Validation croisée et pénalisation pour l'estimation de densité

Nelo Molter Magalhães Magalhães

UNIVERSITÉ PARIS SUD

Comprendre le monde, construire l'avenir®

UNIVERSITÉ PARIS-SUD

ECOLE DOCTORALE 142 : MATHÉMATIQUES DE LA RÉGION PARIS-SUD

LABORATOIRE DE MATHÉMATIQUES D'ORSAY

LABORATOIRE DE PROBABILITÉS ET MODÈLES ALÉATOIRES

DISCIPLINE : MATHÉMATIQUES

THÈSE DE DOCTORAT

Soutenue le 26 mai 2015 par

Nelo MAGALHÃES

Validation croisée et pénalisation pour l'estimation de densité

Composition du jury :

Directeur de thèse :	Pascal MASSART	Université Paris-Sud
Co-directeur de thèse :	Lucien BIRGÉ	Université Pierre et Marie Curie
Président du jury :	Yannick BARAUD	Université Nice Sophia Antipolis
Rapporteurs :	Vincent RIVOIRARD	Université Paris Dauphine
	Nicolas VAYATIS	ENS Cachan
Examinateur :	Guillaume LECUÉ	CNRS et École Polytechnique

(...)

Fais ton bonheur d'augmenter celui de tous. Travaille et lutte et n'accepte de mal rien de ce que tu pourrais changer. Sache te répéter sans cesse : il ne tient qu'à moi. On ne prend point son parti sans lâcheté de tout le mal qui dépend des hommes. Cesse de croire, si tu l'as jamais cru, que la sagesse est dans la résignation ; ou cesse de prétendre à la sagesse.

Camarade, n'accepte pas la vie telle que te la proposent les hommes. Ne cesse point de te persuader qu'elle pourrait être plus belle, la vie ; la tienne et celle des autres hommes ; non point une autre, future, qui nous consolerait de celle-ci et qui nous aiderait à accepter sa misère.

N'accepte pas. Du jour où tu commenceras à comprendre que le responsable de presque tous les maux de la vie, ce n'est pas Dieu, ce sont les hommes, tu ne prendras plus ton parti de ces maux. Ne sacrifie pas aux idoles.

(...)

Gide

Je dédie cette thèse à mes amis, aux vagabonds désespérément optimistes, et à tous ceux qui *n'acceptent pas*.

Remerciements

Merci à tous les membres du jury pour l'honneur qu'ils me font d'être présents aujourd'hui. Je suis très reconnaissant envers les rapporteurs d'avoir pris le temps de lire de façon critique mon travail. Comme j'étais inquiet de vous le soumettre, je dois reconnaître que j'étais soulagé et sincèrement touché en lisant vos commentaires encourageants. En particulier, je remercie mes deux directeurs d'avoir pris le pari de me lancer dans une thèse. Si le chemin fut parfois difficile, ce manuscrit doit nous convaincre qu'il n'y a pas de regret à avoir pour s'être engagés dans cette longue aventure.

L'image qui me vient à l'esprit quand je pense aux années de thèse, c'est cette tour humaine - célèbre en Catalogne - que l'on nomme *castells*. Cette construction se compose de *castellers* qui se juchent sur les épaules les uns des autres en une succession d'étages dans le but de faire monter un enfant au sommet. Aujourd'hui le castells s'achève et cet enfant, là-haut, c'est moi. Mais je n'ai aucune gloire personnelle à en tirer. C'est vous, les castellers, qui m'avez porté, soutenu, donné confiance et encouragé à escalader les étages sans crainte de tomber - car vous seriez là pour me rattraper. Vous êtes parvenus à me convaincre, moi qui ai si peur de l'imposture, que j'avais le droit de grimper en haut de la tour, puisque j'y mettais de la volonté et du travail. En vous côtoyant, j'ai compris que l'histoire du génie solitaire qui produit des merveilles dans son coin est un mythe qu'il faut définitivement écarter en sciences. Tout ce qui suit est, par conséquent, issu d'un travail commun : tout ce qui est écrit doit effectivement se lire à la première personne du pluriel.

Le dernier étage est composé de mes co-auteurs qui ont rendu ce manuscrit concret : c'est la partie la plus visible de l'iceberg - pardon, du castells.

Merci à toi, Matthieu. Ton importance a été capitale cette dernière année de thèse. Je valide ton stage de directeur de thèse avec la plus haute mention : pour ta patience héroïque face à mes questions répétées, tes encouragements et ton enthousiasme sincère à la moindre de mes hypothèses (ou début de morceau d'idée mal formulée). C'est vraiment avec toi que j'ai (enfin) compris ce qu'était la recherche mathématique et c'est en ta compagnie que j'ai appris à mettre les mains dans le cambouis : me tromper, recommencer, ne rien comprendre, comprendre un peu, avoir une intuition, la démolir, reprendre l'idée en la modifiant, voir tout qui s'agence dans une preuve, etc. Tu aurais pu tout faire seul, car tu vois mieux que quiconque dans la Matrix de la sélection optimale L_2, mais tu as naturellement préféré qu'on le fasse ensemble. Merci aussi pour les invitations à Nice, les verres partagés dans la bonne humeur, les concours d'imitations, et j'en passe !

Merci à toi, Sylvain, dont la thèse (souvent imitée, jamais égalée) constitue sans aucun doute mon oracle personnel. Tu parviens toujours à mettre les explications précises sur des concepts qui tournent dans ma tête à la recherche des bons mots. Tu rends limpides mes propres idées : quand je te lis je me dis "c'est exactement ça que je voulais dire.. mais en mieux écrit". Ta lucidité, ta compréhension-éclair des preuves, ta sérénité à toute épreuve malgré tes 143 travaux en cours ont renforcé ma sincère admiration, née lors du cours de M2 que tu donnais à Orsay (j'avais alors présenté un papier de... Matthieu pour l'examen !). Je salue aussi ton éthique irréprochable qui t'a poussé à m'inviter à collaborer avec vous - puisque Matthieu et moi avions eu la même idée. Merci pour les ballades à Cargèse ainsi que nos discussions à midi au restaurant végétarien. Merci pour tes relectures critiques en toute fin de thèse. Pouvoir te côtoyer est un réel privilège.

Merci à toi, l'incroyable Patricia. Si tu le pouvais, je suis sûr que tu aiderais tous les thésards de la terre (et dans tous les domaines) tant tu es généreuse et sincèrement curieuse. Outre ton caractère naturellement rassurant, tu m'as littéralement ébloui (oui, ça brille quand tu parles !) à chaque fois que nous avons partagé un moment ensemble, par ton engouement pour la Science, tes connaissances hallucinantes, ton aptitude à toucher les sommets dans la joie et sans effort, et tes 200 idées lumineuses à la seconde. Un immense merci pour ta relecture de dernière minute, tes conseils de rédaction et ta réelle bienveillance de grande sœur.

Merci à toi, Yves, pour m'avoir récupéré à un moment où je coulais et pour m'avoir initié à la programmation (en partant de rien !). Ensemble nous avons passé de longues journées à développer ce joli package R. S'il arrivait qu'elles soient frustrantes (auquel cas ton petit whisky se pointait pour nous détendre en fin de journée), je dois dire qu'elles étaient souvent très enrichissantes.

Quand on fait un petit zoom arrière, on aperçoit les visages du quotidien : ceux qui m'ont supporté, qui m'ont vu (souvent) souffrir et (un peu) vibrer. Merci à toi, collègue de Jussieu, avec qui j'ai toujours plaisir à discuter opéra, voyages, politique, art, sport, littérature, pédagogie, végétarisme, vélo, etc. Merci à toi, collègue du bureau 16-26-131 : le fantasque Reda dont le titre de thèse fait encore trembler les murs du laboratoire, Cyril (dont j'admire secrètement la régularité et la concentration stakhanoviste) pour son soutien moral (et le foot ! et les bières !), Olga pour la salsa et ses encouragements, Liping dont le geste de fin de journée est le plus mignon de l'histoire du bureau, Wangru pour sa sempiternelle bonne humeur, Pablo mi compañero de Sudamérica, Sandro dont je me sens absolument proche sur la façon d'appréhender la vie, Bastien notre incroyable (et incroyablement modeste) "réponse à tout", et "papa" Guillaume dont je jalouse la capacité à rester imperméable à la pression et au stress. Merci aux bibliothèques publiques qui m'ont hébergé (la BPI et surtout la MIR), à tout le personnel administratif ainsi qu'aux secrétaires de Jussieu et d'Orsay à qui la recherche en mathématique doit énormément.

Je tiens aussi à remercier quelques mathématiciens, rencontrés lors des divers colloques et conférences, qui m'ont accordé du temps à une époque où je végétais oisivement au fond du trou. En particulier merci à toi Karim Lounici, Sébastien Gerchinovitz, Gaëlle Chagny, Emilien Joly, Maud Thomas, Mathieu Sart, Laure Sansonnet et Ilaria Giulini. Une mention spéciale pour Mathieu qui m'a soutenu pendant des mois sur mes différents projets et qui a relu un chapitre et l'introduction de cette thèse. Tu as été un grand frère inespéré pour moi en cette fin de parcours.

La base du castells contient tout ce qui est non-académique, sans qui rien ne serait possible. Il y a là une kyrielle de faits et gestes du quotidien - capturés par mon carnet de notes, mon appareil photo ou, plus simplement, ma mémoire - qui me font sentir vivant et m'inspirent une réelle fureur de vivre. La Seine qui dort tel un lac le matin, la complicité avec mes élèves qui s'émancipent, le saxo de Redman, les solos de Roach, mes indispensables empêcheurs de tourner en rond (au hasard : Lordon, Gorz, Chomsky, Bourdieu, Stengers, Friot, Debord, Rancière), le ciel d'Ouzbékistan (qui m'a tant promis que je vais le rejoindre bientôt), Ernesto et ses visites nocturnes, Mermet et le Diplo, les pensées existentielles de Kundera, l'argot hilarant de Céline, l'âme vagabonde de Cendrars et Gary, l'absurde de Beckett, le bleu de Chagall, les subtiles répétitions de Philipp Glass, le "moment décisif" de Cartier-Bresson, la danse cyclique de ATDK ou les pas envoûtants des poupées mécaniques de Lucinda Childs.

Ma famille est naturellement présente, au pied du castells. Merci à toi, Cosme, pour ces 900 km à vélo et pour notre complicité absolue. Merci à toi, papa Anthony, pour tes mots encourageants, ta confiance, ton intérêt pour mes espoirs et mes craintes. Merci pour ta générosité hors-norme qui me permet, depuis toujours, d'assouvir mon immense curiosité dans bien des domaines. Obrigado a você, pai, por ter feito a viagem do Brasil e por me fazer sentir o que é "arte à flor da pele". Merci à toi, maman, d'être si sereine face à un fils si tourmenté, et de me répéter sans cesse que la vie n'est pas si grave, au fond. Merci de me pousser calmement, dans le dos.

Enfin, ce sont mes amis qui m'ont accompagné toutes ces années que je voudrais remercier. Mon groupe d'amis belges mais aussi la chaleureuse communauté de Nanterre ainsi que Alexandre, Aurélie et Jean-Louis, John, Maelle, Margot, Marion et Mélanie. Merci à BTMY - Mélik, Quentin et Yvan - pour l'escapade visuelle autour du monde, et pour les vélos. Merci à mon formidable "tonton" pour toutes nos aventures, depuis 9 ans déjà, hors (ou sous) les sentiers battus. Yann et Jonas : vous avez été extraordinaires durant cette période. Nous avons tant bu de rouge, tant écouté de jazz et tant discuté de sciences sociales et de politique, qu'il faudrait une autre thèse pour tout évoquer. Merci pour le soutien permanent et pour cette amitié totale.

À vous tous, sachez-le : moi aussi je vous porterai en haut d'un castells !

Je conclus en te remerciant, jolie Aphirom, toi qui me fais encore mieux sentir que la vie est bel et bien ailleurs. Merci pour ce qu'il y a de meilleur : le rire, l'aventure, la tendresse et l'amour. Merci à l'avenir qui est tout entier à nous.

FIGURE 1 – Castell "3 de 10" avec "folre" et avec "manilles" par les Castellers de Vilafranca. Eric Sala & Tània García (Creative Commons BY-SA).

Avant-propos et organisation de la thèse

Depuis une trentaine d'années, le développement des techniques informatiques - et en particulier l'augmentation spectaculaire de la puissance de l'ordinateur comme instrument de simulation - a fortement modifié le paysage de la recherche en mathématique, notamment dans les domaines qui nous concernent des probabilités et statistiques (où il est possible de simuler l'aléa). L'image caricaturale ou fantasmée[(1)] du mathématicien travaillant dans un monde détaché du nôtre muni uniquement d'un crayon et de papier, est doucement en train de s'effriter. Avec elle s'effondre aussi la représentation de la Mathématique présentée comme une branche scientifique à part : une discipline fondée uniquement sur des concepts abstraits, qui serait un simple (mais néanmoins puissant et sophistiqué) outil au service des sciences empiristes et expérimentales classiques, ce qui la situerait en conséquence quelque part au-dessus de celles-ci. Le débat sur le caractère scientifique de la Mathématique a ainsi souvent été tranché comme suit : oui, elle relève de la Science ne fût-ce que parce qu'elle est nécessaire à la formalisation des autres disciplines (pas de Science sans elle) ; non, ce ne peut être une science comme les autres car la partie observation et expérimentation y est totalement absente. Décider si les mathématiques satisfont aux critères de réfutabilité de Karl Popper[(2)] est un autre point qui fait l'objet de nombreux débats. Si la Mathématique peut être réfutable au sens où une hypothèse peut être rejetée (par exemple au moyen d'un raisonnement par l'absurde), sa nature abstraite impliquerait qu'elle ne peut satisfaire ces critères.

Selon la philosophe des sciences Isabelle Stengers[(3)], le chercheur en mathématique *ne vise plus une Vérité absolue qui ferait taire les fictions*, mais s'évertue plutôt à *construire pour tout phénomène la fiction mathématique qui le reproduit*. Selon elle, la simulation modifie la hiérarchie entre phénomène purifié et complications anecdotiques. Elle met en effet sur un même plan ce qu'elle prend en compte : les "lois" deviennent des contraintes dont les effets n'ont aucun intérêt indépendamment des circonstances qui font de chaque simulation un nouveau cas. L'art du simulateur est alors *de définir la manière dont une multiplicité disparate d'éléments jouent ensemble, pour ensuite suivre les histoires qu'est susceptible d'engendrer cette matrice narrative*. Ce sont ces histoires qui font de la simulation une expérimentation sur nos énoncés : *elles les mettent en acte, sans nous donner la possibilité d'intervenir, d'infléchir dans le sens de ce que nous désirons ou jugeons plausible. L'explication d'un processus peut révéler qu'elle impliquait certes ce qu'elle visait mais peut-être tout aussi bien, dans des circonstances légèrement différentes, un processus très différent*[(4)].

[(1)] notamment par l'auteur de ces lignes, au début de ses études universitaires.

[(2)] qui peuvent, grossièrement, se résumer par sa célèbre phrase "Une théorie qui n'est réfutable par aucun événement qui se puisse concevoir est dépourvue de caractère scientifique." (Conjectures et réfutations, ch.1, section 1). Ainsi, selon son principe de "falsification", la démarche du savant doit consister non pas à prouver le bien-fondé d'une théorie mais à essayer de la démolir, de multiplier les expériences susceptibles de démontrer qu'elle est fausse. Ce n'est que si la théorie résiste à ces tests qu'elle peut être considérée comme scientifiquement vraie - du moins jusqu'à la prochaine théorie, plus générale encore, qui la remplacera dans la succession des mises à l'épreuve.

[(3)] Les extraits qui suivent sont tirés de son livre *L'invention des sciences modernes*, Paris, La Découverte, 1993.

[(4)] Elle ajoute au passage que ce changement de paradigme engendre inévitablement de nombreuses questions auquel doit se soumettre le scientifique - que veut dire l'énoncé "l'expérience montre que" lorsqu'il ne s'agit plus d'un événement, lien conquis entre les mots et les choses, mais d'une scène qui est définie toute entière en termes de représentations ? - et nécessite une *éthique de la simulation* (car la manière dont un programme "trafique" les lois, en négocie la portée au lieu d'en traduire le pouvoir, met en question le mode d'engagement mutuel entre démarche, vérité et réalité).

Cette thèse peut se lire comme un argument supplémentaire visant à intégrer pleinement la Mathématique dans la grande famille des sciences qui suivent une démarche théorico-expérimentale. Les différents chapitres du manuscrit constituent en effet autant de va-et-vient entre théorie et expérience, et par conséquent autant d'illustrations de cette démarche. Nous ne soutenons pas qu'elle est strictement égale aux autres[5] mais qu'elle fonctionne effectivement selon le même *modus operandi*. La Partie I illustre (en quelque sorte) le critère de réfutabilité de Popper puisque nous "testons" la solidité d'une théorie, *l'heuristique de pente*, dans un cadre plus large que celui pour lequel elle a été initialement formulée[6]. Comme dans les sciences expérimentales, on partira parfois d'un travail purement théorique pour ensuite regarder en pratique si les simulations confirment les théorèmes énoncés (Partie II). À l'inverse, on proposera une procédure qui semble prometteuse en pratique, pour ensuite expliquer son comportement par des preuves mathématiques (Partie III). Enfin, certaines simulations dans la Conclusion peuvent motiver de nouvelles procédures qui soient algorithmiquement moins coûteuses, créant ainsi un appel à de nouvelles recherches fondamentales pour les garantir théoriquement. Ce processus de recherche met en évidence de grandes disparités entre théorie et pratique : d'une part de solides résultats théoriques peuvent afficher certaines limites en pratique, d'autre part des procédures plébiscitées par les praticiens manquent d'une base théorique solide.

Cette thèse porte sur le problème de la sélection d'estimateur dans le cadre de l'estimation d'une densité, traité d'une part sans rééchantillonnage (Partie I) et d'autre part en utilisant celui-ci (la Partie II avec la perte des moindres carrés et la Partie III avec la perte Hellinger). L'objectif principal est de proposer des procédures de sélection optimales, par pénalisation ou validation croisée, pour ce problème. Nous cherchons également à comparer de telles procédures, étudions le rôle du paramètre V dans les procédures de validation croisée V-fold (VCVF) ainsi que la validité de l'heuristique de pente dans ce cadre. Chaque partie doit pouvoir se lire indépendamment du reste, excepté la Conclusion qui nécessite les notions et notations présentées dans l'Introduction et le Chapitre 6.

- Le premier chapitre est une introduction générale qui sert de présentation des procédures de validation croisée et de pénalisation qui sont considérées dans cette thèse.
- La Partie I ne contient qu'un seul chapitre, le Chapitre 2, qui traite de la sélection optimale d'estimateurs par pénalisation, de la présence d'une pénalité minimale ainsi que de l'heuristique de pente.
- La Partie II concerne des procédures rééchantillonnées (pénalité rééchantillonnée, pénalité V-fold, leave-p-out, VCVF) reposant sur l'estimation sans biais du risque dans le cadre des moindres carrés. Le Chapitre 3 étudie l'optimalité au premier ordre de ces procédures. Le Chapitre 4 est un travail en cours qui vise à les comparer du point de vue théorique afin de mieux les utiliser en pratique.

[5] Par exemple lorsqu'un théorème mathématique est rigoureusement démontré rien ne peut plus le remettre en question. Ainsi, à l'inverse de la Physique ou de la Chimie, aucune expérience ne pourra venir démolir une théorie.

[6] Celle-ci résiste n'est plus vraie telle qu'elle : ceci ne démolit pas ce qui précède mais montre qu'une généralisation nécessitera une re-formulation.

- La Partie III concerne le cadre Hellinger. Elle présente une alternative aux précédentes procédures, par rééchantillonnage également, qui repose sur des tests robustes. Le Chapitre 5 propose un algorithme pratique pour implémenter cette procédure dans le cas de la validation simple. Le Chapitre 6 est un travail en cours dans lequel nous étudions, du point de vue théorique et pratique, la procédure de type V-fold qui en est le prolongement naturel.
- Le dernier chapitre rappelle les principales conclusions de cette thèse et suggère, à partir d'arguments théoriques, une nouvelle procédure qui jouisse des avantages des différents critères. Il présente enfin une série de questions ouvertes qui constituent autant de recherches à mener pour rapprocher davantage théorie et pratique.

Un résumé figure à la toute fin du manuscrit.

Table des matières

Introduction

Sommaire

Le postulat de base, quand on fait de la Statistique, consiste à supposer que ce que l'on observe est le résultat d'une expérience aléatoire. Le but est alors de déterminer quelle est la *loi* (par nature inconnue) qui régit les observations.

Une stratégie d'estimation s'est développée de façon intense ces dernières décennies tant pour la souplesse qu'elle offre pour la modélisation que pour la variété d'algorithmes efficaces qui peuvent être déployés pour sa mise en œuvre. Elle consiste à construire, dans un premier temps, des estimateurs de la loi et de sélectionner, ensuite, un candidat selon une procédure bien définie. Deux grandes tendances existent pour cette stratégie générale. L'une, considérée dans la Partie I, utilise "en bloc" toutes les données pour les deux étapes. L'autre, que nous suivrons dans les Parties II et III, consiste à segmenter l'échantillon et à itérer successivement les deux étapes sur deux sous-échantillons distincts.

Dans cette thèse nous allons essentiellement étudier, du point de vue théorique et pratique, des procédures – par *pénalisation* ou par *validation croisée V-fold* – dont l'objectif commun est de sélectionner un candidat dans une famille donnée de *méthodes d'estimation*. Nous examinerons deux approches sensiblement différentes pour évaluer la qualité de chacun des concurrents à l'étape de sélection. L'une vise à estimer leurs *risques*, l'autre à leur attribuer un *indice de plausibilité* en les comparant paire par paire au moyen de *tests robustes*.

Ce chapitre est dédié à l'introduction du sujet de thèse et aux définitions de ces différentes notions en partant du cadre général de l'estimation d'une densité (Section 1.1.1) pour se diriger vers le problème précis qui nous intéresse, celui de la sélection d'une méthode d'estimation, exposé en détail en Section 1.1.3. La Section 1.2 présente les procédures qui seront analysées dans cette thèse pour résoudre ce problème, et rappelle de façon succincte la bibliographie les concernant. Enfin, le chercheur en Statistique déjà coutumier du domaine pourra directement se référer à la Section 1.3 qui présente les principaux résultats ainsi que les recherches originales de cette thèse.

1.1 Cadre et problèmes de sélection

1.1.1 Cadre statistique : estimation de la densité

Dans toute la suite, on dispose d'un échantillon $\mathbf{X} = \{X_1, \dots, X_n\}$ de variables aléatoires X_i, à valeurs dans un espace mesurable $(\Xi, \mathcal{Z})$, indépendantes et identiquement distribuées selon une loi P que l'on cherche à estimer. Lorsque P est absolument continue par rapport à une mesure μ sur Ξ, nous notons $s = dP/d\mu$ la **densité** et supposons qu'elle appartient à un sous-ensemble $\mathcal{S}$ de $\mathbb{L}_1(\mu)$.

Répondre au problème de l'estimation d'une densité consiste à proposer, à partir des observations, un **estimateur** $\widehat{s} = \widehat{s}(\mathbf{X})$ dans $\mathcal{S}$ qui soit "le plus proche" possible de s. La qualité d'un candidat $t \in \mathcal{S}$ pour approcher s est jugée au moyen d'une **fonction de perte**[7] $\ell : \mathcal{S} \to \mathbb{R},\ t \mapsto \ell(s,t)$, qui est supposée minimale pour $t = s$. Comme l'estimateur $\widehat{s}$ est aléatoire, car il dépend des données, sa perte $\ell(s, \widehat{s})$ l'est aussi. Pour mesurer la qualité d'un estimateur par une quantité déterministe, on définit son **risque** $\mathbb{E}_P\left[\ell(s, \widehat{s})\right]$, où $\mathbb{E}_P$ désigne l'espérance mathématique par rapport à la loi de probabilité $P^{\otimes n}$ de l'échantillon $\mathbf{X}$ (dans la suite nous noterons de façon abusive $\mathbb{E}\left[\ell(s, \widehat{s})\right]$). Dans l'absolu le but est donc de construire un estimateur avec la plus petite perte, ou

[7] En apprentissage statistique, on l'appelle *perte relative*.

le plus petit risque, "possible". Le problème qui nous intéresse particulièrement dans ce travail est celui de la sélection du meilleur estimateur *possible* dans une collection donnée de candidats. Un autre objectif, secondaire dans ce manuscrit, est la construction d'un estimateur qui ait le plus petit risque *possible* pour une famille donnée de lois. Afin de restreindre un peu la présentation, nous allons adopter une approche qui présente deux caractéristiques principales.

- Elle est *non-asymptotique* : le nombre d'observations n est à prendre pour ce qu'il est ! Notre objectif est principalement d'obtenir des résultats théoriques avec un nombre d'observations ne tendant pas vers l'infini, de sorte que tous les paramètres du problème considéré apparaissent explicitement dans les bornes obtenues, ce qui permet en général une compréhension plus fine de leurs rôles en pratique[(8)].

- Elle est *non-paramétrique adaptative* : on ne fait pas d'hypothèses a priori sur P. En particulier, on ne suppose pas que P appartient à une famille $(P_\beta)_{\beta\in B}$ paramétrée par un ensemble $B \subset \mathbb{R}^d$. Toutefois, il est fréquent de supposer, a posteriori, que la densité possède une certaine régularité, ou appartient à un espace fonctionnel (de dimension infinie) par exemple. On cherche ainsi à construire un estimateur qui *s'adapte* aux caractéristiques de la loi.

1.1.2 Une histoire de perte

Différentes pertes peuvent être utilisées en estimation de densité : selon le goût et les préférences du statisticien mais aussi selon des considérations plus mathématiques. Un point crucial est le suivant. Si P et $P_{\widehat{s}}$ sont absolument continues par rapport à μ, que devient la perte de l'estimateur $\widehat{s} = dP_{\widehat{s}}/d\mu$ de s lorsqu'on modifie la mesure dominante ?

Selon la réponse apportée à cette question, on trouve deux groupes de perte dans la littérature. Le premier concerne les pertes définies sur $\mathcal{P}$, l'espace des lois de probabilités sur Ξ, donc qui *ne dépendent pas* de la mesure dominante. Le second contient les distances classiques entre les fonctions de $\mathcal{S}$, et sont sensibles à tout changement de mesure dominante.

En outre, l'hypothèse d'absolue continuité de la loi P par rapport à la mesure μ peut ne pas être satisfaite dans la réalité – auquel cas il est possible qu'il n'existe pas de densité qui régisse les observations. Si nous acceptons que cette hypothèse puisse être fausse, ce qui paraît le plus raisonnable, alors nous utilisons une perte sur $\mathcal{P}$. Si nous considérons cette hypothèse comme un *fait* alors nous nous focalisons uniquement sur la fonction s et les pertes sur $\mathcal{S}$ peuvent également être considérées. Nous mentionnons dans cette section les fonctions de perte les plus étudiées.

Commençons par les distances définies sur $\mathcal{P}$. Soient $P, Q \in \mathcal{P}$ et dP et dQ les densités de P et Q par rapport à n'importe quelle mesure dominante.

- La **distance en variation**,

$$D(P,Q) = \frac{1}{2}\int |dP - dQ| = \sup_{A\in\mathcal{Z}} |P(A) - Q(A)| \quad .$$

[(8)] Seule la Section 2.4 fait exception à la règle.

- Le carré de la **distance de Hellinger**,

$$h^2(P,Q) = \frac{1}{2}\int \left(\sqrt{dP} - \sqrt{dQ}\right)^2 \ . \tag{1.1}$$

L'affinité de Hellinger[(9)] se définit simplement par $\rho(P,Q) = 1 - h^2(P,Q) = \int \sqrt{dPdQ}$.

- La **divergence de Kullback-Leibler**,

$$\mathrm{KL}(P,Q) = \int \log\left(\frac{dP}{dQ}\right) dP \qquad \text{si} \quad P << Q \quad (\mathrm{KL}(P,Q) = +\infty \ \text{autrement}) \ .$$

Cette perte est plus contraignante à utiliser puisqu'elle nécessite une hypothèse de domination, au contraire de la distance en variation et la distance de Hellinger.

Le deuxième groupe contient les distances entre des fonctions de $\mathcal{S}$. Toute puissance $\ell = d^p$ d'une distance d définie sur $\mathcal{S}$ avec $p \geq 1$ peut être envisagée comme perte.

- Si $\mathcal{S} = \mathbb{L}_\infty(\mu) \cap \mathbb{L}_1(\mu)$, on peut considérer la *perte ponctuelle* définie pour tout $x_0 \in \Xi$ par $\ell(s,t) = |t(x_0) - s(x_0)|$. Dans cette optique, certains auteurs (Korostelev & Nussbaum, 1999; Giné & Nickl, 2009, 2010; Gach *et al.*, 2013) préfèrent la distance $\mathbb{L}_\infty$: $\ell(s,t) = \|s-t\|_\infty = \operatorname{ess\,sup}_{x\in\Xi} |s(x) - t(x)|$.

- Soit $p \geqslant 1$ un indice quelconque et $\mathcal{S} = \mathbb{L}_p(\mu) \cap \mathbb{L}_1(\mu)$. Dans ce cas, on définit naturellement la perte associée à la norme $\mathbb{L}_p$:

$$\ell(s,t) = \|s-t\|_p^p = \int_\Xi |s(x) - t(x)|^p \, d\mu(x) \ .$$

En particulier la perte $\mathbb{L}_2$ est appelée la *perte des moindres carrés* et le risque correspondant **risque quadratique intégré**[(10)].

- Les distances définies auparavant sur $\mathcal{P}$ peuvent également servir si $\mathcal{S}$ est l'ensemble des densités par rapport à la mesure μ sur Ξ. Dans ce cas, on a pour tout $t \in \mathcal{S}$

$$D(s,t) = \frac{1}{2}\int |s-t| \, d\mu, \quad h^2(s,t) = \frac{1}{2}\int \left(\sqrt{s} - \sqrt{t}\right)^2 d\mu, \quad \mathrm{KL}(s,t) = \int s \log\left(\frac{s}{t}\right) d\mu \ .$$

[(9)] Analyste, Hellinger a introduit dans la continuité de ses travaux de thèse une nouvelle intégrale (Hellinger, 1909), qu'on appelle aujourd'hui *l'intégrale de Hellinger*. Celle-ci a été utilisée vraisemblablement pour la première fois dans le domaine des probabilités par Kakutani (1948) qui se pose la question de l'équivalence de deux mesures produits $\mu = \prod \mu_n$ et $\nu = \prod \nu_n$. Son principal théorème affirme que les mesures sont équivalentes ou orthogonales selon qu'un certain produit $\prod \rho(\mu_n, \nu_n)$ soit positif ou nul. Cette fonction ρ se définit par une intégrale de Hellinger et deviendra plus tard *l'affinité de Hellinger* (dans son article, Kakutani remercie von Neumann d'avoir fait le lien avec l'intégrale de Hellinger : c'est ce dernier qui définit alors la distance de Hellinger – sans la nommer – par la relation donnée en (1.1)). Le nom de distance de Hellinger apparaît semble-t-il pour la première fois dans la thèse de Kraft (1955). Selon Le Cam (Le Cam & Yang, 2000, Chapter 3) une distance similaire était utilisée auparavant en mécanique quantique - il apparaît en effet que les probabilités dans ce domaine sont données par la racine carrée de fonctions d'onde. Ainsi, il semble quasi certain que Hellinger lui-même n'ait jamais considéré cette distance (surtout qu'il ne pouvait pas connaître le théorème de Radon-Nikodym) et que l'on doit au processus de la recherche mathématique, l'utilisation par des statisticiens (et les physiciens !) 40 ans après sa définition d'un outil initialement pensé et développé pour l'analyse spectrale.

[(10)] en anglais, M.I.S.E. : *Mean Integrated Squared Error*.

En raison de son agrément lié à la structure hilbertienne de l'espace $\mathbb{L}_2(\mu)$, de nombreux travaux en estimation de densité font usage de la perte quadratique. En effet, on peut facilement décomposer le risque quadratique d'un estimateur $\widehat{s}_m = \widehat{s}_m(\mathbf{X})$ de s, en notant $s_m(x) = \mathbb{E}[\widehat{s}_m(x)]$, comme la somme suivante

$$\mathbb{E}\left[\|\widehat{s}_m - s\|_2^2\right] = \|s - s_m\|_2^2 + \mathbb{E}\left[\|\widehat{s}_m - s_m\|_2^2\right] \quad . \tag{1.2}$$

Dans (1.2), le premier terme est le terme de *biais*, ou erreur d'approximation, que l'on commet en estimant s_m plutôt que s. Cette erreur est inconnue, puisque la densité est inconnue, et son traitement nécessite généralement des hypothèses supplémentaires sur s. Le second est le terme de *variance*, ou erreur d'estimation, et représente le prix à payer pour avoir remplacé s_m par $\widehat{s}_m$. Ces deux termes évoluent en général en sens opposé comme nous le verrons en Section 1.1.4. Lorsque nous avons une collection d'estimateurs $(\widehat{s}_m)_{m\in\mathcal{M}}$ dans $\mathbb{L}_2$, il est donc nécessaire, pour minimiser leurs erreurs quadratiques, de trouver un équilibre entre ces deux termes, connu sous le nom de *compromis biais-variance*.

Il convient toutefois de relativiser les conclusions qui pourraient être tirées des performances statistiques d'un estimateur relatif à cette perte en raison, précisément, de son caractère non-intrinsèque. Sans rentrer dans les détails[(11)], on peut rappeler que cette perte varie selon la mesure dominante (le fait même de savoir si $dP/d\mu \in \mathbb{L}_2(\mu)$ dépend de μ) et perd son sens si P n'est pas absolument continue par rapport à μ. De plus, il est souvent indispensable, pour contrôler le risque quadratique d'un estimateur donné, de rajouter des conditions supplémentaires sur la norme $\mathbb{L}_\infty$ de s (voir la Proposition 4 de Birgé (2006b) et la Proposition 3 de Birgé (2014)). Enfin, on peut ajouter qu'elle n'est pas invariante par des transformations monotones des axes de coordonnées (Devroye & Györfi, 1985).

Il existe un autre point de vue, relié aux deux précédents, qui est fort répandu dans la théorie statistique de l'apprentissage. On définit, suivant (Birgé & Massart, 1993, Definition 1), une **fonction de contraste** γ comme une application de $\mathcal{S}\times\Xi$ dans $\mathbb{R}$ telle que la fonction $t \mapsto \mathbb{E}[\gamma(t, X)]$ (où la variable aléatoire X est de loi P) soit minimale en s. Dans ce cas, une perte ℓ en t peut être définie par

$$\ell(s,t) = \mathbb{E}[\gamma(t,X) - \gamma(s,X)] \geqslant 0 \quad \text{pour tout } t \in \mathcal{S} \quad . \tag{1.3}$$

La perte $\mathbb{L}_2$ se déduit du choix $\mathcal{S} = \mathbb{L}_2(\mu)\cap\mathbb{L}_1(\mu)$ et $\gamma(t,x) = \|t\|_2^2 - 2t(x)$. La perte de Kullback-Leibler provient de la fonction de contraste $\gamma(t,x) = -\log(t(x))$ où $\mathcal{S}$ est l'ensemble des densités de probabilité par rapport à μ. De nombreux liens existent entre ces différentes solutions : on renvoie le lecteur intéressé aux travaux de Le Cam (1973, 1986) ainsi qu'au premier chapitre de Devroye (1987) pour une discussion plus poussée sur les relations entre celles-ci.

Puisque la perte de référence modifie le cadre du problème auquel on s'attaque, la thèse est subdivisée selon son choix : les deux premières parties se font dans le cadre des moindres carrés alors que la troisième partie se situe dans le cadre Hellinger.

[(11)] on pourra consulter, par exemple, la Section 6.5 dans (Devroye & Lugosi, 2001), "$\mathbb{L}_2$-distances are to be avoided", la Section 5.4.1 de Birgé (2006b), "Problems connected with the use of the $\mathbb{L}_2$-distance in density estimation", et la Section 1.3 de (Birgé, 2014), "Some negative results for the $\mathbb{L}_2$-loss".

1.1.3 Problème du choix d'une méthode d'estimation

En plus de l'échantillon $\mathbf{X}$, nous supposerons avoir à notre disposition des "méthodes d'estimation" capables de construire des candidats à l'estimation de s à l'aide des données. Celles-ci sont parfois appelées algorithmes statistiques (Arlot & Celisse, 2010), mais nous évitons cette dernière appellation pour ne pas créer de confusion avec les algorithmes (au sens classique) que nous introduirons dans la thèse.

Définition 1.1. *Une **méthode d'estimation** désigne toute application mesurable $\mathcal{A} : \bigcup_{k\geqslant 1} \Xi^k \to \mathcal{S}$ qui associe à tout échantillon aléatoire $\mathbf{Y}^k \in \Xi^k$ un estimateur $\widehat{s} = \mathcal{A}(\mathbf{Y}^k)$ de s.*

Nous mesurons naturellement la qualité d'une méthode d'estimation $\mathcal{A}$ par la perte $\ell(s, \mathcal{A}(\mathbf{X}))$ (resp. le risque $\mathbb{E}\left[\ell(s, \mathcal{A}(\mathbf{X}))\right]$) de l'estimateur qu'elle construit avec toutes les données disponibles. Comme auparavant, plus petite est la perte (resp. le risque) meilleure est la méthode d'estimation.

Inégalités d'oracle. Soit $(\mathcal{A}_m)_{m\in\mathcal{M}}$ une collection finie ou dénombrable de méthodes d'estimation. On note $\{\widehat{s}_m = \mathcal{A}_m(\mathbf{X}), m \in \mathcal{M}\}$ l'ensemble des estimateurs obtenus en allouant toutes les observations aux méthodes $(\mathcal{A}_m)_{m\in\mathcal{M}}$. Le but est de choisir $\widehat{m} = \widehat{m}(\mathbf{X}) \in \mathcal{M}$, à partir de l'échantillon $\mathbf{X}$, de sorte que l'estimateur correspondant soit, parmi tous les candidats, le plus proche de s. Ainsi, pour une perte ℓ donnée, le meilleur choix possible consiste à sélectionner *l'oracle*[12] $m^* \in \mathcal{M}$ qui vérifie $\ell(s, \widehat{s}_{m^*}) = \inf_{m\in\mathcal{M}} \ell(s, \widehat{s}_m)$. Malheureusement, comme la loi P est inconnue, la perte $\ell(s, \widehat{s}_m)$ l'est également et on ne peut sélectionner la *méthode d'estimation oracle* $\mathcal{A}_{m^*}$. On voudrait que le choix $\widehat{m}$ mène à un estimateur qui fasse aussi bien que l'oracle, à constante près. Autrement dit on voudrait que $\widehat{s}_{\widehat{m}}$ satisfasse, avec grande probabilité, une *inégalité oracle trajectorielle*

$$\ell(s, \widehat{s}_{\widehat{m}}) \leqslant C_n \inf_{m\in\mathcal{M}} \left\{ \ell(s, \widehat{s}_m) \right\} + R_n \ , \tag{1.4}$$

où $C_n \geq 1$ est la constante dominante (indépendante de P) et $R_n \geq 0$ est le terme de reste. Ces quantités déterministes doivent être de taille raisonnable (par exemple C_n plus petit qu'une puissance de $\log n$, et R_n petit par rapport à $\inf_{m\in\mathcal{M}} \ell(s, \widehat{s}_m)$) pour s'assurer que la perte de $\widehat{s}_{\widehat{m}}$ soit aussi petite que celle de l'oracle. Ainsi, une inégalité oracle présente *de facto* un caractère non-asymptotique au sens où elle permet de comparer les performances d'un choix $\widehat{m} \in \mathcal{M}$ à un estimateur idéal, et ce, quelque soit le nombre d'observations. Ce point de vue non-asymptotique permet également d'observer le comportement du risque de l'estimateur pour $n \to \infty$. En effet, si on trouve, dans ce cas, que R_n est négligeable devant la perte oracle $\inf_{m\in\mathcal{M}} \ell(s, \widehat{s}_m)$ et $C_n \to 1$[13] uniformément par rapport à P, alors on peut en déduire une inégalité oracle asymptotiquement optimale, ou *inégalité oracle optimale au premier ordre*.

L'oracle peut aussi être défini comme le meilleur estimateur choisi de manière déterministe, c'est-à-dire $\operatorname{argmin}_{m\in\mathcal{M}} \mathbb{E}\left[\ell(s, \widehat{s}_m)\right]$. Dans ce cas, on souhaite montrer une *inégalité oracle en espérance* pour $\widehat{s}_{\widehat{m}}$, c'est-à-dire

$$\mathbb{E}\left[\ell(s, \widehat{s}_{\widehat{m}})\right] \leqslant C_n \inf_{m\in\mathcal{M}} \left\{ \mathbb{E}\left[\ell(s, \widehat{s}_m)\right] \right\} + R_n \ . \tag{1.5}$$

[12] appellation d'origine contrôlée, introduite par Donoho & Johnstone (1994).

