsf{C}_{E\bar{Z},k}^2} \frac{(\log n)^a}{n} \geq \frac{4^2 \mathsf{C}_\lambda^2}{\mathsf{C}_{E\bar{Z},j}^2} \frac{(\log n)^a}{n}. \tag{II.30}$$

Therefore, on $\mathcal{E}_Z$, the component $j$ remains active until the iteration $\lfloor \tau_n \rfloor$.

3. Let us now prove that on $\mathcal{E}_Z$, each relevant component $j$ deactivates at last at iteration $\lceil T_n \rceil$. In particular, by definition of $\mathcal{H}_{\mathsf{hp}}$, $\hat{h}$ belongs to $\mathcal{H}_{\mathsf{hp}}$ on $\mathcal{E}_Z$.
Assume $\mathcal{E}_Z$ happens.
We fix $j \in \mathcal{R}$. It suffices to prove that if $j$ is still active at iteration $\lceil T_n \rceil$, then on $\mathcal{E}_Z$, $j$ deactivates at the end of this iteration. We assume $j$ is still active and we denote $h$ the state of the bandwidth at iteration $\lceil T_n \rceil$.
By the first point, for any $k \notin \mathcal{R}$, $h_k = h_0$.
Given the second point, each relevant component $k$ was still active at the beginning of the iteration $\lfloor \tau_n \rfloor + 1$, *ie*: for any $k \in \mathcal{R}$, $h_k \leq \beta^{\lfloor \tau_n \rfloor + 1} h_0 \leq \beta^{\tau_n} h_0$.
Moreover, since $j$ is still active, $h_j = \beta^{\lceil T_n \rceil} h_0$. Let us prove that: $\mathbb{1}_{\mathcal{E}_Z} |Z_{hj}| < \lambda_{hj}$. Using the decomposition (II.29):

$$\mathbb{1}_{\mathcal{E}_Z} |Z_{hj}| \leq \left| \mathbb{E}\left[\bar{Z}_{hj}\right] \right| + \mathbb{1}_{\mathcal{E}_Z} \left| \bar{Z}_{hj} - \mathbb{E}\left[\bar{Z}_{hj}\right] \right| + \mathbb{1}_{\mathcal{E}_Z} \left| \Delta_{Z,hj} \right|.$$

Using the points 1. and 2. of Lemma II.6 and 1. Lemma II.7, since $\mathcal{E}_Z \subset \mathcal{B}_{\bar{Z},hj} \cap \mathcal{B}_{|\bar{Z}|,hj} \cap \widetilde{A}_n$:

$$\begin{aligned}
\mathbb{1}_{\mathcal{E}_Z} |Z_{hj}| &\leq 2\mathsf{C}_{E\bar{Z},j} h_j^{p-1} + \frac{1}{2}\lambda_{hj} + \frac{\mathsf{C}_{\mathsf{M}\Delta Z}}{(\log n)^{\frac{a}{2}}} \lambda_{hj} \\
&\leq \lambda_{hj} \left( \frac{2\mathsf{C}_{E\bar{Z},j} n^{\frac{1}{2}} h_j^p \prod_{k=1}^{d} h_k^{\frac{1}{2}}}{\mathsf{C}_\lambda (\log n)^{\frac{a}{2}}} + \frac{1}{2} + \frac{\mathsf{C}_{\mathsf{M}\Delta Z}}{(\log n)^{\frac{a}{2}}} \right).
\end{aligned}$$

Given the specific form of $h$ :

$$\begin{aligned}
\frac{2\mathsf{C}_{E\bar{Z},j} n^{\frac{1}{2}} h_j^p \prod_{k=1}^{d} h_k^{\frac{1}{2}}}{\mathsf{C}_\lambda (\log n)^{\frac{a}{2}}} &\leq \frac{2\mathsf{C}_{E\bar{Z},j} n^{\frac{1}{2}} h_0^{\frac{2p+d}{2}} \beta^{\frac{(2p+1)}{2}(T_n - \tau_n)} \beta^{\frac{(2p+r)\tau_n}{2}}}{\mathsf{C}_\lambda (\log n)^{\frac{a}{2}}} \\
&= \sqrt{\frac{4^3 \mathsf{C}_{E\bar{Z},j}^2 \beta^{(2p+1)(T_n - \tau_n)}}{\min_{k\in\mathcal{R}} \mathsf{C}_{E\bar{Z},k}^2}}, \text{ by definition of } \tau_n \\
&\leq \frac{1}{3}, \text{ by definition of } T_n.
\end{aligned}$$

Moreover, for $n$ large enough:

$$\frac{\mathsf{C}_{\mathsf{M}\Delta Z}}{(\log n)^{\frac{a}{2}}} < \frac{1}{6}.$$

Therefore:

$$\mathbb{1}_{\mathcal{E}_Z} |Z_{hj}| < \lambda_{hj}.$$

In other words, when $\mathcal{E}_Z$ happens, any active component at iteration $\lceil T_n \rceil$ deactivates.

So we have proved that on $\mathcal{E}_Z$, $\hat{h} \in \mathcal{H}_{\mathsf{hp}}$.

It remains to show that $\mathcal{E}_Z$ holds with high probability.

$$\mathbb{P}(\mathcal{E}_Z^c) \leq \mathbb{P}\left(\widetilde{A}_n^c\right) + \sum_{k=1}^{d}\left\{\mathbb{P}\left(\mathcal{B}^c_{\bar{Z},h^{(0)}k}\right) + \mathbb{P}\left(\mathcal{B}^c_{|\bar{Z}|,h^{(0)}k}\right)\right\}$$
$$+ \sum_{j\in\mathcal{R}}\left[\sum_{h\in\mathcal{H}_{\mathsf{hp}}}\left(\mathbb{P}\left(\mathcal{B}^c_{\bar{Z},hj}\right) + \mathbb{P}\left(\mathcal{B}^c_{|\bar{Z}|,hj}\right)\right) + \sum_{t=1}^{\lfloor\tau_n\rfloor}\left(\mathbb{P}\left(\mathcal{B}^c_{\bar{Z},h^{(t)}j}\right) + \mathbb{P}\left(\mathcal{B}^c_{|\bar{Z}|,h^{(t)}j}\right)\right)\right].$$

By choice of $\widetilde{\mathrm{f}}_X$:

$$\mathbb{P}\left(\widetilde{A}_n^c\right) \leq C_X e^{-(\log n)^{\frac{5}{4}}}.$$

We want to apply 2. and 3. of Lemma II.6 for any $h \in \mathcal{H}_{\mathsf{hp}}$ and any $h^{(t)}$ with $t = 1 : \lfloor\tau_n\rfloor$. These bandwidths satisfy:

$$\prod_{k=1}^{d} h_k^{(t)} \geq \prod_{k=1}^{d} h_k \geq h_0^d \beta^{r\lceil T_n\rceil} \geq \mathsf{C}_T^{\frac{r}{2p+1}} \mathsf{C}_\tau^{\frac{r}{2p+r}} (\log n)^{\frac{ra-2p(d-r)}{2p+r}} n^{-\frac{r}{2p+r}},$$

which ensures that for $n$ large enough, $\mathrm{Cond}_{\bar{Z}}(h^{(t)})$ and $\mathrm{Cond}_{\bar{Z}}(h^{(t)})$ hold for any $h \in \mathcal{H}_{\mathsf{hp}}$ and any $t = 0 : \lfloor\tau_n\rfloor$. Note in particular that $\mathcal{H}_{\mathsf{hp}} \subset \{h^{(t),t=0:\lceil T_n\rceil}\}$.
Since, for any component $k = 1 : d$,

$$\mathbb{P}\left(\mathcal{B}^c_{\bar{Z},h^{(0)}k}\right) \leq 2e^{-\gamma_{Z,n}},$$

by induction, for any $h \in \mathcal{H}_{\mathsf{hp}}$ and any $t = 0 : \lceil T_n\rceil$ ,

$$\begin{aligned}\mathbb{P}\left(\mathcal{B}^c_{|\bar{Z}|,h^{(t)}j}\right) &\leq 2\exp\left(-\mathsf{C}_{\gamma|\bar{Z}|} n \prod_{k=1}^{d} h_k\right)\\ &\leq 2\exp\left(-\mathsf{C}_{\gamma|\bar{Z}|}\mathsf{C}_T^{\frac{r}{2p+1}}\mathsf{C}_\tau^{\frac{r}{2p+r}}(\log n)^{\frac{ra-2p(d-r)}{2p+r}} n^{\frac{2p}{2p+r}}\right)\\ &\leq 2e^{-\gamma_{Z,n}}, \text{ for } n \text{ large enough.}\end{aligned}$$

To conclude, note that $|\mathcal{H}_{\mathsf{hp}}| = (\lceil T_n\rceil - \lfloor\tau_n\rfloor)^r \leq (T_n - \tau_n + 2)^r = \left(\frac{\log(\mathsf{C}_T^{-1})}{(2p+1)\log\frac{1}{\beta}} + 2\right)^r$ is finite, so for any $q > 0$:

$$\begin{aligned}\mathbb{P}(\mathcal{E}_Z^c) &\leq C_X e^{-(\log n)^{\frac{5}{4}}} + 2(d + r|\mathcal{H}_{\mathsf{hp}}| + r\tau_n)2e^{-\gamma_{Z,n}}\\ &= o(n^{-q}),\end{aligned}$$

by definition $\gamma_{Z,n} := \frac{\delta}{\|f\|_{\infty,\,\mathcal{U}_n(w)}}(\log n)^a$.

### 5.d.ii Proof of Lemma II.4

1. We control the bias $\left\|\mathbb{E}\left[\tilde{\mathrm{f}}_X^K\right] - \mathrm{f}_X\right\|_{\infty,\,\mathcal{U}_n(x)}$. We write for any $u \in \mathcal{U}_n(x)$:

$$\mathbb{E}\left[\tilde{\mathrm{f}}_X^K(u)\right] - \mathrm{f}_X(u) = \frac{1}{h_X^{d_1}}\int_{u'\in\mathbb{R}^{d_1}}\left(\prod_{j=1}^{d_1} K_X\left(\frac{u_j - u'_j}{h_X}\right)\right)\mathrm{f}_X(u')du' - \mathrm{f}_X(u)\int_{\mathbb{R}^{d_1}}\left(\prod_{j=1}^{d_1} K_X(z_j)\right)dz$$

The kernel $K_X$ is of order $p_X$ and $\mathrm{f}_X$ is assumed of class $\mathcal{C}^{p'}$ on $\mathcal{U}'_n(x)$, with in particular $p'-1 \leq p_X - 1$, then we can apply Lemma II.9 with the settings $u = u$, $d' = d_1$, $f_0 = \mathrm{f}_X$, $p = p'-1$, $K = K_X$ and for $j = 1 : d'$, $h_k = h_X$. We obtain:

$$\mathbb{E}\left[\tilde{\mathrm{f}}_X^K(u)\right] - \mathrm{f}_X(u) = \sum_{k=1}^{d_1} (\mathsf{I}_k + \mathsf{II}_k). \tag{II.31}$$

with

$$\begin{aligned}
\mathsf{I}_k &:= \int_{z\in\mathbb{R}^{d_1}} \left(\prod_{k'=1}^{d_1} K_X(z_{k'})\right) \rho_k dz,\\
\rho_k &:= \rho_k(z, h_X, u)\\
&= (-h_X z_k)^{p'-1} \int\limits_{0\leq t_{p'-1}\leq\cdots\leq t_1\leq 1} \left(\partial_k^{p'-1}\mathrm{f}_X(\bar{z}_{k-1} - t_{p'-1}h_X z_k e_k) - \partial_k^{p'-1}\mathrm{f}_X(\bar{z}_{k-1})\right) dt_{1:(p'-1)},\\
\mathsf{II}_k &:= (-h_X)^{p'-1} \int_{t\in\mathbb{R}} \frac{t^{p'-1}}{(p'-1)!} K_X(t)dt \int_{z_{-k}\in\mathbb{R}^{d_1-1}} \partial_k^{p'-1}\mathrm{f}_X(\bar{z}_{k-1}) \left(\prod_{k'\neq k} K_X(z_{k'})\right) dz_{-k}.
\end{aligned}$$

Let us control $\rho_k$. First we write:

$$\partial_k^{p'-1}\mathrm{f}_X(\bar{z}_{k-1} - t_{p'-1}h_X z_k e_k) - \partial_k^{p'-1}\mathrm{f}_X(\bar{z}_{k-1}) = -h_X z_k \int_{t_{p'}=0}^{t'_{p-1}} \partial_k^{p'}\mathrm{f}_X(\bar{z}_{k-1} - t_{p'}h_X z_k e_k)dt_{p'}.$$

Therefore:

$$\rho_k = (-h_X z_k)^{p'} \int\limits_{0\leq t_{p'}\leq\cdots\leq t_1\leq 1} \partial_k^{p'}\mathrm{f}_X(\bar{z}_{k-1} - t_{p'}h_X z_k e_k)dt_{1:p'}.$$

Hence:

$$\begin{aligned}
|\rho_k| &\leq |h_X z_k|^{p'} \int\limits_{0\leq t_{p'}\leq\cdots\leq t_1\leq 1} \left|\partial_k^{p'}\mathrm{f}_X(\bar{z}_{k-1} - t_{p'}h_X z_k e_k)\right| dt_{1:p'}\\
&= \frac{|z_k|^{p'}}{p'!} \|\partial_k^{p'}\mathrm{f}_X\|_{\infty,\,\mathcal{U}'_n(x)} h_X^{p'}.
\end{aligned}$$

Then:

$$\begin{aligned}
|\mathsf{I}_k| &\leq \int_{z\in\mathbb{R}^{d_1}} \left|\prod_{k'=1}^{d_1} K_X(z_{k'})\right| |\rho_k|\, dz\\
&\leq \|\partial_k^{p'}\mathrm{f}_X\|_{\infty,\,\mathcal{U}'_n(x)} h_X^{p'} \int_{z\in\mathbb{R}^{d_1}} \frac{|z_k|^{p'}}{p'!} \left|\prod_{k'=1}^{d_1} K_X(z_{k'})\right| dz\\
&= \frac{\|K_X\|_1^{d_1-1} \|(\cdot)^{p'} K_X(\cdot)\|_1}{p'!} \|\partial_k^{p'}\mathrm{f}_X\|_{\infty,\,\mathcal{U}'_n(x)} h_X^{p'}
\end{aligned} \tag{II.32}$$

Besides, $K_X$ is of order $p_X$ and $p'-1 < p_X$ and so:

$$\mathsf{II}_k := \frac{(-h_X)^{p'-1}}{(p'-1)!} \int_{t\in\mathbb{R}} t^{p'-1} K_X(t)dt \int_{z_{-k}\in\mathbb{R}^{d_1-1}} \partial_k^{p'-1}\mathrm{f}_X(\bar{z}_{k-1}) \left(\prod_{k'\neq k} K_X(z_{k'})\right) dz_{-k} = 0.$$

Therefore the terms $\mathrm{II}_k$ vanish in the equation (II.31), and with the upper bound of $\mathrm{I}_k$ (II.32), we obtain:

$$\begin{aligned}
\left\|\mathbb{E}\left[\tilde{\mathrm{f}}_X^K\right] - \mathrm{f}_X\right\|_{\infty,\,\mathcal{U}_n(x)} &= \sup_{u\in\,\mathcal{U}_n(x)} \left|\mathbb{E}\left[\tilde{\mathrm{f}}_X^K(u)\right] - \mathrm{f}_X(u)\right| \\
&\leq \sup_{u\in\,\mathcal{U}_n(x)} \sum_{k=1}^{d_1} |\mathrm{I}_k| \\
&\leq \frac{\|K_X\|_1^{d_1-1}\|(\cdot)^{p'}K_X(\cdot)\|_1}{p'!} h_X^{p'} \sum_{k=1}^{d_1} \|\partial_k^{p'}\mathrm{f}_X\|_{\infty,\,\mathcal{U}_n'(x)} \\
&= \mathsf{C}_{\mathsf{bias}_X} h_X^{p'},
\end{aligned}$$

with $\mathsf{C}_{\mathsf{bias}_X} := \frac{\|K_X\|_1^{d_1-1}\|.^{p'}K_X(\cdot)\|_1}{p'!} d_1 \max\limits_{k=1:d_1} \|\partial_k^{p'}\mathrm{f}_X\|_{\infty,\,\mathcal{U}_n'(x)}$.

2. We apply Bernstein's inequality (see Lemma II.10). We define for any $u \in \mathcal{U}_n(x)$ and any $i = 1 : n_X$:

$$\tilde{\mathrm{f}}_{Xi}^K(u) := \frac{1}{h_X^{d_1}} \prod_{j=1}^{d_1} K_X\left(\frac{u_j - \widetilde{X}_{ij}}{h_X}\right).$$

Then we control $\tilde{\mathrm{f}}_{X1}^K$ a.s.: for any $u \in \mathcal{U}_n(x)$,

$$\left|\tilde{\mathrm{f}}_{X1}^K(u)\right| \leq \mathsf{M}_{h_X} := \|K_X\|_\infty^{d_1} h_X^{-d_1}.$$

and its variance:

$$\begin{aligned}
\mathrm{Var}\left(\tilde{\mathrm{f}}_{X1}^K(u)\right) &\leq \mathbb{E}\left[(\tilde{\mathrm{f}}_{X1}^K)^2\right] \\
&= h_X^{-2d_1} \int_{u'\in\mathbb{R}^{d_1}} \left(\prod_{j=1}^{d_1} K_X\left(\frac{u_j - u_j'}{h_X}\right)\right)^2 \mathrm{f}_X(u')du' \\
&= h_X^{-d_1} \int_{z\in\mathbb{R}^{d_1}} \left(\prod_{j=1}^{d_1} K_X(z_j)\right)^2 \mathrm{f}_X(u - h_X z)dz \\
&\leq \mathsf{v}_{h_X}
\end{aligned}$$

with $\mathsf{v}_{h_X} := \mathsf{C}_{\mathsf{v}X} h_X^{-d_1}$ and $\mathsf{C}_{\mathsf{v}X} := \|K_X\|_2^{2d_1}\|\mathrm{f}_X\|_{\infty,\,\mathcal{U}_n'(x)}$.
Then we apply Lemma II.10: for any $\lambda > 0$,

$$\mathbb{P}\left(\left|\tilde{\mathrm{f}}_X^K(u) - \mathbb{E}\left[\tilde{\mathrm{f}}_X^K(u)\right]\right| > \lambda\right) \leq 2\exp\left(-\min\left(\frac{n_X\lambda^2}{4\mathsf{v}_{h_X}}, \frac{3n_X\lambda}{4\mathsf{M}_{h_X}}\right)\right).$$

We set $\lambda$ at $\lambda_X := \sqrt{\frac{4\mathsf{v}_{h_X}}{n_X}(\log n)^{\frac{3}{2}}}$ such that $(\log n)^{\frac{3}{2}} = \frac{n_X\lambda_X^2}{4\mathsf{v}_{h_X}}$. Then we compare the rates:

$$\begin{aligned}
& \frac{n_X\lambda_X^2}{4\mathsf{v}_{h_X}} \leq \frac{3n_X\lambda_X}{4\mathsf{M}_{h_X}} \\
\iff \quad & \lambda_X^2 \leq \frac{3^2\mathsf{C}_{\mathsf{v}X}^2}{\|K_X\|_\infty^{2d_1}} \\
\iff \quad & h_X^{d_1} \geq \frac{4\|K_X\|_\infty^{2d_1}}{9\mathsf{C}_{\mathsf{v}X}} \frac{(\log n)^{\frac{3}{2}}}{n_X}, \\
\iff \quad & \mathsf{Cond}_X(h_X).
\end{aligned}$$

Hence, when $\text{Cond}_X(h_X)$ is satisfied :

$$\begin{aligned}\mathbb{P}\left(\left|\tilde{\mathrm{f}}_X^K(u) - \mathbb{E}\left[\tilde{\mathrm{f}}_X^K(u)\right]\right| > \lambda_X\right) &\leq 2\exp\left(-\frac{n_X\lambda_X^2}{4\mathsf{v}_{h_X}}\right)\\ &\leq 2\exp\left(-(\log n)^{\frac{3}{2}}\right).\end{aligned}$$

### 5.d.iii Proof of Lemma II.5

1. We recall the notation $\cdot$ for the multiplication terms by terms of two vectors. Then:

$$\begin{aligned}\left|\mathbb{E}\left[\bar{f}_{h1}(w)\right]\right| &\leq \mathbb{E}\left[\left|\bar{f}_{h1}(w)\right|\right]\\ &= \int_{u\in\mathbb{R}^d}\left|\prod_{k=1}^d \frac{K(h_k^{-1}(w_k-u_k))}{h_k}\right| f(u)du\\ &= \int_{z\in\mathbb{R}^d}\left|\prod_{k=1}^d K(z_k)\right| f(w-h\cdot z)dz\\ &\leq \|f\|_{\infty,\,\mathcal{U}_n(w)}\|K\|_1^d =: \mathsf{C}_{\bar{\mathsf{E}}}.\end{aligned}$$

Now let us give an upper bound on the bias of $\bar{f}_h(w)$:

$$\overline{B}_h = \mathbb{E}\left[\bar{f}_{h1}(w)\right] - f(w) = \int_{u\in\mathbb{R}^d}\left(\prod_{k=1}^d \frac{K(h_k^{-1}(w_k-u_k))}{h_k}\right) f(u)du - f(w)\int_{\mathbb{R}^d}\prod_{k'=1}^d K(z_{k'})dz,$$

since $\int_{\mathbb{R}} K(t)dt = 1$. Then we apply the Lemma II.9 with the settings $d' = d$, $u = w$, $h = h$, $f_0 = f$, $p = p$ and $K = K$. We obtain:

$$\overline{B}_h = \sum_{k=1}^d (\mathsf{I}_k + \mathsf{II}_k),$$

where

$$\begin{aligned}\mathsf{I}_k &:= \int_{z\in\mathbb{R}^d}\left(\prod_{k'=1}^d K(z_{k'})\right)\rho_k dz,\\ \rho_k &:= (-h_k z_k)^p \int_{0\leq t_p\leq\cdots\leq t_1\leq 1}\left(\partial_k^p f(\bar{z}_{k-1} - t_p h_k z_k e_k) - \partial_k^p f(\bar{z}_{k-1})\right)dt_{1:p},\\ \mathsf{II}_k &:= (-h_k)^p \int_{t\in\mathbb{R}}\frac{t^p}{p!}K(t)dt\int_{z_{-k}\in\mathbb{R}^{d-1}}\partial_k^p f(\bar{z}_{k-1})\left(\prod_{k'\neq k}K(z_{k'})\right)dz_{-k}.\end{aligned}$$

Notice that for $k\notin\mathcal{R}$, $\partial_k^p f(u) = 0$ for any $u\in\mathcal{U}_n(x)$, thus $\mathsf{I}_k$ and $\mathsf{II}_k$ vanish. Therefore:

$$\overline{B}_h = \sum_{k\in\mathcal{R}}(\mathsf{I}_k + \mathsf{II}_k).$$

Now let us give an equivalent of the bias. First, using Assumption II.3, for any $k\in\mathcal{R}$, we can define the modulus of continuity of $\partial_k^p f$ on $\mathcal{U}_n(w)$ by:

$$\Omega_{nk} := \sup_{z,z'\in\mathcal{U}_n(w)}\left|\partial_k^p f(z') - \partial_k^p f(z)\right|.$$

Then we decompose $\mathrm{II}_k$ as follows:

$$\mathrm{II}_k = \frac{(-h_k)^p \int_{t\in\mathbb{R}} t^p K(t)dt}{p!} \partial_k^p f(w) + R_k,$$

with $R_k := \frac{(-h_k)^p \int_{t\in\mathbb{R}} t^p K(t)dt}{p!} \int_{z_{-k}\in\mathbb{R}^{d-1}} (\partial_k^p f(\bar{z}_{k-1}) - \partial_k^p f(w)) \left( \prod_{k'\neq k} K(z_{k'}) \right) dz_{-k}$ such that:

$$|R_k| \leq h_k^p \left| \int_{t\in\mathbb{R}} \frac{t^p}{p!} K(t)dt \right| \Omega_{nk} \|K\|_1^{d-1} \tag{II.33}$$

since $\left|\partial_k^p f(\bar{z}_{k-1}) - \partial_k^p f(w)\right| \leq \Omega_{nk}$.
It remains to bound $\mathrm{I}_k$. From the definition of $\rho_k$ in (II.50), we write:

$$\begin{aligned} |\rho_k| &\leq |h_k z_k|^p \left| \int_{0\leq t_p \leq \cdots \leq t_1 \leq 1} \left[ \partial_k^p f(\bar{z}_{k-1} - t_p h_k z_k e_k) - \partial_k^p f(\bar{z}_{k-1}) \right] dt_{1:p} \right| \\ &\leq |h_k z_k|^p \frac{\Omega_{nk}}{p!}. \end{aligned}$$

Therefore:

$$\begin{aligned} |\mathrm{I}_k| &= \left| \int_{z\in\mathbb{R}^d} \left( \prod_{k'=1}^d K(z_{k'}) \right) \rho_k dz \right| \\ &\leq \frac{h_k^p}{p!} \Omega_{nk} \int_{z\in\mathbb{R}^d} \left| z_k^p \prod_{k'=1}^d K(z_{k'}) \right| dz \\ &\leq \|K\|_1^{d-1} \int_{t\in\mathbb{R}} \left| \frac{t^p}{p!} K(t) \right| dt \times h_k^p \Omega_{nk}. \end{aligned} \tag{II.34}$$

Since $\mathcal{U}_n(w) \underset{n\to\infty}{\longrightarrow} \{w\}$, by continuity of $\partial_k^p f$:

$$\Omega_{nk} \underset{n\to\infty}{\longrightarrow} 0.$$

Therefore for $n$ large enough (when $\Omega_{nk} \leq \frac{1}{2\|K\|_1^{d-1}}$), combining (II.33) and (II.34):

$$|\mathrm{I}_k| + |R_k| \leq \frac{|\int_{t\in\mathbb{R}} t^p K(t)dt|}{p!} \max_{k\in\mathcal{R}} |\partial_k^p f(w)| \times h_k^p.$$

Therefore, since:

$$\overline{B}_h = \sum_{k\in\mathcal{R}} (\mathrm{II}_k + \mathrm{I}_k) = \sum_{k\in\mathcal{R}} \left( \frac{(-h_k)^p \int_{t\in\mathbb{R}} t^p K(t)dt \; \partial_k^p f(w)}{p!} + R_k + \mathrm{I}_k \right),$$

we obtain:

$$\left|\overline{B}_h\right| \leq \mathrm{C}_{\overline{\mathrm{bias}}} \sum_{k\in\mathcal{R}} h_k^p,$$

with $\mathrm{C}_{\overline{\mathrm{bias}}} := \frac{2|\int_{t\in\mathbb{R}} t^p K(t)dt|}{p!} \max_{k\in\mathcal{R}} |\partial_k^p f(w)|$.

2. We want to apply Bernstein's inequality (cf Lemma II.10) to $\bar{f}_h(w)$. We first obtain an almost sure upper bound:

$$\begin{aligned} |\bar{f}_{h1}(w)| &= \frac{1}{\mathrm{f}_X(X_1)} \prod_{k=1}^{d} \frac{\left| K\left(\frac{w_k - W_{1k}}{h_k}\right)\right|}{h_k} \\ &\leq \overline{\mathsf{M}}_h, \end{aligned} \tag{II.35}$$

where $\overline{\mathsf{M}}_h := \frac{\mathsf{C}_{\bar{\mathsf{M}}}}{\prod_{k=1}^{d} h_k}$ with $\mathsf{C}_{\bar{\mathsf{M}}} := \frac{\|K\|_\infty^d}{\delta}$.
Then we control the variance:

$$\begin{aligned} \mathrm{Var}\left(\bar{f}_{h1}(w)\right) &= \mathrm{Var}\left( \frac{1}{\mathrm{f}_X(X_1)} \prod_{k=1}^{d} \frac{K\left(\frac{w_k - W_{1k}}{h_k}\right)}{h_k} \right) \\ &\leq \mathbb{E}\left[ \left( \frac{1}{\mathrm{f}_X(X_1)} \prod_{k=1}^{d} \frac{K\left(\frac{w_k - W_{1k}}{h_k}\right)}{h_k} \right)^2 \right] \\ &= \int_{u\in\mathbb{R}^d} \left\{ \prod_{k=1}^{d} \frac{1}{h_k^2} K\left(\frac{w_k - u_k}{h_k}\right)^2 \right\} \frac{f(u)}{\mathrm{f}_X(u_{1:d_1})} du \\ &\leq \frac{1}{\delta \prod_{k=1}^{d} h_k} \int_{z\in\mathbb{R}^d} \left\{ \prod_{k=1}^{d} K(z_k)^2 \right\} f(w - Hz) dz \\ &\leq \bar{\mathsf{v}}_h, \end{aligned} \tag{II.36}$$

where $\bar{\mathsf{v}}_h := \frac{\mathsf{C}_\sigma^2}{4\prod_{k=1}^{d} h_k}$.Therefore we obtain from Bernstein's inequality (cf Lemma II.10):

$$\mathbb{P}\left(\overline{\mathcal{B}}_h^c\right) \leq 2\exp\left( -\min\left( \frac{n\sigma_h^2}{4\bar{\mathsf{v}}_h}, \frac{3n\sigma_h}{4\overline{\mathsf{M}}_h} \right)\right).$$

We compare the rates:

$$\begin{aligned} & \frac{n\sigma_h^2}{4\bar{\mathsf{v}}_h} \leq \frac{3n\sigma_h}{4\overline{\mathsf{M}}_h} \\ \iff \quad & \mathsf{C}_\sigma \sqrt{\frac{(\log n)^a}{n\prod_{k=1}^{d} h_k}} = \sigma_h \leq \frac{3\bar{\mathsf{v}}_h}{\overline{\mathsf{M}}_h} = \frac{3\mathsf{C}_\sigma^2}{4\mathsf{C}_{\bar{\mathsf{M}}}} \\ \iff \quad & \prod_{k=1}^{d} h_k \geq \frac{4^2\mathsf{C}_{\bar{\mathsf{M}}}^2}{9\mathsf{C}_\sigma^2} \frac{(\log n)^a}{n} \\ \iff \quad & \mathsf{Cond}(h). \end{aligned}$$

Therefore, if $\mathsf{Cond}(h)$ is satisfied:

$$\mathbb{P}\left(\overline{\mathcal{B}}_h^c\right) \leq 2e^{-\frac{n\sigma_h^2}{4\bar{\mathsf{v}}_h}} = 2e^{-(\log n)^a}.$$

3. We now apply Bernstein's inequality (cf Lemma II.10) to $\frac{1}{n}\sum_{i=1}^{n}|\bar{f}_{hi}(w)|$. From the upper bounds (II.35) and (II.36), we obtain:

$$\mathbb{P}\left(\mathcal{B}^c_{|\bar{f}|h}\right) \leq 2\exp\left(-\min\left(\frac{n\mathsf{C}^2_{\bar{\mathsf{E}}}}{4\overline{\mathsf{v}}_h}, \frac{3n\mathsf{C}_{\bar{\mathsf{E}}}}{4\overline{\mathsf{M}}_h}\right)\right).$$

We calculate the rates: by definition of $\overline{\mathsf{v}}_h$ and $\overline{\mathsf{M}}_h$,

$$\frac{n\mathsf{C}^2_{\bar{\mathsf{E}}}}{4\overline{\mathsf{v}}_h} = \frac{\mathsf{C}^2_{\bar{\mathsf{E}}}}{\mathsf{C}^2_{\sigma}} n\prod_{k=1}^{d} h_k$$

and

$$\frac{3n\mathsf{C}_{\bar{\mathsf{E}}}}{4\overline{\mathsf{M}}_h} = \frac{3\mathsf{C}_{\bar{\mathsf{E}}}}{4\mathsf{C}_{\bar{\mathsf{M}}}} n\prod_{k=1}^{d} h_k.$$

Hence:

$$\mathbb{P}\left(\mathcal{B}^c_{|\bar{f}|h}\right) \leq 2e^{-\mathsf{C}_{\gamma|f|} n\prod_{k=1}^{d} h_k},$$

with $\mathsf{C}_{\gamma|f|} := \min\left(\frac{\mathsf{C}^2_{\bar{\mathsf{E}}}}{\mathsf{C}^2_{\sigma}}; \frac{3\mathsf{C}_{\bar{\mathsf{E}}}}{4\mathsf{C}_{\bar{\mathsf{M}}}}\right)$.

### 5.d.iv Proof of Lemma II.6

1. First, we write $\bar{Z}_{hij}$ more explicitly: for any bandwidth $h$, any observation $i = 1:n$ and any direction $j$,

$$\begin{aligned}
\bar{Z}_{hij} &= \frac{\partial}{\partial h_j}\left(\frac{K(\frac{w_j - W_{ij}}{h_j})}{h_j}\right)\frac{\prod_{k\neq j} K(\frac{w_k - W_{ik}}{h_k})}{\mathrm{f}_X(X_i)\prod_{k\neq j} h_k} \\
&= \frac{-\left(K(\frac{w_j - W_{ij}}{h_j}) + \frac{w_j - W_{ij}}{h_j}K'(\frac{w_j - W_{ij}}{h_j})\right)\prod_{k\neq j} K(\frac{w_k - W_{ik}}{h_k})}{\mathrm{f}_X(X_i)h_j\prod_{k=1}^{d} h_k} \\
&= \frac{-J(\frac{w_j - W_{ij}}{h_j})\prod_{k\neq j} K(\frac{w_k - W_{ik}}{h_k})}{\mathrm{f}_X(X_i)h_j\prod_{k=1}^{d} h_k}
\end{aligned}$$

where we recall $J : \mathbb{R} \to \mathbb{R}$ is the function $t \mapsto tK'(t) + K(t)$.
Note then that the support of $J$ is included in the support of $K$, and by integration by part, we obtain for any $l \in \mathbb{N}$:

$$\int_{\mathbb{R}} t^l J(t)dt = \int_{\mathbb{R}} t^l (tK(t))'dt = -l\int_{\mathbb{R}} t^l K(t)dt. \tag{II.37}$$

In particular, since $K$ is of order $p$, for $l = 0 : p-1$, $\int_{\mathbb{R}} t^l J(t)dt = 0$ and $\int_{\mathbb{R}} t^p J(t)dt \neq 0$.
We recall the notation $\cdot$ for the multiplication terms by terms of two vectors. We also denote, for any $z \in \mathbb{R}^d$ and any $j \in \{1,\dots,d\}$,

$$\tilde{z}_{-j} := w - (Hz)_{-j} = w - \sum_{k\neq j} h_k z_k e_k$$

with $\{e_k\}_{k=1}^d$ is the canonic basis of $\mathbb{R}^d$.
Using Assumption II.2, if $j \notin \mathcal{R}$, $f(w - h\cdot z) - f(\tilde{z}_{-j}) = 0$ for any $z \in \mathbb{R}^d$. Thus we obtain:

$$\begin{aligned}\mathbb{E}\left[\bar{Z}_{h1\,j}\right] &= -\frac{1}{h_j \prod_{k=1}^d h_k} \int_{u\in\mathbb{R}^d} J(\tfrac{w_j-u_j}{h_j}) \left(\prod_{k\neq j} K(\tfrac{w_k-u_k}{h_k})\right) f(u)du \\ &= -\frac{1}{h_j} \int_{z_j\in\mathbb{R}} J(z_j)dz_j \int_{z_{-j}\in\mathbb{R}^{d-1}} \left(\prod_{k\neq j} K(z_k)\right) f(w - h\cdot z)dz_{-j} = 0.\end{aligned}$$

Therefore $\mathbb{E}\left[\bar{Z}_{h1\,j}\right] = 0$ for $j \notin \mathcal{R}$.
Now, we deal with the case $j \in \mathcal{R}$. Let us fix $j \in \mathcal{R}$. Then we write:

$$\mathbb{E}\left[\bar{Z}_{h1\,j}\right] = \frac{-1}{h_j \prod_{k=1}^d h_k} \int_{u_{-j}\in\mathbb{R}^{d-1}} \left(\prod_{k\neq j} K(\tfrac{w_k-u_k}{h_k})\right) \left[\int_{u_j\in\mathbb{R}} J(\tfrac{w_j-u_j}{h_j}) f(u)du_j - f(\tilde{z}_{-j}) \int_{\mathbb{R}} J(z_j)dz_j\right] du_{-j}.$$

Then for fixed $\{z_k\}_{k\neq j}$, denoting $f_j : z_j \mapsto f(w - h\cdot z)$, we apply Lemma II.9 with the settings $d' = 1$, $u = \tilde{z}_{-j}$, $h = h_j$, $f_0 = f_j$, $p = p$, $K = J$, then

$$\begin{aligned}\mathbb{E}\left[\bar{Z}_{h1\,j}\right] &= \frac{-1}{h_j \prod_{k=1}^d h_k} \int_{u_{-j}\in\mathbb{R}^{d-1}} \left(\prod_{k\neq j} K(\tfrac{w_k-u_k}{h_k})\right) [\mathrm{I}_1 + \mathrm{II}_1]\, du_{-j} \\ &= \widetilde{\mathrm{I}}_j + \tilde{\mathrm{II}}_j, \end{aligned} \tag{II.38}$$

where

$$\tilde{\mathrm{I}}_j := (-h_j)^{-1} \int_{z\in\mathbb{R}^d} \left(\prod_{k\neq j} K(z_k)\right) J(z_j)\tilde{\rho}_j dz, \tag{II.39}$$

$$\text{with } \tilde{\rho}_j := (-h_j z_j)^p \int_{0\le t_p\le\cdots\le t_1\le 1} \left(\partial_j^p f(\tilde{z}_{-j} - t_p h_j z_j e_j) - \partial_j^p f(\tilde{z}_{-j})\right) dt_{1:p}, \tag{II.40}$$

$$\text{and } \widetilde{\mathrm{II}}_j := (-h_j)^{p-1} \int_{t\in\mathbb{R}} \frac{t^p}{p!} J(t)dt \int_{z_{-j}\in\mathbb{R}^{d-1}} \partial_j^p f(\tilde{z}_{j-1}) \left(\prod_{k'\neq j} K(z_{k'})\right) dz_{-j}.$$

Now let us determine an equivalent of $\mathbb{E}\left[\bar{Z}_{hj}\right]$. For this purpose, let us introduce the modulus of continuity of $\partial_j^p f$ on $\mathcal{U}_n(w)$ (which is well defined by Assumption II.3):

$$\Omega_{nj} := \sup_{z,z'\in\,\mathcal{U}_n(w)} \left|\partial_j^p f(z') - \partial_j^p f(z)\right|.$$

Then we write:

$$\widetilde{\mathrm{II}}_j = (-h_j)^{p-1} \partial_j^p f(w) \int_{t\in\mathbb{R}} \frac{t^p}{p!} J(t)dt + \widetilde{R}_j, \tag{II.41}$$

with

$$\widetilde{R}_j := (-h_j)^{p-1} \int_{t\in\mathbb{R}} \frac{t^p}{p!} J(t)dt \int_{z_{-j}\in\mathbb{R}^{d-1}} \left(\partial_j^p f(\tilde{z}_{-j}) - \partial_j^p f(w)\right) \left(\prod_{k\neq j} K(z_k)\right) dz_{-j}.$$

In particular:

$$|\widetilde{R}_j| \leq h_j^{p-1} \left| \int_{t\in\mathbb{R}} \frac{t^p}{p!} J(t)dt \right| \int_{z_{-k}\in\mathbb{R}^{d-1}} \Omega_{nj} \prod_{k\neq j} |K(z_k)| dz_{-j}$$
$$= h_j^{p-1} \Omega_{nj} \left| \int_{t\in\mathbb{R}} \frac{t^p}{p!} J(t)dt \right| \|K\|_1^{d-1}. \tag{II.42}$$

Now let us bound $\tilde{\mathrm{I}}_j$ defined in (II.39). First, we bound $\tilde{\rho}_j$, defined in (II.40):

$$|\tilde{\rho}_j| = (h_j|z_j|)^p \left| \int_{0\leq t_p\leq\cdots\leq t_1\leq 1} \left( \partial_j^p f(\tilde{z}_{-j} - t_p h_j z_j e_j) - \partial_j^p f(\tilde{z}_{-j}) \right) dt_{1:p} \right|$$
$$\leq h_j^p |z_j|^p \frac{\Omega_{nj}}{p!},$$

which leads to:

$$\left|\tilde{\mathrm{I}}_j\right| = h_j^{-1} \left| \int_{z\in\mathbb{R}^d} \left( \prod_{k\neq j} K(z_k) \right) J(z_j) \tilde{\rho}_j dz \right|$$
$$\leq h_j^{p-1} \Omega_{nj} \|K\|_1^{d-1} \int_{z_j\in\mathbb{R}} \frac{|z_j|^p}{p!} |J(z_j)| dz_j. \tag{II.43}$$

Therefore using (II.41) then (II.42) and (II.43):

$$\left|\mathbb{E}\left[\bar{Z}_{h1j}\right]\right| \leq \left|\widetilde{\mathrm{II}}_j\right| + \left|\tilde{\mathrm{I}}_j\right| \quad \leq \quad h_j^{p-1} \left| \partial_j^p f(w) \int_{t\in\mathbb{R}} \frac{t^p}{p!} J(t)dt \right| + \left|\widetilde{R}_j\right| + \left|\tilde{\mathrm{I}}_j\right|$$
$$\leq \mathsf{C}_{E\bar{Z},j} h_j^{p-1} + h_j^{p-1} \Omega_{nj} \left( \left| \int_{t\in\mathbb{R}} \frac{t^p}{p!} J(t)dt \right| \|K\|_1^{d-1} + \|K\|_1^{d-1} \int_{\mathbb{R}} \frac{|t|^p}{p!} |J(t)| dt \right)$$

with $\mathsf{C}_{E\bar{Z},j} := \left| \partial_j^p f(w) \int_{t\in\mathbb{R}} \frac{t^p}{p!} J(t)dt \right|$.

Finally, notice that by continuity of $\partial_j^p f$ (Assumption II.3), since $\mathcal{U}_n(w) \underset{n\to\infty}{\longrightarrow} \{w\}$:

$$\Omega_{nj} \underset{n\to\infty}{\longrightarrow} 0.$$

Thus for $n$ large enough:

$$\Omega_{nj} \left( \left| \int_{t\in\mathbb{R}} \frac{t^p}{p!} J(t)dt \right| \|K\|_1^{d-1} + \|K\|_1^{d-1} \int_{z_j\in\mathbb{R}} \frac{|z_j|^p}{p!} |J(z_j)| dz_j \right) \leq \frac{1}{2} \mathsf{C}_{E\bar{Z},j},$$

which leads to the result (II.12) of Lemma II.6:

$$\frac{1}{2} \mathsf{C}_{E\bar{Z},j} h_j^{p-1} \leq \left|\mathbb{E}\left[\bar{Z}_{hj}\right]\right| \leq \frac{3}{2} \mathsf{C}_{E\bar{Z},j} h_j^{p-1}.$$

To obtain the result (II.13) of Lemma II.6, just note that:

$$\mathbb{E}\left[|\bar{Z}_{h1j}|\right] = \frac{1}{h_j \prod_{k=1}^d h_k} \int_{u\in\mathbb{R}^d} \left| J\left(\tfrac{w_j - u_j}{h_j}\right) \left( \prod_{k\neq j} K\left(\tfrac{w_k - u_k}{h_k}\right) \right) \right| f(u)du$$
$$= h_j^{-1} \int_{z\in\mathbb{R}^d} \left| J(z_j) \left( \prod_{k\neq j} K(z_k) \right) \right| f(w - Hz)dz$$
$$\leq \mathsf{C}_{E|\bar{Z}|} h_j^{-1},$$

with $\mathsf{C}_{E|\bar{Z}|} := \|f\|_{\infty,\,\mathcal{U}_n(w)} \|J\|_1 \|K\|_1^{d-1}$.

2. We first bound $\bar{Z}_{hij}$ a.s. and its variance.

$$|\bar{Z}_{hij}| = \frac{\left|J(\frac{w_j - W_{ij}}{h_j})\right| \prod_{k\neq j}\left|K(\frac{w_k - W_{ik}}{h_k})\right|}{\mathrm{f}_X(X_i) h_j \prod_{k=1}^d h_k}$$
$$\leq \frac{\|J\|_\infty \|K\|_\infty^{d-1}}{\delta h_j \prod_{k=1}^d h_k} = \frac{\mathsf{C}_{\mathsf{M}\bar{Z}}}{h_j \prod_{k=1}^d h_k} =: \mathsf{M}_{\bar{Z},hj}. \tag{II.44}$$

For the variance:

$$\begin{aligned}\mathrm{Var}\left(\bar{Z}_{hij}\right) &\leq \mathbb{E}\left[\bar{Z}_{hij}^2\right] \\ &= \int_{\mathbb{R}^d} J(\tfrac{w_j - u_j}{h_j})^2 \left(\prod_{k\neq j} K(\tfrac{w_k - u_k}{h_k})^2\right) \frac{f_{XY}(u)}{\mathrm{f}_X(u_{1:d_1})^2 h_j^2 \prod_{k=1}^d h_k^2} du \\ &= \frac{1}{h_j^2 \prod_{k=1}^d h_k} \int_{\mathbb{R}^d} J(z_j)^2 \left(\prod_{k\neq j} K(z_k)^2\right) \frac{f(w - Hz)}{\mathrm{f}_X(x - (Hz)_{1:d_1})} dz \\ &\leq \frac{\|f\|_{\infty,\, \mathcal{U}_n(w)} \|J\|_2^2 \|K\|_2^{2(d-1)}}{\delta h_j^2 \prod_{k=1}^d h_k} = \frac{\mathsf{C}_{\mathsf{v}\bar{Z}}}{h_j^2 \prod_{k=1}^d h_k} =: \mathsf{v}_{\bar{Z},hj}. \end{aligned} \tag{II.45}$$

We apply Bernstein's inequality (cf Lemma II.10) to $\bar{Z}_{hj}$:

$$\mathbb{P}\left(\mathcal{B}^c_{\bar{Z},hj}\right) \leq 2\exp\left(-\min\left(\frac{n(\frac{\lambda_{hj}}{2})^2}{4\mathsf{v}_{\bar{Z},hj}}, \frac{3n\frac{\lambda_{hj}}{2}}{4\mathsf{M}_{\bar{Z},hj}}\right)\right).$$

Let us compare the rates:

$$\begin{aligned} & \frac{n(\frac{\lambda_{hj}}{2})^2}{4\mathsf{v}_{\bar{Z},hj}} \leq \frac{3n\frac{\lambda_{hj}}{2}}{4\mathsf{M}_{\bar{Z},hj}} \\ \iff \quad & \mathsf{C}_\lambda \sqrt{\frac{(\log n)^a}{n h_j^2 \prod_{k=1}^d h_k}} = \lambda_{hj} \leq \frac{6\mathsf{v}_{\bar{Z},hj}}{\mathsf{M}_{\bar{Z},hj}} = \frac{6\mathsf{C}_{\mathsf{v}\bar{Z}}}{\mathsf{C}_{\mathsf{M}\bar{Z}} h_j} \\ \iff \quad & \prod_{k=1}^d h_k \geq \frac{\mathsf{C}_{\mathsf{M}\bar{Z}}^2 \mathsf{C}_\lambda^2}{6^2 \mathsf{C}_{\mathsf{v}\bar{Z}}^2} \frac{(\log n)^a}{n} \\ \iff \quad & \mathsf{Cond}_{\bar{Z}}(h). \end{aligned}$$

So, if $\mathsf{Cond}_{\bar{Z}}(h)$ is satisfied:

$$\mathbb{P}\left(\mathcal{B}^c_{\bar{Z},hj}\right) \leq 2e^{\frac{-n(\lambda_{hj}/2)^2}{4\mathsf{v}_{\bar{Z},hj}}} = 2e^{\frac{-\delta}{\|f\|_{\infty,\, \mathcal{U}_n(w)}}(\log n)^a} = 2e^{-\gamma_{Z,n}}.$$

3. We apply Bernstein's inequality (cf Lemma II.10) to $\frac{1}{n}\sum_{i=1}^n |\bar{Z}_{hij}|$ using the upper bounds (II.44) and (II.45):

$$\mathbb{P}\left(\mathcal{B}^c_{|\bar{Z}|,h}\right) \leq 2\exp\left(-\min\left(\frac{n(\mathsf{C}_{E|\bar{Z}|} h_j^{-1})^2}{4\mathsf{v}_{\bar{Z},hj}}, \frac{3n\mathsf{C}_{E|\bar{Z}|} h_j^{-1}}{4\mathsf{M}_{\bar{Z},hj}}\right)\right).$$

Let us calculate the rate: by definition of $\mathsf{C}_{\mathsf{v}\bar{Z}}hj$ and $\mathsf{M}_{\bar{Z},hj}$,

$$\frac{n(\mathsf{C}_{E|\bar{Z}|}h_j^{-1})^2}{4\mathsf{v}_{\bar{Z},hj}} = \frac{\mathsf{C}_{E|\bar{Z}|}^2}{4\mathsf{C}_{\mathsf{v}\bar{Z}}} n \prod_{k=1}^{d} h_k$$

and

$$\frac{3n\mathsf{C}_{E|\bar{Z}|}h_j^{-1}}{4\mathsf{M}_{\bar{Z},hj}} = \frac{3\mathsf{C}_{E|\bar{Z}|}}{4\mathsf{C}_{\mathsf{M}\bar{Z}}} n \prod_{k=1}^{d} h_k.$$

Hence:

$$\mathbb{P}\left(\mathcal{B}^c_{|\bar{f}|h}\right) \leq 2e^{-\mathsf{C}_{\gamma|\bar{Z}|} n \prod_{k=1}^{d} h_k},$$

with $\mathsf{C}_{\gamma|\bar{Z}|} := \min\left(\frac{\mathsf{C}^2_{E|\bar{Z}|}}{4\mathsf{C}_{\mathsf{v}\bar{Z}}}; \frac{3\mathsf{C}_{E|\bar{Z}|}}{4\mathsf{C}_{\mathsf{M}\bar{Z}}}\right)$.

### 5.d.v Proof of Lemma II.7

1. We decompose $\Delta_{Z,hj}$ as follows:

$$\Delta_{Z,hj} := Z_{hj} - \bar{Z}_{hj} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\mathrm{f}_X(X_i) - \widetilde{\mathrm{f}}_X(X_i)}{\widetilde{\mathrm{f}}_X(X_i)}\right)\bar{Z}_{hij}.$$

Using $\bar{Z}_{hij} = 0$ when $X_i \notin \mathcal{U}_h(x)$:

$$\left|\Delta_{Z,hj}\right| \leq \left\|\frac{\mathrm{f}_X - \widetilde{\mathrm{f}}_X}{\widetilde{\mathrm{f}}_X}\right\|_{\infty,\,\mathcal{U}_n(x)} \frac{1}{n}\sum_{i=1}^{n}\left|\bar{Z}_{hij}\right|. \tag{II.46}$$

First we deal with $\left\|\frac{\mathrm{f}_X - \widetilde{\mathrm{f}}_X}{\widetilde{\mathrm{f}}_X}\right\|_{\infty,\,\mathcal{U}_n(x)}$. By definition of $\widetilde{A}_n$:

$$\mathbb{1}_{\widetilde{A}_n}\left\|\frac{\mathrm{f}_X - \widetilde{\mathrm{f}}_X}{\widetilde{\mathrm{f}}_X}\right\|_{\infty,\,\mathcal{U}_n(x)} \leq M_X\left(\frac{(\log n)^d}{n}\right)^{1/2} \quad \text{for } n \text{ large enough.} \tag{II.47}$$

Now let us give an upper bound of $\frac{1}{n}\sum_{i=1}^{n}\left|\bar{Z}_{hij}\right|$. Using Lemma II.6,

$$\begin{aligned}\mathbb{1}_{\mathcal{B}_{|\bar{Z}|,hj}}\frac{1}{n}\sum_{i=1}^{n}\left|\bar{Z}_{hij}\right| &\leq \mathbb{1}_{\mathcal{B}_{|\bar{Z}|,hj}}\left|\frac{1}{n}\sum_{i=1}^{n}\left|\bar{Z}_{hij}\right| - \mathbb{E}\left[\left|\bar{Z}_{h1j}\right|\right]\right| + \mathbb{E}\left[\left|\bar{Z}_{h1j}\right|\right] \\ &\leq 2\mathsf{C}_{E|\bar{Z}|}h_j^{-1}.\end{aligned}$$

To conclude, combining this last result with (II.47) and (II.46):

$$\begin{aligned}\mathbb{1}_{\mathcal{B}_{|\bar{Z}|,hj}\cap\widetilde{A}_n}\left|\Delta_{Z,hj}\right| &\leq 2\mathsf{C}_{E|\bar{Z}|}M_X h_j^{-1}\left(\frac{(\log n)^d}{n}\right)^{1/2} \\ &\leq \frac{2\mathsf{C}_{E|\bar{Z}|}M_X}{\mathsf{C}_\lambda(\log n)^{\frac{a}{2}}}\lambda_{hj} = \frac{\mathsf{C}_{\mathsf{M}\Delta Z}}{(\log n)^{\frac{a}{2}}}\lambda_{hj},\end{aligned}$$

since $\prod_{k=1}^{d} h_k \leq h_0^d = \frac{1}{(\log n)^d}$.

2. We decompose $\Delta_h$ as follows:

$$\Delta_h := \hat{f}_h(w) - \bar{f}_h(w) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\mathrm{f}_X(X_i) - \widetilde{\mathrm{f}}_X(X_i)}{\widetilde{\mathrm{f}}_X(X_i)}\right)\bar{f}_{hi}(w).$$

Using $\bar{f}_{hi}(w) = 0$ when $X_i \notin \mathcal{U}_h(x)$:

$$|\Delta_h| \leq \left\|\frac{\mathrm{f}_X - \widetilde{\mathrm{f}}_X}{\widetilde{\mathrm{f}}_X}\right\|_{\infty,\,\mathcal{U}_n(x)} \frac{1}{n}\sum_{i=1}^{n}\left|\bar{f}_{hi}(w)\right|.$$

We have proved in (II.47): $\mathbb{1}_{\widetilde{A}_n}\left\|\frac{\mathrm{f}_X - \widetilde{\mathrm{f}}_X}{\widetilde{\mathrm{f}}_X}\right\|_{\infty,\,\mathcal{U}_n(x)} \leq M_X\left(\frac{(\log n)^d}{n}\right)^{1/2}$.

Let us now give an upper bound of $\frac{1}{n}\sum_{i=1}^{n}\left|\bar{f}_{hi}(w)\right|$. Using Lemma II.5,

$$\begin{aligned}\mathbb{1}_{\mathcal{B}_{|\bar{f}|h}}\frac{1}{n}\sum_{i=1}^{n}\left|\bar{f}_{hi}(w)\right| &\leq \mathbb{1}_{\mathcal{B}_{|\bar{f}|h}}\left|\frac{1}{n}\sum_{i=1}^{n}\left|\bar{f}_{hi}(w)\right| - \mathbb{E}\left[\left|\bar{f}_{hi}(w)\right|\right]\right| + \mathbb{E}\left[\left|\bar{f}_{h1}(w)\right|\right]\\ &\leq 2\mathsf{C}_{\bar{\mathsf{E}}}.\end{aligned}$$

Therefore:

$$\mathbb{1}_{\widetilde{A}_n\cap\mathcal{B}_{|\bar{f}|h}}|\Delta_h| \leq 2\mathsf{C}_{\bar{\mathsf{E}}}M_X\left(\frac{(\log n)^d}{n}\right)^{1/2} \leq \frac{2\mathsf{C}_{\bar{\mathsf{E}}}M_X}{\mathsf{C}_\sigma}(\log n)^{-\frac{a}{2}}\sigma_h,$$

since $\prod_{k=1}^{d} h_k \leq h_0^d = (\log n)^{-d}$.

### 5.d.vi Proof of Lemma II.9

We first denote

$$B := \int_{\mathbb{R}^{d'}}\left(\prod_{j=1}^{d'} h_j^{-1}K\left(\tfrac{u_j - u'_j}{h_j}\right)\right)f_0(u')du' - f_0(u)\int_{\mathbb{R}^{d'}}\left(\prod_{j=1}^{d'}K\left(z_j\right)\right)dz.$$

Then we obtain by integration by parts:

$$B := \int_{z\in\mathbb{R}^{d'}}\left(\prod_{j=1}^{d'}K(z_j)\right)(f_0(u - h\cdot z) - f_0(u))dz. \tag{II.48}$$

For any $z \in \mathbb{R}^{d'}$, we denote $\bar{z}_0 := u$ and for $k = 1 : d'$, $\bar{z}_k := u - \sum_{j=1}^{k} h_j z_j e_j$ (where $\{e_j\}_{j=1}^{d'}$ is the canonical basis of $\mathbb{R}^{d'}$). Then, we write:

$$f_0(u - h\cdot z) - f_0(u) = \sum_{k=1}^{d'} f_0(\bar{z}_k) - f_0(\bar{z}_{k-1}). \tag{II.49}$$

Then we apply Taylor's theorem (cf Lemma II.11) to the functions $g_k : t \in [0,1] \mapsto f_0(\bar{z}_{k-1} - t h_k z_k e_k)$, $k \in (1 : d')$:

$$f_0(\bar{z}_k) - f_0(\bar{z}_{k-1}) = g_k(1) - g_k(0) = \sum_{l=1}^{p}\frac{(-z_k h_k)^l}{l!}\partial_k^l f_0(\bar{z}_{k-1}) + \rho_k,$$

where we denote for short:

$$\rho_k := \rho_k(z,h,u) = (-h_k z_k)^p \int\limits_{0\le t_p\le\cdots\le t_1\le 1} \left(\partial_k^p f_0(\bar z_{k-1} - t_p h_k z_k e_k) - \partial_k^p f_0(\bar z_{k-1})\right) dt_{1:p}. \tag{II.50}$$

We introduce the notation

$$\mathsf{I}_k := \int_{z\in\mathbb{R}^{d'}} \left(\prod_{j=1}^{d'} K(z_j)\right) \rho_k dz$$

and for any $z \in \mathbb{R}^{d'}$, we denote $z_{-k} \in \mathbb{R}^{d'-1}$ the vector $z$ without its $k^{th}$ variable, then (II.48) becomes:

$$\begin{aligned} B &= \int_{z\in\mathbb{R}^{d'}} \left(\prod_{j=1}^{d'} K(z_j)\right) \left(\sum_{k=1}^{d'}\sum_{l=1}^{p} \frac{(-h_k)^l}{l!} z_k^l \partial_k^l f_0(\bar z_{k-1}) + \rho_k\right) dz \\ &= \sum_{k=1}^{d'} \left(\mathsf{I}_k + \sum_{l=1}^{p} \frac{(-h_k)^l}{l!} \int_{z_{-k}\in\mathbb{R}^{d-1}} \partial_k^l f_0(\bar z_{k-1}) \left(\prod_{j\neq k} K(z_j)\right) \int_{z_k\in\mathbb{R}} z_k^l K(z_k) dz_k dz_{-k}\right). \end{aligned}$$

Since $K$ has at least $p-1$ zero moments, the terms with $l \le p-1$ vanish, leading to:

$$\begin{aligned} B &= \sum_{k=1}^{d'} \left(\mathsf{I}_k + \frac{(-h_k)^p \int_{t\in\mathbb{R}} t^p K(t) dt}{p!} \int_{z_{-k}\in\mathbb{R}^{d'-1}} \partial_k^p f_0(\bar z_{k-1}) \left(\prod_{j\neq k} K(z_j)\right) dz_{-k}\right) \\ &=: \sum_{k=1}^{d'} (\mathsf{I}_k + \mathsf{II}_k), \end{aligned} \tag{II.51}$$

with $\mathsf{II}_k := (-h_k)^p \int_{t\in\mathbb{R}} \frac{t^p}{p!} K(t) dt \int_{z_{-k}\in\mathbb{R}^{d'-1}} \partial_k^p f_0(\bar z_{k-1}) \left(\prod_{j\neq k} K(z_j)\right) dz_{-k}$.