[13] le plus souvent on cherche à écrire $C_n = 1 + \delta_n$, avec $\delta_n \to 0$ quand $n \to \infty$.

Celle-ci est plus faible que la précédente puisque (1.4) implique (1.5) dès que la perte est bornée. Ajoutons, pour être complet, que dans un cadre asymptotique on peut montrer *l'optimalité asymptotique* de l'estimateur $\widehat{s}_{\widehat{m}}$ en prouvant que

$$\mathbb{P}\left(\frac{\ell\left(s, \widehat{s}_{\widehat{m}}\right)}{\inf_{m \in \mathcal{M}} \ell\left(s, \widehat{s}_m\right)} \xrightarrow[n\to\infty]{} 1 \right) = 1 \ .$$

Dans le cas le plus général, les méthodes d'estimation sont inconnues (cas des "boîtes noires") et nous pouvons les juger uniquement en observant ce qu'elles fabriquent (en "output") avec des sous-échantillons de $\mathbf{X}$ (en "input"). Dans cette thèse, nous allons nous restreindre à un type particulier de méthodes d'estimation.

Définition 1.2. *Une **méthode d'estimation à noyau** associe à tout échantillon $\mathbf{Y}^k \in \Xi^k$ un **estimateur linéaire** de s, c'est-à-dire un estimateur de la forme*

$$\widehat{s}(x) = \mathcal{A}\left(\mathbf{Y}^k\right)(x) = \frac{1}{k} \sum_{X_i \in \mathbf{Y}^k} \mathcal{K}(x, X_i) \qquad \textit{pour tout } x \in \Xi \tag{1.6}$$

*pour une fonction $\mathcal{K} : \Xi \times \Xi \mapsto \mathbb{R}$ symétrique, qu'on appellera **noyau** dans la suite.*

Méthode d'estimation ou estimateur, quelle différence ? Par cette hypothèse, nous disposons donc du "secret de fabrication" des méthodes, puisque nous savons quel type d'estimateur elles construisent. Dans notre cadre, choisir une méthode d'estimation à noyau dans $(\mathcal{A}_m)_{m\in\mathcal{M}}$ revient exactement à choisir un noyau dans $(\mathcal{K}_m)_{m\in\mathcal{M}}$, ou encore un estimateur linéaire dans $(\widehat{s}_m)_{m\in\mathcal{M}}$ (puisque tout estimateur linéaire est défini par son noyau uniquement). Ainsi, le problème du choix d'un candidat $\widehat{m} \in \mathcal{M}$ dans chacune de ces familles constitue *un seul et même problème* dont une solution mène dans tous les cas à un estimateur $\widehat{s}_{\widehat{m}}$ de s de la forme (1.6).

La distinction que nous faisons se situe dans l'utilisation de techniques de *rééchantillonnage* de l'échantillon $\mathbf{X}$ qui permettent de construire de nouveaux échantillons à partir de $\mathbf{X}$ (de même taille, ou de taille plus petite). Comme nous le mettrons en évidence en Section 1.2, la propriété qui nous intéresse particulièrement dans la définition d'une méthode d'estimation est la suivante : à chaque sous-échantillon de $\mathbf{X}$ la méthode d'estimation nous renvoie un estimateur de s, auquel nous avons *accès*. Nous pouvons dès lors allouer une partie des données afin de construire un estimateur, pour ensuite le juger sur de *nouvelles données*. Cette nuance est subtile mais elle a son importance : nous ne sommes pas contraints à réutiliser les *mêmes données* pour construire un estimateur et évaluer sa qualité. En résumé, la notion de méthode d'estimation apporte une souplesse et une précision sur l'utilisation des données par le statisticien. Voici la règle que nous respecterons concernant la terminologie.

- Si nous n'utilisons pas le rééchantillonnage, alors la notion de méthode d'estimation est obsolète et nous parlons du *problème de la sélection d'un estimateur*.
- Si nous utilisons le rééchantillonnage, nous parlons du *problème du choix d'une méthode d'estimation*.

- Enfin, le *problème du choix du noyau* se ramène à l'un des deux cas précédents (toujours selon que l'on utilise ou pas le rééchantillonnage).

1.1.4 Estimateurs linéaires

Nous rappelons ici les définitions des estimateurs linéaires les plus connus, que nous rencontrerons dans chaque chapitre, et qui serviront à illustrer nos résultats théoriques sur des simulations : les estimateurs par projection et les estimateurs de Parzen-Rosenblatt. Il existe une littérature abondante sur ces estimateurs qui font, à notre connaissance, leur première apparition sous cette forme dans Whittle (1958). Pour la perte des moindres carrés, les premiers résultats théoriques sur leurs propriétés asymptotiques ont été prouvés par Watson & Leadbetter (1964a,b), suivis de Winter (1975) et Walter & Blum (1979) (ceux-ci les nomment *"delta-sequence estimators"*) qui ont établi des vitesses d'estimation. Ils ont été introduits dans le contexte de la validation croisée par Rudemo (1982) et utilisés par Marron (1987) pour comparer ce type de techniques. Dans leur livre dédié à l'estimation d'une densité avec la distance en variation, Devroye & Lugosi (2001) considèrent ce même type d'estimateurs (ils les nomment *estimateurs additifs*). Enfin, le terme "estimateur linéaire" que nous adoptons apparaît, sous cette même forme, dans (Goldenshluger & Lepski, 2011, Section 2.6).

Estimateur de Parzen-Rosenblatt. Considérons ici la perte $\ell = D$, l'espace $\Xi = \mathbb{R}$ muni de la tribu $\mathcal{Z}$ de l'ensemble des boréliens sur $\mathbb{R}$[(14)] et μ la mesure de Lebesgue[(15)]. Si on note $P_n := n^{-1} \sum_{i=1}^n \delta_{X_i}$ la mesure empirique de l'échantillon, nous avons pour toute loi P de densité s par rapport à μ, $D(P, P_n) = 1$, de sorte qu'une densité ne peut être approchée dans $\mathbb{L}_1$ par la mesure empirique. Il est donc nécessaire que l'approximation de la vraie loi possède également une densité, c'est l'idée de l'estimateur de Parzen-Rosenblatt (du nom des mathématiciens qui l'ont introduit : Rosenblatt (1956) et Parzen (1962)). Étant donné $m = (k, h)$, où k est une fonction symétrique, intégrable sur $\mathbb{R}$[(16)], et $h > 0$ une *fenêtre*, on note pour tout $x \in \mathbb{R}$, $k_h(x) := h^{-1}k(x/h)$. La famille $(k_h)_{h>0}$ forme une *approximation de l'unité* pour le produit de convolution et nous avons le résultat suivant (Devroye & Lugosi, 2001, Theorem 9.1)

$$\lim_{h \to 0} \int \left| s \star k_h - s \int k \right| = 0 \ ,$$

où $k_h \star s(x) := \int_{\mathbb{R}} k_h(x - y)s(y)dy = \mathbb{E}\left[k_h(x - X)\right]$. Si on suppose $\int k = 1$, on peut donc approcher s par $k_h \star s$ dans $\mathbb{L}_1$[(17)]. L'estimateur de Parzen-Rosenblatt de la densité s estime sans

[(14)] le cas $\mathbb{R}^d$ se traite de la même façon.

[(15)] Soulignons que cet estimateur est spécifique à cette mesure au sens où son traitement nécessite que l'on considère des densités par rapport à celle-ci.

[(16)] Celle-ci est souvent appelée *noyau* dans la littérature. Nous évitons cette terminologie pour ne pas créer de confusion avec les méthodes d'estimation à noyau.

[(17)] en particulier, $k_h \star s$ converge vers s, quand h tend vers 0 en norme $\mathbb{L}_1$ et d'autant plus rapidement que s est régulière.

biais la quantité $k_h \star s(x)$. Il s'écrit, pour tout $x \in \mathbb{R}$,

$$\widehat{s}_m(x) = \frac{1}{n}\sum_{i=1}^{n} k_h(x - X_i) \ . \tag{1.7}$$

C'est donc un estimateur linéaire où le **noyau d'approximation** $\mathcal{K}_m = \mathcal{K}_{(k,h)}$ est défini pour tout $x, y \in \mathbb{R}$ par $\mathcal{K}_{(k,h)}(x, y) = k_h(x - y)$.

Pour la perte quadratique, des travaux de nature asymptotique ont montré que, parmi les fonctions k positive, la fonction $k(x) = \max(3/4(1 - x^2), 0)$[(18)] est optimale dans $\mathbb{R}$ (Watson & Leadbetter, 1963) et $k(x) = \max(1, (1 - \|x\|^2)^d)$ dans $\mathbb{R}^d$ (Deheuvels, 1977). Aussi la plupart des auteurs fixent la fonction k (dans ce cas nous notons $m = h$) pour se concentrer sur un seul hyperparamètre, la fenêtre ("*bandwidth*" en anglais), dont le choix est crucial pour assurer la qualité de $\widehat{s}_{\widehat{m}} = \widehat{s}_{\widehat{h}}$. Voir la Section 1.2.4 de Tsybakov (2009) pour une discussion sur le choix de k.

Estimateur par projection. La méthode d'estimation par projection repose fortement sur l'hypothèse selon laquelle la fonction s appartient à $\mathbb{L}_2$. Dans ce cadre on notera, pour toute fonction $f : \Xi \to \mathbb{R}$, $P_n(f) = n^{-1}\sum_{i=1}^{n} f(X_i)$ et $P(f) = \mathbb{E}\,[f(X)]$. On pose $\mathcal{S} = \mathbb{L}_2(\mu) \cap \mathbb{L}_1(\mu)$ et on se donne comme modèle S un sous-espace linéaire de $\mathcal{S}$ (par exemple un sous-espace vectoriel de dimension $D > 0$ finie). Le candidat le plus proche de s dans S, ici la projection de s sur S, s'écrit pour une base orthonormée $(\psi_\lambda)_{\lambda\in\Lambda}$ de S, $s_S = \sum_{\lambda\in\Lambda} P\,(\psi_\lambda)\,\psi_\lambda$.

L'estimateur par projection estime sans biais s_S. Il s'écrit, pour tout $x \in \Xi$,

$$\widehat{s}_S(x) = \sum_{\lambda\in\Lambda}\left(\frac{1}{n}\sum_{i=1}^{n}\psi_\lambda(X_i)\right)\psi_\lambda(x) \ . \tag{1.8}$$

Cette idée peut se traduire en utilisant une fonction de contraste. Il suit en effet de la définition de la perte $\mathbb{L}_2$ par la relation (1.3) que minimiser $\ell(s, t)$ pour $t \in S$ revient à chercher le minimum de $t \mapsto \mathbb{E}\,[\gamma(t, X)]$ sur S. Puisque cette quantité est inconnue, on la remplace par sa version empirique sans biais, le *contraste empirique des moindres carrés* $\gamma_n(t) := \|t\|_2^2 - 2P_n(t)$ de sorte que $\widehat{s}_S = \mathrm{argmin}_{t\in S}\{\gamma_n(t)\}$. Celui-ci est l'estimateur linéaire associé au **noyau de projection** $\mathcal{K}_m = \mathcal{K}_S$, défini pour tout $(x, y) \in \Xi^2$ par $\mathcal{K}_S(x, y) = \sum_{\lambda\in\Lambda}\psi_\lambda(x)\psi_\lambda(y)$. En particulier *l'estimateur histogramme* est également un estimateur linéaire si on prend S l'espace des fonctions constantes par morceau sur une partition donnée $\{I_\lambda,\ \lambda \in \Lambda\}$ de Ξ. Dans ce cas, l'ensemble $\{\psi_\lambda = (\mu(I_\lambda))^{-1/2}\mathbb{1}_{I_\lambda},\ \lambda \in \Lambda\}$ est une base orthonormée de S et $\mathcal{K}_S$ est le **noyau histogramme**.

Il existe une généralisation des estimateurs par projection, les *estimateurs par projection avec poids*. On pose, pour tout $n \in \mathbb{N}^\star = \mathbb{N}\backslash\{0\}$, $[n] := \{1, \ldots, n\}$. Soit $(\varphi_j)_{j\in[p]}$ un système orthonormal dans $\mathbb{L}_2$ où $p \in \mathbb{N}^\star \cup \{\infty\}$ et soit un vecteur de poids $w = (w_1, \ldots, w_p) \in [0, 1]^p$ avec

[(18)] appelé noyau Epanechnikov.

$\sum_{i=1}^{p} w_i > 0$. L'estimateur par projection avec poids est défini par

$$\widehat{s}_w(x) = \sum_{i=1}^{p} w_i \left(\frac{1}{n} \sum_{j=1}^{n} \varphi_i(X_j) \right) \varphi_i(x) \ . \tag{1.9}$$

L'estimateur par projection se retrouve naturellement en posant $w_i = 1$ pour tout $i \in [p]$. Un exemple moins trivial de ce type d'estimateur est l'estimateur de Pinsker (1980)[(19)] qui possède de bonnes propriétés pour estimer des fonctions dans les classes de Sobolev (Dalelane, 2005a,b). C'est donc un estimateur linéaire, où le noyau $\mathcal{K}_m = \mathcal{K}_w$ est le **noyau de projection avec poids** w qui s'écrit pour tout $(x, y) \in \Xi^2$, $\mathcal{K}_w(x, y) = \sum_{i=1}^{p} w_i \varphi_i(x)\varphi_i(y)$.

Enfin, les *estimateurs par série orthogonale* (Walter & Blum, 1979) sont aussi des estimateurs linéaires. Étant donnée une fonction de poids positive w et un ensemble de fonctions $\{\psi_\lambda,\ \lambda \in \Lambda\}$ qui est orthonormal et complet par rapport au produit scalaire

$$\langle \psi_\lambda, \psi_{\lambda'} \rangle = \int \psi_\lambda(x)\psi_{\lambda'}(x)w(x)d\mu(x) \ ,$$

les estimateurs par série orthogonale sont définis en posant, pour tout $(x, y) \in \Xi^2$, $\mathcal{K}(x, y) = \sum_{\lambda\in\Lambda} \psi_\lambda(x)\psi_\lambda(y)w(x)$. Cet exemple inclut les fonctions trigonométriques, les polynômes de Legendre et les fonctions Hermitiennes.

Retour au compromis biais-variance. Afin de comprendre les enjeux de la sélection d'un noyau, revenons au compromis biais-variance et étudions le risque quadratique des estimateurs par projection et de Parzen-Rosenblatt (pour k fixé). Par le Théorème de Pythagore, et en utilisant $\mathbb{E}\left[\widehat{s}_h(x)\right] = k_h \star s(x)$, on trouve par (1.2)

$$\mathbb{E}\left[\|\widehat{s}_S - s\|_2^2 \right] = \|s - s_S\|_2^2 + \mathbb{E}\left[\|\widehat{s}_S - s_S\|_2^2 \right] \ , \tag{1.10}$$

$$\mathbb{E}\left[\|\widehat{s}_h - s\|_2^2 \right] = \|s - k_h \star s\|_2^2 + \mathbb{E}\left[\|\widehat{s}_h - k_h \star s\|_2^2 \right] \ . \tag{1.11}$$

Les deux termes dans (1.10) évoluent de façon opposée en fonction de la taille du modèle. Plus le modèle est gros, mieux s est approchée par sa projection s_{S} et le biais décroît. Mais, dans le même temps, le terme de variance augmente car il devient difficile d'estimer correctement s_S dans un modèle de grande dimension à cause du grand nombre de paramètres qui doivent être estimés. Le constat est identique pour l'estimateur de Parzen-Rosenblatt avec k fixé. Dans (1.11), le terme de biais est d'autant plus petit que la fenêtre h est petite (puisque $k_h \star s$ tend vers s en norme $\mathbb{L}_2$ quand h tend vers 0) et le terme de variance est, à l'opposé, croissant quand h tend vers 0[(20)]. Une bonne procédure de sélection d'estimateurs $\widehat{m}$ doit donc réaliser le meilleur compromis entre ces deux termes. Nous reviendrons sur cette difficulté en Section 1.2.2.

[(19)] Celui-ci a d'abord été introduit dans le cadre de la régression gaussienne, pour être ensuite considéré dans le cadre de l'estimation d'une densité par Efroĭmovich (1985). On peut trouver une vue d'ensemble de cette théorie dite de "Pinsker-Efromovich" dans une série d'articles plus récents Efromovich (2000, 2005, 2008).

[(20)] Si on met en parallèle (1.10) et (1.11), on observe que la fenêtre $1/h$ joue pour les estimateurs de Parzen-Rosenblatt le rôle de la dimension du modèle qui sert à définir les estimateurs par projection.

1.1.5 Liens avec d'autres problèmes statistiques

Le principal but de cette thèse est donc de construire, à partir des observations, une *procédure* $\mathcal{C} : \mathcal{M} \to \mathbb{R}$ telle que, si on note

$$\widehat{m}(\mathcal{C}) \in \operatorname*{argmin}_{m \in \mathcal{M}} \mathcal{C}(m) \ , \tag{1.12}$$

alors l'estimateur $\widehat{s}_{\widehat{m}(\mathcal{C})}$ vérifie une inégalité oracle (si possible optimale au premier ordre). Nous utilisons systématiquement la même terminologie, de sorte qu'une procédure désigne uniquement une manière de choisir une méthode d'estimation (ou un estimateur).

Relions à présent notre cadre à quelques problèmes statistiques connus, selon que nous utilisons le rééchantillonnage des données ou pas. Le problème du choix du noyau est utilisé comme terme générique lorsque nous voulons désigner le problème général sans préciser la contrainte d'accès aux données (voir la synthèse dans la Table 1.1). Certains problèmes sont commentés dans le texte ci-dessous. Nous rappelons aussi dans cette section d'autres problèmes proches ou reliés au nôtre, l'agrégation et l'estimation adaptative.

	Rééchantillonnage	
	Avec	Sans
Pas de noyau **(I)**	Méthode d'estimation	Estimateur
Noyau quelconque **(II)**	Méthode d'estimation à noyau	Estimateur linéaire
Noyau d'approximation **(III)**	Méthode d'estimation à noyau d'approximation	Estimateur de Parzen-Rosenblatt
Noyau de projection **(IV)**	Méthode d'estimation à noyau de projection	Estimateur par projection
Noyau histogramme **(V)**	Méthode d'estimation à noyau histogramme	Estimateur histogramme

TABLE 1.1 – Connexions avec d'autres problèmes de sélection.

Les Problèmes **(III)**, **(IV)** et **(V)** se ramènent tous à la sélection d'un hyperparamètre (fenêtre, modèle et partition). Ce sont des cas particuliers du Problème **(II)** qui provient lui-même du Problème **(I)**.

- **(I)** Cadre le plus général où aucune hypothèse n'est faite a priori sur la nature des estimateurs. La *sélection d'un estimateur* (voir par exemple Baraud (2011)) est l'équivalent du problème général du *choix d'une méthode d'estimation* présenté en Section 1.1.3, en n'utilisant pas de rééchantillonnage. L'objectif est de construire à partir des mêmes données une procédure qui sélectionne un candidat dans la collection d'estimateurs $(\widehat{s}_m(\mathbf{X}))_{m \in \mathcal{M}}$ qui vérifie une inégalité oracle (comme (1.4) ou (1.5)).

- **(II)** *Choix du noyau* : c'est le principal problème considéré dans cette thèse. Nous traiterons les deux cas correspondants, *sélection d'un estimateur linéaire* (Goldenshluger & Lepski, 2011, Section 2.6) et *sélection d'une méthode d'estimation à noyau* (Rudemo, 1982).

- **(III)** C'est l'équivalent du problème de la *sélection de la fenêtre* (voir le livre de Tsybakov (2009)). Soit $\mathcal{H} = \{h_m, m \in \mathcal{M}\}$ une collection finie de fenêtres h_m. On fixe k et on associe, grâce au noyau d'approximation, à chaque fenêtre dans $\mathcal{H}$ l'estimateur Parzen-Rosenblatt correspondant. Le problème consiste alors à choisir la "meilleure" fenêtre dans $\mathcal{H}$, c'est-à-dire celle qui réalise le meilleur compromis biais-variance.

- **(IV)** C'est l'équivalent de la *sélection de modèles*[(21)] (voir le livre de Massart (2007)). Si on dispose d'une collection de modèles $(S_m)_{m\in\mathcal{M}}$ dans $\mathcal{S}$, on peut associer à chacun d'eux, grâce au noyau de projection, un estimateur par projection. Le problème de la sélection de modèles consiste donc à sélectionner à partir des observations le "meilleur modèle"[(22)] dans la famille. Nous verrons à la Section 1.2 que la grande difficulté d'un tel choix, encore illustrée par le compromis biais-variance, peut être affrontée par différentes techniques selon la façon dont les données sont utilisées.

- **(V)** C'est l'équivalent du problème de la *sélection d'une partition*. C'est un cas particulier du Problème **(IV)**, où les modèles sont des espaces de fonctions constantes par morceau. La plupart des auteurs (Rudemo, 1982; Daly, 1988; Hall, 1990; Wand, 1997) s'attaquent à ce problème en utilisant l'idée de rééchantillonnage. D'autres (Castellan, 2000; Birgé & Rozenholc, 2006) suggèrent des solutions sans diviser les données.

Agrégation. Il existe d'autres procédures pour estimer s à partir d'une collection d'estimateurs $(\widehat{s}_m(\mathbf{X}))_{m\in\mathcal{M}}$. Plutôt que de sélectionner un candidat, on peut aussi agréger les estimateurs en leur donnant plus ou moins d'importance grâce à une suite de poids $(w_m = w_m(\mathbf{X}))_{m\in\mathcal{M}}$ de somme 1 que le statisticien détermine à partir des données. Cette procédure s'appelle *l'agrégation* (une bonne référence générale sur le sujet est le livre de Nemirovski (2000)) et mène à l'estimateur final $\widetilde{s} = \sum_{m\in\mathcal{M}} w_m \widehat{s}_m$. L'objectif est sensiblement différent puisqu'il s'agit de construire un estimateur dont le risque soit similaire à la meilleure combinaison (convexe ou linéaire) des estimateurs. Des questions similaires à celles précédemment évoquées se posent alors : comment construire une inégalité oracle avec constante C_n proche de 1, quelles sont les vitesses d'agrégation optimales (voir Tsybakov (2003)), etc. Il faut être très précautionneux avec le véritable sens de la constante C_n dans ce cadre. Ainsi, certains auteurs (Rigollet & Tsybakov, 2007; Lecué, 2007) obtiennent la constante 1 pour de nombreuses procédures d'agrégation. Cependant, celles-ci supposent que les estimateurs sont construits sur un échantillon indépendant de sorte que l'estimateur final (construit, lui, avec toutes les données) est comparé à l'oracle sur les estimateurs qui reposent sur une partie des données uniquement. De manière générale, rien ne garantit que ces procédures avec $C_n = 1$ sont préférables aux autres en pratique.

Il existe une littérature riche sur les différents types d'agrégation (convexe, linéaire ou de type sélection de modèles). En densité, voir notamment Catoni (1999), Yang (2000), Devroye & Lugosi

[(21)] introduite par Akaike (1973) et Mallows (1973), ensuite formalisée par Birgé & Massart (1997) et Barron *et al.* (1999).

[(22)] Nous avons, depuis le départ, comme objectif la prédiction. On peut aussi supposer que $s \in \bigcup_{m\in\mathcal{M}} S_m$ et vouloir *identifier* le vrai modèle m_{ident} : le but est alors de proposer une procédure $\widehat{m}$ telle que la probabilité $\mathbb{P}\left(\widehat{m} = m_{\mathrm{ident}}\right)$ soit maximale.

(2001), Birgé (2006a), Rigollet & Tsybakov (2007), Samarov & Tsybakov (2004) et Bunea *et al.* (2010).

Inégalité oracle et adaptation. Lorsque l'oracle possède des propriétés statistiques intéressantes, l'inégalité oracle permet de les transmettre à l'estimateur sélectionné. En particulier elle peut servir à construire un estimateur adaptatif à la régularité de la densité, c'est-à-dire un estimateur qui approche la densité "aussi bien" qu'un estimateur qui connaitrait la vraie régularité de s (Birgé & Massart, 1997; Barron *et al.*, 1999).

Étant donnée une famille $\mathcal{F}$ de lois de probabilité, le **risque minimax** de $\mathcal{F}$ est défini par

$$\mathcal{R}_{\text{minimax}}(\mathcal{F}) := \inf_{\widetilde{s}} \sup_{P \in \mathcal{F}} \mathbb{E}\left[\ell\left(s, \widetilde{s}\right)\right] \ ,$$

où l'infimum est pris sur tous les estimateurs[23]. Un **estimateur minimax** pour $\mathcal{F}$ est simplement un estimateur $\widehat{s}$ tel que $\mathbb{E}\left[\ell(s, \widehat{s})\right] \leq L\mathcal{R}_{\text{minimax}}(\mathcal{F})$ pour une constante $L < \infty$ indépendante de n. Ce risque mesure donc le pire cas possible pour la classe $\mathcal{F}$, de sorte qu'un estimateur minimax pour une très grande classe $\mathcal{F}$ n'est pas forcément un "bon" estimateur de s.

Souvent, on suppose que la densité s possède une certaine régularité α (inconnue) et on écrit $s \in \mathcal{F}_\alpha$. Les exemples de référence classiquement utilisés sont les espaces de Besov, de Hölder, de Sobolev ou encore de Nikol'ski (voir leurs définitions dans DeVore & Lorentz (1993) ou Tsybakov (2009)). Des estimateurs minimax sont "connus" pour une grande diversité de fonctions de perte et de classes de fonctions[24]. Cependant, la plupart du temps, ces estimateurs dépendent précisément du paramètre α et s'avèrent inutilisables en pratique. On suppose alors que s appartient à une classe de fonctions $\mathcal{F} = \bigcup_{\alpha \in \aleph} \mathcal{F}_\alpha$ pour un ensemble $\aleph$ donné, c'est-à-dire qu'il existe un $\alpha_0 \in \aleph$ tel que $s \in \mathcal{F}_{\alpha_0}$. Le but de l'adaptativité est de construire un estimateur $\widehat{s}$ qui n'utilise pas la connaissance de α_0 et qui soit aussi bon que tout estimateur qui connaitrait α_0. Un estimateur est **adaptatif au sens du minimax** pour $\mathcal{F}$ s'il est simultanément minimax sur chaque $\mathcal{F}_\alpha$, c'est-à-dire que pour tout $\alpha \in \aleph$ il existe une constante C_α indépendante de P telle que

$$\sup_{P \in \mathcal{F}_\alpha} \mathbb{E}\left[\ell\left(s, \widehat{s}\right)\right] \leq C_\alpha \mathcal{R}_{\text{minimax}}(\mathcal{F}_\alpha) \ .$$

En sélection de modèles, Birgé & Massart (1997) ont montré que si on dispose d'un estimateur $\widehat{s}_{\widehat{m}}$ qui satisfait une inégalité oracle du type (1.5), on peut prouver qu'il est adaptatif au paramètre α en choisissant une famille de modèles $\{S_m, m \in \mathcal{M}\}$ de sorte que chaque $\mathcal{F}_\alpha$ puisse être bien approché par un modèle dans la famille. De manière générale, les inégalités d'oracle permettent de construire des estimateurs adaptatifs (voir par exemple la Section 3.3 dans l'Introduction de la thèse de Rigollet (2006b)). Ainsi, de nombreuses procédures énoncées ci-dessus permettent d'obtenir des estimateurs adaptatifs, notamment les procédures de sélection de modèles et les procédures d'agrégation d'estimateurs. Citons aussi les méthodes dites de Efroimovich-Pinsker

[23] Le plus souvent on suppose que $\mathcal{F}$ est une famille dominée par la mesure μ. Ainsi, le risque minimax est défini pour la classe de fonctions associées, les estimateurs appartiennent à $\mathcal{S}$ et l'infimum est pris sur les estimateurs $\widetilde{s}$ de s.

[24] Par exemple, les estimateurs par projection et les estimateurs Parzen-Rosenblatt sont minimax pour les classes de Hölder et de Nikol'ski, l'estimateur de Pinsker est minimax pour la classe de Sobolev, etc.

(initialement présentée dans Efroĭmovich & Pinsker (1984) et actualisée dans Efromovich (2008)), de Lepski (dont l'originale est présentée dans les papiers (Lepskiĭ, 1991, 1992a,b)) et ses variantes (voir par exemple Goldenshluger & Lepski (2014) pour le problème de l'adaptation en densité et les résultats connus sur les vitesses de convergence pour la norme $\mathbb{L}_p$ dans les espace de Besov) et le seuillage de coefficients en base d'ondelettes (voir Donoho & Johnstone (1994), Donoho *et al.* (1995) ou Reynaud-Bouret *et al.* (2011)). Notons que la plupart de ces résultats adaptatifs concernent la perte quadratique. Enfin, il est parfois nécessaire de payer un prix dans la borne de droite pour l'adaptation (par exemple, une puissance de $\log n$ pour la norme $\mathbb{L}_\infty$).

Remarque. Si nous nous focalisons sur l'adaptation à la régularité de la densité, ce terme peut prendre diverses significations et mener à des objectifs bien différents dans d'autres cadres statistiques (par exemple l'adaptation au niveau de bruit dans un cadre de régression hétéroscédastique ou à la condition de marge en classification).

1.1.6 Comparaison de procédures

Le deuxième objectif de cette thèse, qui occupe une place nettement moins importante dans ce manuscrit, consiste à franchir un pas supplémentaire dans la comparaison de procédures[(25)] de sélection d'estimateurs.

Au vu de ce qui précède, les inégalités d'oracle démontrent (au moins) un double intérêt : elles garantissent que le candidat sélectionné est le meilleur possible dans une famille donnée, à constante et terme de reste près, et constituent un outil pratique pour obtenir des résultats adaptatifs. Elles forment par conséquent un *graal* à atteindre pour toute procédure en quête de légitimité. Vu le nombre croissant[(26)] de procédures satisfaisant de telles inégalités, une question primordiale s'impose sur la garantie *effective* qu'offrent celles-ci : avoir une inégalité oracle théorique implique-t-il forcément une bonne procédure pratique ? De plus, parmi deux procédures optimales, sait-on dire théoriquement laquelle est la meilleure ?

Supposons que nous faisons face à un des problèmes de la Table 1.1 et que nous ayons pu prouver pour deux procédures distinctes $\mathcal{C}_1$ et $\mathcal{C}_2$ des inégalités d'oracle avec grande probabilité. Supposons aussi que ces deux procédures reposent sur une estimation du risque de chaque méthode d'estimation[(27)]. Comment savoir laquelle des deux procédures est la meilleure ? La comparaison des bornes obtenues sur $\mathbb{F}_n(\widehat{m}(\mathcal{C}_1))$ et $\mathbb{F}_n(\widehat{m}(\mathcal{C}_2))$, où

$$\mathbb{F}_n(\widehat{m}) := \frac{\ell\left(s, \widehat{s}_{\widehat{m}}\right)}{\inf_{m\in\mathcal{M}} \ell\left(s, \widehat{s}_m\right)} \ ,$$

ne suffit pas à déterminer un gagnant. D'abord parce que ce ne sont que des majorants, ensuite parce que les constantes C_n (parfois gigantesques) qui apparaissent ne reflètent pas la réalité. Par ailleurs, la limite $\lim_{n\to\infty} \mathbb{F}_n(\widehat{m}(\mathcal{C}_1))/\mathbb{F}_n(\widehat{m}(\mathcal{C}_2))$ dépend dans certains cas du biais de $\mathcal{C}_1(m)$ et $\mathcal{C}_2(m)$ comme estimateurs de $\mathbb{E}\left[\ell(s, \widehat{s}_m)\right]$, de sorte qu'on ne peut en tirer aucune information si les procédures ont même biais (ce qui est le cas pour beaucoup de procédures rééchantillonnées

[(25)] plus spécifiquement des procédures par rééchantillonnage.

[(26)] en densité, voir par exemple Rigollet (2006a), Massart (2007), Bunea *et al.* (2010), Goldenshluger & Lepski (2011) et Birgé (2014).

[(27)] Cette situation est très fréquente, mais d'autres possibilités existent comme nous le verrons dans la Section 1.2.1.

comme nous le verrons en Section 1.2.4). Ainsi, alors que l'on s'évertue à prouver des inégalités d'oracle pour toutes les procédures connues, on est encore loin de pouvoir dire au praticien laquelle est la meilleure. Pire, il semblerait que certaines procédures dites "optimales" du point de vue de l'oracle, ne fonctionnent pas bien en pratique ! En partant de ces constats, et pour tenter de répondre (partiellement) à cette question, Arlot & Lerasle (2014) proposent une heuristique pour aller au-delà du premier ordre, c'est-à-dire de l'objectif d'obtenir des inégalités d'oracle, pour s'assurer de la qualité d'une procédure. Nous la décrivons très brièvement ici.

Leur point de départ est le suivant. Si on pose $\overline{m} = \operatorname{argmin}_{m\in\mathcal{M}} \mathbb{E}\left[\ell(s,\widehat{s}_m)\right]$, alors plus la loi de $\widehat{m}(\mathcal{C})$ est concentrée autour de $\overline{m}$, meilleur est l'estimateur $\widehat{s}_{\widehat{m}(\mathcal{C})}$. Ce qui signifie qu'une procédure $\mathcal{C}$ est bonne si la probabilité $\mathbb{P}\left(m = \widehat{m}(\mathcal{C})\right)$ est petite pour tout $m \neq \overline{m}$. Ils soulignent que ceci paraît en tout cas souhaitable pour les $m \in \mathcal{M}$ dont le risque $\mathbb{E}\left[\ell(s,\widehat{s}_m)\right]$ est beaucoup plus grand que $\mathbb{E}\left[\ell(s,\widehat{s}_{\overline{m}})\right]$. En effet, la performance de $\widehat{m}(\mathcal{C})$ est jugée mauvaise si un tel candidat est choisi, alors qu'elle n'est pas trop détériorée si un $m \in \mathcal{M}$ proche de $\overline{m}$ en terme de risque est sélectionné.

Ils définissent ensuite pour tout $m, m' \in \mathcal{M}$, $\Delta_{\mathcal{C}}(m,m') = \mathcal{C}(m) - \mathcal{C}(m')$ et font l'approximation[28] suivante. Pour tout $m \in \mathcal{M}$, on a

$$\mathbb{P}\left(\widehat{m}(\mathcal{C}) = m\right) \approx \varphi\left(\mathrm{SNR}_{\mathcal{C}}(m)\right) \qquad \text{où} \quad \mathrm{SNR}_{\mathcal{C}}(m) := \max_{m'\neq m} \frac{\mathbb{E}\left[\Delta_{\mathcal{C}}(m,m')\right]}{\sqrt{\mathrm{Var}\left[\Delta_{\mathcal{C}}(m,m')\right]}} \ ,$$

et $1-\varphi(t) = \mathbb{P}\left(\mathcal{N}(0,1) \leq t\right)$ est la fonction de répartition d'une variable gaussienne centrée réduite. Ceci les mène à écrire l'implication suivante

$$\text{si} \qquad \mathrm{SNR}_{\mathcal{C}_1}(m) > \mathrm{SNR}_{\mathcal{C}_2}(m) \quad \forall m \neq \overline{m}, \quad \text{alors} \qquad \mathcal{C}_1 \text{ est meilleure que } \mathcal{C}_2 \ . \tag{1.13}$$

Arlot & Lerasle (2014) montrent empiriquement que cette heuristique est vérifiée pour des procédures de rééchantillonnage (qui seront présentées dans la suite) sur une collection d'estimateurs histogrammes. Du point de vue théorique, ils mettent en évidence le rôle clé que joue $\mathrm{Var}\left[\Delta_{\mathcal{C}}(m,m')\right]$ pour comparer des procédures ayant même biais. Leur travail est un premier pas dans la comparaison de procédures qui appelle à dépasser la "simple" recherche d'inégalités d'oracle et jette un nouveau regard sur l'évaluation d'une procédure de sélection d'estimateurs. Nous étudierons dans le Chapitre 4 le comportement de $\mathrm{Var}\left[\Delta_{\mathcal{C}}(m,m')\right]$ pour des procédures optimales au premier ordre et validerons empiriquement cette heuristique pour le Problème **(III)**.

1.2 Rééchantillonnage et pénalisation

Nous présentons dans cette section les procédures $\mathcal{C} = \mathcal{C}(\mathbf{X})$ qui seront étudiées et comparées dans cette thèse comme solutions au problème qui nous intéresse, celui de la sélection d'un noyau. Pour rappel celui-ci se décline, selon que l'on rééchantillonne ou pas les données, en sélection d'une méthode d'estimation à noyau ou sélection d'un estimateur linéaire (voir Problème **(II)** dans la Table 1.1). Nous verrons deux philosophies très différentes qui servent à définir ces procédures. L'une, plus connue, repose sur le principe d'estimation sans biais du risque des estimateurs.

[28] ils n'affirment pas que ces quantités sont proches, mais qu'elles se comportent de la même façon par rapport à $\mathcal{C}$.

L'autre, illustrée par la T-estimation, évalue la qualité des estimateurs en les comparant deux à deux par des tests robustes. Elle confirme, si nécessaire, qu'il faut distinguer estimation du risque et sélection d'un estimateur (Breiman & Spector, 1992).

Ainsi, il existe quatre types de procédure que l'on peut envisager pour s'attaquer aux problèmes du choix d'un noyau décrits à la Table 1.1.

		Rééchantillonnage	
		Avec	Sans
Façon de choisir	Estimation sans biais	Procédure $\mathcal{C}^{(i)}$	Procédure $\mathcal{C}^{(ii)}$
	Choix par T-estimation	Procédure $\mathcal{C}^{(iii)}$	Procédure $\mathcal{C}^{(iv)}$

TABLE 1.2 – Types de procédures étudiées dans cette thèse.

Dans cette thèse, nous étudions uniquement des procédures de type $\mathcal{C}^{(ii)}$ (Partie I), $\mathcal{C}^{(i)}$ (Partie II), et $\mathcal{C}^{(iii)}$ (Partie III).

1.2.1 Choisir dans un ensemble de candidats déterministes

Nous allons illustrer, dans un cas simple, deux façons de choisir une densité dans un ensemble de candidats déterministes. Soit $\mathcal{S}$ l'ensemble des densités de probabilité par rapport à μ muni de la perte Hellinger $\ell = h^2$ ou de la perte de Kullback-Leibler définie par le contraste $\gamma(t,x) = -\log(t(x))$. On se donne un modèle fini $\{P_t, t \in S\}$ de lois de probabilité qu'on suppose dominées par la mesure μ, et on note la perte d'un candidat dans le modèle $\ell(s,t) = \ell(P, P_t)$. On estime alors P à partir de $P_{\widehat{s}}$, ou encore s par $\widehat{s} \in S$. La question que l'on se pose est la suivante : comment choisir, à partir de l'échantillon $\mathbf{X}$ un point dans S dont la perte soit la plus petite possible ?

Estimation sans biais de la perte

Chercher le point $t \in S$ qui minimise $\ell(s,t) = \mathrm{KL}(s,t)$, revient à chercher le point $t \in S$ qui maximise $\int_\Xi \log(t(x))s(x)d\mu(x) = P(\log(t))$. L'idée naturelle est d'estimer, pour tout $t \in S$, cette fonction par l'estimateur sans biais correspondant $P_n(\log(t))$. La perte $\ell(s,t)$ de chaque candidat $t \in S$ est alors "estimée" par $\mathfrak{L}_t = P(\log(s)) - P_n(\log(t))$. Enfin, on sélectionne le point correspondant à la plus petite valeur dans $(\mathfrak{L}_t)_{t\in S}$. Ceci mène naturellement à l'estimateur qui maximise la fonction de vraisemblance sur le modèle S

$$\widehat{s}^{\mathrm{MV}} = \operatorname*{argmax}_{t\in S} \frac{1}{n}\sum_{i=1}^{n} \log\left(t(X_i)\right) \quad .$$

Cet estimateur peut aussi se définir comme le vainqueur des tests de rapport de vraisemblance successifs entre les points de S.

Remarque. En Section 1.1.4 nous avons utilisé le même raisonnement, remplacer la fonction de contraste par le *contraste empirique*, pour définir l'estimateur par projection à l'aide du contraste des moindres carrés.