# Chapter III

# Adaptation to the smoothness

This chapter is a paper in preparation [Nguyen et al. 2019].

**Abstract :** • *We estimate the conditional density $f(x,\cdot)$ of $Y_i$ given $X_i = x$, from the observation of an i.i.d. sample $(X_i, Y_i) \in \mathbb{R}^d$, $i = 1,\dots,n$. We assume that $f$ depends only on $r$ unknown components with typically $r \ll d$. We provide an adaptive fully-nonparametric strategy based on kernel rules to estimate $f$. To select the bandwidth of our kernel rule, we propose a new fast iterative algorithm inspired by the Rodeo algorithm [Wasserman and Lafferty 2006] to detect the sparsity structure of $f$. More precisely, in the minimax setting, our pointwise estimator, which is adaptive to both the regularity and the sparsity, achieves the quasi-optimal rate of convergence. Its computational complexity is only $O(dn\log n)$.*

***Keywords:*** *conditional density, high dimension, minimax rates, kernel density estimators, greedy algorithm, sparsity, nonparametric inference.* •

## Table of contents

# 1 Introduction

## 1.a Motivations

Consider $W = (W_1, \dots, W_n)$ a sample of a couple $(X, Y)$ of multivariate random vectors: for $i = 1, \dots, n$,

$$W_i = (X_i, Y_i),$$

with $X_i$ valued in $\mathbb{R}^{d_1}$ and $Y_i$ in $\mathbb{R}^{d_2}$. We denote $d := d_1 + d_2$ the joint dimension. We assume that the marginal distribution of $X$ and the conditional distribution of $Y$ given $X$ are absolutely continuous with respect to the Lebesgue measure, and we denote by $\mathrm{f}_X$ the marginal density of $X$. Let us define $f : \mathbb{R}^d \to \mathbb{R}_+$ such that for any $x \in \mathbb{R}^{d_1}$, $f(x, \cdot)$ is the conditional density of $Y$ conditionally to $X = x$:

$$f(x, y)dy = d\mathbb{P}_{Y|X=x}(y).$$

We aim at estimating the conditional density $f$ at a set point $w = (x, y)$ in $\mathbb{R}^d$.

The issue of estimating a conditional density may arise as soon as we observe a (possibly multidimensional) response $Y$ associated with a (possibly multidimensional) covariate $X$. We often study the regression function $\mathbb{E}(Y|X = x)$, but this information is restrictive, and the entire distribution is more informative than the mean (think in particular to the case of an asymetric or multimodal distribution). Thus the problem of estimating the conditional distribution is found in various application fields: Meteorology, Insurance, Medical studies, Geology, Astronomy. See Nguyen [2018] and references therein. Moreover, the ABC methods (Approximate Bayesian Computation) are actually dedicated to find a conditional distribution (of the parameter given observations) in the case where the likelihood is not computable but simulable: see Izbicki et al. [2018] (and references therein) where the link between conditional density estimation and ABC is studied.

Several nonparametric methods have been proposed for estimating a conditional density: Hyndman et al. [1996] and Fan et al. [1996] have improved the seminal Nadaraya-Watson-type estimator of Rosenblatt [1969] and Lincheng and Zhijun [1985], as well as De Gooijer and Zerom [2003] who introduced another weighted kernel estimator. For these kernel estimators, different methods have been advocated to tackle the bandwidth selection issue: bootstrap approach [Bashtannyk and Hyndman 2001] or cross-validation variants [Fan and Yim 2004 ; Holmes et al. 2010 ; Ichimura and Fukuda 2010]. Later, adaptive-in-smoothness estimators have been introduced: Brunel et al. [2007] with piecewise polynomial representation, Chagny [2013] with wraped base method, Le Pennec and Cohen [2013] with penalized maximum likelihood estimator, Bertin et al. [2016] with Lepski-type method, Sart [2017] with tests-based histograms.

All above references do not really deal with the curse of dimensionality. From a theoretical point of view, the minimax rate of convergence for such nonparametric statistical problems is known to be $n^{-s/(2s+d)}$ (possibly up to a logarithmic term), where $s$ is the smoothness of the target function. This illustrates that estimation gets increasingly hard when $d$ is large. Moreover the computational complexity of above methods is often intractable as soon as $d$ is larger than 3 or 4. A first answer to overcome this limitation is to consider the single-index model, as Fan et al. [2009] or Bouaziz and Lopez [2010], but this implies a strong structural assumption. A more general advance has been made by Hall et al. [2004] who assume that some components of $X$ can be irrelevant, i.e. that they contain no information about $Y$ and should be dropped before conducting inference. Their cross-validation approach allows them to obtain a minimax rate for a $r_1$-dimensional $C^2$ function, where $r_1$ is the number of relevant $X$-components. Efromovich [2010a] has improved these non-adaptive results by using thresholding and Fourier series and achieves the minimax rate $n^{-s/(2s+r_1)}$ without any knowledge of $r_1$ nor $s$. Note that above rates were established for the $\mathbb{L}^2$-loss whereas we shall consider the pointwise loss. Moreover these combinatorial

approaches make their computation cost prohibitive when both $n$ and $d$ are large. In the same framework, Shiga et al. [2015] assume that the dependence of $Y$ on the relevant components is additive. Another way is paved by Otneim and Tjøstheim [2018] who estimate the dependence structure in a Gaussian parametric way while estimating marginal distributions nonparametrically. More recently, Izbicki and Lee [2016; 2017] have proposed two attractive methodologies using orthogonal series estimators in the context of an eventual smaller unknown intrinsic dimension of the support of the conditional density. In particular, the Flexcode method originally proposes to transfer successful procedures for high dimensional regression to the conditional density estimation setting by interpreting the coefficients of the orthogonal series estimator as regression functions, which allows to adapt to data with different features (mixed data, smaller intrinsic dimension, relevant variables) in function of the regression method. However, the optimal tuning parameters depend in fact on the unknown intrinsic dimension. Furthermore, optimal minimax rates are not achieved, revealing the specific nature of the problem of conditional density estimation, more intricate, in full generality, than regression.

## 1.b Objectives, methodology and contributions

We consider the estimation of a sparse conditional density $f$ by assuming that only $r \in [0,d]$ components are *relevant*, i.e. that there exists a subset $\mathcal{R} \subset \{1,\dots,d\}$ with cardinal $r$, such that for any fixed $\{z_j\}_{j\in\mathcal{R}}$, the function $\{z_k\}_{k\in\mathcal{R}^c} \mapsto f(z_1,\dots,z_d)$ is constant on the neighborhood of $w$, with $\mathcal{R}^c = \{1,\dots,d\} \setminus \mathcal{R}$. Assuming that $f$ is $s$-Hölderian, our goal is to provide an estimation procedure such that it achieves the best adaptive rate. The meaning of *adaptation* is twofold in this chapter: The first meaning corresponds to adaptation with respect to the smoothness, which is the classical meaning of adaptation. The second one corresponds to adaptation with respect to the sparsity. So our goal is to propose an optimal procedure in this context, meaning that it does not depend on the knowledge of $s$ and $r$. Furthermore, for practical purposes in moderate large dimensions, it should be implemented with low computational time.

For this purpose, we consider a particular kernel estimator depending on a bandwidth $h \in \mathbb{R}_+^d$ to be selected. To circumvent the curse of dimensionality, we consider an iterative algorithm on a special path of bandwidths inspired by the *RODEO* procedures proposed by Lafferty and Wasserman [2008] for nonparametric regression, Liu et al. [2007] for density estimation and Nguyen [2018] for conditional density estimation. More precisely, our new procedure, called RevDir CDRODEO, is a variation of the CDRODEO proposed by Nguyen [2018] (and called Direct CDRODEO in the sequel). Each iteration step of this new algorithm is based on comparisons between partial derivatives of our kernel rule, denoted $Z_{hj}$, and specific thresholds $\lambda_{hj}$, respectively defined in (III.3) and (III.5). Let us mention that for variable selection in the regression model with very high ambient dimension, Comminges and Dalalyan [2012] used similar ideas to select the relevant variables by comparing some quadratic functionals of empirical Fourier coefficients to prescribed significance levels. Consistency of this (non-greedy) procedure is established by Comminges and Dalalyan [2012].

We establish that, up to a logarithmic term whose exponent is positive but as close to 0 as desired, RevDir CDRODEO achieves the rate $((\log n)/n)^{s/(2s+r)}$, which is the optimal adaptive minimax rate on Hölder balls $\mathcal{H}_d(s,L)$, when the conditional density depends on $r$ components. When $r$ is much smaller than $d$, this rate is much faster than the usual rate $((\log n)/n)^{s/(2s+d)}$ achieved by classical kernel rules. Furthermore, unlike previous RODEO-type procedures, our procedure is adaptive with respect to both the smoothness and the sparsity. To the best of our knowledge, our RevDir CDRODEO procedure is the first algorithm achieving quasi-minimax rates for conditional density estimation in this setting where both sparsity and smoothness are unknown. Furthermore,

tuning RevDir CDRODEO is very easy (see Section 3.b) and we show that the total worst-case complexity of RevDir CDRODEO algorithm is only $O(dn\log n)$. This last result is very important for modern statistics where many problems deal with very large datasets.

### 1.c Plan of the chapter and notations

The plan of the chapter is the following. First we describe in Section 2 the estimation procedure. We give heuristic ideas based on the oracle approach and explain why some modifications of the Direct CDRODEO procedure are necessary. Then a detailed presentation of our algorithm is provided in Section 2.b.iv. Next the main result is stated in Section 3. The complexity of the algorithm is computed in Section 3.d. The proofs are gathered in Section 4.

In the sequel, we denote by $\star$ the convolution product. For a function $g : (u_1,\dots,u_d) \mapsto g(u_1,\dots,u_d)$, we denote $\partial_j g$ the partial derivative $\frac{\partial}{\partial u_j} g$ when there is no ambiguity. We introduce the following partial order on the bandwidths:

$$h \preceq h' \quad \Leftrightarrow \quad \forall k \in \{1,\dots,d\} \quad h_k \leq h'_k.$$

## 2 Estimation procedure

### 2.a Kernel rule

Our estimation procedure of the conditional density $f$ is based on a kernel rule, namely the kernel estimator introduced in [Bertin et al. 2016]. So, let $K : \mathbb{R} \to \mathbb{R}$ be a kernel function, namely $K$ satisfies $\int_{\mathbb{R}} K(t)dt = 1$. Then, for any bandwidth $h = (h_j)_{j=1,\dots,d} \in (\mathbb{R}_+^*)^d$, the estimator of $f$ associated with $K$ and $h$ is defined for any $w \in \mathbb{R}^d$, by

$$\hat{f}_h(w) := \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\widetilde{\mathrm{f}}_X(X_i)} \mathrm{K}_h(w - W_i), \tag{III.1}$$

where for any $v \in \mathbb{R}^d$,

$$\mathrm{K}_h(v) = \prod_{j=1}^{d} h_j^{-1} K(v_j/h_j)$$

and $\widetilde{\mathrm{f}}_X$ is an estimator of $\mathrm{f}_X$, built from a sample $\tilde{X}$ not necessarily independent of $W$.

**Remark III.1.** *Note that (non conditional) density estimation is a special case of our problem. It corresponds to the setting where $d_1 = 0$ and $\mathrm{f}_X \equiv 1$ ($\equiv \widetilde{\mathrm{f}}_X$). In this case, $\hat{f}_h(w)$ is the classical kernel density estimator extensively studied in the literature.*

Since $f$ can be expressed as the ratio

$$f(x,y) = \frac{f_{XY}(x,y)}{\mathrm{f}_X(x)},$$

the class of rules defined as the ratio of two density estimates has intensively been studied. The estimate $\hat{f}_h(w)$ does not belong to this class. Actually, our goal is to take into account the specific nature of the conditional density $f$, not the nature of $f_{XY}$ and $\mathrm{f}_X$. In particular, a relevant component both for the joint density $f_{XY}$ and the marginal density $\mathrm{f}_X$ may be irrelevant for the

conditional density; this occurs if a component of $X$ is independent of $Y$ and in this case relevance may be not detected by a ratio of two density estimates. Similarly, the smoothness of $f$ can be different from the smoothness of the functions $f_{XY}$ and $\mathrm{f}_X$. Remark that if we could take $\widetilde{\mathrm{f}}_X = \mathrm{f}_X$, then

$$\mathbb{E}[\hat{f}_h(w)] = \iint \frac{1}{\mathrm{f}_X(u)} \mathrm{K}_h(w-(u,v)) f_{XY}(u,v)dudv = \int \mathrm{K}_h(w-z)f(z)dz = (\mathrm{K}_h \star f)(w), \qquad \text{(III.2)}$$

which ensures that $\mathbb{E}[\hat{f}_h(w)]$ is a good approximation of $f$ when $h$ is small enough under mild assumptions on $K$ and $f$. These arguments justify the introduction of $\hat{f}_h(w)$. The choice of $\widetilde{\mathrm{f}}_X$ is essential and will be discussed in Section 3.c. Equality (III.2) shows that the selection of $h$ will be essentially dictated by the intrinsic properties of the conditional density $f$.

Now, as explained in Introduction, the principal issue is to choose an appropriate bandwidth $h$ which adapts simultaneously to the unknown sparsity and smoothness of $f$. In particular, large values of the components of the bandwidths will correspond to irrelevant components of $f$, namely $\mathcal{R}^c = \{1,\dots,d\} \setminus \mathcal{R}$. Several estimation kernel procedures based on optimization over an exhaustive grid of bandwidths have been proposed in the literature. But the larger the class of bandwidths, the larger the computational time. So, most of them have to face with large running times, leading to intractable procedures, even for moderately large dimensions. Furthermore, as explained in Introduction, very few are able to deal with the two-fold adaptive objective.

These are the reasons why, unlike classical methods involving criteria minimization over a large class of smoothing parameters, we propose an algorithm generating an iterative smooth path through the set of bandwidths in the same spirit as Wasserman and Lafferty [2006] and Lafferty and Wasserman [2008] for nonparametric regression, Liu et al. [2007] for density estimation and Nguyen [2018] for non-adaptive conditional density estimation. The greediness of our procedure, which is presented in the next paragraph, leans on the selection of this path of bandwidths. It enables us to address adaptive conditional density estimation in high dimensions.

## 2.b From the Direct CDRODEO procedure to the RevDir CDRODEO procedure

In the sequel, to describe our algorithm, we fix $w = (x,y)$, the estimation point, and we assume that $K$ is of class $\mathcal{C}^1$.

### 2.b.i The Direct CDRODEO procedure

To select the bandwidth, we would like to use local variations of $f$. Indeed, heuristically, the larger the local variations of $f$, the smaller the bandwidth. So, we naturally rely on partial derivatives of $f$, which are, of course, not observed. So, as a proxy of $\frac{\partial}{\partial w_j} f$, we consider $Z_{hj}$, the partial derivatives of the estimator with respect to the components of the bandwidths, defined for $h \in (\mathbb{R}_+^*)^d$ and $j \in \{1,\dots,d\}$ by:

$$Z_{hj} := \frac{\partial}{\partial h_j} \hat{f}_h(w). \qquad \text{(III.3)}$$

Denoting $J : t \mapsto K(t) + tK'(t)$, $Z_{hj}$ can be easily expressed, which constitutes a key step to obtain algorithms with low computational time. We obtain:

$$Z_{hj} = \frac{-1}{nh_j^2} \sum_{i=1}^{n} \frac{1}{\widetilde{\mathrm{f}}_X(X_i)} J\left(\tfrac{w_j - W_{ij}}{h_j}\right) \prod_{k \neq j}^{d} h_k^{-1} K\left(\frac{w_k - W_{ik}}{h_k}\right). \qquad \text{(III.4)}$$

The CDRODEO procedure proposed by Nguyen [2018], called the *Direct* CDRODEO procedure in the sequel, involves the $Z_{hj}$'s as follows:

1. We start from a bandwidth $h = (h_1, \ldots, h_d)$ whose components $h_j$ are all equal to $h_0 > 0$ quite large (typically, $h_0$ is close to 1).
2. At each step, for all $j$, if $j$ is not deactivated, we compare $|Z_{hj}|$ to a threshold $\lambda_{hj}$, where

$$\lambda_{hj} := \mathsf{C}_\lambda \sqrt{\frac{(\log n)^a}{n h_j^2 \prod_{k=1}^d h_k}}, \tag{III.5}$$

   with $\mathsf{C}_\lambda = 4\|J\|_2 \|K\|_2^{d-1}$ and $a > 1$ a tuning parameter. Observe that $\lambda_{hj}^2$ is a good proxy of $\mathrm{Var}(Z_{hj})$ up to the logarithmic term.
   - If $|Z_{hj}| > \lambda_{hj}$, then $h_j \leftarrow \beta h_j$ for $\beta \in (0,1)$ a constant fixed in advance, and $j$ is still active.
   - If $|Z_{hj}| \leq \lambda_{hj}$, $j$ is deactivated and $h_j$ remains unchanged for the next steps of the path.
3. We stop when all components are deactivated or if $\prod_{j=1}^n h_j < \frac{\log n}{n}$.

The next paragraph provides heuristic arguments explaining why such an algorithm is able, simultaneously, to detect irrelevant components and provide suitable bandwidths for relevant components.

#### 2.b.ii Heuristic arguments

Introducing

$$\bar{Z}_{hj} = \frac{-1}{n h_j^2} \sum_{i=1}^n \frac{1}{\mathrm{f}_X(X_i)} J\left(\tfrac{w_j - W_{ij}}{h_j}\right) \prod_{k \neq j}^d h_k^{-1} K\left(\frac{w_k - W_{ik}}{h_k}\right), \tag{III.6}$$

which is close to $Z_{hj}$ if $\widetilde{\mathrm{f}}_X$ is a good estimate of $\mathrm{f}_X$, we easily obtain that $\mathbb{E}[\bar{Z}_{hj}] = 0$ if $j \in \mathcal{R}^c$, which means that, with high probability, $j$ is rapidly deactivated by the Direct CDRODEO procedure. Indeed, $\lambda_{hj}$ is tuned (via the Bernstein concentration inequality) so that with high probability, $|\bar{Z}_{hj} - \mathbb{E}[\bar{Z}_{hj}]| \leq \lambda_{hj}$. We then obtain large smoothing parameters for irrelevant components.

To explain heuristically why the Direct CDRODEO procedure is suitable for relevant components, we use the oracle approach. For the sake of simplicity, we assume that $\widetilde{\mathrm{f}}_X = \mathrm{f}_X$. Given a bandwidth $h$, we have:

$$\mathbb{E}[(\hat{f}_h(w) - f(w))^2] = B^2(h) + \mathrm{Var}(\hat{f}_h(w)),$$

where $B(h) := \mathbb{E}[\hat{f}_h(w)] - f(w)$ is the bias term and

$$\mathrm{Var}(\hat{f}_h(w)) = \frac{1}{n}\mathrm{Var}\left(\frac{\mathrm{K}_h(w - W_1)}{\mathrm{f}_X(X_1)}\right) \approx \frac{1}{n}\|\mathrm{K}_h\|^2 \approx \frac{1}{n} \times \prod_{j=1}^d \frac{1}{h_j}, \tag{III.7}$$

where previous approximations are justified if $f$ is bounded from above and $\mathrm{f}_X$ bounded from below in the neighborhood of $w$. Then, the ideal bandwidth should be a global minimizer of the function

$$h \mapsto \tilde{R}(h) := B^2(h) + \frac{1}{n} \times \prod_{j=1}^d \frac{1}{h_j}.$$

Denoting $h^*$ such a global minimizer, we assume that the sign of $B$ is constant in the neighborhood of $h^*$. Without loss of generality, we then assume that $B$ is positive in the neighborhood of $h^*$. So $h^*$ will be a minimizer of

$$h \mapsto R(h) := B(h) + \frac{1}{\sqrt{n \times \prod_{j=1}^d h_j}}. \tag{III.8}$$

Then, if $B$ is of class $\mathcal{C}^1$, $h^*$ should satisfy for any $j$,

$$\frac{\partial}{\partial h_j} B(h^*) = \frac{1}{2}\sqrt{\frac{1}{n(h_j^*)^2 \prod_{k=1}^d h_k^*}}.$$

Ideally, a good algorithm would select a bandwidth satisfying this property. Of course, partial derivatives of the bias are unknown but for any $h$, under mild assumptions,

$$\frac{\partial}{\partial h_j} B(h) = \mathbb{E}\left[\frac{\partial}{\partial h_j}\hat{f}_h(w)\right] = \mathbb{E}[Z_{hj}],$$

so $Z_{hj}$ is an unbiased estimate of $\frac{\partial}{\partial h_j} B(h)$. Finally, heuristically, an ideal bandwidth should satisfy

$$Z_{h^* j} \approx \sqrt{\frac{1}{n(h_j^*)^2 \prod_{k=1}^d h_k^*}},$$

which is the case for the Direct CDRODEO procedure up to a logarithmic term, since CDRODEO stops as soon as $|Z_{hj}| = \lambda_{hj}$ (observe that similar arguments can be used if $B$ remains negative in the neighborhood of $h^*$ and in this case, we have to replace $Z_{hj}$ with $-Z_{hj}$).

Note that if previous arguments are only heuristic ones, several issues can be pointed out:

1. Some singular points of the risk function $R$ (defined in (III.8)) can correspond to non-global minimizers. In particular, the larger the distance between the initial bandwidth of the algorithm and the minimizer of $R$, the larger the probability to stop at a local minimizer of $R$. To circumvent this problem, we can take a small value for $h_0$. But taking a too small value for $h_0$ may be inappropriate for irrelevant components.
2. If $\text{card}(\mathcal{R})$ is large, many components of the optimal bandwidth are small, which leads to many steps of the *Direct* CDRODEO procedure and then to a larger computational cost.

The first point shows that initialization appears as a key point of the algorithm. In view of these issues, it is natural to consider some variations of the *Direct* CDRODEO procedure. They are described in the next paragraph.

#### 2.b.iii The Reverse CDRODEO procedure

The first variation which could be considered is the *Reverse* CDRODEO procedure in the same spirit as Liu et al. [2007] (see Section 4.2 therein). We start with a small bandwidth and use a sequence of non-decreasing bandwidths to select the optimal value, still by comparing the $Z_{hj}$'s with the $\lambda_{hj}$'s. As illustrated by Liu et al. [2007], this approach is very useful for image data. However, the choice of the initial bandwidth is very sensitive. In particular, assume that $f$ has a very low regularity and has only one relevant component, say the first one for instance. In this case, if $h^*$ is the ideal bandwidth, $h_1^* = 1/n$ (up to a logarithmic term). So, since $\mathcal{R}$ is unknown, the initialization of the bandwidth must be not larger than $h_{0,\text{rev}} = (1/n, \ldots, 1/n)$. However, such a small bandwidth leads to instability problem. In particular, the variance of $\hat{f}_{h_{0,\text{rev}}}(w)$ is of order $n^{d-1}$ (see Equation (III.7)).

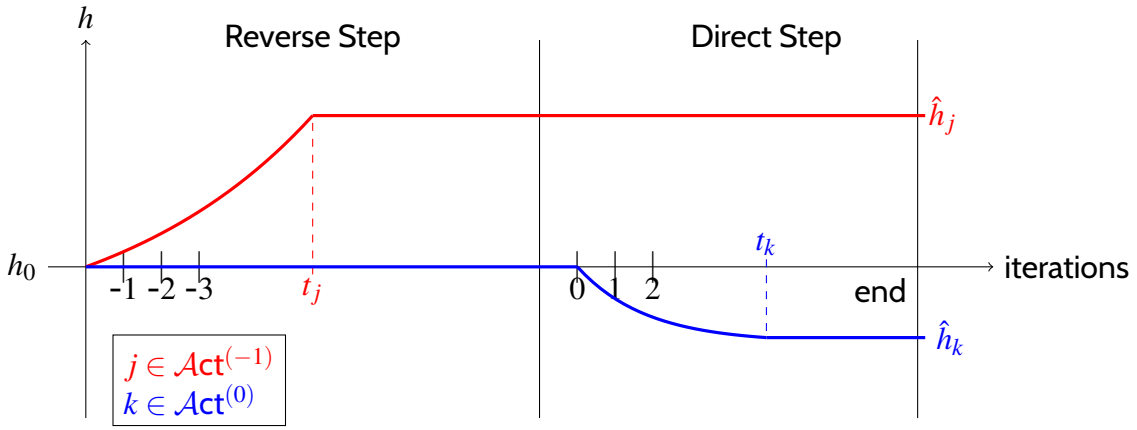



Figure III.1 – The two patterns of bandwidth path: in red for a component $j \in \mathcal{A}\text{ct}^{(-1)}$ with a deactivation time $t_j \leq 0$, and in blue for a component $k \in \mathcal{A}\text{ct}^{(0)}$ with a deactivation time $t_k \geq 0$.

#### 2.b.iv Our procedure: The RevDir CDRODEO procedure

Previous arguments show that to circumvent previous issues, we have to combine Direct and Reverse CDRODEO procedures, leading to the *RevDir* CDRODEO procedure. This new procedure, precisely described by Algorithm 2, comprises two steps after fixing the initial bandwidth whose components are all equal to $h_0$, where $h_0$ is assumed to be larger than all relevant components of the optimal bandwidth.

1. The first step is the Reverse CDRODEO algorithm with a sequence of non-decreasing bandwidths to estimate $\mathcal{R}^c$.
2. The second step, which concerns only components $j$ such that after the Reverse Step $h_j = h_0$, is the Direct CDRODEO algorithm. Its goal is to deal with components associated with $\mathcal{R}$.

The output bandwidth of the algorithm is denoted $\hat{h}$. The function $f$ is finally estimated by $\hat{f} := \hat{f}_{\hat{h}}$. Figure III.1 illustrates the two kinds of path for the bandwidth components. If the component belongs to $\mathcal{A}\text{ct}^{(-1)}$ (resp. $\mathcal{A}\text{ct}^{(0)}$), it is deactivated during the Reverse Step (resp. the Direct Step) and is larger (resp. smaller) than the initial bandwidth value $h_0$. Note that the RevDir procedure generalizes both the Direct and Reverse procedures in function of the choice of $h_0$. Indeed, if we set $h_0 = 1$, the RevDir procedure behaves as a Direct procedure with the same initialization. Conversely, setting $h_0 = 1/n$ brings us back on the Reverse procedure. Nonetheless, note that the tuning of $h_0$, as well as of the parameters $a$ and $\beta$, needs a careful attention, which is discussed in the next section.

# 3 Theoretical results

## 3.a Sparsity and smoothness classes of functions

This section is devoted to the theoretical results satisfied by the RevDir CDRODEO procedure. We consider a kernel function $K : \mathbb{R} \to \mathbb{R}$ of class $\mathcal{C}^1$, with compact support denoted $\text{supp}(K)$. We shall also assume that $K$ is of order $p$, *i.e.*: for $\ell = 1, \ldots, p-1$, $\int_{\mathbb{R}} t^\ell K(t)dt = 0$. Taking a kernel of order $p$ is usual for the control of the bias of the estimator. Then, we define the neighborhood $\mathcal{U}$

**Algorithm 2** RevDir CDRODEO algorithm

1. *Input:* the estimation point $w$, the observations $W$, the bandwidth decreasing factor $\beta \in (0,1)$, the bandwidth initialization value $h_0 > 0$, a tuning parameter $a > 1$.
2. *Initialization:*
   - ▷ Initialize the trial bandwidth: *for* $k = 1:d$, $H_k^{(0)} \leftarrow h_0$.
   - ▷ Determine which variables are active for the Reverse Step or for the Direct Step:
     $\mathcal{A}\text{ct}^{(-1)} \leftarrow \{k = 1:d, |Z_{H^{(0)}k}| \leq \lambda_{H^{(0)}k}\}$
     $\mathcal{A}\text{ct}^{(0)} \leftarrow \{1:d\} \setminus \mathcal{A}\text{ct}^{(-1)}$
3. *Reverse Step:*
   - ▷ Initialize the counter: $t \leftarrow -1$
   - ▷ Initialize the current bandwidth: $\hat{h}^{(-1)} \leftarrow H^{(0)}$
   - ▷ While $(\mathcal{A}\text{ct}^{(t)} \neq \emptyset)$ & $(\max \hat{h}_k^{(t)} \leq \beta)$ :
     - ▶ Set the current trial bandwidth: $H_k^{(t)} = \begin{cases} \beta^{-1}\hat{h}_k^{(t)} & \text{if } k \in \mathcal{A}\text{ct}^{(t)} \\ \hat{h}_k^{(t)} & \text{else.} \end{cases}$
     - ▶ Set the next active set: $\mathcal{A}\text{ct}^{(t-1)} \leftarrow \{k \in \mathcal{A}\text{ct}^{(t)}, |Z_{H^{(t)}k}| \leq \lambda_{H^{(t)}k}\}$
     - ▶ Update the current bandwidth: $\hat{h}_k^{(t)} \leftarrow \begin{cases} H_k^{(t)} & \text{if } k \in \mathcal{A}\text{ct}^{(t-1)} \\ \hat{h}_k^{(t)} & \text{else.} \end{cases}$
     - ▶ Initialize the next bandwidth: $\hat{h}^{(t-1)} \leftarrow \hat{h}^{(t)}$
     - ▶ Decrement the counter: $t \leftarrow t - 1$
4. *Direct Step:*
   - ▷ Initialize the current bandwidth: $\hat{h}^{(0)} \leftarrow \hat{h}^{(t)}$
   - ▷ Reinitialize the counter: $t \leftarrow 0$
   - ▷ While $\left(\mathcal{A}\text{ct}^{(t)} \neq \emptyset\right)$ & $\left(\prod_{k=1}^{d} \hat{h}_k^{(t)} \geq \frac{(\log n)^{1+a}}{n}\right)$:
     - ▶ Increment the counter: $t \leftarrow t + 1$
     - ▶ Set the current active set: $\mathcal{A}\text{ct}^{(t)} \leftarrow \{k \in \mathcal{A}\text{ct}^{(t-1)}, |Z_{\hat{h}^{(t-1)}k}| > \lambda_{\hat{h}^{(t-1)}k}\}$
     - ▶ Set the current bandwidth: $\hat{h}_k^{(t)} \leftarrow \begin{cases} \beta.\hat{h}_k^{(t-1)} & \text{if } k \in \mathcal{A}\text{ct}^{(t)} \\ \hat{h}_k^{(t-1)} & \text{else.} \end{cases}$
5. *Output:* $\hat{h} \leftarrow \hat{h}^{(t)}$ (and compute $\hat{f}_{\hat{h}}(w)$).

of the point $w \in \mathbb{R}^d$ as follows:

$$\mathcal{U} := \left\{u \in \mathbb{R}^d : w - u \in (\text{supp}(K))^d\right\}.$$

In the sequel, we denote

$$\|f\|_{\infty,\mathcal{U}} := \sup_{x \in \mathcal{U}} |f(x)|.$$

**Remark III.2.** *The size of $\mathcal{U}$ is fixed. But $\mathcal{U}$ could be chosen so that its size goes to 0. In this case, we have to modify the stopping rule of the Reverse Step, namely $\max \hat{h}_k^{(t)} \leq \beta$, to force $\max \hat{h}_k^{(t)} \overset{n\to\infty}{\longrightarrow} 0$. For instance, if we impose $\max \hat{h}_k^{(t)} \leq \frac{1}{\log n}$, the rates of convergence of our estimate would typically be deteriorated by logarithmic factors.*

The notion of relevant components has already been introduced in Section 1.b but subsequent results only need that the function $f$ is locally sparse, so we shall consider the following definition depending on $w = (x, y)$ and $\mathcal{U}$.

**Definition III.1.** *We denote $\mathcal{R}$ the subset of $\{1, \dots, d\}$ with cardinal $r$ such that for any fixed $\{z_j\}_{j\in\mathcal{R}}$, the function $\{z_k\}_{k\in\mathcal{R}^c} \mapsto f(z_1, \dots, z_d)$ is constant on $\mathcal{U}$. We call relevant any component in $\mathcal{R}$.*

The previous definition means that on $\mathcal{U}$, $f$ depends only on $r$ of its $d$ variables. In the sequel, we consider the minimax point of view and we derive rates on Hölder balls defined as follows.

**Definition III.2.** *Let $L > 0$ and $s > 0$. We say that the conditional density $f$ belongs to the Hölder ball of smoothness $s$ and radius $L$, denoted $\mathcal{H}_d(s, L)$, if $f$ is of class $\mathcal{C}^q$ and if it satisfies for all $z \in \mathcal{U}$ and for all $t \in \mathbb{R}$ such that $z + te_k \in \mathcal{U}$*

$$\left|\partial_k^q f(z + te_k) - \partial_k^q f(z)\right| \leq L|t|^{s-q},$$

*where $q = \lceil s - 1 \rceil = \max\{l \in \mathbb{N} : l < s\}$ and $e_k$ is the vector where all coordinates are null except the $k$th one which is equal to 1.*

In the sequel, we investigate adaptive results in terms of sparsity and smoothness properties on Hölder balls $\mathcal{H}_d(s, L)$. It means that our procedure will not depend on the knowledge of $\mathcal{R}$ nor $(s, L)$.

## 3.b Tuning the RevDir CDRODEO procedure

The RevDir CDRODEO procedure depends on three tuning parameters, namely $h_0$, $\beta$ and $a$.

In the sequel, we take $\beta \in (0, 1)$. Its value has no influence on rates of convergence. But of course, the larger $\beta$, the more accurate the procedure, but the larger the computational time. In practice, we set $\beta$ close to 1.

The parameter $a$ will be assumed to be larger than 1. Its value does not affect the polynomial rate of convergence but the smaller $a$, the smaller the exponent of the logarithmic factor of the rate. In practice, $a$ will be larger but close to 1.

Finally, to initialize the procedure, we take $h_0$ such that

$$C_\lambda^{2/d} \left(\frac{(\log n)^a}{n}\right)^{\frac{1}{d(2p+1)}} \leq h_0 \leq 1, \tag{III.9}$$

where $C_\lambda$, only depending on the kernel $K$, is defined in Section 2.b.i. Note in particular that the lower bound does not depend on any unknown value, and thus can be implemented as the bandwidth initialization. Besides, observe that each component of the ideal bandwidth for estimating $f$ on $\mathcal{H}_d(s, L)$ is of order $n^{-1/(2s+r)}$ for relevant components and are constant for irrelevant ones. So, if $s \leq p$ as assumed in Theorem III.2, then $h_0$ is larger than all relevant components of the optimal bandwidth, as required by the RevDir CDRODEO procedure.

## 3.c Assumptions and main result

To derive rates of convergence for $\hat{f}(w)$, we need three assumptions. The first two ones are related to $\mathrm{f}_X$, the density of the $X_i$'s.

**Assumption** $\mathcal{L}_X$ *[Lower bound on* $\mathrm{f}_X$*]*
*The density* $\mathrm{f}_X$ *is bounded away from 0 in the neighborhood of* $x$ *(from the evaluation point* $w=(x,y)$*):*

$$\delta := \inf_{u\in\mathcal{U}_1} \mathrm{f}_X(u) > 0,$$

*where* $\mathcal{U}_1 := \left\{ u\in\mathbb{R}^{d_1} : x-u \in (\mathit{supp}(K))^{d_1} \right\}$.

**Remark III.3.** *Similarly, to Remark III.2, the size of* $\mathcal{U}_1$ *is fixed but it could decrease to 0 if we modify the stopping rule of the Reverse Step.*

This assumption is classical in the regression setting or for conditional density estimation. Indeed, if $\mathrm{f}_X$ is equal or close to 0 in the neighborhood of $x$, we shall have no or very few observations to estimate the distribution of $Y$ given $X=x$. Thus, this assumption is required in all of the aforementioned works about conditional density estimation.

The next assumption specifies that we can estimate $\mathrm{f}_X$ very precisely.

**Assumption** $\mathcal{E}\mathrm{f}_X$ *[Estimation of* $\mathrm{f}_X$*]*
*The estimator of* $\mathrm{f}_X$ *in (III.1) satisfies the following two conditions:*

*Condition (i) a positive lower bound:* $\tilde{\delta}_X := \inf_{u\in\mathcal{U}_1} \widetilde{\mathrm{f}}_X(u) > n^{-1/2}$,

*Condition (ii) a concentration inequality in local sup norm:*

$$\mathbb{P}\left( \sup_{u\in\mathcal{U}_1} \left| \mathrm{f}_X(u) - \widetilde{\mathrm{f}}_X(u) \right| > M_X \frac{(\log n)^{\frac{a}{2}}}{\sqrt{n}} \right) \leq \exp(-(\log n)^{1+\frac{a-1}{2}}),$$

*with* $M_X := \frac{\delta \|J\|_2 \|K\|_2^{d-1}}{4\|f\|_{\infty,\mathcal{U}} \|J\|_1 \|K\|_1^{d-1}}$.

**Remark III.4.** *For the simpler problem of density estimation, since* $\mathrm{f}_X \equiv 1 \equiv \widetilde{\mathrm{f}}_X$*, Assumption* $\mathcal{E}\mathrm{f}_X$ *is obviously satisfied.*

The following proposition shows that conditions of Assumption $\mathcal{E}\mathrm{f}_X$ are feasible if we have at hand a sample, with same distribution as $X$, whose size is large enough. Furthermore, $\widetilde{\mathrm{f}}_X$, the estimator provided by the proof of Proposition III.1, is easily implementable.

**Proposition III.1.** *Given a sample* $\tilde{X}$ *with same distribution as* $X$ *and of size* $n_X = n^c$ *with* $c>1$*, if* $\mathrm{f}_X$ *is of class* $\mathcal{C}^{p'}$ *with* $p' \geq \frac{d_1}{2(c-1)}$*, there exists an estimator* $\widetilde{\mathrm{f}}_X$ *which satisfies Assumption* $\mathcal{E}\mathrm{f}_X$*.*

To prove Proposition III.1, we build $\widetilde{\mathrm{f}}_X$ as a truncated kernel estimator with a fixed bandwidth, but other methods can be used in practice, as, for instance, a Rodeo algorithm for density estimation. Actually any reasonable nonparametric estimator would have a rate of convergence in sup norm of the form $n_X^{-\kappa}$ (typically $\kappa = p'/(2p'+d_1)$). Then Condition (ii) of Assumption $\mathcal{E}\mathrm{f}_X$ is verified as soon as $n_X^{-\kappa} \leq n^{-1/2}$ and we need $c \geq 1 + d_1/(2p')$. Then, observe that if $\mathrm{f}_X$ is of class $\mathcal{C}^\infty$, then we just need $c=1$ and we can take $\tilde{X}=X$. If we know that $\mathrm{f}_X$ is at least of class $\mathcal{C}^1$ but its precise smoothness is unknown, taking $c \geq 1+d_1/2$ is sufficient to satisfy assumptions of Proposition III.1.

The next assumption is necessary to control the bias.

**Assumption** $\mathcal{M}$ *[Monotonicity]*
*For all* $j \in \mathcal{R}$*, for all* $h$ *and* $h' \in (\mathbb{R}_+^*)^d$ *such that* $h \preceq h'$*,* $|\mathbb{E}[\bar{Z}_{h,j}]| \leq |\mathbb{E}[\bar{Z}_{h',j}]|$*, where* $\bar{Z}_{h,j}$ *is defined as* $Z_{h,j}$ *in* (III.3) *but with true* $\mathrm{f}_X$ *replacing* $\widetilde{\mathrm{f}}_X$*.*
Let us comment Assumption $\mathcal{M}$ that requires monotony of a specific bias term. Indeed, denoting $M_j$ the pseudo-kernel defined by $M_j(z) = J(z_j)\prod_{k\neq j} K(z_k)$, we have

$$\mathbb{E}[\bar{Z}_{hj}] = \frac{\partial}{\partial h_j}(\mathrm{K}_h \star f - f)(w) = -\frac{1}{h_j}\int M_j(z)[f(w - h.z) - f(w)]dz,$$

which is, under mild assumptions, of order $\sum_{k=1}^d h_k^s h_j^{-1} \approx h_j^{s-1}$ if the smoothness of $f$ at $w$ is exactly $s$ in each direction. In this case, Assumption $\mathcal{M}$ is satisfied (we assume $s > 1$ subsequently). This assumption is needed to control the bias term $B(h) := (\mathrm{K}_h \star f - f)(w)$ to prevent the algorithm from stopping at bandwidths for which $\frac{\partial}{\partial h_j}B(h)$ vanishes. Remember that this term plays a key role for the RevDir CDRodeo procedure (see Section 2.b.ii). It means that the RevDir CDRodeo procedure is not suitable for too irregular functions. Anyway, estimating non-smooth functions in large dimensions is a very intricate problem. Actually, this assumption is the price to pay for not exploring all possible bandwidths and only focusing on special paths and is the counterpart of the competitive computational time of the RevDir CDRodeo algorithm. Such conditions are shared by many iterative procedures. See the stopping time procedure proposed by Blanchard et al. [2016] and their Section 1.2 for instance or more generally, gradient descent algorithms that use convexity conditions. Observe that Assumption $\mathcal{M}$ looks like a convexity condition.

**Remark III.5.** *Assumption* $\mathcal{M}$ *is not always required : in particular, see the alternative assumption in the Chapter II for non adaptive estimation :* $f$ *is assumed of class* $C^p$ *with* $\frac{\partial^p}{\partial h_j^p} f(h) \neq 0$ *where* $p$ *is the exact order of the kernel.*

We now derive the main result of this chapter proved in Section 4 in which we show that $\hat{h}$ is closed to the ideal bandwidth $h^*$ defined in Section 2.b.ii.

**Theorem III.2.** *We assume that* $f$ *has only* $r$ *relevant components with* $r \in \{0, \dots, d\}$ *and belongs to* $\mathcal{H}_d(s, L)$ *where* $L > 0$ *and* $1 < s \leq p$*. Then, under Assumptions* $\mathcal{L}_X$*,* $\mathcal{E}\mathrm{f}_X$*,* $\mathcal{M}$*, the pointwise risk of the RevDir CDRodeo estimator* $\hat{f}_{\hat{h}}(w)$ *is bounded as follows: for any* $l \geq 1$*, for* $n$ *large enough,*

$$\mathbb{E}\left[\left|\hat{f}_{\hat{h}}(w) - f(w)\right|^l\right]^{1/l} \leq C\left(\frac{(\log n)^a}{n}\right)^{\frac{s}{2s+r}} \tag{III.10}$$

*where C only depends on* $d, r, K, \beta, \delta, L, s, \|f\|_{\infty,\mathcal{U}}$*.*

We can compare the obtained rate with the classical pointwise adaptive minimax rate for estimating a $s$-regular $r$-dimensional density, which is $((\log n)/n)^{s/(2s+r)}$ (see Rebelles [2015]). Our procedure achieves this rate up to the term $(\log n)^{s(a-1)/(2s+r)}$. In Section 3.b, we specify that any value $a > 1$ is suitable. So, our procedure is nearly optimal. Actually, we need $a > 1$ to ensure that for $n$ large enough,

$$(\log n)^{a-1} \geq \frac{\|f\|_{\infty,\mathcal{U}}}{\delta}$$

but if an upper bound (or a pre-estimator) of $\frac{\|f\|_{\infty,\mathcal{U}}}{\delta}$ were known, we could obtain the similar result with $a = 1$, and our procedure would be rate-optimal without any additional logarithmic term. Remember that the term $(\log n)^{s/(2s+r)}$ is the price to pay for adaptation with respect to the smoothness (see Tsybakov [1998]). Theorem III.2 shows that, in our setting, there is no additive price for not knowing the sparsity, i.e. the value of $r$. This result is new.

**Remark III.6.** *We need $s > 1$, which means that $f$ has to be at least $\mathcal{C}^1$. This technical assumption is related to our methodology based on derivatives of $\hat{f}_h(w)$ as proxies of derivatives of $f$ to detect relevant components.*

### 3.d Algorithm complexity

We now discuss the complexity of CDRODEO without taking into account the pre-computation cost of $\widetilde{\mathrm{f}}_X$ at the points $X_i$, $i = 1:n$ (used for computing the $Z_{hj}$). Regarding the computation cost of $\widetilde{\mathrm{f}}_X$, the estimator built for the proof of Proposition III.1 has complexity $O(d_1 n^c)$ but in practice we use a RODEO estimator with the same sample size $n$, which has a complexity $O(d_1 n \log n)$ for each computation of $\widetilde{\mathrm{f}}_X(X_i)$ which causes an additional cost in $O(d_1 n^2 \log n)$.

During the Reverse Step, $|\mathcal{A}\mathrm{ct}^{(-1)}|$ components are updated, and, for fixed $h$, the computation of all $Z_{hj}$'s and the comparisons to the thresholds $\lambda_{hj}$ need $O(|\mathcal{A}\mathrm{ct}^{(-1)}|n)$ operations. In the same way, during the Direct Step, $|\mathcal{A}\mathrm{ct}^{(0)}|$ components are updated and each update needs $O(|\mathcal{A}\mathrm{ct}^{(0)}|n)$ operations. Since the number of updates is at worse of order $\log(n)$ (because of the stopping conditions), and $|\mathcal{A}\mathrm{ct}^{(-1)}| + |\mathcal{A}\mathrm{ct}^{(0)}| \leq d$, we obtain the following proposition. More details can be found in the proof (see Section 4.f).

**Proposition III.3.** *Apart from the computation of $\widetilde{\mathrm{f}}_X$, the total worst-case complexity of RevDir CDRODEO algorithm is*

$$O(dn \log n).$$

Notice that for classical methods with optimization on a bandwidths grid, the complexity is of ordrer $dn|H|^d$, where $|H|$ denotes the size of the grid for each component. In practice, the grid has to include at least $\log n$ points, which leads to a computational cost $O(dn(\log n)^d)$. For $d = 5$ and $n = 10^5$, the ratio of complexities is already $\frac{dn(\log n)^d}{dn \log n} > 1.7 \times 10^4$.

## 4 Proofs

The proof of the theorem uses some intermediate lemmas. See Appendix for their statements and proofs.

### 4.a Notations

First, we define some general notations: We denote

- $\partial_j g$ the partial derivative of a function $g$ with respect to its $j$-th component;
- $v \cdot v'$ the multiplication term by term of two vectors $v$ and $v'$;
- $l:m$ the set of consecutive integers from $l$ to $m$;
- $v_{\mathcal{I}}$ the vector $v$ restricted to its components indexed in $\mathcal{I}$;
- $b \vee c = \max(b,c)$ the maximum value of two reals $b$ and $c$.

Let us now introduce the key quantities of the proofs. For any bandwidth $h \in (\mathbb{R}_+^*)^d$ and any component $k \in \{1:d\}$, we consider the estimator $\bar{f}_h(w)$ that we would have use if the density $\mathrm{f}_X$ were known:

$$\bar{f}_h(w) := \frac{1}{n}\sum_{i=1}^{n} \bar{f}_{hi}(w), \quad \bar{f}_{hi}(w) := \frac{\mathrm{K}_h(w - W_i)}{\mathrm{f}_X(X_i)}$$

and we denote $\Delta_h$ its difference with the real estimator:

$$\Delta_h := \hat{f}_h(w) - \bar{f}_h(w).$$

We denote $\bar{B}_h := \mathbb{E}\left[\bar{f}_h(w)\right] - f(w)$ the bias of $\bar{f}_h(w)$. We also consider its partial derivative $\bar{Z}_{hk}$:

$$\bar{Z}_{hk} := \frac{\partial}{\partial h_k}\bar{f}_h(w).$$

We can write

$$\bar{Z}_{hk} := \frac{1}{n}\sum_{i=1}^{n}\bar{Z}_{hik}, \quad \bar{Z}_{hik} := \frac{1}{\mathrm{f}_X(X_i)}\frac{\partial}{\partial h_k}\left(\prod_{k=1}^{d} h_k^{-1} K\left(\frac{w_k - W_{ik}}{h_k}\right)\right).$$

We shall consider $\Delta_{Z,hk}$ the difference between $Z_{hk}$ and $\bar{Z}_{hk}$:

$$\Delta_{Z,hk} := Z_{hk} - \bar{Z}_{hk}.$$

Note that the value of the final bandwidth of our procedure provides the value of the bandwidth at each iteration. More precisely, if a bandwidth $h$ is the output of the RevDir procedure, we denote $(h^{(t)})_{t\in\mathbb{Z}}$, the different values of the bandwidth for all iterations $t$.
- On the one hand, if $h_k > h_0$, it means that at Initialization, the component $k$ was in $\mathcal{A}\mathrm{ct}^{(-1)}$ and then the bandwidth path of this component has increased during the Reverse Step according to the following path $h_0\beta^{-1}, h_0\beta^{-2}, \ldots$ until $h_k := h_0\beta^{-|t_k|}$, and remains fixed during the whole Direct Step ($t \geq 0$).
- On the other hand, if $h_k < h_0$, the component $k$ was in $\mathcal{A}\mathrm{ct}^{(0)}$ at Initialization. Thus the value of the bandwidth component was fixed and equals to $h_0$ during the Reverse Step (i.e for every $t < 0$ ). Then, it decreases during the Direct step: $h_0\beta, h_0\beta^2, \ldots$ until $h_k := h_0\beta^{t_k}$ is achieved (see Figure III.1). This gives the following formula: for any $k = 1 : d$, during the Reverse Step (when $t < 0$),

$$h_k^{(t)} := \max(h_0, \min\left(h_k, \beta^t h_0\right)) = \begin{cases} h_0 & \text{if } k \text{ is active during the Direct Step,} \\ \beta^t h_0 & \text{if } k \text{ is active during the Reverse Step and not} \\ & \text{deactivated yet,} \\ h_k & \text{if } k \text{ has already been deactivated during the} \\ & \text{Reverse Step,} \end{cases}$$

and during Direct Step (when $t \geq 0$),

$$h_k^{(t)} := \max\left(h_k, \beta^t h_0\right) = \begin{cases} \beta^t h_0 & \text{if } k \text{ is active during the Direct Step and not deactivated yet,} \\ h_k & \text{if } k \text{ has already been deactivated (during the Reverse or the} \\ & \text{Direct Step).} \end{cases}$$

Now we can define the set of bandwidths $\mathcal{H}_{\mathsf{hp}}$ which contains with high probability the bandwidth selected by the RevDir procedure:

$$\begin{aligned} \mathcal{H}_{\mathsf{hp}} := \{h \in \left(\mathbb{R}_+^*\right)^d : \forall k = 1 : d, h_k = \beta^{t_k} h_0 \leq 1 \text{ with } t_k \in \mathbb{Z}, \\ \text{and } \prod_{k=1}^{d} h_k \geq \beta^r \frac{(\log n)^{a+1}}{n}, \\ \text{and } \forall k \in \mathcal{R}^c, h_k = h_{\mathsf{irr}}\}, \end{aligned}$$

where $\beta^{t_{\text{irr}}}$ and $h_{\text{irr}}$ are uniquely defined by: $t_{\text{irr}} \in \mathbb{Z}$ and $\beta < h_{\text{irr}} := \beta^{t_{\text{irr}}} h_0 \leq 1$. We also denote $\mathcal{H}_{\text{hp}}^{\text{Rev}}$ (respectively $\mathcal{H}_{\text{hp}}^{\text{Dir}}$) the set which contains the different states of the bandwidth during the Reverse Step (respectively the Direct Step) provided that the selected bandwidth is in $\mathcal{H}_{\text{hp}}$:

$$\mathcal{H}_{\text{hp}}^{\text{Rev}} := \{h^{(t)} : h \in \mathcal{H}_{\text{hp}}, t < 0\} \tag{III.11}$$

$$\mathcal{H}_{\text{hp}}^{\text{Dir}} := \{h^{(t)} : h \in \mathcal{H}_{\text{hp}}, t \geq 0\}. \tag{III.12}$$

Finally, we introduce the high probability event $\mathcal{E}_{\text{hp}}$ on which $\hat{h}$ systematically belongs to $\mathcal{H}_{\text{hp}}$:

$$\mathcal{E}_{\text{hp}} := \widetilde{\mathcal{A}}_n \cap \bigcap_{h \in \mathcal{H}_{\text{hp}}} \left( \mathcal{B}\text{ern}_{\bar{f}}(h) \cap \mathcal{B}\text{ern}_{|\bar{f}|}(h) \right) \cap \bigcap_{h \in (\mathcal{H}_{\text{hp}}^{\text{Rev}} \cup \mathcal{H}_{\text{hp}}^{\text{Dir}})} \bigcap_{k=1}^{d} \left( \mathcal{B}\text{ern}_{\bar{Z}}(h,k) \cap \mathcal{B}\text{ern}_{|\bar{Z}|}(h,k) \right), \tag{III.13}$$

where $\widetilde{\mathcal{A}}_n$ is the high probability event of (ii) in Assumption $\mathcal{E}\text{f}_X$:

$$\widetilde{\mathcal{A}}_n = \left\{ \sup_{u \in \mathcal{U}_1} \left| \text{f}_X(u) - \widetilde{\text{f}}_X(u) \right| \leq M_X \frac{(\log n)^{\frac{a}{2}}}{\sqrt{n}} \right\},$$

and $\mathcal{B}\text{ern}_{\dagger}(\ddagger)$ is the high probability event resulting of Bernstein's Inequality applied on the random variable $\dagger$ with parameter(s) $\ddagger$. More formally:

$$\mathcal{B}\text{ern}_{\bar{f}}(h) := \{|\bar{f}_h(w) - \mathbb{E}[\bar{f}_h(w)]| \leq \sigma_h\},$$

$$\mathcal{B}\text{ern}_{|\bar{f}|}(h) := \left\{ \left| \frac{1}{n} \sum_{i=1}^{n} |\bar{f}_{hi}(w)| - \mathbb{E}[|\bar{f}_h(w)|] \right| \leq \mathsf{C}_{\bar{\mathsf{E}}} \right\},$$

$$\mathcal{B}\text{ern}_{\bar{Z}}(h,k) := \left\{ |\bar{Z}_{hk} - \mathbb{E}\bar{Z}_{hk}| \leq \frac{1}{2} \lambda_{hk} \right\},$$

$$\mathcal{B}\text{ern}_{|\bar{Z}|}(h,k) := \left\{ |\frac{1}{n} \sum_{i=1}^{n} |\bar{Z}_{hik}| - \mathbb{E}|\bar{Z}_{h1k}|| \leq \mathsf{C}_{E|\bar{Z}|} h_k^{-1} \right\},$$

where

$$\sigma_h = \mathsf{C}_\sigma \sqrt{\frac{(\log n)^a}{n \prod_{k=1}^{d} h_k}}$$

with $\mathsf{C}_\sigma = \frac{2\|K\|_2^d \|f\|_{\infty,\mathcal{U}}^{\frac{1}{2}}}{\delta^{\frac{1}{2}}}$. See Lemmas II.5 and II.6 in Appendix for the details and definitions of constants $\mathsf{C}_{\bar{\mathsf{E}}}, \mathsf{C}_{E|\bar{Z}|}$.

## 4.b Main steps of the proof

Proposition III.4 describes the form of the bandwidth selected by the RevDir procedure with high probability. Given this selection, Proposition III.5 gives upper bounds on the bias and the deviation of the estimator $\bar{f}_{\hat{h}}(w)$.

**Proposition III.4.** *The selected bandwidth belongs to $\mathcal{H}_{hp}$ with high probability. More precisely:*

$$\mathcal{E}_{hp} \subset \{\hat{h} \in \mathcal{H}_{hp}\} \tag{III.14}$$

*and for $n$ large enough:*

$$\mathbb{P}\left(\mathcal{E}_{hp}^c\right) \leq 2e^{-(\log n)^{1+\frac{a-1}{2}}}. \tag{III.15}$$

Note in particular that with high probability the irrelevant components of the selected bandwidth are equal to $h_{\text{irr}}$.

Recall that $\bar{B}_h := \mathbb{E}\left[\bar{f}_h(w)\right] - f(w)$ is the bias of $\bar{f}_h(w)$.

**Proposition III.5.** *The following upper bounds are satisfied for all $h \in \mathcal{H}_{hp}$, and any constants $A \in \mathbb{R}$ and $C_A > 0$:*

$$\mathbb{1}_{\{\hat{h}=h\}\cap\mathcal{E}_{hp}} |\bar{B}_h| \leq r C_{\bar{B}} {C_A}^s \frac{(\log n)^{As}}{n^{\frac{s}{2s+r}}} + r \max\left(\frac{7C_\lambda}{4\beta^{\frac{d-r}{2}} {C_A}^{\frac{r}{2}}} \frac{(\log n)^{\frac{a-Ar}{2}}}{n^{\frac{s}{2s+r}}}, \frac{7}{4}\left(\frac{(\log n)^a}{n}\right)^{\frac{p}{2p+1}}\right), \tag{III.16}$$

$$\begin{aligned}\mathbb{1}_{\{\hat{h}=h\}\cap\mathcal{E}_{hp}} \left|\bar{f}_h(w) - \mathbb{E}\left[\bar{f}_h(w)\right]\right| &\leq \mathbb{1}_{\{\hat{h}=h\}\cap\mathcal{E}_{hp}} \sigma_h \\ &\leq \max\left(\frac{C_\sigma}{\beta^{(d-r)/2} {C_A}^{r/2}} (\log n)^{(a-Ar)/2}, \frac{4{C_A}^s C_{E\bar{Z}} C_\sigma \beta^{-\frac{r}{2}-s}}{C_\lambda} (\log n)^{sA}\right) n^{-\frac{s}{2s+r}},\end{aligned} \tag{III.17}$$

*where $C_\lambda$ is the constant defined in (III.5) and $C_{\bar{B}}, C_\sigma, C_{E\bar{Z}}$ are constants defined in Lemmas II.5 and II.6 in Appendix.*

## 4.c Proof of Theorem III.2

Let us fix $l > 1$. From Proposition III.4: $\mathcal{E}_{\text{hp}} \subset \{\hat{h} \in \mathcal{H}_{\text{hp}}\}$, thus:

$$\mathbb{E}\left[\left|\hat{f}_{\hat{h}}(w) - f(w)\right|^l\right] = \mathbb{E}\left[\mathbb{1}_{\mathcal{E}_{\text{hp}}^c} \left|\hat{f}_{\hat{h}}(w) - f(w)\right|^l\right] + \sum_{h\in\mathcal{H}_{\text{hp}}} \mathbb{E}\left[\mathbb{1}_{\{\hat{h}=h\}\cap\mathcal{E}_{\text{hp}}} \left|\hat{f}_h(w) - f(w)\right|^l\right]. \tag{III.18}$$

We first control the terms $\mathbb{E}\left[\mathbb{1}_{\{\hat{h}=h\}\cap\mathcal{E}_{\text{hp}}} \left|\hat{f}_h(w) - f(w)\right|^l\right]$. We fix $h \in \mathcal{H}_{\text{hp}}$. Then, we decompose the difference $\hat{f}_h(w) - f(w)$ as follows:

$$\hat{f}_h(w) - f(w) = \Delta_h + \left(\bar{f}_h(w) - \mathbb{E}\left[\bar{f}_h(w)\right]\right) + \bar{B}_h, \tag{III.19}$$

where we recall the notations $\Delta_h := \hat{f}_h(w) - \bar{f}_h(w)$ and $\bar{B}_h := \mathbb{E}\left[\bar{f}_h(w)\right] - f(w)$. Remark that $\prod_{k=1}^d h_k \leq 1$, since $h \in \mathcal{H}_{\text{hp}}$. We apply 2. of Lemma 3 and 3. of Lemma 1: Since $\mathcal{E}_{\text{hp}} \subset \left(\widetilde{\mathcal{A}}_n \cap \mathcal{B}\text{ern}_{|\bar{f}|}(h)\right) \cap \mathcal{B}\text{ern}_{\bar{f}}(h)$:

$$\mathbb{1}_{\mathcal{E}_{\text{hp}}} |\Delta_h| \leq \mathsf{C}_{\mathsf{M}\Delta} \sigma_h$$

and

$$\mathbb{1}_{\mathcal{E}_{\text{hp}}} \left|\bar{f}_h(w) - \mathbb{E}\left[\bar{f}_h(w)\right]\right| \leq \sigma_h.$$

Therefore:

$$\mathbb{1}_{\{\hat{h}=h\}\cap\mathcal{E}_{\text{hp}}} \left|\hat{f}_h(w) - f(w)\right| \leq \mathbb{1}_{\{\hat{h}=h\}\cap\mathcal{E}_{\text{hp}}} \left((\mathsf{C}_{\mathsf{M}\Delta} + 1)\sigma_h + |\bar{B}_h|\right). \tag{III.20}$$

From Proposition III.5 which controls both $\sigma_h$ and $|\bar{B}_h|$, we deduce:

$$\begin{aligned}
&\mathbb{1}_{\{\hat{h}=h\}\cap\mathcal{E}_{\mathsf{hp}}}\left|\hat{f}_h(w)-f(w)\right| \\
&\qquad\leq (\mathsf{C}_{\mathsf{M}\Delta}+1)\max\left(\frac{\mathsf{C}_\sigma}{\beta^{(d-r)/2}{C_A}^{r/2}}(\log n)^{\frac{a-Ar}{2}},\frac{4{C_A}^s\mathsf{C}_{E\bar{Z}}\mathsf{C}_\sigma\beta^{-\frac{r}{2}-s}}{\mathsf{C}_\lambda}(\log n)^{sA}\right)n^{-\frac{s}{2s+r}} \\
&\qquad\quad + r\mathsf{C}_{\bar{B}}{C_A}^s(\log n)^{As}n^{-\frac{s}{2s+r}} + r\max\left(\frac{7\mathsf{C}_\lambda}{4\beta^{\frac{d-r}{2}}{C_A}^{\frac{r}{2}}}\frac{(\log n)^{\frac{a-Ar}{2}}}{n^{\frac{s}{2s+r}}},\frac{7}{4}\left(\frac{(\log n)^a}{n}\right)^{\frac{p}{2p+1}}\right).
\end{aligned}$$

We optimize in $A$ and $C_A$: With $A=\frac{a}{2s+r}$, we obtain

$$\mathbb{1}_{\{\hat{h}=h\}\cap\mathcal{E}_{\mathsf{hp}}}\left|\hat{f}_h(w)-f(w)\right|\leq\max\left(C_1\left(\frac{(\log n)^a}{n}\right)^{\frac{s}{2s+r}},\frac{7}{4}r\left(\frac{(\log n)^a}{n}\right)^{\frac{p}{2p+1}}\right),$$

where $C_1$ depends on $\beta,d,r,s,\mathsf{C}_{\bar{B}},\mathsf{C}_{E\bar{Z}},\mathsf{C}_\sigma,\mathsf{C}_{\mathsf{M}\Delta},\mathsf{C}_\lambda$. If $r=0$, the last term in the right hand side vanishes, otherwise $p/(2p+1)\geq s/(2s+r)$ (since $p\geq s$). Therefore, for $n$ large enough:

$$\mathbb{1}_{\{\hat{h}=h\}\cap\mathcal{E}_{\mathsf{hp}}}\left|\hat{f}_h(w)-f(w)\right|\leq\mathsf{C}'\left(\frac{(\log n)^a}{n}\right)^{\frac{s}{2s+r}}. \tag{III.21}$$

To prove the theorem, it then remains to control $\left|\hat{f}_h(w)-f(w)\right|$ on $\mathcal{E}_{\mathsf{hp}}^c$. Recall that:

$$\prod_{k=1}^d\hat{h}_k\geq\beta^r\frac{(\log n)^{1+a}}{n},$$

and (i):

$$\tilde{\delta}_X:=\inf_{u\in\mathcal{U}_1}\widetilde{\mathrm{f}}_X(u)>n^{-1/2},$$

then we can roughly bound $\hat{f}_{\hat{h}}(w)$ by:

$$\left|\hat{f}_{\hat{h}}(w)\right|\leq\frac{\|K\|_\infty^d n}{\tilde{\delta}_X\beta^r(\log n)^{1+a}}=o(n^2).$$

So:

$$\left|\hat{f}_h(w)-f(w)\right|^l=o(n^{2l})=o(e^{2l\log n}).$$

Besides, from Proposition III.4:

$$\mathbb{P}\left(\mathcal{E}_{\mathsf{hp}}^c\right)\leq 2e^{-(\log n)^{1+\frac{a-1}{2}}}.$$

Note that, since $a>1$,

$$2l\log n+l\log(n^{\frac{1}{2}})=o((\log n)^{1+\frac{a-1}{2}}), \tag{III.22}$$

therefore:

$$\mathbb{E}\left[\mathbb{1}_{\mathcal{E}_{\mathsf{hp}}^c}\left|\hat{f}_{\hat{h}}(w)-f(w)\right|^l\right]^{1/l}\leq\left(\mathbb{P}\left(\mathcal{E}_{\mathsf{hp}}^c\right)e^{2l\log n}\right)^{1/l}=o(n^{-\frac{1}{2}}).$$

To conclude, we combine Equation (III.18) with the above upper bound and Inequality (III.21):

$$\begin{aligned}
\mathbb{E}\left[\left|\hat{f}_{\hat{h}}(w)-f(w)\right|^l\right]^{1/l}&\leq o(n^{-\frac{1}{2}})+\left\{\left(\mathsf{C}'\left(\frac{(\log n)^a}{n}\right)^{\frac{s}{2s+r}}\right)^l\sum_{h\in\mathcal{H}_{\mathsf{hp}}}\mathbb{E}[\mathbb{1}_{\hat{h}=h}]\right\}^{1/l} \\
&\leq\mathsf{C}\left(\frac{(\log n)^a}{n}\right)^{\frac{s}{2s+r}},
\end{aligned}$$

with C depending on $d,r,\|f\|_\infty,\mathcal{U},\delta,L,s,K,\beta$.

## 4.d Proof of Proposition III.4

By definition of the procedure, any selected bandwidth $\hat{h}$ satisfies

$$\exists (t_1,\dots,t_d) \in \mathbb{Z}^d, \forall k = 1:d, \hat{h}_k = \beta^{t_k} h_0.$$

The loop condition in the Reverse Step imposes for any active component $k$ that at the beginning of an iteration $t \in \mathbb{Z}_-$ :

$$\hat{h}_k^{(t)} \leq \beta.$$

At most, $\hat{h}_k^{(t)}$ is multiplied by $\beta^{-1}$. Then after the last update of the component $\hat{h}_k$:

$$\hat{h}_k \leq 1 = \beta^{-1}\beta.$$

Now let us prove that on $\mathcal{E}_{\text{hp}}$, the irrelevant components are deactivated at value $h_{\text{irr}}$. It suffices to show that during the initialization, the irrelevant components activate for Reverse Step, *i.e.*:

$$\mathcal{R}^c \subset \mathcal{A}\text{ct}^{(-1)},$$

and in the case where $h_0 \leq \beta$, it suffices to prove that they remain active at all iterations $t = -1 : t_{\text{irr}}$. Remember that $t_{\text{irr}} \in \mathbb{Z}$ is defined such that: $h_{\text{irr}} = \beta^{t_{\text{irr}}} h_0$.

Note that if the irrelevant components remain active at all iteration $t = -1 : t_{\text{irr}}$, then for $k \in \mathcal{R}^c$, $\hat{h}_k^{(t)} = H_k^{(t)} = \beta^t h_0$. It corresponds to the definition of $\mathcal{H}_{\text{hp}}$, since for all $h \in \mathcal{H}_{\text{hp}}$, $t = -1 : t_{\text{irr}}$ and $k \in \mathcal{R}^c$,

$$h_k^{(t)} = \beta^t h_0.$$

Therefore, there exists $h \in \mathcal{H}_{\text{hp}}$ such that $\hat{h}^{(t)} = h^{(t)}$ for all iterations $t = -1 : t_{\text{irr}}$. We will then prove that for any $h \in \mathcal{H}_{\text{hp}}$, $t = -1 : t_{\text{irr}}$ and $k \in \mathcal{R}^c$,

$$\mathbb{1}_{\mathcal{E}_{\text{hp}}} |Z_{h^{(t)}k}| \leq \lambda_{h^{(t)}k}.$$

Let us fix $h \in \mathcal{H}_{\text{hp}}$, $t \in \{-1,\dots,t_{\text{irr}}\}$ and $k \in \mathcal{R}^c$. We decompose $Z_{h^{(t)}k}$ as follows:

$$Z_{h^{(t)}k} = \left(Z_{h^{(t)}k} - \bar{Z}_{h^{(t)}k}\right) + \left(\bar{Z}_{h^{(t)}k} - \mathbb{E}\bar{Z}_{h^{(t)}k}\right) + \mathbb{E}\bar{Z}_{h^{(t)}k}. \tag{III.23}$$

We use:

- 1. of Lemma 3: Recall the notation $\Delta_{Z,h^{(t)}k} := Z_{h^{(t)}k} - \bar{Z}_{h^{(t)}k}$, then remark that $\forall h' \in \mathcal{H}_{\text{hp}}^{\text{Rev}} \cup \mathcal{H}_{\text{hp}}^{\text{Dir}}$, $\prod_{k=1}^d h'_k \leq 1$, and $\mathcal{E}_{\text{hp}} \subset \mathcal{B}\text{ern}_{|\bar{Z}|}(h^{(t)},k) \cap \widetilde{\mathcal{A}}_n$, therefore:

$$\mathbb{1}_{\mathcal{E}_{\text{hp}}} \left(Z_{h^{(t)}k} - \bar{Z}_{h^{(t)}k}\right) \leq \frac{1}{4}\lambda_{h^{(t)}k},$$

- the definition of $\mathcal{B}\text{ern}_{\bar{Z}}(h^{(t)},k)$: since $\mathcal{E}_{\text{hp}} \subset \mathcal{B}\text{ern}_{\bar{Z}}(h^{(t)},k)$,

$$\mathbb{1}_{\mathcal{E}_{\text{hp}}} \left|\bar{Z}_{h^{(t)}k} - \mathbb{E}\bar{Z}_{h^{(t)}k}\right| \leq \frac{1}{2}\lambda_{h^{(t)}k},$$

- 2. of Lemma 2: since $k \in \mathcal{R}^c$,

$$\mathbb{E}\bar{Z}_{h^{(t)}k} = 0.$$

Therefore:

$$\mathbb{1}_{\mathcal{E}_{\mathsf{hp}}}|Z_{h^{(t)}k}| \leq \frac{3}{4}\lambda_{h^{(t)}k} \leq \lambda_{h^{(t)}k},$$

and so, every irrelevant component is active during Reverse Step until Iteration $t_{\mathsf{irr}}$. In particular, we have proved that:

$$\mathcal{E}_{\mathsf{hp}} \subset \{\forall k \in \mathcal{R}^c : \hat{h}_k = h_{\mathsf{irr}}\}.$$

Let us now prove that on $\mathcal{E}_{\mathsf{hp}}$,

$$\prod_{k=1}^{d} \hat{h}_k \geq \beta^r \frac{(\log n)^{1+a}}{n}.$$

The loop condition in the Direct Step imposes that at the beginning of any iteration $t \geq 0$:

$$\prod_{k=1}^{d} \hat{h}_k^{(t)} \geq \frac{(\log n)^{1+a}}{n}.$$

For our algorithm, the bandwidth can only decrease during the Direct Step. Since on $\mathcal{E}_{\mathsf{hp}}$, the irrelevant components are active the during Reverse Step, they are inactive during the Direct Step. This is the reason why during the last iteration, only relevant components could decrease and be multiplied by $\beta$. Therefore:

$$\prod_{k=1}^{d} \hat{h}_k \geq \beta^r \frac{(\log n)^{1+a}}{n},$$

which ends the proof of the inclusion (III.14) of Proposition III.4.

Finally, we control $\mathbb{P}\left(\mathcal{E}_{\mathsf{hp}}^c\right)$. We first control the cardinal of $\mathcal{H}_{\mathsf{hp}}$ by enumerating the possible values for a component of a bandwidth in $\mathcal{H}_{\mathsf{hp}}$. For $h \in \mathcal{H}_{\mathsf{hp}}$ and $k \in \mathcal{R}$,

$$\beta(\log n)^{1+a} n^{-1} \leq h_k \leq 1,$$

thus:

$$\left|\left\{h_k : h \in \mathcal{H}_{\mathsf{hp}}\right\}\right| = \left|\left\{\beta^t h_0 \in [\beta(\log n)^{1+a} n^{-1}, 1], t \in \mathbb{Z}\right\}\right| \leq 1 + \log_{\frac{1}{\beta}}\left(\frac{1}{\beta(\log n)^{1+a} n^{-1}}\right) \leq \log_{\frac{1}{\beta}} n$$

(for $n$ large enough). For $k \in \mathcal{R}^c$,

$$h_k = h_{\mathsf{irr}},$$

thus, we have

$$\left|\left\{h_k : h \in \mathcal{H}_{\mathsf{hp}}\right\}\right| = 1.$$

Therefore:

$$\left|\mathcal{H}_{\mathsf{hp}}\right| \leq \left(\log_{\frac{1}{\beta}} n\right)^r. \tag{III.24}$$

Let us also control the cardinal of $\mathcal{H}_{\mathsf{hp}}^{\mathsf{Rev}} \cup \mathcal{H}_{\mathsf{hp}}^{\mathsf{Dir}}$. The only supplementary bandwidths are the ones whose irrelevant components are smaller than $h_{\mathsf{irr}}$. We consider the irrelevant components as the relevant ones, and we obtain the rough bound

$$\left|\mathcal{H}_{\mathsf{hp}}^{\mathsf{Rev}} \cup \mathcal{H}_{\mathsf{hp}}^{\mathsf{Dir}}\right| \leq \left(\log_{\frac{1}{\beta}} n\right)^d. \tag{III.25}$$

By Assumption $\mathcal{E}f_X$, (ii):

$$\mathbb{P}\left(\widetilde{\mathcal{A}}_n^c\right) \leq \exp(-(\log n)^{1+\frac{a-1}{2}}).$$

We bound the events $\mathcal{B}\mathrm{ern}_{\bar{f}}(h)^c$'s and $\mathcal{B}\mathrm{ern}_{|\bar{f}|}(h)^c$'s using Lemma 1. Since for all $h \in \mathcal{H}_{\mathsf{hp}}$,

$$\prod_{k=1}^{d} h_k \geq \beta^r \frac{(\log n)^{a+1}}{n},$$

note that:

- Cond($h$): $\prod_{k=1}^{d} h_k \geq \frac{4^2\|K\|_\infty^{2d}}{9\delta^2 \mathsf{C}_\sigma^2} \frac{(\log n)^a}{n}$ is satisfied for any $h \in \mathcal{H}_{\mathsf{hp}}$ for $n$ large enough (when $\log n \geq \frac{4^2\|K\|_\infty^{2d}}{9\beta^r\delta^2\mathsf{C}_\sigma^2}$). So, we have
$$\mathbb{P}\left(\mathcal{B}\mathrm{ern}_{\bar{f}}(h)^c\right) \leq 2e^{-(\log n)^a}.$$
- Moreover,
$$\mathbb{P}\left(\mathcal{B}\mathrm{ern}_{|\bar{f}|}(h)^c\right) \leq 2e^{-\mathsf{C}_{\gamma|f|} n \prod_{k=1}^{d} h_k} \leq 2e^{-\mathsf{C}_{\gamma|f|}\beta^r(\log n)^{a+1}}.$$

Similarly, we bound the probability of events $\mathcal{B}\mathrm{ern}_{\bar{Z}}(h)^c$'s and $\mathcal{B}\mathrm{ern}_{|\bar{Z}|}(h)^c$'s using Lemma 2. Note that for all $h \in \mathcal{H}_{\mathsf{hp}}^{\mathsf{Rev}} \cup \mathcal{H}_{\mathsf{hp}}^{\mathsf{Dir}}$:

- Cond$_{\bar{Z}}(h)$: $\prod_{k=1}^{d} h_k \geq \mathsf{cond}_{\bar{Z}} \frac{(\log n)^a}{n}$ is satisfied for $n$ large enough (when $\log n \geq \frac{\mathsf{cond}_{\bar{Z}}}{\beta^r}$). So, we have
$$\mathbb{P}\left(\mathcal{B}\mathrm{ern}_{\bar{Z}}(h,j)^c\right) \leq 2e^{-\frac{\delta}{\|f\|_\infty,\, \mathcal{U}}(\log n)^a}.$$
- Moreover,
$$\mathbb{P}\left(\mathcal{B}\mathrm{ern}_{|\bar{Z}|}(h,j)^c\right) \leq 2e^{-\mathsf{C}_{\gamma|\bar{Z}|} n \prod_{k=1}^{d} h_k} \leq 2e^{-\mathsf{C}_{\gamma|\bar{Z}|}\beta^r(\log n)^{a+1}}.$$

Therefore,

$$\begin{aligned}
\mathbb{P}\left(\mathcal{E}_{\mathsf{hp}}^c\right) &\leq \mathbb{P}\left(\widetilde{\mathcal{A}}_n^c\right) + \sum_{h\in\mathcal{H}_{\mathsf{hp}}} \left(\mathbb{P}\left(\mathcal{B}\mathrm{ern}_{\bar{f}}(h)^c\right) + \mathbb{P}\left(\mathcal{B}\mathrm{ern}_{|\bar{f}|}(h)^c\right)\right) \\
&\quad + \sum_{h\in(\mathcal{H}_{\mathsf{hp}}^{\mathsf{Rev}}\cup\mathcal{H}_{\mathsf{hp}}^{\mathsf{Dir}})} \sum_{k=1}^{d} \left(\mathbb{P}\left(\mathcal{B}\mathrm{ern}_{\bar{Z}}(h,k)^c\right) + \mathbb{P}\left(\mathcal{B}\mathrm{ern}_{|\bar{Z}|}(h,k)^c\right)\right) \\
&\leq e^{-(\log n)^{1+\frac{a-1}{2}}} + \sum_{h\in\mathcal{H}_{\mathsf{hp}}} \left(2e^{-(\log n)^a} + 2e^{-\mathsf{C}_{\gamma|f|}\beta^r(\log n)^{a+1}}\right) \\
&\quad + \sum_{h\in(\mathcal{H}_{\mathsf{hp}}^{\mathsf{Rev}}\cup\mathcal{H}_{\mathsf{hp}}^{\mathsf{Dir}})} \sum_{k=1}^{d} \left(2e^{-\frac{\delta}{\|f\|_\infty,\, \mathcal{U}}(\log n)^a} + 2e^{-\mathsf{C}_{\gamma|\bar{Z}|}\beta^r(\log n)^{a+1}}\right) \\
&\leq e^{-(\log n)^{1+\frac{a-1}{2}}} \left(1 + 4\left(\log_{\frac{1}{\beta}} n\right)^r e^{-(\log n)^{\frac{a-1}{2}}} + 4d\left(\log_{\frac{1}{\beta}} n\right)^d e^{-\frac{\delta}{\|f\|_\infty,\, \mathcal{U}}(\log n)^{\frac{a-1}{2}}}\right) \\
&\leq 2e^{-(\log n)^{1+\frac{a-1}{2}}},
\end{aligned}$$

for $n$ large enough.

## 4.e Proof of Proposition III.5

We fix $h \in \mathcal{H}_{\mathsf{hp}}$ and consider the event $\{\hat{h} = h\} \cap \mathcal{E}_{\mathsf{hp}}$. Let $(t_1, \dots, t_d) \in \mathbb{Z}^d$ such that for all $k = 1 : d$,

$$h_k = \beta^{t_k} h_0.$$

For fixed $A$ and $C_A$, we define $t(A, C_A) \in \mathbb{R}$ such that

$$\beta^{t(A,C_A)} h_0 = C_A \, (\log n)^A \, n^{-\frac{1}{2s+r}}.$$

Using (III.9), observe that $t(A, C_A) > 0$ (for $n$ large enough). To simplify the notations, we assume:

$$\mathcal{R} = 1 : r$$

and

$$t_1 \geq t_2 \geq \cdots \geq t_r. \tag{III.26}$$

### 4.e.i Proof of Inequality (III.16)

The bias of $\bar{f}_h(w)$ is denoted $\bar{B}_h$. Note that it does not depend on $\{h_k\}_{k \in \mathcal{R}^c}$. Indeed, we have

$$\begin{aligned}
\bar{B}_h &:= \mathbb{E}\left[\bar{f}_h(w)\right] - f(w) \\
&= \int_{u \in \mathbb{R}^d} \mathrm{K}_h(w-u) \frac{f_{XY}(u)}{\mathrm{f}_X(u_{1:d_1})} du - f(w) \\
&= \int_{u \in \mathbb{R}^d} \mathrm{K}_h(w-u) f(u) du - f(w) \\
&= \int_{z \in \mathbb{R}^d} \left( \prod_{k=1}^{d} K(z_k) \right) [f(w - h \cdot z) - f(w)] \, dz \qquad \text{(III.27)} \\
&= \int_{z' \in \mathbb{R}^r} \left( \prod_{k=1}^{r} K(z'_k) \right) \left[ f_{\mathcal{R}}\left(w_{1:r} - h_{1:r} \cdot z'\right) - f_{\mathcal{R}}(w_{1:r}) \right] dz'.
\end{aligned}$$

We consider the following disjunction of cases:

(Case A) $\mathcal{R} = \emptyset$

(Case B) $\min_{j \in \mathcal{R}} t_j \geq t(A, C_A)$

(Case C) $\exists j \in \mathcal{R}, t_j < t(A, C_A)$.

Then we control the bias in each case.

(Case A) Assume $\mathcal{R} = \emptyset$. In particular, $f$ is constant on the neighborhood $\mathcal{U}$. Note that for any $z \in \mathrm{supp}\,(K)^d$, $w - h \cdot z \in \mathcal{U}$. We then derive from Equation (III.27):

$$\bar{B}_h = 0.$$

(Case B) Assume $\min_{j \in \mathcal{R}} t_j \geq t(A, C_A)$. We apply 2. of Lemma 1 :

$$\begin{aligned}
|\bar{B}_h| &\leq \mathsf{C}_{\bar{B}} \sum_{j \in \mathcal{R}} h_j^s = \mathsf{C}_{\bar{B}} \sum_{j \in \mathcal{R}} \left(\beta^{t_j} h_0\right)^s \\
&\leq \mathsf{C}_{\bar{B}} \times r \left(\beta^{t(A,C_A)} h_0\right)^s = r \mathsf{C}_{\bar{B}} {C_A}^s \, (\log n)^{As} \, n^{-\frac{s}{2s+r}}.
\end{aligned}$$

(Case C) Assume $\exists j \in \mathcal{R}, t_j < t(A, C_A)$. Then we consider

$$j_A = \min\left(j \in \mathcal{R} : t_j < t(A, C_A)\right).$$

In particular, for all $j \geq j_A$,

$$h_j \geq C_A (\log n)^A n^{-\frac{1}{2s+r}}. \tag{III.28}$$

For the previously fixed bandwidth $h$ (and its relevant deactivation times $(t_1, \ldots, t_r)$), we define the following intermediate bandwidths $h^{(\mathsf{int},t)}$, $t \in \mathbb{R}$:

$$h_k^{(\mathsf{int},t)} = \begin{cases} \beta^{t \vee t_k} h_0 & \text{if } k \in \mathcal{R} \\ h_k & \text{else.} \end{cases}$$

Then we decompose the bias by splitting $f(w - h \cdot z) - f(w)$ (note that $h^{(\mathsf{int},t_r)} = h$):

$$\begin{aligned} \bar{B}_h = \int_{z \in \mathbb{R}^d} \left( \prod_{k=1}^d K(z_k) \right) [ & f(w - h^{(\mathsf{int},t(A,C_A))} \cdot z) - f(w) \\ & + f(w - h^{(\mathsf{int},t_{j_A})} \cdot z) - f(w - h^{(\mathsf{int},t(A,C_A))} \cdot z) \\ & + \sum_{j_0 = j_A + 1}^{r} f(w - h^{(\mathsf{int},t_{j_0})} \cdot z) - f(w - h^{(\mathsf{int},t_{j_0 - 1})} \cdot z)] dz \\ = \bar{B}_{h^{(\mathsf{int},t(A,C_A))}} + (\bar{B}_{h^{(\mathsf{int},t_{j_A})}} - \bar{B}_{h^{(\mathsf{int},t(A,C_A))}}) & + \sum_{j_0 = j_A + 1}^{r} \left( \bar{B}_{h^{(\mathsf{int},t_{j_0})}} - \bar{B}_{h^{(\mathsf{int},t_{j_0 - 1})}} \right). \end{aligned} \tag{III.29}$$

For the first term, note that $h^{(\mathsf{int},t(A,C_A))}$ satisfies the condition of (Case B), thus:

$$\left| \bar{B}_{h^{(\mathsf{int},t(A,C_A))}} \right| \leq r \mathsf{C}_{\bar{B}} {C_A}^s (\log n)^{As} n^{-\frac{s}{2s+r}}. \tag{III.30}$$

Let us now control the other terms. The same arguments are used to control the second term $\bar{B}_{h^{(\mathsf{int},t_{j_A})}} - \bar{B}_{h^{(\mathsf{int},t(A,C_A))}}$ or the terms in the sum $\bar{B}_{h^{(\mathsf{int},t_{j_0})}} - \bar{B}_{h^{(\mathsf{int},t_{j_0-1})}}$ for $j_0 = (j_A + 1) : r$. To shorten the proof, the followings lines are applied to the control of the second term by identifying $h^{(\mathsf{int},t_{j_A - 1})}$ to $h^{(\mathsf{int},t(A,C_A))}$ by a slight abuse of notation.
Then, let us fix $j_0 \in \{j_A, \ldots, r\}$. We consider the path between $h_j^{(\mathsf{int},t_{j_0-1})}$ and $h_j^{(\mathsf{int},t_{j_0})}$, namely for $u \in [0,1]$, we denote $h^{[j_0,u]} := h^{(\mathsf{int},t_{j_0-1})} + u \left( h^{(\mathsf{int},t_{j_0})} - h^{(\mathsf{int},t_{j_0-1})} \right)$. Remark that, for any $j = 1 : d$,

$$h_j^{(\mathsf{int},t_{j_0})} - h_j^{(\mathsf{int},t_{j_0-1})} \neq 0 \Rightarrow \left( j \in \mathcal{R} \text{ and } t_j \leq t_{j_0} \right).$$

Indeed, given the definition of $h^{(\mathsf{int},t)}$ for all $t$, each irrelevant component $j$ keeps the value $h_j$. For $j \in \mathcal{R}$, note that $\beta^{t_j \vee t_{j_0}} \neq \beta^{t_j \vee t_{j_0 - 1}} \Rightarrow t_j \leq t_{j_0}$.
Then, we introduce the function $g : u \in [0,1] \mapsto f(w - h^{[j_0,u]} \cdot z)$ (for a fixed $z \in \mathbb{R}^d$). In particular, using the above remark:

$$g'(u) = \sum_{\substack{j \in \mathcal{R} \\ t_j \leq t_{j_0}}} \left( h_j^{(\mathsf{int},t_{j_0})} - h_j^{(\mathsf{int},t_{j_0-1})} \right) \times z_j \partial_j f(w - h^{[j_0,u]} \cdot z).$$

Then we write:

$$\begin{aligned} f(w-h^{(\mathsf{int},t_{j_0})}\cdot z)-f(w-h^{(\mathsf{int},t_{j_0-1})}\cdot z) \\ &= g(1)-g(0)=\int_{u=0}^{1} g'(u)du \\ &= \sum_{\substack{j\in\mathcal{R}\\ t_j\le t_{j_0}}} \int_{u=0}^{1} \left(h_j^{(\mathsf{int},t_{j_0})}-h_j^{(\mathsf{int},t_{j_0-1})}\right)\times z_j\partial_j f(w-h^{[j_0,u]}\cdot z)du. \end{aligned}$$

Hence, we obtain

$$\begin{aligned} \bar{B}_{h^{(\mathsf{int},t_{j_0})}}-\bar{B}_{h^{(\mathsf{int},t_{j_0-1})}} &= \int_{z\in\mathbb{R}^d}\left(\prod_{k=1}^{d}K(z_k)\right)[f(w-h^{(\mathsf{int},t_{j_0})}\cdot z)-f(w-h^{(\mathsf{int},t_{j_0-1})}\cdot z)]dz \\ &= \sum_{\substack{j\in\mathcal{R}\\ t_j\le t_{j_0}}} \int_{u=0}^{1}\left(h_j^{(\mathsf{int},t_{j_0})}-h_j^{(\mathsf{int},t_{j_0-1})}\right)\int_{z\in\mathbb{R}^d}\left(\prod_{k=1}^{d}K(z_k)\right)z_j\partial_j f(w-h^{[j_0,u]}\cdot z)dz\,du \\ &= \sum_{\substack{j\in\mathcal{R}\\ t_j\le t_{j_0}}} \int_{u=0}^{1}\left(h_j^{(\mathsf{int},t_{j_0})}-h_j^{(\mathsf{int},t_{j_0-1})}\right)\mathbb{E}\left[\bar{Z}_{h^{[j_0,u]},j}\right]du, \end{aligned} \tag{III.31}$$

using Equation (III.37):

$$\mathbb{E}\left[\bar{Z}_{h^{[j_0,u]},j}\right]=\int_{\mathbb{R}^d}\left(\prod_{k=1}^{d}K(z_k)\right)z_j\partial_j f(w-h^{[j_0,u]}\cdot z)dz.$$

Now the idea is to control $\left|\mathbb{E}\left[\bar{Z}_{h^{[j_0,u]},j}\right]\right|$ with the test at the iteration $t_j$ on $|Z_{h^{(t_j)},j}|$. More precisely, we will first apply Assumption $\mathcal{M}$ to move from $\left|\mathbb{E}\left[\bar{Z}_{h^{[j_0,u]},j}\right]\right|$ to $\left|\mathbb{E}\left[\bar{Z}_{h^{(t_j)},j}\right]\right|$. Then, we will apply Bernstein's inequality to convert the control on $\left|Z_{h^{(t_j)},j}\right|$ to a control on $\left|\mathbb{E}\left[\bar{Z}_{h^{(t_j)},j}\right]\right|$. Let us fix $j\in\mathcal{R}$ such that $t_j\le t_{j_0}$. We distinguish the cases where the component $j$ is deactivated during the Reverse Step or when it happens during the Direct Step.

Subcase (C.a) $t_j\ge 0$, *i.e.*: $j$ is deactivated during the Direct Step.

Let us show $h^{[j_0,u]}\preccurlyeq h^{(t_j)}$:

- for $k\in\mathcal{R}^c$, since $h_k^{(\mathsf{int},t_{j_0-1})}=h_k=h_k^{(\mathsf{int},t_{j_0})}$,

$$h_k^{[j_0,u]}=h_k.$$

  Remember that the irrelevant components deactivate during the Reverse Step, therefore they already have their final value during the Direct Step. Formally, since $t_k<0\le t_j$, we have

$$h_k^{[j_0,u]}=h_k=\beta^{t_k}h_0=\beta^{t_j\wedge t_k}h_0=h_k^{(t_j)}.$$

- for $k\in\mathcal{R}$, notice $h^{(\mathsf{int},t_{j_0-1})}\preccurlyeq h^{(\mathsf{int},t_{j_0})}$. Therefore:

$$\begin{aligned} h_k^{[j_0,u]} &\le h_k^{(\mathsf{int},t_{j_0})}=\beta^{t_{j_0}\vee t_k}h_0 \\ &\le \beta^{t_j\wedge t_k}h_0=h_k^{(t_j)}. \end{aligned}$$

Then, we have proved $h^{[j_0,u]} \preccurlyeq h^{(t_j)}$. Using Assumption $\mathcal{M}$:

$$\left|\mathbb{E}\left[\bar{Z}_{h^{[j_0,u]},j}\right]\right| \leq \left|\mathbb{E}\left[\bar{Z}_{h^{(t_j)},j}\right]\right|.$$

Subcase (C.b) $t_j < 0$, *i.e.*: $j$ is deactivated during Reverse Step.
As well as $h' \mapsto \bar{B}_{h'}$, $h' \mapsto \mathbb{E}\left[\bar{Z}_{h',j}\right]$ is independent of the irrelevant components of the bandwidth (see for instance Equation (III.37)).
Then we modify the irrelevant components of $h^{[j_0,u]}$ and use the value of the irrelevant components of $h^{(t_j)}$. Formally, we introduce the notation $h^{\{j_0,u\}}$ such that

$$h_k^{\{j_0,u\}} = \begin{cases} h_k^{[j_0,u]} & \text{if } k \in \mathcal{R} \\ h_k^{(t_j)} & \text{else,} \end{cases}$$

so that:

$$\mathbb{E}\left[\bar{Z}_{h^{[j_0,u]},j}\right] = \mathbb{E}\left[\bar{Z}_{h^{\{j_0,u\}},j}\right].$$

Now we just have to verify $h^{\{j_0,u\}} \preccurlyeq h^{(t_j)}$:

- for $k \in \mathcal{R}^c$, by definition of $h^{\{j_0,u\}}$:

$$h_k^{\{j_0,u\}} = h_k^{(t_j)}$$

- for $k \in \mathcal{R}$,

$$\begin{aligned} h_k^{\{j_0,u\}} &= h_k^{[j_0,u]} \\ &\leq h_k^{(\mathsf{int},t_{j_0})} = \beta^{t_{j_0} \vee t_k} h_0 \\ &\leq \beta^{t_j \vee t_k} h_0, \text{ since } t_j \leq t_{j_0}, \\ &\leq \max\left(h_k, \beta^{t_j} h_0\right) =: h_k^{(t_j)}. \end{aligned}$$

Then we have proved $h^{\{j_0,u\}} \preccurlyeq h^{(t_j)}$. Using Assumption $\mathcal{M}$:

$$\left|\mathbb{E}\left[\bar{Z}_{h^{[j_0,u]},j}\right]\right| = \left|\mathbb{E}\left[\bar{Z}_{h^{\{j_0,u\}},j}\right]\right| \leq \left|\mathbb{E}\left[\bar{Z}_{h^{(t_j)},j}\right]\right|.$$

In each case (C.a and C.b), we have proved $\left|\mathbb{E}\left[\bar{Z}_{h^{[j_0,u]},j}\right]\right| \leq \left|\mathbb{E}\left[\bar{Z}_{h^{(t_j)},j}\right]\right|$, then we apply this inequality in Equation (III.31):

$$\begin{aligned} \left|\bar{B}_{h^{(\mathsf{int},t_{j_0})}} - \bar{B}_{h^{(\mathsf{int},t_{j_0-1})}}\right| &\leq \sum_{\substack{j \in \mathcal{R} \\ t_j \leq t_{j_0}}} \int_{u=0}^{1} \left(h_j^{(\mathsf{int},t_{j_0})} - h_j^{(\mathsf{int},t_{j_0-1})}\right) \left|\mathbb{E}\left[\bar{Z}_{h^{[j_0,u]},j}\right]\right| du \\ &\leq \sum_{\substack{j \in \mathcal{R} \\ t_j \leq t_{j_0}}} \int_{u=0}^{1} \left(h_j^{(\mathsf{int},t_{j_0})} - h_j^{(\mathsf{int},t_{j_0-1})}\right) \left|\mathbb{E}\left[\bar{Z}_{h^{(t_j)},j}\right]\right| du \\ &\leq \sum_{\substack{j \in \mathcal{R} \\ t_j \leq t_{j_0}}} \left(h_j^{(\mathsf{int},t_{j_0})} - h_j^{(\mathsf{int},t_{j_0-1})}\right) \left|\mathbb{E}\left[\bar{Z}_{h^{(t_j)},j}\right]\right|. \end{aligned}$$

Then, the previous decomposition of the bias (III.29) leads to:

$$\begin{aligned}
|\bar{B}_h| &\leq \left|\bar{B}_{h^{(\mathsf{int},t(A,C_A))}}\right| + \sum_{j_0=j_A}^{r} \left|\bar{B}_{h^{(\mathsf{int},t_{j_0})}} - \bar{B}_{h^{(\mathsf{int},t_{j_0-1})}}\right| \\
&\leq r\mathsf{C}_{\bar{B}} {C_A}^s (\log n)^{As} n^{-\frac{s}{2s+r}} + \sum_{j_0=j_A}^{r} \sum_{\substack{j\in\mathcal{R} \\ t_j \leq t_{j_0}}} \left(h_j^{(\mathsf{int},t_{j_0})} - h_j^{(\mathsf{int},t_{j_0-1})}\right) \left|\mathbb{E}\left[\bar{Z}_{h^{(t_j)},j}\right]\right| \\
&\leq r\mathsf{C}_{\bar{B}} {C_A}^s (\log n)^{As} n^{-\frac{s}{2s+r}} + \sum_{j=j_A}^{r} \left|\mathbb{E}\left[\bar{Z}_{h^{(t_j)},j}\right]\right| \sum_{j_0=j_A}^{j} \left(h_j^{(\mathsf{int},t_{j_0})} - h_j^{(\mathsf{int},t_{j_0-1})}\right) \\
&\leq r\mathsf{C}_{\bar{B}} {C_A}^s (\log n)^{As} n^{-\frac{s}{2s+r}} + \sum_{j=j_A}^{r} \left|\mathbb{E}\left[\bar{Z}_{h^{(t_j)},j}\right]\right| h_j^{(t_j)},
\end{aligned}$$

since the sum is telescoping, and by noticing that: $h_j^{(\mathsf{int},t_j)} = h_j^{(t_j)}$.

Now, it remains to control $\left|\mathbb{E}\left[\bar{Z}_{h^{(t_j)},j}\right]\right|$ for $j = j_A : r$ using the test at the iteration $t_j$ on $Z_{h^{(t_j)},j}$:

$$\begin{aligned}
\mathbb{1}_{\mathcal{E}_{\mathsf{hp}}\cap\{\hat{h}=h\}} \left|\mathbb{E}\left[\bar{Z}_{h^{(t_j)},j}\right]\right| \leq \mathbb{1}_{\hat{h}=h} \left|Z_{h^{(t_j)},j}\right| + \mathbb{1}_{\widetilde{\mathcal{A}}_n\cap\mathcal{B}\mathsf{ern}_{|\bar{Z}|}(h^{(t_j)},j)} \left|Z_{h^{(t_j)},j} - \bar{Z}_{h^{(t_j)},j}\right| \\
+ \mathbb{1}_{\mathcal{B}\mathsf{ern}_{\bar{Z}}(h^{(t_j)},j)} \left|\bar{Z}_{h^{(t_j)},j} - \mathbb{E}\left[\bar{Z}_{h^{(t_j)},j}\right]\right|.
\end{aligned}$$

By construction of the CDRODEO procedure, if $\hat{h} = h$, then $j$ is deactivated at iteration $t_j$, in other words:

$$\mathbb{1}_{\mathcal{E}_{\mathsf{hp}}\cap\{\hat{h}=h\}} \left|Z_{h^{(t_j)},j}\right| \leq \lambda_{h^{(t_j)},j}.$$

We also apply:

- the definition of $\mathcal{B}\mathsf{ern}_{\bar{Z}}(h^{(t_j)}, j)$:

$$\mathbb{1}_{\mathcal{B}\mathsf{ern}_{\bar{Z}}(h^{(t_j)},j)} \left|\bar{Z}_{h^{(t_j)},j} - \mathbb{E}\left[\bar{Z}_{h^{(t_j)},j}\right]\right| \leq \frac{1}{2}\lambda_{h^{(t_j)},j},$$

- 1. of Lemma 3 (note in particular $\prod_{k=1}^{d} h_k^{(t_j)} \leq 1$):

$$\mathbb{1}_{\widetilde{\mathcal{A}}_n\cap\mathcal{B}\mathsf{ern}_{|\bar{Z}|}(h^{(t_j)},j)} \left|Z_{h^{(t_j)},j} - \bar{Z}_{h^{(t_j)},j}\right| = \mathbb{1}_{\widetilde{\mathcal{A}}_n\cap\mathcal{B}\mathsf{ern}_{|\bar{Z}|}(h^{(t_j)},j)} \left|\Delta_{Z,h^{(t_j)}j}\right| \leq \frac{1}{4}\lambda_{h^{(t_j)},j}.$$

Therefore:

$$\mathbb{1}_{\mathcal{E}_{\mathsf{hp}}\cap\{\hat{h}=h\}} \left|\mathbb{E}\left[\bar{Z}_{h^{(t_j)},j}\right]\right| \leq \mathbb{1}_{\mathcal{E}_{\mathsf{hp}}\cap\{\hat{h}=h\}} \frac{7}{4}\lambda_{h^{(t_j)},j}.$$

Hence:

$$\begin{aligned}
\mathbb{1}_{\mathcal{E}_{\mathsf{hp}}\cap\{\hat{h}=h\}} |\bar{B}_h| &\leq \mathbb{1}_{\{\hat{h}=h\}} \left(r\mathsf{C}_{\bar{B}} {C_A}^s (\log n)^{As} n^{-\frac{s}{2s+r}} + \sum_{j=j_A}^{r} \frac{7}{4}\lambda_{h^{(t_j)},j} \times h_j^{(t_j)}\right), \\
&\leq \mathbb{1}_{\{\hat{h}=h\}} \left(r\mathsf{C}_{\bar{B}} {C_A}^s (\log n)^{As} n^{-\frac{s}{2s+r}} + \sum_{j=j_A}^{r} \frac{7\mathsf{C}_\lambda (\log n)^{a/2}}{4\left(n\prod_{k=1}^{d} h_k^{(t_j)}\right)^{1/2}}\right). \qquad \text{(III.32)}
\end{aligned}$$

Then we control $\prod_{k=1}^{d} h_k^{(t_j)}$ using the same disjunction of subcases as above:

Subcase (C.a) $t_j \geq 0$. At the iteration $t_j \geq 0$, the Direct Step has begun, thus the Reverse Step is over. Since $h \in \mathcal{H}_{\mathsf{hp}}$, the irrelevant components have already their final value: for all $k \in \mathcal{R}^c$,

$$1 \geq h_k^{(t_j)} = h_k = h_{\mathsf{irr}} > \beta.$$

Moreover, during the Direct Step, at iteration $t_j$, all components are lower bounded by the current active bandwidth value $\beta^{t_j} h_0$, *i.e.*: for any $k \in \mathcal{R}$,

$$h_k^{(t_j)} \geq \beta^{t_j} h_0.$$

Recall that $j \geq j_A$, thus:

$$t_j \leq t_{j_A} \leq t(A, C_A).$$

It follows:

$$h_k^{(t_j)} \geq \beta^{t(A,C_A)} h_0 = C_A (\log n)^A n^{-\frac{1}{2s+r}}.$$

Therefore:

$$\prod_{k=1}^{d} h_k^{(t_j)} \geq \beta^{d-r} \left( C_A (\log n)^A n^{-\frac{1}{2s+r}} \right)^r.$$

Then the upper bound in Equation (III.32) becomes:

$$\begin{aligned} \frac{7\mathsf{C}_\lambda (\log n)^{a/2}}{4 \left( n \prod_{k=1}^{d} h_k^{(t_j)} \right)^{1/2}} &\leq \frac{7\mathsf{C}_\lambda}{4\beta^{\frac{d-r}{2}} {C_A}^{\frac{r}{2}}} (\log n)^{\frac{a-Ar}{2}} n^{-\frac{1}{2}\left(1-\frac{r}{2s+r}\right)} \\ &= \frac{7\mathsf{C}_\lambda}{4\beta^{\frac{d-r}{2}} {C_A}^{\frac{r}{2}}} (\log n)^{\frac{a-Ar}{2}} n^{-\frac{s}{2s+r}}. \end{aligned}$$

Subcase (C.b) $t_j < 0$. At iteration $t_j$, only iterations of the Reverse Step have been performed. Thus, the current bandwidth has only been increased. Therefore:

$$\frac{7\mathsf{C}_\lambda (\log n)^{a/2}}{4 \left( n \prod_{k=1}^{d} h_k^{(t_j)} \right)^{1/2}} \leq \frac{7\mathsf{C}_\lambda (\log n)^{a/2}}{4 \left( n h_0^d \right)^{1/2}}.$$

Remark that $h_0$'s lower bound (III.9) is exactly defined so, we have

$$\frac{7\mathsf{C}_\lambda (\log n)^{a/2}}{4 \left( n h_0^d \right)^{1/2}} \leq \frac{7}{4} \left( \frac{(\log n)^a}{n} \right)^{\frac{p}{2p+1}}.$$

Note that $n^{-\frac{p}{2p+1}}$ is smaller than the minimax optimal rate for any regularity and any sparsity structure (except for the degenerate case where $r = 0$ and which is solved separately: cf (Case A)):

$$n^{-\frac{p}{2p+1}} = \min_{\substack{1 \leq r' \leq d \\ 1 \leq s' \leq p}} \left( n^{-\frac{s'}{2s'+r'}} \right).$$

When we reunite the two subcases, Inequality (III.32) becomes:

$$\begin{aligned} \mathbb{1}_{\mathcal{E}_{\mathsf{hp}} \cap \{\hat{h}=h\}} |\bar{B}_h| &\leq r \mathsf{C}_{\bar{B}} {C_A}^s (\log n)^{As} n^{-\frac{s}{2s+r}} \\ &\quad + r \times \max \left( \frac{7\mathsf{C}_\lambda}{4\beta^{\frac{d-r}{2}} {C_A}^{\frac{r}{2}}} \frac{(\log n)^{\frac{a-Ar}{2}}}{n^{\frac{s}{2s+r}}}, \frac{7}{4} \left( \frac{(\log n)^a}{n} \right)^{\frac{p}{2p+1}} \right), \end{aligned}$$

which concludes the proof of Inequality (III.16).

### 4.e.ii Proof of Inequality (III.17)

Let us now prove the second inequality (III.17). By definition: $\mathcal{E}_{\mathsf{hp}} \subset \mathcal{B}\mathrm{ern}_{\bar{f}}(h)$. Thus, we have

$$\mathbb{1}_{\{\hat{h}=h\}\cap\mathcal{E}_{\mathsf{hp}}} \left|\bar{f}_h(w) - \mathbb{E}\left[\bar{f}_h(w)\right]\right| \leq \sigma_h := \mathsf{C}_\sigma \sqrt{\frac{(\log n)^a}{n\prod_{k=1}^d h_k}}.$$

Two cases occur: in the first case, the deviation is controlled by a concentration inequality; in the second case, we control the deviation by $\mathbb{E}Z_{hj}$ thanks to the tests on the $Z_{hj}$'s.

1. $\max_{k\in\mathcal{R}} t_k \leq t(A,C_A)$. Then, $\forall k \in \mathcal{R}$:

$$h_k = \beta^{t_k} h_0 > \beta^{t(A,C_A)} h_0 = C_A (\log n)^A n^{-\frac{1}{2s+r}}.$$

Besides, for $k \in \mathcal{R}^c$:

$$h_k = h_{\mathsf{irr}} > \beta.$$

Therefore:

$$\sigma_h \leq \mathsf{C}_\sigma \sqrt{\frac{(\log n)^a}{n\beta^{d-r}\left(C_A(\log n)^A n^{-\frac{1}{2s+r}}\right)^r}} = \frac{\mathsf{C}_\sigma}{\beta^{(d-r)/2}{C_A}^{r/2}} (\log n)^{(a-Ar)/2} n^{-\frac{s}{2s+r}}.$$

2. $\max_{k\in\mathcal{R}} t_k > t(A,C_A)$. First remark that for any $k = 1:d$,

$$\sigma_h = \frac{\mathsf{C}_\sigma}{\mathsf{C}_\lambda} h_k \, \lambda_{hk}.$$

Hence, it suffices to control the threshold in order to bound the deviation. Let us consider $j_0 \in \arg\max_{k\in\mathcal{R}} t_k$ (actually assuming (III.26) means that $j_0 = 1$). In particular, when $\hat{h} = h$, the component $j_0$ is deactivated during the last iteration, and during the Direct Step (recall that $t(A,C_A) > 0$). Let us consider the penultimate iteration, i.e. Iteration $t_{j_0} - 1$. At this iteration, $j_0$ is not deactivated, *i.e.*:

$$\mathbb{1}_{\hat{h}=h} \left|Z_{h^{(t_{j_0}-1)}\, j_0}\right| > \mathbb{1}_{\hat{h}=h} \lambda_{h^{(t_{j_0}-1)}\, j_0}.$$

Then we use 1. of Lemma 3. Note that $\prod_{k=1}^d h_k^{(t_{j_0}-1)} \leq 1$, thus:

$$\mathbb{1}_{\mathcal{E}_{\mathsf{hp}}} \left|\Delta_{Z,h^{(t_{j_0}-1)}\, j_0}\right| \leq \frac{1}{4} \lambda_{h^{(t_{j_0}-1)}\, j_0}.$$

Remember the definition of $\mathcal{B}\mathrm{ern}_{\bar{Z}}(h,j)$, thus

$$\mathbb{1}_{\mathcal{E}_{\mathsf{hp}}} \left|\bar{Z}_{h^{(t_{j_0}-1)}\, j_0} - \mathbb{E}\left[\bar{Z}_{h^{(t_{j_0}-1)}\, j_0}\right]\right| \leq \frac{1}{2} \lambda_{h^{(t_{j_0}-1)}\, j_0}.$$

Therefore:

$$\mathbb{1}_{\{\hat{h}=h\}\cap\mathcal{E}_{\mathsf{hp}}} \left|\mathbb{E}\left[\bar{Z}_{h^{(t_{j_0}-1)}\, j_0}\right]\right| > \mathbb{1}_{\{\hat{h}=h\}\cap\mathcal{E}_{\mathsf{hp}}} \frac{1}{4} \lambda_{h^{(t_{j_0}-1)}\, j_0}. \tag{III.33}$$

Let us compare $h^{(t_{j_0}-1)}$ to $h$. Recall $h = h^{(t_{j_0})}$, since $t_{j_0}$ is the final iteration of our algorithm. We have:

– for $k \in \mathcal{R}^c$, $h_k^{(t_{j_0}-1)} = h_k$. Indeed, $t_k < 0$, hence the components $k$ have been deactivated before Iteration $t_{j_0} - 1$, and have the same value for the last two iterations.

– for $k \in \mathcal{R}$, $h_k \geq \beta h_k^{(t_{j_0}-1)}$. Indeed, at worst, the component $k$ was active during Iteration $t_{j_0} - 1$ and have been multiplied by $\beta$.

Therefore:

$$\prod_{k=1}^{d} h_k \geq \beta^r \prod_{k=1}^{d} h_k^{(t_{j_0}-1)}$$

and

$$h_{j_0}\lambda_{hj_0} = \mathsf{C}_\lambda \sqrt{\frac{(\log n)^a}{n\prod_{k=1}^d h_k}} \leq \beta^{-\frac{r}{2}} h_{j_0}^{(t_{j_0}-1)} \lambda_{h^{(t_{j_0}-1)}j_0}.$$

To summarize, we have

$$\begin{aligned}
\mathbb{1}_{\{\hat{h}=h\}\cap\mathcal{E}_{\mathsf{hp}}} \left|\bar{f}_h(w) - \mathbb{E}\left[\bar{f}_h(w)\right]\right| &\leq \mathbb{1}_{\{\hat{h}=h\}\cap\mathcal{E}_{\mathsf{hp}}} \sigma_h = \mathbb{1}_{\{\hat{h}=h\}\cap\mathcal{E}_{\mathsf{hp}}} \frac{\mathsf{C}_\sigma}{\mathsf{C}_\lambda} h_{j_0}\lambda_{hj_0} \\
&\leq \mathbb{1}_{\{\hat{h}=h\}\cap\mathcal{E}_{\mathsf{hp}}} \beta^{-\frac{r}{2}} \frac{\mathsf{C}_\sigma}{\mathsf{C}_\lambda} h_{j_0}^{(t_{j_0}-1)} \lambda_{h^{(t_{j_0}-1)}j_0} \\
&\leq \mathbb{1}_{\{\hat{h}=h\}\cap\mathcal{E}_{\mathsf{hp}}} 4\beta^{-\frac{r}{2}} \frac{\mathsf{C}_\sigma}{\mathsf{C}_\lambda} h_{j_0}^{(t_{j_0}-1)} \left|\mathbb{E}\left[\bar{Z}_{h^{(t_{j_0}-1)}j_0}\right]\right|.
\end{aligned}$$

Then we apply 2. of Lemma 2:

$$\left|\mathbb{E}\left[\bar{Z}_{h^{(t_{j_0}-1)}j_0}\right]\right| \leq \mathsf{C}_{E\bar{Z}} \left(h_{j_0}^{(t_{j_0}-1)}\right)^{s-1}.$$

Therefore:

$$\begin{aligned}
\mathbb{1}_{\{\hat{h}=h\}\cap\mathcal{E}_{\mathsf{hp}}} \left|\bar{f}_h(w) - \mathbb{E}\left[\bar{f}_h(w)\right]\right| &\leq \mathbb{1}_{\{\hat{h}=h\}\cap\mathcal{E}_{\mathsf{hp}}} 4\beta^{-\frac{r}{2}} \frac{\mathsf{C}_\sigma}{\mathsf{C}_\lambda} h_{j_0}^{(t_{j_0}-1)} \times \mathsf{C}_{E\bar{Z}} \left(h_{j_0}^{(t_{j_0}-1)}\right)^{s-1} \\
&\leq \frac{4\mathsf{C}_{E\bar{Z}}\mathsf{C}_\sigma\beta^{-\frac{r}{2}}}{\mathsf{C}_\lambda} \left(\beta^{t_{j_0}-1}h_0\right)^s = \frac{4\mathsf{C}_{E\bar{Z}}\mathsf{C}_\sigma\beta^{-\frac{r}{2}-s}}{\mathsf{C}_\lambda} \left(\beta^{t_{j_0}}h_0\right)^s \\
&\leq \frac{4\mathsf{C}_{E\bar{Z}}\mathsf{C}_\sigma\beta^{-\frac{r}{2}-s}}{\mathsf{C}_\lambda} \left(\beta^{t(A,C_A)}h_0\right)^s = \frac{4\mathsf{C}_{E\bar{Z}}\mathsf{C}_\sigma\beta^{-\frac{r}{2}-s}}{\mathsf{C}_\lambda} \left(C_A(\log n)^A n^{-\frac{1}{2s+r}}\right)^s \\
&= \frac{4{C_A}^s\mathsf{C}_{E\bar{Z}}\mathsf{C}_\sigma\beta^{-\frac{r}{2}-s}}{\mathsf{C}_\lambda} (\log n)^{sA} n^{-\frac{s}{2s+r}}.
\end{aligned}$$

Reuniting the two cases, we obtain Inequality (III.17):

$$\begin{aligned}
\mathbb{1}_{\{\hat{h}=h\}\cap\mathcal{E}_{\mathsf{hp}}} \left|\bar{f}_h(w) - \mathbb{E}\left[\bar{f}_h(w)\right]\right| &\leq \mathbb{1}_{\{\hat{h}=h\}\cap\mathcal{E}_{\mathsf{hp}}} \sigma_h \\
&\leq \max\left(\frac{\mathsf{C}_\sigma}{\beta^{(d-r)/2}{C_A}^{r/2}} (\log n)^{(a-Ar)/2}, \frac{4{C_A}^s\mathsf{C}_{E\bar{Z}}\mathsf{C}_\sigma\beta^{-\frac{r}{2}-s}}{\mathsf{C}_\lambda} (\log n)^{sA}\right) n^{-\frac{s}{2s+r}}.
\end{aligned}$$

## 4.f Proof of Proposition III.3

Let us evaluate the number of operations of our procedure. During the Reverse Step, each bandwidth of $\mathcal{A}\mathrm{ct}^{(-1)}$ can be multiplied by $\beta^{-1}$ several times until the loop condition is achieved:

$$(\mathcal{A}\mathrm{ct}^{(t)} \neq \emptyset) \& (\max \hat{h}_k^{(t)} \leq \beta).$$

In particular, $\max \hat{h}_k^{(t)} \leq 1$. Since $\hat{h}_k^{(t)} = h_0 \beta^{-|t_k|}$,

$$|t_k| = \log\left(\frac{\hat{h}_k^{(t)}}{h_0}\right) / \log\left(\beta^{-1}\right) \leq \frac{\log(h_0^{-1})}{\log(\beta^{-1})} = O\left(\frac{\log(n)}{d(2p+1)}\right)$$

using the lower bound on $h_0$ (III.9). Thus, during this Reverse Step, note that only $|\mathcal{A}\text{ct}^{(-1)}|$ components are updated and:

- the number of updates of the $Z_{hj}$'s is of order $\frac{|\mathcal{A}\text{ct}^{(-1)}|}{d(2p+1)}\log(n)$ given the above remark,
- the computation of the $Z_{hj}$'s and the comparison to the threshold cost $\mathcal{O}(|\mathcal{A}\text{ct}^{(-1)}|n)$ operations.