T-estimation

Un autre estimateur peut être construit en comparant les différents candidats dans S à partir de tests : le T-estimateur (T pour "test")[29]. Celui-ci a été proposé par Birgé (1983) (alors appelé "d-estimateur") à la suite des travaux de Le Cam (1973, 1975) faisant le lien entre test et estimation. Rappelons d'abord sa construction suivant une version actualisée (Birgé, 2006a). Pour toute paire de densités $(t,u) \in S^2$, avec $t \neq u$, soit $\psi_{t,u} = \psi_{u,t}$ un test symétrique qui choisit entre t et u ($\psi_{t,u} = t$ signifie que le test accepte t). L'idée est d'utiliser l'échantillon $\mathbf{X}$ pour tester les points deux par deux. Pour tout $t \in S$, on note $\mathcal{R}_t$ l'ensemble des points préférés à t par ce test, c'est-à-dire $\mathcal{R}_t := \{u \in S, u \neq t \text{ t.q. } \psi_{t,u}(\mathbf{X}) = u\}$. On attribue alors à chaque t un *indice de plausibilité* $\mathcal{D}(t)$ défini par

$$\mathcal{D}(t) := \sup_{u \in \mathcal{R}_t} h(t,u) \qquad (\text{on pose} \quad \mathcal{D}(t) = 0 \quad \text{si} \quad \mathcal{R}_t = \varnothing) \ . \tag{1.14}$$

Si on note, pour $r > 0$ et $t \in S$, $\overline{\mathcal{B}}(t,r) = \{v \in \mathcal{S} \text{ t.q. } h(t,v) \leq r\}$, on remarque immédiatement que $\mathcal{R}_t \subset S \bigcap \overline{\mathcal{B}}(t, \mathcal{D}(t))$, puisque

$$u \in \mathcal{R}_t \quad \Longrightarrow \quad h(u,t) \leq \mathcal{D}(t) \ . \tag{1.15}$$

L'idée qui mène à ce critère est la suivante. Supposons que le test soit "bien construit", au sens où si la vraie densité est plus proche de l'un des deux points il choisit le plus éloigné, donc se trompe, avec faible probabilité. Dans ce cas, si $\mathcal{D}(t)$ est grand, alors il existe un point u éloigné de t qui est préféré à t par le test $\psi_{t,u}(\mathbf{X})$, ce qui suggère que u est un meilleur candidat que t pour estimer s. A l'inverse, si $\mathcal{D}(t)$ est petit, tous les points appartenant à $\mathcal{R}_t$ sont forcément proches de t, par (1.15), ce qui indique que ce point ne devrait pas être pire que les autres. Ainsi $\mathcal{D}(t)$ apparaît comme un bon indicateur pour jauger de la qualité de t comme estimation de s et le **T-estimateur** se définit simplement comme un point $\widehat{s} \in S$ qui minimise $\mathcal{D}$ (comme S est un ensemble fini, il existe au moins un tel minimiseur). En particulier, on a pour tout $t, u \in S$, $h(t,u) \leq \max(\mathcal{D}(t), \mathcal{D}(u))$ et on trouve que le T-estimateur vérifie $h(\widehat{s}, s) \leq \inf_{t \in S}(h(s,t) + \mathcal{D}(t))$. Il faut à présent expliquer précisément ce que l'on entend par test "robuste", ou "bien construit", l'ingrédient fondamental dans la définition d'un T-estimateur.

Robustesse. Si la signification du terme "robuste" varie selon l'objet étudié et le point de vue de l'auteur, elle se réfère en général au fait qu'une petite perturbation des hypothèses initiales ne change pas le comportement général d'un phénomène. Une question importante posée en ce sens concerne le cas paramétrique. Si nous avons supposé, pour construire un estimateur, qu'il existe un $s \in S$ tel que $P = P_s$, comment réagit-il si cette hypothèse s'avère être fausse (même très légèrement) ? Ceci a motivé de nombreuses recherches à la fin des années 70 et au début des années 80 (voir par exemple les travaux de Beran (Beran, 1977, 1978, 1980, 1981b,a) ou le livre de Bickel (1981)). En ce qui concerne la T-estimation, et la robustesse d'un test, nous suivons Huber (1965) qui propose de dire qu'une procédure est "robuste" si ses performances se dégradent peu

[29] La procédure dite "de Lepski" utilise aussi la comparaison entre les candidats pour majorer le terme de variance.

lorsque les hypothèses du modèle théorique idéal ne sont pas exactement satisfaites mais que l'on s'en éloigne peu.

Nous ne considérerons que des tests qui s'écrivent pour une statistique de test $T_{t,u}(\mathbf{X})$ et un réel z sous la forme

$$\psi_{t,u}(\mathbf{X}) = \begin{cases} t & \text{si} \quad T_{t,u}(\mathbf{X}) > z \ ; \\ u & \text{si} \quad T_{t,u}(\mathbf{X}) < z \ . \end{cases} \tag{1.16}$$

Le test fait un choix arbitraire lorsque $T_{t,u}(\mathbf{X}) = z$. La propriété de robustesse requise pour le test $\psi_{t,u}$ entre les densités t et u est résumée par l'hypothèse suivante.

Hypothèse (TEST). *Soit $\theta \in (0, 1/2)$ et $a > 0$. Il existe $\psi_{t,u}(\mathbf{X})$ qui dépend de t, u et $\mathbf{X}$ tel que*

$$\sup_{\{\, P \in \mathcal{P} \,|\, h(s,t) \leq \theta h(t,u) \,\}} \mathbb{P}\left(\, \psi_{t,u}(\mathbf{X}) = u \,\right) \leq \exp\left[-n \left(a h^2(t,u) + z \right) \right] \tag{1.17}$$

et

$$\sup_{\{\, P \in \mathcal{P} \,|\, h(s,u) \leq \theta h(t,u) \,\}} \mathbb{P}\left(\, \psi_{t,u}(\mathbf{X}) = t \,\right) \leq \exp\left[-n \left(a h^2(t,u) - z \right) \right] \ . \tag{1.18}$$

Le but est de s'assurer que si la vraie loi P appartient à l'une des boules de Hellinger, $\overline{\mathcal{B}}(t,r)$ ou $\overline{\mathcal{B}}(u,r)$ avec $r = \theta h(t,u)$, la probabilité que le test se trompe (choisisse u alors que $h(t,s) < h(u,s)$) diminue de façon proportionnelle à la distance entre les deux centres des boules. Cette propriété, qui semble pourtant essentielle, n'est pas vérifiée par le test de rapport de vraisemblance (Birgé, 2006a, Section 3.2).

Pour conclure, mettons en évidence la propriété de robustesse de cet estimateur en considérant un modèle paramétrique $S = \{s_\eta, \eta \in \Upsilon\}$, pour un sous-ensemble $\Upsilon \subset \mathbb{R}$ sur lequel on trouve un lien entre la distance euclidienne et la distance de Hellinger au sens où il existe des constantes c_1, c_2 et κ telles que

$$c_1 \left| \eta - \eta' \right|^\kappa \leq h^2(s_\eta, s_{\eta'}) \leq c_2 \left| \eta - \eta' \right|^\kappa \qquad \forall \eta, \eta' \in \Upsilon \ .$$

Dans cette situation, il est possible de prouver la borne suivante sur le risque Hellinger du T-estimateur

$$\mathbb{E}\left[h^2\left(s, \widehat{s} \right) \right] \leq C \left(\inf_{\eta \in \Upsilon} \ h^2\left(s, s_\eta \right) + \frac{\max\left(1/2, \kappa^{-1}(2\log(5) + \log(c_2/c_1)) \right)}{n} \right) \ .$$

En particulier, si $s \in S$ alors on trouve la vitesse, attendue dans le cas paramétrique, en n^{-1} (Le Cam, 1973). Plus intéressant, si la densité s n'est pas exactement dans S mais est tout de même proche du modèle au sens où $\inf_{\eta \in \Upsilon} h^2(s, s_\eta) \leq n^{-1}$, la borne ne change pas et reste en n^{-1}. C'est précisément la propriété de robustesse que n'a pas l'estimateur par maximum de vraisemblance $\widehat{s}^{\mathrm{MV}}$ (connu pour avoir de très bonnes propriétés sous des hypothèses fortes, mais dont la qualité peut s'avérer catastrophique si la vraie loi n'appartient pas au modèle, même si elle en est très proche).

Dans le cadre de l'estimation d'une densité, des procédures de type $\mathcal{C}^{(iv)}$ dans la Table 1.2, c'est-à-dire par T-estimation et sans rééchantillonnage, ont été essentiellement étudiées par Birgé

(2006a, 2014); Baraud (2011) et Sart (2013) dans un cadre paramétrique. Récemment, Baraud *et al.* (2014) ont proposé une nouvelle approche, appelée ρ-estimation, qui s'inscrit dans la lignée des procédures basées sur des tests. Tout comme le T-estimateur, le ρ-estimateur possède de très bonnes propriétés sous des hypothèses très faibles[(30)]. Il existe néanmoins plusieurs différences entre ρ et T-estimation. En particulier le risque du ρ-estimateur peut se contrôler sur des modèles de dimension métrique non-bornée (et éventuellement non-compact). Notons enfin que les procédures d'estimation de Sart (2013) et Baraud *et al.* (2014) coïncident avec le maximum de vraisemblance lorsque le modèle est paramétrique, contient la densité et est suffisamment régulier. Nous renvoyons le lecteur à ces articles pour plus d'informations sur ce sujet.

Remarque. En utilisant la perte Hellinger nous pouvons également exhiber un compromis biais-variance pour les T-estimateurs. Pour un modèle S de *dimension métrique* (voir Birgé (2006a) pour une définition précise) bornée par D, il est en effet possible de majorer son risque par la somme de $h^2(s,S)$, qui représente l'erreur de modélisation (terme de biais), avec le terme de complexité D/n, qui équivaut à l'erreur d'estimation (terme de variance). La discussion est par conséquent sensiblement la même que pour les estimateurs linéaires.

1.2.2 Choisir dans un ensemble de candidats aléatoires

Le problème auquel nous faisons face est plus complexe que le précédent puisqu'il s'agit de choisir, à partir de l'échantillon $\mathbf{X}$, entre des candidats aléatoires et non plus déterministes. Nous allons voir que ces deux façons de faire peuvent être encore utilisées, mais que des aménagements sont nécessaires. Pour chacune d'elle, il existe, selon la perte sous-jacente au problème, une procédure idéale (mais inconnue) $\mathcal{C}_{\mathrm{id}}$ qu'il est tentant de vouloir imiter.

- Si l'on considère une perte définie à partir d'une fonction de contraste (1.3), le risque d'une méthode d'estimation $\mathcal{A}_m$ s'écrit $\mathbb{E}\left[\ell(s,\widehat{s}_m)\right] = P\gamma(\widehat{s}_m) - P\gamma(s)$. Le critère idéal dans ce cadre statistique est donc donné par $\mathcal{C}_{\mathrm{id}} : m \mapsto P\gamma(\widehat{s}_m)$. Le principe d'estimation sans biais du risque cherche alors à construire une procédure $\mathcal{C}$ tel que $\mathbb{E}\left[\mathcal{C}(m)\right] = \mathbb{E}\left[\mathcal{C}_{\mathrm{id}}(m)\right]$.

- Soit $\ell = h^2$ la perte Hellinger et ρ l'affinité de Hellinger. Notons pour les densités t, u et $r = (t+u)/2$ la statistique

$$T_{t,u}(P) := \frac{1}{2}\left(\int \frac{\sqrt{t(x)} - \sqrt{u(x)}}{\sqrt{r(x)}} s(x) d\mu(x) + \left(\rho\left(t,r\right) - \rho\left(u,r\right)\right)\right) \ .$$

  Pour décider qui de t et u est plus proche de s, on sait par Baraud (2011) que

$$T_{t,u}(P) \geqslant 0 \quad \Longrightarrow \quad h^2(s,t) \leqslant \frac{\sqrt{2}+1}{\sqrt{2}-1}\, h^2(s,u) \ .$$

  Le test $\psi_{t,u}$ qui choisit t quand $T_{t,u}(P) \geqslant 0$ n'est pas "idéal" au sens strict (ce dernier donnerait $h^2(s,t) < h^2(s,u)$ si t est choisi et l'inégalité inverse autrement), mais considérons-le comme tel. Cette statistique est inconnue et Baraud l'estime par l'estimateur sans biais

[(30)] à notre connaissance il s'agit de la solution qui s'approche au plus près du souhait de Le Cam d'obtenir un estimateur "universel" pour remplacer l'estimateur par maximum de vraisemblance.

correspondant. Ainsi, pour choisir entre deux estimateurs $\widehat{s}_l$ et $\widehat{s}_m$, un test idéal pour la T-estimation est donné par $T_{\widehat{s}_l,\widehat{s}_m}(P)$. Le critère idéal correspondant serait $\mathcal{C}_{\mathrm{id}}(m) = \mathcal{D}(\widehat{s}_m)$ où la fonction $\mathcal{D}$ est donnée par (1.14) avec le test robuste $\psi_{\widehat{s}_l,\widehat{s}_m}$ défini au travers de $T_{\widehat{s}_l,\widehat{s}_m}(P)$.

Un problème important (connu au moins depuis les années 30 (Larson, 1931)) lorsqu'on a affaire à des candidats aléatoires est le suivant : utiliser les mêmes données pour construire les estimateurs $(\widehat{s}_m(\mathbf{X}))_{m\in\mathcal{M}}$ et évaluer leurs qualités mène à un résultat trop optimiste, au sens où l'estimateur qui "colle" le plus aux données sera favorisé. Ceci peut se voir facilement sur l'exemple de sélection de modèles suivant. Soit ℓ la perte des moindres carrés et $(\widehat{s}_m(\mathbf{X}))_{m\in\mathcal{M}}$ une collection d'estimateurs par projection où chaque $\widehat{s}_m$ est défini sur un modèle S_m de dimension D_m. Supposons aussi que les modèles sont emboîtés au sens où, pour tout $m, m' \in \mathcal{M}$, on a $S_m \subset S_{m'}$ si $D_m < D_{m'}$. Comme le choix idéal est $\mathcal{C}_{\mathrm{id}}(m) = P\gamma(\widehat{s}_m)$, on pourrait être tenté de procéder comme en Section 1.2.1, c'est-à-dire estimer une nouvelle fois P par P_n et proposer $\mathcal{C}(m) = P_n\gamma(\widehat{s}_m)$. Mais dans ce cas, le critère $\mathcal{C}$ sera toujours minimal pour le plus grand modèle étant donné que $P_n\gamma(\widehat{s}_{m_1}) < P_n\gamma(\widehat{s}_{m_2})$ si $D_{m_1} > D_{m_2}$. On parle alors de *sur-apprentissage*, en opposition au *sous-apprentissage* qui touche au cas où le modèle est trop petit et l'erreur d'approximation trop grande. Le même problème surgit si on utilise le test $T_{\widehat{s}_l,\widehat{s}_m}(P_n)$ pour faire un choix entre deux estimateurs.

Nous voyons que les difficultés apparaissent quand on estime une fonctionnelle $\mathcal{L}(P, P_n)$[31] qui dépend de la vraie loi P et de la mesure empirique P_n par $\mathcal{L}(P_n, P_n)$. Discutons à présent une solution générale qui nous vient du rééchantillonnage (sans rentrer dans les détails par souci de clarté) avant de décrire dans les deux sous-sections qui suivent, les deux types de procédures que l'on considère dans cette thèse, la sélection par pénalisation et la méthode dite de validation croisée, ainsi que les liens entre celles-ci.

Rééchantillonnage et bootstrap. L'objectif du rééchantillonnage (voir Arlot (2007) dont nous nous inspirons) est de construire de nouveaux échantillons à partir de $\mathbf{X}$. On peut par exemple générer un n-échantillon $\mathbf{X}^*$ en tirant aléatoirement des données dans $\mathbf{X}$. Cette technique s'appelle le *bootstrap* et consiste à construire un cadre statistique, miroir du monde réel, dans lequel tout est connu : la loi P est remplacée par P_n alors que P_n est elle-même remplacée par une loi aléatoire P_n^*. Les processus d'échantillonnage et de rééchantillonnage sont identiques, ce qui implique que les données rééchantillonnées $\mathbf{X}^* = (X_1^*, \dots, X_n^*)$[32] sont i.i.d. de loi P_n conditionnellement à $\mathbf{X}$. On estime alors, selon *l'heuristique d'Efron* (Efron, 1979), la fonctionnelle $\mathcal{L}(P, P_n)$ par $\mathbb{E}^*\left[\mathcal{L}(P_n, P_n^*)\right]$, où $\mathbb{E}^*$ désigne l'espérance conditionnelle à $\mathbf{X}$.

Il existe un nombre important de possibilités pour définir P_n^*. Dans le bootstrap *à poids* (Mason & Newton, 1992; Præstgaard & Wellner, 1993), on définit P_n^* en se donnant un vecteur $W = (W_1, \dots, W_n)$ de poids aléatoires, indépendants de $\mathbf{X}$ et tels que $\mathbb{E}\left[W_i\right] = 1$ pour tout $i \in [n]$. La mesure qui remplace P_n dans le monde bootstrap est alors définie par

$$P_n^* = P_n^W = \frac{1}{n}\sum_{i=1}^{n} W_i \delta_{X_i} \ . \tag{1.19}$$

[31] dans les exemples cités ici, cette fonctionnelle $\mathcal{L}(P, P_n)$ est $P\gamma(\widehat{s}_m)$ ou $T_{\widehat{s}_l,\widehat{s}_m}(P)$.

[32] ces données ne sont pas "nouvelles", elles proviennent de ce qu'on a, c'est à dire de l'échantillon $\mathbf{X}$.

La quantité $\mathcal{L}(P, P_n)$ est alors estimée par $C_W \mathbb{E}_W \left[\mathcal{L}(P_n, P_n^W) \right]$, où C_W est une constante qui ne dépend que de la variabilité des poids et $\mathbb{E}_W$ représente l'espérance par rapport à l'aléa du rééchantillon uniquement. Il est courant de supposer en plus que les coordonnées du vecteur W sont échangeables au sens où pour toute permutation τ de $[n]$, le vecteur W a même loi que $(W_{\tau(1)}, \ldots, W_{\tau(n)})$.

1.2.3 Validation croisée

Lorsque nous disposons de méthodes d'estimation, une solution pour éviter cette sous-estimation du risque consiste à allouer une partie seulement des observations pour construire des estimateurs et juger de leurs qualités sur les données restantes (par exemple selon une des deux manières présentées à la Section 1.2.1). C'est l'idée sur laquelle repose la validation croisée (VC). Pour tout ensemble fini A, on note $|A|$ son cardinal. De plus, pour tout sous-ensemble $A \subset [n]$, on pose $A^c := [n] \backslash A$ et

$$P_n^{(A)} f = \frac{1}{|A|} \sum_{i \in A} f(X_i) \qquad \text{et} \qquad P_n^{(-A)} f = \frac{1}{|A^c|} \sum_{i \in A^c} f(X_i) \ .$$

Hold-Out. La procédure *hold-out* (HO) ou de validation simple (VS) correspond au schéma le plus simple de validation croisée. Elle peut se décrire dans le cadre général comme suit. Pour un sous-ensemble $E \subset [n-1]$ donné, on divise l'échantillon $\mathbf{X}$ en deux sous-échantillons disjoints, $\mathbf{X}^{(E)}$ et $\mathbf{X}^{(-E)} = \mathbf{X} \backslash \mathbf{X}^{(E)}$, qu'on appelle respectivement échantillon *d'entraînement* et de *validation*. L'échantillon $\mathbf{X}^{(E)}$ est d'abord utilisé pour construire l'ensemble $\{\widehat{s}_m^{(E)} := \mathcal{A}_m(\mathbf{X}^{(E)}), m \in \mathcal{M}\} \subset \mathcal{S}$ d'estimateurs préliminaires[(33)]. Ensuite la qualité de chaque méthode $\mathcal{A}_m$ est évaluée à l'aide d'un critère $\mathrm{crit}^{(-E)} : \mathcal{S} \to \mathbb{R}$ qui ne repose que sur les données appartenant à l'échantillon de validation. Le critère final, lui, dépend de toutes les données et s'écrit en toute généralité

$$\mathcal{C}_E^{\mathrm{VS}}(m) := \mathrm{crit}^{(-E)} \left(\widehat{s}_m^{(E)} \right) \ . \tag{1.20}$$

Pour un sous-ensemble E, la procédure hold-out se déduit de ce critère en sélectionnant $\widehat{m} = \widehat{m}(\mathcal{C}_E^{\mathrm{VS}})$. Selon les auteurs, l'estimateur final peut être $\mathcal{A}_{\widehat{m}}(\mathbf{X}^{(E)})$ (comme dans Devroye & Lugosi (2001),Blanchard & Massart (2006)) ou $\mathcal{A}_{\widehat{m}}(\mathbf{X})$ (comme dans Arlot & Lerasle (2014)). Donnons les définitions des deux HO qui se déduisent de l'estimation sans biais du risque et de la T-estimation.

- Si l'on veut estimer le critère $\mathcal{C}_{\mathrm{id}}(m) = P\gamma(\widehat{s}_m)$, le critère HO classique[(34)] s'écrit

$$\mathcal{C}_E^{\mathrm{HO}}(m) := P_n^{(-E)} \gamma \left(\widehat{s}_m^{(E)} \right) = \frac{1}{|E^c|} \sum_{i \in E^c} \gamma \left(\widehat{s}_m^{(E)}, X_i \right) \ . \tag{1.21}$$

- Pour la perte Hellinger, Birgé (2006a) propose, comme application possible de la construction générale des T-estimateurs, le critère T-hold-out (THO dans la suite). Il définit une

[(33)] dans notre cadre ils s'écrivent $\widehat{s}_m^{(E)}(x) = \frac{1}{|E|} \sum_{i \in E} \mathcal{K}_m(X_i, x)$.

[(34)] nous appellerons HO le critère classique lié au cadre où la perte s'écrit via une fonction de contraste.

collection de poids $(\Delta_m)_{m\in\mathcal{M}}$, associée à la famille de méthodes d'estimation $(\mathcal{A}_m)_{m\in\mathcal{M}}$, telle que

$$\Delta_m \geqslant 0 \quad \text{pour tout } m \in \mathcal{M}, \qquad \text{et} \qquad \Gamma = \sum_{m\in\mathcal{M}} \exp(-\Delta_m) < \infty \ . \tag{1.22}$$

Le critère s'écrit alors à l'aide des notations introduites à la Section 1.2.1 et en posant $z = \Delta_l - \Delta_m$ dans (1.16)

$$\mathcal{C}_E^{\mathrm{THO}}(m) := \sup_{\widehat{s}_l^{(E)}\in\mathcal{R}_m} h\left(\widehat{s}_l^{(E)}, \widehat{s}_m^{(E)}\right) \ , \tag{1.23}$$

où $\mathcal{R}_m = \{\widehat{s}_l^{(E)},\ l \neq m \mid \psi_{\widehat{s}_l^{(E)},\widehat{s}_m^{(E)}}(\mathbf{X}^{(-E)}) = l\}$.

La procédure de la validation simple, d'apparence très naïve, a été abondamment étudiée du point de vue théorique, essentiellement parce qu'elle présente l'avantage de pouvoir considérer les estimateurs comme "gelés", étant donné qu'ils sont construits sur un échantillon indépendant. Parmi les travaux théoriques dédiés au HO dans le cadre de la densité[35], nous pouvons mentionner (Arlot & Lerasle, 2014, Section 8.1) (pour les estimateurs par projection), Devroye & Lugosi (2001) (pour les estimateurs de Parzen-Rosenblatt) et Birgé (2006a) (pour les T-estimateurs). Enfin, notons que les procédures d'agrégation (Section 1.1.5) reposent également sur l'idée du hold-out et cherchent à définir la meilleure façon $\mathrm{crit}^{(-E)}$ d'agréger les estimateurs $\{\widehat{s}_m^{(E)}, m \in \mathcal{M}\}$ avec l'échantillon de validation $\mathbf{X}^{(-E)}$. Ceci soulève le problème du choix du découpage optimal pour celles-ci.

Validation croisée V-fold. Malgré sa simplicité et ses propriétés théoriques enviables en apparence, le HO n'est pas utilisé par les praticiens car il souffre d'une trop grande variabilité qui vient du choix arbitraire du sous-ensemble E. Pour l'améliorer, on peut considérer différentes partitions dans l'espoir d'obtenir une évaluation plus précise de la qualité de chaque méthode d'estimation. Toute procédure de VC peut se déduire du HO dès lors qu'on se donne une collection $\mathcal{E}$ d'échantillons d'entraînement. En effet, le critère de validation croisée est la moyenne des critères de validation simple $\mathcal{C}_E^{\mathrm{VS}}$ obtenus pour chaque $E \in \mathcal{E}$

$$\mathcal{C}_{\mathcal{E}}^{\mathrm{VC}}(m) = \frac{1}{|\mathcal{E}|}\sum_{E\in\mathcal{E}} \mathcal{C}_E^{\mathrm{VS}}(m) \ . \tag{1.24}$$

Il existe deux grandes tendances selon que l'on souhaite (ou non) l'exploitation exhaustive des données, c'est-à-dire de tous les découpages possibles pour un échantillon d'entraînement d'une certaine taille donnée. Parmi les procédures de VC les plus connues, le *leave-p-out* (LPO) consiste à allouer, pour chaque partition, $p \in [n-1]$ données à l'échantillon de validation[36], ce qui revient à considérer $\mathcal{E}_p = \{E \subset [n-1] \text{ t.q. } |E| = n-p\}$. Dans le cas classique, quand $\mathcal{C}_E^{\mathrm{HO}}$ est donné

[35] pour être complet, citons aussi les travaux en classification (Bartlett *et al.*, 2002; Blanchard & Massart, 2006), et en régression (Lugosi & Nobel, 1999; Juditsky & Nemirovski, 2000; Wegkamp, 2003).

[36] ceci explique le "p-out" en anglais.

par (1.21), on peut écrire pour tout $p \in [n-1]$

$$\mathcal{C}_p^{\mathrm{LPO}}(m) := \mathcal{C}_{\mathcal{E}_p}^{\mathrm{VC}}(m) = \frac{1}{|\mathcal{E}_p|} \sum_{E \in \mathcal{E}_p} P_n^{(-E)} \gamma\left(\widehat{s}_m^{(E)}\right) \tag{1.25}$$

Le célèbre critère du *leave-one-out* (LOO) est le cas particulier où $p = 1$ (Stone, 1974; Rudemo, 1982). Nous conseillons l'article de survol de Arlot & Celisse (2010) pour avoir un aperçu complet des différentes procédures de VC et l'article de Celisse & Robin (2008) pour le LPO.

Dans cette thèse nous allons principalement nous focaliser sur une exploration non-exhaustive des données. Soit $V \in \{2, \ldots, n\}$ tel que $n/V \in \mathbb{N}$, et $(B_1, \ldots, B_V)$ une partition de $[n]$ en sous-ensembles disjoints[37] de même taille n/V[38]. Pour chaque découpage $j \in [V]$, l'échantillon d'entraînement $\mathbf{X}^{(-B_j)}$ sert à la construction des "estimateurs partiels"

$$\left\{ \widehat{s}_m^{(-j)} := \mathcal{A}_m\left(\mathbf{X}^{(-B_j)}\right),\ m \in \mathcal{M} \right\}$$

alors que l'échantillon de validation correspondant $\mathbf{X}^{(B_j)}$ est dédié à la fabrication du critère d'évaluation $\mathrm{crit}^{(j)}$ des méthodes $\mathcal{A}_m$.

Remarque. Pour toute méthode d'estimation à noyau $m \in \mathcal{M}$, on a $\widehat{s}_m = V^{-1} \sum_{j=1}^V \widehat{s}_m^{(-j)}$. Ceci revient à dire que l'estimateur construit avec toutes les observations est exactement la combinaison convexe des V estimateurs partiels.

Si on pose $\mathcal{E}_{B_{[V]}} = \{B_1^c, \ldots, B_V^c\}$, l'estimateur obtenu par validation croisée V-fold (Geisser, 1975) vient du choix $\widehat{m}(\mathcal{C}^{\mathrm{VF}})$ où

$$\mathcal{C}^{\mathrm{VF}}(m) := \mathcal{C}_{\mathcal{E}_{B_{[V]}}}^{\mathrm{VC}}(m) = \frac{1}{V} \sum_{j=1}^V \mathcal{C}_{B_j^c}^{\mathrm{VS}}(m) = \frac{1}{V} \sum_{j=1}^V \mathrm{crit}^{(j)}\left(\widehat{s}_m^{(-j)}\right) \quad .$$

Ainsi, il existe autant de procédures de VCVF qu'il y a de façon de définir $\mathrm{crit}^{(j)}$, c'est-à-dire qu'il y a de manières d'évaluer la qualité d'un estimateur dans une collection donnée. Dans le cadre classique, lorsque la perte est définie par (1.3), le critère VF a été introduit par Breiman *et al.* (1984). Il se déduit directement de la définition de $\mathcal{C}_{B_j^c}^{\mathrm{HO}}$ donnée par (1.21)

$$\mathcal{C}_V^{\mathrm{VFCV}}(m) := \frac{1}{V} \sum_{j=1}^V \frac{1}{|B_j|} \sum_{i \in B_j} \gamma\left(\widehat{s}_m^{(-j)}, X_i\right) \quad . \tag{1.26}$$

Nous désignons par $\mathcal{C}_V^{\mathrm{LSVF}}$ et $\mathcal{C}_V^{\mathrm{KLVF}}$[39] les procédures définies respectivement par les contrastes des moindres carrés et du maximum de vraisemblance. Ainsi, nous avons une collection d'estimation des risques des méthodes $\{\mathcal{C}_V^{\mathrm{VFCV}}(m), m \in \mathcal{M}\}$ et l'estimateur sélectionné est naturellement celui dont le risque estimé est le plus petit. Voyons ce que vaut le biais de $\mathcal{C}_V^{\mathrm{VFCV}}$ comme estima-

[37] de sorte que $\mathbf{X} = \bigcup_{j=1}^V \mathbf{X}^{(B_j)}$ avec les sous-échantillons de validation $\mathbf{X}^{(B_j)}$ qui sont indépendants.

[38] Aussi, il faudrait plutôt parler de sous-échantillonnage car l'échantillon résultant n'est plus de taille n.

[39] pour "Least-Squares V-fold" et "Kullback-Leibler V-fold" en anglais.

teur de $\mathcal{C}_{\mathrm{id}}$. On trouve, par indépendance de $\mathbf{X}^{(B_j)}$ et $\mathbf{X}^{(-B_j)}$,

$$\mathbb{E}\left[\frac{1}{V}\sum_{j=1}^{V} P_n^{(j)}\gamma\left(\widehat{s}_m^{(-j)}\right) - \mathcal{C}_{\mathrm{id}}(m)\right] = \mathbb{E}\left[P_n^{(1)}\gamma\left(\widehat{s}_m^{(-1)}\right) - \mathcal{C}_{\mathrm{id}}(m)\right]$$
$$= \mathbb{E}\left[P\gamma\left(\widehat{s}_m^{(-1)}\right) - \mathcal{C}_{\mathrm{id}}(m)\right] = \mathbb{E}\left[P\gamma\left(\widehat{s}_m^{(E)}\right) - P\gamma(\widehat{s}_m)\right] \ ,$$

où $\widehat{s}_m^{(E)}$ est un estimateur construit avec $n - n/V$ données. Ainsi, le biais s'écrit comme la différence des risques de la méthode $\mathcal{A}_m$ utilisant respectivement $n - n/V$ et n données de sorte que la VCVF estime le risque d'un estimateur construit avec $n(V-1)/V$ données au lieu de n. Cette quantité est souvent positive et décroît quand V augmente, ce qui signifie que l'estimateur est biaisé mais asymptotiquement sans biais. Burman (1989) a proposé de corriger ce défaut[40] et introduit le critère de correction suivant

$$\mathcal{C}_V^{\mathrm{corr,VFCV}}(m) := \mathcal{C}_V^{\mathrm{VFCV}}(m) + P_n\gamma\left(\widehat{s}_m\right) - \frac{1}{V}\sum_{j=1}^{V} P_n\gamma\left(\widehat{s}_m^{(-j)}\right) \ . \tag{1.27}$$

Les procédures $\mathcal{C}_p^{\mathrm{LPO}}$, $\mathcal{C}_V^{\mathrm{VFCV}}$ et $\mathcal{C}_V^{\mathrm{corr,VFCV}}$ sont toutes du type $\mathcal{C}^{(i)}$ dans la Table 1.2. Le tableau suivant récapitule les différentes procédures de rééchantillonnage qui reposent sur un contraste. Notons que les équivalents de type validation croisée (en particulier VF ou LPO) de $\mathcal{C}_E^{\mathrm{THO}}$ n'existent pas dans la littérature.

	Procédure $\mathcal{C}^{(i)}$		
	Cadre général	Type de découpage	
		Exhaustif	Non-exhaustif
	VS $\longrightarrow$ VC	LPO	VCVF $\longrightarrow$ Burman
	$\mathcal{C}_E^{\mathrm{VS}} \longrightarrow \mathcal{C}_{\mathcal{E}}^{\mathrm{VC}}$	$\mathcal{C}_p^{\mathrm{LPO}}$	$\mathcal{C}_V^{\mathrm{VFCV}} \longrightarrow \mathcal{C}_V^{\mathrm{corr,VFCV}}$
Définition	(1.20) $\longrightarrow$ (1.24)	(1.25)	(1.26) $\longrightarrow$ (1.27)

TABLE 1.3 – Procédures de validation croisée mentionnées dans cette section.

Remarque. Si on revient au problème initial, celui de l'estimation d'une certaine fonctionnelle $\mathcal{L}(P, P_n)$, l'idée du critère HO, décrit par (1.20), revient à l'estimer par $\mathcal{L}(P_n^{(-E)}, P_n^{(E)})$, où $P_n^{(-E)}$ et $P_n^{(E)}$ sont respectivement les mesures empiriques des échantillons $\mathbf{X}^{(-E)}$ et $\mathbf{X}^E$. En toute généralité, nous pouvons donc résumer l'heuristique de la VC et de la VCVF comme suit :

$$\text{estimer} \quad \mathcal{L}(P, P_n) \quad \text{par} \quad \frac{1}{|\mathcal{E}|}\sum_{E\in\mathcal{E}} \mathcal{L}\left(P_n^{(-E)}, P_n^{(E)}\right) \quad \text{et} \quad \frac{1}{V}\sum_{j=1}^{V} \mathcal{L}\left(P_n^{(j)}, P_n^{(-j)}\right) \ , \tag{1.28}$$

où $P_n^{(j)}$ et $P_n^{(-j)}$ sont les mesures empiriques des échantillons $\mathbf{X}^{(B_j)}$ et $\mathbf{X}^{(-B_j)}$.

[40] initialement dans le cadre de la régression.

Quelles garanties théoriques pour la VC ? Le principal avantage de la validation croisée, outre le fait que sa définition semble très intuitive, est qu'elle ne requiert que l'indépendance des observations. Ce caractère universel en fait une méthode particulièrement appréciée par les chercheurs. Une brève recherche des mots clés "cross-validation" sur certains sites spécialisés permet de se faire une idée de l'utilisation massive[(41)] de techniques de validation croisée dans bien des domaines : en mathématiques, en ingénierie mais aussi en biologie, en physique et même en sciences sociales. Ce succès considérable auprès de la communauté scientifique (dans son ensemble, donc) contraste avec le faible nombre de garanties théoriques non-asymptotiques qui existent dans la littérature[(42)]. De nombreuses modifications du critère (1.26) ont été suggérées, pour corriger certains inconvénients et améliorer son application en pratique, menant à une ribambelle de nouvelles procédures[(43)] dont personne ne peut dire laquelle est la plus performante puisqu'elles reposent essentiellement sur des considérations empiriques dont on ne peut tirer de conclusions définitives.

Ainsi, de nombreux paradoxes existent. Un exemple révélateur concerne la procédure hold-out. Alors qu'elle est rejetée par les utilisateurs car jugée trop instable, celle-ci a été très étudiée théoriquement et de nombreuses inégalités d'oracle ont pu être obtenues dans divers cadres statistiques[(44)]. La VCVF, à l'inverse, semble être nettement plus appréciée par les praticiens alors que les garanties théoriques font défaut. Depuis quelques années, des inégalités de type oracle ont été prouvées (pour des estimateurs linéaires particuliers) en régression pour la VCVF (Arlot, 2007) et en densité pour la VCVF (Arlot & Lerasle, 2014) et pour le LPO (Celisse & Robin, 2008; Celisse, 2014). Le problème de comparaison des procédures évoqué en Section 1.1.6 s'avère dès lors particulièrement pertinent. Il cherche, entre autre, à répondre théoriquement à cette question : pourquoi est-il préférable de découper l'échantillon en plusieurs blocs plutôt que d'utiliser le HO ? Le travail de Arlot & Lerasle (2014) vise ainsi à aider, grâce à des arguments théoriques non-asymptotiques (notamment l'heuristique (1.13)), le praticien à comparer deux procédures par rééchantillonnage et à les calibrer (choisir V, dans le cas de la VCVF).

1.2.4 Pénalisation

Revenons au cadre où la perte se définit à partir d'une fonction de contraste et à l'idée de l'estimation sans biais du risque, évoquée en Section 1.2.1. Nous allons définir des procédures de type $\mathcal{C}^{(ii)}$ et $\mathcal{C}^{(i)}$ (voir la Table 1.2) qui seront étudiées dans le détail respectivement dans la Partie I et II de cette thèse.

Au vu du problème de sélection de modèles exposé en Section 1.2.2, une idée naturelle pour éviter le sur-apprentissage consiste à pénaliser les modèles en fonction de leur dimension, ce qui légitime l'introduction de la fonction *pénalité* $\mathrm{pen} : \mathcal{M} \to \mathbb{R}$. On définit la procédure pénalisée $\mathcal{C}_{\mathrm{pen}}$ par

$$\mathcal{C}_{\mathrm{pen}}(m) := P_n\gamma(\widehat{s}_m) + \mathrm{pen}(m) \ . \tag{1.29}$$

[(41)]743 000 résultats sur Google Scholar, contre 130 000 recensés en 2007 (Arlot, 2007, p.20).

[(42)]Il existe néanmoins des travaux de nature asymptotique sur la consistance et les vitesses de convergence des estimateurs sélectionnés, voir par exemple Hall (1983); Stone (1984); Hall (1987); Hall & Marron (1987); Scott & Terrell (1987).

[(43)]et autant de préfixes devant 'CV' : 'trimmed CV' (Feluch & Koronacki, 1992), 'modified CV' (Stute, 1992), 'indirect CV' (Savchuk *et al.*, 2010), 'biased CV' (Scott & Terrell, 1987).

[(44)]Il "suffit" pour cela de considérer un échantillon d'entraînement de taille $|E| = n(1 - 1/\log(n))$, voir Blanchard & Massart (2006) et van der Laan & Dudoit (2003).

Ceci mène à *l'estimateur par minimum de contraste pénalisé*. Il existe autant de procédures pénalisées que de définitions pour pen, dont le choix s'avère être crucial pour obtenir un bon estimateur final $\widehat{s}_{\widehat{m}(\mathcal{C}_{\mathrm{pen}})}$. La pénalité idéale[45] dans ce contexte est donnée par

$$\mathrm{pen}_{\mathrm{id}}(m, \mathbf{X}) := (P - P_n)\gamma\left(\mathcal{A}_m(\mathbf{X})\right) \quad , \tag{1.30}$$

et on trouve, en utilisant la définition de la perte ℓ et $\widehat{m}(\mathcal{C}_{\mathrm{pen}})$, pour tout $m \in \mathcal{M}$

$$\ell\left(s, \widehat{s}_{\widehat{m}(\mathcal{C}_{\mathrm{pen}})}\right) \leq \ell\left(s, \widehat{s}_m\right) + \left(\mathrm{pen}(m) - \mathrm{pen}_{\mathrm{id}}(m, \mathbf{X})\right) \\ - \left(\mathrm{pen}(\widehat{m}(\mathcal{C}_{\mathrm{pen}})) - \mathrm{pen}_{\mathrm{id}}(\widehat{m}(\mathcal{C}_{\mathrm{pen}}), \mathbf{X})\right) \quad . \tag{1.31}$$

De nombreuses propositions (déterministes ou aléatoires) pour pen ont été introduites afin de gérer les fluctuations de $\mathrm{pen} - \mathrm{pen}_{\mathrm{id}}$ pour obtenir une inégalité oracle en partant de (1.31). Dans les années 70, Akaike (1973) et Mallows (1973) suggèrent, selon le principe d'estimation sans biais du risque, qu'une bonne pénalité pen doit vérifier

$$\mathbb{E}\left[\mathrm{pen}(m)\right] = \mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m, \mathbf{X})\right] \quad \text{pour tout } m \in \mathcal{M} \quad .$$

Dans cette optique, la connaissance de l'espérance de la pénalité idéale apparaît être un atout important pour calibrer pen. Or on peut souvent[46] écrire,

$$\mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m, \mathbf{X})\right] = \mathbb{E}\left[(P - P_n)\gamma\left(\mathcal{A}_m(\mathbf{X})\right)\right] = \frac{1}{|\mathbf{X}|} f(m) \quad , \tag{1.32}$$

pour une certaine fonction $f : \mathcal{M} \to \mathbb{R}$. Par exemple, pour le contraste des moindres carrés et pour un estimateur linéaire quelconque défini par (1.6), on peut montrer que

$$f(m) = 2\left(\mathbb{E}\left[\mathcal{K}_m(X, X)\right] - \mathbb{E}\left[\mathcal{K}_m(X, Y)\right]\right) \quad .$$

Des procédures du type $\mathcal{C}^{(i)} = \mathcal{C}_{\mathrm{pen}^{(i)}}$ et $\mathcal{C}^{(ii)} = \mathcal{C}_{\mathrm{pen}^{(ii)}}$ peuvent être proposées pour estimer $\mathrm{pen}_{\mathrm{id}}$ avec pour but de vérifier $\mathbb{E}\left[\mathrm{pen}^{(i)}(m)\right] = \mathbb{E}\left[\mathrm{pen}^{(ii)}(m)\right] = \mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m, \mathbf{X})\right]$. Dans un cas une solution est proposée par l'heuristique de pente, dans l'autre le bootstrap à poids donne de nombreuses possibilités.