Therefore at worst, there are $\mathcal{O}\left(\frac{|\mathcal{A}\text{ct}^{(-1)}|^2}{d}\log(n)n\right)$ operation during the Reverse Step.

For the Direct Step, the stopping condition is $\left(\prod_{k=1}^{d} \hat{h}_k^{(t)} > \frac{(\log n)^{1+a}}{n}\right)$, which is satisfied for the penultimate iteration, hence:

$$\prod_{k=1}^{d} \hat{h}_k > \beta^d \frac{(\log n)^{1+a}}{n}.$$

We denote $t_k$ the deactivation times of $\hat{h}$, then

$$h_0^d \beta^{\Sigma_{k=1}^d t_k} > \beta^d \frac{(\log n)^{1+a}}{n},$$

which gives

$$\sum_{k=1}^{d} t_k < \frac{\log(\beta^{-d}(\log n)^{-(1+a)} n h_0^d)}{\log(1/\beta)}.$$

Thus, during the Direct Step, note that only $|\mathcal{A}\text{ct}^{(0)}|$ components are updated and

- the total number of updates of the $Z_{hj}$'s is of order $\log_{\frac{1}{\beta}}(n)$ given the above remark,
- the computation of the $Z_{hj}$'s and the comparison to the threshold cost $\mathcal{O}(|\mathcal{A}\text{ct}^{(0)}|n)$ operations.

Therefore at worst, there are $\mathcal{O}(|\mathcal{A}\text{ct}^{(-1)}|\log(n)n)$ operations during the Direct Step. Using $|\mathcal{A}\text{ct}^{(-1)}| + |\mathcal{A}\text{ct}^{(0)}| \leq d$, the sum of these two steps leads to the proposition.

# 5 Appendix

## 5.a Lemmas

The following lemmas are mainly proved in Nguyen [2018]. Note that some adjustments have been made from their initial versions. In particular, we have refined points 2. of Lemma 1 and of Lemma 2 to take into account the extension of our results to Hölder smoothness.
In the sequel, we only prove results of subsequent lemmas which were not established in Nguyen [2018].

**Lemma 1** (Lemma 5 of Nguyen [2018]: $\bar{f}_h(w)$ behaviour). *Under Assumption $\mathcal{L}_X$, for any bandwidth $h \in (0,1]^d$, and any $i = 1:n$,*

1. *Let* $C_{\bar{E}} := \|f\|_{\infty,\mathcal{U}}\|K\|_1^d$. *Then*

$$\left|\mathbb{E}\bar{f}_{h1}(w)\right| \leq \mathbb{E}\left|\bar{f}_{h1}(w)\right| \leq C_{\bar{E}}.$$

2. *If* $f$ *has only* $r$ *relevant components* $\mathcal{R}$ *and belongs to* $\mathcal{H}_d(s,L)$ *and if the order* $p$ *of the kernel* $K$ *is larger than or equal to* $s$,

$$|\bar{B}_h| \leq C_{\bar{B}} \sum_{k\in\mathcal{R}} h_k^s, \tag{III.34}$$

*with* $C_{\bar{B}} > 0$ *a constant only depending on* $L$, $s$ *and* $K$.

3. *Let* $\mathcal{B}ern_{\bar{f}}(h) := \{|\bar{f}_h(w) - \mathbb{E}[\bar{f}_h(w)]| \leq \sigma_h\}$, *where* $\sigma_h := C_\sigma \sqrt{\frac{(\log n)^a}{n\prod_{k=1}^d h_k}}$ *with* $C_\sigma = \frac{2\|K\|_2^d \|f\|_{\infty,\mathcal{U}}^{\frac{1}{2}}}{\delta^{\frac{1}{2}}}$. *If*

*Cond(h):* $\prod_{k=1}^d h_k \geq \frac{4^2\|K\|_\infty^{2d}}{9\delta^2 C_\sigma^2}\frac{(\log n)^a}{n}$ *is satisfied, then:*

$$\mathbb{P}\left(\mathcal{B}ern_{\bar{f}}(h)^c\right) \leq 2e^{-(\log n)^a}.$$

4. *Let* $\mathcal{B}ern_{|\bar{f}|}(h) := \{\left|\frac{1}{n}\sum_{i=1}^n |\bar{f}_{hi}(w)| - \mathbb{E}[|\bar{f}_h(w)|]\right| \leq C_{\bar{E}}\}$. *Then*

$$\mathbb{P}\left(\mathcal{B}ern_{|\bar{f}|}(h)^c\right) \leq 2e^{-C_{\gamma|f|}n\prod_{k=1}^d h_k},$$

*with* $C_{\gamma|f|} := \min\left(\frac{C_{\bar{E}}^2}{C_\sigma^2}; \frac{3\delta C_{\bar{E}}}{4\|K\|_\infty^d}\right)$.

**Lemma 2** (Lemma 6 of Nguyen [2018]: $\bar{Z}_{hj}$ behaviour). *If* $K$ *is chosen as in Section 3.a, and under Assumption* $\mathcal{L}_X$, *for any* $j \in \{1,\dots,d\}$ *and any bandwidth* $h \in (0,h_0]^d$, *we have the following results.*

1. *Let* $C_{E|\bar{Z}|} := \|f\|_{\infty,\mathcal{U}}\|J\|_1\|K\|_1^{d-1}$. *We have*

$$\mathbb{E}|\bar{Z}_{h1j}| \leq C_{E|\bar{Z}|}h_j^{-1}.$$

2. *If* $f$ *has only* $r$ *relevant components* $\mathcal{R}$, *for* $j \notin \mathcal{R}$:

$$\mathbb{E}\bar{Z}_{hj} = 0,$$

*and if in addition* $f$ *belongs to* $\mathcal{H}_d(s,L)$, *for* $j \in \mathcal{R}$:

$$|\mathbb{E}[\bar{Z}_{h,j}]| \leq C_{E\bar{Z}}h_j^{s-1}, \tag{III.35}$$

*where* $C_{E\bar{Z}} := (\int |z^s K(z)|dz)\frac{\|K\|_1^{r-1}L}{(s-1)!}$ *denoting* $(s-1)! := (s-q+1)(s-q+2)\dots(s-1)$.

3. *Let* $\mathcal{B}ern_{\bar{Z}}(h,j) := \{|\bar{Z}_{hj} - \mathbb{E}\bar{Z}_{hj}| \leq \frac{1}{2}\lambda_{hj}\}$. *If the bandwidth satisfies:*

*Cond*$_{\bar{Z}}(h)$: $\prod_{k=1}^d h_k \geq cond_{\bar{Z}}\frac{(\log n)^a}{n}$, *with* $cond_{\bar{Z}} := \frac{4\|J\|_\infty^2\|K\|_\infty^{2(d-1)}}{3^2\|f\|_{\infty,\mathcal{U}}\|J\|_2^2\|K\|_2^{2(d-1)}}$,

*then:*

$$\mathbb{P}\left(\mathcal{B}ern_{\bar{Z}}(h,j)^c\right) \leq 2e^{-\frac{\delta}{\|f\|_{\infty,\mathcal{U}}}(\log n)^a}.$$

4. *Let* $\mathcal{B}ern_{|\bar{Z}|}(h,j) := \{|\frac{1}{n}\sum_{i=1}^{n}|\bar{Z}_{hij}| - \mathbb{E}|\bar{Z}_{h1j}|| \leq C_{E|\bar{Z}|}h_j^{-1}\}$. *Then,*

$$\mathbb{P}\left(\mathcal{B}ern_{|\bar{Z}|}(h,j)^c\right) \leq 2e^{-C_{\gamma|\bar{Z}|}n\prod_{k=1}^{d}h_k},$$

*with* $C_{\gamma|\bar{Z}|} := \min\left(\frac{\delta C_{E|\bar{Z}|}^2}{4\|f\|_{\infty,\,\mathcal{U}}\|J\|_2^2\|K\|_2^{2(d-1)}}; \frac{3\delta C_{E|\bar{Z}|}}{4\|K\|_\infty^{d-1}\|J\|_\infty}\right)$.

**Lemma 3.** *For any* $h \in \mathcal{H}_{hp}^{Rev} \cup \mathcal{H}_{hp}^{Dir}$ *and any component* $j \in \{1:d\}$, *under Assumptions* $\mathcal{L}_X$ *and* $\mathcal{E}f_X$, *if* $\sqrt{\prod_{k=1}^{d}h_k} \leq 1$, *then*

1. *we have:*

$$\mathbb{1}_{\mathcal{B}ern_{|\bar{Z}|}(hj)\cap\widetilde{\mathcal{A}}_n}\left|\Delta_{Z,hj}\right| \leq \frac{1}{4}\lambda_{hj}$$

2. *for* $C_{M\Delta} := \frac{4M_X C_{\bar{E}}}{\delta C_\sigma}$:

$$\mathbb{1}_{\widetilde{\mathcal{A}}_n\cap\mathcal{B}ern_{|\bar{f}|}(h)}\left|\Delta_h\right| \leq C_{M\Delta}\sigma_h.$$

**Lemma 4** (Taylor's theorem). *Let* $g:[0,1]\to\mathbb{R}$ *be a function of class* $\mathcal{C}^q$. *Then we have:*

$$g(1)-g(0) = \sum_{l=1}^{q}\frac{g^{(l)}(0)}{l!} + \int_{t_1=0}^{1}\int_{t_2=0}^{t_1}\dots\int_{t_q=0}^{t_{q-1}}(g^{(q)}(t_q)-g^{(q)}(0))dt_q dt_{q-1}\dots dt_1.$$

## 5.b Proof of Inequality (III.34) in Lemma 1

We recall that the notation $\cdot$ means the multiplication term by term of two vectors, then we have:

$$\begin{aligned}\bar{B}_h = \mathbb{E}\bar{f}_h(w) - f(w) &= \int_{u\in\mathbb{R}^d}\left(\prod_{k=1}^{d}\frac{K(h_k^{-1}(w_k-u_k))}{h_k}\right)f(u)du - f(w)\\ &= \int_{z\in\mathbb{R}^d}\left(\prod_{k=1}^{d}K(z_k)\right)(f(w-h\cdot z)-f(w))dz.\end{aligned}$$

For any $z\in\mathbb{R}^d$, let us introduce the notations $\bar{z}_0 := w$ and for $k=1,\dots,d$, $\bar{z}_k := w - \sum_{j=1}^{k}h_j z_j e_j$, where $\{e_j\}_{j=1}^d$ is the canonical basis of $\mathbb{R}^d$. Then, we write:

$$f(w-h.z)-f(w) = \sum_{k=1}^{d}f(\bar{z}_k)-f(\bar{z}_{k-1}) = \sum_{k\in\mathcal{R}}f(\bar{z}_k)-f(\bar{z}_{k-1}),$$

since for $k\notin\mathcal{R}$, $f(\bar{z}_k)-f(\bar{z}_{k-1})=0$. We apply Taylor's theorem (cf Lemma 4) to the functions $g_k : t\in[0,1]\mapsto f(\bar{z}_{k-1}-th_k z_k e_k)$, $k\in\mathcal{R}$:

$$f(\bar{z}_k)-f(\bar{z}_{k-1}) = g_k(1)-g_k(0) = \sum_{l=1}^{q}\frac{(-z_k h_k)^l}{l!}\partial_k^l f(\bar{z}_{k-1}) + J_k,$$

where we recall that $q$ is the largest integer smaller than $s$ and with

$$\begin{aligned} J_k &:= \int_{0\le t_q\le\cdots\le t_1\le 1} \left(g_k^{(q)}(t_q) - g_k^{(q)}(0)\right) dt_{1:q} \\ &= (-h_k z_k)^q \int_{0\le t_q\le\cdots\le t_1\le 1} \left(\partial_k^q f(\bar{z}_{k-1} - t_q h_k z_k e_k) - \partial_k^q f(\bar{z}_{k-1})\right) dt_{1:q}. \end{aligned}$$

We denote $\mathsf{I}_k := \int_{z\in\mathbb{R}^d} \left(\prod_{k'=1}^d K(z_{k'})\right) J_k dz$ and for any $z \in \mathbb{R}^d$, we denote $z_{-k} \in \mathbb{R}^{d-1}$ the vector $z$ without its $k^{th}$ variable, then we obtain:

$$\begin{aligned} \bar{B}_h &= \sum_{k\in\mathcal{R}} \int_{z\in\mathbb{R}^d} \left(\prod_{k'=1}^d K(z_{k'})\right) \left(J_k + \sum_{l=1}^q \frac{(-h_k)^l}{l!} \partial_k^l f(\bar{z}_{k-1}) z_k^l\right) dz \\ &= \sum_{k\in\mathcal{R}} \left(\mathsf{I}_k + \sum_{l=1}^q \mathsf{II}_{k,l}\right), \end{aligned}$$

where

$$\begin{aligned} \mathsf{II}_{k,l} :&= \int_{z_{-k}\in\mathbb{R}^{d-1}} \left(\prod_{k'\neq k} K(z_{k'})\right) \frac{(-h_k)^l}{l!} \partial_k^l f(\bar{z}_{k-1}) \int_{z_k\in\mathbb{R}} z_k^l K(z_k) dz_k dz_{-k} \\ &= \frac{(-h_k)^l}{l!} \int_{z_{-k}\in\mathbb{R}^{d-1}} \partial_k^l f(\bar{z}_{k-1}) \left(\prod_{k'\neq k} K(z_{k'})\right) dz_{-k} \times \int_{t\in\mathbb{R}} t^l K(t) dt = 0, \end{aligned}$$

since $K$ is of order $p \ge s > q$. So,

$$\bar{B}_h = \sum_{k\in\mathcal{R}} \mathsf{I}_k.$$

Now we control $|J_k|$:

$$\begin{aligned} |J_k| &\le |h_k z_k|^q \left| \int_{0\le t_q\le\cdots\le t_1\le 1} \left[\partial_k^q f(\bar{z}_{k-1} - t_q h_k z_k e_k) - \partial_k^q f(\bar{z}_{k-1})\right] dt_{1:q} \right| \\ &\le |h_k z_k|^q \int_{0\le t_q\le\cdots\le t_1\le 1} L|t_q h_k z_k|^{s-q} dt_{1:q} = \frac{L(h_k|z_k|)^s}{s(s-1)\dots(s-q)}. \end{aligned}$$

So:

$$|\mathsf{I}_k| = \left| \int_{z\in\mathbb{R}^d} \left(\prod_{k'=1}^d K(z_{k'})\right) J_k dz \right| \le \frac{L\|K\|_1^{d-1} \|(\cdot)^s K(\cdot)\|_1}{s(s-1)\dots(s-q)} {h_k}^s.$$

Finally,

$$|\bar{B}_h| \le \mathsf{C}_{\bar{B}} \sum_{k\in\mathcal{R}} h_k^s, \tag{III.36}$$

with $\mathsf{C}_{\bar{B}} := \frac{L\|K\|_1^{d-1} \|(\cdot)^s K(\cdot)\|_1}{s(s-1)\dots(s-q)}$.

## 5.c Proof of Inequality (III.35) in Lemma 2

Let $j \in \mathcal{R}$. Denoting $J : \mathbb{R} \to \mathbb{R}$ the function $t \mapsto tK'(t) + K(t)$, we can write

$$\bar{Z}_{h,j} = \frac{1}{n}\sum_{i=1}^{n} \frac{-J(\frac{w_j - W_{ij}}{h_j}) \prod_{k\neq j} K(\frac{w_k - W_{ik}}{h_k})}{\mathrm{f}_X(X_i) h_j \prod_{k=1}^{d} h_k}.$$

Then, taking the expectation,

$$\mathbb{E}[\bar{Z}_{hj}] = -\frac{1}{h_j}\int_{\mathbb{R}^d} J(z_j)\left(\prod_{k\neq j} K(z_k)\right) f(w - h\cdot z)dz.$$

To simplify the notations, we assume $\mathcal{R} = \{1,\dots,r\}$. Then, by integration by part

$$\begin{aligned}\mathbb{E}[\bar{Z}_{h,j}] &= \int_{\mathbb{R}^d} \left(z_j K(z_j)\right)\left(\prod_{k\neq j} K(z_k)\right)\partial_j f(w - h\cdot z)dz \\ &= \int_{\mathbb{R}^r}\left(\prod_{k\in\mathcal{R}} K(z_k)\right) z_j \partial_j f_{\mathcal{R}}(w_{1:r} - (h.z)_{1:r})dz_{1:r}, \end{aligned} \tag{III.37}$$

where $f_{\mathcal{R}}$ is the restriction of $f$ to the first $r$ components (remember that for any $u \in \mathbb{R}^r$ and any $v \in \mathbb{R}^{d-r}$ $f_{\mathcal{R}}(u) := f_{\mathcal{R}}(u,v)$ does not depend on $v$). Let us denote by $G_{j,z,h} : [0,1] \to \mathbb{R}$ the function

$$t \mapsto \partial_j f_{\mathcal{R}}(w_1 - h_1 z_1, \dots, w_j - t h_j z_j, \dots, w_r - h_r z_r).$$

Then

$$\begin{aligned}\mathbb{E}[\bar{Z}_{h,j}] &= \int_{\mathbb{R}^r}\left(\prod_{k\in\mathcal{R}} K(z_k)\right) z_j G_{j,z,h}(1)dz_{1:r} \\ &= \int_{\mathbb{R}^r}\left(\prod_{k\in\mathcal{R}} K(z_k)\right) z_j \{G_{j,z,h}(1) - G_{j,z,h}(0)\}dz_{1:r},\end{aligned}$$

since the order $p$ of $K$ satisfies: $p \geq s > q \geq 1$. Next we use the Taylor expansion given by Lemma 4:

$$G_{j,z,h}(1) - G_{j,z,h}(0) = \sum_{l=1}^{q-1} \frac{G^{(l)}_{j,z,h}(0)}{l!} + R'_{j,z,h,q-1}, \tag{III.38}$$

where $R'_{j,z,h,q-1} := \int_{t_1=0}^{1}\int_{t_2=0}^{t_1}\dots\int_{t_{q-1}=0}^{t_{q-2}}(G^{(q-1)}_{j,z,h}(t_{q-1}) - G^{(q-1)}_{j,z,h}(0))dt_{q-1}dt_{q-2}\dots dt_1$. But

$$G^{(l)}_{j,z,h}(t) = (-h_j z_j)^l \partial_j^{l+1} f_{\mathcal{R}}(w_1 - h_1 z_1, \dots, w_j - t h_j z_j, \dots, w_r - h_r z_r).$$

Then, the first $q-1$ terms in the r.h.s. of (III.38) vanish since $\int z_j^{l+1} K(z_j)dz_j = 0$. Now, we will bound the integral remainder of (III.38). Using that $f$ belongs to $\mathcal{H}_d(s,L)$, for all $t \in [0,1]$,

$$\left|G^{(q-1)}_{j,z,h}(t) - G^{(q-1)}_{j,z,h}(0)\right| \leq |h_j z_j|^{q-1} L |t h_j z_j|^{s-q},$$

since $w - h\cdot z + (1-t)h_j z_j e_j \in \mathcal{U}$. Hence

$$|R'_{j,z,h,q-1}| \leq \int_{t_1=0}^{1}\int_{t_2=0}^{t_1}\cdots\int_{t_{q-1}=0}^{t_{q-2}}\left|G_{j,z,h}^{(q-1)}(t_{q-1}) - G_{j,z,h}^{(q-1)}(0)\right| dt_{q-1}dt_{q-2}\ldots dt_1$$
$$\leq L(h_j|z_j|)^{s-1}\int_{t_1=0}^{1}\int_{t_2=0}^{t_1}\cdots\int_{t_{q-1}=0}^{t_{q-2}} t_{q-1}^{s-q}dt_{q-1}dt_{q-2}\ldots dt_1 = \frac{L(h_j|z_j|)^{s-1}}{(s-1)!},$$

denoting $(s-1)! := (s-q+1)(s-q+2)\ldots(s-1)$. Finally,

$$|\mathbb{E}[\bar{Z}_{h,j}]| = \left|\int_{\mathbb{R}^r}\left(\prod_{k\in\mathcal{R}}K(z_k)\right)z_j R'_{j,z,h,q-1}dz_{1:r}\right| \leq \int_{\mathbb{R}^r}\left(\prod_{k\in\mathcal{R}}|K(z_k)|\right)|z_j|\frac{L(h_j|z_j|)^{s-1}}{(s-1)!}dz_{1:r}$$
$$\leq \frac{Lh_j^{s-1}}{(s-1)!}\left(\prod_{k\in\mathcal{R}\setminus\{j\}}\|K\|_1\right)\int_{\mathbb{R}}|z_j|^s|K(z_j)|dz_{1:r} \leq \mathsf{C}_{E\bar{Z}}h_j^{s-1},$$

denoting $\mathsf{C}_{E\bar{Z}} := \left(\int_{\mathbb{R}}|z|^s|K(z)|dz\right)\|K\|_1^{r-1}L/(s-1)!$.

## 5.d Proof of Lemma 3

Before establishing the upper bounds, let us control $\mathbb{1}_{\widetilde{\mathcal{A}}_n}\left\|\frac{\mathrm{f}_X-\widetilde{\mathrm{f}}_X}{\widetilde{\mathrm{f}}_X}\right\|_{\infty,\,\mathcal{U}_1}$. First, using Assumption $\mathcal{L}_X$:

$$\delta := \inf_{u\in\mathcal{U}_1}\mathrm{f}_X(u) > 0,$$

remark that: for any $u\in\mathcal{U}_1$,

$$\mathbb{1}_{\widetilde{\mathcal{A}}_n}\widetilde{\mathrm{f}}_X(u) \geq \mathbb{1}_{\widetilde{\mathcal{A}}_n}\left(\mathrm{f}_X(u) - \|\mathrm{f}_X - \widetilde{\mathrm{f}}_X\|_{\infty,\mathcal{U}_1}\right)$$
$$\geq \mathbb{1}_{\widetilde{\mathcal{A}}_n}\left(\delta - M_X\frac{(\log n)^{\frac{a}{2}}}{\sqrt{n}}\right) \quad \text{by (ii)},$$
$$\geq \mathbb{1}_{\widetilde{\mathcal{A}}_n}\frac{\delta}{2} \quad \text{(for } n \text{ large enough)}.$$

Therefore:

$$\tilde{\delta}_X := \inf_{u\in\,\mathcal{U}_1}\widetilde{\mathrm{f}}_X(u) \geq \mathbb{1}_{\widetilde{\mathcal{A}}_n}\frac{\delta}{2},$$

which leads to:

$$\mathbb{1}_{\widetilde{\mathcal{A}}_n}\left\|\frac{\mathrm{f}_X-\widetilde{\mathrm{f}}_X}{\widetilde{\mathrm{f}}_X}\right\|_{\infty,\,\mathcal{U}_1} \leq \mathbb{1}_{\widetilde{\mathcal{A}}_n}\frac{\left\|\mathrm{f}_X-\widetilde{\mathrm{f}}_X\right\|_{\infty,\,\mathcal{U}_1}}{\tilde{\delta}_X}$$
$$\leq \frac{2M_X}{\delta}\frac{(\log n)^{a/2}}{n^{1/2}}. \tag{III.39}$$

Let us now prove the first upper bound.

1. We still denote, for any bandwidth $h$, any component $k$ and any observation $i$,

$$\bar{Z}_{hik} := \frac{\partial}{\partial h_k}\left(\frac{\mathrm{K}_h(w-W_i)}{\mathrm{f}_X(X_i)}\right),$$

such that $\bar{Z}_{hk} = \frac{1}{n}\sum_{i=1}^{n} \bar{Z}_{hik}$, with $\{\bar{Z}_{hik}\}_{i=1}^{n}$ i.i.d.. Then we can write:

$$\Delta_{Z,hk} := Z_{hk} - \bar{Z}_{hk} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\mathrm{f}_X}{\widetilde{\mathrm{f}}_X}(X_i) - 1\right)\bar{Z}_{hik} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\mathrm{f}_X - \widetilde{\mathrm{f}}_X}{\widetilde{\mathrm{f}}_X}(X_i)\right)\bar{Z}_{hik}.$$

Note that since $K$ is compactly supported, if $X_i \notin \mathcal{U}_1$,

$$\bar{Z}_{hik} = 0.$$

Hence:

$$\begin{aligned}\left|\Delta_{Z,hk}\right| &\leq \left\|\frac{\mathrm{f}_X - \widetilde{\mathrm{f}}_X}{\widetilde{\mathrm{f}}_X}\right\|_{\infty,\,\mathcal{U}_1} \times \frac{1}{n}\sum_{i=1}^{n}\left|\bar{Z}_{hik}\right| \\ &\leq \left\|\frac{\mathrm{f}_X - \widetilde{\mathrm{f}}_X}{\widetilde{\mathrm{f}}_X}\right\|_{\infty,\,\mathcal{U}_1} \times \left(\mathbb{E}\left[\left|\bar{Z}_{h1k}\right|\right] + \frac{1}{n}\sum_{i=1}^{n}\left|\bar{Z}_{hik}\right| - \mathbb{E}\left[\left|\bar{Z}_{hik}\right|\right]\right).\end{aligned}$$

Using the above Inequality (III.39) and the upper bounds 1. and 4. of Lemma 2:

$$\begin{aligned}\mathbb{1}_{\widetilde{\mathcal{A}}_n \cap \mathcal{B}\mathsf{ern}_{|\bar{Z}|}(h,k)}\left|\Delta_{Z,hk}\right| &\leq \left(\frac{2M_X}{\delta}\frac{(\log n)^{a/2}}{n^{1/2}}\right) \times 2\mathsf{C}_{E|\bar{Z}|}h_k^{-1} \\ &\leq \frac{1}{4}\lambda_{h,k} := \frac{\mathsf{C}_\lambda}{4}\frac{(\log n)^{a/2}}{n^{1/2}h_k\left(\prod_{k'=1}^{d} h_{k'}\right)^{1/2}},\end{aligned}$$

if $\left(\prod_{k'=1}^{d} h_{k'}\right)^{1/2} \leq \frac{\delta \mathsf{C}_\lambda}{16 M_X \mathsf{C}_{E|\bar{Z}|}}$. Note that $M_X$ is determined in order to satisfy:

$$\frac{\delta \mathsf{C}_\lambda}{16 M_X \mathsf{C}_{E|\bar{Z}|}} = 1.$$

Hence the condition on the bandwidth becomes:

$$\left(\prod_{k'=1}^{d} h_{k'}\right)^{1/2} \leq 1.$$

2. We still denote, for any bandwidth $h$ and any observation $i$,

$$\bar{f}_{hi}(w) := \frac{\mathrm{K}_h(w - W_i)}{\mathrm{f}_X(X_i)},$$

such that $\bar{f}_h(w) = \frac{1}{n}\sum_{i=1}^{n} \bar{f}_{hi}(w)$, with $\{\bar{f}_{hi}(w)\}_{i=1}^{n}$ i.i.d. Then we can write:

$$\Delta_h := \hat{f}_h(w) - \bar{f}_h(w) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\mathrm{f}_X}{\widetilde{\mathrm{f}}_X}(X_i) - 1\right)\bar{f}_{hi}(w) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\mathrm{f}_X - \widetilde{\mathrm{f}}_X}{\widetilde{\mathrm{f}}_X}(X_i)\right)\bar{f}_{hi}(w).$$

Note that since $K$ is compactly supported, if $X_i \notin \mathcal{U}_1$,

$$\bar{f}_{hi}(w) = 0.$$

Hence:

$$|\Delta_h| \leq \left\| \frac{\mathrm{f}_X - \widetilde{\mathrm{f}}_X}{\widetilde{\mathrm{f}}_X} \right\|_{\infty, \mathcal{U}_1} \times \frac{1}{n} \sum_{i=1}^{n} |\bar{f}_{hi}(w)|$$
$$\leq \left\| \frac{\mathrm{f}_X - \widetilde{\mathrm{f}}_X}{\widetilde{\mathrm{f}}_X} \right\|_{\infty, \mathcal{U}_1} \times \left( \mathbb{E}\left[ \left| \bar{f}_{h1}(w) \right| \right] + \frac{1}{n} \sum_{i=1}^{n} \left| \bar{f}_{hi}(w) \right| - \mathbb{E}\left[ \left| \bar{f}_{hi}(w) \right| \right] \right).$$

Using the above Inequality (III.39) and the upper bounds 1. and 4. of Lemma 1:

$$\mathbb{1}_{\widetilde{\mathcal{A}}_n \cap \mathcal{B}\mathrm{ern}_{|\bar{f}|}(h)} |\Delta_h| \leq \left( \frac{2M_X}{\delta} \frac{(\log n)^{a/2}}{n^{1/2}} \right) \times 2\mathsf{C}_{\bar{\mathsf{E}}}$$
$$= \frac{4M_X \mathsf{C}_{\bar{\mathsf{E}}}}{\delta \mathsf{C}_\sigma} \sigma_h \left( \prod_{k'=1}^{d} h_{k'} \right)^{1/2} \leq \mathsf{C}_{\mathsf{M}\Delta} \sigma_h.$$

## 5.e Proof of Proposition II.1

The proof is very similar to the Proposition 1 of [Nguyen 2018]. The main modification is due to the tighter $\log$ exponent in (ii) and the enlarged neighborhood $\mathcal{U}_1$ of $x$.
We introduce the classical kernel density estimator $\tilde{\mathrm{f}}_X^{\mathcal{K}}$: for any $u \in \mathbb{R}^{d_1}$ and a bandwidth $h_X \in \mathbb{R}_+^*$ to be specified later,

$$\tilde{\mathrm{f}}_X^{\mathcal{K}}(u) := \frac{1}{n_X . h_X^{d_1}} \sum_{i=1}^{n_X} \prod_{j=1}^{d_1} \mathcal{K}\left( \frac{u_j - \widetilde{X}_{ij}}{h_X} \right), \tag{III.40}$$

where $\mathcal{K} : \mathbb{R} \to \mathbb{R}$ is a kernel which is compactly supported, of class $\mathcal{C}^1$ and of order $p_X \geq \frac{d_1}{2(c-1)}$, where we recall that $c > 1$ is defined by $n_X = n^c$. We first show that there exists $C_X > 0$ such that for any $\xi > 0$:

$$\mathbb{P}\left( \|\mathrm{f}_X - \tilde{\mathrm{f}}_X^{\mathcal{K}}\|_{\infty, \mathcal{U}_1} > C_X \frac{(\log n)^{\frac{1+\xi}{2}}}{\sqrt{n}} \right) \leq \mathcal{O}\left( n_X^{d_1+1} \exp\left( -(\log n)^{1+\xi} \right) \right). \tag{III.41}$$

Then we set

$$\widetilde{\mathrm{f}}_X \equiv \tilde{\mathrm{f}}_X^{\mathcal{K}} \vee n^{-\frac{1}{2}},$$

and we shall prove that this estimator satisfies (i) and (ii) for $\widetilde{\mathrm{f}}_X$.

Let us prove Inequality (III.41). Let us first explicit $\tilde{\mathrm{f}}_X^{\mathcal{K}}$'s behaviour. Following Lemma 5 gives a pointwise concentration inequality and a control of the bias of $\tilde{\mathrm{f}}_X^{\mathcal{K}}$ on $\mathcal{U}_1$. We introduce an enlarged neighborhood of $\mathcal{U}_1$:

$$\mathcal{U}_1' := \left\{ u' = u - h_X z : u \in \mathcal{U}_1, z \in \mathrm{supp}(\mathcal{K}) \right\}.$$

**Lemma 5** ($\tilde{\mathrm{f}}_X^{\mathcal{K}}$ behaviour). *The estimator $\tilde{\mathrm{f}}_X^{\mathcal{K}}$ satisfies the following results:*

1. *If there exists $q_X \in \mathbb{N}$ such that $\mathrm{f}_X$ is $\mathcal{C}^{q_X}$ on $\mathcal{U}_1'$ and such that $\mathcal{K}$ has $q_X - 1$ zero moments, then there exists a positive constant $\mathsf{C}'_{\mathrm{bias}_X}$ such that*

$$\left\| \mathbb{E}\tilde{\mathrm{f}}_X^{\mathcal{K}} - \mathrm{f}_X \right\|_{\infty, \mathcal{U}_1} \leq \mathsf{C}'_{\mathrm{bias}_X} h_X^{q_X}.$$

2. *For any $\xi > 0$, any $u \in \mathcal{U}_1$ and any $\lambda > 0$ such that:*

$$4\mathsf{C}_{\mathsf{var}_X}\frac{(\log n)^{1+\xi}}{n_X h_X^{d_1}} \leq \lambda^2 \leq \frac{9{\mathsf{C}_{\mathsf{var}_X}}^2}{\|\mathcal{K}\|_\infty^{2d_1}},$$

*where* $\mathsf{C}_{\mathsf{var}_X} := \|\mathcal{K}\|_2^{d_1}\|f_X\|_{\infty,\mathcal{U}_1'}^{\frac{1}{2}}$,

$$\mathbb{P}\left(\left|\tilde{\mathrm{f}}_X^{\mathcal{K}}(u) - \mathbb{E}\tilde{\mathrm{f}}_X^{\mathcal{K}}(u)\right| > \lambda\right) \leq 2\exp\left(-(\log n)^{1+\xi}\right).$$

This lemma is proved in Section 5.f.

We define $p_X' = \min(p', p_X)$, so that: $\mathrm{f}_X$ is of class $\mathcal{C}^{p_X'}$ and the first $p_X' - 1$ moments of $\mathcal{K}$ vanish. Therefore, we can apply 1. of Lemma 5:

$$\left\|\mathbb{E}\tilde{\mathrm{f}}_X^{\mathcal{K}} - \mathrm{f}_X\right\|_{\infty,\mathcal{U}_1} \leq \mathsf{C}'_{\mathsf{bias}_X} h_X^{p_X'}.$$

Therefore:

$$\begin{aligned}\left\|\tilde{\mathrm{f}}_X^{\mathcal{K}} - \mathrm{f}_X\right\|_{\infty,\mathcal{U}_1} &\leq \left\|\tilde{\mathrm{f}}_X^{\mathcal{K}} - \mathbb{E}\tilde{\mathrm{f}}_X^{\mathcal{K}}\right\|_{\infty,\mathcal{U}_1} + \left\|\mathbb{E}\tilde{\mathrm{f}}_X^{\mathcal{K}} - \mathrm{f}_X\right\|_{\infty,\mathcal{U}_1} \\ &\leq \left\|\tilde{\mathrm{f}}_X^{\mathcal{K}} - \mathbb{E}\tilde{\mathrm{f}}_X^{\mathcal{K}}\right\|_{\infty,\mathcal{U}_1} + \mathsf{C}'_{\mathsf{bias}_X} h_X^{p_X'},\end{aligned}$$

and we have for any threshold $\lambda$:

$$\mathbb{P}\left(\left\|\tilde{\mathrm{f}}_X^{\mathcal{K}} - \mathrm{f}_X\right\|_{\infty,\mathcal{U}_1} \geq \lambda\right) \leq \mathbb{P}\left(\left\|\tilde{\mathrm{f}}_X^{\mathcal{K}} - \mathbb{E}\tilde{\mathrm{f}}_X^{\mathcal{K}}\right\|_{\infty,\mathcal{U}_1} \geq \lambda - \mathsf{C}'_{\mathsf{bias}_X} h_X^{p_X'}\right). \tag{III.42}$$

We have then reduced the problem to a concentration inequality of $\tilde{\mathrm{f}}_X^{\mathcal{K}}$ in sup norm. In order to move from a supremum on $\mathcal{U}_1$ to a maximum on a finite set of elements of $\mathcal{U}_1$, let us construct an $\varepsilon$-net $\{u_{(l)}\}_l$ of $\mathcal{U}_1$, in the meaning that for any $u \in \mathcal{U}_1$, there exists $l$ such that $\|u - u_{(l)}\|_\infty := \max_{k=1:d_1} |u_k - u_{(l)k}| \leq \varepsilon$. We denote $A > 0$ such that:

$$\mathsf{supp}(\mathcal{K}) \cup \mathsf{supp}(K) \subset \left[-\tfrac{A}{2}, \tfrac{A}{2}\right].$$

Set $N(\varepsilon)$ is the smallest integer such that $2\varepsilon N(\varepsilon) \geq A$, and for $l \in \{1 : N(\varepsilon)\}^{d_1}$, $u_{(l)}$ such that its $j$-th component is equal to:

$$u_{(l)j} := x_j - \frac{A}{2} + (2l_j - 1)\varepsilon.$$

Then $\{u_{(l)}\}_{l\in\{1:N(\varepsilon)\}^{d_1}}$ is an $\varepsilon$-net of $\mathcal{U}_1$.

Therefore in order to obtain Inequality (III.41), we only need to obtain the concentration inequality for each point of $\{u_{(l)} : l \in (1 : N(\varepsilon))^{d_1}\}$ and to control the difference of the function $\tilde{\mathrm{f}}_X^{\mathcal{K}} - \mathbb{E}\tilde{\mathrm{f}}_X^{\mathcal{K}}$ evaluated at the point $u$ and at the nearest point of $u$ in the $\varepsilon$-net. More formally, we have to control the following supremum

$$\sup_{u\in\mathcal{U}_1} \min_{l\in(1:N(\varepsilon))^{d_1}} \left|\tilde{\mathrm{f}}_X^{\mathcal{K}}(u) - \mathbb{E}\tilde{\mathrm{f}}_X^{\mathcal{K}}(u) - \tilde{\mathrm{f}}_X^{\mathcal{K}}(u_{(l)}) + \mathbb{E}\tilde{\mathrm{f}}_X^{\mathcal{K}}(u_{(l)})\right|.$$

For this purpose, we obtain (from Taylor's Inequality): for any $u, v \in \mathbb{R}^{d_1}$,

$$\left| \prod_{k=1}^{d_1} \mathcal{K}(u_k) - \prod_{k=1}^{d_1} \mathcal{K}(v_k) \right| \leq d_1 \|\mathcal{K}'\|_\infty \|\mathcal{K}\|_\infty^{d_1-1} \|u - v\|_\infty.$$

Therefore, for any $u, v \in \mathcal{U}_1$:

$$\begin{aligned} \left| \tilde{\mathrm{f}}_X^{\mathcal{K}}(u) - \tilde{\mathrm{f}}_X^{\mathcal{K}}(v) \right| &\leq \frac{1}{n_X . h_X^{d_1}} \sum_{i=1}^{n_X} \left| \prod_{k=1}^{d_1} \mathcal{K}\left(\tfrac{u_k - \widetilde{X}_{ik}}{h_X}\right) - \prod_{k=1}^{d_1} \mathcal{K}\left(\tfrac{v_k - \widetilde{X}_{ik}}{h_X}\right) \right| \\ &\leq d_1 \|\mathcal{K}'\|_\infty \|\mathcal{K}\|_\infty^{d_1-1} \frac{\|u - v\|_\infty}{h_X^{d_1+1}}. \end{aligned}$$

Since $\{u_{(l)} : l \in (1 : N(\varepsilon))^{d_1}\}$ is an $\varepsilon$-net of $\mathcal{U}_1$:

$$\sup_{u \in \mathcal{U}_1} \min_{l \in (1:N(\varepsilon))^{d_1}} \left| \tilde{\mathrm{f}}_X^{\mathcal{K}}(u) - \tilde{\mathrm{f}}_X^{\mathcal{K}}(u_{(l)}) \right| \leq d_1 \|\mathcal{K}'\|_\infty \|\mathcal{K}\|_\infty^{d_1-1} \frac{\varepsilon}{h_X^{d_1+1}},$$

and also:

$$\sup_{u \in \mathcal{U}_1} \min_{l \in (1:N(\varepsilon))^{d_1}} \left| \mathbb{E}\tilde{\mathrm{f}}_X^{\mathcal{K}}(u) - \mathbb{E}\tilde{\mathrm{f}}_X^{\mathcal{K}}(u_{(l)}) \right| \leq d_1 \|\mathcal{K}'\|_\infty \|\mathcal{K}\|_\infty^{d_1-1} \frac{\varepsilon}{h_X^{d_1+1}}.$$

Therefore:

$$\sup_{u \in \mathcal{U}_1} \min_{l \in (1:N(\varepsilon))^{d_1}} \left| \tilde{\mathrm{f}}_X^{\mathcal{K}}(u) - \mathbb{E}\tilde{\mathrm{f}}_X^{\mathcal{K}}(u) - \tilde{\mathrm{f}}_X^{\mathcal{K}}(u_{(l)}) + \mathbb{E}\tilde{\mathrm{f}}_X^{\mathcal{K}}(u_{(l)}) \right| \leq 2d_1 \|\mathcal{K}'\|_\infty \|\mathcal{K}\|_\infty^{d_1-1} \frac{\varepsilon}{h_X^{d_1+1}}.$$

We denote $\mathsf{C}_{\mathsf{diff}} := 2d_1 \|\mathcal{K}'\|_\infty \|\mathcal{K}\|_\infty^{d_1-1}$. We then obtain the following inequality:

$$\begin{aligned} \left\| \tilde{\mathrm{f}}_X^{\mathcal{K}} - \mathbb{E}\tilde{\mathrm{f}}_X^{\mathcal{K}} \right\|_{\infty,\, \mathcal{U}_1} &\leq \max_{l \in (1:N(\varepsilon))^{d_1}} \left| \tilde{\mathrm{f}}_X^{\mathcal{K}}(u_{(l)}) - \mathbb{E}\tilde{\mathrm{f}}_X^{\mathcal{K}}(u_{(l)}) \right| \\ &\quad + \sup_{u \in \mathcal{U}_1} \min_{l \in (1:N(\varepsilon))^{d_1}} \left| \tilde{\mathrm{f}}_X^{\mathcal{K}}(u) - \mathbb{E}\tilde{\mathrm{f}}_X^{\mathcal{K}}(u) - \tilde{\mathrm{f}}_X^{\mathcal{K}}(u_{(l)}) + \mathbb{E}\tilde{\mathrm{f}}_X^{\mathcal{K}}(u_{(l)}) \right| \\ &\leq \max_{l \in (1:N(\varepsilon))^{d_1}} \left| \tilde{\mathrm{f}}_X^{\mathcal{K}}(u_{(l)}) - \mathbb{E}\tilde{\mathrm{f}}_X^{\mathcal{K}}(u_{(l)}) \right| + \mathsf{C}_{\mathsf{diff}} \frac{\varepsilon}{h_X^{d_1+1}}. \end{aligned}$$

Then the inequality (III.42) becomes: for any threshold $\lambda$,

$$\begin{aligned} \mathbb{P}\left( \left\| \tilde{\mathrm{f}}_X^{\mathcal{K}} - \mathrm{f}_X \right\|_{\infty,\, \mathcal{U}_1} \geq \lambda \right) &\leq \mathbb{P}\left( \left\| \tilde{\mathrm{f}}_X^{\mathcal{K}} - \mathbb{E}\tilde{\mathrm{f}}_X^{\mathcal{K}} \right\|_{\infty,\, \mathcal{U}_1} \geq \lambda - \mathsf{C}'_{\mathsf{bias}_X} h_X^{p'_X} \right) \\ &\leq \mathbb{P}\left( \max_{l \in (1:N(\varepsilon))^{d_1}} \left| \tilde{\mathrm{f}}_X^{\mathcal{K}}(u_{(l)}) - \mathbb{E}\tilde{\mathrm{f}}_X^{\mathcal{K}}(u_{(l)}) \right| \geq \lambda - \mathsf{C}'_{\mathsf{bias}_X} h_X^{p'_X} - \mathsf{C}_{\mathsf{diff}} \frac{\varepsilon}{h_X^{d_1+1}} \right) \\ &\leq N(\varepsilon)^{d_1} \max_{l \in (1:N(\varepsilon))^{d_1}} \mathbb{P}\left( \left| \tilde{\mathrm{f}}_X^{\mathcal{K}}(u_{(l)}) - \mathbb{E}\tilde{\mathrm{f}}_X^{\mathcal{K}}(u_{(l)}) \right| \geq \lambda - \mathsf{C}'_{\mathsf{bias}_X} h_X^{p'_X} - \mathsf{C}_{\mathsf{diff}} \frac{\varepsilon}{h_X^{d_1+1}} \right). \end{aligned} \tag{III.43}$$

It then remains to apply 2. of Lemma 5 for each $u_{(l)}$, $l \in (1 : N(\varepsilon))^{d_1}$. We set the following settings:

- $h_X := n_X^{-\frac{c-1}{c.d_1}}$;

- $\varepsilon := h_X^{1+\frac{d_1}{2}} n_X^{-\frac{1}{2}}$;
- $\lambda := 2\lambda_X$, where $\lambda_X$ is defined by:

$$\lambda_X := 2\sqrt{\mathsf{C}_{\mathsf{var}_X}}(\log n)^{\frac{1+\xi}{2}} h_X^{-\frac{d_1}{2}} n_X^{-\frac{1}{2}} = 2\sqrt{\mathsf{C}_{\mathsf{var}_X}}(\log n)^{\frac{1+\xi}{2}} n_X^{-\frac{1}{2c}},$$

where we recall that $\mathsf{C}_{\mathsf{var}_X} := \|\mathcal{K}\|_2^{d_1} \|f_X\|_{\infty,\mathcal{U}_1'}^{\frac{1}{2}}$.

In particular, since we take $p_X \geq \frac{d_1}{2(c-1)}$ and we assume $p' \geq \frac{d_1}{2(c-1)}$, then $p_X' = \min(p', p_X) \geq \frac{d_1}{2(c-1)}$. Hence we obtain for $n$ large enough:

$$\begin{aligned}\mathsf{C}'_{\mathsf{bias}_X} h_X^{p_X'} &= \mathsf{C}'_{\mathsf{bias}_X} n_X^{-\frac{p_X'(c-1)}{c.d_1}} \\ &\leq \mathsf{C}'_{\mathsf{bias}_X} n_X^{-\frac{1}{2c}} \\ &\leq \frac{1}{2}\lambda_X = \sqrt{\mathsf{C}_{\mathsf{var}_X}}(\log n)^{\frac{1+\xi}{2}} n_X^{-\frac{1}{2c}}.\end{aligned}$$

and also, since $c > 1$:

$$\begin{aligned}\mathsf{C}_{\mathsf{diff}} \frac{\varepsilon}{h_X^{d_1+1}} &= \mathsf{C}_{\mathsf{diff}} h_X^{-\frac{d_1}{2}} n_X^{-\frac{1}{2}} = \mathsf{C}_{\mathsf{diff}}\, n_X^{-\frac{1}{2c}} \\ &\leq \frac{1}{2}\lambda_X = \sqrt{\mathsf{C}_{\mathsf{var}_X}}(\log n)^{\frac{1+\xi}{2}} n_X^{-\frac{1}{2c}}.\end{aligned}$$

Hence, we have

$$\lambda - \mathsf{C}'_{\mathsf{bias}_X} h_X^{p_X'} - \mathsf{C}_{\mathsf{diff}} \frac{\varepsilon}{h_X^{d_1+1}} \geq \lambda_X,$$

and the inequality (III.43) becomes:

$$\mathbb{P}\left(\left\|\tilde{\mathrm{f}}_X^{\mathcal{K}} - \mathrm{f}_X\right\|_{\infty,\mathcal{U}_1} \geq \lambda\right) \leq N(\varepsilon)^{d_1} \max_{l\in(1:N(\varepsilon))^{d_1}} \mathbb{P}\left(\left|\tilde{\mathrm{f}}_X^{\mathcal{K}}(u_{(l)}) - \mathbb{E}\tilde{\mathrm{f}}_X^{\mathcal{K}}(u_{(l)})\right| \geq \lambda_X\right). \tag{III.44}$$

We apply 2. of Lemma 5: we verify (since $n_X = n^c$)

$$\begin{aligned}4\mathsf{C}_{\mathsf{var}_X}\frac{(\log n)^{1+\xi}}{n_X h_X^{d_1}} = \lambda_X^2 &= 4\mathsf{C}_{\mathsf{var}_X}(\log n)^{1+\xi} n^{-1} \\ &\leq \frac{9{\mathsf{C}_{\mathsf{var}_X}}^2}{\|\mathcal{K}\|_\infty^{2d_1}}, \quad \text{(for } n \text{ large enough)},\end{aligned}$$

then we obtain

$$\mathbb{P}\left(\left|\tilde{\mathrm{f}}_X^{\mathcal{K}}(u_{(l)}) - \mathbb{E}\tilde{\mathrm{f}}_X^{\mathcal{K}}(u_{(l)})\right| > \lambda_X\right) \leq 2\exp\left(-(\log n)^{1+\xi}\right).$$

Thus the inequality (III.44) becomes:

$$\mathbb{P}\left(\left\|\tilde{\mathrm{f}}_X^{\mathcal{K}} - \mathrm{f}_X\right\|_{\infty,\mathcal{U}_1} \geq \lambda\right) \leq 2N(\varepsilon)^{d_1}\exp\left(-(\log n)^{1+\xi}\right). \tag{III.45}$$

Let us control $2N(\varepsilon)^{d_1}$:

$$2N(\varepsilon)^{d_1} = 2\left\lceil \frac{A}{2\varepsilon} \right\rceil^{d_1} = 2\left\lceil \frac{A}{2h_X^{1+\frac{d_1}{2}} n_X^{-\frac{1}{2}}} \right\rceil^{d_1} = o\left(n_X^{d_1+1}\right).$$

Therefore, we have obtained the desired concentration inequality (III.41). Now we consider $\widetilde{\mathrm{f}}_X \equiv \tilde{\mathrm{f}}_X^{\mathcal{K}} \vee n^{-1/2}$, therefore $\widetilde{\mathrm{f}}_X$ satisfies (i). Let us show it also satisfies (ii), for $n$ large enough. We first show:

$$\left\{ \left\| \tilde{\mathrm{f}}_X^{\mathcal{K}} - \mathrm{f}_X \right\|_{\infty,\,\mathcal{U}_1} < \lambda \right\} \quad \Rightarrow \quad \left\{ \left\| \widetilde{\mathrm{f}}_X - \mathrm{f}_X \right\|_{\infty,\,\mathcal{U}_1} < \lambda \right\}. \tag{III.46}$$

Assume that for any $u \in \mathcal{U}_1$, $\left|\tilde{\mathrm{f}}_X^{\mathcal{K}}(u) - \mathrm{f}_X(u)\right| < \lambda$. Let us fix $u \in \mathcal{U}_1$. Three cases occurs:

(a) When $\tilde{\mathrm{f}}_X^{\mathcal{K}}(u) \geq n^{-\frac{1}{2}}$, then $\widetilde{\mathrm{f}}_X(u) := \tilde{\mathrm{f}}_X^{\mathcal{K}}(u)$, and obviously:

$$\left|\widetilde{\mathrm{f}}_X(u) - \mathrm{f}_X(u)\right| < \lambda.$$

(b) When $\tilde{\mathrm{f}}_X^{\mathcal{K}}(u) < n^{-\frac{1}{2}}$ and $\mathrm{f}_X(u) \geq n^{-\frac{1}{2}}$, then since $\widetilde{\mathrm{f}}_X(u) = n^{-\frac{1}{2}} > \tilde{\mathrm{f}}_X^{\mathcal{K}}(u)$,

$$\left|\widetilde{\mathrm{f}}_X(u) - \mathrm{f}_X(u)\right| \leq \left|\tilde{\mathrm{f}}_X^{\mathcal{K}}(u) - \mathrm{f}_X(u)\right| < \lambda.$$

(c) When $\tilde{\mathrm{f}}_X^{\mathcal{K}}(u) < n^{-\frac{1}{2}}$ and $\mathrm{f}_X(u) < n^{-\frac{1}{2}}$, then $\widetilde{\mathrm{f}}_X(u) = n^{-\frac{1}{2}}$, so for $n$ large enough:

$$\left|\widetilde{\mathrm{f}}_X(u) - \mathrm{f}_X(u)\right| \leq n^{-\frac{1}{2}} < \lambda.$$

Therefore these three cases show Implication (III.46), and thus, from Equation (III.45), we obtain:

$$\mathbb{P}\left( \left\| \widetilde{\mathrm{f}}_X - \mathrm{f}_X \right\|_{\infty,\,\mathcal{U}_1} \geq \lambda \right) \leq \mathbb{P}\left( \left\| \tilde{\mathrm{f}}_X^{\mathcal{K}} - \mathrm{f}_X \right\|_{\infty,\,\mathcal{U}_1} \geq \lambda \right) \leq 2N(\varepsilon)^{d_1} \exp\left(-(\log n)^{1+\xi}\right).$$

Now, to obtain (ii), for $\xi$ such that $1 + \frac{a-1}{2} < 1 + \xi < a$,

$$\lambda = 4\sqrt{\mathsf{C}_{\mathrm{var}_X}} (\log n)^{\frac{1+\xi}{2}} n^{-\frac{1}{2}} \leq M_X (\log n)^{\frac{a}{2}} n^{-\frac{1}{2}} \text{ (for } n \text{ large enough).} \tag{III.47}$$

Therefore:

$$\begin{aligned}
\mathbb{P}\left( \left\| \widetilde{\mathrm{f}}_X - \mathrm{f}_X \right\|_{\infty,\,\mathcal{U}_1} \geq M_X (\log n)^{\frac{a}{2}} n^{-\frac{1}{2}} \right) &\leq \mathbb{P}\left( \left\| \widetilde{\mathrm{f}}_X - \mathrm{f}_X \right\|_{\infty,\,\mathcal{U}_1} \geq \lambda \right) \\
&\leq 2N(\varepsilon)^{d_1} \exp\left(-(\log n)^{1+\xi}\right) \\
&\leq \exp\left(-(\log n)^{1+\frac{a-1}{2}}\right),
\end{aligned}$$

that is (ii).