Heuristique de pente

Jusque dans les années 90, seuls des résultats asymptotiques (Shibata, 1981) donnent des garanties théoriques sur les choix de Akaike et Mallows. Les outils nécessaires à un traitement général et non-asymptotique des procédures pénalisées, les inégalités de concentration et en particulier celles pour les suprema de processus empiriques, sont apparus dans les années 90 à la suite des travaux de Ledoux et Talagrand (voir notamment Ledoux & Talagrand (1991); Talagrand (1996); Ledoux (2001)) sur le phénomène de concentration de la mesure. Ainsi, dans le sillage des travaux de Birgé et Massart (Birgé & Massart, 1997; Barron *et al.*, 1999), des principes généraux sont apparus sur la

[45] Le pénalité qui mène au critère idéal $\mathcal{C}_{\mathrm{id}}(m) = P\gamma(\widehat{s}_m)$.

[46] en densité mais aussi en régression.

calibration de la pénalité dans différents cadres statistiques. Outre la mise en évidence du rôle que joue la complexité de la collection $\mathcal{M}$ dans le choix de la pénalité, ils ont prouvé qu'il peut être désastreux de sous-pénaliser (voir Proposition 4.3 et Theorem 7.10 dans Massart (2007)), alors que légèrement sur-pénaliser peut être avantageux. De plus, lorsque la famille est trop grande, par exemple exponentielle[(47)] ou avec beaucoup de modèles de même dimension, Birgé & Massart (2007) montrent que le principe d'estimation sans biais du risque ne fonctionne plus. Il faut alors prendre une pénalité pen telle que $\mathbb{E}\left[\mathrm{pen}(m)\right] \geq \mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m, \mathbf{X})\right]$ pour compenser les fluctuations de $\mathrm{pen} - \mathrm{pen}_{\mathrm{id}}$ uniformément en $m \in \mathcal{M}$.

C'est à la suite de tous ces travaux qu'est née *l'heuristique de pente* qui sert à calibrer de façon optimale la pénalité. Dans le cadre de la régression gaussienne homoscédastique avec des modèles linéaires S_m de dimension finie, Birgé & Massart (2007) ont montré qu'il existe une *pénalité minimale* $\mathrm{pen}_{\mathrm{min}}$ et une *pénalité optimale* $\mathrm{pen}_{\mathrm{opt}}$ telles que :

- si pour un $\varepsilon > 0$ on choisit $\mathrm{pen}(m) \leq (1-\varepsilon)\,\mathrm{pen}_{\mathrm{min}}(m)$ pour tout $m \in \mathcal{M}$, alors $\widehat{m}(\mathcal{C}_{\mathrm{pen}})$ est de grande dimension et son risque est bien plus grand que celui de l'oracle.
- si pour un $\varepsilon > 0$ on a $\mathrm{pen}(m) \geq (1+\varepsilon)\,\mathrm{pen}_{\mathrm{min}}(m)$ pour tout $m \in \mathcal{M}$, alors $\widehat{m}(\mathcal{C}_{\mathrm{pen}})$ est de dimension bien moindre et il est possible d'obtenir une inégalité oracle avec constante $C_n > 1$.
- $2\,\mathrm{pen}_{\mathrm{min}}(m)$ est la pénalité optimale au sens où si $\mathrm{pen}(m) \approx 2\,\mathrm{pen}_{\mathrm{min}}(m)$ pour tout $m \in \mathcal{M}$, alors $\mathcal{C}_{\mathrm{pen}}$ mène à une inégalité oracle avec constante C_n presque 1.

Le "saut" de dimension qui se fait autour de la pénalité minimale est un phénomène de transition de phase. La recherche d'une généralisation de cette heuristique à d'autres cadres statistiques a provoqué un intérêt qui est toujours d'actualité comme l'attestent les récents articles publiés sur le sujet[(48)]. Arlot (2007) a étendu leurs résultats et montré que la pénalité optimale peut s'écrire $2 \times \mathrm{pen}_{\mathrm{min}}$ quelle que soit la forme de la pénalité minimale dans un cadre de régression hétéroscédastique. Plus tard, Arlot & Massart (2009) ont proposé un algorithme qui sert à calibrer la pénalité optimale dans un contexte plus général en cherchant d'abord à déterminer quand se produit la transition de phase et en multipliant ensuite la valeur de la constante devant la pénalité minimale par 2. Cet algorithme a ensuite été utilisé dans divers contextes (surtout dans un cadre gaussien) notamment la détection de ruptures (Lebarbier, 2005), la génétique (Villers, 2007), l'estimation de modèles graphiques (Verzelen, 2010), les modèles de mélanges (Maugis & Michel, 2011a,b), et la classification non-supervisée (Baudry, 2009). Pour l'estimation d'une densité, Lerasle (2012) (pour des estimateurs par projection avec contraste des moindres carrés) et Saumard (2010) (pour des histogrammes avec contraste du maximum de vraisemblance) ont également obtenu des garanties théoriques pour cette heuristique. Mais, à notre connaissance, aucun résultat n'existe pour les estimateurs linéaires en général. Ainsi l'existence d'une transition de phase autour d'une pénalité minimale semble se vérifier dans divers cadres statistiques. Toutefois, le fait que la pénalité optimale soit égale à deux fois la pénalité minimale n'est pas valable en toute généralité comme l'ont montré Arlot & Bach (2009) pour sélectionner parmi des estimateurs linéaires en

[(47)] c'est-à-dire plus grande qu'un nombre polynomial cn^{κ} pour toutes constantes $c, \kappa > 0$.

[(48)] Voir une bibliographie exhaustive dans les articles de survol (Baudry *et al.*, 2012; Arlot, 2015) dédiés à ce sujet.

régression. L'algorithme d'heuristique de pente appliqué tel quel ne fonctionne pas, mais des modifications permettent de le rectifier.

Pénalités rééchantillonnées

Une alternative à l'heuristique de pente nous vient des pénalités rééchantillonnées (Efron, 1983; Arlot, 2009). Le principe est simple : comme la pénalité idéale (1.30) s'écrit comme une fonctionnelle $\mathcal{L}(P, P_n)$, on peut l'estimer via l'heuristique d'Efron et le bootstrap à poids. Soit $W = (W_1, \dots, W_n)$ un n-échantillon indépendant de $\mathbf{X}$, tel que $\mathbb{E}[W_i] = 1$ pour tout i, et P_n^W la mesure empirique correspondante donnée par (1.19). On définit alors la **pénalité rééchantillonnée associée à** W (Arlot, 2007, Chapter 6) par

$$\mathrm{pen}_W(m) := \frac{1}{\mathbb{E}[(W_1 - 1)^2]} \mathbb{E}_W\left[\left(P_n - P_n^W \right) \gamma\left(\widehat{s}_m^W \right) \right] \quad , \tag{1.33}$$

où $\widehat{s}_m^W(x) = n^{-1} \sum_{i=1}^n W_i \mathcal{K}_m(X_i, x)$. Ainsi, lorsque la perte s'écrit au travers d'un contraste γ, le critère pénalisé correspondant est défini pour une constante $F > 0$ par

$$\mathcal{C}_{W,F}(m) := P_n \gamma\left(\widehat{s}_m \right) + F \, \mathrm{pen}_W(m) \ .$$

Outre la constante F, dont le rôle est de permettre une certaine sur-pénalisation en pratique, c'est donc le vecteur W qui définit toute procédure pénalisée de ce type. Une fois encore, il existe un nombre important de possibilités pour définir ce vecteur (dont on peut voir un bel aperçu dans l'introduction de Arlot (2007)). Un exemple important est lié au V-fold. On considère $(B_1, \dots, B_V)$ une partition de $[n]$ en V sous-ensembles de même taille et $J \sim \mathcal{U}(\{1, \dots, V\})$ une variable aléatoire indépendante de $\mathbf{X}$. Dans ce cas, Arlot (2008) définit la **pénalité V-fold** $\mathrm{pen}_{\mathrm{VF}}(m, V) := \mathrm{pen}_{W^{(\mathrm{VF})}}(m)$ en posant pour tout $i \in [n]$, $W_i^{(\mathrm{VF})} := V/(V-1)\mathbf{1}_{i \notin B_J}$. Ces poids sont de même loi et vérifient $\sum_{i=1}^n W_i^{(\mathrm{VF})} = n$ et $\mathbb{E}\left[(W_1 - 1)^2\right] = 1/(V-1)$. On peut l'écrire en toute généralité comme

$$\mathrm{pen}_{\mathrm{VF}}(m, V) := \frac{V-1}{V} \sum_{j=1}^V \left(\left(P_n - P_n^{(-j)} \right) \gamma\left(\widehat{s}_m^{(-j)} \right) \right) \ .$$

Cette pénalité a l'avantage d'estimer sans biais la quantité $\mathbb{E}[\mathrm{pen}_{\mathrm{id}}(m, \mathbf{X})]$ lorsque celle-ci s'écrit comme (1.32). En effet, on a $P_n - P_n^{(-j)} = V^{-1}(P_n^{(j)} - P_n^{(-j)})$, pour tout $j \in [V]$. Et, par indépendance des sous-échantillons $\mathbf{X}^{(B_j)}$ et $\mathbf{X}^{(-B_j)}$, et par (1.32), on trouve

$$\begin{aligned}
\mathbb{E}[\mathrm{pen}_{\mathrm{VF}}(m, V)] &= \frac{V-1}{V^2} \sum_{j=1}^V \mathbb{E}\left[\left(P_n^{(j)} - P_n^{(-j)} \right) \gamma\left(\widehat{s}_m^{(-j)} \right) \right] \\
&= \frac{V-1}{V} \mathbb{E}\left[\left(P_n^{(1)} - P_n^{(-1)} \right) \gamma\left(\widehat{s}_m^{(-1)} \right) \right] \\
&= \frac{V-1}{V} \mathbb{E}\left[\left(P - P_n^{(-1)} \right) \gamma\left(\widehat{s}_m^{(-1)} \right) \right] \\
&= \frac{V-1}{V} \mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}\left(m, \mathbf{X}^{(-B_1)} \right) \right]
\end{aligned}$$

$$= \frac{V-1}{V} \frac{1}{\left|\mathbf{X}^{(-B_1)}\right|} f(m) = \mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m, \mathbf{X})\right] \ . \tag{1.34}$$

Dans la suite, nous considérons la perte des moindres carrés et posons, pour tout $i, j \in [n]$,

$$\rho_{i,j}^{W} := \frac{\mathbb{E}\left[(W_i - 1)(W_j - 1)\right]}{\mathbb{E}\left[(W_1 - 1)^2\right]} \ .$$

La pénalité rééchantillonnée peut se réécrire en toute généralité comme suit

$$\mathrm{pen}_{W}(m) = \frac{2}{n^2} \left(\sum_{i=1}^{n} \mathcal{K}_m(X_i, X_i) + \sum_{1 \leqslant i \neq j \leqslant n} \rho_{i,j}^{W} \, \mathcal{K}_m(X_i, X_j) \right) \ . \tag{1.35}$$

Voici deux cas particuliers qui sont au cœur des travaux de cette thèse.

- On déduit de ce qui précède l'expression suivante pour les poids VF

$$\mathrm{pen}_{\mathrm{VF}}(m, V) := \frac{2}{n^2} \left(\sum_{i=1}^{n} \mathcal{K}_m(X_i, X_i) + \sum_{1 \leqslant i \neq j \leqslant n} \rho_{i,j}^{(\mathrm{VF})} \mathcal{K}_m(X_i, X_j) \right) \ , \tag{1.36}$$

  où si $i \in B_k$ et $j \in B_{k'}$, $\rho_{i,j}^{(\mathrm{VF})} := \rho_{i,j}^{W^{(\mathrm{VF})}} = 1 - \frac{V}{V-1} \mathbf{1}_{k \neq k'}$. Le *critère V-fold pénalisé* est défini pour une constante $F > 0$ par

$$\mathcal{C}_{V,F}^{\mathrm{pen}}(m) := P_n \gamma\left(\widehat{s}_m\right) + F \, \mathrm{pen}_{\mathrm{VF}}(m, V) \ .$$

- En supposant que les poids sont échangeables, on peut facilement montrer que $\rho_{i,j}^{W} = -1/(n-1)$ pour tout $i \neq j$ de sorte que toute pénalité associée à un vecteur échangeable s'écrit

$$\mathrm{pen}_{W}(m) = \frac{2}{n^2} \left(\sum_{i=1}^{n} \mathcal{K}_m(X_i, X_i) - \frac{1}{n-1} \sum_{1 \leqslant i \neq j \leqslant n} \mathcal{K}_m(X_i, X_j) \right) \ . \tag{1.37}$$

  Ceci implique en particulier que toutes les pénalités rééchantillonnées définies à partir d'un vecteur à poids échangeables sont équivalentes (Lerasle, 2012). Notons que les poids $W^{(\mathrm{VF})}$ sont échangeables si et seulement si $V = n$. Ainsi toute pénalité rééchantillonnée définie pour un vecteur à poids échangeables est égale, à constante multiplicative près, à la pénalité n-fold. La **pénalité leave-p-out**, notée $\mathrm{pen}_{\mathrm{LPO}}(m, p)$, provient du choix du poids *leave-p-out* $W_i^p = \frac{n}{n-p} \mathbf{1}_{i \in B}$, où $p \in [n-1]$ et B est un ensemble choisi aléatoirement selon la loi uniforme dans $\mathcal{E}_p$.

Le tableau suivant récapitule les différentes pénalités rééchantillonnées via un vecteur W de poids.

Validation croisée et pénalités rééchantillonnées : une même histoire ? Toute procédure (et en particulier toute procédure rééchantillonnée) $\mathcal{C}$ peut s'écrire comme $\mathcal{C}_{\mathrm{pen}}$ avec $\mathrm{pen}(m) =$

Poids	W quelconque	W échangeable $\longrightarrow W^p$	$W^{(\mathrm{VF})}$
Pénalité	(1.35)	(1.37)	(1.36)
Procédure $\mathcal{C}^{(i)}$	$\mathcal{C}_{W,F}(m)$	$\mathcal{C}_{W,F}(m) \longrightarrow \mathcal{C}^{\mathrm{pen}}_{p,F}(m)$	$\mathcal{C}^{\mathrm{pen}}_{V,F}(m)$

TABLE 1.4 – Pénalités rééchantillonnées mentionnées dans cette section.

$\mathcal{C}(m) - P_n\gamma(\widehat{s}_m)$. Des liens plus explicites existent entre les procédures de validation croisée de la Table 1.3 et les pénalités VF de la Table 1.4. Ainsi, pour une partition régulière de $[n]$ en V sous-échantillons de taille n/V, Arlot (2007) a montré que la relation suivante est vraie quel que soit le contraste et l'estimateur considéré

$$\mathcal{C}_V^{\mathrm{corr,VFCV}}(m) = \mathcal{C}^{\mathrm{pen}}_{V,1}(m) \ . \tag{1.38}$$

Pour les estimateurs par projection, d'autres relations ont été prouvées (Arlot & Lerasle, 2014, Lemme 1)

$$\mathcal{C}_V^{\mathrm{LSVF}}(m) = P_n\gamma\left(\widehat{s}_m\right) + \frac{2V-1}{2(V-1)}\,\mathrm{pen}_{\mathrm{VF}}(m,V) = \mathcal{C}^{\mathrm{pen}}_{V,\frac{2V-1}{2(V-1)}}(m) \tag{1.39}$$

$$\mathcal{C}_p^{\mathrm{LPO}}(m) = P_n\gamma\left(\widehat{s}_m\right) + \frac{2n/p-1}{2(n/p-1)}\,\mathrm{pen}_{\mathrm{VF}}(m,V) = \mathcal{C}^{\mathrm{pen}}_{V,\frac{2n/p-1}{2(n/p-1)}}(m) \ . \tag{1.40}$$

Ainsi, Arlot & Lerasle (2014) parviennent à étudier simultanément la VCVF classique, le LPO et les pénalités rééchantillonnées en étudiant uniquement le critère $\mathcal{C}^{\mathrm{pen}}_{V,F}$. La force de ces relations réside dans l'utilisation des techniques de preuve développées pour la sélection de modèles par pénalisation au profit de l'analyse de critères de validation croisée. Ceci est un atout majeur puisque des résultats fins (comme des inégalités d'oracle optimales) ont été prouvés pour les premières, alors que les secondes, malgré leur usage massif en pratique, n'ont jamais pu être analysées de manière générale du point de vue non-asymptotique. En outre, ceci permet de voir un critère de VC comme un "tout" et non plus comme une moyenne sur différents découpages. Ceci évite le piège[(49)] qui consiste à analyser d'abord ce qui se passe sur seule partition (ce qui est "facile" puisque les deux sous-échantillons sont indépendants) avant de s'essayer à dire quelque chose sur la moyenne (ce qui est nettement plus ardu car l'indépendance se perd, et avec elle l'avantage supposé de la VC sur le HO). Nous reviendrons plus longuement sur ces relations dans la prochaine section.

1.3 Principaux travaux et résultats

Au vu de ce qui précède, reprenons quelques remarques (déjà émises dans les sections précédentes) qui structureront la présentation des résultats de thèse ainsi que les chapitres du manuscrit.

- Pour l'estimation d'une densité, la qualité des procédures de rééchantillonnage qui estiment sans biais le risque n'est assurée que pour des cas particuliers (histogrammes ou estimateurs par projection). Les garanties théoriques manquent pour une classe plus large de candidats.

[(49)] dans lequel nous sommes tombés (voir le Chapitre 6).

- Il existe dans les faits un écart important entre compréhension théorique et utilisation pratique qu'illustre bien le passage du hold-out à la validation croisée. De manière générale, la question se pose de la comparaison théorique de deux procédures quelconques. Dans cette optique, les inégalités d'oracle sont insuffisantes et il paraît nécessaire d'aller au-delà du premier ordre.

- En densité, l'heuristique de pente gagnerait à être étudiée dans un cadre général afin de servir à la résolution de nouveaux problèmes (par exemple le choix d'une fenêtre dans une famille donnée).

- La T-estimation n'est pas utilisée en pratique pour la sélection d'un estimateur linéaire en densité, alors qu'elle repose sur la perte Hellinger qui jouit de nombreux avantages par rapport à la perte $\mathbb{L}_2$.

- Mise à part le cas simple du THO, il n'existe pas de procédure rééchantillonnée du type $\mathcal{C}^{(iii)}$, c'est-à-dire qui utilise la comparaison des candidats pour les juger et qui soit une alternative aux procédures fondées sur l'estimation du risque.

Deux grandes motivations distinctes apparaissent.

- Étudier les deux grandes solutions (de type $\mathcal{C}^{(i)}$ et $\mathcal{C}^{(ii)}$) à la sélection d'une méthode d'estimation qui reposent sur le principe d'estimation sans biais du risque, les procédures de validation croisée et les procédures pénalisées, avec pour objectif d'avoir une compréhension fine de leurs qualités du point de vue théorique afin de garantir (et si possible d'optimiser) leur utilisation en pratique[(50)].

- Développer une procédure rééchantillonnée, dans le cadre de la perte Hellinger (qui est intrinsèque au problème de l'estimation d'une densité), qui ne repose pas sur l'estimation sans biais du risque (de type $\mathcal{C}^{(iii)}$, donc) et qui soit implémentable en pratique.

1.3.1 Vue d'ensemble

Étude des procédures d'estimation sans biais du risque dans le cadre des moindres carrés. Dans la Partie I, nous considérons la procédure $\mathcal{C}_{\mathrm{pen}}$ et proposons une procédure du type $\mathcal{C}^{(ii)} = \mathcal{C}_{\mathrm{pen}_{\mathrm{opt}}}$ (donc sans rééchantillonnage) qui est optimale pour sélectionner un estimateur linéaire. De plus, nous étudions l'heuristique de pente, prouvons l'existence d'une pénalité minimale $\mathrm{pen}_{\mathrm{min}}$ et infirmons la règle "$\mathrm{pen}_{\mathrm{opt}} = 2 \times \mathrm{pen}_{\mathrm{min}}$" en général.

La Partie II est dédiée à l'étude de procédures du type $\mathcal{C}^{(i)}$: $\mathcal{C}_V^{\mathrm{LSVF}}$, $\mathcal{C}_{V,F}^{\mathrm{pen}}$ (et par conséquent $\mathcal{C}_V^{\mathrm{corr,LSVF}}$) et $\mathcal{C}_p^{\mathrm{LPO}}$. Nous obtenons pour celles-ci des inégalités d'oracle optimales au premier ordre, pour une collection quelconque de méthodes d'estimation à noyau. On en déduit un résultat d'adaptation aux classes de Sobolev qui n'a, à notre connaissance, pu être prouvé autrement que par la méthode par blocs de Stein. C'est l'objet du Chapitre 3.

Mais les inégalités d'oracle ne suffisent pas à décider du choix de V pour les méthodes de VCVF ou à comparer deux procédures du point de vue théorique (Arlot & Lerasle, 2014, Section

[(50)] nous nous plaçons ainsi dans la droite lignée des travaux de Sylvain Arlot, qui sont le point de départ de cette thèse.

4). Nous donnons dans le Chapitre 4 les calcul des variances des incréments $\mathrm{Var}\left[\Delta_{\mathcal{C}}(m, m')\right]$ pour chacune de ces procédures $\mathcal{C}^{(i)}$, c'est-à-dire la différence entre les valeurs qu'elle prend en deux estimateurs quelconques, et étudions l'heuristique (1.13) ainsi que le rôle de V lorsqu'on utilise la VFCV pour choisir une fenêtre.

Si la plupart du temps nous généralisons de récents travaux aux méthodes d'estimation à noyau (c'est-à-dire aux estimateurs linéaires), l'une des valeurs ajoutées de la généralité de notre approche est qu'elle permet de traiter le problème très connu de la sélection de la fenêtre pour un estimateur de Parzen-Rosenblatt par des techniques de rééchantillonnage et par pénalisation.

Développement d'une VCVF pour le cadre Hellinger. La Partie III se place dans un contexte qui semble plus approprié à l'estimation d'une densité, puisque la perte utilisée est la perte Hellinger qui est intrinsèquement liée à ce cadre. Il faut restreindre ici le type d'estimateurs considérés, puisqu'il faut supposer que les méthodes d'estimation à noyau produisent des estimateurs linéaires qui soient des densités. Ainsi de nombreux estimateurs par projection sont *a priori*[51] exclus de ce cadre.

Nous proposons des procédures du type $\mathcal{C}^{(iii)}$, c'est-à-dire rééchantillonnées et qui reposent sur la comparaison des candidats, qui soient implémentables en pratique. Dans le Chapitre 5 nous développons un algorithme qui permet d'implémenter la procédure $\mathcal{C}^{\mathrm{THO}}$, introduite par Birgé (2006a), avec un coût algorithmique raisonnable (la condition *sine qua non* pour évaluer les performances de cette procédure en pratique).

L'étape suivante consiste à introduire et étudier la procédure VF (que nous notons $\mathcal{C}_V^{\mathrm{TVF}}$) qui se déduit de la procédure $\mathcal{C}^{\mathrm{THO}}$ selon l'heuristique (1.28). Nous développons un nouvel algorithme qui calcule exactement l'estimateur final en évitant un coût exorbitant qui semblait condamner l'implémentation de la procédure en pratique. La visée du Chapitre 6 est donc triple. L'étude est similaire à celle entreprise pour le cas des moindres carrés : recherche d'inégalités d'oracle, compréhension du rôle de V du point de vue théorique et pratique, comparaison aux autres VF.

- Le Chapitre 2 est un papier écrit en collaboration avec Matthieu Lerasle et Patricia Reynaud-Bouret. Il généralise une analyse entamée par Lerasle (2012). Nous l'avons soumis au "Proceedings of the High Dimensional Probability VII meeting".

- Les Chapitre 3 et Chapitre 4 résultent d'un travail en cours avec Sylvain Arlot et Matthieu Lerasle. Ils étendent les idées de Arlot & Lerasle (2014) aux estimateurs linéaires (nous retrouvons en effet leurs résultats, concernant les estimateurs par projection, comme cas particuliers des nôtres).

- Le Chapitre 5 est le fruit d'un travail mené entre septembre 2012 et le début de l'année 2014 avec Yves Rozenholc. Pendant cette période nous avons également développé un package R[52] qui implémente notre algorithme. Cet article a été soumis au "Journal of the American Statistical Association" en février 2015.

[51] On peut toutefois espérer les inclure dans notre analyse en utilisant une technique de troncature et en se restreignant à la partie positive.

[52] donc disponible en libre accès sur le site `http://cran.r-project.org/web/packages/Density.T.HoldOut/index.html`.

- Le Chapitre 6 est un travail en cours avec Lucien Birgé et Pascal Massart.
- Nous concluons la thèse par un travail de recherche personnel, complètement ouvert, qui nous incite à ne pas conclure ! Au cours de la rédaction de ce manuscrit, il est apparu en effet qu'on peut lier les deux dernières parties de cette thèse et s'appuyer sur les différentes heuristiques ainsi que sur les forces de chaque chapitre pour introduire une nouvelle procédure rééchantillonnée qui puisse jouir des avantages de chacun. Celle-ci est présentée dans la conclusion de ce manuscrit et illustrée par une étude empirique.

1.3.2 Cadre des moindres carrés (Partie I et Partie II)

Le cadre est le même dans les trois chapitres que contiennent les deux parties. Nous rappelons que $\ell(s,t) = \|t-s\|_2^2$ et $\mathcal{S} = \mathbb{L}_2(\mu) \cap \mathbb{L}_1(\mu)$. Nous disposons d'une famille $(\mathcal{K}_m)_{m\in\mathcal{M}}$ de noyaux à laquelle est naturellement associée, par la relation (1.6), la famille d'estimateurs linéaires $(\widehat{s}_m)_{m\in\mathcal{M}}$. Dans la première partie il faut réutiliser le même échantillon pour sélectionner un estimateur, alors que dans l'autre on peut profiter des techniques de rééchantillonnage (i.e. proposer respectivement des procédures du type $\mathcal{C}^{(ii)}$ et $\mathcal{C}^{(i)}$ selon la Table 1.2). Dans tous les cas, nous faisons les mêmes hypothèses sur les noyaux (voir Section 2.3.1), la principale étant qu'il existe une constante $\Gamma \geq 1$ telle que, pour tout $m \in \mathcal{M}$,

$$\sup_{x\in\Xi} \int_\Xi \mathcal{K}_m(x,y)^2 d\mu(y) \ \vee \sup_{(x,y)\in\Xi^2} |\mathcal{K}_m(x,y)| \leq \Gamma n \ .$$

Cette condition est plus forte que celle utilisée par Devroye & Lugosi (2001) dont les estimateurs (dits *additifs réguliers*) vérifient, pour tout $x \in \Xi$, $\mathbb{E}\left[|\mathcal{K}_m(x,X)|\right] < \infty$.

Au vu de (1.31), nous devons contrôler les déviations de $\mathrm{pen}(m) - \mathrm{pen}_{\mathrm{id}}(m,\mathbf{X})$ pour différentes pénalités pen[(53)], uniformément sur tous les $m \in \mathcal{M}$. Nous n'utilisons pour ce faire que deux inégalités de concentration : une variante de l'inégalité de Bernstein et une inégalité de concentration pour les U-statistiques d'ordre deux (voir Section 2.2.3). Notons pour tout $m \in \mathcal{M}$, pour tout $(x,y) \in \Xi^2$

$$A_m(x,y) := \int_\Xi \mathcal{K}_m(x,z)\mathcal{K}_m(z,y)d\mu(z), \qquad \text{et} \qquad \Theta_m(x) := A_m(x,x) \ .$$

Sélection optimale et pénalité minimale

Le premier chapitre de la thèse est le seul qui soit consacré au problème de la sélection d'estimateurs (et non pas à une méthode d'estimation) et par conséquent le seul où n'apparaît pas l'idée du découpage des données pour mieux évaluer la qualité des estimateurs. Nous étudions la procédure $\mathcal{C}_{\mathrm{pen}}$, telle qu'elle a été définie en (1.29), pour sélectionner parmi une famille finie d'estimateurs linéaires. L'objectif du travail est double. D'une part, donner un choix de pénalité optimale, c'est-à-dire qui mène à une inégalité oracle optimale au premier ordre. D'autre part, montrer l'existence d'une pénalité minimale en mettant en lumière un phénomène de transition de phase autour d'une

[(53)] notamment les pénalités $\mathrm{pen}_{\mathrm{VF}}(m,V)$ données en (1.36) ainsi que $\mathrm{pen}_V^{\mathrm{LSVF}}(m) = \mathcal{C}_V^{\mathrm{LSVF}}(m) - P_n\gamma(\widehat{s}_m)$, et $\mathrm{pen}_p^{\mathrm{LPO}}(m) = \mathcal{C}_p^{\mathrm{LPO}}(m) - P_n\gamma(\widehat{s}_m)$.

certaine valeur de la pénalité. Notons pour tout $m \in \mathcal{M}$ et $x \in \Xi$

$$s_m(x) := \int_\Xi \mathcal{K}_m(y,x)s(y)d\mu(y), \qquad \text{et} \qquad \chi_m(x) := \mathcal{K}_m(x,x) \ .$$

Notre analyse repose sur une compréhension fine du comportement de la pénalité idéale dont on peut montrer qu'elle se concentre autour d'une certaine valeur déterministe :

$$\mathrm{pen}_{\mathrm{id}}(m, \mathbf{X}) \simeq 2\left(\frac{P\chi_m - Ps_m}{n}\right) \ .$$

Comme le terme inconnu Ps_m/n s'avère, sous une hypothèse assez faible, être négligeable, notre premier résultat (Théorème 2.1) montre que le choix

$$\mathrm{pen}_{\mathrm{opt}}(m) = 2\frac{P\chi_m}{n}$$

mène à une procédure de pénalisation optimale. Notons qu'il peut toujours être estimé par l'estimateur sans biais $\mathrm{pen}(m) = 2P_n\chi_m/n$ mais qu'il est explicite dans les cas les plus connus.

Le deuxième théorème majeur (Théorème 2.2) concerne l'existence de la pénalité minimale $\mathrm{pen}_{\min}(m) = (2P\chi_m - P\Theta_m)/n$ pour la sélection d'estimateurs linéaires. Toute la difficulté réside dans la signification d'un tel résultat dans un cadre aussi général. Dans leurs travaux en sélection de modèles, Birgé & Massart (2007) (en régression) et Lerasle (2012) (en densité) ont montré un "saut" de la dimension pour exhiber une transition de phase et faire apparaître la pénalité minimale. Mais ceci est spécifique au cas des estimateurs par projection. *Que* devient la dimension (aussi appelée complexité) en toute généralité ? *Qui* fait ce "saut" dans notre cadre ? Un premier résultat (Proposition 2.6) nous garantit que la décomposition biais-variance mentionnée à la Section 1.1.2 reste vraie avec grande probabilité en général. Comme expliqué auparavant, le terme $\|s - s_m\|_2^2$ est le terme de biais et nous suggérons de voir $P\Theta_m/n$ comme le terme de variance. Nous justifions ce choix par le fait que dans les exemples des estimateurs par projection et de Parzen-Rosenblatt, $P\Theta_m/n$ est une mesure naturelle de la complexité. Pour des estimateurs histogrammes ce terme correspond à la dimension de l'espace considéré, alors que pour les estimateurs du type Parzen-Rosenblatt, il est proportionnel à l'inverse de la fenêtre considérée. On voit que cette interprétation est cohérente avec l'explication classique donnée en Section 1.1.4.

Ainsi, le but est de prouver que la valeur de $P\Theta_{\widehat{m}}$ est sensiblement modifiée lorsque la pénalité passe de $(1-\varepsilon)\,\mathrm{pen}_{\min}(m)$ à $(1+\varepsilon)\,\mathrm{pen}_{\min}(m)$ pour un $\varepsilon > 0$. Pour cela il nous faut supposer qu'il existe au moins un estimateur avec un biais petit. En effet il apparaît que si tous les biais sont prépondérants sur les parties variance correspondantes, l'estimateur $\widehat{s}_{\widehat{m}(\mathcal{C}_{\mathrm{pen}})}$ vérifie forcément une inégalité oracle asymptotiquement optimale (Corollaire 2.1) et il ne peut y avoir de pénalité minimale. Nous montrons que cette hypothèse est vérifiée sur les deux exemples que nous suivons depuis le début. Les conclusions les concernant peuvent se résumer comme suit.

- Pour les estimateurs par projection, nous avons $A_m(x,y) = \mathcal{K}_m(x,y)$ pour tout $x, y \in \Xi$. En particulier $\Theta_m = \chi_m$ et $2P\chi_m - P\Theta_m = P\chi_m$ de sorte que nous validons l'heuristique $\mathrm{pen}_{\mathrm{opt}} = 2 \times \mathrm{pen}_{\min}$.

- Pour les estimateurs de Parzen-Rosenblatt on a, pour tout $x \in \Xi$, $\chi_m(x) = k(0)/h_m$ et $\Theta_m(x) = \|k\|_2^2 / h_m$. Ainsi la pénalité optimale $2P\chi_m/n = 2k(0)/(nh_m)$ est connue, et la pénalité minimale $(2k(0) - \|k\|_2^2)/(nh_m)$ peut être négative si $\|k\|_2^2 > 2k(0)$. Dans ce cas, tout $\widehat{m}$ qui minimise le risque empirique $P_n\gamma(\widehat{s}_m)$ satisfait une inégalité oracle et il n'est plus nécessaire de pénaliser ! L'heuristique de pente n'est donc plus vraie telle quelle.

LSVF et pénalisation V-fold

La Partie III peut se voir comme l'extension du papier de Arlot & Lerasle (2014) au cas des estimateurs linéaires. L'inconvénient de notre généralisation est que nous ne pouvons plus analyser simultanément le LSVF, les pénalités V-fold et les pénalités rééchantillonnées. Autrement dit, si le lien $\mathcal{C}_V^{\mathrm{corr,LSVF}} = \mathcal{C}_{V,1}^{\mathrm{pen}}$ est toujours vrai, nous n'avons plus l'équivalent des formules (1.39) et (1.40). Cependant, on prouve (voir Lemme 3.1) les relations suivantes pour les pénalités issues des critères $\mathcal{C}_V^{\mathrm{LSVF}}$ et $\mathcal{C}_p^{\mathrm{LPO}}$

$$\mathrm{pen}_V^{\mathrm{LSVF}}(m) = \mathrm{pen}_{\mathrm{VF}}(m,V) + \frac{1}{n^2(V-1)} \left(\sum_{i=1}^n \Theta_m(X_i) + \sum_{1 \leqslant i \neq j \leqslant n} \rho_{i,j}^{(\mathrm{VF})} A_m(X_i, X_j) \right) ,$$

$$\mathrm{pen}_p^{\mathrm{LPO}}(m) = \mathrm{pen}_{\mathrm{VF}}(m,n) + \frac{p}{n^2(n-p)} \left(\sum_{i=1}^n \Theta_m(X_i) - \frac{1}{n-1} \sum_{1 \leqslant i \neq j \leqslant n} A_m(X_i, X_j) \right) .$$

Afin de concentrer uniformément $\mathrm{pen} - \mathrm{pen}_{\mathrm{id}}$ pour ces trois pénalités, nous utilisons encore des inégalités de concentration, notamment pour contrôler les U-statistiques d'ordre deux (voir Proposition 2.2) qui se déduisent de

$$\sum_{1 \leqslant i \neq j \leqslant n} B_{i,j}\, K_m(X_i, X_j) \qquad \text{et} \qquad \sum_{1 \leqslant i \neq j \leqslant n} B'_{i,j}\, A_m(X_i, X_j) ,$$

où différents $B_{i,j}$ et $B'_{i,j}$ peuvent apparaître selon la procédure considérée. Nous obtenons ainsi des inégalités d'oracle du type (1.4) pour chacune de ces procédures (Théorème 3.1 et Théorème 3.2). Nous pouvons faire ici quelques commentaires sur ces résultats[(54)].

D'abord on remarque que, dans tous les cas, le terme de reste R_n peut être facilement majoré en utilisant une collection polynomiale d'estimateurs, c'est-à-dire en supposant qu'il existe une constante $a' > 0$ telle que pour tout $n \in \mathbb{N}^\star$, on a $|\mathcal{M}| = |\mathcal{M}_n| \leqslant n^{a'}$. Concernant la constante multiplicative C_n, la discussion est plus délicate.

- Pour la procédure $\mathcal{C}_{V,F}^{\mathrm{pen}}$, on trouve $C_n = (1 + (\delta_{\widehat{m}})_+)/(1 - (\delta_m)_-) + \mathrm{o}(1)$, avec $\delta_m = 2(F-1)\gamma_m$ et $\gamma_m := \mathbb{E}\left[K_m(X,X)\right]/P\Theta_m$, de sorte que le comportement au premier ordre de la majoration de la perte s'explique par la quantité δ_m qui dépend de m en général[(55)]. Un résultat asymptotiquement optimal n'est donc possible que si $\delta_m = \mathrm{o}(1)$. Le critère de

[(54)] Lorsqu'on se restreint aux noyaux par projection, on a $K_m = A_m$ de sorte que le Lemme 3.1 implique les formules de Arlot & Lerasle (2014) pour les estimateurs par projection et il n'y a plus qu'une seule U-statistique à traiter. Nous retrouvons par conséquent l'intégralité de leurs résultats comme cas particulier de notre analyse.

[(55)] pour les estimateurs par projection et pour les estimateurs avec noyau k fixé, γ_m ne dépend pas de m : on trouve respectivement $\gamma_S = 1$ et $\gamma_h = k(0)/\|k\|_2^2$.

Burman, qui est le cas particulier où $F = 1$ et $\delta_m = 0$, satisfait par conséquent une inégalité oracle non-asymptotique optimale au premier ordre (pour une collection polynomiale par exemple).

- Les procédures $\mathcal{C}_V^{\mathrm{LSVF}}$ et $\mathcal{C}_p^{\mathrm{LPO}}$ sont toujours biaisées de sorte que le terme γ_m doit être contrôlé au cas par cas. Si une majoration est possible sur $\max_{m \in \mathcal{M}} \gamma_m$, alors on peut déduire une inégalité oracle asymptotiquement optimale quand $V = V_n \to \infty$ pour $\widehat{s}_{\widehat{m}(\mathcal{C}^{\mathrm{LSVF}})}$ et quand $p = p_n$ vérifie $p_n/(n - p_n) \to \infty$ pour $\widehat{s}_{\widehat{m}(\mathcal{C}^{\mathrm{LPO}})}$.

Nous illustrons l'attrait de l'inégalité oracle optimale obtenue pour la procédure $\mathcal{C}_{V,1}^{\mathrm{pen}}$ en prouvant des résultats adaptatifs pour l'estimateur qu'elle sélectionne. Nous supposons, comme première application, que la vraie densité appartient à une classe de Sobolev $\mathcal{F}_{\alpha_0}$ où $\alpha_0 = (\beta_0, Q_0)$ appartient à un ensemble $\aleph = (1/2, a_n) \times (0, b_n)$, avec a_n et b_n qui dépendent uniquement de n. Nous rappelons d'abord (Proposition 3.2) que pour tout $\alpha = (\beta, Q)$ dans $\aleph$, l'estimateur de Pinsker $\widehat{s}_\alpha$ est minimax pour $\mathcal{F}_\alpha$. Celui-ci dépend des paramètres inconnus β et Q et ne peut être utilisé en pratique. Nous définissons ensuite $\mathcal{M}$ en discrétisant les intervalles $(1/2, a_n)$ et $(0, b_n)$ pour que l'estimateur $\widehat{s}_{\widehat{m}}$ soit bien minimax pour $\mathcal{F}_{\alpha_0}$. La procédure $\mathcal{C}_{V,1}^{\mathrm{pen}}$ sert finalement à sélectionner les paramètres (β, Q) dans la grille $(\beta_j, Q_k)_{(j,k) \in \mathcal{M}}$ et résulte en un estimateur adaptatif (Corollaire 3.1) sur les classes de Sobolev $\bigcup_{\alpha \in \aleph} \mathcal{F}_\alpha$. Comme deuxième application, nous considérons une famille d'estimateurs par projection à poids et montrons que l'estimateur sélectionné est adaptatif sur un ensemble défini par des poids décroissants (voir Corollaire 3.2). Si ces résultats sont légèrement moins forts que ceux obtenus par Rigollet (2006a), ils n'avaient à notre connaissance pu être prouvés à l'aide de procédures de sélection d'estimateurs par rééchantillonnage.

Enfin, nous effectuons une étude de simulations (Section 3.6) afin d'analyser la qualité des différentes procédures VF pour le problème du choix de la fenêtre ainsi que pour le choix d'un estimateur dans une famille d'estimateurs linéaires provenant de noyaux d'approximation et par projection. Nous donnons, au passage, des formules exactes pour les pénalités leave-one-out pour les estimateurs histogrammes et de Parzen-Rosenblatt avec noyau gaussien.

Malheureusement, ces inégalités d'oracle nous donnent uniquement une information au premier ordre (Arlot & Lerasle, 2014) et s'avèrent inutiles lorsqu'on veut comparer deux procédures du type V-fold. En effet, par (1.34) on a pour tout $V \in \{2, \dots, n\}$,

$$\mathbb{E}\left[\mathcal{C}_{V,1}^{\mathrm{pen}}(m)\right] = \mathbb{E}\left[\mathcal{C}_{\mathrm{pen}_{\mathrm{id}}}(m)\right], \quad \text{pour tout } m \in \mathcal{M} \ .$$

Ainsi, nous avons prouvé que les procédures $\mathcal{C}_{2,1}^{\mathrm{pen}}$ et $\mathcal{C}_{n,1}^{\mathrm{pen}}$ sont optimales mais nous ne pouvons les distinguer au premier ordre puisqu'elles ont même biais. Le Chapitre 4 cherche à comprendre, à partir des travaux de Arlot et Lerasle, pourquoi et à quel point $\mathcal{C}_{n,1}^{\mathrm{pen}}$ est meilleur que $\mathcal{C}_{2,1}^{\mathrm{pen}}$, et plus généralement quel V il faut choisir pour optimiser la qualité de $\mathcal{C}_{V,1}^{\mathrm{pen}}$.

Si on suppose que $\mathcal{C}_1$ et $\mathcal{C}_2$ ont même biais, et

$$\overline{m} = \operatorname*{argmin}_{m \in \mathcal{M}} \mathbb{E}\left[\mathcal{C}_1(m)\right] = \operatorname*{argmin}_{m \in \mathcal{M}} \mathbb{E}\left[\mathcal{C}_2(m)\right] \ ,$$

alors on trouve d'après l'heuristique (1.13)

$$\text{si } \mathrm{Var}\left[\Delta_{\mathcal{C}_1}(m,m')\right] < \mathrm{Var}\left[\Delta_{\mathcal{C}_2}(m,m')\right] \ \forall m \neq m', \ \text{ alors } \mathcal{C}_1 \text{ est meilleure que } \mathcal{C}_2 \ . \quad (1.41)$$

Dans cette optique, le but du Chapitre 4 est de donner les expressions exactes (voir Théorème 4.3 et Théorème 4.1) des variances de $\Delta_{\mathcal{C}}$ pour $\mathcal{C} = \mathcal{C}_{V,F}^{\mathrm{pen}}, \mathcal{C}_V^{\mathrm{corr,LSVF}}$ et $\mathcal{C}_V^{\mathrm{LSVF}}$. En particulier nous retrouvons le Théorème 2 de Arlot & Lerasle (2014) qui concerne les estimateurs par projection. Dans le cas des histogrammes réguliers ils montrent que pour les m, m' qui "comptent"[(56)], on a, pour $\mathcal{C} = \mathcal{C}_{V,1}^{\mathrm{pen}}$ et $\mathcal{C} = \mathcal{C}_V^{\mathrm{LSVF}}$, $\mathrm{Var}\left[\Delta_{\mathcal{C}}(m,m')\right] \simeq (1 + 4/(V-1) - 1/n)\kappa_1 + \kappa_2$ où κ_1, κ_2 ne dépend pas de V. Ainsi, si augmenter V améliore les performances du critère $\mathcal{C}_{V,1}^{\mathrm{pen}}$, cette amélioration se limite au gain d'une constante multiplicative dans la variance. Dans un cadre aussi général que celui des estimateurs linéaires, il est plus difficile d'interpréter l'ordre de grandeur des différents termes qui apparaissent dans l'expression des variances. Néanmoins, il est remarquable que pour le critère $\mathcal{C}_{V,F}^{\mathrm{pen}}$ (Théorème 4.3) l'influence du facteur V ne dépende que du signe de

$$\mathrm{Var}\left[\mathcal{K}_m(X,Y) - \mathcal{K}_{m'}(X,Y)\right] - 2\,\mathrm{Cov}\left[\mathcal{K}_m(X,Y) - \mathcal{K}_{m'}(X,Y), \mathcal{K}_m(Y,Z) - \mathcal{K}_{m'}(Y,Z)\right] \ ,$$

qui s'avère être toujours positif (voir Lemme 4.1). Ainsi la variance $\mathrm{Var}\left[\Delta_{V,1}(m,m')\right]$ est une fonction décroissante en V ce qui garantit que la performance du critère de Burman s'améliore avec V. De plus, comme pour les estimateurs par projection, l'amélioration ne se fait qu'au second ordre dans la variance. En ce qui concerne la procédure $\mathcal{C}_V^{\mathrm{LSVF}}$, on trouve le même type de conclusion (sous des conditions techniques qui imposent que le premier terme dans le Théorème 4.1 soit bien le terme dominant) : asymptotiquement le biais et la variance diminuent avec V. Nous retrouvons également les calculs de variance de Celisse (2014) pour la procédure $\mathcal{C}_p^{\mathrm{LPO}}$ (voir Théorème 4.2).

Nous illustrons nos calculs (Section 4.4) sur une collection d'estimateurs de Parzen-Rosenblatt en remplaçant la dimension D_m d'un modèle par la fenêtre h_m^{-1}. Les conclusions de notre étude confirment l'heuristique (1.41) pour ces estimateurs et affirment une fois de plus qu'il n'est pas nécessaire de prendre V supérieur à 10 pour avoir une bonne procédure V-fold.

Une compréhension fine des différents termes qui apparaissent dans les théorèmes constitue la prochaine étape de ce travail. Des simulations supplémentaires seront nécessaires pour comprendre le rôle de V. Au moins deux scénarios prennent forme : soit l'analyse de Arlot & Lerasle (2014) peut se généraliser, et on pourra garantir à l'utilisateur que dans tous les cas les procédures VF s'améliorent quand V augmente, soit on trouvera un cadre particulier dans lequel la variance des incréments augmente avec V, ce qui poussera à affiner davantage la théorie.

1.3.3 Cadre Hellinger (Partie III)

Dans cette sous-section, la perte $\ell = h^2$ est donnée par (1.1) et $\mathcal{S}$ est l'ensemble des densités de probabilité par rapport à la mesure μ. Ce cadre a initialement été choisi afin d'observer la différence, en pratique, entre le T-estimateur et l'estimateur par maximum de vraisemblance qui n'est pas robuste (car le test de rapport de vraisemblance ne l'est pas, voir Birgé (2006a)). Voici

[(56)] les paires importantes sont les $(m,m') = (m,\overline{m})$ avec $\mathbb{E}\left[\ell(s,\widehat{s}_m) - \ell(s,\widehat{s}_{\overline{m}})\right]$ qui soit ni trop grand ni trop petit.

les deux statistiques de test que nous considérons dans ce travail pour choisir entre deux densités par le test (1.16).

- **Test de boule.** Le test de boule est apparu à la suite des travaux de Huber (1965) et Le Cam (1973) qui ont montré qu'il est possible d'obtenir un test robuste pour tester deux ensembles convexes disjoints dans $\mathcal{P}$ (et en particulier, deux boules) en effectuant un test entre la paire de points la moins favorable qui est la paire de probabilités la plus proche entre ces deux ensembles (intuitivement la plus difficile à tester). Pour tester $\mathcal{B}(t,r)$ contre $\mathcal{B}(u,r)$, Birgé (1984b) donne l'expression explicite de ce test

$$T_{t,u}(\mathbf{X}) = \sum_{i=1}^{n} \log\left(\frac{\sin(\omega(1-\theta))\sqrt{t}(X_i) + \sin(\omega\theta)\sqrt{u}(X_i)}{\sin(\omega(1-\theta))\sqrt{u}(X_i) + \sin(\omega\theta)\sqrt{t}(X_i)}\right) \quad \text{où } \omega = \arccos\rho(t,u) \ . \tag{1.42}$$

  Le cas $\theta = 0$ correspond ainsi au test de rapport de vraisemblance entre t et u.

- **Formulation variationnelle.** Soient $t, u \in S$ et $r = 1/2(t+u)$. En partant d'une formulation variationnelle de l'affinité de Hellinger, Baraud (2011) propose d'utiliser la statistique de test suivante

$$T_{t,u}(\mathbf{X}) = \frac{1}{2}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\sqrt{t}(X_i) - \sqrt{u}(X_i)}{\sqrt{r}(X_i)} + \int\left(\sqrt{t}(x) - \sqrt{u}(x)\right)\sqrt{r}(x)d\mu(x)\right) \ . \tag{1.43}$$

La preuve de la propriété de robustesse du test de boule (c'est-à-dire le fait qu'il vérifie (1.17) et (1.18)) se trouve dans Birgé (1984b) ainsi que dans le Corollaire 1 de Birgé (2014). Le test de Baraud satisfait aussi ces inégalités (à condition toutefois que θ et a soient petits). Ce résultat, prouvé par Baraud et Sart, n'a jamais été publié dans le cadre de la densité (une inégalité similaire a toutefois été prouvée par Sart (2011) dans un cadre poissonien).

Il nous faut ajouter une hypothèse supplémentaire sur les méthodes d'estimation à noyau pour utiliser les tests robustes. Effectivement ceux-ci sont définis pour des densités uniquement et ne peuvent être utilisés pour des estimateurs pouvant prendre des valeurs négatives ou ne s'intégrant pas à un. Nous avons donc supposé que pour tout $m \in \mathcal{M}$, $\widehat{s}_m$ est un estimateur du type (1.6) avec

$$\mathcal{K}_m(x,y) \geq 0 \quad \forall x, y \in \Xi \qquad \text{et} \qquad \int_{\Xi} \mathcal{K}_m(x,y)d\mu(x) = 1 \quad \forall y \in \Xi \ . \tag{1.44}$$

Ceci est une restriction relativement importante étant donné que de nombreux estimateurs par projection ne sont pas positifs. Néanmoins les histogrammes et les estimateurs de Parzen-Rosenblatt avec $k \geq 0$ peuvent encore être considérés dans la collection initiale. Devroye & Györfi (1985) font la même hypothèse[57], ainsi que Goldenshluger & Lepski (2011), lorsqu'ils généralisent leur procédure aux estimateurs linéaires en général.

[57] on peut lire au tout début de leur livre "we believe that a density should be estimated by a density".

Implémentation de la procédure THO

La motivation première du Chapitre 5 est de se faire une idée des performances réelles de procédures du type T-estimation en pratique[58]. Malheureusement, le calcul des T-estimateurs en un temps raisonnable paraît utopique lorsqu'on se place dans le cadre général de la sélection de modèles. Il faut en effet discrétiser des espaces (potentiellement très "gros"), calibrer le pas de discrétisation ainsi que le paramètre z dans (1.16), pour ensuite tester deux par deux les points des réseaux et sélectionner l'estimateur ayant le plus petit critère (1.14). C'est pourquoi nous avons décidé de nous focaliser sur la mise en pratique du cas le plus simple : la procédure $\mathcal{C}^{\mathrm{THO}}$. Comme c'est souvent le cas pour le HO, cette procédure possède de bonnes garanties théoriques. Ainsi, le Corollaire 9 de Birgé (2006a) prouve que l'estimateur provenant de $\widehat{m}(\mathcal{C}^{\mathrm{THO}})$ satisfait une inégalité oracle dans un cadre très général qui couvre l'estimation d'une densité. Cependant, même dans cette situation très simple, le coût algorithmique reste problématique. En effet, pour une collection $S = \{\widehat{s}_m^{(E)}, m \in \mathcal{M}\}$ de $M = |\mathcal{M}|$ estimateurs, il faut potentiellement effectuer de l'ordre de M^2 tests alors que les procédures $\mathcal{C}^{\mathrm{HO}}$ demandent au pire M opérations[59].

Le premier résultat est l'implémentation, pour tout sous-ensemble d'entraînement $E \subset [n-1]$, de la procédure $\mathcal{C}_E^{\mathrm{THO}}$ décrite par (1.23). L'idée de notre algorithme (voir Figure 5.1) est très simple. En partant de (1.15), nous en déduisons que le T-estimateur vérifie la propriété suivante : pour tout sous-ensemble $J \subset [M]$,

$$\widehat{s}^{(E)}_{\widehat{m}(\mathcal{C}^{\mathrm{THO}})} \in \bigcap_{l \in J} \left(S \bigcap \overline{\mathcal{B}} \left(\widehat{s}_l^{(E)}, \mathcal{C}_E^{\mathrm{THO}}(l)) \right) \right) \ ,$$

où $\overline{\mathcal{B}}(\widehat{s}_l^{(E)}, r) = \{t \in \mathcal{S} \text{ t.q. } h(t, \widehat{s}_l^{(E)}) \leq r\}$. Ainsi, l'algorithme intersecte des boules de Hellinger $\overline{\mathcal{B}}(\widehat{s}_l^{(E)}, \mathcal{C}_E^{\mathrm{THO}}(l)))$ jusqu'à ce qu'il n'y ait plus qu'un seul point dans l'intersection, le T-estimateur. En procédant de la même façon, celui-ci pourrait, en théorie, être appliqué à la T-estimation en général[60]. Nous n'avons pu démontrer rigoureusement que le nombre de tests calculés est sous-quadratique, mais il ressort de nos milliers de simulations que c'est très nettement le cas (voir Section 5.6).

Par ailleurs, nous avons effectué une étude empirique approfondie de la procédure $\mathcal{C}^{\mathrm{THO}}$, en proposant de comprendre l'influence de différents paramètres sur le risque Hellinger. La taille de l'échantillon d'entraînement (Section 5.5.3), le rôle du paramètre $\theta \in (0, 1/2)$ (Section 5.5.1) sur l'estimateur provenant du test (1.42), ainsi que l'influence du test (Section 5.5.2) lui-même (à savoir s'il existe une différence dans la qualité de la procédure lorsqu'on utilise le test (1.43) plutôt que le test (1.42)). Le meilleur choix pour θ semble se situer autour de 3/8 mais la différence est faible avec $\theta = 1/4$ ou 7/16. Le test de Baraud paraît être légèrement meilleur et moins couteux que celui de Birgé lorsqu'on regarde le nombre de tests calculés. Toutefois, il faut remarquer que

[58] Lorsque nous avons commencé ce projet, aucune étude pratique n'avait encore été menée. Depuis, Sart (2013, 2014) a implémenté des procédures similaires dans d'autres cadres statistiques.

[59] Dans notre cadre, nous partons du principe que le coût pour construire les estimateurs dans S est le même pour toutes les procédures HO de sorte que nous n'en tenons pas compte. Il faut remarquer que ceci n'est pas forcément le cas dans le cadre des moindres carrés où il est possible de construire les estimateurs et évaluer leur qualité de façon dynamique, voire la discussion dans Arlot & Lerasle (2014).

[60] dès que l'on a affaire à des points, c'est-à-dire si la discrétisation des modèles a déjà été effectuée.

ce dernier nécessite le calcul d'une intégrale supplémentaire que celui de Birgé et mène donc à une procédure plus lente.

Nous avons enfin comparé (Section 5.5.4) la procédure $\mathcal{C}^{\mathrm{THO}}$ par rapport à des procédures spécifiques à certains problèmes de sélection et à $\mathcal{C}^{\mathrm{HO}}$ (avec les deux fonctions de contraste classique), en calculant les risques Hellinger, $\mathbb{L}_1$ et $\mathbb{L}_2$ des estimateurs correspondants pour 18 densités et différentes collections d'estimateurs. On peut observer que le THO est meilleur que les autres candidats, quelle que soit la perte, et que la différence s'accentue lorsqu'on "mélange" différents types de noyau dans S.

Une procédure de VCVF à partir de tests robustes

Le Chapitre 6 vise à améliorer la procédure $\mathcal{C}^{\mathrm{THO}}$, qui souffre des mêmes maux que le HO classique, notamment d'une trop grande variabilité causée par l'unique découpage des données. En suivant l'itinéraire emprunté à la Section 1.2.3, on peut, dès lors que l'on se donne une collection $\mathcal{E}$ d'échantillons d'entraînement, définir toutes les procédures de VC via l'heuristique (1.28) et plus précisément la formule (1.24). En particulier nous pouvons définir les procédures alternatives à la VCVF et aux LPO et LOO de la Table 1.3 à partir de tests. Puisque l'exploitation exhaustive des données mènerait à un coût algorithmique trop important, nous nous focalisons sur le prolongement V-fold uniquement.

Comme pour le THO, on associe à la famille de méthodes d'estimation $(\mathcal{A}_m)_{m\in\mathcal{M}}$ une collection de poids $(\Delta_m)_{m\in\mathcal{M}}$ qui satisfait (1.22). En reprenant les notations de la Section 1.2.3, la seule fonction à définir s'avère être $\mathrm{crit}^{(j)}$ qui évalue la qualité des estimateurs au découpage $j \in [V]$. Pour ce faire, nous effectuons, pour tout $l, m \in \mathcal{M}$, $l \neq m$, les tests

$$\psi_{l,m}\left(\mathbf{X}^{(B_j)}\right) = \psi_{\widehat{s}_l^{(-j)},\widehat{s}_m^{(-j)}}\left(\mathbf{X}^{(B_j)}\right) \ ,$$

entre les densités $\widehat{s}_l^{(-j)}$ et $\widehat{s}_m^{(-j)}$, en posant $z = \Delta_l - \Delta_m$ dans (1.16). Nous utilisons ensuite naturellement le critère $\mathcal{C}^{\mathrm{THO}}$, qui s'écrit ici

$$\mathcal{D}_j(m) := \sup_{l\in\mathcal{R}_{m,j}} h^2\left(\widehat{s}_l^{(-j)}, \widehat{s}_m^{(-j)}\right) \ ,$$

où $\mathcal{R}_{m,j} = \{l \in \mathcal{M}, l \neq m \mid \psi_{l,m}(\mathbf{X}^{(B_j)}) = l\}$. La procédure T-$V$-fold (TVF) s'écrit alors pour tout $m \in \mathcal{M}$

$$\mathcal{C}_V^{\mathrm{TVF}}(m) := \frac{1}{V}\sum_{j=1}^{V} \mathcal{D}_j(m) \ . \tag{1.45}$$

Ce critère est fondamentalement différent du critère classique du fait qu'il n'estime pas le risque d'une méthode $\mathcal{A}_m$, mais fournit un indice de plausibilité qui repose sur des comparaisons entre les différents candidats dans la famille initiale. A notre connaissance il s'agit de la première procédure VF qui repose sur la distance de Hellinger.

Pour étudier théoriquement cette procédure nous sommes partis de la propriété remarquable suivante. Toute méthode d'estimation à noyau $\mathcal{A}_m$ vérifie

$$h^2(s, \widehat{s}_m) \leqslant \frac{1}{V} \sum_{j=1}^{V} h^2 \left(s, \widehat{s}_m^{(-j)} \right) \quad .$$

Par conséquent, pour tout $V \in \{2, \ldots, n\}$, le risque Hellinger d'une méthode d'estimation à noyau construit avec n observations est plus petit que celui construit avec $n(V-1)/V$ données. Ceci nous a poussé à d'abord travailler sur un découpage $j \in [V]$ fixé, pour lequel nous prouvons une inégalité de déviation pour $\mathcal{D}_j(m)$ (Proposition 6.2) pour en déduire, dans un deuxième temps, un résultat sur la moyenne des espérances. Nous ne parvenons pas, en revanche, à prouver un résultat de concentration pour $\mathcal{C}_V^{\mathrm{TVF}}(m)$ à cause de problèmes de dépendance entre les critères pour les différents découpages et par conséquent de "montrer" clairement le gain supposé de la moyenne sur V découpages[(61)].

Nous obtenons toutefois une inégalité oracle qui offre une garantie sur le risque Hellinger sans d'autre hypothèse (voir le Théorème 6.1). Cette borne montre qu'il y a un certain compromis à faire entre deux termes pour décider du choix du nombre optimal V. Nous illustrons ce résultat pour les estimateurs histogrammes (Section 6.2.5) pour lesquels l'estimateur choisi atteint la borne oracle (qui est la somme d'un terme de biais $h^2(s, \overline{s}_m)$ et d'un terme de variance) à constante près. Dans ce cas, nous considérons $\mathcal{M} = \{1, \ldots, n\}$ et $1 \leq 2\Gamma \leq \exp(3\Delta_m - 1)$ pour tout $m \in \mathcal{M}$, de sorte qu'une optimisation dans la borne mène à choisir V de l'ordre de

$$V \sim \sqrt{\frac{m^0 - 1}{2\Delta_{m^0}}} \quad \text{où} \quad m^0 = \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ h^2 \left(s, \overline{s}_m \right) + \frac{(m-1)}{2n} \right\} \quad .$$

La plupart du temps cette valeur optimale dépend de la valeur de m^0 qui est inconnue puisque le terme de biais est inconnu. Une densité qui est difficile à estimer par un histogramme avec peu de cases mène à une grande valeur de m^0 et donc une grande valeur optimale pour V, alors qu'une densité simple pour laquelle m^0 est plutôt petit ($m^0 = 1$ pour la loi uniforme sur $[0, 1]$) est estimée avec plus de précision pour une petite valeur de V.

Mais, à l'inverse du Chapitre 3, les risques des deux côtés de l'inégalité ne sont pas du même type : on compare le risque de $\widehat{s}_{\widehat{m}(\mathcal{C}_V^{\mathrm{TVF}})}$, donc construit avec tout l'échantillon, au meilleur estimateur construit avec $n(V-1)/V$ observations. De plus, notre analyse se fait au premier ordre et nous sommes incapables de dire théoriquement à quel point la qualité s'améliore ou se détériore quand V augmente et encore moins de comparer deux procédures $\mathcal{C}^{\mathrm{TVF}}$ pour deux V différents.

Du point de vue pratique, nous effectuons d'abord des simulations (Section 6.3.2) pour comprendre l'influence du paramètre θ quand on utilise le test (1.42). Les observations sont sensiblement les mêmes que celles présentées pour la procédure $\mathcal{C}^{\mathrm{THO}}$. Nos simulations, pour des familles d'histogrammes réguliers et d'estimateurs de Parzen-Rosenblatt, confirment l'interprétation du théorème principal quant au rôle de V (Section 6.3.3). Elles confirment que dans le cas où D_{m^0} est petit $V = 2$ peut constituer le meilleur choix (c'est le cas pour l'uniforme puisque l'histo-

[(61)] c'est le gros avantage d'être passé, dans le Chapitre 3, de l'étude d'un critère VF, souvent difficile à manier, à un critère pénalisé plus commun à traiter.

gramme avec une seule case appartient à la famille). Autrement la situation paraît s'améliorer quand V augmente, mais le gain devient infime lorsque V dépasse 10. Comme dans le chapitre précédent, nous comparons (Section 6.3.4) les procédures $\mathcal{C}^{\mathrm{TVF}}$, $\mathcal{C}^{\mathrm{LSVF}}$ et $\mathcal{C}^{\mathrm{KLVF}}$ pour les pertes Hellinger, $\mathbb{L}_1$ et $\mathbb{L}_2$. En général la procédure qui repose sur les tests est nettement plus lente mais présente de meilleurs résultats en termes de risque.

Nous proposons un algorithme (Section 6.4) qui accélère le calcul naïf de l'estimateur final qui a un coût prohibitif. En effet, alors que les procédures VF usuelles demandent, au pire, $V \times |\mathcal{M}|$ opérations, la procédure $\mathcal{C}^{\mathrm{TVF}}$ nécessiterait de tester pour chaque $j \in [V]$ tous les estimateurs $(\widehat{s}_m^{(-j)})_{m\in\mathcal{M}}$ entre eux, ce qui mènerait à un nombre de tests de l'ordre de $V \times |\mathcal{M}|^2$. Comme dans le Chapitre 5, il permet au mieux (modulo le temps de calcul d'une intégrale) de se rapprocher du coût des méthodes classiques, en ne réalisant "que" $V \times |\mathcal{M}|$ tests.

1.3.4 Perspectives de recherche

Dans cette thèse, nous avons présenté et étudié différentes procédures reposant sur deux pertes différentes, qui ne peuvent être comparées que sur des simulations. Nous trouvons encore une fois un gouffre entre la théorie et la pratique. Il ressort en effet des différentes études empiriques que le critère pénalisé $\mathcal{C}^{\mathrm{pen}}_{V,1}$, théoriquement bien compris et pour lequel nous savons que la situation s'améliore quand V augmente, est sensiblement moins bon (du moins tant qu'on ne sur-pénalise pas) que la procédure $\mathcal{C}^{\mathrm{TVF}}_V$. Alors que nous n'avons pas pu prouver une inégalité oracle optimale pour celle-ci, elle semble être la meilleure procédure à condition de ne pas être regardant sur le coût algorithmique ! Effectivement, bien que nous ayons réduit le coût calculatoire grâce à notre algorithme, le TVF reste une procédure bien plus gourmande que les VF classiques[(62)].

La Conclusion de ce manuscrit invite.. à ne pas conclure le travail. En effet, alors que la rédaction de ce manuscrit avançait, il est apparu qu'on pouvait combiner les différentes idées phares pour construire une nouvelle procédure qui garderait la qualité du TVF tout en étant plus rapide en temps de calcul. Celle-ci est une procédure du type $\mathcal{C}^{(iii)}$ dans la Table 1.2 et constitue une alternative à $\mathcal{C}^{\mathrm{TVF}}_V$ et à la procédure (du type $\mathcal{C}^{(iv)}$) de sélection d'estimateurs proposée dans Baraud (2011).

Nous développons notre idée en utilisant quelques points présentés précédemment dans la thèse. D'abord, l'existence d'un test idéal dans la procédure de sélection d'estimateurs par T-estimation. Ensuite l'heuristique de VCVF ainsi que celle d'Efron via le bootstrap avec poids V-fold. Enfin, afin de la mettre en pratique, l'algorithme pour le calcul d'un T-estimateur introduit au Chapitre 5. Plaçons nous dans le cadre Hellinger avec $\mathcal{S}$ l'ensemble des densités de probabilité par rapport à la mesure μ et supposons que nous disposons d'une collection de méthodes d'estimation à noyau $\{\mathcal{A}_m, m \in \mathcal{M}\}$ qui fournit la collection d'estimateurs linéaires $\{\widehat{s}_m, m \in \mathcal{M}\}$ qui satisfont à (1.44). Nous avons vu à la Section 1.2.2 que le test idéal pour choisir entre deux estimateurs $\widehat{s}_l$ et $\widehat{s}_m$, pouvait s'écrire $T_{\widehat{s}_l,\widehat{s}_m}(P)$. Le critère "idéal" donné par la T-estimation s'écrit alors comme une fonctionnelle $\mathcal{L}(P, P_n)$ qui dépend de P et P_n, plus précisément

$$\sup_{l\in\mathcal{R}_m} h^2(\widehat{s}_l, \widehat{s}_m) \quad \text{où} \quad \mathcal{R}_m = \{l \in \mathcal{M},\ l \neq m \mid T_{\widehat{s}_l,\widehat{s}_m}(P) \geqslant 0\} \ .$$

[(62)] Si on prend en compte cette contrainte, le conseil que l'on donnerait au praticien serait d'effectuer un 5-fold légèrement sur-pénalisé, voir Chapitre 3.

La procédure $\mathcal{C}_V^{\mathrm{TVF}}$, telle qu'elle est définie en (1.45), vise précisément à estimer ce critère $\mathcal{L}(P, P_n)$ en moyennant sur V découpages les critères $\mathcal{L}(P_n^{(j)}, P_n^{(-j)})$ selon l'heuristique de la validation croisée V-fold présentée en (1.28). Le coût algorithmique de cette moyenne est malheureusement élevé à cause de la moyenne sur les critères qui nous oblige à tester, pour chaque nouveau j, tous les estimateurs entre eux. Afin de diminuer ce coût, une idée naturelle est d'approcher le test plutôt que le critère, puisque ce dernier est "idéal" parce que le test l'est. Ceci permet de se ramener à une procédure du type T-estimation (avec un test différent de ceux de Baraud et Birgé) et par conséquent d'utiliser l'algorithme du Chapitre 5. Nous allons mettre à profit les heuristiques de rééchantillonnage et proposer deux manières d'estimer le test idéal $T_{\widehat{s}_l,\widehat{s}_m}(P)$. D'abord, en appliquant directement l'heuristique (1.28) sur le test, ensuite en utilisant l'heuristique d'Efron. La *procédure avec test rééchantillonné* se définit naturellement par

$$\mathcal{C}^{\mathrm{res.test}}(m) := \sup_{l \in \mathcal{R}_m} h^2\left(\widehat{s}_l, \widehat{s}_m\right) \tag{1.46}$$

où le test $\mathcal{T}_{l,m}(P_n)$ dans l'ensemble $\mathcal{R}_m = \{l \in \mathcal{M},\ l \neq m \mid \mathcal{T}_{l,m}(P_n) \geqslant 0\}$ est un test rééchantillonné selon l'une des deux alternatives pour estimer le test idéal. On peut noter que nous profitons à nouveau du fait que nous disposons de méthodes d'estimation et non pas d'estimateurs figés. Cette procédure peut se voir comme une procédure de T-estimation avec "test V-fold", ou comme une nouvelle procédure de sélection d'estimateurs avec test rééchantillonné via des poids VF.

Dans le chapitre de conclusion, nous proposons de l'étudier avec les deux tests et de la comparer empiriquement, grâce à l'algorithme de la Section 5.3.1, aux procédures étudiées dans cette thèse. Du point de vue pratique, les deux tests rééchantillonnés présentent des performances similaires en termes de risque. Ils donnent des résultats légèrement moins bons que ceux de la procédure $\mathcal{C}^{\mathrm{TVF}}$ mais restent sensiblement plus performants que les procédures VF classiques. De manière générale, la procédure $\mathcal{C}^{\mathrm{res.test}}$ diminue drastiquement le coût algorithmique du TVF (Section 7.1.4) et présente un coût similaire (en nombre de tests calculés) à celui du THO.

Des résultats théoriques devraient être prouvés pour mieux comprendre les simulations et rendre encore plus attrayante cette utilisation du rééchantillonnage sur le test. Il nous semble que les outils théoriques développés par Baraud (2011); Baraud *et al.* (2014) pourraient mener à un résultat similaire à celui obtenu pour la procédure TVF. Mais ceci pourrait ne pas suffire pour obtenir une compréhension fine du rôle de V dans la procédure.

Enfin, nous présentons dans la Section 7.2 quelques défis et problèmes ouverts pour toutes les problématiques évoquées dans cette thèse. Ceux-ci constituent autant de travaux de recherche possibles et confirment, s'il le fallait, qu'il y a toujours plus à comprendre que ce qui a déjà été compris.

Part I

Optimal selection by penalization

Chapter 2

Optimal kernel selection in density estimation

Abstract. We provide new general kernel selection rules thanks to penalized least-squares criteria. We derive optimal oracle inequalities using adequate concentration tools. We also investigate the problem of minimal penalty as described in Birgé & Massart (2007). This simple approach provides sharp results in a very general setting.

NOTA: Ce chapitre est une version légèrement modifiée d'un travail en collaboration avec Matthieu Lerasle[(1)] et Patricia Reynaud-Bouret[(1)], soumis au "Proceedings of the High Dimensional Probability VII meeting" en février 2015.

Contents

[(1)]Université Nice Sophia Antipolis

2.1 Introduction

Concentration inequalities are central tools in adaptive nonparametric statistics. They allow to build sharp penalized criteria for model selection (Massart, 2007), to select bandwidths and even approximation kernels in high dimension (Goldenshluger & Lepski, 2011), to aggregate estimates (Rigollet & Tsybakov, 2007) and to properly calibrate threshold estimators (Donoho *et al.*, 1996).

In the present work, we use similar tools to select a general kernel estimate in the least-squares density estimation framework. Similar problems for $\mathbb{L}_1$-loss have been considered by Devroye & Lugosi (2001), leading to estimators that are difficult to compute in practice. Here we propose least-squares penalized criteria leading to easily computable estimates. Sharp concentration inequalities for U-statistics (Giné *et al.*, 2000; Adamczak, 2006; Houdré & Reynaud-Bouret, 2003) help us to control the variance term of the kernel estimates, a term which has already been very carefully controlled in an asymptotic way for instance in Mason & Swanepoel (2011) or Deheuvels & Ouadah (2013). We derive from these bounds (see Proposition 2.6) a penalization method to select a kernel which satisfies an asymptotically optimal oracle inequality, i.e. with leading constant asymptotically equal to 1.

In the spirit of Giné & Nickl (2009), we use an extended definition of kernels that allows to deal simultaneously with classical collections of linear estimators as projection estimators, weighted projection estimators, or classical kernel estimators. This method can be used for example to select an optimal model in model selection (in accordance with Massart (2007)) or to select an optimal bandwidth together with an optimal approximation kernel among a finite collection. In this sense, our method in particular deals with the same problem as that of Goldenshluger & Lepski (2011) and we prove in this framework that a leading constant 1 in the oracle inequality is indeed possible.

In addition, there is a sharp phase transition in the dimension of the selected models allowing an estimate of the optimal penalty in their case (which is known up to a multiplicative constant). Indeed, starting from the idea that in many models the optimal penalty is twice the minimal one (this is the *slope heuristic*), Arlot & Massart (2009) propose to detect the minimal penalty by the phase transition and to apply the rule "$\times 2$" (this is the *slope algorithm*). They prove that this algorithm works at least in some regression settings.

In the present work, we also show that minimal penalties exist in the density estimation setting. In particular, we exhibit a sharp "phase transition" of the behavior of the selected estimator around this minimal penalty.

The analysis of this last result is not standard here. First, the "slope heuristic" of Birgé & Massart (2007) only holds in particular cases as the selection of projection estimators, see also Lerasle (2012). As in the selection of a linear estimator in a regression setting (Arlot & Bach, 2009), the heuristic can sometimes be corrected, for example, for the selection of a bandwidth when the approximation kernel is fixed. In general though there is no simple relation between the minimal penalty and the optimal one and the slope algorithm of Arlot & Massart (2009) shall therefore only be carefully used for kernel selection.

Surprisingly our work reveals that the minimal penalty can be negative; in this case minimizing an unpenalized criterion leads to oracle estimators. Up to our knowledge, such phenomenon has only been noted once in a very particular classification setting (Fromont & Tuleau, 2006). We illustrate all these different behaviors in a simulation study.

After giving the main notation, providing some examples and defining the framework, we explain our goal, detail what *oracle inequality* means and give the precise exponential inequalities that we need in Section 2.2. Then we provide optimal penalties in Section 2.3 and study the problem of minimal penalties in Section 2.4, all those results being illustrated in three main examples : projection kernels, approximation kernels and weighted projection kernels. In Section 2.5, some simulations are performed in the approximation kernel case. The main proofs are in Section 2.6 and the technical results are discussed in the appendix.

2.2 Kernel selection for least-squares density estimation

2.2.1 Settings

Let $X, Y, X_1, \ldots, X_n$ denote i.i.d. random variables taking values in the measured space $(\Xi, \mathcal{Z}, \mu)$, with common distribution P. Assume P has density s with respect to μ and s is uniformly bounded. Hence, s belongs to $\mathbb{L}_2$, where, for any $p \geq 1$,

$$\mathbb{L}_p := \left\{ t : \Xi \to \mathbb{R}, \text{ s.t. } \|t\|_p^p := \int |t|^p \, d\mu < \infty \right\} \ .$$

Moreover, $\|\cdot\| = \|\cdot\|_2$ and $\langle \cdot, \cdot \rangle$ denote respectively the $\mathbb{L}_2$-norm and the associated inner product and $\|\cdot\|_\infty$ is the infinite norm. We systematically use $x \vee y$ and $x \wedge y$ for $\max(x, y)$ and $\min(x, y)$ respectively, and denote $|A|$ the cardinality of the set A. Recall that $x_+ = x \vee 0$ and, for any $y \in \mathbb{R}^+$, $\lfloor y \rfloor = \sup\{n \in \mathbb{N} \text{ s.t. } n \leq y\}$.

Let $\{\mathcal{K}_m\}_{m \in \mathcal{M}}$ denote a collection of symmetric functions $\mathcal{K}_m : \Xi^2 \to \mathbb{R}$ such that

$$\sup_{x \in \Xi} \int_\Xi \mathcal{K}_m(x, y)^2 d\mu(y) \ \vee \sup_{(x,y) \in \Xi^2} |\mathcal{K}_m(x, y)| < \infty \ .$$

A function $\mathcal{K}_m$ satisfying these assumptions is called *kernel*, in the sequel. Any kernel $\mathcal{K}_m$ is associated with a *linear estimator* $\widehat{s}_m$ of s defined for any $x \in \Xi$ by

$$\widehat{s}_m(x) := \frac{1}{n} \sum_{i=1}^n \mathcal{K}_m(X_i, x) \ .$$

Our aim is to select a "good" estimator $\widehat{s}_{\widehat{m}}$ in the family $\{\widehat{s}_m, m \in \mathcal{M}\}$. Our results are expressed in terms of a constant $\Gamma \geq 1$ such that, for all $m \in \mathcal{M}$,

$$\sup_{x \in \Xi} \int_\Xi \mathcal{K}_m(x, y)^2 d\mu(y) \ \vee \sup_{(x,y) \in \Xi^2} |\mathcal{K}_m(x, y)| \leq \Gamma n \ . \tag{2.1}$$

This condition plays the same role as the milder condition $\int |\mathcal{K}_m(x, y)| s(y) d\mu(y) < \infty$ used in Devroye & Lugosi (2001) when working with $\mathbb{L}_1$-loss. Before describing the method, let us give three examples of those estimators, repeatedly used for density estimation, and see how they can naturally be associated to some kernels (see Section 2.7 in the appendix for the evaluation of the corresponding Γ's).