## 5.f Proof of Lemma 5

The result 1. of Lemma 5 is proved in Lemma 4 of Nguyen [2018]. To prove 2. of Lemma 5, let us fix $\xi > 0$. Then, we simply apply Bernstein's Inequality (see Lemma 10 in Nguyen [2018]). We define for any $u \in \mathcal{U}_1$ and for $i = 1 : n$

$$\tilde{\mathrm{f}}^K_{X,i}(u) := \frac{1}{h_X^{d_1}} \prod_{j=1}^{d_1} K\left(\frac{u_j - \widetilde{X}_{ij}}{h_X}\right).$$

Observe that the $\tilde{\mathrm{f}}^K_{X,i}(u)$'s are *i.i.d.* Then we pick up the following bounds from [Nguyen 2018, p. 23]:

$$\begin{aligned} \left|\tilde{\mathrm{f}}^K_{X,1}(u)\right| &\leq \mathsf{M}_{h_X} := \|K\|_\infty^{d_1} h_X^{-d_1}. \\ \mathrm{Var}\left(\tilde{\mathrm{f}}^K_{X,1}(u)\right) &\leq \mathsf{v}_{h_X} := \mathsf{C}_{\mathsf{var}_X} h_X^{-d_1}, \end{aligned}$$

(we recall $\mathsf{C}_{\mathsf{var}_X} := \|K\|_2^{2d_1} \|\mathrm{f}_X\|_{\infty,\mathcal{U}_1'}$). Therefore: for any $\lambda > 0$,

$$\mathbb{P}\left(\left|\tilde{\mathrm{f}}^K_X(u) - \mathbb{E}\tilde{\mathrm{f}}^K_X(u)\right| > \lambda\right) \leq 2\exp\left(-\min\left(\frac{n_X\lambda^2}{4\mathsf{v}_{h_X}}, \frac{3n_X\lambda}{4\mathsf{M}_{h_X}}\right)\right).$$

Let us show that when

$$4\mathsf{C}_{\mathsf{var}_X}\frac{(\log n)^{1+\xi}}{n_X h_X^{d_1}} \leq \lambda^2 \leq \frac{9{\mathsf{C}_{\mathsf{var}_X}}^2}{\|K\|_\infty^{2d_1}},$$

then, we have

$$(\log n)^{1+\xi} \leq \frac{n_X\lambda^2}{4\mathsf{v}_{h_X}} \leq \frac{3n_X\lambda}{4\mathsf{M}_{h_X}}.$$

Indeed,

$$\begin{aligned} \frac{n_X\lambda^2}{4\mathsf{v}_{h_X}} \leq \frac{3n_X\lambda}{4\mathsf{M}_{h_X}} \quad &\Leftrightarrow \quad \lambda \leq \frac{3\mathsf{v}_{h_X}}{\mathsf{M}_{h_X}} = \frac{3\mathsf{C}_{\mathsf{var}_X}}{\|K\|_\infty^{d_1}} \\ &\Leftrightarrow \quad \lambda^2 \leq \frac{9\mathsf{C}^2_{\mathsf{var}_X}}{\|K\|_\infty^{2d_1}} \end{aligned}$$

and

$$(\log n)^{1+\xi} \leq \frac{n_X\lambda^2}{4\mathsf{v}_{h_X}} \quad \Leftrightarrow \quad \frac{4\mathsf{C}_{\mathsf{var}_X}(\log n)^{1+\xi}}{n_X h_X^{d_1}} \leq \lambda^2.$$

Therefore when

$$4\mathsf{C}_{\mathsf{var}_X}\frac{(\log n)^{1+\xi}}{n_X h_X^{d_1}} \leq \lambda^2 \leq \frac{9{\mathsf{C}_{\mathsf{var}_X}}^2}{\|K\|_\infty^{2d_1}},$$

$$\begin{aligned} \mathbb{P}\left(\left|\tilde{\mathrm{f}}^K_X(u) - \mathbb{E}\tilde{\mathrm{f}}^K_X(u)\right| > \lambda\right) &\leq 2\exp\left(-\min\left(\frac{n_X\lambda^2}{4\mathsf{v}_{h_X}}, \frac{3n_X\lambda}{4\mathsf{M}_{h_X}}\right)\right) = 2\exp\left(-\frac{n_X\lambda^2}{4\mathsf{v}_{h_X}}\right) \\ &\leq 2\exp\left(-(\log n)^{1+\xi}\right). \end{aligned}$$

# Chapitre IV

# Étude numérique

Je m'intéresse dans ce chapitre aux performances numériques des procédures CDRodeo. Après quelques remarques sur le coût algorithmique, l'étude se fait en deux temps. On commence par la calibration des paramètres principaux $a$ et $\beta$, puis on s'intéresse à proprement parler à la qualité d'estimation des deux procédures CDRodeo en étudiant l'erreur absolue des estimées, la détection de variables pertinentes, le comportement en cas de discontinuité, les temps d'exécution.

## Table des matières

Il est connu des statisticiens qu'il y a souvent de grandes différences entre la pratique et la théorie en terme de performance optimale. En particulier, pour contrôler le risque, certaines précautions (par exemple de régularité) peuvent être non nécessaires dans le cas particulier d'un jeu de données, ou encore, l'enchaînement de majorations peuvent rendre le contrôle du risque lâche. Par ailleurs, les résultats théoriques reposent souvent sur la mise à disposition d'un échantillon assez grand pour pouvoir utiliser des résultats de convergence ou des inégalités de concentration. En pratique, rien ne nous garantit qu'il n'y a pas des constantes plus grandes que la taille $n$ de l'échantillon ou, plus facile encore, que son logarithme $\log n$ qui, pour donner un ordre de grandeur, n'atteint pas $15$ pour $n = 3.10^6$.

# 1 Complexité et subtilités algorithmiques

Nos procédures sont particulièrement rapides grâce à leur complexité linéarithmique $\mathcal{O}(\log(n)n)$. Précisons dans notre cas les jalons essentiels pour obtenir cette vitesse.
Premièrement, les composantes des fenêtres explorables évoluent de manière exponentielle, au sens où ne sont explorées que des fenêtres de la forme $(\beta^{t_1}h_0, \dots, \beta^{t_d}h_0)$ avec $t = (t_1, \dots, t_d) \in \mathbb{Z}^d$, assurant une complexité quasi-linéaire $\mathcal{O}(\log(n)^d n)$. Pour atteindre la complexité linéarithmique, c'est l'exploration gloutonne de chemins uniquement monotones qui permet à nos procédures de l'obtenir : cela diminue la taille $\mathcal{O}(\log(n)^d)$ de la grille de fenêtres à seulement $\mathcal{O}(d\log(n))$ fenêtres le long d'un chemin, et c'est ce qui assure la caractère glouton de nos procédures pour contrer le fléau de la dimension.
De manière plus détaillée, pour faire varier une composante $h_j$ de la fenêtre de $1$ à $\frac{1}{n}$ (les valeurs qu'il suffit d'explorer dans la recherche du compromis biais-variance) par multiplication par un facteur $\beta \in ]0,1[$, $\log_{\frac{1}{\beta}} n$ valeurs suffisent par direction. De plus, le nombre de fenêtres par chemin décroissant de $(1, \dots, 1)$ à $(\frac{1}{n}, \dots, \frac{1}{n})$ est $d\log_{\frac{1}{\beta}} n$.

# 2 Les exemples considérés

Avant de décrire les modèles, donnons (ou rappelons) quelques notations : on considère un échantillon $W$ de $n$ observations i.i.d. de dimension $d$ sous forme matricielle $n \times d$. Chaque observation $W_i$ (qui correspond à la ligne $i$ de $W$) se réécrit comme le couple $(X_i, Y_i)$ de vecteurs aléatoires de dimensions respectives $d_1$ et $d_2$ (de sorte que $d = d_1 + d_2$) et on estime $f$ la densité conditionnelle de $Y_i$ sachant $X_i$.
Pour les lois usuelles, on utilise les notations suivantes :

- $\mathcal{N}(\mu, \sigma^2)$ pour la loi gaussienne d'espérance $\mu$ et de variance $\sigma^2$,
- $\mathcal{U}_{[a,b]}$ pour la loi uniforme sur l'intervalle $[a,b]$,
- $\mathcal{IG}(\alpha, \beta)$ la loi inverse-gamma de paramètre de forme $\alpha$ et de paramètre d'échelle $\beta$.

Trois modèles seront considérés et sont décrits comme suit :

**Modèle 1** : avec $d_2 = 1$ et $d_1$ variant dans $\{1, \dots, 12\}$,

$$X_{ij} \overset{\text{iid}}{\sim} \mathcal{U}_{[-1,1]}, \quad Y_i | X_i \sim \mathcal{N}(3X_{i1}^3, 0.5^2).$$

La densité conditionnelle à estimer est donc

$$f : (x,y) \mapsto \frac{1}{\sqrt{\pi}} e^{-2(y-3x^3)^2} \mathbb{1}_{x \in [-1,1]^{d_1}}.$$

Ainsi, $r = 2$ sauf aux voisinages des bords $x_j = \pm 1$, $j \geq 2$, et en dehors du support.

**Modèle 2** : avec $d_2 = 1$ et $d_1$ variant dans $\{1, \dots, 12\}$,

$$X_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0,1), \quad Y_i | X_i \sim \mathcal{N}(3X_{i1}^3, 0.5^2).$$

La densité conditionnelle à estimer est donc

$$f : (x,y) \mapsto \frac{1}{\sqrt{\pi}} e^{-2(y-3x^3)^2}.$$

Ainsi, $r = 2$.

**Modèle 3** : avec $d_2 = 2$ et $d_1$ variant dans $\{1, \dots, 4\}$,

$$Y_{i2} \sim \mathcal{IG}(4,3), \quad Y_{i1} | Y_{i2} \sim \mathcal{N}(0, Y_{i2}), \quad X_{ij} | Y_i \overset{\text{iid}}{\sim} \mathcal{N}(Y_{i1}, Y_{i2}).$$

La densité conditionnelle à estimer est

$$f : (x,y) \mapsto \mathbb{1}_{y_2>0} \frac{\sqrt{d_1+1}}{\sqrt{2\pi}\,\Gamma(4+\frac{d_1}{2})} (\beta_1(x))^{4+\frac{d_1}{2}} y_2^{-(5+\frac{d_1+1}{2})} e^{-\frac{\beta_1(x)}{y_2}} \exp\left( - \frac{\left(y_1 - \frac{\Sigma_{j=1}^{d_1} x_j}{d_1+1}\right)^2}{2y_2/(d_1+1)} \right)$$

avec $\beta_1(x) := \frac{1}{2}(6 + \sum_{j=1}^{d_1} x_j^2 - \frac{(\Sigma_{j=1}^{d_1} x_j)^2}{d_1+1})$. Ainsi, $r = d$ (sur le support $\{y_2 \geq 0\}$). Pour l'obtention de cette dernière densité, voir les calculs en Appendice 5.a de ce chapitre.

Les deux premiers modèles sont des modèles-tests dans le sens où ils présentent une grande similarité (on change uniquement la loi des variables auxiliaires) tout en restant simples (les composantes de $X$ sont i.i.d., $Y$ est de dimension $1$ et ne dépend que de la première composante de $X$) : ils vont donc me permettre de tester les performances de CDRODEO en cas de parcimonie ou de discontinuité.
Le troisième modèle est plus complexe. Noter en particulier que les composantes de $X$ ne sont pas indépendantes, qu'il n'y a pas de composante non pertinente et que $Y$ est de dimension 2. Ce modèle reçoit le fléau de de la dimension de plein fouet, et la difficulté croissante pour estimer quand la dimension augmente nous conforte dans le besoin de détection de parcimonie.

# 3 Calibration

Calibrer une méthode statistique consiste à déterminer le bon jeu de paramètres de l'algorithme de manière à optimiser ses performances. La calibration est souvent déterminante pour obtenir de bons résultats et peut représenter un travail préliminaire conséquent dans l'objectif de fournir au praticien une procédure clés en main, avec idéalement aucun paramètre à régler pour que l'application sur jeux de données réelles en soient facilitée. Mais il arrive que cette calibration soit particulièrement ardue, en dépendant par exemple d'inconnues du modèle, et nécessite alors de séparer une partie de l'échantillon pour en déduire les paramètres propres au problème.
Dans la suite, on calibre les deux paramètres principaux de nos deux procédures CDRODEO Direct et RevDir : la constante de seuil $a$ et le pas de décroissance $\beta$.

D'autres paramètres ne vont pas changer, que ce soit pour la calibration ou l'étude des performances, et seront dans la suite systématiquement fixés comme il suit :

- le noyau est la densité gaussienne (d'ordre $2$) ;
- pour l'initialisation de Direct, $h_0 = 1$ (et on rappelle que pour RevDir, $h_0 = \left(\frac{3(\log n)^a}{2^d \pi^{d/2} n}\right)^{\frac{1}{5d}}$) ;
- on suppose $\mathrm{f}_X$ connue et on prend $\widetilde{\mathrm{f}}_X = \mathrm{f}_X$ pour éviter l'ajout de difficulté qu'occasionnerait un estimateur aléatoire ;

## 3.a Calibration du seuil : sensibilité du paramètre $a$

La constante du seuil est en pratique un paramètre fondamental car sa calibration impacte directement le compromis biais-variance. Pour rappel, dans nos procédures CDRodeo, la valeur des $|Z_{hj}|$ est testée contre le seuil

$$\lambda_{hj} := \mathsf{C}_\lambda \sqrt{\frac{(\log n)^a}{n h_j^2 \prod_{k=1}^d h_k}}.$$

Le paramètre $a$ est donc l'exposant du terme $\log n$ intervenant dans le seuil. Noter que la calibration de $a$ englobe celle de la constante $\mathsf{C}_\lambda$.

**Procédure de la calibration.** Pour distinguer d'éventuelles relations entre $a$ et les méta-données fixant le modèle (c'est-à-dire la taille de l'échantillon $n$, la régularité locale $(L,s)$, la dimension des données $d$ et, plus pernicieusement, la dimension pertinente $r$), on calcule, pour chaque modèle et nos deux procédures Direct et RevDir, l'erreur absolue de l'estimateur et la vraie densité conditionnelle en fonction de $a$ pour plusieurs jeux de méta-données :

- selon différentes tailles d'échantillon $n$ variant dans $\{10\,000, 25\,000, 50\,000, 100\,000, 200\,000\}$ (dans les figures, chaque graphe étant à taille d'échantillon fixée),
- selon différentes dimensions $d_1$ : de $1$ à $8$ dans les modèles $1$ et $2$, et de $1$ à $4$ dans le modèle $3$ (chaque sous-graphe étant à dimension fixée et titré sous la forme "$d = d_1 + d_2$" ),
- selon différents points d'évaluation $\{w^k\}_{k=1}^{16}$, tirés aléatoirement selon la loi jointe $f_W$ (en couleurs pastels du bleu au rouge, ordonnée par valeur croissante des $f(w^k)$),
- selon $B = 3$ échantillons différents (différentiés par le type de ligne : ligne pleine, en tirets ou en pointillés).

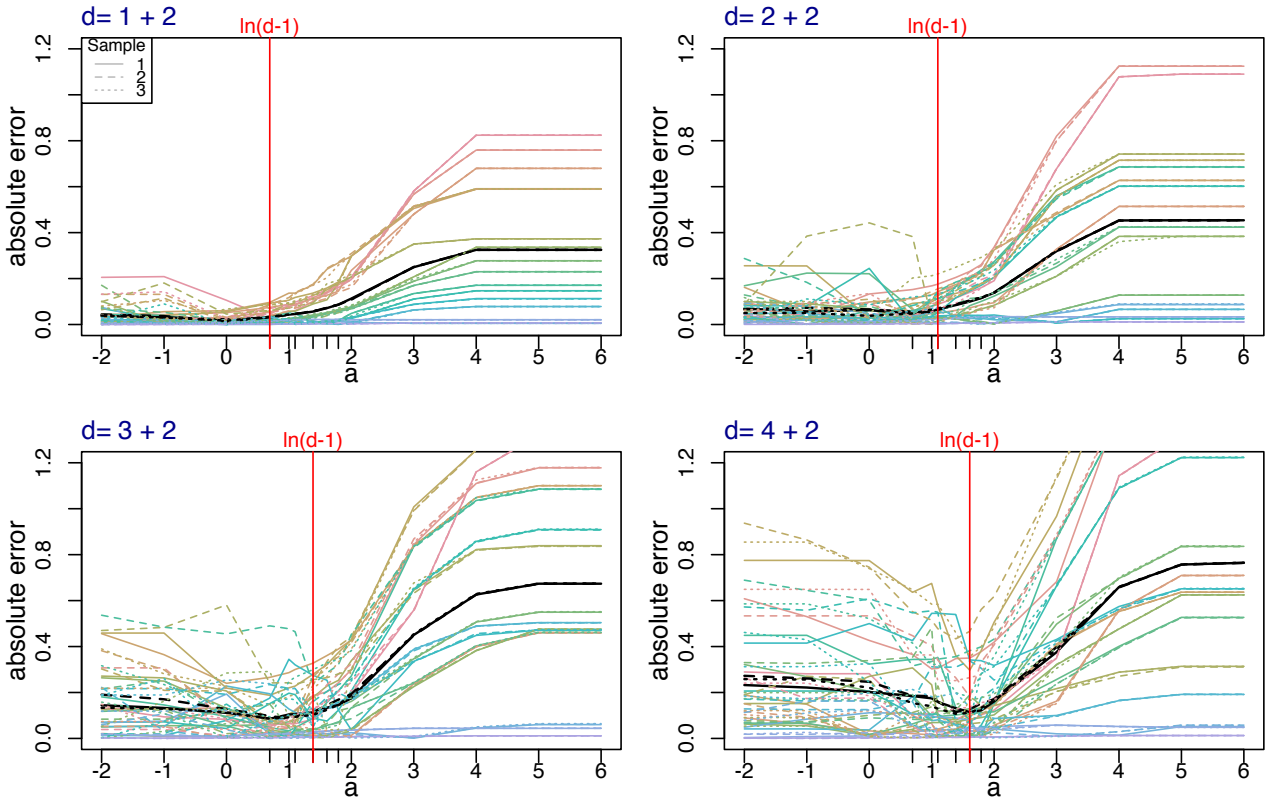
Précisons la grille de valeurs pour $a$ : $\{-2, \dots, d_{\max}\} \bigcup \{\ln(2), \dots, \ln(d_{\max})\}$, où $d_{\max}$ est la dimension maximale considérée en fonction des modèles (c'est-à dire $d_{\max} = 9$ pour les modèles $1$ et $2$ et $d_{\max} = 6$ pour le modèle $3$). L'ajout de valeurs logarithmes est plutôt naturel au vu du déplacement sous-linéaire du minimum d'erreur quand on fait croître $d_1$. Cela permet aussi d'affiner la grille autour du lieu de ce minimum d'erreur.

Le pas itératif est fixé à $\beta = 0.9$.

L'ensemble des figures est affiché en Appendices de ce chapitre, Section 5.b. Pour illustrer la calibration et faciliter la compréhension des remarques qui suivent, nous affichons ici la figure de la calibration de RevDir sur le modèle $3$ pour un échantillon de taille $n = 200\,000$, Figure IV.1 .

**Choix de $a$ et quelques remarques.** Remarquons que comme les points d'évaluation sont tirés aléatoirement selon la loi jointe $f_W$, chaque point a ainsi ses propres degrés de régularité et de pertinence locales. Noter aussi qu'un tel tirage aléatoire est particulièrement utile pour limiter le nombre de points tests à considérer en grande dimension car cela proscrit de faire des tests sur

FIGURE IV.1 – Calibration de $a$ pour le modèle $3$ pour la procédure RevDir sur des échantillons de taille $n = 200\,000$.




une grille $d$-dimensionnelle complète et évite ainsi des temps de calculs excessifs. Cela crée malgré tout un biais en faveur de la qualité d'estimation, puisque les points tirés sont plus probables que ceux d'une grille et auront donc potentiellement plus d'observations de l'échantillon à proximité, rendant l'estimation ponctuelle plus aisée.
Par ailleurs, la répétition de l'expérience pour trois échantillons différents permet de se rendre compte de la stabilité des procédures en fonction de $a$.

Cherchons une calibration unique pour les trois modèles. Noter que la difficulté majeure de cette calibration réside dans le fait que prendre différents points d'évaluation conduit à différentes vitesses de convergence en lien avec la régularité et la parcimonie locales.
Pour faciliter l'interprétation des $16$ courbes par échantillon représentant chacune un point d'évaluation, on a ajouté à chaque sous-graphe et pour chaque échantillon leur moyenne (représentée par une ligne noire plus épaisse).
Notre but est de trouver une expression de $a$ en fonction des méta-données qui atteignent le minimum des moyennes pour chaque cas de figures (modèle, dimension, point d'évaluation, changement d'échantillon).
Précisons l'effet de $a$ sur le compromis biais-variance : plus $a$ est choisi grand, plus la fenêtre sélectionnée est grande, plus la procédure est stable, car les estimateurs sont de plus petite variance, mais plus la borne du biais est grande. Remarquez le comportement chaotique de l'erreur pour les petites valeurs de $a$, surtout quand l'échantillon est petit et la dimension grande. Ces valeurs sont donc à éviter dès que différents échantillons (de même couleur mais types de ligne différents sur les graphes) présentent une différence significative dans l'erreur absolue.

Inversement, pour les grandes valeurs de $a$, l'erreur stochastique disparaît et chaque échantillon a le même comportement. Cependant, pour ces valeurs de $a$, le risque croît rapidement avec le biais, à l'exception des faibles valeurs de $f$ (courbes bleues ou violettes) mais il est généralement facile d'estimer des valeurs nulles ou quasi-nulles de densités. En particulier, c'est automatique pour les estimateurs à noyau : quand la densité est très faible localement en $w$, avec très grande probabilité notre échantillon ne continent pas d'observations dans le voisinage de $w$ qui est le support de $K_h(w-\cdot)$ et donc l'estimateur à noyau qui est la moyenne sur les $\{K_h(w-W_i)\}_{i=1}^n$ devient nul.

Au vu des figures, je choisis pour toute la suite (sauf indication contraire) la calibration

$$a = \ln(d-1),$$

ce qui favorise les Modèles $2$ et $3$ en terme de compromis biais-variance. Mon choix est de ne pas s'ajuster aux caractéristiques inconnues de ma fonction cible, telles que la discontinuité, qui est hors du cadre de mon étude théorique. Le Modèle $1$ va donc de manière générale être sur-lissé.

## 3.b Calibration du pas itératif : sensibilité du paramètre $\beta$

La calibration du pas d'un algorithme itératif revient principalement à faire un compromis entre la précision et le temps d'exécution.

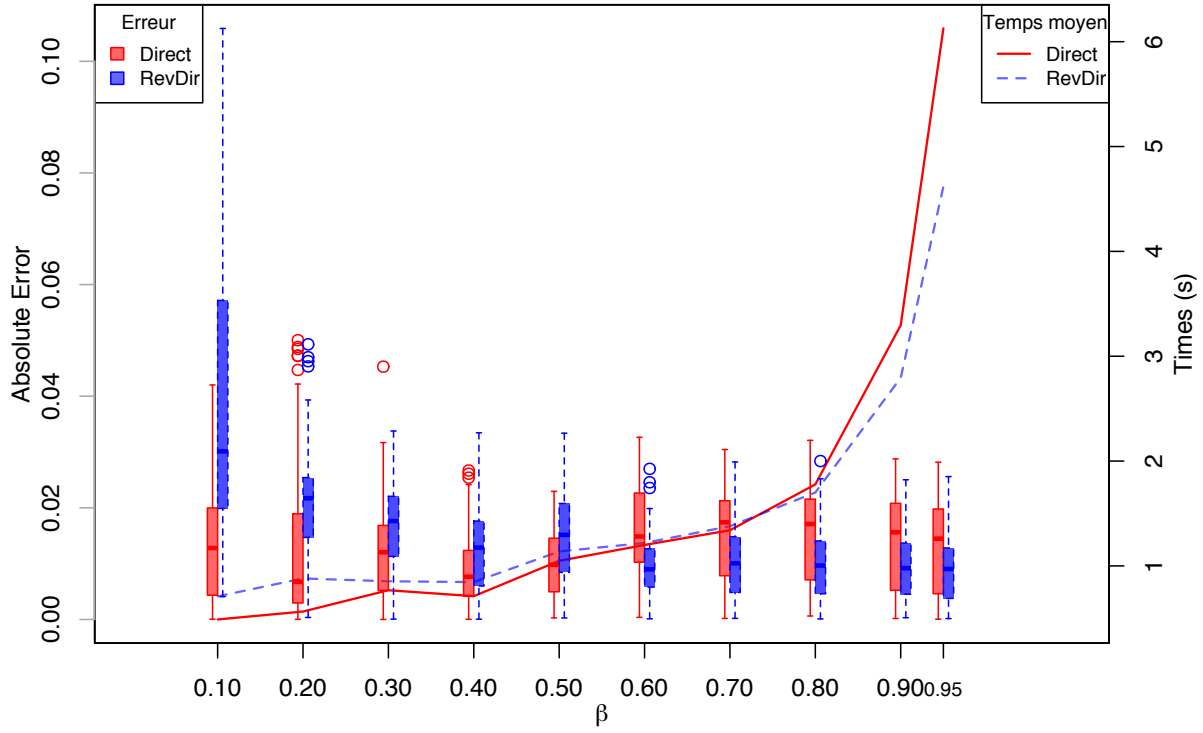
**Procédure de la calibration.** On va comparer en fonction de $\beta$ l'erreur absolue de nos deux procédures (sous forme de boxplots dans la Figure IV.2) versus leurs temps moyens d'exécution pour sélectionner la fenêtre (sous forme de lignes brisées).
On considère la grille de valeurs de $\beta$ suivante : $\{0.1, 0.2, \ldots, 0.9, 0.95\}$.
Pour constater l'effet de la dimension sur les temps, on répète la simulation pour deux dimensions de $x$ : $d_1 = 1$ et $d_1 = 3$ (sur deux sous-graphes différents).
Le paramètre $\beta$ étant moins sensible, on limite le nombre de cas : on prend le Modèle $3$ (car les deux autres ont une structure de parcimonie qui pourrait biaiser la comparaison pour des dimensions différentes), ainsi que la taille d'échantillon à $n = 100\,000$.

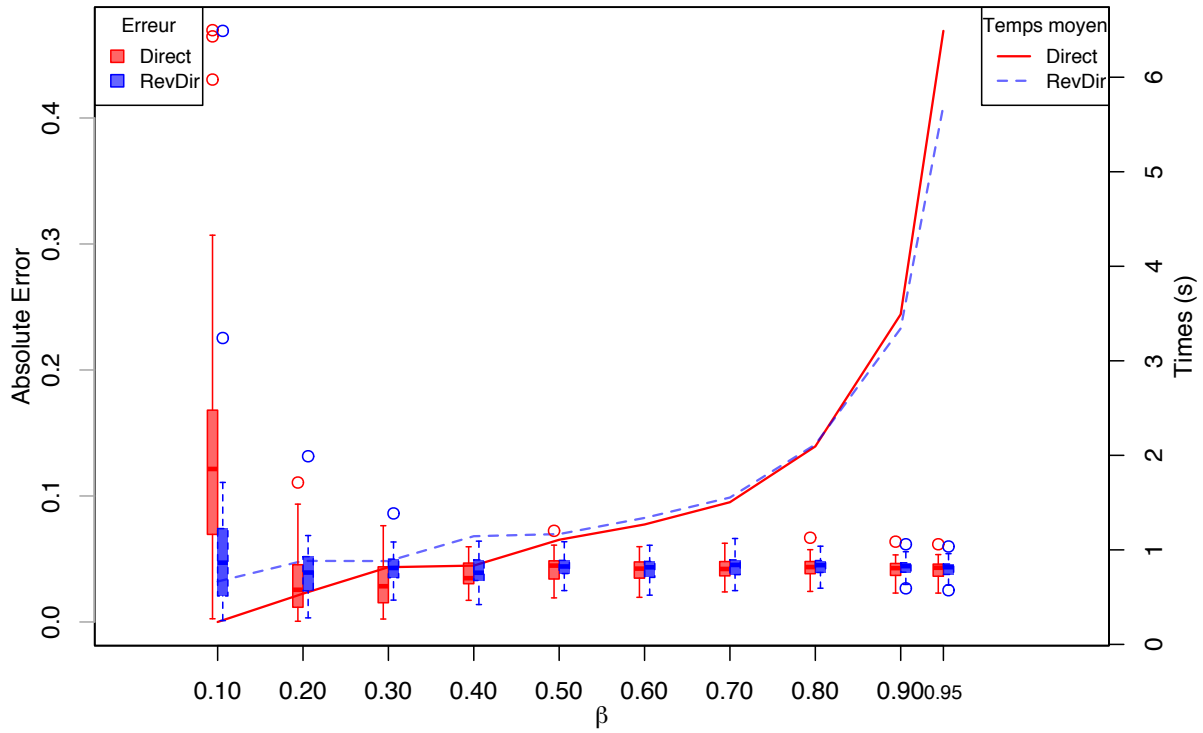
Pour évaluer la variance de l'erreur, on simule $B = 50$ échantillons pour construire un boxplot par procédure et par valeur de $\beta$, tandis les temps d'exécution sont moyennés sur ces $50$ simulations.

Le paramètre $a$ est sa valeur calibrée : $a = \log(d-1)$.

**Choix de $\beta$ et quelques remarques.** Précisons d'abord que $\beta$ est un pas multiplicatif : contrairement aux grilles régulières, deux valeurs consécutives sont proportionnelles par le facteur $\beta$. Ainsi, pour des valeurs de $\beta$ de décroissance raisonnable (c'est-à-dire pas trop proche de $0$), deux valeurs consécutives de la fenêtre sont du même ordre de grandeur. La décroissance jusqu'à la valeur sélectionnée prend donc quelques itérations, ce qui allonge le temps d'exécution mais assure de s'arrêter aux alentours du compromis $Z_{hj} \approx \lambda_{hj}$. Notons qu'aussi proche de $1$ que $\beta$ soit pris, l'erreur ne tends pas vers $0$ puisque changer $\beta$ ne diminue pas l'erreur stochastique (à taille d'échantillon $n$ fixée).
Inversement, pour des valeurs très petites de $\beta$ (ici, $0.1$ ou $0.2$), en une itération, les composantes actives dépasseront leurs valeurs idéales et la procédure s'arrête. Dans ce modèle où toutes les composantes sont pertinentes, toutes les composantes décroissent alors à des valeurs très basses.

FIGURE IV.2 – Calibration de $\beta$ pour le modèle 3 sur $B = 50$ échantillons de taille $n = 100\,000$ et de dimension $d = 1 + 2$ en haut et $d = 3 + 2$ en bas.



Donc on sur-apprend pour ces petites valeurs de $\beta$ : la quantité d'intérêt est donc l'écart type de l'estimation représenté par la longueur des boxplots, plus que la valeur moyenne des boxplots.

Analysons la Figure IV.2.
Concernant les temps d'exécution, il sont sans surprise exponentiellement croissants avec le pas multiplicatif $\beta$. On déconseille donc les valeurs $0.9$ et $0.95$.
Remarquons que les deux procédures Direct et RevDir ont des temps similaires.
Noter par ailleurs que les deux sous-graphes ont des courbes de temps très similaires : la dimension dans cet exemple impacte donc peu le temps d'exécution.
Concernant la précision d'estimation, les erreurs sont comme attendues plus petites en dimension inférieure. On déconseille les valeurs $0.1$, $0.2$ et $0.3$ du fait de la non stabilité des procédures.

Pour les valeurs intermédiaires de $\beta$ (entre $0.4$ et $0.8$), les procédures sont peu influencées par $\beta$. Plus pour fixer le paramètre que par choix spécifique, on prend pour la suite

$$\beta = 0.8.$$

# 4 Performances

## 4.a Reconstruction visuelles

Dans cette partie, on propose une visualisation de l'estimation de nos deux procédures contre la vraie densité dans chacun de nos modèles avec la dimension $d = 4$.

### 4.a.i Procédure de la reconstruction

Face à l'impossibilité de représenter un objet de dimension $4$, on ne fait varier qu'une composante à la fois, les autres composantes étant fixées selon un point d'évaluation de référence.
Ces points de référence pour chaque modèle ont été choisis de sorte que la force du signal $f$ soit suffisante. Plus précisément : le point de référence pour les Modèles $1$ et $2$ est $w = (0,0,0,0)$, tandis que pour le Modèle $3$, $w = (0.25, 0.25, 0.25, 0.5)$.
La composante qui varie est prise dans une grille régulière de $30$ points et centrée de manière à contenir la majorité de la masse de la loi jointe (évitant ainsi les zones plates sans observation).
Noter à ce sujet que le nombre local d'observations servant à estimer dépend de la densité jointe $f_W$, et non de notre cible $f$. Ainsi une valeur élevée de $f$ ne signifie pas qu'il y a beaucoup d'observations localement, et réciproquement l'annulation de $f$ ne signifie pas qu'il n'y en a pas.
On a pris les valeurs calibrées pour les paramètres $a$ et $\beta$, c'est-à-dire $a = \ln(d-1)$ et $\beta = 0.8$ pour nos deux procédures.
Les reconstructions pour chacun des trois modèles se trouvent respectivement en Figures IV.3, IV.4 et IV.5 selon chaque direction (indiquée en haut à gauche de chaque sous-graphe). Dans chaque graphique, la vraie densité est tracée en ligne noire pleine, chaque estimée de la procédure Direct (respectivement RevDir) pour chacun des $30$ points de la grille est représentée par un cercle rouge (respectivement un triangle bleu), éventuellement reliée par une corde en cas d'estimées voisines éloignées.

Avant de passer à l'analyse de ces résultats, noter que bien que nos trois modèles proposées ici ne présentent pas de bimodalité, l'exemple du Chapitre II (dont la reconstruction est en Figure II.2 mais seulement pour la procédure Directe) est bimodal dans la direction $x_1$, et CDRODEO ne semble pas gêné par cette structure.

FIGURE IV.3 – Comparaison des estimées Direct (cercles rouges) et RevDir (triangles bleus) contre la vraie densité (ligne pleine noire) pour le modèle 1.

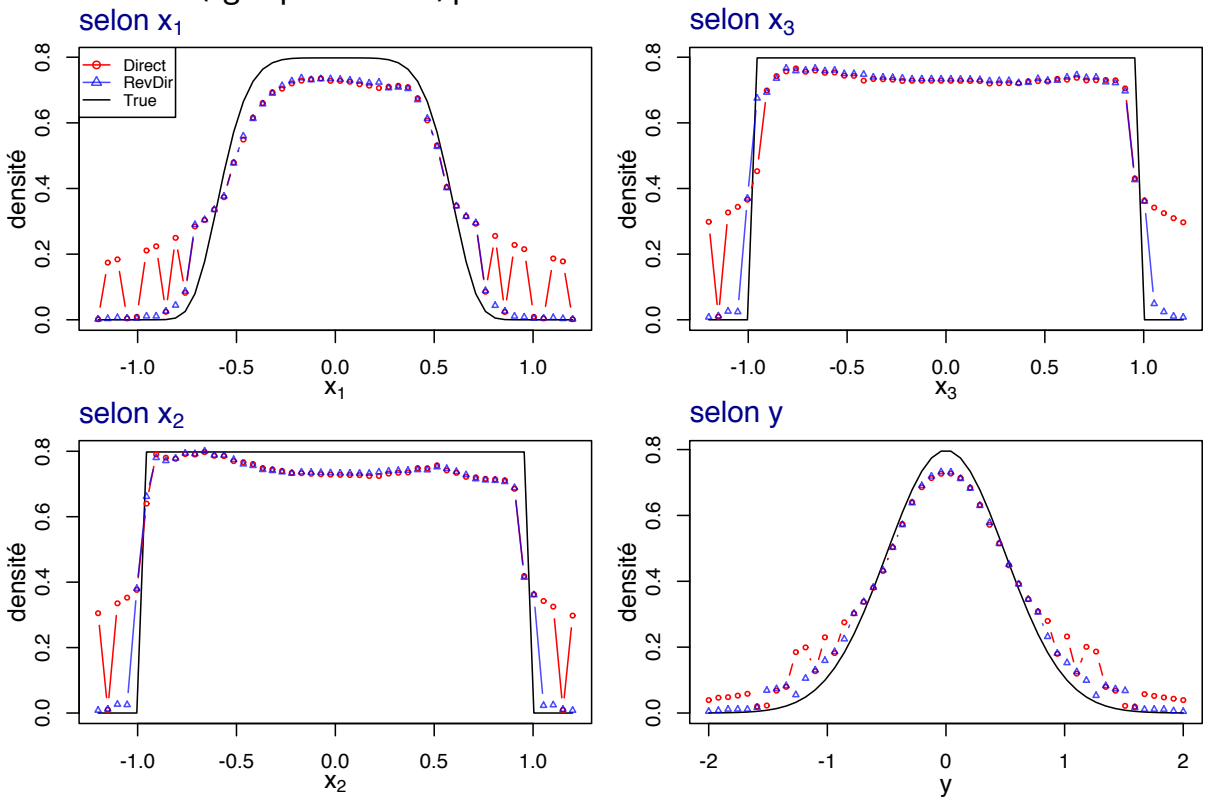



FIGURE IV.4 – Comparaison des estimées Direct (cercles rouges) et RevDir (triangles bleus) contre la vraie densité (ligne pleine noire) pour le modèle 2.

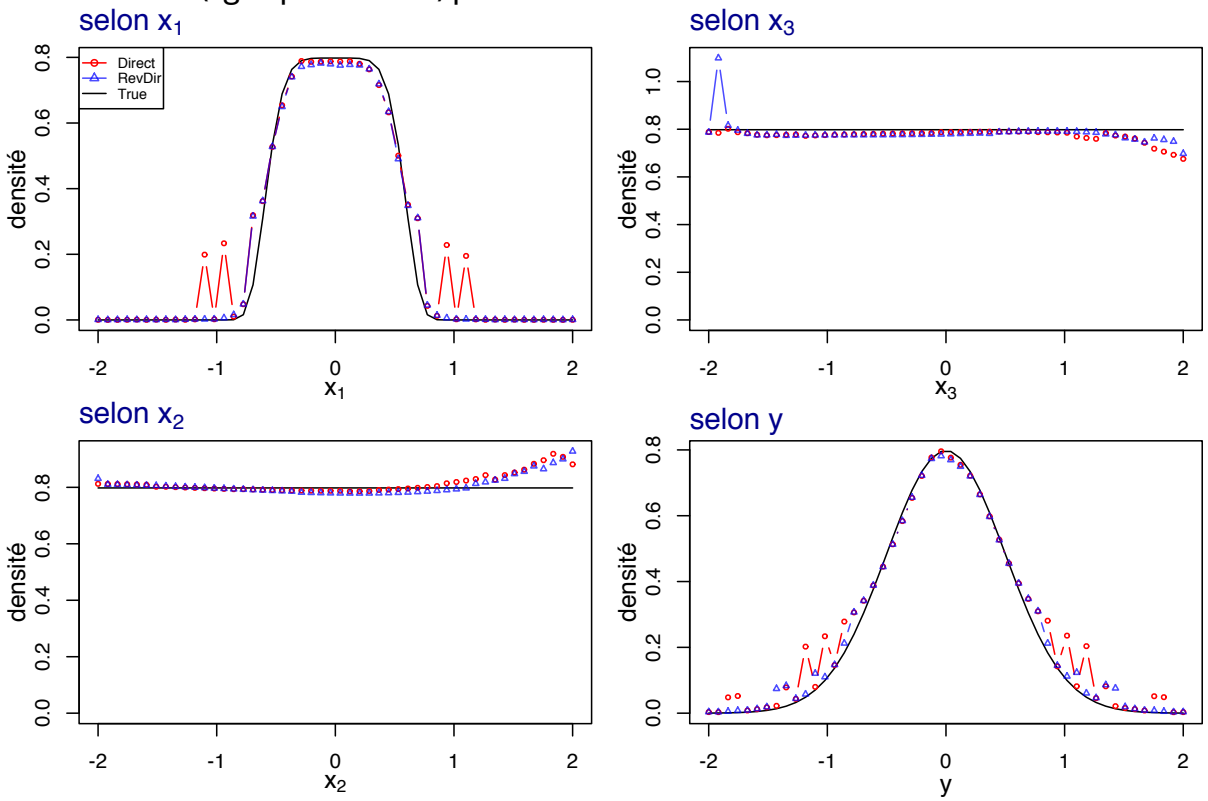


FIGURE IV.5 – Comparaison des estimées Direct (cercles rouges) et RevDir (triangles bleus) contre la vraie densité (ligne pleine noire) pour le modèle 3.

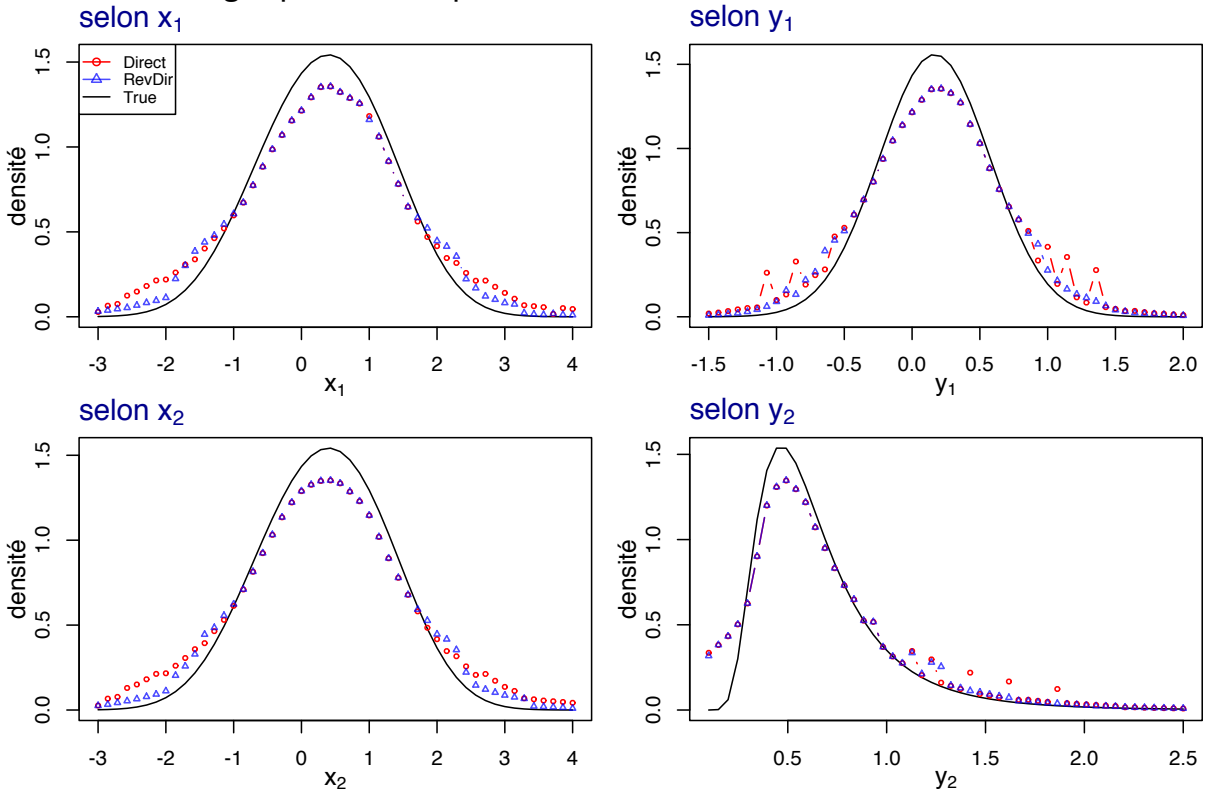
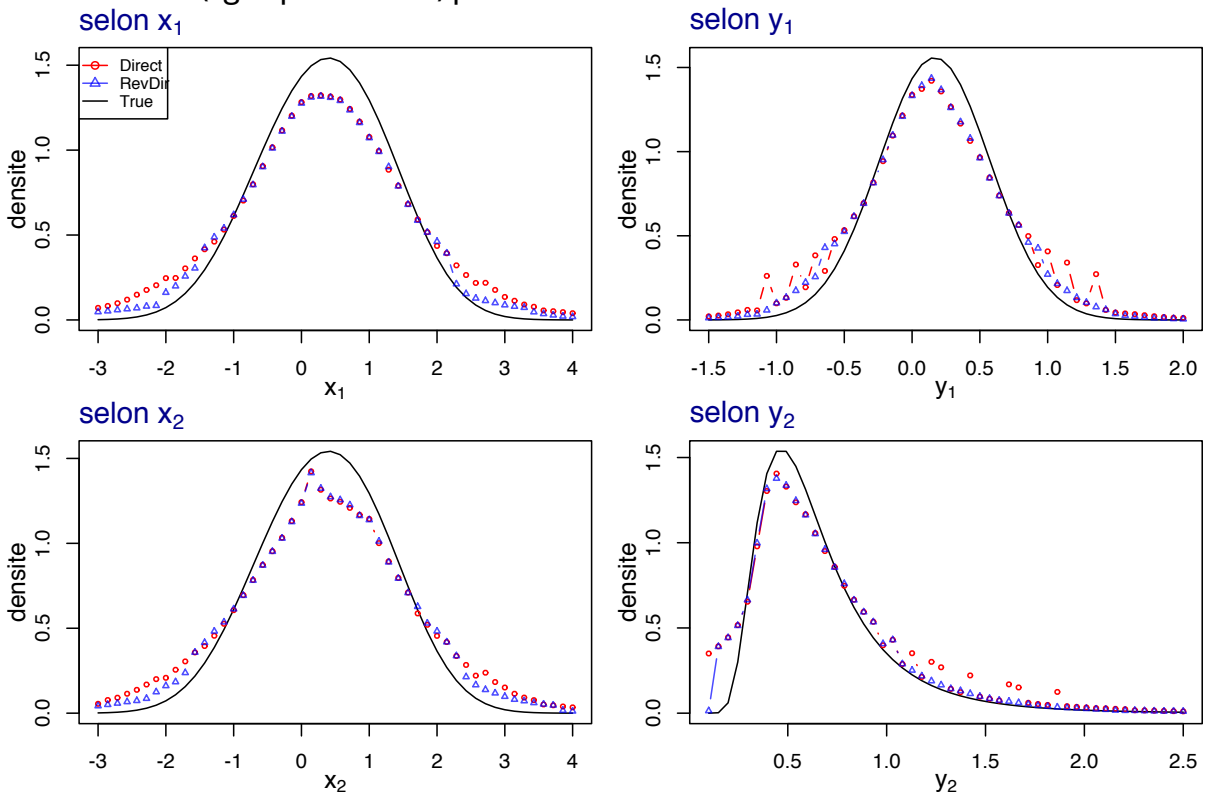



FIGURE IV.6 – Comparaison des estimées Direct (cercles rouges) et RevDir (triangles bleus) contre la vraie densité (ligne pleine noire) pour le modèle 3 avec une autre calibration de $a = 1$.



### 4.a.ii Analyse des reconstructions

De manière générale, les reconstructions sont informatives quant à la vraie densité, sauf dans des cas bien spécifiques dont j'explique précisément les raisons ci-après, en Section 4.a.iii.

En dehors de ces zones critiques, les deux procédures sont vraiment très similaires, ce qui est rassurant car construites à partir du même échantillon et grâce aux mêmes statistiques de test $Z_{hj}$, leurs seules différences sont l'intialisation et le chemin de fenêtres parcouru. Cela nous conforte dans l'intérêt d'étudier des algorithmes gloutons, puisque même en initialisant ailleurs et en construisant un chemin d'optimisation complètement différent, on continue à estimer aussi bien pourvu que quelques conditions soient réalisées.

Par ailleurs, on pourrait reprocher à notre calibration de $a$ de renvoyer un estimateur trop lissé. En particulier, dans les Modèles $1$ et $3$, on peut remarquer que nos estimées peinent à atteindre les modes et reportent l'excès de masse sur les voisinages des ces modes, syndromes typiques d'un sur-lissage. Bien que les Modèles $1$ et $2$ aient la même densité conditionnelle (en dehors des directions non pertinentes $x_2$ et $x_3$), il semblerait que le Modèle $2$ soit préservé de ce sur-lissage, sans doute grâce à sa grande parcimonie qui facilite son estimation en réduisant la dimension (alors que le Modèle $1$ doit gérer des points de discontinuité).
Bien que l'objectif initial d'un estimateur à noyau soit de proposer une version lisse de l'histogramme, regardons tout de même une calibration de $a$ légèrement inférieure : on fait la même procédure de reconstruction du Modèle $3$ en remplaçant uniquement $a = \log(d-1) \approx 1.1$ par $a = 1$ : voir Figure IV.6. La reconstruction dans les directions $y_1$ et $y_2$ en est améliorée significativement au niveau de leurs modes, en particulier en respectant mieux l'asymétrie dans la direction $y_2$. Cependant, de manière étonnante, le gain d'apprentissage n'est pas obtenue dans les directions $x_1$ et $x_2$. De plus, on peut noter un point de décrochage pour $x_2 = 0.2$, ce qui est suspect pour une méthode à noyau et est peut-être le signe de sur-apprentissage.

### 4.a.iii En dehors des hypothèses de régularité ou de « convexité » - le problème d'initialisation des méthodes gloutonnes.

On s'intéresse maintenant aux zones estimées « étrangement ». Plus spécifiquement, que se passe-t-il à mi-pente des courbes et pourquoi obtient-on ces estimées clairement au-dessus de la vraie densité et dont on devine prolongement sur quasiment une estimée sur $2$ dans ces zones pour la procédure Direct ?

Rappelons d'abord les hypothèses qu'on suppose pour nos résultats théoriques.

Dans le Chapitre II, on demande à ce que la densité conditionnelle $f$ soit $C^p$, $p$ l'ordre du noyau, et que dans les directions pertinentes $j$, la dérivée partielle d'ordre $p$ $\partial_j f^{(p)}$ soit non nulle au point d'évaluation $w$ (cf Assumption II.3) . Ici pour le noyau gaussien, $p = 2$. Et les zones étranges sont soit au niveau d'une discontinuité pure et dure (Modèle 1), soit justement aux points d'inflexion de $f$ pour lesquels la dérivée seconde s'annule.

Dans le Chapitre III, il faut que $h \mapsto |\bar{Z}_{hj}|$ soient monotones (cf Assumption $\mathcal{M}$). Noter que c'est très relié à ce que vérifie Assumption II.3 pour la régularité $s = p$ au vu de l'encadrement II.12 en notant l'intervention de $\partial_j f^{(p)}$ dans la constante des bornes.

C'est aussi là qu'est l'explication : l'annulation de $\partial_j f^{(p)}$ entraîne une annulation de $Z_{hj}$ à un certain niveau $h_\bullet$ de la composante $j$ de la fenêtre. Quand le chemin de CDRODEO atteint $h_j \approx h_\bullet$, la composante $j$ s'arrête de décroître, alors que le compromis biais-variance est sans doute réalisé à un niveau plus bas.

**Comment RevDir contourne le problème ?** Rappelons qu'une motivation majeure pour la recherche de la seconde procédure que l'on a nommée RevDir était de régler ce problème, que l'on peut aussi voir comme un problème d'initialisation de la procédure Direct. En effet, si l'on initialise $h_0 < h_\bullet$, il n'y a pas de problème, car le chemin ne passe pas par $h_\bullet$.
Par ailleurs, la théorie préconise de prendre $h_0 = \frac{1}{\log n}$, mais en pratique cette valeur est trop basse par rapport à celle du compromis biais-variance surtout en grande dimension (si l'on souhaite conserver une taille $n$ d'échantillon raisonnable). Par exemple, pour $n = 100\,000$, quand $s = p = 2$, $\frac{1}{\log n} \approx 0.087$, et une composante pertinente de la fenêtre doit être sélectionnée de l'ordre de $h_\star \approx n^{-\frac{1}{4+d}}$, mais dès la dimension $d = 1$, $n^{-\frac{1}{4+d}} > \frac{1}{\log n}$ et l'écart empire quand $d$ croît. De plus, noter que l'on ne connaît a priori pas $s$.

Le principe de RevDir d'autoriser aussi les fenêtres à croître permet de commencer en dessous du niveau $h_\star$ et de l'atteindre par en-dessous. Notons de plus qu'en Step Reverse, le test est inversé, l'annulation de $|Z_{hj}|$ ne désactive pas la composante $j$ à croître. Enfin, un dernier avantage de RevDir est qu'en explorant des fenêtres plus petites, les voisinages en tant que support de $K_h(w-\cdot)$ et sur lesquels est basée la majorité des hypothèses sont plus petits et donc moins contraignants. Ainsi, une irrégularité de type discontinuité, comme c'est le cas dans le Modèle $1$ en $x_j = \pm 1$ pour $j = 2, 3$, n'affecte que très localement les performance de RevDir.

Pour conclure, il y un côté rassurant à ce problème, car cela vérifie empiriquement la nécessité des hypothèses de régularité, qui aurait pu simplement être des artefacts des preuves théoriques. D'autre part, comme toute densité régulière (conditionnelle ou non) possède au moins deux points d'annulation de sa dérivée seconde, aux voisinages desquels CDRODEO rencontrera des problèmes d'estimation, cette analyse est plutôt contraignante pour l'algorithme glouton CDRODEO, même si la procédure RevDir limite grandement son effet.

## 4.b Impact de la dimension et détection de parcimonie

Avant de considérer des modèles parcimonieux, noter que l'ensemble des figures de la calibration de $a$ (voir Section 5.b) sur le Modèle $3$ (dans lequel $r = d$) met en lumière le fléau de la dimension. Pour les petites dimensions, la calibration est simple car le risque est faible sur une grande plage de valeurs de $a$ et de petits échantillons sont suffisants pour obtenir une bonne précision d'estimation. En revanche, à chaque ajout de variables (pertinentes), l'erreur stochastique augmente, rétrécissant la plage de de « bonnes » valeurs de $a$ jusqu'à la faire disparaître. On peut ainsi dresser une table de taille minimum $n_{\min}$ d'échantillon en fonction de la dimension pertinente :

| $r$ | 3 | 4 | 5 | 6 |
|---|---|---|---|---|
| $n_{\min}$ | $\ll 10\,000$ | $\lesssim 10\,000$ | $\approx 50\,000$ | $\gtrsim 200\,000$ |

La croissance de $n_{\min}$ semble exponentielle. Ce phénomène est à relier celle de la taille minimale $n_d$ d'échantillon en fonction de la dimension $d$ telle que pour une marge d'erreur $\varepsilon > 0$ fixée, la vitesse minimax optimale soit inférieure à $\varepsilon$, *i.e.* : pour une régularité $s$,

$$n_d{}^{-\frac{s}{2s+d}} < \varepsilon \quad \Leftrightarrow \quad n_d > \varepsilon^{-\frac{2s+d}{s}}.$$

### 4.b.i Robustesse à l'ajout de variables non pertinentes : les Modèles $1$ et $2$

Un des atouts majeurs de CDRODEO est la détection de variables non pertinentes permettant de contrer le fléau de la dimension en cas de parcimonie comme pour les Modèles $1$ et $2$.

FIGURE IV.7 – Modèle $1$ : boxplot de $50$ estimées par CDRODEO Direct (rouge) et CDRODEO RevDir (bleu) en fonction de la dimension $d$ et en comparaison avec la vraie valeur $f(w)$ (ligne violette).

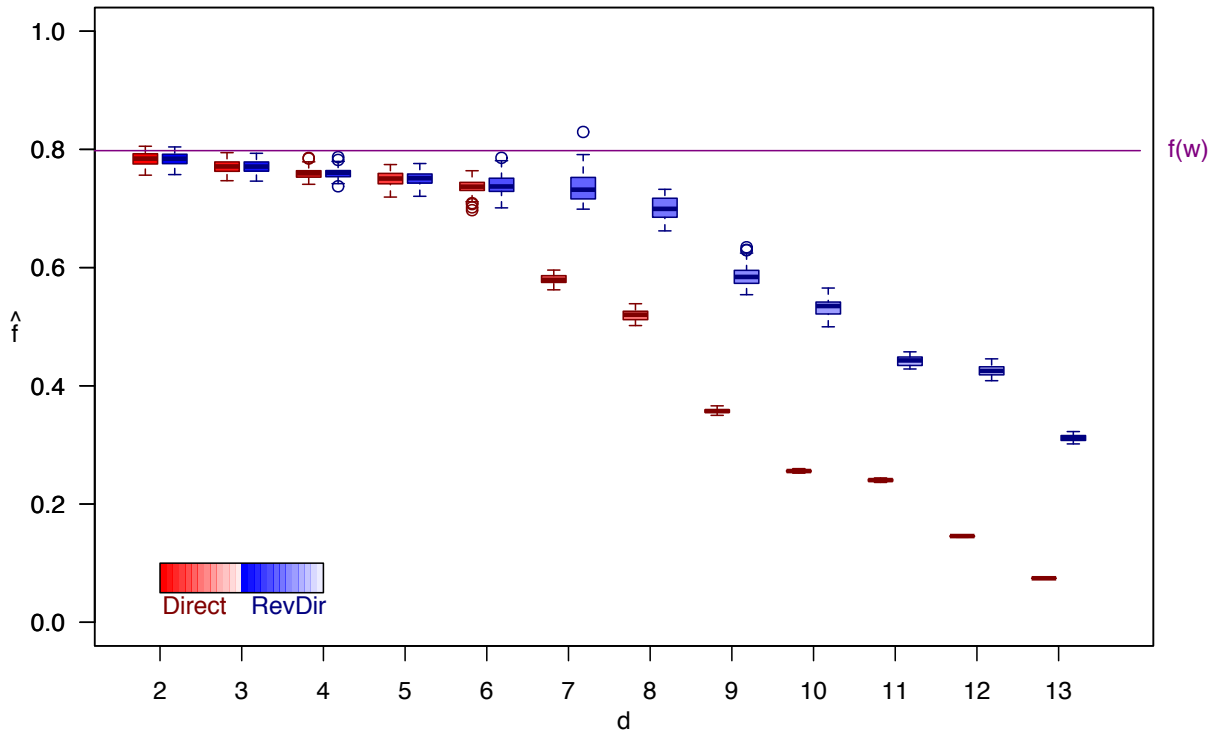



FIGURE IV.8 – Modèle $2$ : boxplot de $50$ estimées par CDRODEO Direct (rouge) et CDRODEO RevDir (bleu) en fonction de la dimension $d$ et en comparaison avec la vraie valeur $f(w)$ (ligne violette).

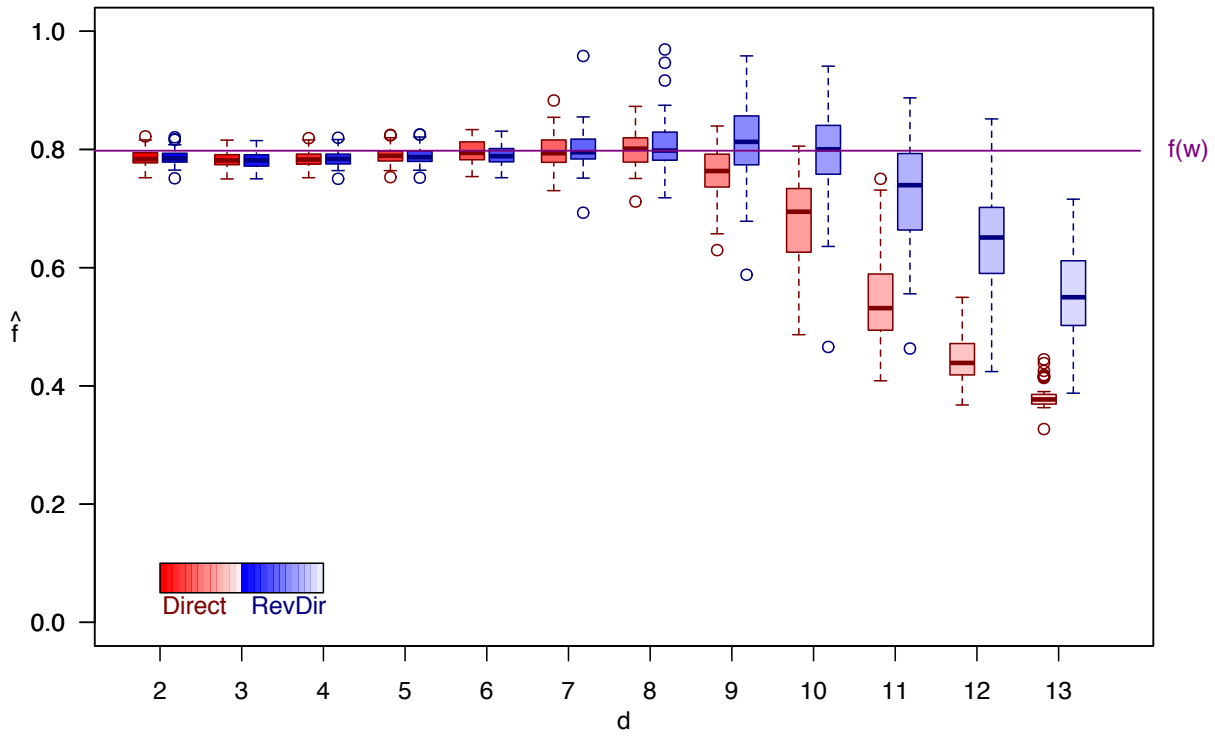


FIGURE IV.9 – Modèle 1 : boxplot des 50 fenêtres sélectionnées par CDRODEO Direct (à droite en couleurs chaudes) et CDRODEO RevDir (à gauche en couleurs froides) associé à la Figure IV.7 pour les dimensions $d = 2, 7, 12$.