Example 1: Projection estimators. Projection estimators are among the most classical density estimators. Given a linear subspace $S \subset \mathbb{L}_2$, the projection estimator on S is defined by

$$\widehat{s}_S = \operatorname*{argmin}_{t \in S} \left\{ \|t\|^2 - \frac{2}{n} \sum_{i=1}^{n} t(X_i) \right\} \ .$$

Let $\mathbb{S}$ be a family of linear subspaces S of $\mathbb{L}_2$. For any $S \in \mathbb{S}$, let $(\varphi_\ell)_{\ell \in \mathcal{I}_S}$ denote an orthonormal basis of S. The projection estimator $\widehat{s}_S$ can be computed and is equal to

$$\widehat{s}_S = \sum_{\ell \in \mathcal{I}_S} \left(\frac{1}{n} \sum_{i=1}^{n} \varphi_\ell(X_i) \right) \varphi_\ell \ .$$

It is therefore easy to see that it is the estimator associated to the *projection kernel* $\mathcal{K}_S$ defined for any x and y in Ξ by

$$\mathcal{K}_S(x,y) := \sum_{\ell \in \mathcal{I}_S} \varphi_\ell(x) \varphi_\ell(y) \ .$$

Notice that $\mathcal{K}_S$ actually depends on the basis $(\varphi_\ell)_{\ell \in \mathcal{I}_S}$ even if $\widehat{s}_S$ does not. In the sequel, we always assume that some orthonormal basis $(\varphi_\ell)_{\ell \in \mathcal{I}_S}$ is given with S. Given a finite collection $\mathbb{S}$ of linear subspaces of $\mathbb{L}_2$, one can choose the following constant Γ (see (2.1)) for the collection $(\mathcal{K}_S)_{S \in \mathbb{S}}$

$$\Gamma = 1 \vee \frac{1}{n} \sup_{S \in \mathbb{S}} \sup_{f \in S, \|f\|=1} \|f\|_\infty^2 \ . \tag{2.2}$$

Example 2: Approximation kernel estimators. Another important family of density estimators is derived from approximation kernels. Given a bounded symmetric integrable function $k : \mathbb{R} \to \mathbb{R}$ such that $\int_{\mathbb{R}} k(u)du = 1$, $k(0) > 0$ and a bandwidth $h > 0$, the approximation kernel estimator is defined for any $x \in \mathbb{R}$ by

$$\widehat{s}_{(k,h)}(x) = \frac{1}{nh} \sum_{i=1}^{n} k\left(\frac{x - X_i}{h} \right) \ .$$

It can also naturally be seen as a linear estimator, associated to the *approximation kernel* $\mathcal{K}_{(k,h)}$ defined for any x and y in $\mathbb{R}$ by

$$\mathcal{K}_{(k,h)}(x,y) := \frac{1}{h} k\left(\frac{x-y}{h} \right) \ .$$

Given a finite collection of couples $m = (k,h) \in \mathcal{M} = \mathcal{H}$, one can choose (see (2.1)) $\Gamma = 1$ if,

$$h \geq \frac{\|k\|_\infty \|k\|_1}{n} \quad \text{for any } (k,h) \in \mathcal{H} \ . \tag{2.3}$$

Example 3: Weighted projection estimators. Let $(\varphi_i)_{i=1,\dots,p}$ denote an orthonormal system in $\mathbb{L}_2$ and let $w = (w_i)_{i=1,\dots,p}$ denote real numbers in $[0,1]$. The associated weighted kernel

projection estimator of s is defined by

$$\widehat{s}_w = \sum_{i=1}^{p} w_i \left(\frac{1}{n} \sum_{j=1}^{n} \varphi_i(X_j) \right) \varphi_i \ .$$

These estimators are used to prove very sharp adaptive results. In particular, Pinsker's estimators are weighted kernel projection estimators (see for example Rigollet (2006a)). When $w \in \{0,1\}^p$, we recover a classical projection estimator. A weighted projection estimator is associated to the *weighted projection kernel* defined for any x and y in Ξ by

$$\mathcal{K}_w(x,y) := \sum_{i=1}^{p} w_i \varphi_i(x) \varphi_i(y) \ .$$

Given any finite collection $\mathcal{W}$ of weights, one can choose (see (2.1))

$$\Gamma = 1 \vee \left(\frac{1}{n} \sup_{x \in \Xi} \sum_{i=1}^{p} \varphi_i(x)^2 \right) \ . \tag{2.4}$$

2.2.2 Oracle inequalities and penalized criterion

The goal is to estimate s in the best possible way using a finite collection of linear estimators $(\widehat{s}_m)_{m \in \mathcal{M}}$. In other words, the purpose is to select among $(\widehat{s}_m)_{m \in \mathcal{M}}$ an estimate $\widehat{s}_{\widehat{m}}$ from the data such that $\|\widehat{s}_{\widehat{m}} - s\|^2$ is as close as possible to $\inf_{m \in \mathcal{M}} \|\widehat{s}_m - s\|^2$. More precisely our aim is to select $\widehat{m}$ such that, with large probability,

$$\|\widehat{s}_{\widehat{m}} - s\|^2 \leq C_n \inf_{m \in \mathcal{M}} \|\widehat{s}_m - s\|^2 + R_n \ , \tag{2.5}$$

where $C_n \geqslant 1$ and $R_n > 0$. In this case, $\widehat{s}_{\widehat{m}}$ is said to satisfy an *oracle inequality*, as long as R_n is small compared to $\inf_{m \in \mathcal{M}} \|\widehat{s}_m - s\|^2$ and C_n does not explode, that is C_n is smaller than a power of $\log n$. This means that the selected estimate does as well as the best estimate in the family up to some multiplicative constant. The best case one can expect is to get C_n close to 1. This is why, when $C_n \to_{n \to \infty} 1$, the corresponding oracle inequality is called *asymptotically optimal*. To do so, we study minimizers of *penalized least-squares criteria*. Note that in our three examples choosing $\widehat{m}$ amounts to choose the smoothing parameter, that is respectively $\widehat{S}$, $(\widehat{k}, \widehat{h})$ or $\widehat{w}$.

Let P_n denote the empirical measure, that is, for any real valued function t,

$$P_n(t) := \frac{1}{n} \sum_{i=1}^{n} t(X_i) \ .$$

For any $t \in \mathbb{L}_2$, let also

$$P(t) := \mathbb{E}\left[t(X)\right] = \int_{\Xi} t(x) s(x) d\mu(x) \ .$$

The *least-squares contrast* is defined, for any $t \in \mathbb{L}_2$, by

$$\gamma(t) := \|t\|^2 - 2t \ .$$

Then for any given function $\mathrm{pen} : \mathcal{M} \to \mathbb{R}$, the *least-squares penalized criterion* is defined by

$$\mathcal{C}_{\mathrm{pen}}(m) := P_n\gamma(\widehat{s}_m) + \mathrm{pen}(m) \ . \tag{2.6}$$

Finally the selected $\widehat{m} \in \mathcal{M}$ is given by any minimizer of $\mathcal{C}_{\mathrm{pen}}(m)$, that is,

$$\widehat{m} \in \operatorname*{argmin}_{m\in\mathcal{M}} \{\mathcal{C}_{\mathrm{pen}}(m)\} \ . \tag{2.7}$$

As $P\gamma(t) = \|t-s\|^2 - \|s\|^2$, it is equivalent to minimize $\|\widehat{s}_m - s\|^2$ or $P\gamma(\widehat{s}_m)$. As our goal is to select $\widehat{s}_{\widehat{m}}$ satisfying an oracle inequality, an ideal penalty $\mathrm{pen}_{\mathrm{id}}$ should satisfy $\mathcal{C}_{\mathrm{pen}_{\mathrm{id}}}(m) = P\gamma(\widehat{s}_m)$, i.e. criterion (2.6) with

$$\mathrm{pen}_{\mathrm{id}}(m) := (P - P_n)\gamma(\widehat{s}_m) = 2(P_n - P)(\widehat{s}_m) \ .$$

To identify the main quantities of interest, let us introduce some notation and develop $\mathrm{pen}_{\mathrm{id}}(m)$. For all $m \in \mathcal{M}$, let

$$s_m(x) := \int_\Xi \mathcal{K}_m(y,x)s(y)d\mu(y) = \mathbb{E}\left[\mathcal{K}_m(X,x)\right], \qquad \forall x \in \Xi \ ,$$

and

$$U_m := \sum_{i\neq j=1}^{n} \left(\mathcal{K}_m(X_i,X_j) - s_m(X_i) - s_m(X_j) + \mathbb{E}\left[\mathcal{K}_m(X,Y)\right]\right) \ .$$

Because those quantities will be fundamental in the sequel, let us also define $\Theta_m(x) = A_m(x,x)$ where for $(x,y) \in \Xi^2$

$$A_m(x,y) := \int_\Xi \mathcal{K}_m(x,z)\mathcal{K}_m(z,y)d\mu(z) \ . \tag{2.8}$$

Denoting $\chi_m(x) = \mathcal{K}_m(x,x)$, the ideal penalty is then equal to

$$\begin{aligned}\mathrm{pen}_{\mathrm{id}}(m) &= 2(P_n - P)(\widehat{s}_m - s_m) + 2(P_n - P)s_m \\ &= 2\left(\frac{P\chi_m - Ps_m}{n} + \frac{(P_n - P)\chi_m}{n} + \frac{U_m}{n^2} + \left(1 - \frac{2}{n}\right)(P_n - P)s_m\right) \ . \end{aligned} \tag{2.9}$$

The main point is that by using concentration inequalities (Bernstein and concentration of the totally degenerate U-statistic of order two, U_m) detailed in Section 2.2.3, we obtain:

$$\mathrm{pen}_{\mathrm{id}}(m) \simeq 2\left(\frac{P\chi_m - Ps_m}{n}\right) \ .$$

The term Ps_m/n depends on s which is unknown. Fortunately, it can be easily controlled as detailed in the sequel. Therefore one can hope that the choice

$$\mathrm{pen}(m) = 2\frac{P\chi_m}{n}$$

is convenient. In general, this choice still depends on the unknown density s but it can be easily estimated in a data-driven way by

$$\mathrm{pen}(m) = 2\frac{P_n\chi_m}{n} \ .$$

The goal of Section 2.3 is to prove this heuristic and to show that $2P\chi_m/n$ and $2P_n\chi_m/n$ are optimal choices for the penalty, that is, they lead to an asymptotically optimal oracle inequality.

2.2.3 Concentration tools

To derive sharp oracle inequalities, we only need two fundamental concentration tools, namely weak Bernstein's inequality and the concentration bounds for degenerate U-statistics of order two. We cite them here under their most suitable form for our purpose.

Weak Bernstein's inequality.

Proposition 2.1. *For any bounded real valued function f and any $X_1, \dots, X_n$ i.i.d. with distribution P, for any $u > 0$,*

$$\mathbb{P}\left((P_n - P)f \geqslant \sqrt{\frac{2P\left(f^2\right)u}{n}} + \frac{\|f\|_\infty u}{3n} \right) \leqslant \exp(-u) \ .$$

The proof is straightforward and can be derived from either Bennett's or Bernstein's inequality (Boucheron *et al.*, 2013).

Concentration of degenerate U-statistics of order 2.

Proposition 2.2. *Let $X, X_1, \dots X_n$ be i.i.d. random variables defined on a Polish space Ξ equipped with its Borelian σ-algebra and let $(f_{i,j})_{1\leqslant i\neq j\leq n}$ denote bounded real valued symmetric measurable functions defined on Ξ^2, such that for any $i \neq j$, $f_{i,j} = f_{j,i}$ and*

$$\forall\, i, j \text{ s.t. } 1 \leqslant i \neq j \leq n, \qquad \mathbb{E}\left[f_{i,j}(x, X)\right] = 0 \qquad \text{for a.e. } x \text{ in } \Xi \ . \tag{2.10}$$

Let U be the following totally degenerate U-statistic of order 2,

$$U = \sum_{1\leqslant i\neq j\leq n} f_{i,j}(X_i, X_j) \ .$$

Let A be an upper bound of $|f_{i,j}(x,y)|$ for any i,j,x,y and

$$B^2 = \max\left(\sup_{i,x\in\Xi} \sum_{j=1}^{i} \mathbb{E}\left[f_{i,j}(x,X_j)^2 \right], \sup_{j,t\in\Xi} \sum_{i=j+1}^{n} \mathbb{E}\left[f_{i,j}(X_i,t)^2 \right] \right)$$
$$C^2 = \sum_{1\le i\neq j\le n} \mathbb{E}\left[f_{i,j}(X_i,X_j)^2 \right]$$
$$D = \sup_{(a,b)\in\mathcal{A}} \mathbb{E}\left[\sum_{1\le i<j\le n} f_{i,j}(X_i,X_j) a_i(X_i) b_j(X_j) \right] ,$$

where $\mathcal{A} = \left\{ (a,b),\ \text{s.t. } \mathbb{E}\left[\sum_{i=1}^{n-1} a_i(X_i)^2\right] \le 1,\ \mathbb{E}\left[\sum_{j=2}^{n} b_j(X_j)^2\right] \le 1 \right\}$. *Then for any* $u > 0$ *and* $\varepsilon > 0$,

$$\mathbb{P}\left(U \ge 2\sqrt{2}(1+\varepsilon)^{3/2} C\sqrt{u} + 8\left(1+\frac{1}{\varepsilon}\right)^{1/2} Du + 24\left(1+\varepsilon+\frac{1}{\varepsilon}\right)^{2} (Bu^{3/2} + Au^2) \right) \leqslant 2.7e^{-u} .$$

This inequality dates back from Giné, Latala and Zinn (Giné *et al.*, 2000). Exact constants have been provided in Houdré & Reynaud-Bouret (2003) but the result therein has been stated only for real variables. This result has been further generalized by Adamczak to U-statistics of any order (Adamczak, 2006), though the constants are not explicit. We provide a more ready-to-use result for arbitrary space, which basically follows the proof of Houdré & Reynaud-Bouret (2003) with updated constants, thanks to the improvement of Talagrand's inequality for empirical processes (Talagrand, 1996) by Rio (2012). The complete proof can be found in Section 2.9.

2.3 Optimal penalties for kernel selection

The main objective of this section is to show that $2P\chi_m/n$ is a theoretical optimal penalty for kernel selection, which means that if we choose $\mathrm{pen}(m)$ close to $2P\chi_m/n$, the selected kernel $\mathcal{K}_{\widehat{m}}$ satisfies an asymptotically optimal oracle inequality.

2.3.1 Main assumptions

To express our results in a simple form, a positive constant Υ is assumed to control all the following quantities.

$$(\Gamma(1+\|s\|_\infty)) \vee \sup_{m\in\mathcal{M}} \|s_m\|^2 \le \Upsilon , \tag{2.11}$$
$$\forall m\in\mathcal{M}, \quad P\left(\chi_m^2\right) \le \Upsilon n P\Theta_m , \tag{2.12}$$
$$\forall (m,m')\in\mathcal{M}^2, \quad \|s_m - s_{m'}\|_\infty \le \Upsilon \vee \sqrt{\Upsilon n}\, \|s_m - s_{m'}\| , \tag{2.13}$$
$$\forall m\in\mathcal{M}, \quad \mathbb{E}\left[A_m(X,Y)^2\right] \le \Upsilon P\Theta_m , \tag{2.14}$$
$$\forall m\in\mathcal{M}, \quad \sup_{x\in\Xi} \mathbb{E}\left[A_m(X,x)^2\right] \le \Upsilon n , \tag{2.15}$$

$$\forall m \in \mathcal{M}, \qquad v_m^2 := \sup_{t \in \mathbb{B}_m} Pt^2 \quad \leq \quad \Upsilon \vee \sqrt{\Upsilon P \Theta_m} \ , \tag{2.16}$$

where $\mathbb{B}_m$ is the set of functions t that can be written $t(x) = \int a(z)\mathcal{K}_m(z,x)d\mu(z)$ for some $a \in \mathbb{L}_2$ with $\|a\| \leq 1$.

These assumptions may seem very intricate. They are actually fulfilled by our three main examples under very mild conditions (see Section 2.3.3).

2.3.2 The optimal penalty theorem

In the following, $\square$ denotes a positive absolute constant whose value may change from line to line and if there are indices such as $\square_\theta$, it means that this is a positive function of θ and only θ whose value may change from line to line.

Theorem 2.1. *If Assumptions* (2.11), (2.12), (2.13), (2.14) (2.15), (2.16) *hold, then, for any $x \geq 1$, with probability larger than $1 - \square|\mathcal{M}|^2 e^{-x}$, for any $\theta \in (0,1)$, any minimizer $\widehat{m}$ of the penalized criterion* (2.6) *satisfies the following inequality*

$$\forall m \in \mathcal{M}, \qquad (1-4\theta)\left\|s - \widehat{s}_{\widehat{m}}\right\|^2 \leq (1+4\theta)\left\|s - \widehat{s}_m\right\|^2 + \left(\mathrm{pen}(m) - 2\frac{P\chi_m}{n}\right) - \left(\mathrm{pen}\left(\widehat{m}\right) - 2\frac{P\chi_{\widehat{m}}}{n}\right) + \square\frac{\Upsilon x^2}{\theta n} \ . \tag{2.17}$$

Assume moreover that there exists $C > 0$, $\delta' \geqslant \delta > 0$ and $r \geq 0$ such that for any $x \geq 1$, with probability larger than $1 - Ce^{-x}$

$$\forall m \in \mathcal{M}, \quad (\delta - 1)\frac{P\Theta_m}{n} - \square r\frac{\Upsilon x^2}{n} \leq \mathrm{pen}(m) - \frac{2P\chi_m}{n} \leq (\delta' - 1)\frac{P\Theta_m}{n} + \square r\frac{\Upsilon x^2}{n} \ . \tag{2.18}$$

Then for all $\theta \in (0,1)$ and all $x \geq 1$, with probability at least $1 - \square(C + |\mathcal{M}|^2)e^{-x}$,

$$\frac{(\delta \wedge 1) - 5\theta}{(\delta' \vee 1) + (4+\delta')\theta}\left\|s - \widehat{s}_{\widehat{m}}\right\|^2 \leq \inf_{m \in \mathcal{M}}\left\|s - \widehat{s}_m\right\|^2 + \square\left(r + \frac{1}{\theta^3}\right)\frac{\Upsilon x^2}{n} \ .$$

Let us make some remarks.

- First, this is an oracle inequality (see (2.5)) with leading constant

$$C_n = \frac{(\delta' \vee 1) + (4+\delta')\theta}{(\delta \wedge 1) - 5\theta}$$

and remainder term

$$R_n = \square C_n\left(r + \theta^{-3}\right)\frac{\Upsilon x^2}{n} \ ,$$

as long as

  - θ is small enough for C_n to be positive,
  - x is large enough for the probability to be large and
  - n is large enough for R_n to be negligible.

Typically, $r, \delta, \delta', \theta$ and Υ are constant w.r.t. n and x has to be of the order of $\log(|\mathcal{M}| \vee n)$ for the remainder to be negligible. In particular, $\mathcal{M}$ may grow with n as long as $\log(|\mathcal{M}| \vee n)^2$ remains negligible with respect to n and Υ does not depend on n.

- If $\mathrm{pen}(m) = 2P\chi_m/n$, that is if $\delta = \delta' = 1$ and $r = C = 0$ in (2.18), the estimator $\widehat{s}_{\widehat{m}}$ satisfies an asymptotically optimal oracle inequality i.e. $C_n \to_{n\to\infty} 1$ since θ can be chosen as close to 0 as desired. Take for instance, $\theta = (\log n)^{-1}$.

- In general $P\chi_m$ depends on the unknown s and this last penalty cannot be used in practice. Fortunately, its empirical counterpart $\mathrm{pen}(m) = 2P_n\chi_m/n$ satisfies (2.18) with $\delta = 1 - \theta$, $\delta' = 1 + \theta$, $r = 1/\theta$ and $C = 2|\mathcal{M}|$ for any $\theta \in (0, 1)$ and in particular $\theta = (\log n)^{-1}$ (see (2.34) in Proposition 2.7). Hence, the estimator $\widehat{s}_{\widehat{m}}$ selected with this choice of penalty also satisfies an asymptotically optimal oracle inequality, by the same argument.

- Finally, we only get an oracle inequality when $\delta > 0$, that is when $\mathrm{pen}(m)$ is larger than $(2P\chi_m - P\Theta_m)/n$ up to some residual term. We discuss the necessity of this condition in Section 2.4.

2.3.3 Main examples

This section shows that Theorem 2.1 can be applied in the examples. In addition, it provides the computation of $2P\chi_m/n$ in some specific cases of special interest.

Example 1 (continued).

Proposition 2.3. *Let $\{\mathcal{K}_S, S \in \mathbb{S}\}$ be a collection of projection kernels. Assumptions* (2.11)*,* (2.12)*,* (2.14)*,* (2.15) *and* (2.16) *hold for any $\Upsilon \geq \Gamma(1 + \|s\|_\infty)$, where Γ is given by* (2.2)*. In addition, Assumption* (2.13) *is satisfied under either of the following classical assumptions (see (Massart, 2007, Chapter 7)):*

$$\forall S, S' \in \mathbb{S}, \qquad \textit{either } S \subset S' \textit{ or } S' \subset S \ , \tag{2.19}$$

or

$$\forall S \in \mathbb{S}, \qquad \|s_S\|_\infty \leqslant \frac{\Upsilon}{2} \ . \tag{2.20}$$

These particular kernels satisfy for all $(x, y) \in \Xi^2$

$$A_S(x,y) = \int_\Xi \mathcal{K}_S(x,z)\mathcal{K}_S(y,z)d\mu(z) = \sum_{(i,j)\in\mathcal{I}_S^2} \varphi_i(x)\varphi_j(y)\int_\Xi \varphi_i(z)\varphi_j(z)d\mu(z) = \mathcal{K}_S(x,y) \ .$$

In particular, $\Theta_S = \chi_S = \sum_{i\in\mathcal{I}_S} \varphi_i^2$ and $2P\chi_S - P\Theta_S = P\chi_S$.

Moreover, it appears that the function Θ_S is constant in some examples of linear spaces S of interest (see Lerasle (2012) for more details). Let us mention one particular case studied further

on in the sequel. Suppose $\mathbb{S}$ is a collection of regular histogram spaces S on Ξ, that is, any $S \in \mathbb{S}$ is a space of piecewise constant functions on a partition $\mathcal{I}_S$ of Ξ such that $\mu(i) = 1/D_S$ for any i in $\mathcal{I}_S$. Assumption (2.20) is satisfied for this collection as soon as $\Upsilon \geq 2\|s\|_\infty$. The family $(\varphi_i)_{i\in\mathcal{I}_S}$, where $\varphi_i = \sqrt{D_S}\mathbf{1}_i$ is an orthonormal basis of S and

$$\chi_S = \sum_{i\in\mathcal{I}_S} \varphi_i^2 = D_S \ .$$

Hence, $P\chi_S = D_S$ and $2D_S/n$ can actually be used as a penalty to ensure that the selected estimator satisfies an asymptotically optimal oracle inequality. Moreover, in this example it is actually necessary to choose a penalty larger than D_S/n to get an oracle inequality (see Lerasle (2012) or Section 2.4 for more details).

Example 2 (continued).

Proposition 2.4. *Let* $\left\{\mathcal{K}_{(k,h)}, (k,h) \in \mathcal{H}\right\}$ *be a collection of approximation kernels. Assumptions* (2.11)*,* (2.12)*,* (2.13)*,* (2.14)*,* (2.15) *and* (2.16) *hold with* $\Gamma = 1$*, for any*

$$\Upsilon \geq \max_k \left\{ \frac{k(0)}{\|k\|^2} \vee \left(1 + 2\|s\|_\infty \|k\|_1^2\right) \right\} \ ,$$

as soon as (2.3) *is satisfied.*

These approximation kernels satisfy, for all $x \in \mathbb{R}$,

$$\chi_{(k,h)}(x) = \mathcal{K}_{(k,h)}(x,x) = \frac{k(0)}{h} \ ,$$
$$\Theta_{(k,h)}(x) = A_{(k,h)}(x,x) = \frac{1}{h^2}\int_{\mathbb{R}} k\left(\frac{x-y}{h}\right)^2 dy = \frac{\|k\|^2}{h} \ .$$

Therefore, the optimal penalty $2P\chi_{(k,h)}/n = 2k(0)/(nh)$ can be computed in practice and yields an asymptotically optimal selection criterion. Surprisingly, the lower bound $2P\chi_{(k,h)}/n - P\Theta_{(k,h)}/n = (2k(0) - \|k\|^2)/(nh)$ can be negative if $\|k\|^2 > 2k(0)$. In this case, a minimizer of (2.6) satisfies an oracle inequality, even if this criterion is not penalized. This remarkable fact is illustrated in the simulation study in Section 2.5.

Example 3 (continued).

Proposition 2.5. *Let* $\{\mathcal{K}_w, w \in \mathcal{W}\}$ *be a collection of weighted projection kernels. Assumption* (2.11) *is valid for* $\Upsilon \geq \Gamma(1 + \|s\|_\infty)$*, where* Γ *is given by* (2.4)*. Moreover* (2.11) *and* (2.1) *imply* (2.12)*,* (2.13)*,* (2.14)*,* (2.15) *and* (2.16)*.*

For these weighted projection kernels, for all $x \in \Xi$

$$\chi_w(x) = \sum_{i=1}^p w_i\varphi_i(x)^2, \qquad \text{hence} \qquad P\chi_w = \sum_{i=1}^p w_i P\varphi_i^2 \ ,$$

$$\Theta_w(x) = \sum_{i,j=1}^{p} w_i w_j \varphi_i \varphi_j \int_\Xi \varphi_i(x)\varphi_j(x) d\mu(x) = \sum_{i=1}^{p} w_i^2 \varphi_i(x)^2 \leq \chi_w(x) \ .$$

In this case, the optimal penalty $2P\chi_w/n$ has to be estimated in general. However, in the following example it can still be directly computed.

Let $\Xi = [0,1]$, let μ be the Lebesgue measure. Let $\varphi_0 \equiv 1$ and, for any $j \geq 1$,

$$\varphi_{2j-1}(x) = \sqrt{2}\cos(2\pi jx), \qquad \varphi_{2j}(x) = \sqrt{2}\sin(2\pi jx) \ .$$

Consider some odd p and a family of weights $\mathcal{W} = \{w_i, i = 0, \ldots, p\}$ such that, for any $w \in \mathcal{W}$ and any $i = 1, \ldots, p/2$, $w_{2i-1} = w_{2i} = \tau_i$. In this case, the values of the functions of interest do not depend on x

$$\chi_w = w_0 + \sum_{j=1}^{p/2} \tau_j, \qquad \Theta_w = w_0^2 + \sum_{j=1}^{p/2} \tau_j^2 \ .$$

In particular, this family includes Pinsker's and Tikhonov's weights.

2.4 Minimal penalties for kernel selection

The purpose of this section is to see whether the lower bound $\mathrm{pen}_{\min}(m) := (2P\chi_m - P\Theta_m)/n$ is sharp in Theorem 2.1. To do so we first need the following result which links $\|s - \widehat{s}_m\|$ to deterministic quantities, thanks to concentration tools.

2.4.1 Bias-Variance decomposition with large probability

Proposition 2.6. *Assume* $\{\mathcal{K}_m\}_{m\in\mathcal{M}}$ *is a finite collection of kernels satisfying Assumptions* (2.11)*,* (2.12)*,* (2.13)*,* (2.14) (2.15) *and* (2.16)*. For all* $x > 1$*, for all* η *in* $(0,1]$*, with probability larger than* $1 - \square|\mathcal{M}|e^{-x}$

$$\|s_m - \widehat{s}_m\|^2 \leqslant (1+\eta)\frac{P\Theta_m}{n} + \square\frac{\Upsilon x^2}{\eta n} \ ,$$

$$\frac{P\Theta_m}{n} \leqslant (1+\eta)\|s_m - \widehat{s}_m\|^2 + \square\frac{\Upsilon x^2}{\eta n} \ .$$

Moreover, for all $x > 1$ *and for all* η *in* $(0,1)$*, with probability larger than* $1 - \square|\mathcal{M}|e^{-x}$*, for all* $m \in \mathcal{M}$*, each of the following inequalities hold*

$$\|s - \widehat{s}_m\|^2 \leqslant (1+\eta)\left(\|s - s_m\|^2 + \frac{P\Theta_m}{n}\right) + \square\frac{\Upsilon x^2}{\eta^3 n} \ ,$$

$$\|s - s_m\|^2 + \frac{P\Theta_m}{n} \leqslant (1+\eta)\|s - \widehat{s}_m\|^2 + \square\frac{\Upsilon x^2}{\eta^3 n} \ .$$

This means that not only in expectation but also with high probability can the term $\|s - \widehat{s}_m\|^2$ be decomposed in a bias term $\|s - s_m\|^2$ and a "variance" term $P\Theta_m/n$. The bias term measures the capacity of the kernel $\mathcal{K}_m$ to approximate s whereas $P\Theta_m/n$ is the price to pay for replacing

s_m by its empirical version $\widehat{s}_m$. In this sense, $P\Theta_m/n$ measures the complexity of the kernel $\mathcal{K}_m$ in a way which is completely adapted to our problem of density estimation. Even if it does not seem like a natural measure of complexity at first glance, note that in the previous examples, it is indeed always linked to a natural complexity. When dealing with regular histograms defined on $[0,1]$, $P\Theta_S$ is the dimension of the considered space S, whereas for approximation kernels $P\Theta_{(k,h)}$ is proportional to the inverse of the considered bandwidth h.

2.4.2 Some general results about the minimal penalty

In this section, we assume that we are in the asymptotic regime where the number of observations $n \to \infty$. In particular, the asymptotic notations refer to this regime.

From now on, the family $\mathcal{M} = \mathcal{M}_n$ may depend on n as long as both Γ and Υ remain absolute constants that do not depend on it. Indeed, on the previous examples, this seem a reasonable regime. Since $\mathcal{M}_n$ now depends on n, our selected $\widehat{m} = \widehat{m}_n$ also depends on n.

To prove that the lower bound $\mathrm{pen}_{\min}(m)$ is sharp, we need to show that the estimate chosen by minimizing (2.6) with a penalty smaller than $\mathrm{pen}_{\min}$ does not satisfy an oracle inequality. Intuitively, this is only possible if the $\|s - \widehat{s}_m\|^2$'s are not of the same order and if they are larger than the remainder term $\square(r + \theta^{-3})\Upsilon x^2/n$. From an asymptotic point of view, we rewrite this thanks to Proposition 2.6 as for all $n \geq 1$, there exist $m_{0,n}$ and $m_{1,n}$ in $\mathcal{M}_n$ such that

$$\left\|s - s_{m_{1,n}}\right\|^2 + \frac{P\Theta_{m_{1,n}}}{n} >> \left\|s - s_{m_{0,n}}\right\|^2 + \frac{P\Theta_{m_{0,n}}}{n} >> \square\left(r + \frac{1}{\theta^3}\right)\frac{\Upsilon x^2}{n} \ , \tag{2.21}$$

where $a_n >> b_n$ means that $b_n/a_n \to_{n\to\infty} 0$. More explicitly, denoting by $\mathrm{o}(1)$ a sequence only depending on n and tending to 0 as n tends to infinity and whose value may change from line to line, one assumes that there exists c_s and c_R positive constants such that for all $n \geq 1$, there exist $m_{0,n}$ and $m_{1,n}$ in $\mathcal{M}_n$ such that

$$\left\|s - s_{m_{0,n}}\right\|^2 + \frac{P\Theta_{m_{0,n}}}{n} \leqslant c_s\, \mathrm{o}(1)\left(\left\|s - s_{m_{1,n}}\right\|^2 + \frac{P\Theta_{m_{1,n}}}{n}\right) \tag{2.22}$$

$$\frac{(\log(|\mathcal{M}_n| \vee n))^3}{n} \leqslant c_R\, \mathrm{o}(1)\left(\left\|s - s_{m_{0,n}}\right\|^2 + \frac{P\Theta_{m_{0,n}}}{n}\right) \ . \tag{2.23}$$

We put a log-cube factor in the remainder term to allow some choices of $\theta = \theta_n \to_{n\to\infty} 0$ and $r = r_n \to_{n\to\infty} +\infty$.

But (2.22) and (2.23) (or (2.21)) are not sufficient. Indeed, the following result explain what happens when the bias terms are always the leading terms.

Corollary 2.1. *Let $(\mathcal{M}_n)_{n\geqslant 1}$ be a sequence of finite collections of kernels $\mathcal{K}_m$ satisfying Assumptions* (2.11)*,* (2.12)*,* (2.13)*,* (2.14) (2.15)*,* (2.16) *for a positive constant Υ independent of n and such that*

$$\frac{1}{n} = c_b\, \mathrm{o}(1) \inf_{m\in\mathcal{M}_n} \frac{\|s - s_m\|^2}{P\Theta_m} \ , \tag{2.24}$$

for some positive constant c_b.

Assume that there exist real numbers of any sign $\delta' \geq \delta$ and a sequence $(r_n)_{n\geqslant 1}$ of nonnegative real numbers such that, for all $n \geq 1$, with probability larger than $1 - \square/n^2$, for all $m \in \mathcal{M}_n$,

$$\delta \frac{P\Theta_m}{n} - \square_{\delta,\delta',\Upsilon} \frac{r_n \log(n \vee |\mathcal{M}_n|)^2}{n} \leq \mathrm{pen}(m) - \frac{2P\chi_m - P\Theta_m}{n} \leq \delta' \frac{P\Theta_m}{n} + \square_{\delta,\delta',\Upsilon} \frac{r_n \log(n \vee |\mathcal{M}_n|)^2}{n} \ .$$

Then, with probability larger than $1 - \square/n^2$,

$$\|s - \widehat{s}_{\widehat{m}_n}\|^2 \leq (1 + \square_{\delta,\delta',\Upsilon,c_b}\, \mathrm{o}(1)) \inf_{m\in\mathcal{M}_n} \|s - \widehat{s}_m\|^2 + \square_{\delta,\delta',\Upsilon} \left(r_n + \log n\right) \frac{\log(n \vee |\mathcal{M}_n|)^2}{n} \ .$$

The proof easily follows by taking $\theta = (\log n)^{-1}$ in (2.17), $\eta = 2$ for instance in Proposition 2.6 and by using Assumption (2.24) and the bounds on $\mathrm{pen}(m)$. This result shows that the estimator $\widehat{s}_{\widehat{m}_n}$ satisfies an asymptotically optimal oracle inequality when condition (2.24) holds, whatever the values of δ and δ' even when they are negative. This proves that the lower bound $\mathrm{pen}_{\min}$ is not sharp in this case.

Therefore, we have to assume that at least one bias $\|s - s_m\|^2$ is negligible with respect to $P\Theta_m/n$. Actually, to conclude, we assume that this happens for $m_{1,n}$ in (2.21).

Theorem 2.2. *Let $(\mathcal{M}_n)_{n\geqslant 1}$ be a sequence of finite collections of kernels satisfying Assumptions* (2.11)*,* (2.12)*,* (2.13)*,* (2.14) (2.15)*,* (2.16)*, with Υ not depending on n. The sequence is also assumed to satisfy* (2.22) *and* (2.23) *such that the kernel $m_{1,n} \in \mathcal{M}_n$ in* (2.22) *satisfies*

$$\left\|s - s_{m_{1,n}}\right\|^2 \leqslant c\, \mathrm{o}(1) \frac{P\Theta_{m_{1,n}}}{n} \ , \tag{2.25}$$

for some fixed positive constant c. Suppose that there exist $\delta \geqslant \delta' > 0$ and a sequence $(r_n)_{n\geqslant 1}$ of nonnegative real numbers such that $r_n \leqslant \square \log(|\mathcal{M}_n| \vee n)$ and such that for all $n \geq 1$, with probability larger than $1 - \square/n^2$, for all $m \in \mathcal{M}_n$,

$$\frac{2P\chi_m - P\Theta_m}{n} - \delta \frac{P\Theta_m}{n} - \square_{\delta,\delta',\Upsilon} \frac{r_n \log(|\mathcal{M}_n| \vee n)^2}{n} \leq \mathrm{pen}(m) \leq \frac{2P\chi_m - P\Theta_m}{n} - \delta' \frac{P\Theta_m}{n} + \square_{\delta,\delta',\Upsilon} \frac{r_n \log(|\mathcal{M}_n| \vee n)^2}{n} \ . \tag{2.26}$$

Then, with probability larger than $1 - \square/n^2$, the following holds

$$P\Theta_{\widehat{m}_n} \geq \left(\frac{\delta'}{\delta} + \square_{\delta,\delta',\Upsilon,c,c_s,c_R}\, \mathrm{o}(1) \right) P\Theta_{m_{1,n}} \ , \tag{2.27}$$

$$\|s - \widehat{s}_{\widehat{m}_n}\|^2 \geq \left(\frac{\delta'}{\delta} + \square_{\delta,\delta',\Upsilon,c,c_s,c_R}\, \mathrm{o}(1) \right) \left\|s - \widehat{s}_{m_{1,n}}\right\|^2 >> \left\|s - \widehat{s}_{m_{0,n}}\right\|^2 \geq \inf_{m\in\mathcal{M}_n} \|s - \widehat{s}_m\|^2 \ . \tag{2.28}$$

By (2.28), under the conditions of Theorem 2.2, the estimator $\widehat{s}_{\widehat{m}_n}$ cannot satisfy an oracle inequality, hence, the lower bound $(2P\chi_m - P\Theta_m)/n$ in Theorem 2.1 is sharp. This shows that $(2P\chi_m - P\Theta_m)/n$ is a minimal penalty in the sense of Birgé & Massart (2007) for kernel selection. When

$$\mathrm{pen}(m) = \frac{2P\chi_m - P\Theta_m}{n} + \kappa \frac{P\Theta_m}{n} \ ,$$

the complexity $P\Theta_{\widehat{m}_n}$ presents a sharp phase transition when κ becomes positive. Indeed, when $\kappa < 0$ it follows from (2.27) that the complexity $P\Theta_{\widehat{m}_n}$ is asymptotically larger than $P\Theta_{m_{1,n}}$. But on the other hand, as a consequence of Theorem 2.1, when $\kappa > 0$, this complexity becomes smaller than

$$\square_\kappa n \inf_{k \in \mathcal{M}_n} \left(\|s - s_m\|^2 + \frac{P\Theta_m}{n} \right) \leq \square_\kappa \left(n \left\|s - s_{m_{0,n}}\right\|^2 + P\Theta_{m_{0,n}} \right)$$
$$<< \square_\kappa \left(n \left\|s - s_{m_{1,n}}\right\|^2 + P\Theta_{m_{1,n}} \right) \leq \square_\kappa P\Theta_{m_{1,n}} \ .$$

2.4.3 Examples

Example 1 (continued). Let $\mathbb{S} = \mathbb{S}_n$ be the collection of spaces of regular histograms on $[0,1]$ with dimensions $\{1, \ldots, n\}$ and let $\widehat{S} = \widehat{S}_n$ be the selected space thanks to the penalized criterion. Recall that, for any $S \in \mathbb{S}_n$, the orthonormal basis is defined by $\varphi_i = \sqrt{D_S}\mathbf{1}_i$ and $P\Theta_S = D_S$. Assume that s is α-Hölderian, with $\alpha \in (0,1]$ with α-Hölderian norm L. It is well known (see for instance Section 1.3.3. of Birgé (2006b)) that the bias is upper bounded by

$$\|s - s_S\|^2 \leq \square_L D_S^{-2\alpha} \ .$$

In particular, if $D_{S_1} = n$,

$$\|s - s_{S_1}\|^2 \leq \square_L n^{-2\alpha} << 1 = \frac{D_{S_1}}{n} = \frac{P\Theta_{S_1}}{n} \ .$$

Thus, (2.25) holds for kernel $\mathcal{K}_{S_1}$. Moreover, if $D_{S_0} = \lfloor \sqrt{n} \rfloor$,

$$\frac{(\log(n \vee |\mathbb{S}_n|)^3}{n} << \|s - s_{S_0}\|^2 + \frac{D_{S_0}}{n} \leq \square_L \left(\frac{1}{n^\alpha} + \frac{1}{\sqrt{n}} \right) << \|s - s_{S_1}\|^2 + \frac{D_{S_1}}{n} \ .$$

Hence, (2.21) holds with $m_{0,n} = S_0$ and $m_{1,n} = S_1$. Therefore, Theorem 2.2 and Theorem 2.1 apply in this example. If $\mathrm{pen}(S) = (1-\delta)D_S/n$, the dimension $D_{\widehat{S}_n} \geq \square_\delta n$ and $\widehat{s}_{\widehat{S}_n}$ is not consistent and does not satisfy an oracle inequality. On the other hand, if $\mathrm{pen}(S) = (1+\delta)D_S/n$,

$$D_{\widehat{S}_n} \leq \square_{L,\delta} \left(n^{1-\alpha} + \sqrt{n} \right) << D_{S_1} = n$$

and $\widehat{s}_{\widehat{S}_n}$ satisfies an oracle inequality which implies that, with probability larger than $1 - \square/n^2$,

$$\left\| s - \widehat{s}_{\widehat{S}_n} \right\|^2 \leq \square_{\alpha,L,\delta} n^{-2\alpha/(2\alpha+1)} \ ,$$

by taking $D_S \simeq n^{1/(2\alpha+1)}$. It achieves the minimax rate of convergence over the class of α-Hölderian functions.

From Theorem 2.1, the penalty $\mathrm{pen}(S) = 2D_S/n$ provides an estimator $\widehat{s}_{\widehat{S}_n}$ that achieves an asymptotically optimal oracle inequality. Therefore the optimal penalty is equal to 2 times the minimal one. In particular, the slope heuristics of Birgé & Massart (2007) holds in this example, as already noticed in Lerasle (2012).

Finally to illustrate Corollary 2.1, let us take $s(x) = 2x$ and the collection of regular histograms with dimension in $\{1, \dots, \lfloor n^\beta \rfloor\}$, with $\beta < 1/3$. Simple calculations show that

$$\frac{\|s - s_S\|^2}{D_S} \geqslant \square D_S^{-3} \geqslant \square n^{-3\beta} >> n^{-1}.$$

Hence (2.24) applies and the penalized estimator with penalty satisfying $\mathrm{pen}(S) \simeq \delta \frac{D_S}{n}$ always satisfies an oracle inequality even if $\delta = 0$ or $\delta < 0$. This was actually expected since it is likely to choose the largest dimension which is also the oracle choice in this case.

Example 2 (continued). Let k be a fixed function, let $\mathcal{H} = \mathcal{H}_n$ denote the following grid of bandwidths

$$\mathcal{H} = \left\{ \frac{\|k\|_\infty \|k\|_1}{i} \quad / \quad i = 1, \dots, n \right\}$$

and let $\widehat{h} = \widehat{h}_n$ be the selected bandwidth. Assume as before that s is a density on $[0,1]$ that belongs to the Nikol'ski class $\mathcal{N}(\alpha, L)$ with $\alpha \in (0,1]$ and $L > 0$. By Proposition 1.5 in Tsybakov (2009), if k satisfies $\int |u|^\alpha |k(u)| \, du < \infty$

$$\left\| s - s_{(k,h)} \right\|^2 \leq \square_{\alpha,k,L} h^{2\alpha} \ .$$

In particular, when $h_1 = \|k\|_\infty \|k\|_1 / n$,

$$\left\| s - s_{(k,h_1)} \right\|^2 \leq \square_{\alpha,k,L} n^{-2\alpha} << \frac{P\Theta_{(k,h_1)}}{n} = \frac{\|k\|^2}{\|k\|_\infty \|k\|_1} \ .$$

On the other hand, for $h_0 = \|k\|_\infty \|k\|_1 / \lfloor \sqrt{n} \rfloor$,

$$\begin{aligned} \frac{(\log n \vee |\mathcal{H}_n|)^2}{n} &<< \left\| s - s_{(k,h_0)} \right\|^2 + \frac{P\Theta_{(k,h_0)}}{n} \\ &\leq \square_{k,\alpha,L} \left(\frac{1}{n^\alpha} + \frac{1}{\sqrt{n}} \right) << \left\| s - s_{(k,h_1)} \right\|^2 + \frac{P\Theta_{(k,h_1)}}{n} \ . \end{aligned}$$

Hence, (2.21) and (2.25) hold with kernels $m_{0,n} = (k, h_0)$ and $m_{1,n} = (k, h_1)$. Therefore, Theorem 2.2 and Theorem 2.1 apply in this example. If for some $\delta > 0$ we set $\mathrm{pen}(k,h) = (2k(0) - \|k\|^2 - \delta \|k\|^2)/(nh)$, then $\widehat{h}_n \leq \square_{\delta,k} n^{-1}$ and $\widehat{s}_{(k,\widehat{h}_n)}$ is not consistent and does not satisfy an oracle inequality. On the other hand, if $\mathrm{pen}(k,h) = (2k(0) - \|k\|^2 + \delta \|k\|^2)/(nh)$, then

$$\widehat{h}_n \geq \square_{\delta,k,L} \left(n^{1-\alpha} + \sqrt{n} \right)^{-1} >> \square_{\delta,k,L} n^{-1} \ ,$$

and $\widehat{s}_{(k,\widehat{h}_n)}$ satisfies an oracle inequality which implies that, with probability larger than $1-\square/n^2$,

$$\left\| s - \widehat{s}_{(k,\widehat{h}_n)} \right\|^2 \leq \square_{\alpha,k,L,\delta} n^{-2\alpha/(2\alpha+1)} \ ,$$

for $h = \|k\|_\infty \|k\|_1 / \left\lfloor n^{1/(2\alpha+1)} \right\rfloor$. In particular it achieves the minimax rate of convergence over the class $\mathcal{N}(\alpha, L)$. Finally, if $\mathrm{pen}(k,h) = 2k(0)/(nh)$, $\widehat{s}_{(k,\widehat{h}_n)}$ achieves an asymptotically optimal oracle inequality, thanks to Theorem 2.1.

The minimal penalty is therefore

$$\mathrm{pen}_{\min}(k,h) = \frac{2k(0) - \|k\|^2}{nh} \ .$$

In this case, the optimal penalty $\mathrm{pen}_{\mathrm{opt}}(k,h) = 2k(0)/(nh)$ derived from Theorem 2.1 is not twice the minimal one, but one still has, if $2k(0) \neq \|k\|^2$,

$$\mathrm{pen}_{\mathrm{opt}}(k,h) = \frac{2k(0)}{2k(0) - \|k\|^2} \, \mathrm{pen}_{\min}(k,h) \ ,$$

even if they can be of opposite sign depending on k. This type of nontrivial relationship between optimal and minimal penalty has already been underlined in Arlot & Bach (2009) in regression framework for selecting linear estimators.

Note that if one allows two kernel functions k_1 and k_2 in the family of kernels such that $2k_1(0) \neq \|k_1\|^2$, $2k_2(0) \neq \|k_2\|^2$ and

$$\frac{2k_1(0)}{2k_1(0) - \|k_1\|^2} \neq \frac{2k_2(0)}{2k_2(0) - \|k_2\|^2} \ ,$$

then there is no absolute constant factor linking the minimal penalty and the optimal one.

2.5 Short simulation study

In this section we illustrate on synthetic data Theorem 2.1 and Theorem 2.2. We focus on approximation kernels only since projection kernels have been already discussed in Lerasle (2012).

We observe an $n = 100$ i.i.d. sample of standard Gaussian distribution. For a fixed parameter $a \geq 0$ we consider the family of kernels

$$\mathcal{K}_{k_a,h}(x,y) = \frac{1}{h} k_a\left(\frac{x-y}{h}\right), \qquad \text{with} \quad h \in \mathcal{H} = \left\{ \frac{1}{2i}, \ i = 1, \ldots, 50 \right\} \ ,$$

where for $x \in \mathbb{R}$

$$k_a(x) = \frac{1}{2\sqrt{2\pi}} \left(e^{-\frac{(x-a)^2}{2}} + e^{-\frac{(x+a)^2}{2}} \right) \ .$$

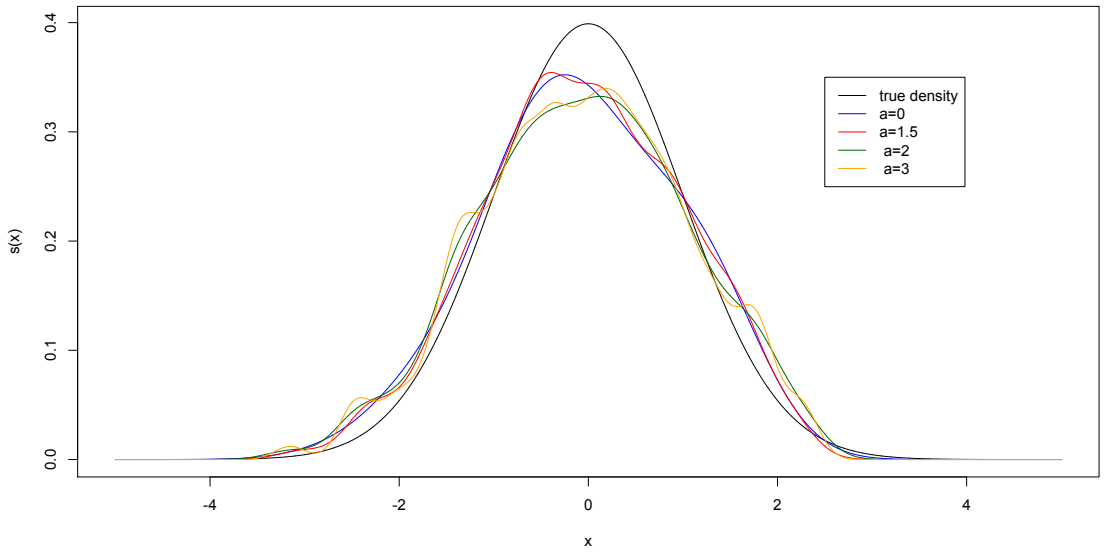

Figure 2.1: Selected approximation kernel estimators when the penalty is the optimal one i.e. $\frac{2k_a(0)}{nh}$.

In particular the kernel estimate with $a = 0$ is the classical Gaussian kernel estimate. Moreover

$$k_a(0) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{a^2}{2}\right), \quad \text{and} \quad \|k_a\|^2 = \frac{1+e^{-a^2}}{4\sqrt{\pi}} \ .$$

Thus, depending on the value of a, the minimal penalty $(2k_a(0) - \|k_a\|^2)/(nh)$ may be negative. We study the behavior of the penalized criterion $\mathcal{C}_{\text{pen}}(k_a, h) = P_n\gamma(\widehat{s}_{(k_a,h)}) + \text{pen}(k_a, h)$ with penalties of the form

$$\text{pen}(k_a, h) = \frac{2k_a(0) - \|k_a\|^2}{nh} + \kappa\frac{\|k_a\|^2}{nh} \ , \tag{2.29}$$

for different values of κ ($\kappa = -1, 0, 1$) and a ($a = 0, 1.5, 2, 3$). On Figure 2.1 are represented the selected estimates by the optimal penalty $2k_a(0)/(nh)$ for the different values of a and on Figure 2.2 one sees the evolution of the different penalized criteria as a function of $1/h$.

The contrast curves for $a = 0$ are classical on Figure 2.2. Without penalization, the criterion decreases and leads to the selection of the smallest bandwidth. At the minimal penalty, the curve is flat and at the optimal penalty one selects a meaningful bandwidth as shown on Figure 2.1.

When $a > 0$, despite the choice of those unusual kernels, the reconstructions on Figure 2.1 for the optimal penalty are also meaningful. However when $a = 2$ or $a = 3$, the criterion with minimal penalty is smaller than the unpenalized criterion, meaning that minimizing the latter criterion leads by Theorem 2.1 to an oracle inequality. On the presented simulation, when $a = 3$, the curves for the optimal criterion and the unpenalized one are so close that the same estimator is selected by both methods.

Finally Figure 2.3 shows that there is indeed in all cases a sharp phase transition around $\kappa = 0$ i.e. at the minimal penalty for the complexity of the selected estimate.

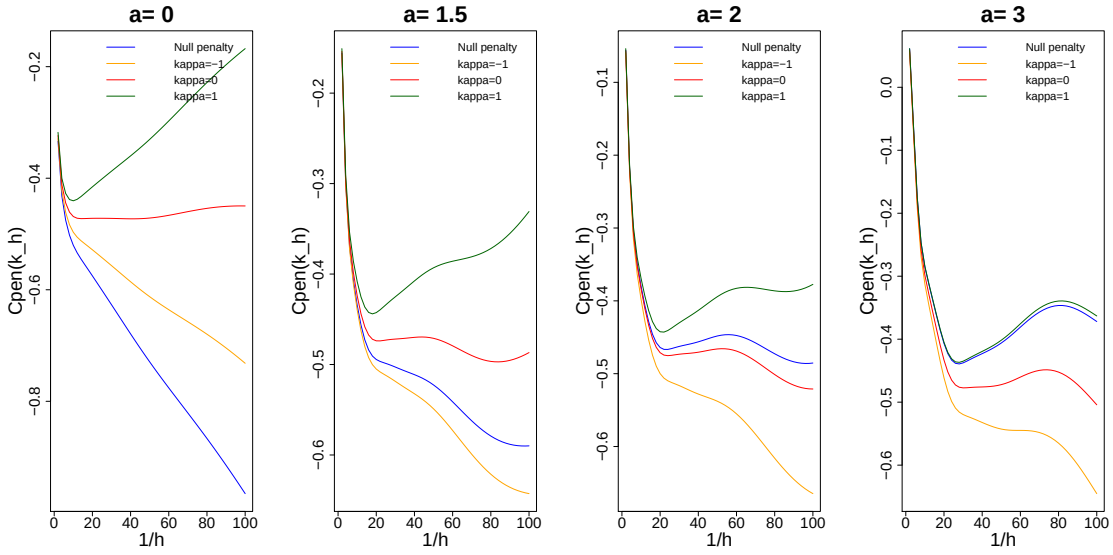

Figure 2.2: Behavior of $P_n\gamma(\widehat{s}_{(k_a,h)})$ (blue line) and $\mathcal{C}_{\mathrm{pen}}\left(k_a,h\right)$ as a function of $1/h$, which is proportional to the complexity $P\Theta_{(k_a,h)}$.

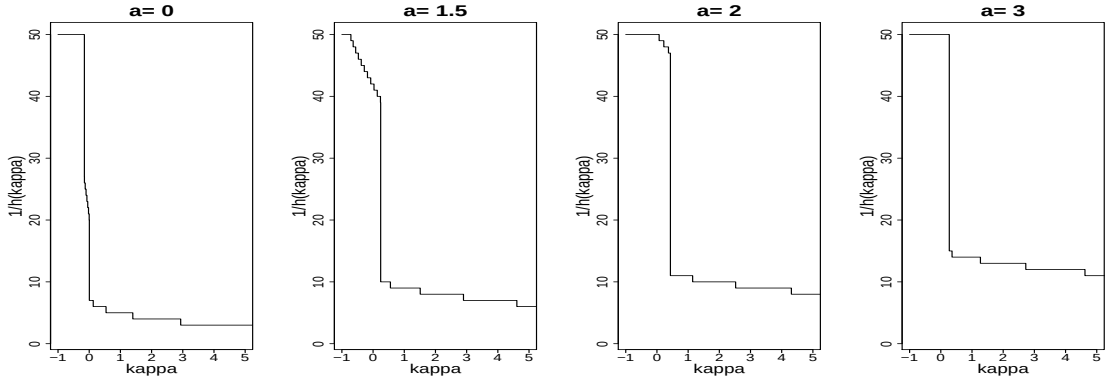

Figure 2.3: Behavior of $1/\widehat{h}$, which is proportional to the complexity $P\Theta_{(k_a,\widehat{h})}$, for the estimate selected by the criterion whose penalty is given by (2.29), as a function of κ.

2.6 Main Proofs

2.6.1 Proof of Theorem 2.1

The starting point to prove the oracle inequality is to notice that any minimizer $\widehat{m}$ of $\mathcal{C}_{\mathrm{pen}}$ satisfies

$$\|s-\widehat{s}_{\widehat{m}}\|^2 \leq \|s-\widehat{s}_m\|^2 + (\,\mathrm{pen}(m)-\mathrm{pen}_{\mathrm{id}}(m)\,) - (\,\mathrm{pen}\,(\widehat{m}\,)-\mathrm{pen}_{\mathrm{id}}\,(\widehat{m}\,)\,) \quad .$$

Using the expression of the ideal penalty (2.9) we find

$$\begin{aligned}\|s-\widehat{s}_{\widehat{m}}\|^2 \leq \|s-\widehat{s}_m\|^2 &+ \left(\mathrm{pen}(m)-2\frac{P\chi_m}{n}\right) - \left(\mathrm{pen}\,(\widehat{m}\,)-2\frac{P\chi_{\widehat{m}}}{n}\right)\\ &+ 2\frac{P(s_m-s_{\widehat{m}})}{n} + 2\left(1-\frac{2}{n}\right)(P_n-P)(s_{\widehat{m}}-s_m)\\ &+ 2\frac{(P_n-P)(\chi_{\widehat{m}}-\chi_m)}{n} + 2\frac{U_{\widehat{m}}-U_m}{n^2} \quad . \end{aligned} \tag{2.30}$$

By Proposition 2.7 (see the appendix), for all $x>1$, for all θ in $(0,1)$, with probability larger than $1-(7.4|\mathcal{M}|+2|\mathcal{M}|^2)e^{-x}$,

$$\begin{aligned}\|s-\widehat{s}_{\widehat{m}}\|^2 &\leq \|s-\widehat{s}_m\|^2 + \left(\mathrm{pen}(m)-2\frac{P\chi_m}{n}\right) - \left(\mathrm{pen}\,(\widehat{m}\,)-2\frac{P\chi_{\widehat{m}}}{n}\right)\\ &+ \theta\,\|s-s_{\widehat{m}}\|^2 + \theta\,\|s-s_m\|^2 + \square\frac{\Upsilon}{\theta n}\\ &+ \left(1-\frac{2}{n}\right)\theta\,\|s-s_{\widehat{m}}\|^2 + \left(1-\frac{2}{n}\right)\theta\,\|s-s_m\|^2 + \square\frac{\Upsilon x^2}{\theta n}\\ &+ \theta\frac{P\Theta_m}{n} + \theta\frac{P\Theta_{\widehat{m}}}{n} + \square\frac{\Upsilon x}{\theta n} + \theta\frac{P\Theta_m}{n} + \theta\frac{P\Theta_{\widehat{m}}}{n} + \square\frac{\Upsilon x^2}{\theta n}\\ &\leqslant \|s-\widehat{s}_m\|^2 + \left(\mathrm{pen}(m)-2\frac{P\chi_m}{n}\right) - \left(\mathrm{pen}\,(\widehat{m}\,)-2\frac{P\chi_{\widehat{m}}}{n}\right)\\ &+ 2\theta\left[\|s-s_{\widehat{m}}\|^2 + \frac{P\Theta_{\widehat{m}}}{n}\right] + 2\theta\left[\|s-s_m\|^2 + \frac{P\Theta_m}{n}\right] + \square\frac{\Upsilon x^2}{\theta n} \quad . \end{aligned}$$

This bound holds using (2.11), (2.12) and (2.13) only. Now by Proposition 2.6 applied with $\eta=1$, we have for all $x>1$, for all $\theta\in(0,1)$, with probability larger than $1-(16.8|\mathcal{M}|+2|\mathcal{M}|^2)e^{-x}$,

$$\begin{aligned}\|s-\widehat{s}_{\widehat{m}}\|^2 \leq \|s-\widehat{s}_m\|^2 &+ \left(\mathrm{pen}(m)-2\frac{P\chi_m}{n}\right) - \left(\mathrm{pen}\,(\widehat{m}\,)-2\frac{P\chi_{\widehat{m}}}{n}\right)\\ &+ 4\theta\,\|s-\widehat{s}_{\widehat{m}}\|^2 + 4\theta\,\|s-\widehat{s}_m\|^2 + \square\frac{\Upsilon x^2}{\theta n} \quad . \end{aligned}$$

This gives the first part of the theorem.

For the second part, by the condition (2.18) on the penalty, we find for all $x > 1$, for all θ in $(0,1)$, with probability larger than $1-(C+16.8|\mathcal{M}|+2|\mathcal{M}|^2)e^{-x}$,

$$(1-4\theta)\left\|s-\widehat{s}_{\widehat{m}}\right\|^2 \leq (1+4\theta)\left\|s-\widehat{s}_m\right\|^2+(\delta'-1)_+\frac{P\Theta_m}{n}+(1-\delta)_+\frac{P\Theta_{\widehat{m}}}{n}+\square\left(r+\frac{1}{\theta}\right)\frac{\Upsilon x^2}{n} \ .$$

By Proposition 2.6 applied with $\eta=\theta$, we have with probability larger than $1-(C+26.2|\mathcal{M}|+2|\mathcal{M}|^2)e^{-x}$,

$$\begin{aligned}(1-4\theta)\left\|s-\widehat{s}_{\widehat{m}}\right\|^2 \leq (1+4\theta)\left\|s-\widehat{s}_m\right\|^2 &+ (\delta'-1)_+(1+\theta)\left\|s-\widehat{s}_m\right\|^2 \\ &+ (1-\delta)_+(1+\theta)\left\|s-\widehat{s}_{\widehat{m}}\right\|^2+\square\left(r+\frac{1}{\theta^3}\right)\frac{\Upsilon x^2}{n} \ ,\end{aligned}$$

that is

$$\begin{aligned}&\left((\delta\wedge 1)-\theta(4+(1-\delta)_+)\right)\left\|s-\widehat{s}_{\widehat{m}}\right\|^2 \\ &\qquad\leq \left((\delta'\vee 1)+\theta(4+(\delta'-1)_+)\right)\left\|s-\widehat{s}_m\right\|^2+\square\left(r+\frac{1}{\theta^3}\right)\frac{\Upsilon x^2}{n} \ .\end{aligned}$$

Hence, because $1\leqslant[(\delta'\vee 1)+(4+(\delta'-1)_+)\theta]\leqslant(\delta'\vee 1)+(4+\delta')\theta$, we obtain for all $m\in\mathcal{M}$

$$\frac{(\delta\wedge 1)-5\theta}{(\delta'\vee 1)+(4+\delta')\theta}\left\|s-\widehat{s}_{\widehat{m}}\right\|^2 \leq \left\|s-\widehat{s}_m\right\|^2+\square\left(r+\frac{1}{\theta^3}\right)\frac{\Upsilon x^2}{n} \ .$$

2.6.2 Proof of Proposition 2.6

Proof: First, let us denote for all $x\in\Xi$

$$F_{A,m}(x):=\mathbb{E}\left[A_m(X,x)\right], \qquad \zeta_m(x):=\int\left(\mathcal{K}_m(y,x)-s_m(y)\right)^2 d\mu(y) \ ,$$

and

$$U_{A,m}:=\sum_{i\neq j=1}^{n}\left(A_m(X_i,X_j)-F_{A,m}(X_i)-F_{A,m}(X_j)+\mathbb{E}\left[A_m(X,Y)\right]\right) \ .$$

Some easy computations then provide the following useful equality

$$\left\|s_m-\widehat{s}_m\right\|^2=\frac{1}{n}P_n\zeta_m+\frac{1}{n^2}U_{A,m} \ .$$

We concentrate both terms on the right-hand side thanks to the probability tools of Section 2.2.3. Using Proposition 2.1, we get, for any $x\geq 1$, with probability larger than $1-2\left|\mathcal{M}\right|e^{-x}$,

$$\left|(P_n-P)\zeta_m\right|\leq\sqrt{\frac{2x}{n}P\zeta_m^2}+\frac{\left\|\zeta_m\right\|_\infty x}{3n} \ .$$

One can then check the following link between ζ_m and Θ_m

$$P\zeta_m = \int \left(\mathcal{K}_m(y,x) - s_m(x)\right)^2 s(y)d\mu(x)d\mu(y) = P\Theta_m - \|s_m\|^2 \ .$$

Next, by (2.1) and (2.11)

$$\|\zeta_m\|_\infty = \sup_{y\in\Xi} \int \left(\mathcal{K}_m(y,x) - \mathbb{E}\left[\mathcal{K}_m(X,x)\right]\right)^2 d\mu(x) \leq 4 \sup_{y\in\Xi} \int \mathcal{K}_m(y,x)^2 d\mu(x) \leq 4\Upsilon n \ .$$

In particular, since $\zeta_m \geq 0$,

$$P\zeta_m^2 \leq \|\zeta_m\|_\infty P\zeta_m \leq 4\Upsilon n P\Theta_m \ .$$

It follows from these computations and from (2.11) that there exists an absolute constant $\square$ such that, for any $x \geq 1$, with probability larger than $1 - 2\left|\mathcal{M}\right| e^{-x}$, for any $\theta \in (0,1)$,

$$|P_n\zeta_m - P\Theta_m| \leq \theta P\Theta_m + \square \frac{\Upsilon x}{\theta} \ .$$

We pursue the proof with the control of $U_{A,m}$. From Proposition 2.2, for any $x \geq 1$, with probability larger than $1 - 5.4\left|\mathcal{M}\right| e^{-x}$,

$$\frac{|U_{A,m}|}{n^2} \leq \frac{\square}{n^2}\left(C\sqrt{x} + Dx + Bx^{3/2} + Ax^2\right) \ .$$

By (2.1), (2.11) and Cauchy-Schwarz inequality,

$$A = 4 \sup_{(x,y)\in\Xi^2} \int \mathcal{K}_m(x,z)\mathcal{K}_m(y,z)d\mu(z) \leq 4 \sup_{x\in\Xi} \int \mathcal{K}_m(x,z)^2 d\mu(z) \leq 4\Upsilon n \ .$$

In addition, by (2.15),

$$B^2 \leq 16 \sup_{x\in\Xi} \mathbb{E}\left[A_m(X,x)^2\right] \leq 16\Upsilon n \ .$$

Moreover, using Assumption (2.14),

$$C^2 \leq \sum_{i\neq j=1}^{n} \mathbb{E}\left[A_m(X_i,X_j)^2\right] = n^2\mathbb{E}\left[A_m(X,Y)^2\right] \leq n^2\Upsilon P\Theta_m \ .$$

Finally, using Cauchy-Schwarz inequality and proceeding as for C^2, the quantity used to define D can be upper bounded as follows:

$$\mathbb{E}\left[\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} a_i(X_i)b_j(X_j)A_m(X_i,X_j)\right] \leq n\sqrt{\mathbb{E}\left[A_m(X,Y)^2\right]} \leq n\sqrt{\Upsilon P\Theta_m} \ .$$

Hence for any $x \geq 1$, with probability larger than $1 - 5.4\left|\mathcal{M}\right| e^{-x}$, for any $\theta \in (0,1)$,

$$\frac{|U_{A,m}|}{n^2} \leq \theta\frac{P\Theta_m}{n} + \square\frac{\Upsilon x^2}{\theta n} \ .$$

Therefore, for all $\theta \in (0,1)$,

$$\left| \|\widehat{s}_m - s_m\|^2 - \frac{P\Theta_m}{n} \right| \leq 2\theta \frac{P\Theta_m}{n} + \square \frac{\Upsilon x^2}{\theta n} \ ,$$

and the first part of the result follows by choosing $\theta = \eta/2$.

Concerning the two remainder inequalities appearing in the proposition, we begin by developing the loss. For all $m \in \mathcal{M}$

$$\|\widehat{s}_m - s\|^2 = \|\widehat{s}_m - s_m\|^2 + \|s_m - s\|^2 + 2\langle \widehat{s}_m - s_m, s_m - s \rangle \ .$$

Then, for all $x \in \Xi$

$$\begin{aligned} F_{A,m}(x) - s_m(x) &= \int s(y) \int \mathcal{K}_m(x,z)\mathcal{K}_m(z,y)d\mu(z)d\mu(y) - \int s(z)\mathcal{K}_m(z,x)d\mu(z) \\ &= \int \left(\int s(y)\mathcal{K}_m(z,y)d\mu(y) - s(z) \right) \mathcal{K}_m(x,z)d\mu(z) \\ &= \int \left(s_m(z) - s(z) \right) \mathcal{K}_m(z,x)d\mu(z) \ . \end{aligned}$$

Moreover, since $PF_{A,m} = \|s_m\|^2$, we find

$$\begin{aligned} \langle \widehat{s}_m - s_m, s_m - s \rangle &= \int \left(\widehat{s}_m(x) \left(s_m(x) - s(x) \right) \right) d\mu(x) + \mathbb{E}\left[s_m(X) \right] - \|s_m\|^2 \\ &= \frac{1}{n} \sum_{i=1}^{n} \int \left(\mathcal{K}_m(x, X_i) \left(s_m(x) - s(x) \right) \right) d\mu(x) + P(s_m - F_{A,m}) \\ &= \frac{1}{n} \sum_{i=1}^{n} \left(F_{A,m}(X_i) - s_m(X_i) \right) + P(s_m - F_{A,m}) \\ &= (P_n - P)(F_{A,m} - s_m) \ . \end{aligned}$$

This expression motivates us to apply again Proposition 2.1 to concentrate this term. We find by (2.1), (2.11) and Cauchy-Schwarz inequality

$$\begin{aligned} \sup_{x \in \Xi} |F_{A,m}(x) - s_m(x)| &\leq \|s - s_m\| \sup_{x \in \Xi} \int \frac{|s(z) - s_m(z)|}{\|s - s_m\|} \mathcal{K}_m(x,z)d\mu(z) \\ &\leq \|s - s_m\| \sqrt{\sup_{x \in \Xi} \int \mathcal{K}_m(x,z)^2 d\mu(z)} \leq \|s - s_m\| \sqrt{\Upsilon n} \ . \end{aligned}$$

Moreover,

$$P\left(F_{A,m} - s_m \right)^2 \leq \|s - s_m\|^2 P \left(\int \frac{|s(z) - s_m(z)|}{\|s - s_m\|} \mathcal{K}_m(.,z)d\mu(z) \right)^2 \leq \|s - s_m\|^2 v_m^2 \ .$$

Thus by (2.16), for any $\theta, u > 0$,

$$\sqrt{\frac{2P\left(F_{A,m}-s_m\right)^2 x}{n}} \leq \theta \left\|s-s_m\right\|^2 + \frac{\left(\Upsilon \vee \sqrt{\Upsilon P \Theta_m}\right) x}{2\theta n}$$
$$\leq \theta \left\|s-s_m\right\|^2 + \frac{\Upsilon x}{\theta n} \vee \left(\frac{u}{\theta}\frac{P\Theta_m}{n} + \frac{\Upsilon x^2}{16\theta u n}\right) \ .$$

Hence, for any $\theta \in (0,1)$ and $x \geq 1$, taking $u = \theta^2$

$$\sqrt{\frac{2P\left(F_{A,m}-s_m\right)^2 x}{n}} \leq \theta\left(\left\|s-s_m\right\|^2 + \frac{P\Theta_m}{n}\right) + \square\frac{\Upsilon x^2}{\theta^3 n} \ .$$

By Proposition 2.1 we have for all θ in $(0,1)$, for all $x > 0$ with probability larger than $1 - 2|\mathcal{M}|e^{-x}$,

$$2\left|\langle \widehat{s}_m - s_m, s_m - s\rangle\right| \leqslant 2\sqrt{\frac{2P\left(F_{A,m}-s_m\right)^2 x}{n}} + 2\left\|s-s_m\right\|\sqrt{\Upsilon n}\frac{x}{3n}$$
$$\leqslant 3\theta\left(\left\|s-s_m\right\|^2 + \frac{P\Theta_m}{n}\right) + \square\frac{\Upsilon x^2}{\theta^3 n} \ .$$

Putting together all of the above, one concludes that for all θ in $(0,1)$, for all $x > 1$, with probability larger than $1 - 9.4|\mathcal{M}|e^{-x}$

$$\left\|\widehat{s}_m - s\right\|^2 - \left\|s_m - s\right\|^2 \leq 3\theta\left\|s-s_m\right\|^2 + (1+4\theta)\frac{P\Theta_m}{n} + \square\frac{\Upsilon x^2}{\theta^3 n}$$

and

$$\left\|\widehat{s}_m - s\right\|^2 - \left\|s_m - s\right\|^2 \geq -3\theta\left(\left\|s-s_m\right\|^2 + \frac{P\Theta_m}{n}\right) + (1-\theta)\frac{P\Theta_m}{n} - \square\frac{\Upsilon x^2}{\theta^3 n} \ .$$

Choosing, $\theta = \eta/4$ leads to the second part of the result.

□

2.6.3 Proof of Theorem 2.2

It comes from (2.17) (applied with $\theta = \square(\log n)^{-1}$ and $x = \square\log(n \vee |\mathcal{M}_n|)$) and Assumption (2.26) that with probability larger than $1 - \square n^{-2}$ we have for any $m \in \mathcal{M}$ and any $n \geqslant 2$

$$\left\|\widehat{s}_{\widehat{m}_n} - s\right\|^2 \leqslant \left(1 + \frac{\square}{\log n}\right)\left\|\widehat{s}_m - s\right\|^2 - (1+\delta')\left(1+\frac{\square}{\log n}\right)\frac{P\Theta_m}{n}$$
$$+ (1+\delta)\left(1+\frac{\square}{\log n}\right)\frac{P\Theta_{\widehat{m}_n}}{n} + \square_{\delta,\delta',\Upsilon}\frac{\log(|\mathcal{M}_n| \vee n)^3}{n} \ . \quad (2.31)$$

Applying this inequality with $m = m_{1,n}$ and using Proposition 2.6 with $\eta = \square(\log n)^{-1/3}$ and $x = \square\log(|\mathcal{M}_n| \vee n)$ as a lower bound for $\left\|\widehat{s}_{\widehat{m}_n} - s\right\|^2$ and as an upper bound for $\left\|\widehat{s}_{m_{1,n}} - s\right\|^2$,

we asymptotically obtain that with probability larger than $1 - \square n^{-2}$,

$$-\delta(1+\square_\delta\,\mathrm{o}(1))\frac{P\Theta_{\widehat{m}}}{n} \leqslant (1+\mathrm{o}(1))\left\|s_{m_{1,n}}-s\right\|^2 - \delta'(1+\square_{\delta'}\,\mathrm{o}(1))\frac{P\Theta_{m_{1,n}}}{n} + \square_{\delta,\delta',\Upsilon}\frac{\log(|\mathcal{M}_n|\vee n)^3}{n} \ .$$

By Assumption (2.25), $\left\|s_{m_{1,n}}-s\right\|^2 \leqslant c\,\mathrm{o}(1)\frac{P\Theta_{m_{1,n}}}{n}$ and by (2.22),

$$\frac{(\log(|\mathcal{M}_n|\vee n))^3}{n} \leqslant c_R c_s\,\mathrm{o}(1)\frac{P\Theta_{m_{1,n}}}{n} \ .$$

This gives (2.27).

In addition, starting with the event where (2.31) holds and using again Proposition 2.6, we finally also have with probability larger than $1-\square n^{-2}$,

$$\|\widehat{s}_{\widehat{m}}-s\|^2 \leqslant \left(1+\frac{\square}{\log n}\right)\left\|\widehat{s}_{m_{1,n}}-s\right\|^2 - (1+\delta')\frac{P\Theta_{m_{1,n}}}{n} + (1+\delta)\,(1+\mathrm{o}(1))\,\|\widehat{s}_{\widehat{m}}-s\|^2 + \square_{\delta,\delta',\Upsilon}\frac{\log(|\mathcal{M}_n|\vee n)^3}{n} \ .$$

This means since $\left\|\widehat{s}_{m_{1,n}}-s\right\|^2 \simeq \frac{P\Theta_{m_{1,n}}}{n}$,

$$(-\delta+\square_\delta\,\mathrm{o}(1))\,\|\widehat{s}_{\widehat{m}}-s\|^2 \leqslant -(\delta'+\square_{\delta',c}\,\mathrm{o}(1))\left\|\widehat{s}_{m_{1,n}}-s\right\|^2 + \square_{\delta,\delta',\Upsilon}\frac{\log(|\mathcal{M}_n|\vee n)^3}{n} \ .$$

This leads to (2.28) by (2.21).

2.7 Proofs for the examples

2.7.1 Computation of the constant Γ for the three examples

We have to show for each family $\{\mathcal{K}_m\}_{m\in\mathcal{M}}$ (see (2.8) and (2.1)) that there exists a constant $\Gamma \geq 1$ such that for all $m \in \mathcal{M}$

$$\sup_{x\in\Xi}|\Theta_m(x)| \leq \Gamma n, \quad \text{and} \quad \sup_{(x,y)\in\Xi^2}|\mathcal{K}_m(x,y)| \leq \Gamma n \ .$$

Example 1: Projection kernels. First, remark that from Cauchy-Schwarz inequality we have for all $(x,y)\in\Xi^2$ $|\mathcal{K}_S(x,y)| \leq \sqrt{\chi_S(x)\chi_S(y)}$ and by orthonormality, for any $(x,x')\in\Xi^2$,

$$A_S(x,x') = \sum_{(i,j)\in\mathcal{I}_S^2}\varphi_i(x)\varphi_j(x')\int_\Xi \varphi_i(y)\varphi_j(y)d\mu(y) = \mathcal{K}_S(x,x') \ .$$

In particular, for any $x \in \Xi$, $\Theta_S(x) = \chi_S(x)$. Hence, projection kernels satisfy (2.1) for $\Gamma = 1 \vee n^{-1} \sup_{S \in \mathbb{S}} \|\chi_S\|_\infty$. We conclude by writing

$$\|\chi_S\|_\infty = \sup_{x \in \Xi} \sum_{i \in \mathcal{I}_S} \varphi_i(x)^2 = \sup_{\substack{(a_i)_{i \in \mathcal{I}} \text{ s.t.} \\ \sum_{i \in \mathcal{I}_S} a_i^2 = 1}} \sup_{x \in \Xi} \left(\sum_{i \in \mathcal{I}_S} a_i \varphi_i(x) \right)^2 .$$

For $f \in S$ we have $\|f\|^2 = \sum_{i \in \mathcal{I}} \langle f, \varphi_i \rangle^2$. Hence with $a_i = \langle f, \varphi_i \rangle$,

$$\|\chi_S\|_\infty = \sup_{f \in S, \|f\| = 1} \|f\|_\infty^2 .$$

Example 2: Approximation kernels. First,

$$\sup_{(x,y) \in \Xi^2} \left| \mathcal{K}_{(k,h)}(x,y) \right| \leq \frac{\|k\|_\infty}{h} .$$

Second, since $k \in \mathbb{L}_1$

$$\Theta_{(k,h)}(x) = \frac{1}{h^2} \int_\Xi k\left(\frac{x-y}{h} \right)^2 dy = \frac{\|k\|^2}{h} \leq \frac{\|k\|_\infty \|k\|_1}{h} .$$

Now $\int k(u)du = 1$ implies $\|k\|_1 \geq 1$, hence (2.1) holds with $\Gamma = 1$ if one assumes that $h \geq \|k\|_\infty \|k\|_1 / n$.

Example 3: Weighted projection kernels. For all $x \in \Xi$

$$\Theta_w(x) = \sum_{i,j=1}^p w_i \varphi_i(x) w_j \varphi_j(x) \int_\Xi \varphi_i(y) \varphi_j(y) d\mu(y) = \sum_{i=1}^p w_i^2 \varphi_i(x)^2 .$$

From Cauchy-Schwarz inequality, for any $(x,y) \in \Xi^2$,

$$|\mathcal{K}_w(x,y)| \leq \sqrt{\Theta_w(x)} \sqrt{\Theta_w(y)} .$$

We thus find that $\mathcal{K}_w$ verifies (2.1) with $\Gamma \geq 1 \vee n^{-1} \sup_{w \in \mathcal{W}} \|\Theta_w\|_\infty$. Since $w_i \leq 1$ we find the announced result which is independent of $\mathcal{W}$.

2.7.2 Proof of Proposition 2.3

Since $\|s_S\|^2 \leq \|s\|^2 \leq \|s\|_\infty$, we find that (2.11) only requires $\Upsilon \geq \Gamma(1 + \|s\|_\infty)$. Assumption (2.12) holds: this follows from $\Upsilon \geq \Gamma$ and

$$\mathbb{E}\left[\chi_S(X)^2 \right] \leq \|\chi_S\|_\infty P\chi_S \leq \Gamma n P \Theta_S .$$

Now for proving Assumption (2.14), we write

$$\begin{aligned}\mathbb{E}\left[A_S(X,Y)^2\right] = \mathbb{E}\left[\mathcal{K}_S(X,Y)^2\right] &= \int_\Xi \mathbb{E}\left[\mathcal{K}_S(X,x)^2\right] s(x)d\mu(x) \\ &\leq \|s\|_\infty \sum_{(i,j)\in\mathcal{I}_S^2} \mathbb{E}\left[\varphi_i(X)\varphi_j(X)\right] \int_\Xi \varphi_i(x)\varphi_j(x)d\mu(x) \\ &= \|s\|_\infty P\Theta_S \leq \Upsilon P\Theta_S \ .\end{aligned}$$

In the same way, Assumption (2.15) follows from $\|s\|_\infty \Gamma \leq \Upsilon$. Suppose (2.19) holds with $S = S+S'$ so that the basis $(\varphi_i)_{i\in\mathcal{I}}$ of S' is included in the one $(\varphi_i)_{i\in\mathcal{J}}$ of S. Since $\|\chi_S\|_\infty \leqslant \Gamma n$ we have

$$\begin{aligned}s_S(x) - s_{S'}(x) = \sum_{j\in\mathcal{J}\setminus\mathcal{I}} (P\varphi_j)\,\varphi_j(x) &\leqslant \sqrt{\sum_{j\in\mathcal{J}\setminus\mathcal{I}} (P\varphi_j)^2 \sum_{j\in\mathcal{J}} \varphi_j(x)^2} \\ &\leqslant \|s_S - s_{S'}\|\,\|\chi_S\|_\infty^{1/2} \leqslant \|s_S - s_{S'}\|\sqrt{\Gamma n} \ .\end{aligned}$$

Hence, (2.13) holds in this case. Assuming (2.20) implies that (2.13) holds since

$$\|s_S - s_{S'}\|_\infty \leqslant \|s_S\|_\infty + \|s_{S'}\|_\infty \leqslant \Upsilon \ .$$

Finally for (2.16), for any $a \in \mathbb{L}_2$,

$$\int_\Xi a(x)\mathcal{K}_S(x,y)d\mu(x) = \sum_{i\in\mathcal{I}} \langle a,\varphi_i\rangle\varphi_i(y) = \Pi_S(a) \ .$$

is the orthogonal projection of a onto S. Therefore, $\mathbb{B}_S$ is the unit ball in S for the $\mathbb{L}_2$-norm and, for any $t \in \mathbb{B}_S$

$$\mathbb{E}\left[t(X)^2\right] \leqslant \|s\|_\infty \|t\|^2 \leqslant \|s\|_\infty \ .$$

2.7.3 Proof of Proposition 2.4

First, since $\|k\|_1 \geq 1$

$$\begin{aligned}\left\|s_{(k,h)}\right\|^2 &= \int_\Xi \left(\int_\Xi s(y)\frac{1}{h}k\left(\frac{x-y}{h}\right)dy\right)^2 dx \\ &= \int_\Xi \left(\int_\Xi s(x+hz)k\,(z)\,dz\right)^2 dx \\ &\leq \|k\|_1^2 \int_\Xi \left(\int_\Xi s(x+hz)\frac{|k\,(z)|}{\|k\|_1}dz\right)^2 dx \\ &\leq \|k\|_1^2 \int_{\Xi^2} s(x+hz)^2\frac{|k\,(z)|}{\|k\|_1}dxdz \leq \|s\|_\infty \|k\|_1^2 \ .\end{aligned}$$

Hence, Assumption (2.11) holds if $\Upsilon \geq 1 + \|s\|_\infty \|k\|_1^2$. Now, we have

$$P\left(\chi^2_{(k,h)}\right) = \frac{k(0)^2}{h^2} = P\Theta_{(k,h)} \frac{k(0)^2}{\|k\|^2 h} \leq nP\Theta_{(k,h)} \frac{k(0)^2}{\|k\|^2 \|k\|_\infty \|k\|_1} \ ,$$

so it is sufficient to have $\Upsilon \geq k(0)/\|k\|^2$ (since $K(0) \leqslant \|K\|_\infty$) to ensure (2.12). Moreover, for any $h \in \mathcal{H}$ and any $x \in \Xi$,

$$s_{(k,h)}(x) = \int_\Xi s(y) \frac{1}{h} k\left(\frac{x-y}{h}\right) dy = \int_\Xi s(x+zh)k(z)dz \leqslant \|s\|_\infty \|k\|_1 \ .$$

Therefore, Assumption (2.13) holds for $\Upsilon \geq 2\|s\|_\infty \|k\|_1$. Then on one hand

$$\begin{aligned} \left|A_{(k,h)}(x,y)\right| &\leq \frac{1}{h^2} \int_\Xi \left| k\left(\frac{x-z}{h}\right) k\left(\frac{y-z}{h}\right)\right| dz \\ &\leq \frac{1}{h} \int_\Xi \left| k\left(\frac{x-y}{h} - u\right) k(u)\right| du \leq \frac{\|k\|^2}{h} \wedge \frac{\|k\|_\infty \|k\|_1}{h} \leq P\Theta_{(k,h)} \wedge n \ . \end{aligned}$$

And on the other hand

$$\begin{aligned} \mathbb{E}\left[\left|A_{(k,h)}(X,x)\right|\right] &\leq \frac{1}{h} \int_{\Xi^2} \left| k\left(\frac{x-y}{h} - u\right) k(u)\right| du \, s(y)dy \\ &= \int_{\Xi^2} |k(v)\, k(u)|\, s(x+h(v-u))dudv \leq \|s\|_\infty \|k\|_1^2 \ . \end{aligned}$$

Therefore,

$$\begin{aligned} \sup_{x\in\Xi} \mathbb{E}\left[A_{(k,h)}(X,x)^2\right] &\leq \sup_{(x,y)\in\Xi^2} \left|A_{(k,h)}(x,y)\right| \sup_{x\in\Xi} \mathbb{E}\left[\left|A_{(k,h)}(X,x)\right|\right] \\ &\leq \left(P\Theta_{(k,h)} \wedge n\right) \|s\|_\infty \|k\|_1^2 \ , \end{aligned}$$

and

$$\mathbb{E}\left[A_{(k,h)}(X,Y)^2\right] \leq \sup_{x\in\Xi} \mathbb{E}\left[A_{(k,h)}(X,x)^2\right] \leq \|s\|_\infty \|k\|_1^2 P\Theta_{(k,h)} \ .$$

Hence Assumption (2.14) and (2.15) hold when $\Upsilon \geq \|s\|_\infty \|k\|_1^2$. Finally let us prove that Assumption (2.16) is satisfied. Let $t \in \mathbb{B}_{(k,h)}$ and $a \in \mathbb{L}_2$ such that $\|a\| = 1$ and $t(y) = \int_\Xi a(x) \frac{1}{h} k\left(\frac{x-y}{h}\right) dx$ for all $y \in \Xi$. Then the following comes from Cauchy-Schwarz inequality

$$t(y) \leqslant \frac{1}{h} \sqrt{\int_\Xi a(x)^2 dx} \sqrt{\int_\Xi k\left(\frac{x-y}{h}\right)^2 dx} \leqslant \frac{\|k\|}{\sqrt{h}} \ .$$

Thus for any $t \in \mathbb{B}_{(k,h)}$

$$Pt^2 \leqslant \|t\|_\infty \langle |t|, s\rangle \leq \frac{\|k\|}{\sqrt{h}} \|s\| = \|s\| \sqrt{P\Theta_{(k,h)}} \leq \sqrt{\Upsilon P\Theta_{(k,h)}} \ .$$

We conclude that all the assumptions hold if $\Upsilon \geq \left(k(0)/\left\|k\right\|^2 \right) \vee \left(1 + 2\left\|s\right\|_\infty \left\|k\right\|_1^2 \right)$.