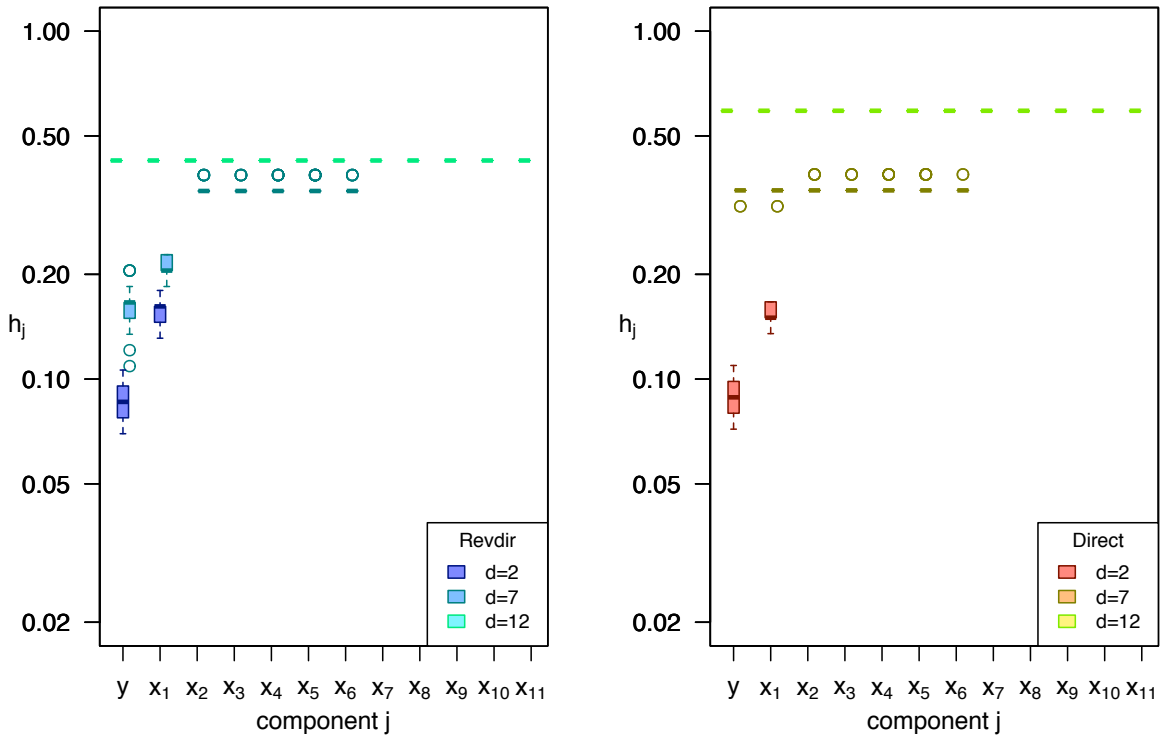
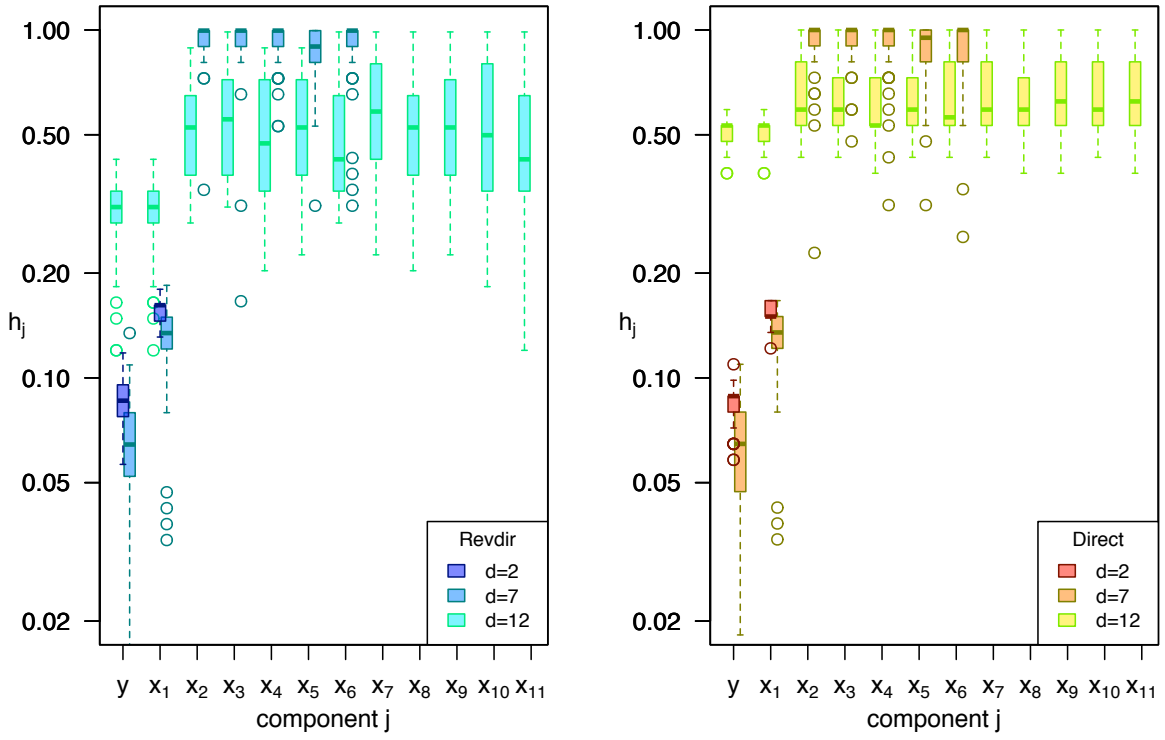



FIGURE IV.10 – Modèle 2 : boxplot des 50 fenêtres sélectionnées par CDRODEO Direct (à droite en couleurs chaudes) et CDRODEO RevDir (à gauche en couleurs froides) associé à la Figure IV.7 pour les dimensions $d = 2, 7, 12$.



**Procédure de la détection.** Pour les Modèles $1$ (voir Figure IV.7) et $2$ (voir Figure IV.8), on fait varier la dimension $d_1$ de $1$ à $12$ et on compare les boxplots de $B = 50$ estimées par nos deux procédures avec la valeur de la vraie densité $f(w)$.Les $B = 50$ échantillons sont de taille $n = 100\,000$, et les paramètres $a$ et $\beta$ sont pris selon la calibration effectuée : $a = \log(d-1)$ et $\beta = 0.8$.
On se place au point d'évaluation $w = 0$, centre de la masse de la loi jointe, et pour lequel la dimension pertinente locale est $r = 2$ pour les deux modèles.

Pour comprendre ce qui se passe au niveau de la sélection de la fenêtre, on trace aussi le boxplot des fenêtres sélectionnées, composante par composante, pour les dimensions $d = 2, 7$ et $12$. En Figure IV.9 pour le Modèle $1$ et Figure IV.10 pour le Modèle $2$, on trouve à gauche la procédure RevDir et à droite la procédure Direct. Sur chaque sous-graphe les fenêtres étant de taille différentes, on trouve les boxplots de $d = 2$ sur les deux premiers points en abscisse (en bleu foncé ou rouge), celle de $d = 7$ sur les $7$ premiers points d'abscisse (en bleu intermédiaire ou orange) et celle de la dimension $d = 12$ sur tout l'axe (en turquoise ou jaune). Noter que la graduation de l'axe des ordonnées est logarithmique.

**Analyse de la détection.** Intéressons-nous d'abord au Modèle $2$ (Figures IV.8 et IV.10), dont la structure de parcimonie est plus franche, car non gênée par des points de discontinuité.
Les deux procédures se montrent robustes à l'ajout de variables non pertinentes dans la limite d'une dimension totale $d = 9$ pour Direct et $d = 10$ pour RevDir (on rappelle que pour le Modèle $3$ sans parcimonie avec un échantillon de taille $100\,000$ n'estime bien que jusqu'à la dimension $5$).

Regardons ce qui se passe au niveau des fenêtres (voir Figure IV.10). On constate que les composantes pertinentes de la fenêtre, celles associées à $y$ et $x_1$, sont de même niveau pour la dimension $d = 2$ ou $d = 7$ (les deux couleurs plus foncées de chaque sous-graphe). Les composantes non pertinentes en dimension $d = 7$ sont sélectionnées majoritairement proche de $1$, ce qui illustre la qualité de détection de parcimonie des deux procédures CDRODEO.
Ce qu'il se passe en dimension $d = 12$ est que la détection de parcimonie est noyée dans le bruit de nombreuses composantes : le fléau de la dimension nous rattrape. En effet, les composantes non pertinentes sont désactivées plus basses et le compromis biais-variance $h_\star = n^{-\frac{1}{2s+d}}$ est aussi plus élevé en grande dimension pour contrôler l'excès de variance. Les composantes pertinentes n'ont alors pas de marge suffisante pour se distinguer des non pertinentes, ce qui détériore complètement l'estimation.
On remarque que c'est plus précoce pour la procédure Direct, et qu'en dimension $d = 12$, les composantes pertinentes sont plus élevées que pour RevDir. Je pense que cette différence vient du fait que RevDir, en cherchant le bon niveau de fenêtre par en-dessous, s'arrête plus tôt que Direct, qui le cherche par au-dessus.

Passons maintenant au Modèle $1$. La sélection de fenêtres et les estimées sur le Modèle $1$ sont très stables. Non seulement les variables auxiliaires du Modèle $1$ sont de variance plus petite ($\frac{1}{3}$ contre $1$ au Modèle $2$), mais surtout, la calibration de $a$ est la moins favorable pour ce modèle en sur-lissant. Ne souhaitant adapter la calibration aux discontinuités de notre fonction cible, je garde pour ces simulations $a = \log(d-1)$. Ainsi le compromis biais-variance n'est pas vraiment réalisé pour ce modèle.

La robustesse est moins bonne que pour le Modèle $2$ : les estimées se détériorent dès la dimension $d = 7$ pour Direct et $d = 9$ pour RevDir, soit une différence de $2$ par rapport au Modèle $2$.
Regardons les fenêtres. En dimension $2$, c'est très similaire au Modèle $2$ (rappelons que l'on estime la même densité conditionnelle).
La différence se passe dans la détection de composantes non pertinentes. En dimension $7$, on voit que leurs niveaux est sous la barre des $0.5$, bien plus loin de $1$ que pour le Modèle $2$. Côté RevDir

(gauche), on voit que pour compenser ses $5$ composantes non pertinentes plus basse, les deux composantes pertinentes sont plus élevées que dans le Modèle $2$. Côté Direct (droite), CDRODEO ne fait quasiment plus la différence entre composantes pertinentes et non pertinentes.
En dimension $d = 12$ du Modèle $1$, plus aucune distinction n'est faite entre composantes pertinentes ou non pertinentes, en RevDir comme en Direct, et l'estimation est très mauvaise.
Si on décalait le point d'évaluation $w$ plus vers les points de discontinuité (par exemple $w = (0.5, \dots, 0.5)$, on pourrait observer des résultats bien pires en terme de résistance au fléau de la dimension. Il semblerait que ce soient vraiment ces discontinuités qui affectent la détection de variables non pertinentes même en $w = 0$. Peut-être qu'en prenant un noyau de support plus petit, on pourrait réduire l'effet de ces discontinuités à de plus petits voisinages.

# 5 Appendices

## 5.a Lois jointes, marginales et conditionnelles du Modèle 3.

On rappelle le modèle hiérarchique considéré :

$$\begin{aligned} Y_{i2} &\sim \mathcal{IG}(\alpha = 4, \beta = 3) \\ Y_{i1}|Y_{i2} &\sim \mathcal{N}(0, Y_{i2}) \\ X_{ij}|Y_i &\overset{\text{iid}}{\sim} \mathcal{N}(Y_{i1}, Y_{i2}), \text{ pour } j = 1, \dots, d_1 \end{aligned}$$

On note $\Gamma$ la fonction gamma définie par $z \in \mathbb{R}_{>0} \mapsto \int_0^\infty t^{z-1} e^{-t} dt$. On définit la loi de Student multivariée, notée $t_\nu(\mu, \Sigma)$ avec $p$ la dimension, $\nu$ le degré de liberté, $\mu \in \mathbb{R}^p$ le paramètre de localisation et $\Sigma \in \mathcal{M}_p(\mathbb{R})$ définie positive le paramètre de forme, par sa densité sur $\mathbb{R}^p$ :

$$u \mapsto \frac{\Gamma(\frac{\nu+p}{2})}{\Gamma(\frac{\nu}{2})\sqrt{(\pi\nu)^p \det(\Sigma)}} \left(1 + \frac{1}{\nu}(u-\mu)\Sigma^{-1}(u-\mu)^\top\right)^{-\frac{\nu+p}{2}}.$$

**- Marginale de $Y_{i2}$** $\sim \mathcal{IG}(\alpha = 4, \beta = 3)$

$$f_{Y_{i2}}(y_2) = \mathbb{1}_{y_2>0} \frac{\beta^\alpha}{\Gamma(\alpha)} y_2^{-(\alpha+1)} e^{-\frac{\beta}{y_2}}$$

**- Loi conditionnelle de $Y_{i1}|Y_{i2}$** $\sim \mathcal{N}(0, Y_{i2})$

$$f_{Y_{i1}|Y_{i2}=y_2}(y_1) = \frac{1}{\sqrt{2\pi\theta_2}} e^{-\frac{y_1^2}{2y_2}}$$

**- Loi conditionnelle de $X_{i1}|Y_i$** $\sim \mathcal{N}(Y_{i1}, Y_{i2})$

$$f_{X_{i1}|Y_i=y}(x) = \frac{1}{\sqrt{2\pi y_2}} e^{-\frac{(x-y_1)^2}{2y_2}}$$

**- Loi jointe de** $(X_i, Y_i)$

$$\begin{aligned} f_{X_iY_i}(x,y) &= f_{Y_{i2}}(y_2) \times f_{Y_{i1}|Y_{i2}=y_2}(y_1) \times \prod_{j=1}^{d_1} f_{X_{ij}|Y_i=y}(x_j) \\ &= \mathbb{1}_{y_2>0} \frac{\beta^\alpha}{\Gamma(\alpha)} y_2^{-(\alpha+1)} e^{-\frac{\beta}{y_2}} \times \frac{1}{\sqrt{2\pi y_2}} e^{-\frac{y_1^2}{2y_2}} \times \prod_{j=1}^{d_1} \left( \frac{1}{\sqrt{2\pi y_2}} e^{-\frac{(x_j-y_1)^2}{2y_2}} \right) \\ &= \mathbb{1}_{y_2>0} \frac{\beta^\alpha}{\Gamma(\alpha)\sqrt{2\pi}^{d_1+1}} y_2^{-(\alpha+1+\frac{d_1+1}{2})} e^{-\frac{1}{2y_2}\left(2\beta+y_1^2+\sum_{j=1}^{d_1}(x_j-y_1)^2\right)} \end{aligned}$$

Remarque :

$$\sum_{j=1}^{d_1} (x_j - y_1)^2 = \sum_{j=1}^{d_1} x_j^2 - 2y_1 \sum_{j=1}^{d_1} x_j + d_1 y_1^2.$$

D'où :

$$\begin{aligned} f_{X_iY_i}(x,y) &= \mathbb{1}_{y_2>0} \frac{\beta^\alpha}{\Gamma(\alpha)\sqrt{2\pi}^{d_1+1}} y_2^{-(\alpha+1+\frac{d_1+1}{2})} e^{-\frac{1}{2y_2}\left(2\beta+(d_1+1)y_1^2+\sum_{j=1}^{d_1}x_j^2-2y_1\sum_{j=1}^{d_1}x_j\right)} \\ &= \mathbb{1}_{y_2>0} \frac{\beta^\alpha y_2^{-(\alpha+1+\frac{d_1+1}{2})}}{\Gamma(\alpha)\sqrt{2\pi}^{d_1+1}} \exp\left( -\frac{1}{2y_2}(2\beta + \sum_{j=1}^{d_1} x_j^2 - \frac{(\sum_{j=1}^{d_1} x_j)^2}{d+1} \right) \exp\left( -\frac{\left(y_1 - \frac{\sum_{j=1}^{d_1} x_j}{d_1+1}\right)^2}{2y_2/(d_1+1)} \right) \\ &= \mathbb{1}_{y_2>0} \frac{\beta^\alpha y_2^{-(\alpha+1+\frac{d_1+1}{2})}}{\Gamma(\alpha)\sqrt{2\pi}^{d_1+1}} e^{-\frac{\beta_1(x)}{y_2}} \exp\left( -\frac{\left(y_1 - \frac{\sum_{j=1}^{d_1} x_j}{d_1+1}\right)^2}{2y_2/(d_1+1)} \right), \end{aligned}$$

où $\beta_1(x) := \frac{1}{2}(2\beta + \sum_{j=1}^{d_1} x_j^2 - \frac{(\sum_{j=1}^{d_1} x_j)^2}{d_1+1})$.

**- Marginale de** $(X_i, Y_{i2})$

$$\begin{aligned} f_{X_iY_{i2}}(x,y_2) &= \int_{y_1} f_{X_iY_i}(x,(y_1,y_2))dy_1 \\ &= \mathbb{1}_{y_2>0} \frac{\beta^\alpha y_2^{-(\alpha+1+\frac{d_1+1}{2})}}{\Gamma(\alpha)\sqrt{2\pi}^{d_1+1}} \exp\left(-\frac{\beta_1(x)}{y_2}\right) \int_{y_1} \exp\left( -\frac{\left(y_1 - \frac{\sum_{j=1}^{d_1} x_j}{d+1}\right)^2}{2y_2/(d_1+1)} \right) dy_1 \\ &= \mathbb{1}_{y_2>0} \frac{\beta^\alpha y_2^{-(\alpha+1+\frac{d_1}{2})}}{\Gamma(\alpha)\sqrt{2\pi}^{d_1}\sqrt{d_1+1}} \exp\left(-\frac{\beta_1(x)}{y_2}\right) \end{aligned}$$

**- Marginale de $X_i$** $\sim t_{2\alpha}\left(0_{d_1}, 2\alpha(I_{d_1} - \frac{1}{(d_1+1)}E)\right)$

$$
\begin{aligned}
f_{X_i}(x) &= \int_{y_2} f_{X_i Y_{i2}}(x, y_2) dy_2 \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)\sqrt{2\pi}^{d_1}\sqrt{d_1+1}} \int_{y_2>0} y_2^{-(\alpha+1+\frac{d_1}{2})} e^{-\frac{\beta_1(x)}{y_2}} dy_2 \\
&= \frac{\beta^\alpha \Gamma(\alpha+\frac{d_1}{2})}{(\beta_1(x))^{\alpha+\frac{d_1}{2}}\Gamma(\alpha)\sqrt{2\pi}^{d_1}\sqrt{d_1+1}} \\
&= \frac{\Gamma(\alpha+\frac{d_1}{2})}{\left(1+\frac{1}{2\beta}\left(\sum_{j=1}^{d_1} x_j^2 - \frac{(\sum_{j=1}^{d_1} x_j)^2}{d_1+1}\right)\right)^{\alpha+\frac{d_1}{2}} \Gamma(\alpha)\sqrt{2\pi\beta}^{d_1}\sqrt{d_1+1}}
\end{aligned}
$$

**- Loi conditionnelle de $Y_i|X_i$**

$$
\begin{aligned}
f_{Y_i|X_i=x}(y_1, y_2) &= \frac{f_{X_i Y_i}(x, y)}{f_{X_i}(x)} \\
&= \mathbb{1}_{y_2>0} \frac{\beta^\alpha y_2^{-(\alpha+1+\frac{d_1+1}{2})}}{\Gamma(\alpha)\sqrt{2\pi}^{d_1+1}} e^{-\frac{\beta_1(x)}{y_2}} \exp\left(-\frac{\left(y_1 - \frac{\sum_{j=1}^{d_1} x_j}{d_1+1}\right)^2}{2y_2/(d_1+1)}\right) \Big/ \left(\frac{\beta^\alpha \Gamma(\alpha+\frac{d_1}{2})}{(\beta_1(x))^{\alpha+\frac{d_1}{2}}\Gamma(\alpha)\sqrt{2\pi}^{d_1}\sqrt{d_1+1}}\right) \\
&= \mathbb{1}_{y_2>0} \frac{\sqrt{d_1+1}}{\sqrt{2\pi}\Gamma(\alpha+\frac{d_1}{2})} (\beta_1(x))^{\alpha+\frac{d_1}{2}} y_2^{-(\alpha+1+\frac{d_1+1}{2})} e^{-\frac{\beta_1(x)}{y_2}} \exp\left(-\frac{\left(y_1 - \frac{\sum_{j=1}^{d_1} x_j}{d_1+1}\right)^2}{2y_2/(d_1+1)}\right)
\end{aligned}
$$

## 5.b Figures supplémentaires de la calibration de $a$

**Modèle 1. Procédure Direct**

FIGURE IV.11 – Calibration de $a$ pour le Modèle $1$ pour la procédure Direct sur des échantillons de taille $n = 10\,000$.

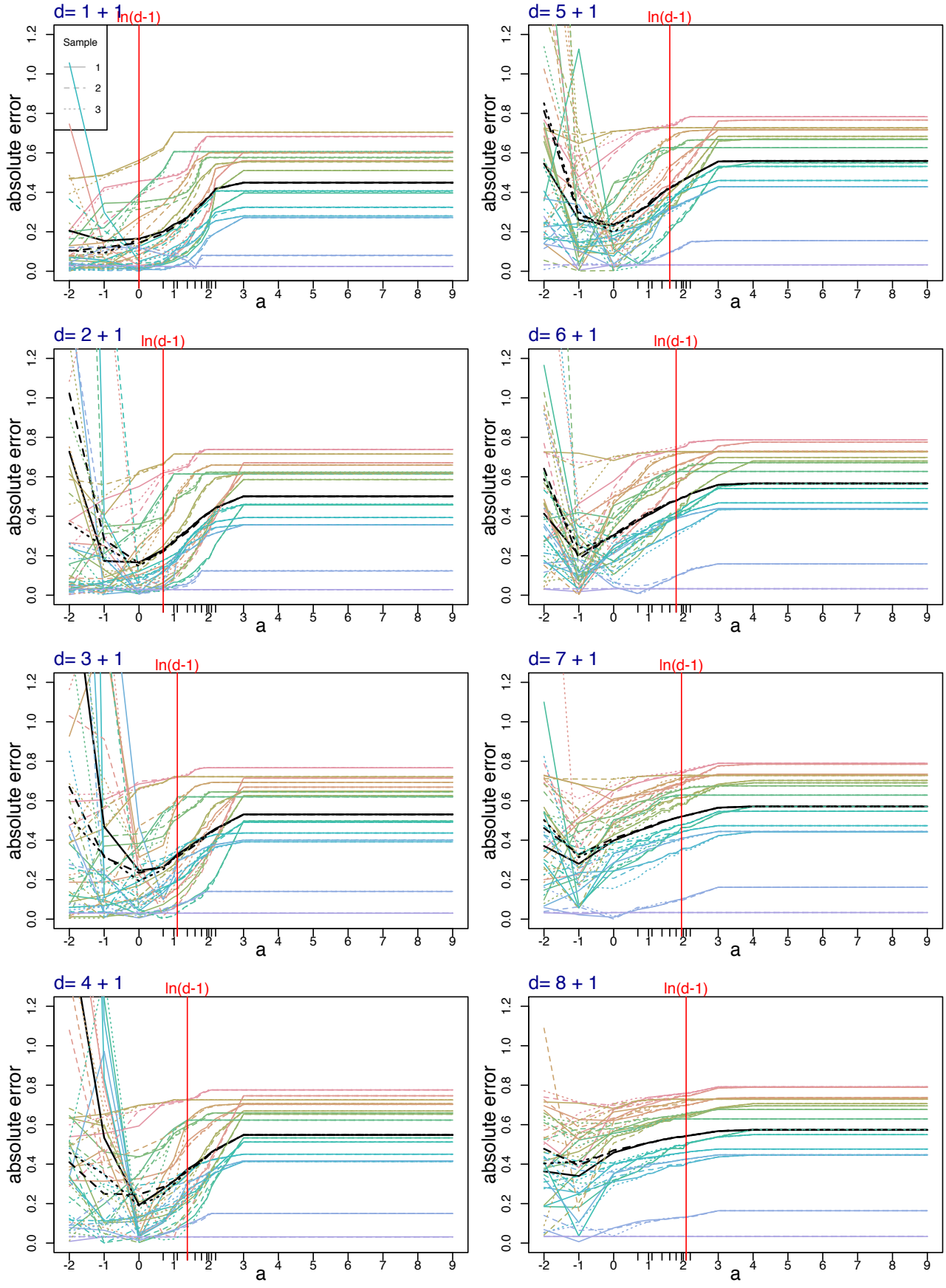


FIGURE IV.12 – Calibration de $a$ pour le Modèle 1 pour la procédure Direct sur des échantillons de taille $n = 25\,000$.

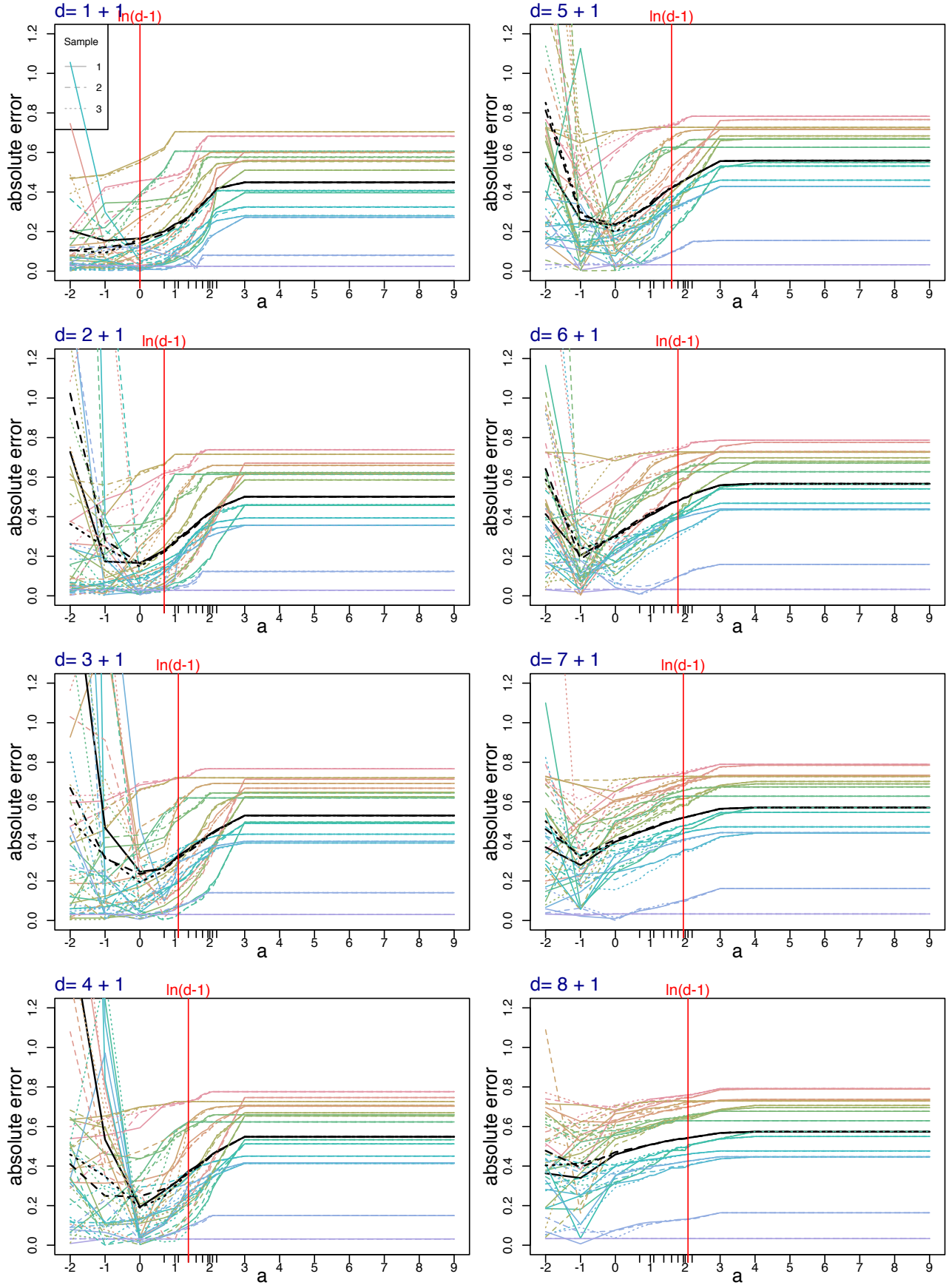


FIGURE IV.13 – Calibration de $a$ pour le Modèle 1 pour la procédure Direct sur des échantillons de taille $n = 50\,000$.

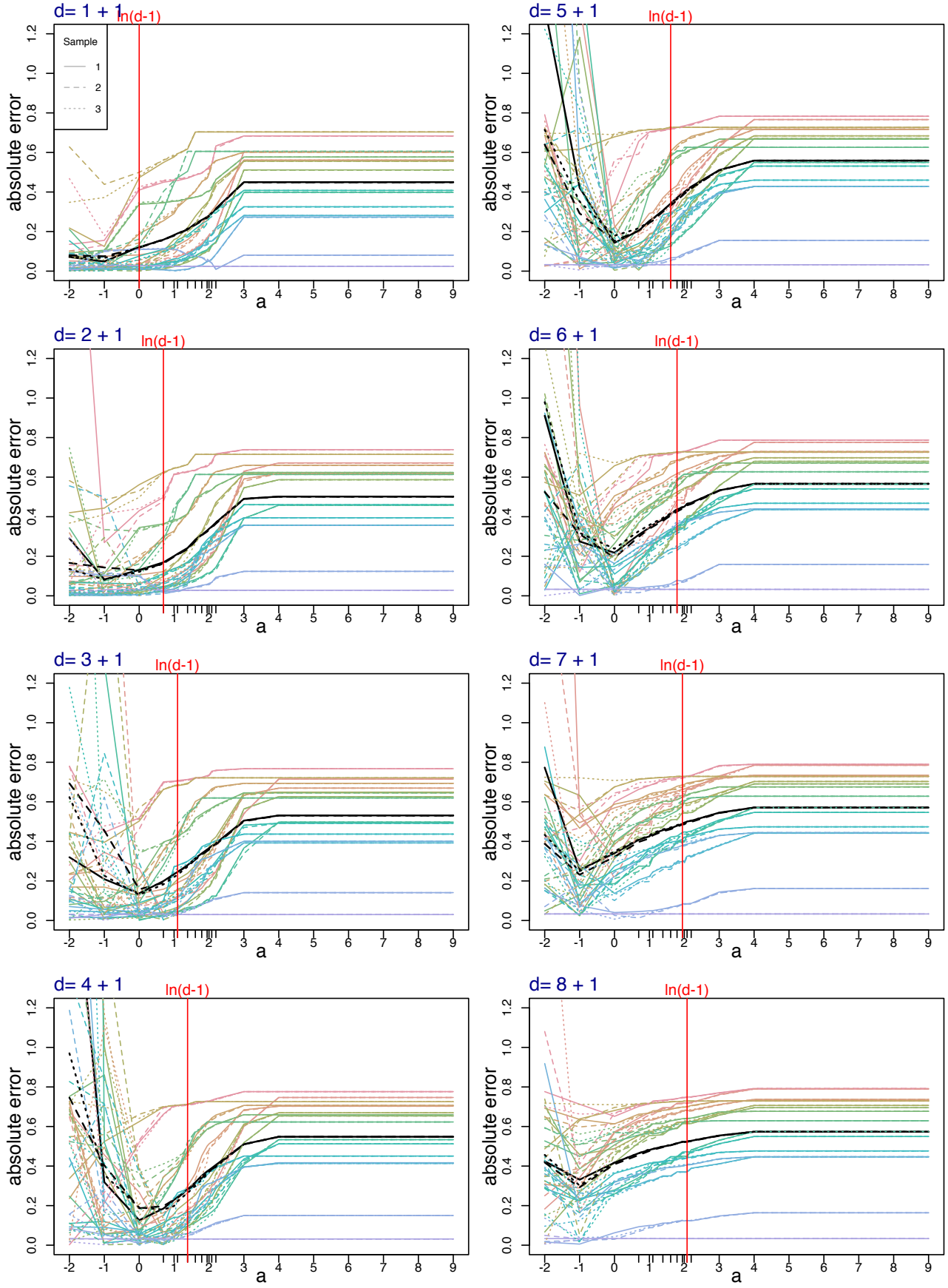


FIGURE IV.14 – Calibration de $a$ pour le Modèle 1 pour la procédure Direct sur des échantillons de taille $n = 100\,000$.

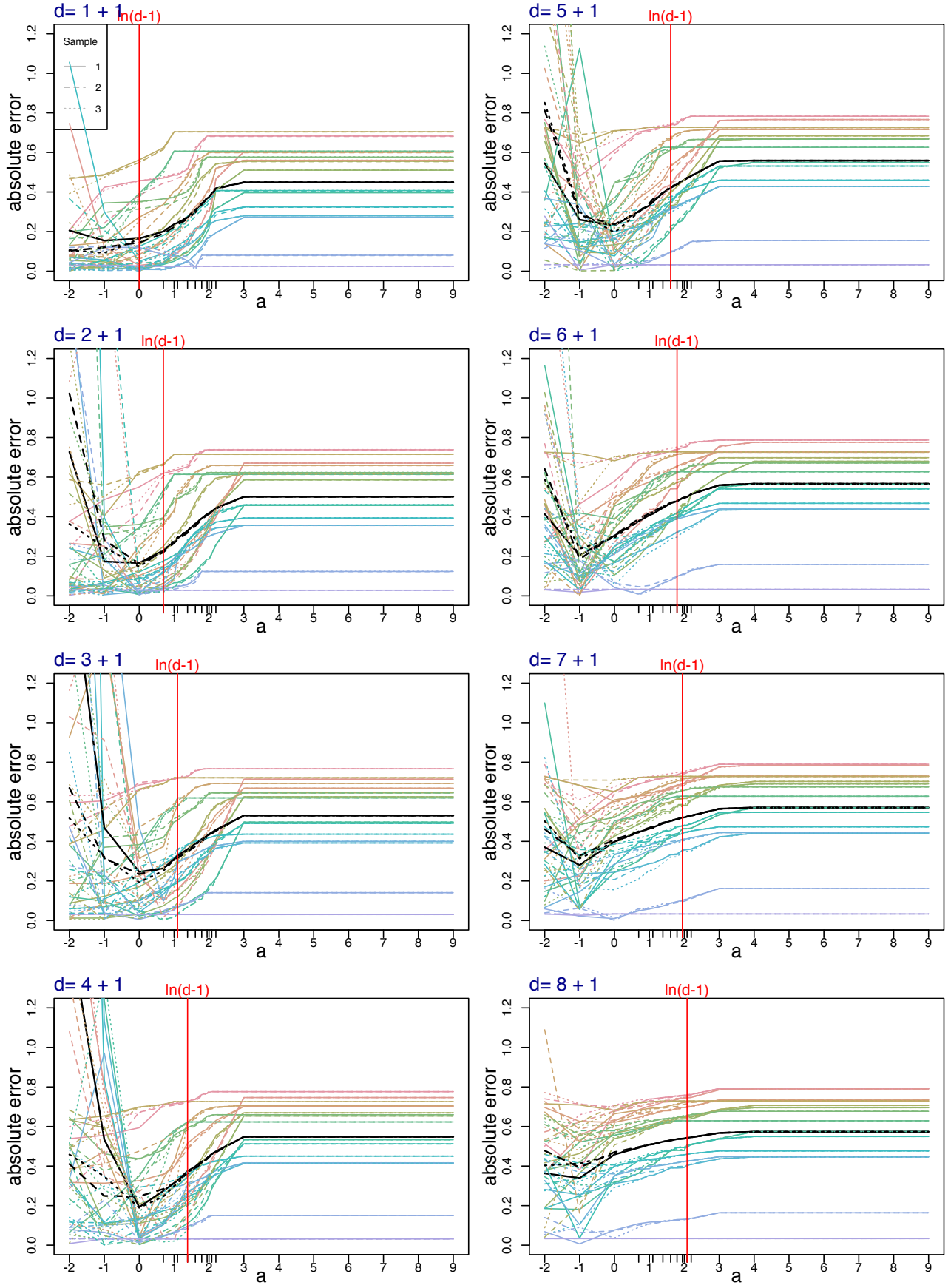
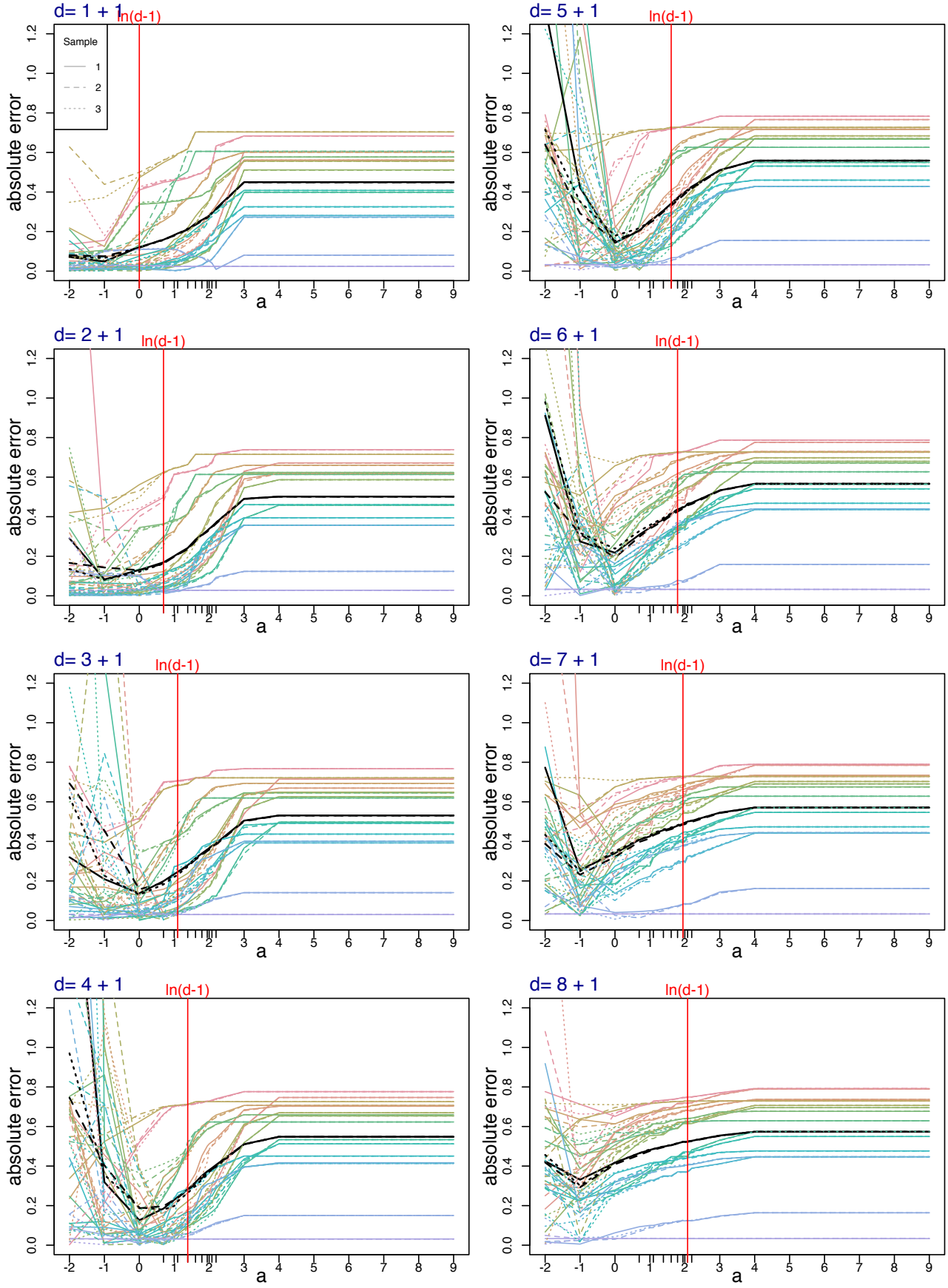


FIGURE IV.15 – Calibration de $a$ pour le Modèle 1 pour la procédure Direct sur des échantillons de taille $n = 200\,000$.



**Procédure Revdir**

FIGURE IV.16 – Calibration de $a$ pour le Modèle $1$ pour la procédure RevDir sur des échantillons de taille $n = 10\,000$.

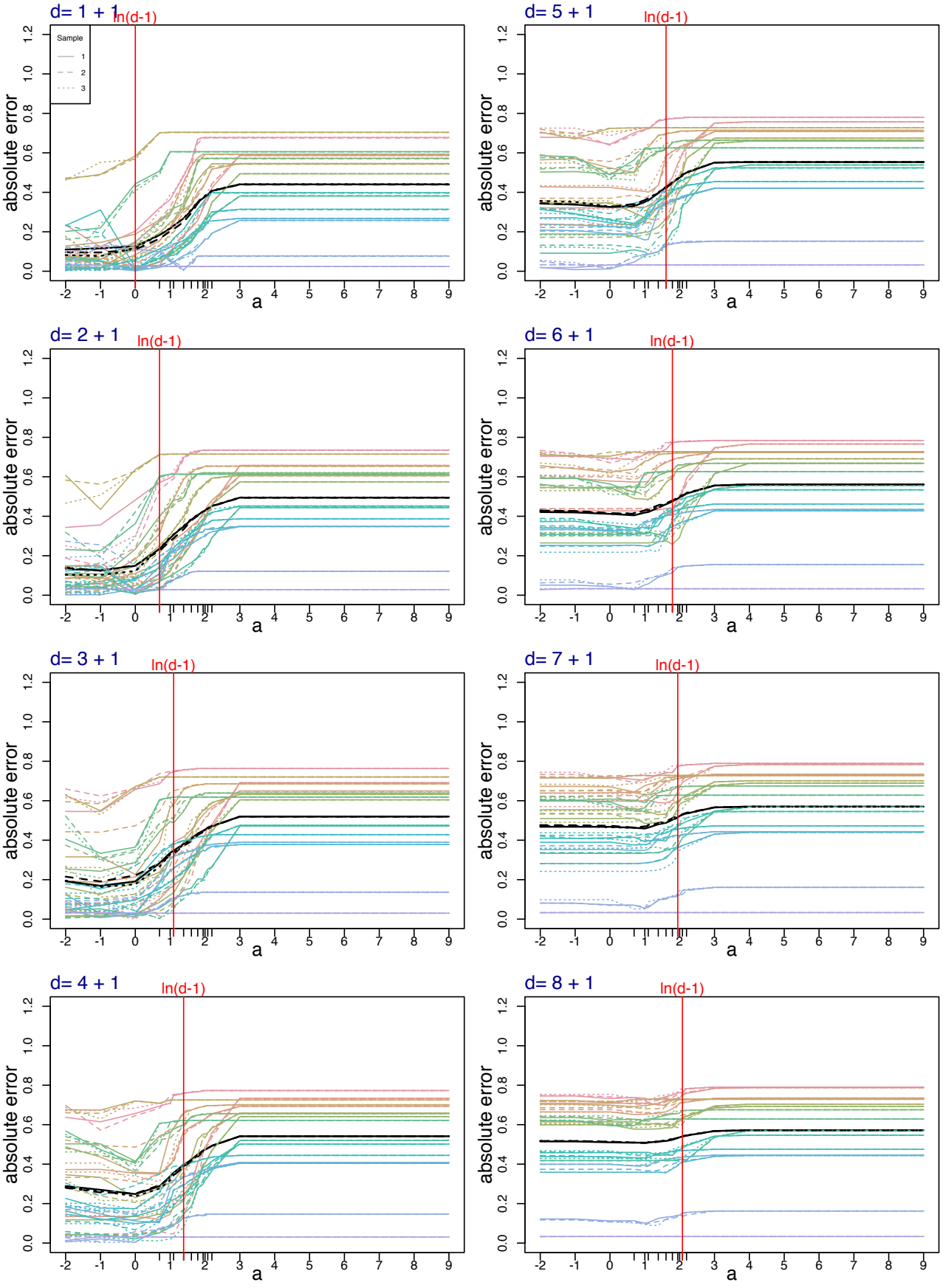


FIGURE IV.17 – Calibration de $a$ pour le Modèle 1 pour la procédure RevDir sur des échantillons de taille $n = 25\,000$.

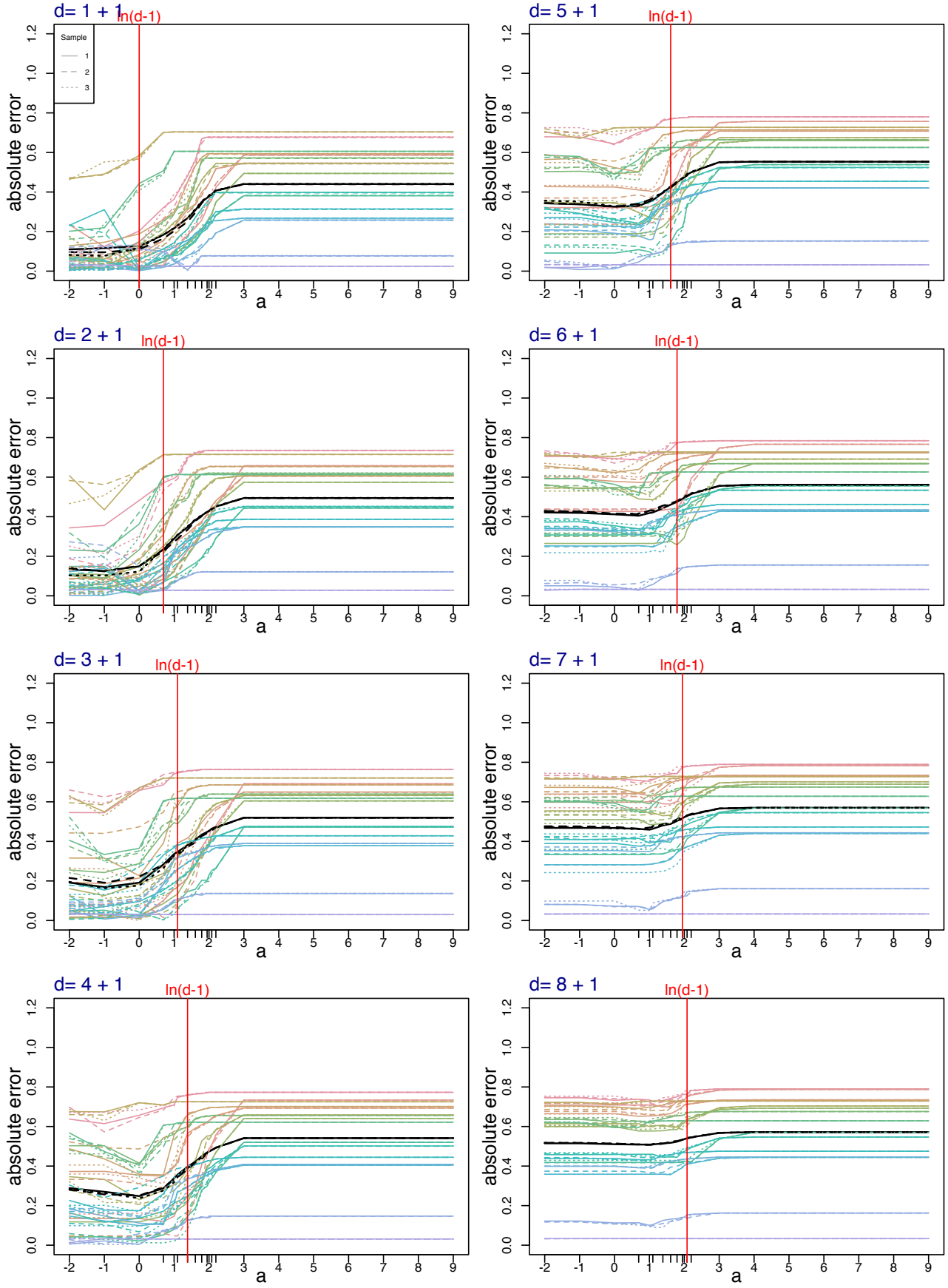


FIGURE IV.18 – Calibration de $a$ pour le Modèle $1$ pour la procédure RevDir sur des échantillons de taille $n = 50\,000$.

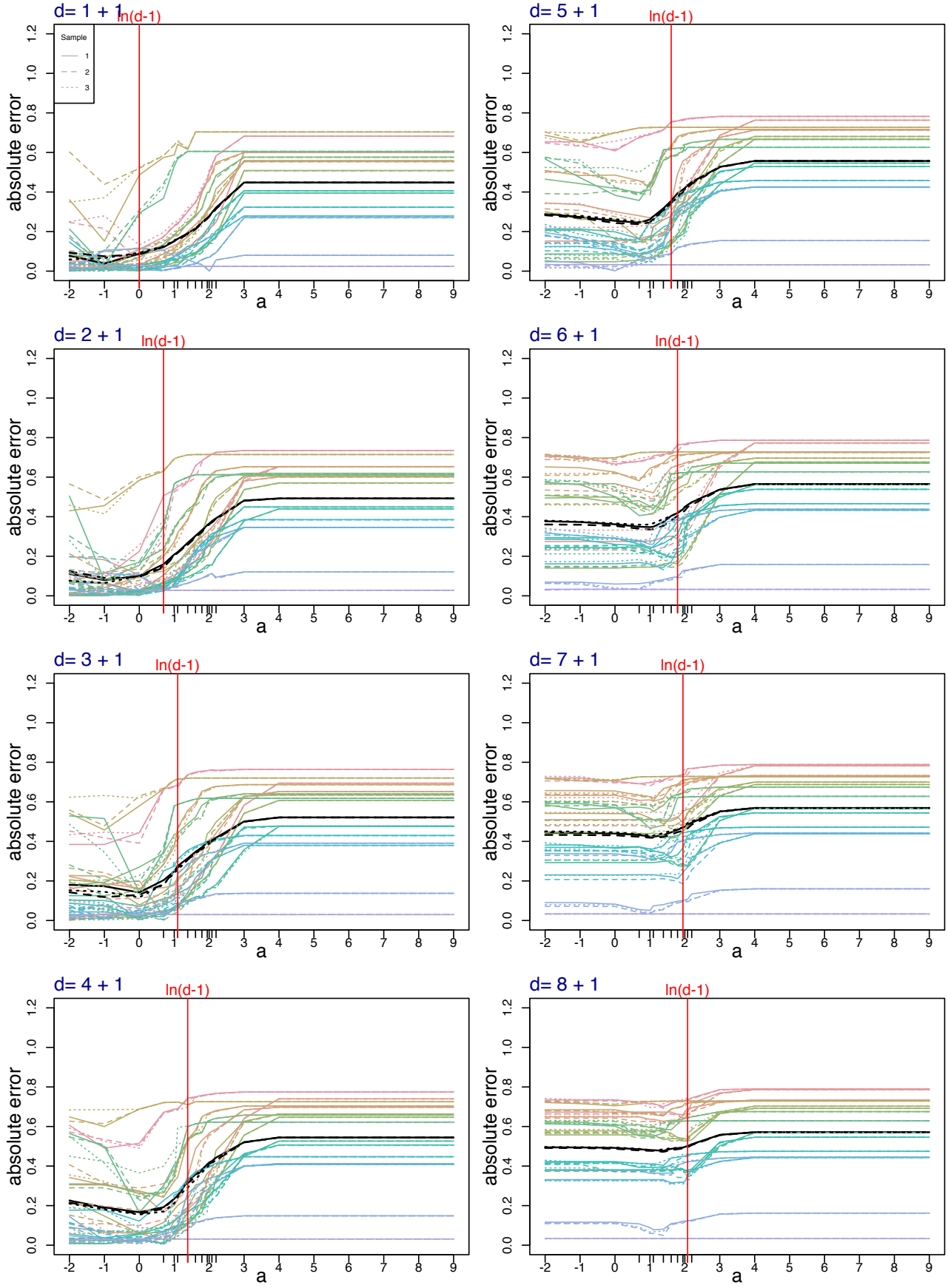


FIGURE IV.19 – Calibration de $a$ pour le Modèle $1$ pour la procédure RevDir sur des échantillons de taille $n = 100\,000$.

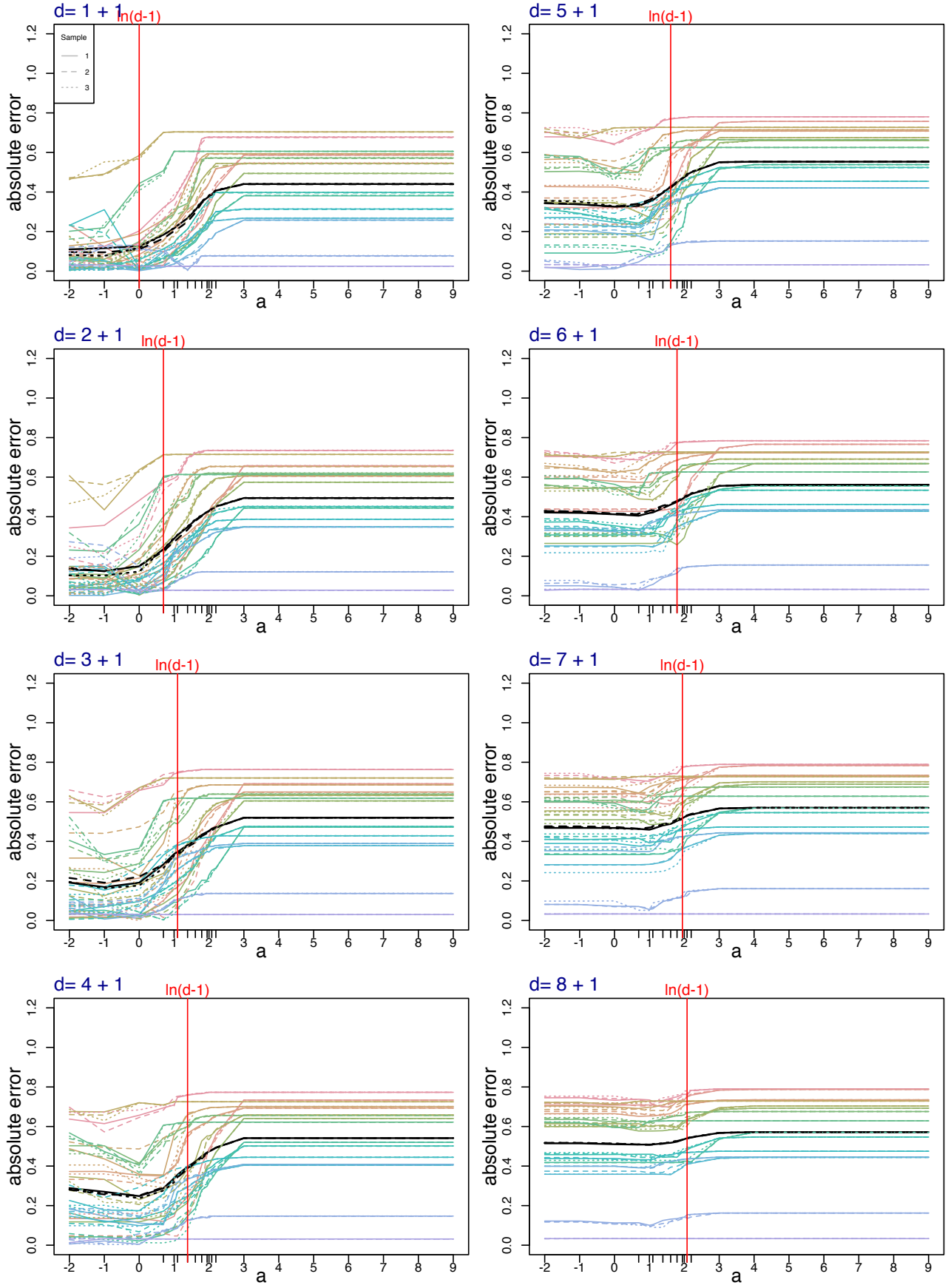
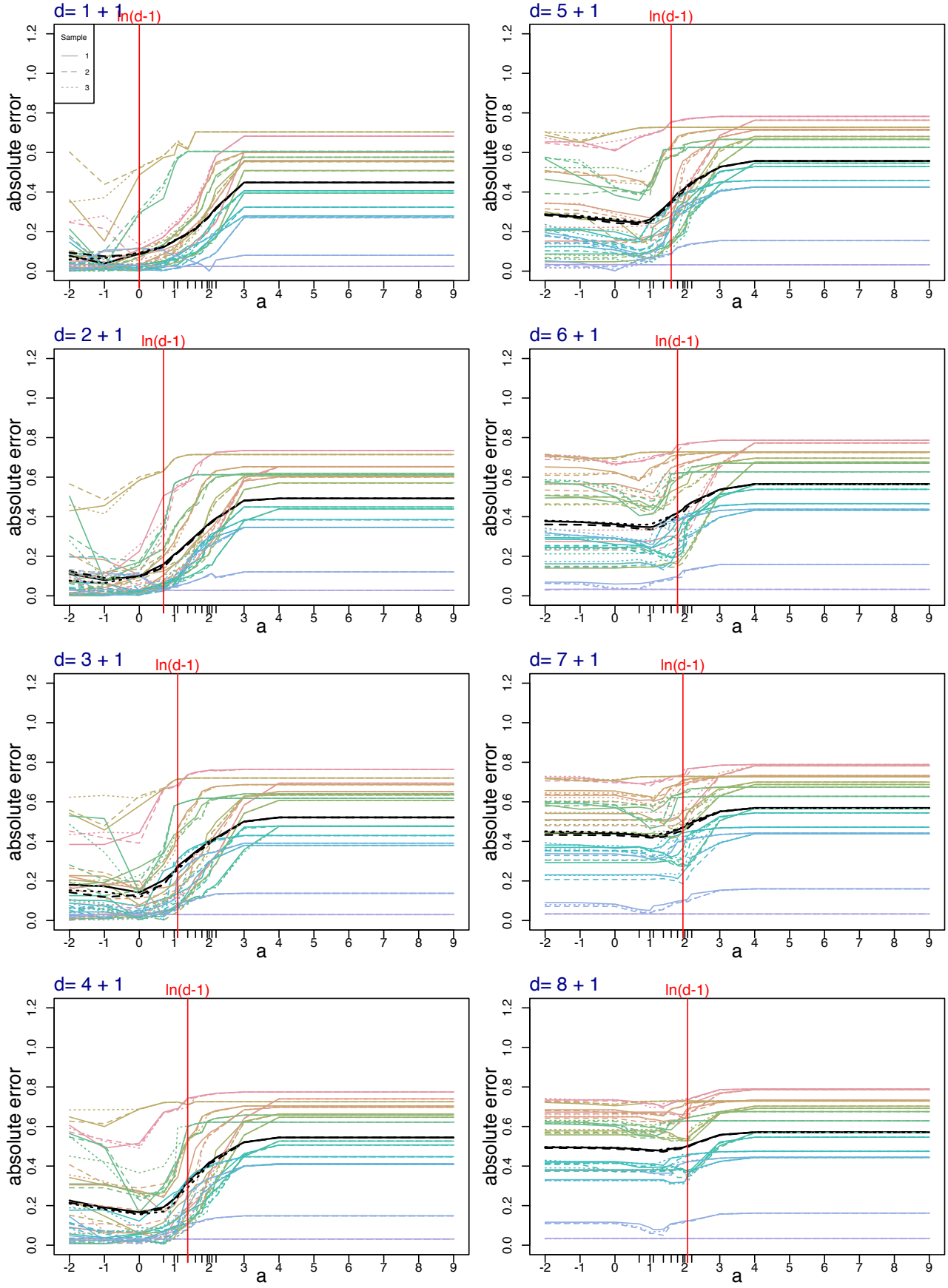


FIGURE IV.20 – Calibration de $a$ pour le Modèle 1 pour la procédure RevDir sur des échantillons de taille $n = 200\,000$.



**Modèle 2. Procédure Direct**

FIGURE IV.21 – Calibration de $a$ pour le Modèle 2 pour la procédure Direct sur des échantillons de taille $n = 10\,000$.

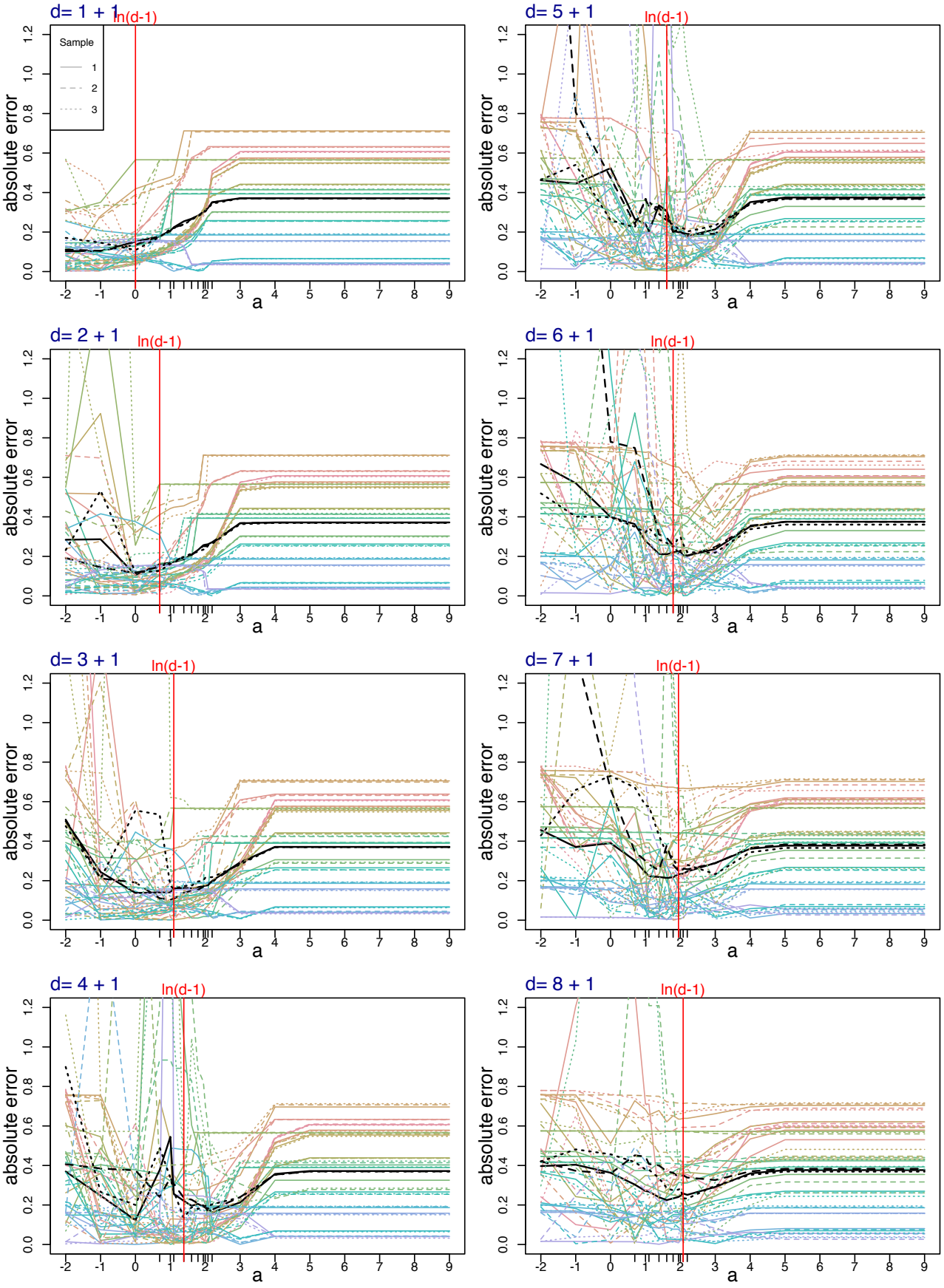


FIGURE IV.22 – Calibration de $a$ pour le Modèle 2 pour la procédure Direct sur des échantillons de taille $n = 25\,000$.

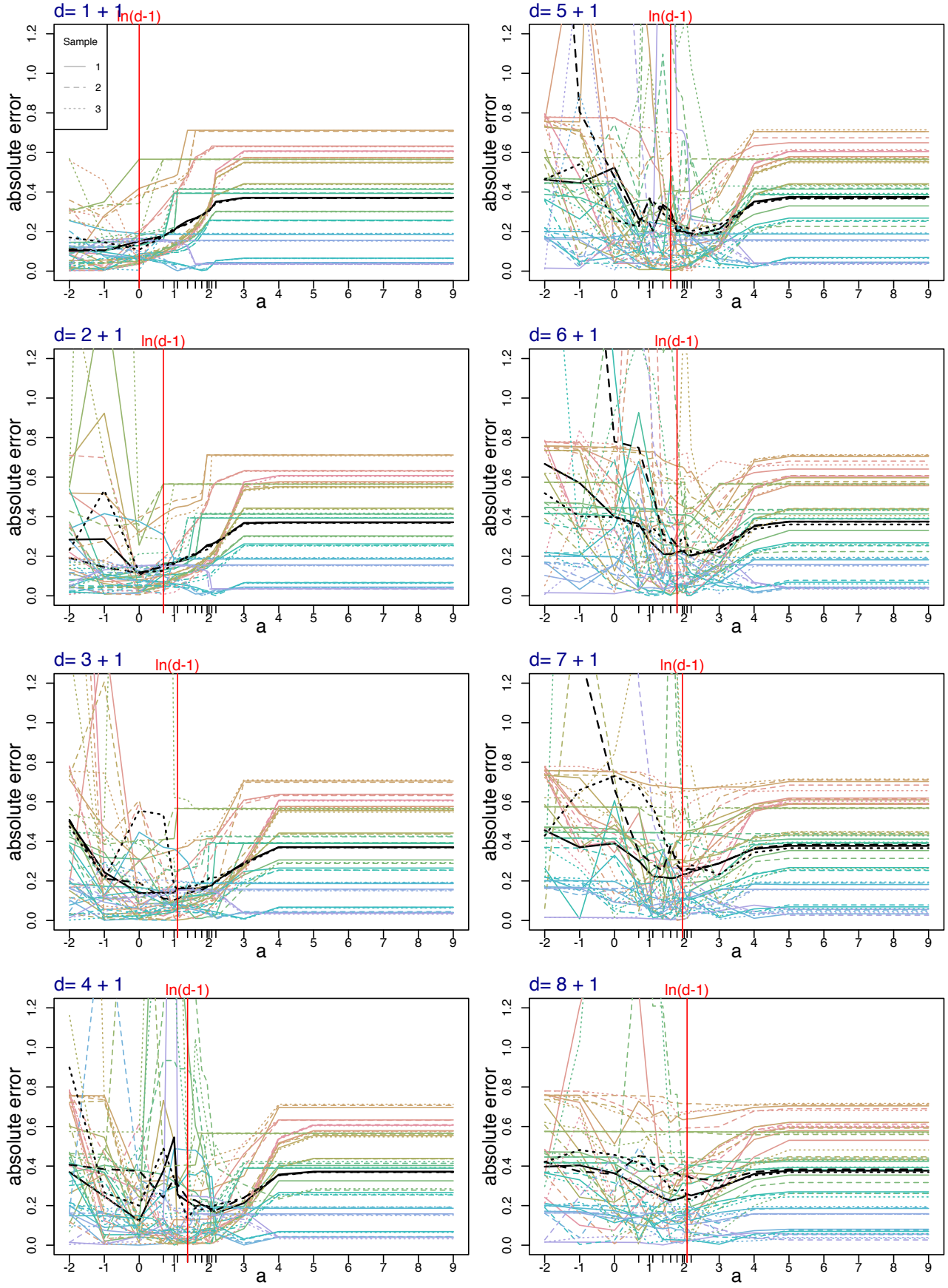


FIGURE IV.23 – Calibration de $a$ pour le Modèle 2 pour la procédure Direct sur des échantillons de taille $n = 50\,000$.

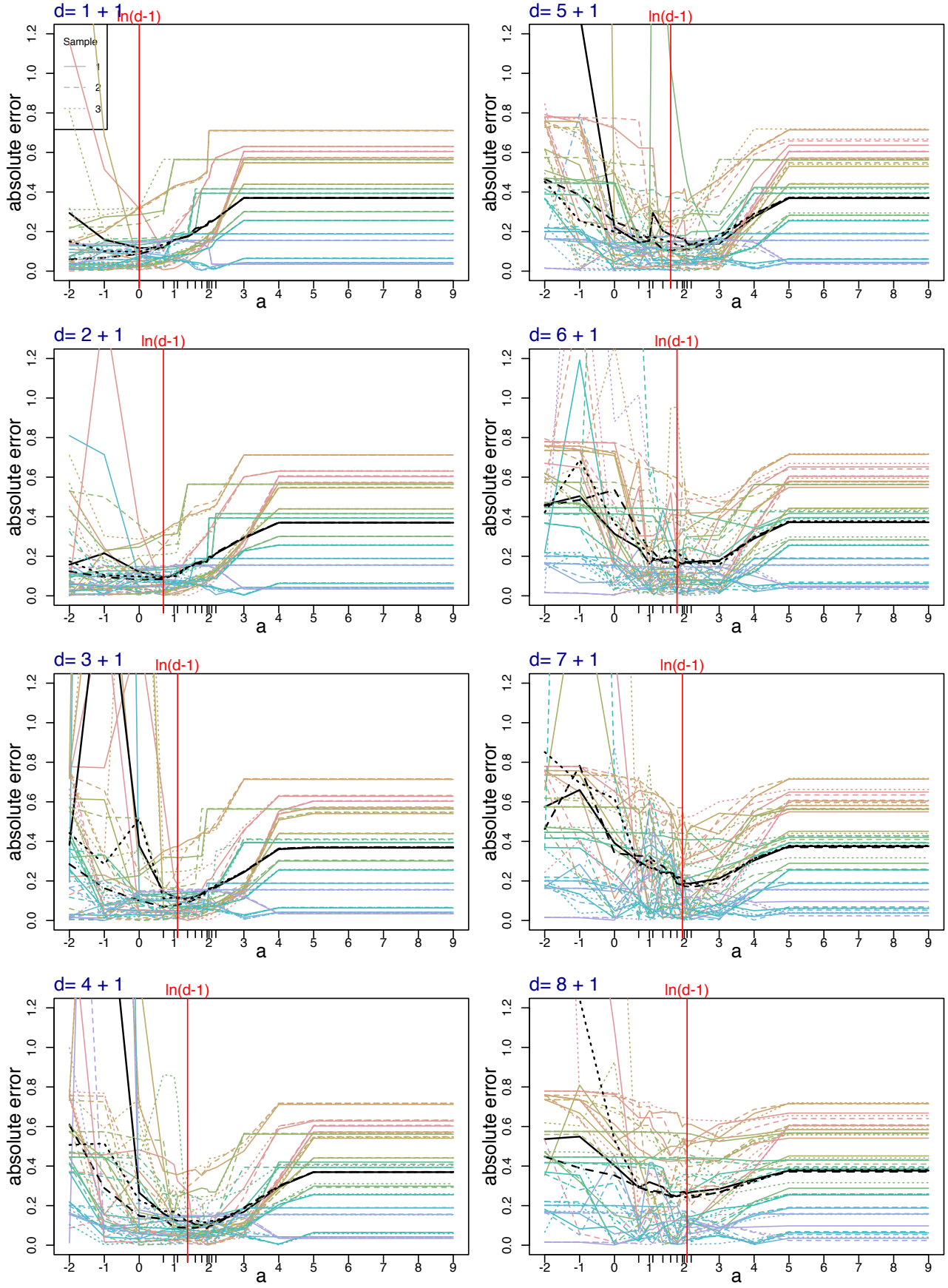


FIGURE IV.24 – Calibration de $a$ pour le Modèle 2 pour la procédure Direct sur des échantillons de taille $n = 100\,000$.

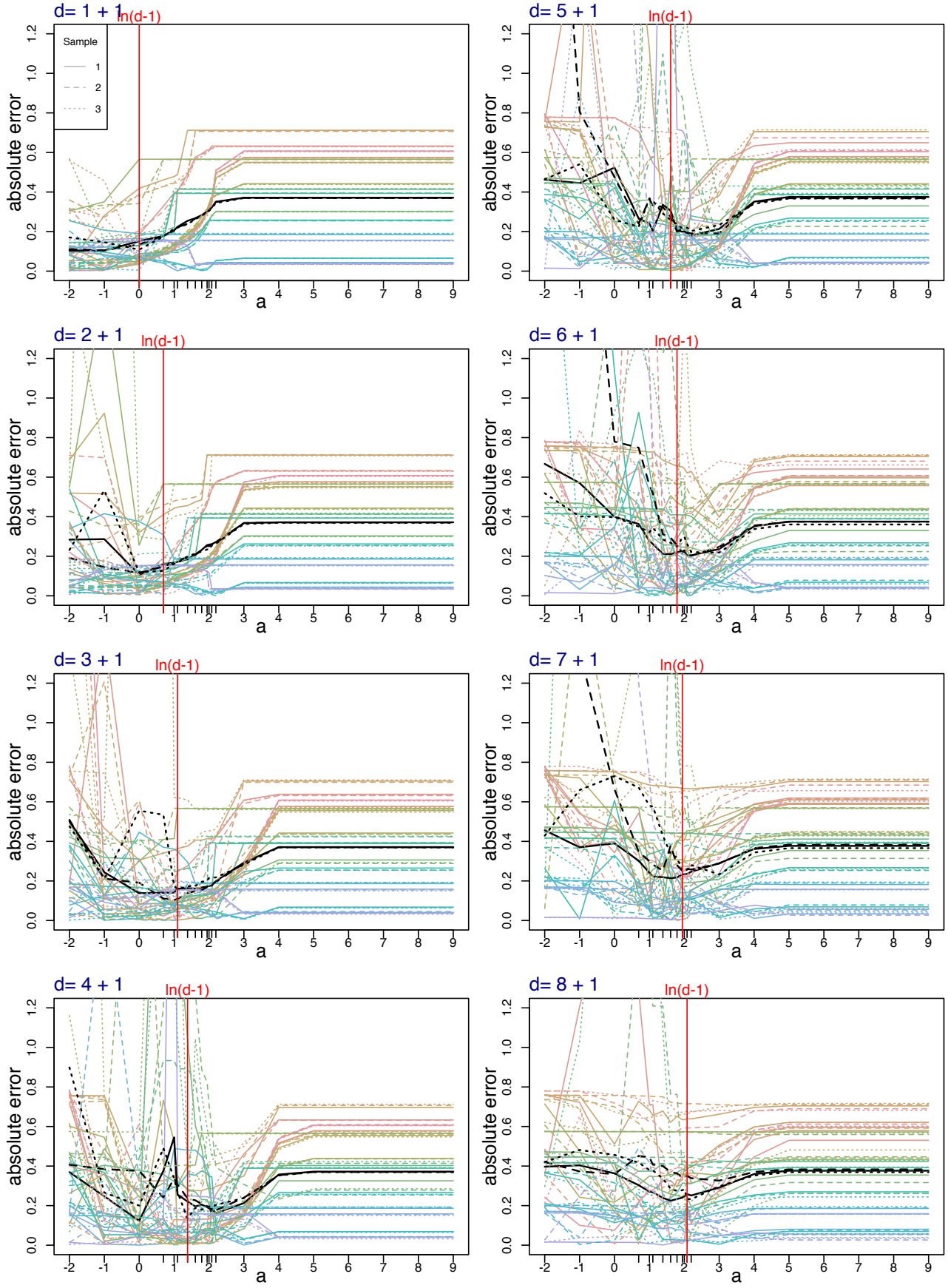


FIGURE IV.25 – Calibration de $a$ pour le Modèle 2 pour la procédure Direct sur des échantillons de taille $n = 200\,000$.

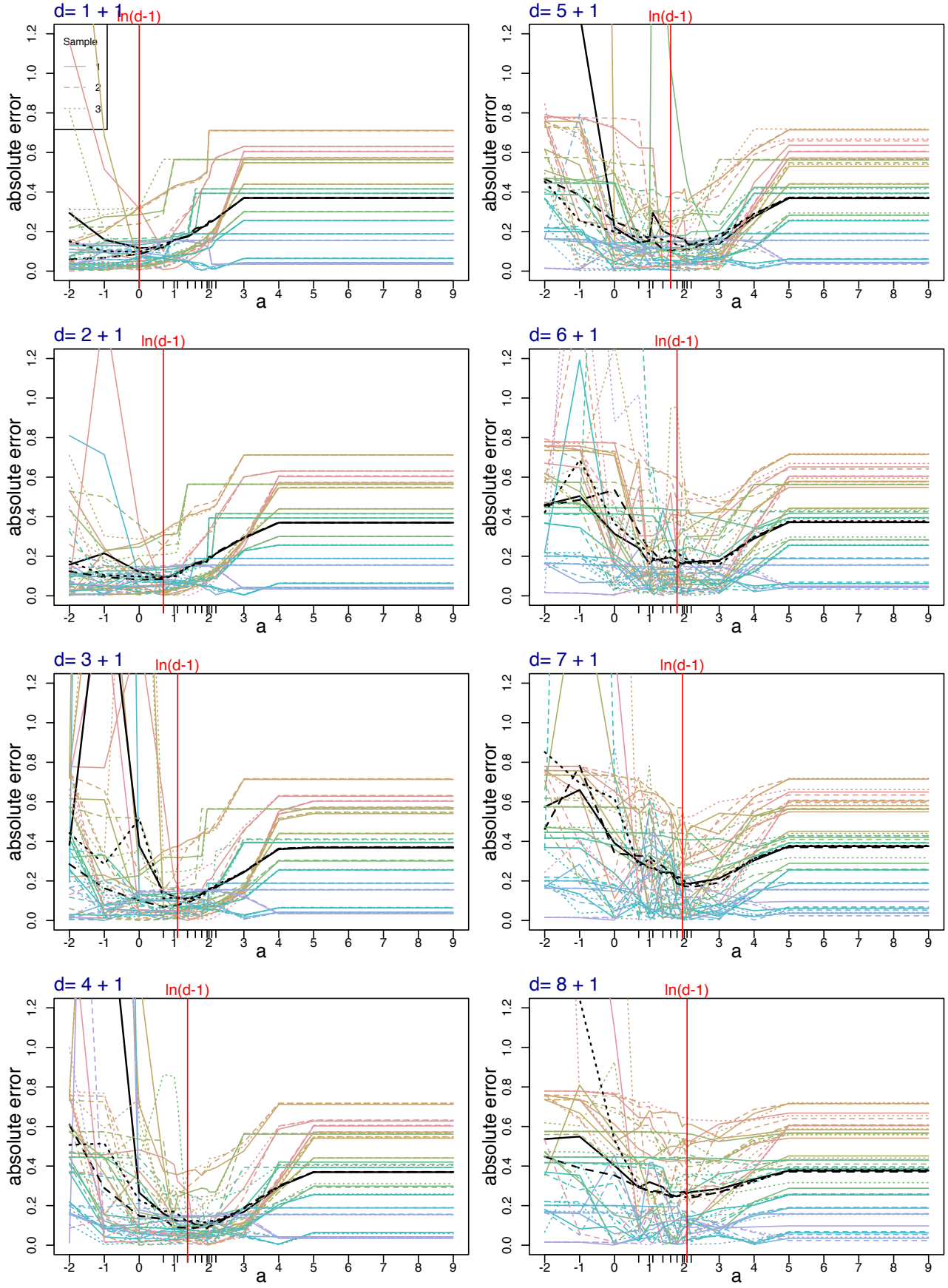


FIGURE IV.26 – Calibration de $a$ pour le Modèle 2 pour la procédure RevDir sur des échantillons de taille $n = 10\,000$.

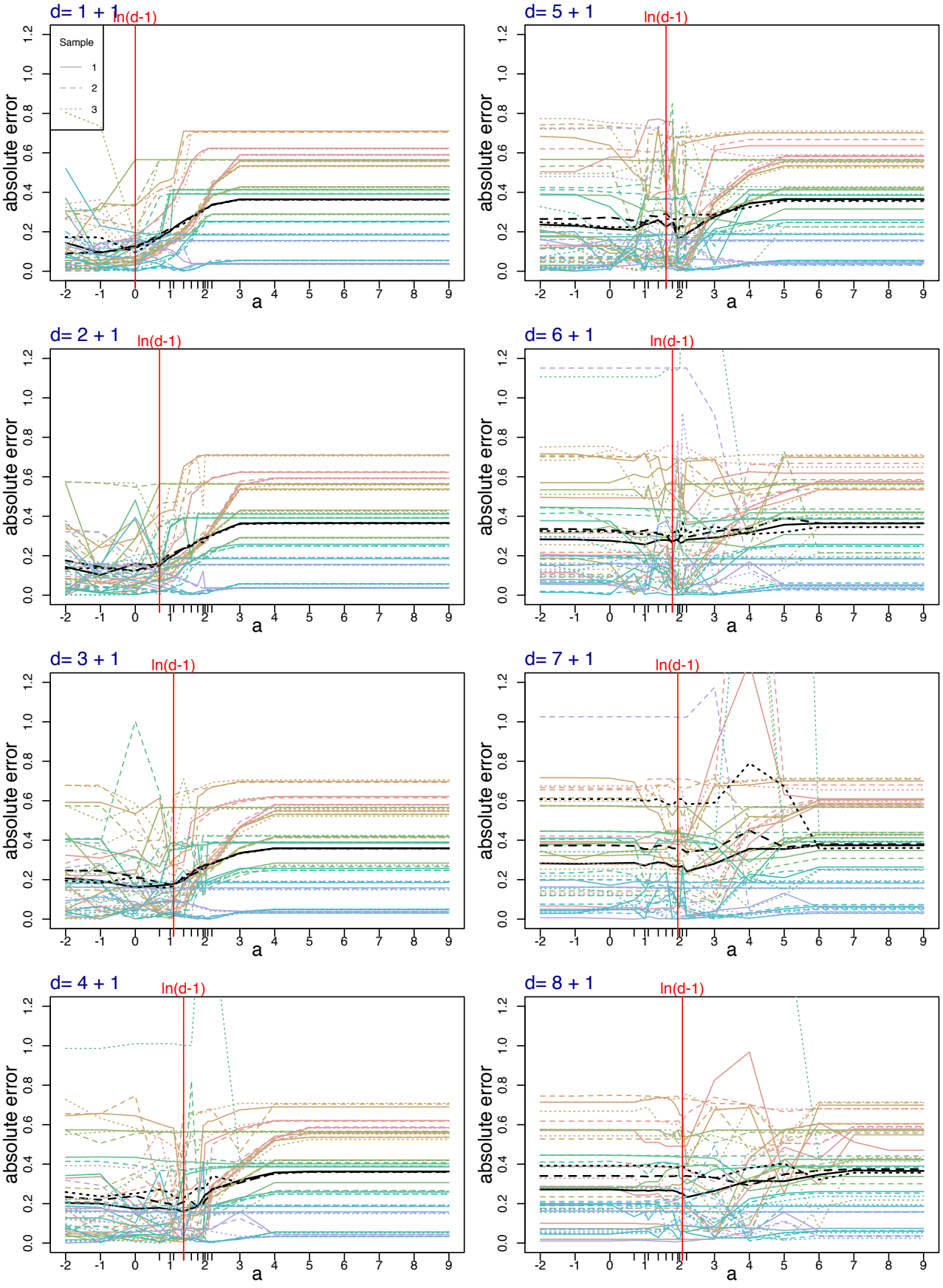


FIGURE IV.27 – Calibration de $a$ pour le Modèle 2 pour la procédure RevDir sur des échantillons de taille $n = 25\,000$.

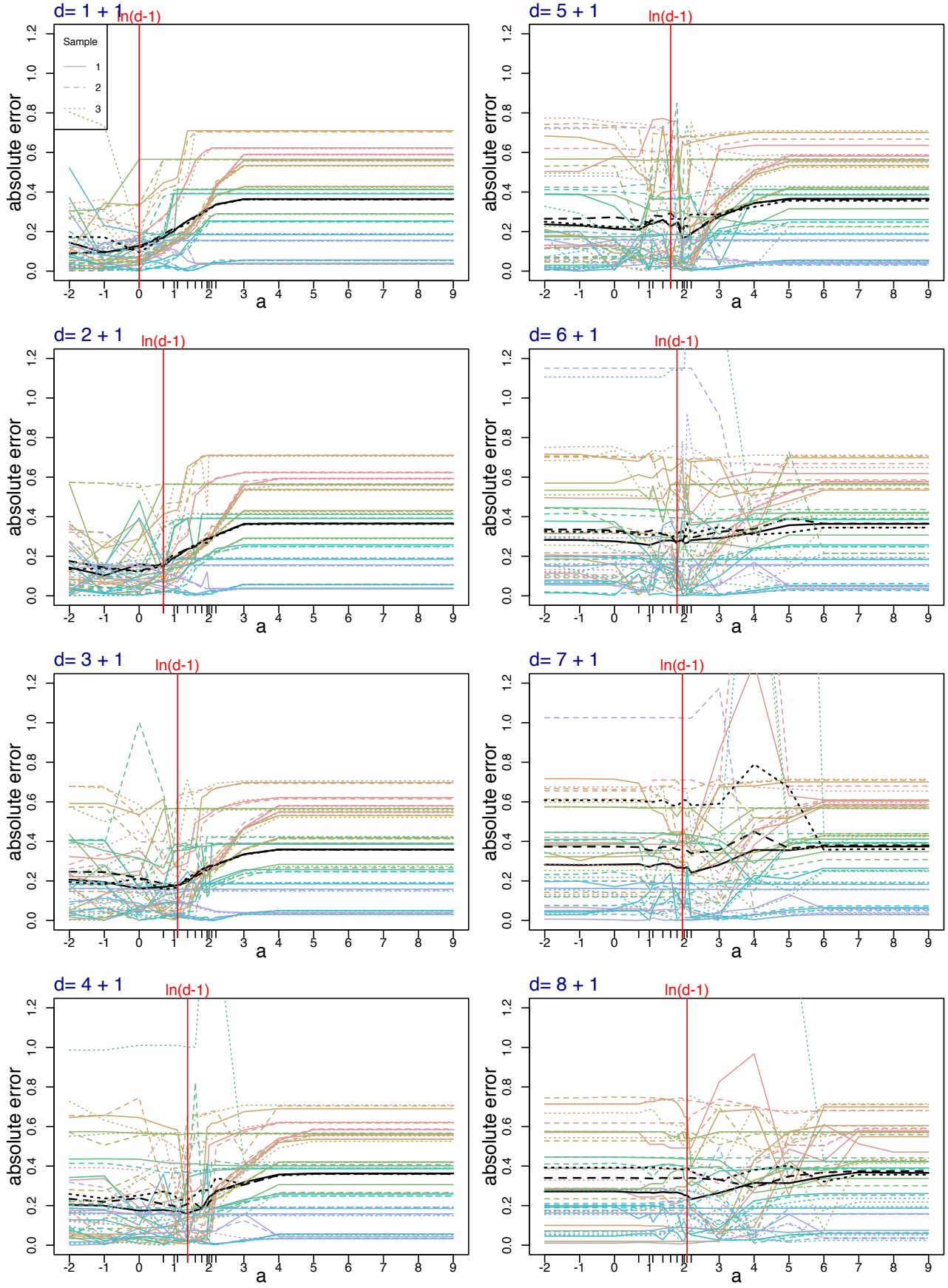


FIGURE IV.28 – Calibration de $a$ pour le Modèle 2 pour la procédure RevDir sur des échantillons de taille $n = 50\,000$.

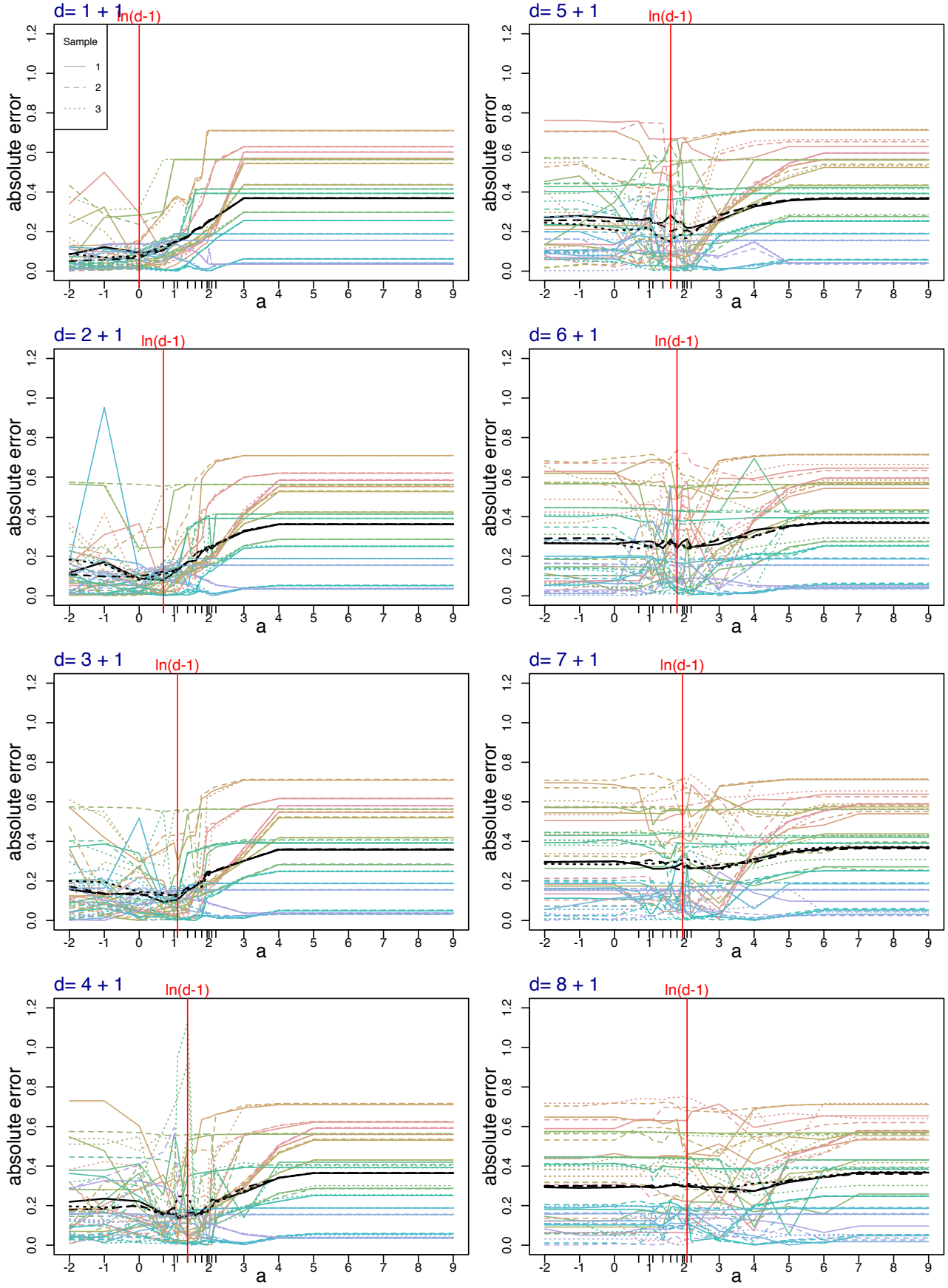


FIGURE IV.29 – Calibration de $a$ pour le Modèle 2 pour la procédure RevDir sur des échantillons de taille $n = 100\,000$.

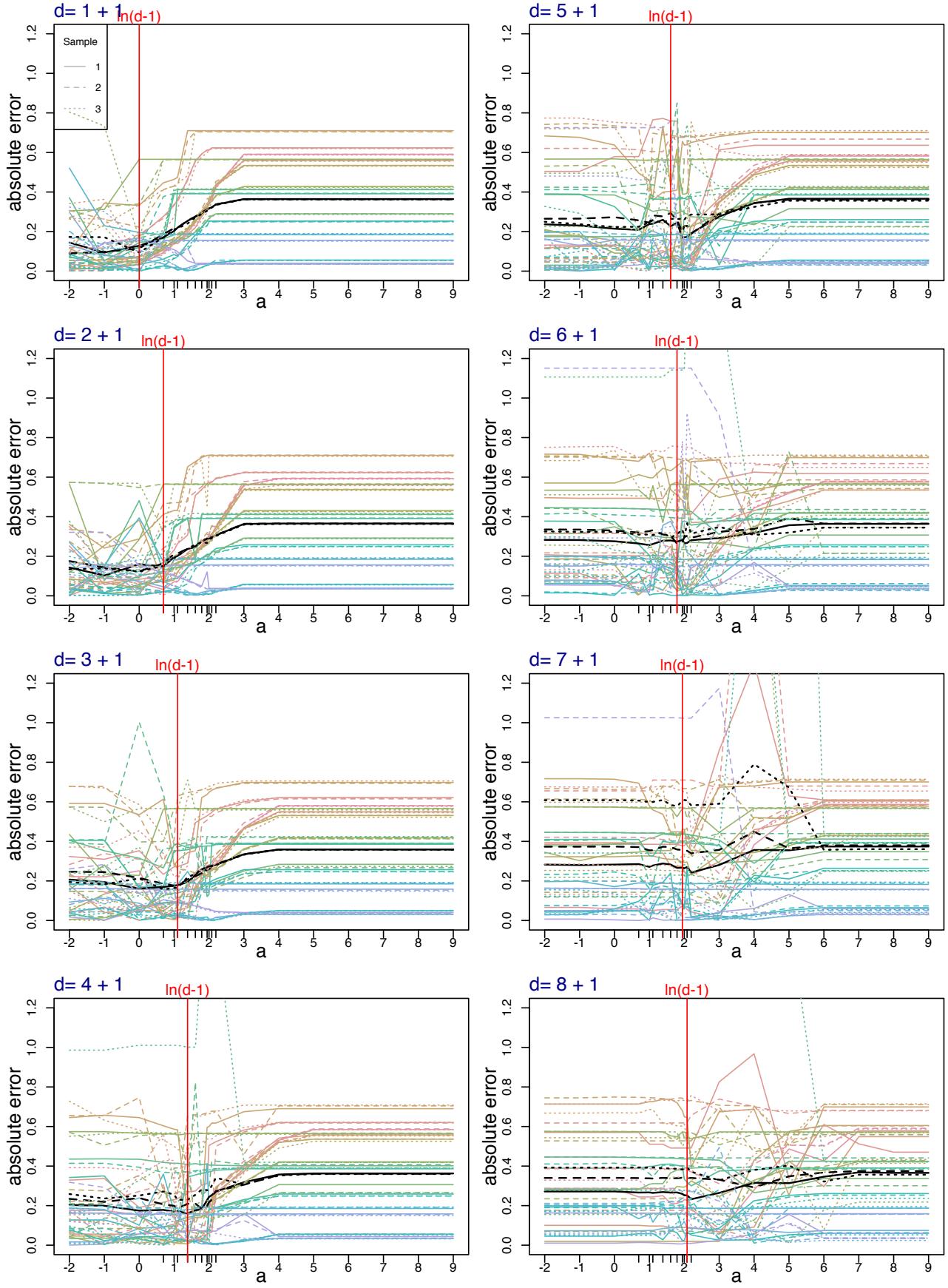
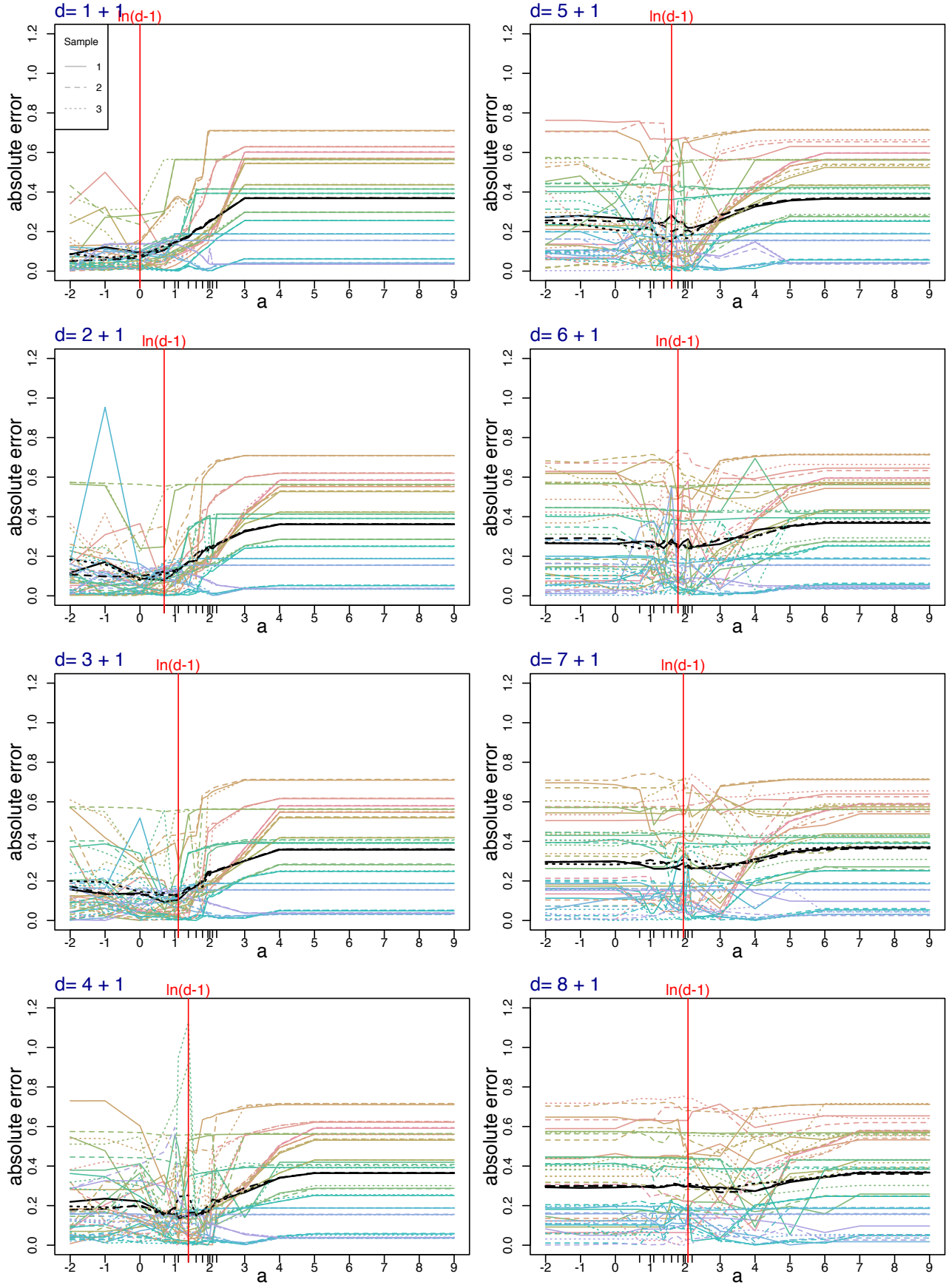


FIGURE IV.30 – Calibration de $a$ pour le Modèle 2 pour la procédure RevDir sur des échantillons de taille $n = 200\,000$.



**Modèle 3. Procédure Direct**

FIGURE IV.31 – Calibration de $a$ pour le Modèle 3 pour la procédure Direct sur des échantillons de taille $n = 10\,000$.

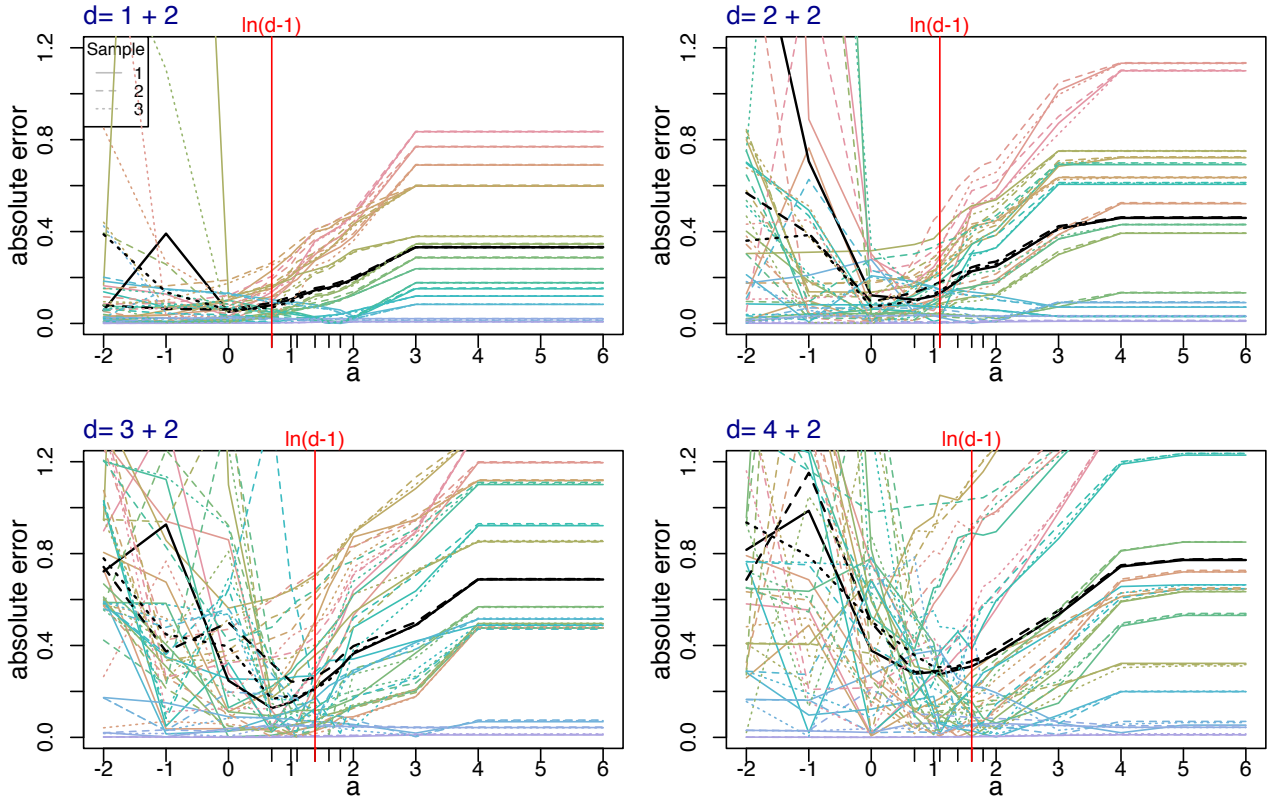



FIGURE IV.32 – Calibration de $a$ pour le Modèle 3 pour la procédure Direct sur des échantillons de taille $n = 25\,000$.

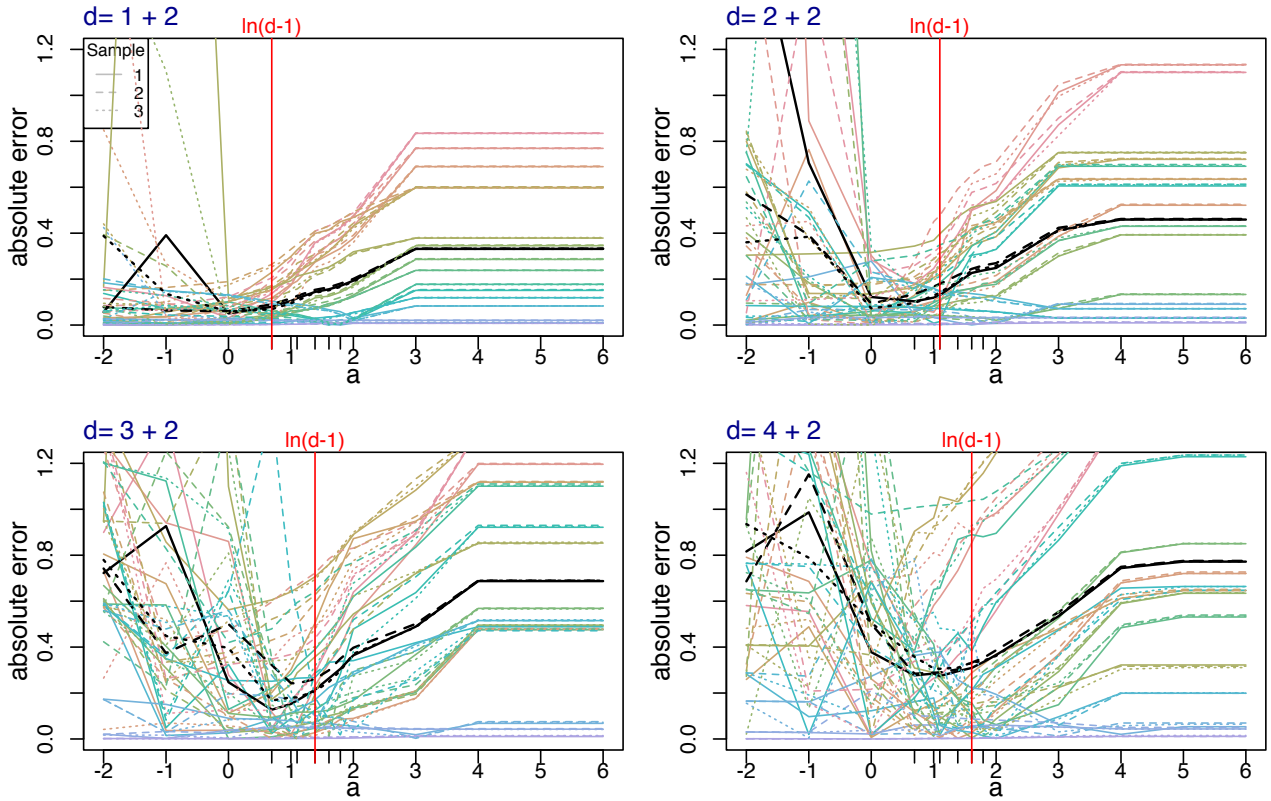


FIGURE IV.33 – Calibration de $a$ pour le Modèle 3 pour la procédure Direct sur des échantillons de taille $n = 50\,000$.

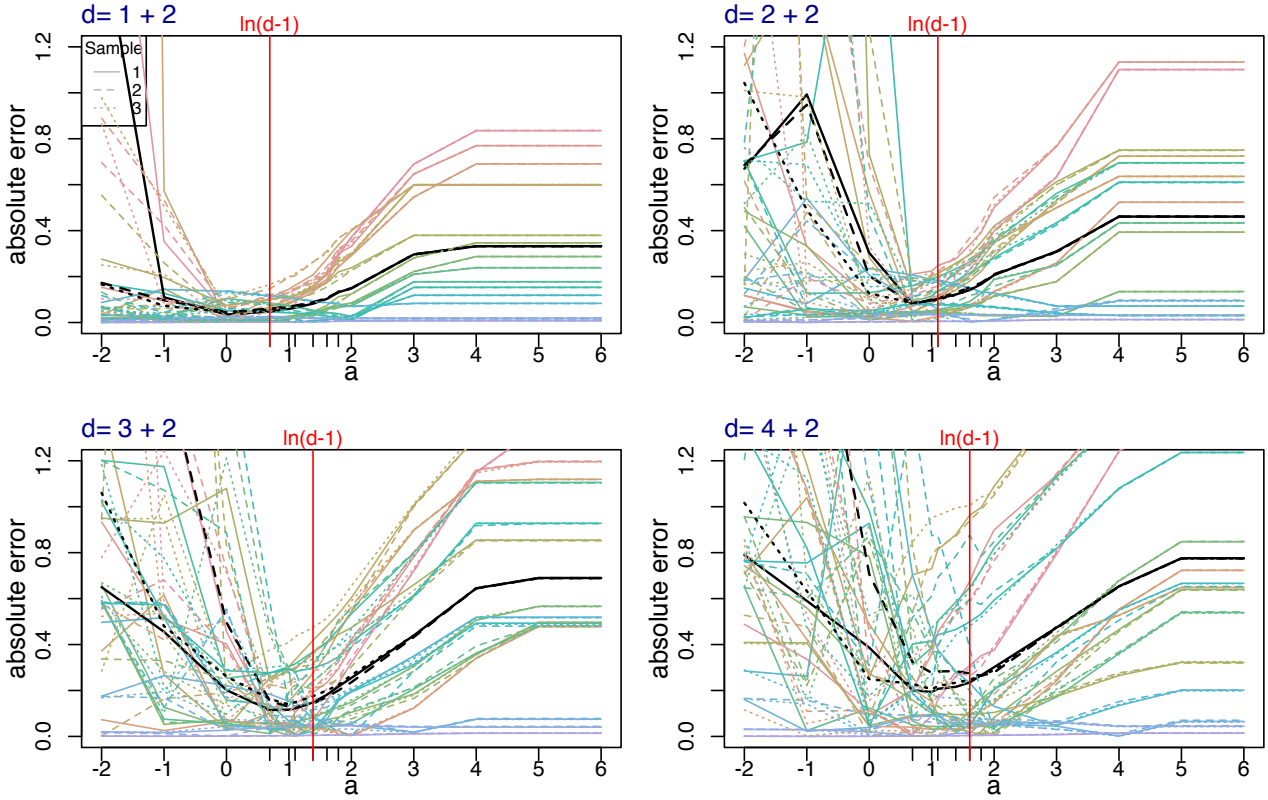



FIGURE IV.34 – Calibration de $a$ pour le Modèle 3 pour la procédure Direct sur des échantillons de taille $n = 100\,000$.

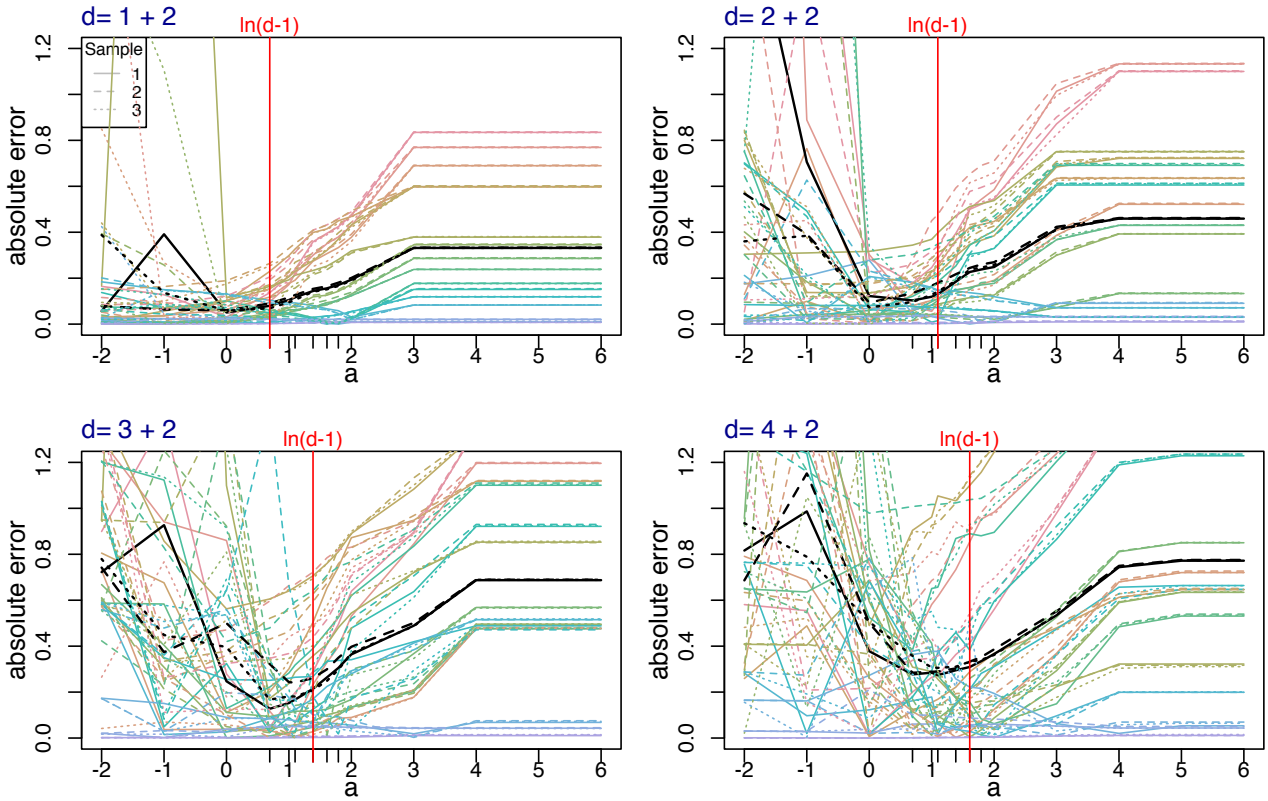
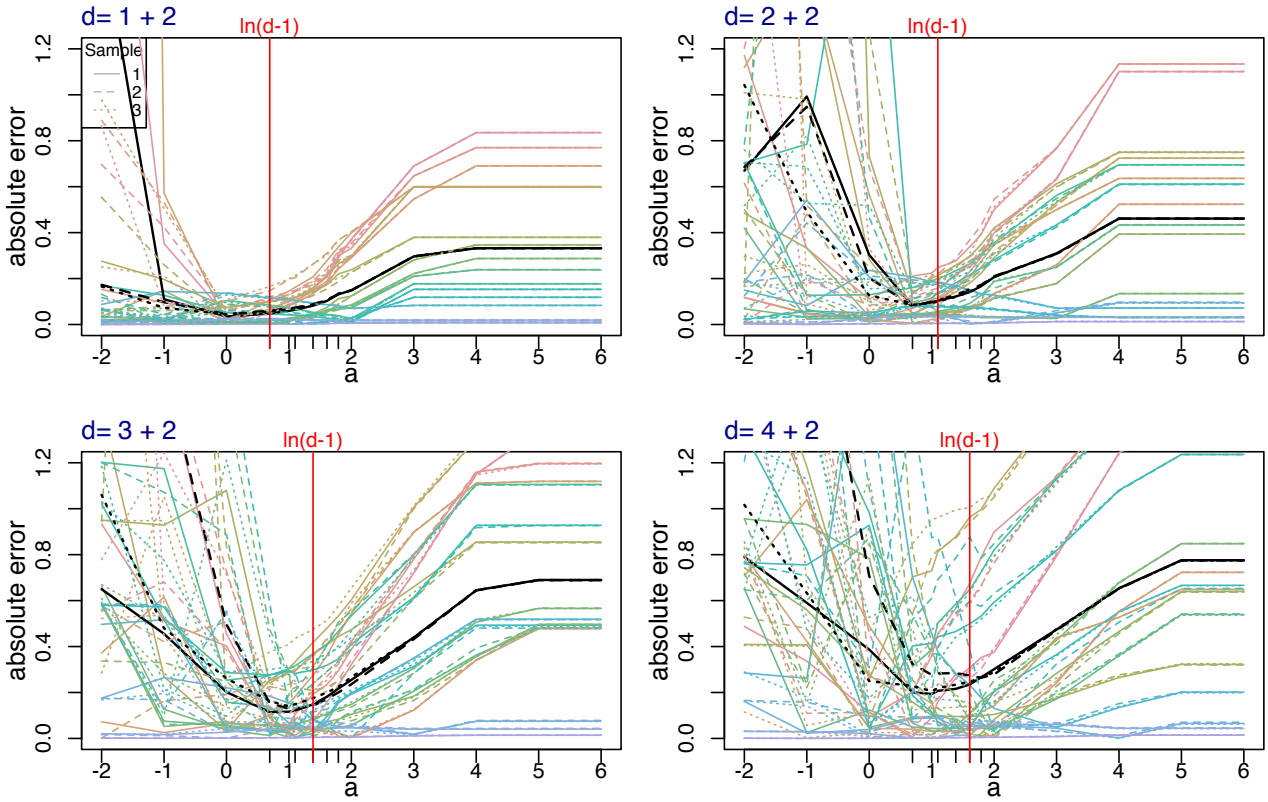


FIGURE IV.35 – Calibration de $a$ pour le Modèle 3 pour la procédure Direct sur des échantillons de taille $n = 200\,000$.




**Procédure Revdir**

FIGURE IV.36 – Calibration de $a$ pour le Modèle 3 pour la procédure RevDir sur des échantillons de taille $n = 10\,000$.

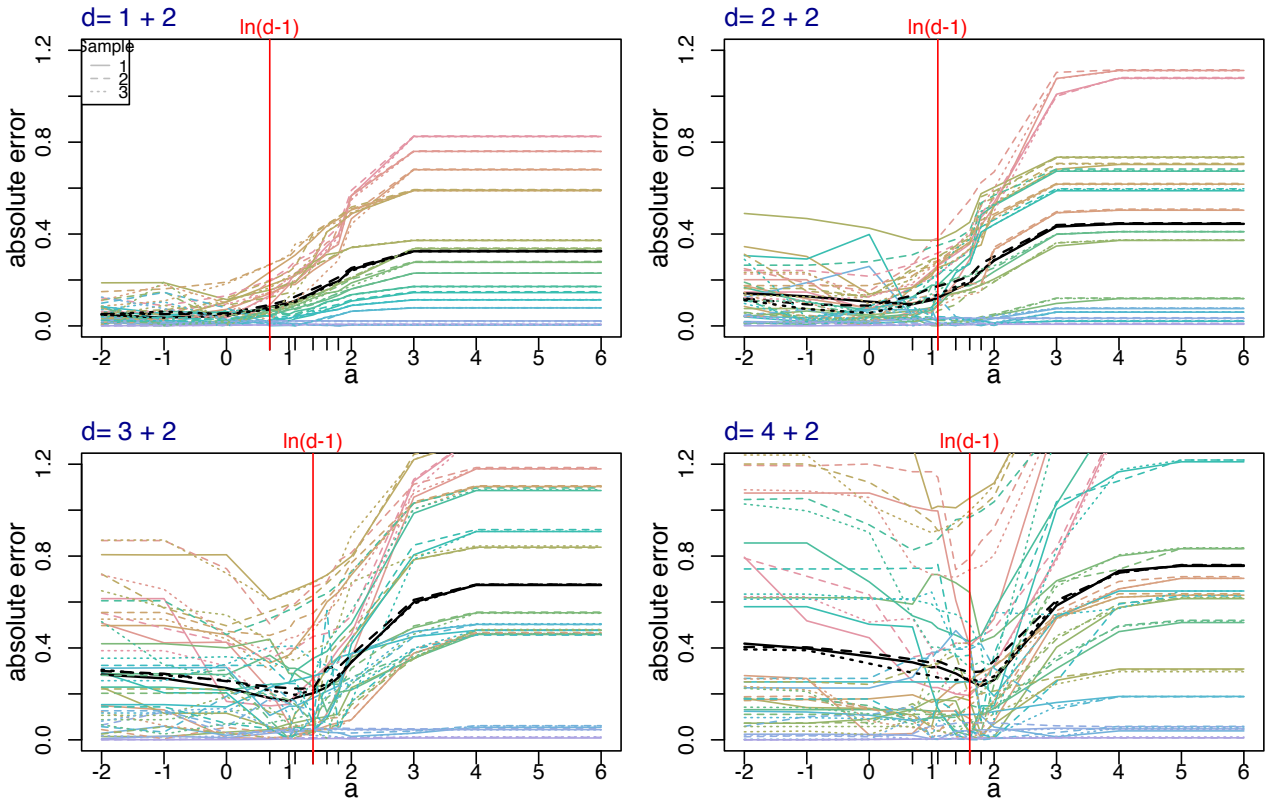


FIGURE IV.37 – Calibration de $a$ pour le Modèle 3 pour la procédure RevDir sur des échantillons de taille $n = 25\,000$.

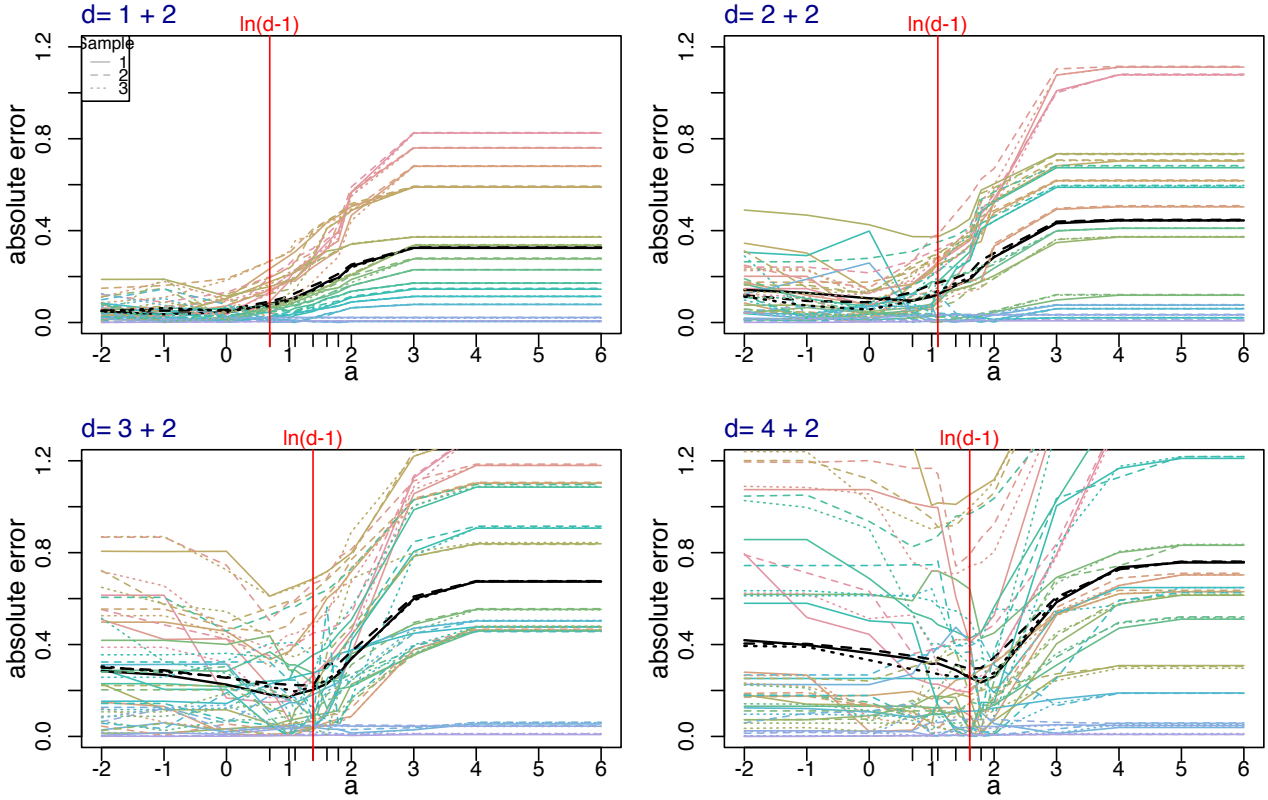



FIGURE IV.38 – Calibration de $a$ pour le modèle 3 pour la procédure RevDir sur des échantillons de taille $n = 50\,000$.

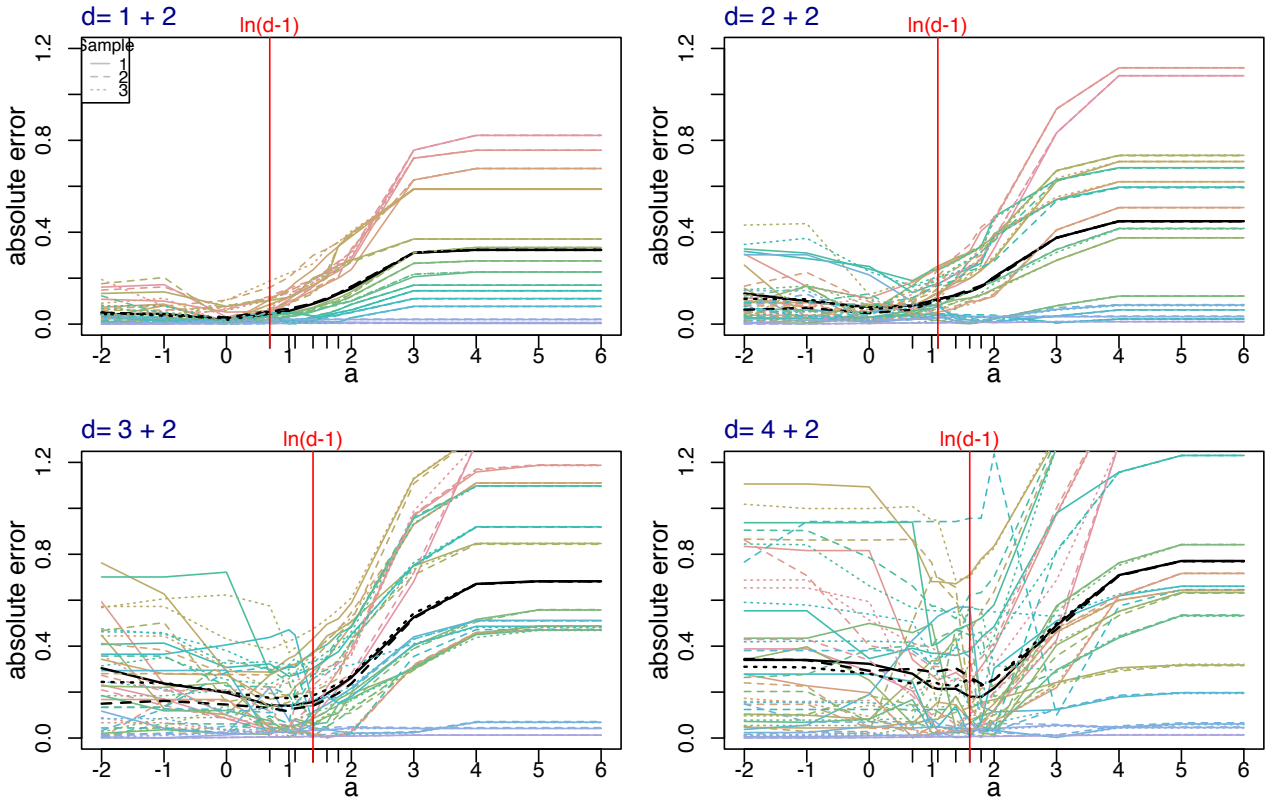
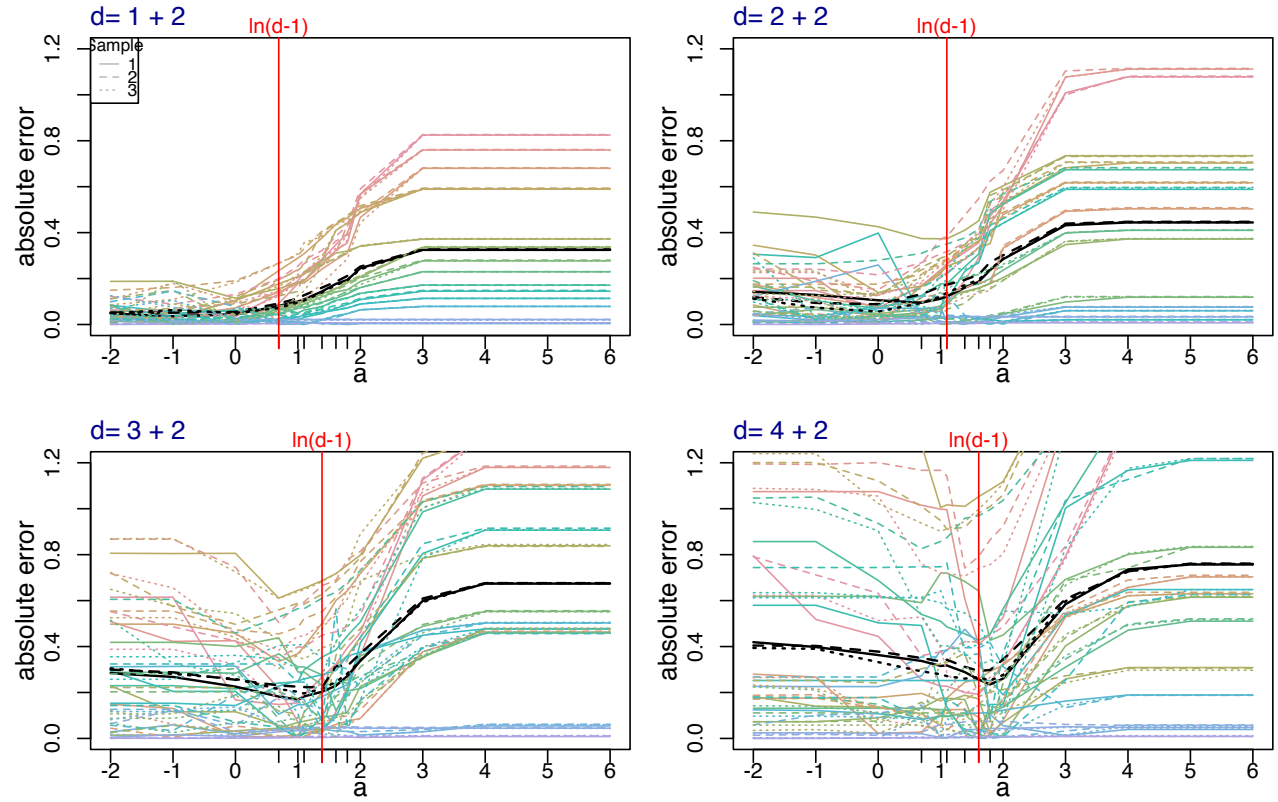
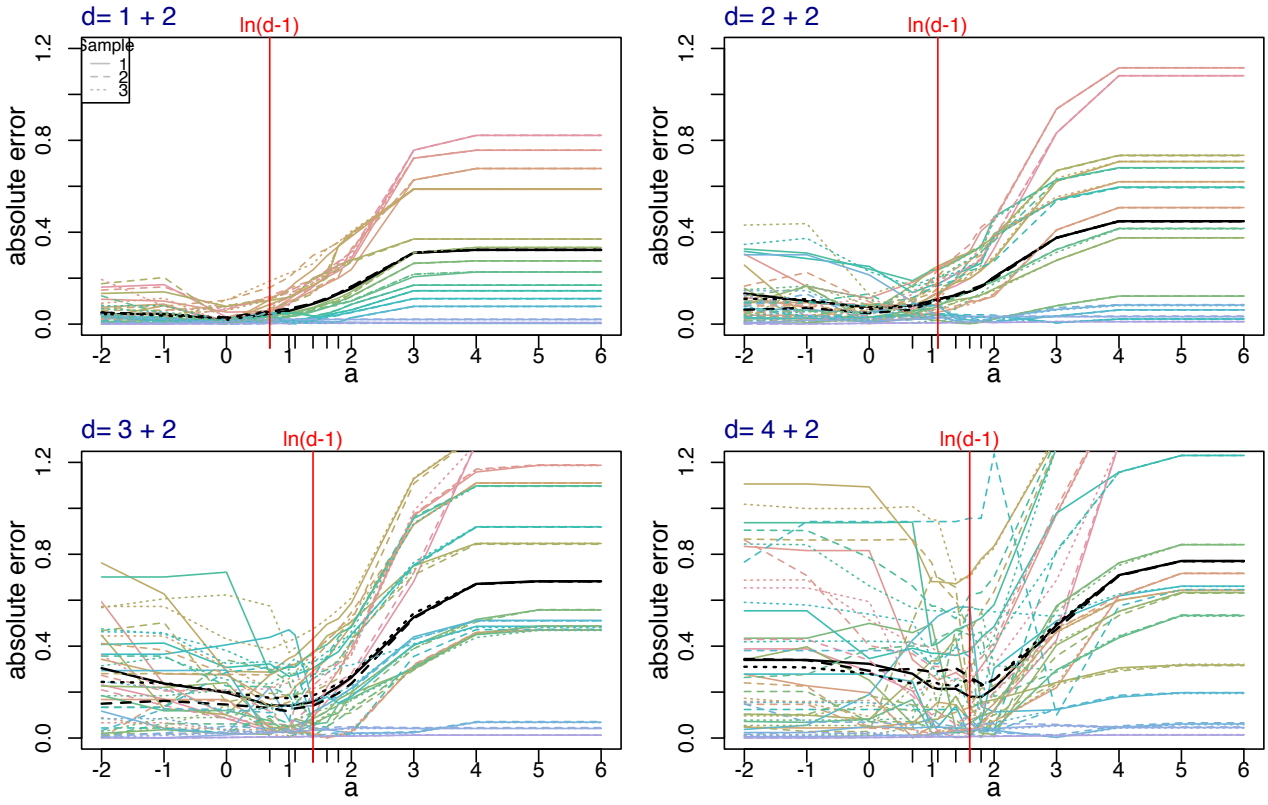


FIGURE IV.39 – Calibration de $a$ pour le Modèle 3 pour la procédure RevDir sur des échantillons de taille $n = 100\,000$.




s

FIGURE IV.40 – Calibration de $a$ pour le Modèle 3 pour la procédure RevDir sur des échantillons de taille $n = 200\,000$.



## 5.c Codes annotés des deux procédures

Les codes sont .R.

### Code de CDRodeo Direct.

```
#CDRODEO Direct

#Fixed tunings/notations:
#K: gaussian product kernel
#n: sample size
#d: joint dimension
#d1: dimension of the vector of the auxiliary variables
#d2: dimension of the vector of the variables of interest
#w: joint focus point
#W: joint data

#Input:
#Y: variables of interest. Matrix of size n.d2
#y: evaluation point (ponctual conditional density estimation). Vector of length d2
#x: focus point (density of Y conditionally to X=x). Vector of length d1.
#Default value=c() (for UNconditional density estimation)
#X: auxiliary variables. Matrix of size n.d1.
#Default value=c() (for UNconditional density estimation)
#fX: vector of {f_X(X_i)} (i=1,...,n) where f_X is the density of X (or an estimator).
#Default value=c() (for UNconditional density estimation)
#beta: decreaing factor of the bandwidth. Strictly include in (0,1).
#Default value=0.9
#a: calibration parameter (have to be >1). Default value=1.1
#h0: bandwidth initialization value. Have to be in |R+. Default value=2

#Output:
#h: bandwidth selected by CDRodeo


CDRodeoDir<-function(Y,y,x=c(),X=c(),fX=1,beta=0.9,a=1.1,h0=1){
  n=nrow(Y) #taille de l'échantillon
  d2=ncol(Y) # dimension de y
  if(length(X)==0){#Case of unconditional density estimation (no auxiliary variables)
    d1=0
    fX.est=1
    W=Y
    w=y
  }else{
    d1=ncol(X)
    w=c(x,y)
    W=cbind(X,Y)
    fX.est=fX
  }
```

```
  d=d1+d2
  h=rep(h0,d) #bandwidth initialisation
  compt=0 #Iteration counter
  Clambda2=3/2^(d-2)/pi^(d/2) #threshold squared constant
  renormalizedlambda2=Clambda2*(log(n))^(a)
  #renormalized squared threshold (by *n prod(h)*hj^2)
  actives=rep(TRUE,d) #initial activation of all components
  prodKinactives=rep(1,n)
  z=t((w-t(W))/h) #matrix of size n.|actives|
  renormalizedKij=t(t(dnorm(z))/sqrt(h[actives]))# renormalized by sqrt(h[actives])

  while(prod(h)>log(n)^a/n && sum(actives)>0){
    A=actives #save actives before updating
    renormalizedprodK=prodKinactives*(apply(matrix(renormalizedKij[,A],n),1,prod))
    deriv=-1+z^2
    renormalizedZ2=apply(deriv*(renormalizedprodK/fX.est),2,mean)^2*n
    #vector of the Zj's renormalized by *n prod(h)*hj^2

    #Updates
    #Deactivation of components below the threshold
    actives[A]=renormalizedZ2>renormalizedlambda2
    deactives=which(A!=actives) #Deactivated components
    if(length(deactives)!=0){
      prodKdeactives=apply(matrix(renormalizedKij[,deactives],n),1,prod)
      prodKinactives=prodKinactives*prodKdeactives
    }
    if(sum(actives)>0){
      h[actives]=beta*h[actives]
      z=matrix(z[,actives[A]],n)/beta
      renormalizedKij[,actives]=t(t(dnorm(z))/sqrt(h[actives]))
      compt<-compt+1
    }
  }
  return(h)
}
```

### Code de CDRodeo RevDir.

```
#CDRodeo Gaussien en 2 étapes.
#Entrées :
#beta: facteur de multiplication < 1 de décroissance de h
#X: MATRICE nxd1 des variables explicatives
#Y: MATRICE nxd2 des variables d'intérêt
#(x,y): point d'intérêt où l'on estime f
#fX: vecteur des {f_X(X_i)}  (i=1,...,n)
#a: paramètre du seuil lambda (puissance du terme log n)
#(doit être >1)
```

```
#Sortie : h, la fenêtre sélectionnée par Rodeo

RevDirCDRodeoGauss<-function(beta=0.95,X,Y,x,y,fX,a=1.1,h.max=1){
  n=nrow(Y) #taille de l'échantillon
  d2=ncol(Y) # dimension de y
  if(length(X)==0){
  #Case of unconditional density estimation (no auxiliary variables)
    d1=0
    fX.est=rep(1,n)
    W=Y
    w=y
  }else{
    d1=ncol(X)
    w=c(x,y)
    W=cbind(X,Y)
    fX.est=fX
  }
  d=d1+d2 # dimension totale
  W=cbind(X,Y) #concaténation de X et Y (matrice nxd)
  w=c(x,y) #concaténation du point d'intérêt
  Ibeta=beta^{-1}#inverse de beta
  chigauss=3/2^(d-2)/pi^(d/2)
  #=4*Clambda^2 où Clambda la constante du seuil pour le noyau gaussien
  lambda2simpl=chigauss*(log(n))^(a)#/min(fX)
  #= lambda_h^2 *n prod(h)*hj^2
  #seuil adapté pour le noyau gaussien
  h0=(chigauss/4*log(n)^a/n)^(1/(d*5))#(1/n)^(1/(4+d))
  h=rep(h0,d)#initialisation de la fenêtre
  actifs=rep(TRUE,d)
  prodKinactif=rep(1,n)
  #matrice de booléens indiquant les comp. actives en fonction
  #de h^(compt) (ligne) et la comp (colonne)
  desactives=c()#indices des composantes désactivées à l'itération courante
  z=t((w[actifs]-t(matrix(W[,actifs],n)))/h[actifs])
  #point d'application du noyau K pour les fenêtres Pertinentes (Ici, toutes)
  # matrice de taille n x card(actifs)
  Kij=t(t(dnorm(z))/sqrt(h)) # K évalué en z_{i,j} divisé par sqrt(h_j)
  #rmq Ici tout le monde est actif
  #l'initialisation de KIj se fait ici (à modifier prudemment)
  #(pour simplifier le rapport Z2simpl sur lambda2)

  #Iteration 0 : comportement séparé pour préparer l'étape Dir
  A=actifs   # A: sauvegarde des comp. actives au début de l'itération
  prodK=prodKinactif*(apply(matrix(Kij[,A],n),1,prod))
  #calcul du noyau produit
  # vecteur nx1 des produits selon j des Kij sur le produit des sqrt(h_k)
  deriv=-1+z^2
  #facteur en plus par dérivation (uniquement pour K gaussien standart)
```

```
Zsimpl=as.vector((t(deriv*prodK))%*%(1/fX.est))
#vecteur des {Zj : j ACTIF} simplifié pour la comparaison avec le seuil
Z2simpl=#apply(deriv*(prodK/fX.est),2,mean)^2*n
  Zsimpl^2/n #=Z^2 *n* prod(h)*hj^2
#lambda2simpl=chigauss*(log(n))^(a)#/min(fX)
#= lambda_h^2 *n prod(h)*hj^2
#seuil adapté pour le noyau gaussien

#actualisations
#désactivation des composantes inférieures au seuil
actifs[A]=Z2simpl<lambda2simpl
#fenêtres désactivées à cette itération pour la mise à jour de prodK
desactives0=which(A!=actifs)
prodKactifDir=rep(1,n) #sauvegarde pour Dir
if(length(desactives0)!=0){
# mise à jour des produits noyaux desactivé et inactif si des
#composantes se sont désactivées
  prodKdesactives=apply(matrix(Kij[,desactives0],n),1,prod)
  prodKactifDir=prodKinactif*prodKdesactives #sauvegarde pour Dir
}

if(sum(actifs)>0){
#mise à jour pour l'itération suivante s'il reste des composantes actives
#mise à jour de la fenêtre
  h[actifs]<-Ibeta*h[actifs]
  #mise à jour du point d'application de K (juste les composantes ACTIVES)
  z=matrix(z[,actifs[A]],n)/Ibeta
  Kij[,actifs]=t(t(dnorm(z))/sqrt(h[actifs]))#mise à jour de Kij
}

#Iterations>0 du Rev
while(max(h)< beta*h.max & sum(actifs)>0){
  A=actifs   # A: sauvegarde des comp. actives en début d'itérations
  prodK=prodKactifDir*prodKinactif*(apply(matrix(Kij[,A],n),1,prod))
  #calcul du produit de noyaux
  # vecteur nx1 des produits selon j des Kij sur le produit des sqrt(h_k)
  #prodKactifDir désigne les facteurs associées aux composantes désactivées
  deriv=-1+z^2
  #facteur en plus par dérivation (uniquement pour K gaussien standart)
  Zsimpl=as.vector((t(deriv*prodK))%*%(1/fX.est))
  #vecteur des {Zj : j ACTIF} simplifié pour la comparaison avec le seuil
  Z2simpl=Zsimpl^2/n #=Z^2 *n prod(h)*hj^2
  #lambda2simpl=chigauss*(log(n))^(a)
  #= 4 lambda_h^2 *n prod(h)*hj^2
  #seuil adapté pour le noyau gaussien

  #actualisations du Rev
  #désactivation des composantes inférieures au seuil
```

```
  actifs[A]=Z2simpl<lambda2simpl
   # mise à jour des fenêtres désactivées à cette itération
  desactives=which(A!=actifs)
  if(length(desactives)!=0){
  # mise à jour du noyau desactivé et inactifs si des comp. se sont désactivées
    prodKdesactives=apply(matrix(Kij[,desactives],n),1,prod)
    prodKinactif=prodKinactif*prodKdesactives
  }

  if(sum(actifs)>0){
  #mise à jour pour l'itération suivante s'il reste des composantes actives
   #mise à jour de la fenêtre (croissance pour reverse)
    h[actifs]<-Ibeta*h[actifs]
    #mise à jour du point d'application de K (juste les composantes ACTIVES)
    z=matrix(z[,(Z2simpl<lambda2simpl)],n)/Ibeta
    Kij[,actifs]=t(t(dnorm(z))/sqrt(h[actifs]))#mise à jour de Kij
  }

}#fin de la boucle while

#Direct
#la fenêtre initiale de Dir est la fenêtre de sortie de Rev,
#mais seules les composantes désactivées à la 1ère itération de Rev sont actives.
#à la fin de rev : actifs= F. Il suffit d'activer les bonnes composantes.
actifs[desactives0]=TRUE
#la mise à jour de prodKinactifs a déjà été prise en
#compte pendant le Rev (avec prodKactifDir):
#la valeur de prodKinactif est la bonne.
#mises à jour de z et Kij pour la fenêtre de sortie de Rev pour
#les composantes de desactives0
z=t((w[actifs]-t(matrix(W[,actifs],n)))/h[actifs])
#point d'application du noyau K pour les fenêtres pertinentes (Ici, toutes)
# matrice de taille n x card(actifs)
Kij[,actifs]=t(t(dnorm(z))/sqrt(h[actifs]))
while(prod(h)>log(n)/n && (sum(actifs)>0)){
  A=actifs   # A: sauvegarde des comp. actives en début d'itérations
  prodK=prodKinactif*(apply(matrix(Kij[,A],n),1,prod))
  #calcul du noyau produit
  # vecteur nx1 des produits selon j des Kij sur le produit des sqrt(h_k)
  deriv=-1+z^2
  #facteur en plus par dérivation (uniquement pour K gaussien standart)
  Zsimpl=as.vector((t(deriv*prodK))%*%(1/fX.est))
  #vecteur des {Zj : j ACTIF} simplifié pour la comparaison avec le seuil
  Z2simpl=Zsimpl^2/n #=Z^2 *n prod(h)*hj^2
  #lambda2simpl=chigauss*(log(n))^(a)#/min(fX)
  #= 4 lambda_h^2 *n prod(h)*hj^2
  #seuil adapté pour le noyau gaussien
```

```
    #actualisation
    s#désactivation des composantes inférieures au seuil
    actifs[A]=Z2simpl>lambda2simpl
    # mise à jour des fenêtres désactivées à cette itération
    desactives=which(A!=actifs)
    if(length(desactives)!=0){
    # mise à jour du noyau desactivé et inactifs si des composantes se sont désactivées
      prodKdesactives=apply(matrix(Kij[,desactives],n),1,prod)
      prodKinactif=prodKinactif*prodKdesactives
    }

    if(sum(actifs)>0){
    #mises à jour pour l'itération suivante s'il reste des composantes actives
    #mise à jour de la fenêtre
      h[actifs]<-beta*h[actifs]
      #mise à jour du point d'application de K (juste les composantes ACTIVES)
      z=matrix(z[,actifs[A]],n)/beta
      Kij[,actifs]=t(t(dnorm(z))/sqrt(h[actifs])) #mise à jour de Kij
    }
    #print(actifs)
    #print(prod(h))#compteur d'itération mis à jour
    #fin de l'actualisation
  }#Fin de croissance des non pert
  return(h)
}
```

# Chapitre V

# Discussions et perspectives

J'apporte ici un point de vue un peu plus discuté des résultats des précédents chapitres, ainsi que des extensions envisageables pour faire suite à ce travail.

## Table des matières

# 1 Raffinements théoriques

## 1.a Inégalités oracles

Dans les problèmes de sélection de modèles (une manière de traduire notre choix de fenêtre), un critère d'optimalité usuel est l'obtention d'inégalités oracles, qui comparent les performances d'une méthode de sélection à la performance du meilleur modèle de notre collection, qui nous serait fourni par un oracle.
Plus formellement :

**Définition V.1** (Inégalité oracle). *Pour une collection $\mathcal{M}$ de modèles et d'estimateurs associés $\{\hat{\theta}_m\}_{m\in\mathcal{M}}$, la sélection de modèles $\hat{m}$ vérifie une inégalité oracle pour une perte $\ell$ s'il existe $C \geq 1$ (de préférence proche de $1$, mais pouvant éventuellement dépendre de $n$) telle que :*

$$\mathbb{E}[\ell(\hat{\theta}_{\hat{m}}, \theta)] \leq C \inf_{m\in\mathcal{M}} \mathbb{E}[\ell(\hat{\theta}_m, \theta)] + r_n,$$

*où $r_n$ est un terme résiduel négligeable.*

Un tel résultat mène souvent à des résultats de convergence minimax optimale (bien que pas systématiquement dans le cas d'une collection de modèles mal adaptée au problème) et assure que la méthode proposée fait aussi bien que possible étant donnée la collection de modèles.
Ce résultat semble réalisable avec la procédure CDRodeo RevDir en changeant notre référence de fenêtre à atteindre. En particulier, dans la preuve de la Proposition III.5 du Chapitre III, on définit un « fenêtre minimax » (voir la ligne 5, Section 4.e du Chapitre III), qui, dépendant des inconnues du modèle, permet d'atteindre la vitesse minimax, à laquelle on compare la performance de l'estimateur que CDRodeo sélectionne. Dans le but d'obtenir une inégalité oracle, on pourrait se comparer plutôt à une « fenêtre oracle », définie comme réalisant le meilleur compromis biais-variance dans notre collection d'estimateurs.

## 1.b Anisotropie

Mon travail s'est concentré sur des fonctions appartenant à des boules de Hölder « isotropes », c'est-à-dire présentant le même degré de régularité dans chaque direction. Certains se sont plutôt intéressés à l'estimation de fonctions anisotropes (voir par exemple [Barron et al. 1999, p. 345] en densités ou [Bertin et al. 2016] en densités conditionnelles) : la régularité est encodée par un vecteur $s \in (\mathbb{R}_+)^d$ contenant le degré de régularité höldérienne dans chaque direction. Cette extension rend l'estimation adaptative en la régularité d'autant plus intéressante.
Cependant nos deux méthodes présentées ici ne peuvent atteindre l'anisotropie. C'est dû au fait que les deux procédures CDRodeo imposent à chaque itération que les composantes actives de la fenêtre (c'est-à-dire qui continuent de croître ou décroître) soient toutes de même valeur. Si l'analyse de l'algorithme en est simplifiée, cela empêche cependant le compromis biais-variance optimal dans le cas anisotrope .
Une variante de CDRodeo pourrait être envisagée où l'on ne mettrait à jour à chaque itération que la composante $j$ la plus éloignée du compromis, éloignée au sens où le rapport $\frac{|Z_{hj}|}{\lambda_{hj}}$ serait loin de $1$. Théorie et mise en pratique de cette procédure restent à faire.

# 2 Procédure locale ou globale ?

On s'est intéressé dans ce travail à une fonction de perte ponctuelle, un point de vue très local en estimation fonctionnelle. Les fonctions de perte discriminent les estimateurs selon différents aspects de la fonction. Donnons quelques arguments.

## 2.a Intérêt de l'estimation ponctuelle de densités conditionnelles.

Le but premier de l'estimation de densités a été de retrouver la distribution du jeu de données fourni. Il est vrai que l'évaluation en un point est assez limitée pour cet objectif : cela ne donne en particulier pas accès aux statistiques descriptives usuelles (moyenne, quantiles). Certains peuvent alors s'interroger sur l'intérêt de l'estimation ponctuelle de densités (conditionnelles ou non), en dehors bien sûr de l'argument « pour la beauté des mathématiques ».

Cependant, l'estimation globale de fonctions est aussi problématique. On a vu que la vitesse de convergence minimax optimale dépend de la régularité, mais sur quel ensemble ? Dans notre cas ponctuel, un voisinage du point d'évaluation suffisait, mais une approche globale oblige à considérer la régularité sur tout l'espace. L'hypothèse de régularité constante restreindrait fortement l'ensemble de fonctions considérées. Dans le cas contraire, je ne crois personnellement pas qu'une méthode purement globale, c'est-à-dire une méthode qui n'aurait à aucun moment besoin de considérer localement les degrés de régularité, puisse qualitativement estimer une fonction à régularité variable à cause de problèmes de régularisation (ou de compromis biais-variance).

Pour nous en convaincre, l'exemple proposé dans [Lafferty and Wasserman 2008, section 4.2 et Figure 5] illustre ce problème pour les estimateurs à noyau : ils estiment la fonction de régression $m(x) = (1/x)\sin(15/x)$ (à laquelle s'ajoute un bruit gaussien) en deux points de régularités différentes. La fenêtre sélectionnée dans chaque cas diffère du fait de la différence de pentes.

**Application dans les algorithmes MCMC.** Pour revenir à l'intérêt des estimateurs ponctuels de densités conditionnelles, des algorithmes d'échantillonnage en nécessitent : en particulier, l'algorithme de Metropolis-Hasting. Le but étant de simuler un échantillon selon une loi, l'algorithme le construit donnée par donnée par chaîne de Markov qui converge vers la loi cible. Plus spécifiquement, avant l'ajout d'une nouvelle donnée $y$ à l'échantillon en cours de construction $(x_1, \dots, x_k)$, l'algorithme de Metropolis-Hasting propose une étape d'acceptation, dont la probabilité d'acceptation dépend d'une densité conditionnelle évaluée au point $(x_k, y)$, à estimer.

## 2.b Vers un CDRODEO global ?

Mais si l'on souhaite quand même faire du global ? Est-il alors possible d'adapter la procédure CDRODEO (avec les avantages qu'on lui connaît) en estimation globale de la densité conditionnelle ? Cette approche n'est pas nouvelle, car les fondateurs de *rodeo* se la posait déjà [Lafferty and Wasserman 2008]. Mais ils ne la traitent que numériquement (sans décrire précisément la procédure qui semble sujette des problèmes d'initialisation et d'arrêt).
De plus, une procédure avec une garantie de performances s'étendant au moins sur un voisinage est nécessaire dans les théorèmes présentés, plus précisément pour satisfaire la condition (ii) en norme $\|\cdot\|_\infty$ sur un voisinage pour l'estimation de la densité de $X$.

Pour répondre à cette problématique, CDRODEO pourrait s'adapter de plusieurs manières. Voici deux idées :

- ⋆ une approche purement globale (avec les défauts qu'on lui connaît) : prendre un estimateur à noyau du biais global, et le dériver pour produire nos statistique de test $Z_{hj}$ pour sélectionner une fenêtre globale ;

- ⋆ ou interpoler nos estimées ponctuelles déjà étudiées. Mais cette approche amène de nombreuses questions.
  Quelle grille prendre ? Une grille régulière ? Rappelons qu'elle est $d$-dimensionnelle. Peut-on limiter le fléau numérique de la dimension avec une longueur de pas adaptative ? Si oui, comment construit-on cette grille : où démarre-t-on et dans quelle direction progresse-t-on ?
  Comment interpoler entre les points pour conserver nos garanties théoriques ? Mieux vaut-il interpoler les estimées ou les fenêtres ?
  Et question renormalisation, notre estimateur final est-il toujours une densité et s'intègre-t-il à $1$ ? Est-ce réellement mieux de le renormaliser en terme de qualité d'estimation ?

# 3 Estimation de densités et de la marginale de $X$

Rappelons que CDRodeo ne nécessite pas de données auxiliaires $X$ (on peut prendre $d_1 = 0$) pour une application directe en estimation de densités non conditionnelles. Ce sujet n'a pas été poussé à son maximum ici. Les Propositions II.1 et III.1 servent surtout à s'assurer que les conditions sur $\widetilde{\mathrm{f}}_X$ étaient réalisables : elles fournissent certes un estimateur de $\mathrm{f}_X$ répondant aux exigences des théorèmes, mais elles demandent un échantillon en $X$ nettement plus grand, de taille $n^c$ avec $c > 1$, ce qui non seulement est contraignant pour le praticien, mais aussi plombe le caractère glouton de toute la procédure. Par ailleurs dans le chapitre numérique, les simulations sont faites à $\mathrm{f}_X$ connue, ne permettant pas de se rendre compte des performances de CDRodeo en conditions réelles.
Néanmoins, la technique de preuve des Propositions II.1 et III.1, en passant par un $\varepsilon$-réseau, donne des résultats en norme $\infty$ sur un compact. Cela ouvre la voie vers une procédure globale où l'on assurerait une convergence minimax optimale sur tout un compact. Noter tout de même que pour l'estimation d'inverse de densité, la Condition (i) est nécessaire (pour ne pas diviser par $0$) et il n'existe pas de densité à support $\mathbb{R}^{d_1}$ tout en entier et minorée par un réel strictement positif.

## 3.a Estimation de $\mathrm{f}_X$ avec CDRodeo

Une simulation est en cours pour utiliser CDRodeo comme estimateur de $\mathrm{f}_X$ : on applique CDRodeo en chaque observation $X_k$ en utilisant l'échantillon $\{X_i\}_{i\neq k}$ privé seulement du point $X_k$ où l'on estime. Cette idée simple de $n$ exécutions est hélas très chronophage : $n$ est pris grand, $X$ est toujours multivarié et il y a peu de parcimonie en densité non conditionnelle. Voir la Section 2.b pour des idées plus sophistiquées.

Néanmoins, noter qu'en densité simple, les résultats théoriques de CDRodeo s'appliquent, garantissant de bonnes propriétés de convergence permettant d'obtenir quasiment la condition (ii) (en considérant plutôt un maximum sur les $X_i$ pour éviter le passage en norme $\infty$ dans la preuve du Lemme 3 : voir Section 5.d, Chapitre III, p. 98). Noter en particulier que l'on ne demande pas d'indépendance entre l'échantillon construisant $\widetilde{\mathrm{f}}_X$ et celui permettant de sélectionner $\hat{h}$.

### 3.b Approfondissements

Dans nos preuves, on décompose notre problème en deux problèmes : l'étude de $\bar{f}_{\hat{h}}$, l'analogue de $\hat{f}_{\hat{h}}$ où l'on remplace $\widetilde{\mathrm{f}}_X$ par la vraie densité marginale $\mathrm{f}_X$, et les conditions sur $\mathrm{f}_X$ et $\widetilde{\mathrm{f}}_X$ pour ne pas influencer les performances de $\bar{f}_{\hat{h}}$. En particulier, $\widetilde{\mathrm{f}}_X$ doit converger plus vite que $\bar{f}_{\hat{h}}$. Delyon and Portier [2016] s'intéressent aux autres cas : voir leur Table 1, p. 5 (nos cas est le (ii), c'est-à-dire $d < 2s < r - d/2$ avec la vitesse de convergence $n^{-\frac{s}{2s+d}}$). Mais Delyon and Portier [2016] va plus loin : pour approximer l'intégrale $\int g(x)dx$, ils montrent qu'en corrigeant bien l'estimateur $\frac{1}{n}\sum_{i=1}^{n}\frac{g(X_i)}{\widetilde{\mathrm{f}}_X(X_i)}$, on obtient de meilleures performances que si on injectait simplement la vraie densité marginale. Pourrait-on faire de même dans CDRODEO ?

## 4 Simulations et données réelles

La partie numérique aurait pu être poussée plus avant sur plusieurs points, en particulier en menant une comparaison factuelle de performances avec d'autres méthodes d'estimation de densités conditionnelles. Ci-après, je précise quelques réflexions qu'il aurait été intéressant d'explorer plus avant.

- ⋆ **Choix du noyau.** Dans les simulations, on a choisit par habitude le noyau gaussien mais son ordre, $2$, est plutôt faible et son support n'est pas compact (même s'il est numériquement nul au-delà de $40$ sur $\mathrm{R}$). Néanmoins, si la théorie demande un ordre supérieur à la régularité pour converger à vitesse optimale, il n'a pas été vraiment établi d'amélioration des performances numériques associées. Lever le doute (dans le cadre de CDRODEO) aurait été intéressant.
- ⋆ **Exemple test spécifique.** Les trois modèles exposés dans le chapitre numérique permettent certaines vérifications, mais il existe des exemples spécifiques pour la détection de variables auxiliaires pertinentes. En particulier, où $Y$ et $X_2$ dépendent de $X_1$ mais telle que $Y$ soit indépendante de $X_2$ sachant $X_1$. La non pertinence de $X_2$ est alors particulièrement dure à détecter. Est-ce que les procédures CDRODEO se laisseraient berner ?
- ⋆ **Accélération de procédures et nombre de données effectif pour estimer.** En dimension modérément grande apparaissent de grandes zones sans observations proches dans lesquels estimer ponctuellement une fonction devient ardu. Le nombre de données $n$ semble peu indicatif de la réelle quantité d'information locale. En particulier pour calculer notre estimateur ponctuel $\hat{f}_h(w)$, de nombreuses données sortent du support de $K_h(w-\cdot)$, allongeant inutilement le temps de calculs en ajoutant des zéros à la moyenne empirique $\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\widetilde{\mathrm{f}}_X(X_i)}K_h(w-W_i)$. Une pré-étape de tri des données selon leur distance au support de $K_h(w-\cdot)$ permettrait de ne considérer que les données effectives pour l'estimation de $f(w)$ et ainsi raccourcir le temps d'exécution. Comment éclaircir l'influence de ce nombre effectif des données bien localisées sur la qualité de l'estimation ?
- ⋆ **Données réelles** : « Enfin, des objectifs concrets ? ». L'application sur jeux de données réelles est un pan important de la statistique, en interaction avec tous les autres domaines de la science pour en apprendre plus sur le monde réel et nous permettant de communiquer et sensibiliser tout un chacun de l'intérêt de notre recherche en statistique. Ne pas avoir appliqué mes méthodes sur un jeu de données réelles reste un regret de mon travail de thèse.

# Bibliographie


Akakpo, N. (2012). Adaptation to anisotropy and inhomogeneity via dyadic piecewise polynomial selection. *Mathematical Methods of Statistics*, 21(1) :1–28.

Arlot, S. and Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statist. Surv.*, 4 :40–79.

Barron, A., Birgé, L., and Massart, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113(3) :301–413.

Bashtannyk, D. M. and Hyndman, R. J. (2001). Bandwidth selection for kernel conditional density estimation. *Comput. Statist. Data Anal.*, 36(3) :279–298.

Beaumont, M., Zhang, W., and Balding, D. (2002). Approximate bayesian computation in population genetics. *Genetics*, 162(4) :2025–2035.

Bertin, K., Lacour, C., and Rivoirard, V. (2016). Adaptive pointwise estimation of conditional density function. *Ann. Inst. H. Poincaré Probab. Statist.*, 52(2) :939–980.

Biau, G., Cérou, F., and Guyader, A. (2015). New insights into approximate bayesian computation. *Ann. Inst. H. Poincaré Probab. Statist.*, 51(1) :376–403.

Birgé, L. and Massart, P. (1998). Minimum contrast estimators on sieves : exponential bounds and rates of convergence. *Bernoulli*, 4(3) :329–375.

Blanchard, G., Hoffmann, M., and Reiß, M. (2016). Optimal adaptation for early stopping in statistical inverse problems . working paper or preprint.

Bouaziz, O. and Lopez, O. (2010). Conditional density estimation in a censored single-index regression model. *Bernoulli*, 16(2) :514–542.

Brunel, E., Comte, F., and Lacour, C. (2007). Adaptive estimation of the conditional density in the presence of censoring. *Sankhya*, 69(4) :734–763.

Chagny, G. (2013). Warped bases for conditional density estimation. *Submitted*.

Comminges, L. and Dalalyan, A. S. (2012). Tight conditions for consistency of variable selection in the context of high dimensionality. *The Annals of Statistics*, 40(5) :2667–2696.

De Gooijer, J. G. and Zerom, D. (2003). On conditional density estimation. *Statist. Neerlandica*, 57(2) :159–176.

Delyon, B. and Portier, F. (2016). Integral approximation by kernel smoothing. *Bernoulli*, 22(4) :2177–2208.

Donoho, D. L. and Low, M. G. (1992). Renormalization exponents and optimal pointwise rates of convergence. *Ann. Statist.*, 20(2) :944–970.

Efromovich, S. (1999). *Nonparametric Curve Estimation : Methods, Theory and Applications*. Springer Science & Business Media.

Efromovich, S. (2007). Conditional density estimation in a regression setting. *Ann. Statist.*, 35(6) :2504–2535.

Efromovich, S. (2010a). Dimension reduction and adaptation in conditional density estimation. *Journal of the American Statistical Association*, 105(490) :761–774.

Efromovich, S. (2010b). Oracle inequality for conditional density estimation and an actuarial example. *Ann. Inst. Statist. Math.*, 62(2) :249–275.

Epanechnikov, V. A. (1969). Non-parametric estimation of a multivariate probability density. *Theory of Probability & Its Applications*, 14(1) :153–158.

Fan, J., Yao, Q., and Tong, H. (1996). Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika*, 83(1) :189–206.

Fan, J. and Yim, T. H. (2004). A crossvalidation method for estimating conditional densities. *Biometrika*, 91(4) :819–834.

Fan, J.-q., Peng, L., Yao, Q.-w., and Zhang, W.-y. (2009). Approximating conditional density functions using dimension reduction. *Acta Mathematicae Applicatae Sinica, English Series*, 25(3) :445–456.

Farrell, R. (1972). On the best obtainable asymptotic rates of convergence in estimation of a density function at a point. *The Annals of Mathematical Statistics*, pages 170–180.

Faugeras, O. P. (2009). A quantile-copula approach to conditional density estimation. *J. Multivariate Anal.*, 100(9) :2083–2099.

Fernández-Soto, A., Lanzetta, K., Chen, H.-W., Levine, B., and Yahata, N. (2002). Error analysis of the photometric redshift technique. *Monthly Notices of the Royal Astronomical Society*, 330(4) :889–894.

Goldenshluger, A. and Lepski, O. (2011a). Bandwidth selection in kernel density estimation : oracle inequalities and adaptive minimax optimality. *Ann. Statist.*, 39(3) :1608–1632.

Goldenshluger, A. and Lepski, O. (2011b). Uniform bounds for norms of sums of independent random functions. *Ann. Probab.*, 39(6) :2318–2384.

Györfi, L. and Kohler, M. (2007). Nonparametric estimation of conditional distributions. *IEEE Trans. Inform. Theory*, 53(5) :1872–1879.

Hall, P., Racine, J., and Li, Q. (2004). Cross-validation and the estimation of conditional probability densities. *J. Amer. Statist. Assoc.*, 99(468) :1015–1026.

Holmes, M. P., Gray, A. G., and Isbell, C. L. (2010). Fast kernel conditional density estimation : A dual-tree monte carlo approach. *Computational Statistics & Data Analysis*, 54(7) :1707 – 1718.

Hyndman, R. J., Bashtannyk, D. M., and Grunwald, G. K. (1996). Estimating and visualizing conditional densities. *J. Comput. Graph. Statist.*, 5(4) :315–336.

Hyndman, R. J. and Yao, Q. (2002). Nonparametric estimation and symmetry tests for conditional density functions. *J. Nonparametr. Stat.*, 14(3) :259–278.

Ichimura, T. and Fukuda, D. (2010). A fast algorithm for computing least-squares cross-validations for nonparametric conditional kernel density functions. *Computational Statistics & Data Analysis*, 54(12) :3404–3410.

Izbicki, R. and Lee, A. B. (2016). Nonparametric conditional density estimation in a high-dimensional regression setting. *Journal of Computational and Graphical Statistics*, 25(4) :1297–1316.

Izbicki, R. and Lee, A. B. (2017). Converting high-dimensional regression to high-dimensional conditional density estimation. *Electron. J. Statist.*, 11(2) :2800–2831.

Izbicki, R., Lee, A. B., and Pospisil, T. (2018). Abc-cde : Towards approximate bayesian computation with complex high-dimensional data and limited simulations. *arXiv preprint arXiv :1805.05480*.

Jeon, J. and Taylor, J. W. (2012). Using conditional kernel density estimation for wind power density forecasting. *J. Amer. Statist. Assoc.*, 107(497) :66–79.

Khas' minskii, R. (1979). A lower bound on the risks of non-parametric estimates of densities in the uniform metric. *Theory of Probability & Its Applications*, 23(4) :794–798.

Lafferty, J. and Wasserman, L. (2008). Rodeo : Sparse, greedy nonparametric regression. *Ann. Statist.*, 36(1) :28–63.

Le Pennec, E. and Cohen, S. (2013). Partition-based conditional density estimation. *ESAIM : Probability and Statistics*, eFirst.

Lepskii, O. V. (1991). Asymptotically minimax adaptive estimation. i. upper bounds. optimally adaptive estimates. *Theory Probab. Appl.*, 36(4) :682–697.

Lincheng, Z. and Zhijun, L. (1985). Strong consistency of the kernel estimators of conditional density function. *Acta Mathematica Sinica*, 1(4) :314–318.

Liu, H., Lafferty, J. D., and Wasserman, L. A. (2007). Sparse nonparametric density estimation in high dimensions using the rodeo. In *International Conference on Artificial Intelligence and Statistics*, pages 283–290.

Marin, J.-M., Pudlo, P., Robert, C. P., and Ryder, R. (2012). Approximate bayesian computation methods. *Statistics and Computing*, 22(6) :1167–1180.

McDonald, D. (2017). Minimax Density Estimation for Growing Dimension. In Singh, A. and Zhu, J., editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 194–203, Fort Lauderdale, FL, USA. PMLR.

Nguyen, M.-L. J. (2018). Nonparametric method for sparse conditional density estimation in moderately large dimensions. In revision.

Nguyen, M.-L. J., Lacour, C., and Rivoirard, V. (2019). Adaptive greedy algorithm for moderately large dimensions in kernel conditional density estimation. In preparation.

Otneim, H. and Tjøstheim, D. (2018). Conditional density estimation using the local gaussian correlation. *Statistics and Computing*, 28(2) :303–321.

Panaretos, V. M. and Konis, K. (2012). Nonparametric construction of multivariate kernels. *Journal of the American Statistical Association*, 107(499) :1085–1095.

Rebelles, G. (2015). Pointwise adaptive estimation of a multivariate density under independence hypothesis. *Bernoulli*, 21(4) :1984–2023.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, pages 832–837.

Rosenblatt, M. (1969). Conditional probability density and regression estimators. In *Multivariate Analysis, II (Proc. Second Internat. Sympos., Dayton, Ohio, 1968)*, pages 25–31. Academic Press, New York.

Rosenblatt, M. (1971). Curve estimates. *Ann. Math. Statist.*, 42(6) :1815–1842.

Sart, M. (2017). Estimating the conditional density by histogram type estimators and model selection. *ESAIM : Probability and Statistics*, 21 :34–55.

Shiga, M., Tangkaratt, V., and Sugiyama, M. (2015). Direct conditional probability density estimation with sparse feature selection. *Machine Learning*, 100(2) :161–182.

Takeuchi, I., Le, Q. V., Sears, T. D., and Smola, A. J. (2006). Nonparametric quantile estimation. *Journal of Machine Learning Research*, 7(Jul) :1231–1264.

Takeuchi, I., Nomura, K., and Kanamori, T. (2009). Nonparametric conditional density estimation using piecewise-linear solution path of kernel quantile regression. *Neural Comput.*, 21(2) :533–559.

Tsybakov, A. B. (1998). Pointwise and sup-norm sharp adaptive estimation of functions on the Sobolev classes. *Ann. Statist.*, 26(6) :2420–2469.

Wasserman, L. and Lafferty, J. D. (2006). Rodeo : Sparse nonparametric regression in high dimensions. In *Advances in Neural Information Processing Systems*, pages 707–714.

université PARIS-SACLAY | École doctorale de mathématiques Hadamard (EDMH)

**Titre :** Estimation non paramétrique de densités conditionnelles : grande dimension, parcimonie et algorithmes gloutons.

**Mots Clefs :** estimation non paramétrique, grande dimension, parcimonie, densité conditionnelle, algorithmes gloutons, estimateurs à noyau.

**Résumé :** Nous considérons le problème d'estimation de densités conditionnelles en modérément grandes dimensions. Beaucoup plus informatives que les fonctions de régression, les densités conditionnelles sont d'un intérêt majeur dans les méthodes récentes, notamment dans le cadre bayésien (étude de la distribution postérieure, recherche de ses modes...). Après avoir rappelé les problèmes liés à l'estimation en grande dimension dans l'introduction, les deux chapitres suivants développent deux méthodes qui s'attaquent au fléau de la dimension en demandant : d'être efficace computationnellement grâce à une procédure itérative gloutonne, de détecter les variables pertinentes sous une hypothèse de parcimonie, et converger à vitesse minimax quasi-optimale. Plus précisément, les deux méthodes considèrent des estimateurs à noyau bien adaptés à l'estimation de densités conditionnelles et sélectionnent une fenêtre multivariée ponctuelle en revisitant l'algorithme glouton RODEO (Regularisation Of Derivative Expectation Operator). La première méthode ayant des problèmes d'initialisation et des facteurs logarithmiques supplémentaires dans la vitesse de convergence, la seconde méthode résout ces problèmes, tout en ajoutant l'adaptation à la régularité. Dans l'avant-dernier chapitre, on traite de la calibration et des performances numériques de ces deux procédures, avant de donner quelques commentaires et perspectives dans le dernier chapitre.

**Title:** Nonparametric estimation of sparse conditional densities in moderately large dimensions by greedy algorithms.

**Keywords:** nonparametric estimation, high dimension, sparsity, conditional density, greedy algorithms, kernel density estimators.

**Abstract:** We consider the problem of conditional density estimation in moderately large dimensions. Much more informative than regression functions, conditional densities are of main interest in recent methods, particularly in the Bayesian framework (studying the posterior distribution, finding its modes...). After recalling the estimation issues in high dimension in the introduction, the two following chapters develop on two methods which address the issues of the curse of dimensionality: being computationally efficient by a greedy iterative procedure, detecting under some suitably defined sparsity conditions the relevant variables, while converging at a quasi-optimal minimax rate. More precisely, the two methods consider kernel estimators well-adapted for conditional density estimation and select a pointwise multivariate bandwidth by revisiting the greedy algorithm RODEO (Regularisation Of Derivative Expectation Operator). The first method having some initialization problems and extra logarithmic factors in its convergence rate, the second method solves these problems, while adding adaptation to the smoothness. In the penultimate chapter, we discuss the calibration and numerical performance of these two procedures, before giving some comments and perspectives in the last chapter.

**Université Paris-Saclay**
Espace Technologique/ Immeuble Discovery
Route de l'Orme aux Merisiers RD 128
91190 Saint-Aubin, France