2.7.4 Proof of Proposition 2.5

Let us define for convenience $\Phi(x) := \sum_{i=1}^p \varphi_i(x)^2$, so $\Gamma \geq 1 \vee n^{-1}\left\|\Phi\right\|_\infty$. Then we have for these kernels: $\Phi(x) \geq \chi_w(x) \geq \Theta_w(x)$ for all $x \in \Xi$. Moreover, denoting by Πs the orthogonal projection of s onto the linear span of $(\varphi_i)_{i=1,\dots,p}$,

$$\left\|s_w\right\|^2 = \sum_{i=1}^p w_i^2 \left(P\varphi_i\right)^2 \leq \left\|\Pi s\right\|^2 \leq \left\|s\right\|^2 \leq \left\|s\right\|_\infty \ .$$

Assumption (2.11) holds for this family if $\Upsilon \geq \Gamma(1 + \left\|s\right\|_\infty)$. We prove in what follows that all the remainder assumptions are valid using only (2.1) and (2.11).

First, it comes from Cauchy-Schwarz inequality that for any $x \in \Xi$, $\chi_w(x)^2 \leq \Phi(x)\Theta_w(x)$. Assumption (2.12) is then automatically satisfied from the definition of Γ

$$\mathbb{E}\left[\chi_w(X)^2\right] \leq \left\|\Phi\right\|_\infty P\Theta_w \leq \Gamma n P\Theta_w \ .$$

Now let w and w' be two vectors in $[0,1]^p$, we have

$$s_w = \sum_{i=1}^p w_i(P\varphi_i)\varphi_i, \qquad s_w - s_{w'} = \sum_{i=1}^p (w_i - w_i')\left(P\varphi_i\right)\varphi_i \ ,$$

hence

$$\left\|s_w - s_{w'}\right\|^2 = \sum_{i=1}^p (w_i - w_i')^2 \left(P\varphi_i\right)^2$$

and, by Cauchy-Schwarz inequality, for any $x \in \Xi$,

$$\left|s_w(x) - s_{w'}(x)\right| \leq \left\|s_w - s_{w'}\right\| \sqrt{\Phi(x)} \leq \left\|s_w - s_{w'}\right\| \sqrt{\Gamma n} \ .$$

Assumption (2.13) follows using (2.11).

Concerning Assumptions (2.14) and (2.15), let us first notice that by orthonormality we have for any $(x, x') \in \Xi^2$

$$A_w(x,x') = \sum_{i=1}^p w_i^2 \varphi_i(x)\varphi_i(x') \ .$$

Therefore, Assumption (2.15) holds since

$$\begin{aligned}\mathbb{E}\left[A_w(X,x)^2\right] &= \int_\Xi \left(\sum_{i=1}^p w_i^2 \varphi_i(y)\varphi_i(x) \right)^2 s(y)d\mu(y) \\ &\leq \left\|s\right\|_\infty \sum_{1 \leq i,j \leq p} w_i^2 w_j^2 \varphi_i(x)\varphi_j(x) \int_\Xi \varphi_i(y)\varphi_j(y)d\mu(y)\end{aligned}$$

$$= \|s\|_\infty \sum_{i=1}^{p} w_i^4 \varphi_i(x)^2 \leq \|s\|_\infty \Phi(x) \leq \|s\|_\infty \Gamma n \ .$$

Assumption (2.14) also holds from similar computations

$$\begin{aligned}
\mathbb{E}\left[A_w(X,Y)^2\right] &= \int_\Xi \mathbb{E}\left[\left(\sum_{i=1}^{p} w_i^2 \varphi_i(X)\varphi_i(x)\right)^2\right] s(x)d\mu(x) \\
&\leq \|s\|_\infty \sum_{1\leq i,j\leq p} w_i^2 w_j^2 \, \mathbb{E}\left[\varphi_i(X)\varphi_j(X)\right] \int_\Xi \varphi_i(x)\varphi_j(x)d\mu(x) \\
&\leq \|s\|_\infty P\Theta_w \ .
\end{aligned}$$

We finish with the proof of (2.16). Let us prove that $\mathbb{B}_w = \mathcal{E}_w$, where

$$\mathcal{E}_w = \left\{ t = \sum_{i=1}^{p} w_i t_i \varphi_i, \ \text{s.t.} \ \sum_{i=1}^{p} t_i^2 \leq 1 \right\} \ .$$

First, notice that any $t \in \mathbb{B}_w$ can be written

$$\int_\Xi a(x)\mathcal{K}_w(x,y)d\mu(x) = \sum_{i=1}^{p} w_i \langle a, \varphi_i\rangle \varphi_i(y) \ .$$

Then, consider some $t \in \mathcal{E}_w$. By definition, there exists some collection $(t_i)_{i=1,\dots,p}$ such that $t = \sum_{i=1}^p w_i t_i \varphi_i$, and $\sum_{i=1}^p t_i^2 \leq 1$. If we set $a = \sum_{i=1}^p t_i \varphi_i$, we immediately observe that $\|a\|^2 = \sum_{i=1}^p t_i^2 \leq 1$, and $\langle a, \varphi_i\rangle = t_i$, meaning that $t \in \mathbb{B}_w$. Conversely, for $t \in \mathbb{B}_w$, there exists some function $a \in \mathbb{L}_2$ such that $\|a\|^2 \leq 1$, and $t = \sum_{i=1}^p w_i \langle a, \varphi_i\rangle \varphi_i$. Since $(\varphi_i)_{i=1,\dots,p}$ is an orthonormal system, one can take $a = \sum_{i=1}^p \langle a, \varphi_i\rangle \varphi_i$. Setting $t_i = \langle a, \varphi_i\rangle$, we find $\|a\|^2 = \sum_{i=1}^p t_i^2$ and $t \in \mathcal{E}_w$. For any $t \in \mathbb{B}_w = \mathcal{E}_w$, we have $\|t\|^2 = \sum_{i=1}^p w_i^2 t_i^2 \leq \sum_{i=1}^p t_i^2 \leq 1$, hence

$$\mathbb{E}\left[t(X)^2\right] \leqslant \|s\|_\infty \|t\|^2 \leq \|s\|_\infty \ .$$

2.8 Concentration of the residual terms

The following proposition gathered the concentration bounds of the remainder terms appearing in (2.30).

Proposition 2.7. *Let $\{\mathcal{K}_m\}_{m\in\mathcal{M}}$ denote a finite collection of kernels satisfying* (2.1) *and suppose that Assumptions* (2.11)*,* (2.12) *and* (2.13) *hold. Then*

$$\forall \theta \in (0,1), \qquad 2\frac{P(s_{\widehat{m}} - s_m)}{n} \leq \theta \|s - s_{\widehat{m}}\|^2 + \theta \|s - s_m\|^2 + \frac{2\Upsilon}{\theta n} \ . \tag{2.32}$$

For any $x \geq 1$, with probability larger than $1 - 2|\mathcal{M}|^2 e^{-x}$, for any $(m, m') \in \mathcal{M}^2$,

$$\forall \theta \in (0,1), \qquad |2(P_n - P)(s_m - s_{m'})| \leq \theta\left(\|s - s_{m'}\|^2 + \|s - s_m\|^2\right) + \square\frac{\Upsilon x^2}{\theta n} \ . \tag{2.33}$$

For any $x \geq 1$, with probability larger than $1 - 2\left|\mathcal{M}\right| e^{-x}$, for any $m \in \mathcal{M}$,

$$\forall \theta \in (0,1), \qquad |2(P_n - P)\chi_m| \leq \theta P\Theta_m + \square\frac{\Upsilon x}{\theta} \ . \tag{2.34}$$

For any $x \geq 1$, with probability larger than $1 - 5.4\left|\mathcal{M}\right| e^{-x}$, for any $m \in \mathcal{M}$,

$$\forall \theta \in (0,1), \qquad \frac{2\left|U_m\right|}{n^2} \leq \theta\frac{P\Theta_m}{n} + \square\frac{\Upsilon x^2}{\theta n} \ . \tag{2.35}$$

Proof:

First for (2.32), notice that, by (2.13), for any $\theta \in (0,1)$

$$\begin{aligned} 2\frac{P(s_{\widehat{m}} - s_m)}{n} \leq 2\frac{\left\|s_{\widehat{m}} - s_m\right\|_\infty}{n} &\leq \frac{2}{n}\left(\Upsilon \vee \left(\frac{\theta}{4} n \left\|s_m - s_{\widehat{m}}\right\|^2 + \frac{\Upsilon}{\theta}\right)\right) \\ &\leq \frac{\theta}{2}\left\|s_m - s_{\widehat{m}}\right\|^2 + \frac{2\Upsilon}{\theta n} \\ &\leq \theta\left\|s - s_{\widehat{m}}\right\|^2 + \theta\left\|s - s_m\right\|^2 + \frac{2\Upsilon}{\theta n} \ . \end{aligned}$$

Then, by Proposition 2.1, with probability larger than $1 - \left|\mathcal{M}\right|^2 e^{-x}$, for any $(m, m') \in \mathcal{M}^2$,

$$(P_n - P)(s_m - s_{m'}) \leq \sqrt{\frac{2P\left(s_m - s_{m'}\right)^2 x}{n}} + \frac{\left\|s_m - s_{m'}\right\|_\infty x}{3n} \ .$$

Since by (2.11)

$$P\left(s_m - s_{m'}\right)^2 \leq \left\|s\right\|_\infty \left\|s_m - s_{m'}\right\|^2 \leq \Upsilon\left\|s_m - s_{m'}\right\|^2 \ ,$$

we deduce that

$$\sqrt{\frac{2P\left(s_m - s_{m'}\right)^2 x}{n}} \leq \frac{\theta}{4}\left\|s_m - s_{m'}\right\|^2 + \frac{2\Upsilon x}{\theta n} \ .$$

Moreover, by (2.13)

$$\frac{\left\|s_m - s_{m'}\right\|_\infty x}{3n} \leq \frac{\theta}{4}\left\|s_m - s_{m'}\right\|^2 + \square\frac{\Upsilon x^2}{\theta n} \ .$$

Hence, for $x \geq 1$, with probability larger than $1 - \left|\mathcal{M}\right|^2 e^{-x}$

$$(P_n - P)(s_m - s_{m'}) \leq \frac{\theta}{2}\left\|s_m - s_{m'}\right\|^2 + \square\frac{\Upsilon x^2}{\theta n} \leq \theta\left(\left\|s - s_{m'}\right\|^2 + \left\|s - s_m\right\|^2\right) + \square\frac{\Upsilon x^2}{\theta n} \ ,$$

which gives (2.33).

Now, using again Proposition 2.1, with probability larger than $1 - \left|\mathcal{M}\right| e^{-x}$, for any $m \in \mathcal{M}$,

$$(P_n - P)\chi_m \leq \sqrt{\frac{2P\left(\chi_m\right)^2 x}{n}} + \frac{\left\|\chi_m\right\|_\infty x}{3n} \ .$$

By (2.1) and (2.11) we find for any $m \in \mathcal{M}$,

$$\|\chi_m\|_\infty \leq \sup_{(x,y)\in\Xi^2} |\mathcal{K}_m(x,y)| \leq \Gamma n \leq \Upsilon n \ .$$

Concerning (2.34), we get by (2.12), $P\chi_m^2 \leq \Upsilon n P\Theta_m$, hence, for any $x \geq 1$ we have with probability larger than $1 - |\mathcal{M}|\, e^{-x}$

$$(P_n - P)\chi_m \leq \theta P\Theta_m + \left(\frac{1}{3} + \frac{1}{2\theta}\right)\Upsilon x \ .$$

For (2.35), we apply Proposition 2.2 to obtain with probability larger than $1 - 2.7\,|\mathcal{M}|\, e^{-x}$, for any $m \in \mathcal{M}$,

$$\frac{U_m}{n^2} \leq \frac{\square}{n^2}\left(C\sqrt{x} + Dx + Bx^{3/2} + Ax^2\right) \ ,$$

where A, B, C, D are defined accordingly to Proposition 2.2. Let us evaluate all these terms. First, $A \leq 4\sup_{(x,y)\in\Xi^2} |\mathcal{K}_m(x,y)| \leq 4\Upsilon n$ by (2.1) and (2.11). Next,

$$C^2 \leq \square n^2 \mathbb{E}\left[\mathcal{K}_m(X,Y)^2\right] \leq \square n^2 \|s\|_\infty P\Theta_m \leq \square n^2 \Upsilon P\Theta_m \ .$$

Using (2.1), we find

$$B^2 \leqslant 4n \sup_{x\in\Xi} \int_\Xi \mathcal{K}_m(x,y)^2 s(y) d\mu(y) \leqslant 4n \|s\|_\infty \Gamma.$$

By (2.11), we consequently have $B^2 \leq 4\Upsilon n$. Finally, using Cauchy-Schwarz inequality and proceeding as for C^2,

$$\mathbb{E}\left[\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} a_i(X_i) b_j(X_j) \mathcal{K}_m(X_i, X_j)\right] \leq n\sqrt{\mathbb{E}\left[\mathcal{K}_m(X,Y)^2\right]} \leq n\sqrt{\Upsilon P\Theta_m} \ .$$

Hence, $D \leq n\sqrt{\Upsilon P\Theta_m}$ which gives (2.35).

□

2.9 Proof of Proposition 2.2

Denote, for any $k \in [n]$, by $\mathcal{F}_k$ the σ-algebra induced by $X_1^k = (X_1, \ldots, X_k)$, by

$$U_k = \sum_{1\leqslant i<j\leq k} f_{i,j}(X_i, X_j) \ ,$$

and by

$$Z_k = \sum_{j=1}^{k-1} f_{k,j}(X_k, X_j) = \sum_{i=1}^{k-1} f_{i,k}(X_i, X_k) = U_k - U_{k-1} \ ,$$

with the convention $\sum_{j=1}^{0} a_j = 0$. Note that U in Proposition 2.2 satisfies $U = 2U_n$. Assumption (2.10) and the independence of the data X_i ensures that, for any $k \geq 2$,

$$\mathbb{E}\left[Z_k \mid \mathcal{F}_{k-1} \right] = 0 \ ,$$

hence U_k is a martingale. Therefore, by Lemma 2.1 below, for any $\lambda > 0$,

$$\mathcal{A}_k = \exp\left(\lambda U_k - \sum_{j \geq 2} \frac{\lambda^j}{j!} \Delta_j^k \right), \text{ where } \Delta_j^k = \sum_{i=1}^{k} \mathbb{E}\left[Z_i^j \,\middle|\, \mathcal{F}_{i-1} \right] \ ,$$

is a supermartingale. In particular

$$\mathbb{E}\left[\mathcal{A}_n \right] \leq 1 = \mathbb{E}\left[\mathcal{A}_0 \right] \ .$$

2.9.1 Evaluation of the $\Delta_j = \Delta_j^n$

We have to evaluate, for any $j \geq 2$,

$$\Delta_j = \sum_{i=1}^{n} \mathbb{E}\left[Z_i^j \,\middle|\, \mathcal{F}_{i-1} \right] \leq \sum_{i=1}^{n} \mathbb{E}\left[|Z_i|^j \,\middle|\, \mathcal{F}_{i-1} \right] = \sum_{i=2}^{n} \mathbb{E}\left[\left| \sum_{\ell=1}^{i-1} f_{i,\ell}(X_i, X_\ell) \right|^j \,\middle|\, \mathcal{F}_{i-1} \right] \ .$$

Denote by

$$H_j = \sum_{i=2}^{n} \mathbb{E}\left[\left| \sum_{\ell=1}^{i-1} f_{i,\ell}(X_i, X_\ell) \right|^j \,\middle|\, \mathcal{F}_{i-1} \right] \ .$$

Hölder's inequality ensures that

$$H_j^{1/j} = \sup_{\sum_{i=2}^{n} \mathbb{E}\left[|a_i(X_i)|^{j/(j-1)} \right] \leq 1} \sum_{\ell=1}^{n-1} \mathbb{E}\left[\sum_{i=\ell+1}^{n} f_{i,\ell}(X_i, X_\ell) a_i(X_i) \,\middle|\, \mathcal{F}_{i-1} \right] \ .$$

We can write $H_j^{1/j}$ as

$$\sup_{(f_1, \ldots, f_{n-1}) \in \mathcal{G}_j} \sum_{\ell=1}^{n-1} f_\ell(X_\ell) \ ,$$

with $\mathbb{E}\left[f_\ell(X_\ell) \right] = 0$ and

$$\mathcal{G}_j = \left\{ x \mapsto \left(\mathbb{E}\left[\sum_{i=\ell+1}^{n} f_{i,\ell}(X_i, x) a_i(X_i) \right] \right)_{\ell = 1, \ldots, n-1} \ \middle/ \ \sum_{i=2}^{n} \mathbb{E}\left[|a_i(X_i)|^{j/(j-1)} \right] \leq 1 \right\} \ .$$

Remark that we cannot apply directly Bousquet (2002) since the $f_\ell(X_\ell)$'s are not i.i.d. but we can apply Lemma 2.2 which is an easy consequence of Rio (2012). However, to apply this lemma we need to take the supremum over a finite set. To do so, we need $\mathcal{G}_j$ to be separable. But by (Bogachev, 2007, Volume II, Chapter 6, p.17), the borelian σ-algebra of a polish space is

countably generated. Therefore by (Billingsley, 1995, Theorem 19.2, p.243), this implies that $\mathbb{L}_1$ is separable and all the functions a^i's are in $\mathbb{L}_1$. Hence $\mathcal{G}_j$ is separable. Hence by taking the limit on a dense subset, we obtain:

$$\mathbb{P}\left(H_j^{1/j} > (1+\varepsilon)\mathbb{E}\left[H_j^{1/j}\right] + \sqrt{2uv_j} + \kappa(\varepsilon)b_j u\right) \leq e^{-u} \ ,$$

with

$$v_j = \sup_{(f_1,\ldots,f_{n-1})\in\mathcal{G}_j} \sum_{\ell=1}^{n-1} \mathbb{E}\left[f_\ell(X_\ell)^2\right], \qquad b_j = \sup_{(f_1,\ldots,f_{n-1})\in\mathcal{G}_j,\ell=1,\ldots,n-1} \|f_\ell\|_\infty \ .$$

Summing over j and applying a union bound, we get

$$\mathbb{P}\left(\exists j \geq 2 : H_j^{1/j} > (1+\varepsilon)\mathbb{E}\left[H_j^{1/j}\right] + \sqrt{2ujv_j} + \kappa(\varepsilon)b_j ju\right) \leq \sum_{j\geq 2} e^{-ju} \ ,$$

As 1 is an obvious upper bound, the above probability is upper bounded by

$$\sum_{j\geq 2} e^{-ju} \wedge 1 = e^{-2u}\frac{1}{1-e^{-u}} \wedge 1 \leq e^{-u} \sup_{v>0}\left(e^v \wedge \frac{1}{e^v-1}\right) \ .$$

Now as e^u is nondecreasing and $1/(e^u-1)$ is nonincreasing, $e^u \wedge \frac{1}{e^u-1}$ is maximized for u such that $e^u = \frac{1}{e^u-1}$, that is for u such that $(e^u - \frac{1+\sqrt{5}}{2})(e^u - \frac{1-\sqrt{5}}{2}) = 0$ and the maximum is then equal to $\aleph = \frac{1+\sqrt{5}}{2}$. We have obtained that

$$\mathbb{P}\left(\exists j \geq 2 : H_j^{1/j} > (1+\varepsilon)\mathbb{E}\left[H_j^{1/j}\right] + \sqrt{2ujv_j} + \kappa(\varepsilon)b_j ju\right) \leq \aleph e^{-u} \ .$$

In order to bound the b_j, we use Hölder's inequality to get

$$b_j \leq \sup_{x\in\Xi,\ell=1,\ldots,n-1} \sum_{i=\ell+1}^{n} \mathbb{E}\left[|f_{i,\ell}(X_i,x)|^j\right]^{1/j} \leq \left(B^2A^{j-2}\right)^{1/j} \ .$$

Let us set $\mathcal{A}_j = \{a \text{ s.t. } \sum_{i=1}^{n-1}\mathbb{E}\left[a_i(X_i)^{j/(j-1)}\right] \leq 1\}$ and $\mathcal{B} = \{b \text{ s.t. } \sum_{\ell=2}^{n}\mathbb{E}\left[b_\ell(X_\ell)^2\right] \leq 1\}$. Then

$$\begin{aligned}
\sqrt{v_j} &= \sup_{a\in\mathcal{A}_j,b\in\mathcal{B}} \sum_{1\leq\ell<i\leq n} \mathbb{E}\left[f_{i,\ell}(X_i,X_\ell)a_i(X_i)b_\ell(X_\ell)\right] \\
&= \sup_{a\in\mathcal{A}_j,b\in\mathcal{B}} \sum_{i=2}^{n} \mathbb{E}\left[\sum_{\ell=1}^{i-1} \mathbb{E}\left[f_{i,\ell}(X_i,X_\ell)b_\ell(X_\ell) \mid X_i\right] a_i(X_i)\right] \\
&= \sup_{b\in\mathcal{B}} \left(\sum_{i=2}^{n} \mathbb{E}\left[\left(\sum_{\ell=1}^{i-1} \mathbb{E}\left[f_{i,\ell}(X_i,X_\ell)b_\ell(X_\ell) \mid X_i\right]\right)^j\right]\right)^{1/j} \\
&\leq (B^{j-2}D^2)^{1/j} \ .
\end{aligned}$$

By convexity of $x \mapsto x^j$ and using that

$$a+b = \frac{1}{1+\varepsilon}(1+\varepsilon)a + \frac{\varepsilon}{1+\varepsilon}(1+\varepsilon^{-1})b \ ,$$

it follows that, with probability larger than $1-\aleph e^{-u}$, for any $j \geq 2$,

$$\begin{aligned}
H_j &\leq \left((1+\varepsilon)\mathbb{E}\left[H_j^{1/j}\right] + (B^{j-2}D^2)^{1/j}\sqrt{2ju} + \kappa(\varepsilon)\left(B^2A^{j-2}\right)^{1/j} ju \right)^j \\
&\leq (1+\varepsilon)^{2j-1}\mathbb{E}\left[H_j^{1/j}\right]^j \\
&\qquad + (1+\varepsilon^{-1})^{j-1}\left((B^{j-2}D^2)^{1/j}\sqrt{2ju} + \kappa(\varepsilon)\left(B^2A^{j-2}\right)^{1/j} ju \right)^j \\
&\leq (1+\varepsilon)^{2j-1}\mathbb{E}\left[H_j\right] \\
&\qquad + [2(1+\varepsilon^{-1})]^{j-1}\left(B^{j-2}D^2\left(\sqrt{2ju}\right)^j + \kappa(\varepsilon)^j B^2 A^{j-2}\,(ju)^j \right) \ .
\end{aligned}$$

Denote by $\tau = \inf\left\{ p \in \mathbb{N} \ : \ \exists j \text{ s.t. } H_j^{(p)} > w_j^{(n)} \right\}$, where

$$H_j^{(p)} = \sum_{i=2}^{p} \mathbb{E}\left[\left| \sum_{\ell=1}^{i-1} f_{i,\ell}(X_i, X_\ell) \right|^j \ \middle| \ \mathcal{F}_{i-1} \right] \ ,$$

$$\begin{aligned}
w_j^{(n)} &= (1+\varepsilon)^{2j-1}\mathbb{E}\left[H_j^{(n)}\right] \\
&\qquad + [2(1+\varepsilon^{-1})]^{j-1}\left(B^{j-2}D^2\left(\sqrt{2ju}\right)^j + \kappa(\varepsilon)^j B^2 A^{j-2}\,(ju)^j \right) \ .
\end{aligned}$$

It is straightforward that τ is a stopping time and the previous computations show $\mathbb{P}(\tau > n) \leq \aleph e^{-u}$. The stopped martingale U_n^τ has brackets Δ_j upper bounded by $w_j^{(n)}$ by definition, therefore by (Pinelis, 1994, Theorem 8.5),

$$\mathbb{E}\left[e^{\lambda U_n^\tau}\right] \leq \exp\left(\sum_{j\geq 2} \frac{\lambda^j}{j!} w_j^{(n)} \right) \ .$$

Now,

$$\begin{aligned}
\sum_{j\geq 2} \frac{\lambda^j}{j!} w_j^{(n)} &= \sum_{j\geq 2} \frac{\lambda^j}{j!}(1+\varepsilon)^{2j-1}\mathbb{E}\left[H_j^{(n)}\right] && (2.36)\\
&\quad + \sum_{j\geq 2} \frac{\lambda^j}{j!}[2(1+\varepsilon^{-1})]^{j-1} B^{j-2}D^2\left(\sqrt{2ju}\right)^j && (2.37)\\
&\quad + \sum_{j\geq 2} \frac{\lambda^j}{j!}[2(1+\varepsilon^{-1})]^{j-1}\kappa(\varepsilon)^j B^2 A^{j-2}\,(ju)^j \ . && (2.38)
\end{aligned}$$

Now, the term (2.38) is upper bounded using that, for any $j \geq 0$, $j! \geq (j/e)^j$:

$$\sum_{j\geq 2} \frac{\lambda^j}{j!}[2(1+\varepsilon^{-1})]^{j-1}\kappa(\varepsilon)^j B^2 A^{j-2}\,(ju)^j \leq \frac{2\delta(\varepsilon)^2 B^2 u^2 \lambda^2}{1-2\delta(\varepsilon)Au\lambda} \ ,$$

for any $\lambda < 1/(2\delta(\varepsilon)Au)$, where $\delta(\varepsilon) = e\kappa(\varepsilon)(1+\varepsilon^{-1})$. The term (2.37) is upper bounded using that, for any $j \geq 1$, $j! \geq j^{j/2}$:

$$\sum_{j\geq 2} \frac{\lambda^j}{j!}[2(1+\varepsilon^{-1})]^{j-1} B^{j-2} D^2 \left(\sqrt{2ju}\right)^j \leq \frac{\eta(\varepsilon) D^2 u \lambda^2}{1-\eta(\varepsilon)B\sqrt{u}\lambda} \ ,$$

for any $\lambda < 1/(\eta(\varepsilon)B\sqrt{u})$, where $\eta(\varepsilon) = 4(1+\varepsilon^{-1})$. Now remark that the right-hand side term in (2.36) is equal to

$$\sum_{j\geq 2} \frac{[\lambda(1+\varepsilon)^2]^j}{j!(1+\varepsilon)} \mathbb{E}\left[H_j^{(n)}\right] = \frac{1}{1+\varepsilon}\sum_{i=2}^n \mathbb{E}\left[\mathbb{E}\left[e^{\mu|C_i|}\,\middle|\,\mathcal{F}_{i-1}\right] - 1 - \mu\mathbb{E}\left[|C_i|\,|\,\mathcal{F}_{i-1}\right]\right] \ ,$$

where $\mu = \lambda(1+\varepsilon)^2$, $C_i = \sum_{\ell=1}^{i-1} f_{i,\ell}(X_i, X_\ell)$. By convexity, for any $\theta \in \mathbb{R}$, $e^\theta - \theta - 1 \geq 0$, hence,

$$\sum_{j\geq 2} \frac{[\lambda(1+\varepsilon)^2]^j}{j!(1+\varepsilon)} \mathbb{E}\left[H_j^{(n)}\right] = \frac{1}{1+\varepsilon}\sum_{i=2}^n \mathbb{E}\left[\mathbb{E}\left[e^{\mu C_i}\,\middle|\,X_i\right] + \mathbb{E}\left[e^{-\mu C_i}\,\middle|\,X_i\right] - 2\right] \ .$$

Given X_i, C_i and $-C_i$ are sums of centered i.i.d. random variables, bounded by A, therefore, by Bernstein's inequality,

$$\mathbb{E}\left[e^{\pm\mu C_i}\,\middle|\,X_i\right] \leq \mathbb{E}\left[e^{\frac{\mu^2 v_i(X_i)}{2-2\mu A/3}}\,\middle|\,X_i\right] \ ,$$

with $v_i(X_i) = \sum_{\ell=1}^{i-1}\mathbb{E}\left[f_{i,\ell}(X_i,X_\ell)^2\,\middle|\,X_i\right] \leq B^2$ and $\sum_{i=2}^n \mathbb{E}\left[v_i(X_i)\right] \leq C^2/2$. Hence,

$$\forall k \geq 1, \qquad \sum_{i=2}^n \mathbb{E}\left[v_i(X_i)^k\right] \leq \frac{C^2}{2}B^{2(k-1)} \ ,$$

and

$$\begin{aligned}
\sum_{j\geq 2} \frac{[\lambda(1+\varepsilon)^2]^j}{j!(1+\varepsilon)} \mathbb{E}\left[H_j^{(n)}\right] &\leqslant \frac{2}{1+\varepsilon}\sum_{i=2}^n \mathbb{E}\left[\sum_{k\geq 1}\frac{1}{k!}\left(\frac{\mu^2 v_i(X_i)}{2-2\mu A/3}\right)^k\right] \\
&\leq \frac{C^2}{(1+\varepsilon)B^2}\sum_{k\geq 1}\frac{1}{k!}\left(\frac{\mu^2 B^2}{2-2\mu A/3}\right)^k \\
&\leq \frac{C^2(1+\varepsilon)^3\lambda^2/2}{1-\lambda(1+\varepsilon)^2 A/3 - \lambda^2(1+\varepsilon)^4 B^2/4} \\
&\leq \frac{C^2(1+\varepsilon)^3\lambda^2/2}{1-\lambda(1+\varepsilon)^2(A/3+B/2)} \ ,
\end{aligned}$$

for any $\lambda < 1/[(1+\varepsilon)^2(A/3+B/2)]$. Therefore,

$$\sum_{j\geq 2}\frac{\lambda^j}{j!}w_j^{(n)} \leq \frac{C^2(1+\varepsilon)^3\lambda^2/2}{1-\lambda(1+\varepsilon)^2(A/3+B/2)} + \frac{\eta(\varepsilon)D^2u\lambda^2}{1-\eta(\varepsilon)B\sqrt{u}\lambda} + \frac{2\delta(\varepsilon)^2B^2u^2\lambda^2}{1-2\delta(\varepsilon)Au\lambda} \ .$$

Let $W = C(1+\varepsilon)^{3/2}/\sqrt{2} + \sqrt{\eta(\varepsilon)}D\sqrt{u} + \sqrt{2}\delta(\varepsilon)Bu$ and

$$c = \max\left((1+\varepsilon)^2(A/3+B/2), \eta(\varepsilon)B\sqrt{u}, 2\delta(\varepsilon)Au\right) \ ,$$

the last inequality implies that

$$\sum_{j\geq 2}\frac{\lambda^j}{j!}w_j^{(n)} \leq \frac{\lambda^2W^2}{1-\lambda c} \ .$$

Hence,

$$\mathbb{E}\left[e^{\lambda U_n^\tau}\right] \leq \exp\left(\frac{\lambda^2W^2}{1-\lambda c}\right) \ .$$

Therefore

$$\mathbb{P}\left(U_n^\tau > 2W\sqrt{u}+cu\right) \leq e^{-u} \ .$$

Finally,

$$\mathbb{P}\left(U_n > 2W\sqrt{u}+cu\right) \leq \mathbb{P}\left(U_n^\tau > 2W\sqrt{u}+cu\right) + \mathbb{P}\left(\tau > n\right) \leq (\aleph+1)e^{-u} \leq 2.7e^{-u} \ .$$

To conclude, let us say that the bound only makes sense when $u \geq \log(C+1)$ and, when $u \geq \log(\aleph+1)$, $u/\sqrt{2} \leq u^{3/2}$ and $u/3 \leq u^2$, hence

$$\begin{aligned} 2W\sqrt{u}+cu &\leq \sqrt{2}(1+\varepsilon)^{3/2}C\sqrt{u} + 2\sqrt{\eta(\varepsilon)}Du \\ &\quad + (2\sqrt{2}\delta(\varepsilon) + (1+\varepsilon)^2 + \eta(\varepsilon))Bu^{3/2} + [(1+\varepsilon)^2 + 2\delta(\varepsilon)]Au^2 \\ &\leq \sqrt{2}(1+\varepsilon)^{3/2}C\sqrt{u} + 4(1+\varepsilon^{-1})^{1/2}Du + [(1+\varepsilon)^2 + 12(1+\varepsilon^{-1})^2](Bu^{3/2}+Au^2) \ . \end{aligned}$$

2.9.2 Technical Lemmas

Lemma 2.1. *Let Y_n be real valued bounded martingale with respect to a filtration $\mathcal{F}_n$, then, for any $\lambda > 0$,*

$$\mathcal{A}_n = \exp\left(\lambda Y_n - \sum_{k\geq 2}\frac{\lambda^k}{k!}\Delta_{n,k}\right), \qquad \textit{where} \qquad \Delta_{n,k} = \sum_{i=1}^n \mathbb{E}\left[\left(Y_i - Y_{i-1}\right)^k \,\middle|\, \mathcal{F}_{i-1}\right]$$

is a supermartingale.

Proof: We have

$$\mathbb{E}\left[\mathcal{A}_n \mid \mathcal{F}_{n-1}\right] = \mathbb{E}\left[e^{\lambda(Y_n - Y_{n-1})} \,\middle|\, \mathcal{F}_{n-1}\right] e^{\lambda Y_{n-1} - \sum_{k\geq 2}\frac{\lambda^k}{k!}\Delta_{n,k}} \ .$$

Moreover,

$$\mathbb{E}\left[e^{\lambda(Y_n-Y_{n-1})}\,\middle|\,\mathcal{F}_{n-1}\right]=1+\sum_{k\geq 2}\frac{\lambda^k}{k!}\mathbb{E}\left[(Y_n-Y_{n-1})^k\,\middle|\,\mathcal{F}_{n-1}\right]\leq e^{\sum_{k\geq 2}\frac{\lambda^k}{k!}(\Delta_{n,k}-\Delta_{n-1,k})}\ .$$

□

Lemma 2.2. *Let $X_1,\ldots,X_n$ be independent variables with values in some polish space Ξ and let $\mathcal{F}$ be a countable class of measurable functions from Ξ into $[-b,b]^n$ for some $b>0$. For $f=(f_1,\ldots,f_n)$ in $\mathcal{F}$, let*

$$S_n(f)=f_1(X_1)+\ldots+f_n(X_n),\quad Z=\sup_{f\in\mathcal{F}}|S_n(f)|\quad \textit{and}\quad v=\sup_{f\in\mathcal{F}}\mathrm{Var}\left[S_n(f)\right]\ .$$

Then for all $x,\varepsilon>0$,

$$\mathbb{P}\left(Z\geqslant(1+\varepsilon)\mathbb{E}\left[Z\right]+\sqrt{2vx}+(1/3+\varepsilon^{-1})bx\right)\leqslant e^{-x}\ .$$

Proof: Without loss of generality and thanks to renormalization, one can assume that $b=1$. We apply Theorem 1 of Rio (2012) to Z reinterpreted with the independent family $Y_1,\ldots,Y_n$ where $Y_i=(Y_i^1,\ldots,Y_i^n)=(0,\ldots,0,X_i,0,\ldots,0)$, i.e. its only nonzero coordinate is its ith coordinate. We consider the family

$$\mathcal{G}=\left\{g\ \text{s.t.}\ g(Y)=f_1(Y^1)+\ldots+f_n(Y^n)\ \text{for}\ f=(f_1,\ldots,f_n)\in\mathcal{F}\right\}\ .$$

Because of the particular vectors Y_i with only one nonzero coordinate, on which the g's act, they actually take values in $[-1,1]$. Therefore denoting $E=\mathbb{E}\left[Z\right]/n$ and $V=v+2E-E^2$, one has that for all $x>0$

$$\mathbb{P}\left(Z-\mathbb{E}\left[Z\right]\geqslant nx\right)\leqslant\exp(-nV(1-E)^{-2}h(x(1-E)/V))\ ,$$

with $h(u)=(1+u)\log(1+u)-u$. But by classical results on Bennett's inequality, it is well known that (Massart, 2007)

$$h(u)\geqslant\frac{u^2}{2+2u/3}\ .$$

Inverting the resulting formula, one obtains for all $z>0$

$$\mathbb{P}\left(Z\geqslant\mathbb{E}\left[Z\right]+\sqrt{2vz+2nE(2-E)z}+\frac{z(1-E)}{3}\right)\leqslant e^{-z}\ .$$

This leads for all $\varepsilon,z>0$ to

$$\mathbb{P}\left(Z\geqslant(1+\varepsilon)\mathbb{E}\left[Z\right]+\sqrt{2vz}+z\left(\frac{(1-E)}{3}+\frac{(2-E)}{2\varepsilon}\right)\right)\leqslant e^{-z}\ .$$

Since $E\geqslant 0$ this leads to the desired inequality. □

Part II

V-fold and risk estimation

Chapter 3

Kernel selection via V-fold penalization

Abstract. This chapter studies V-fold cross-validation, V-fold penalization and leave-p-out procedures for selecting a linear estimator in least-squares density estimation. Thanks to concentration inequalities, we first prove nonasymptotic oracle inequalities for these procedures. In particular, this result implies V-fold penalization is asymptotically optimal for bandwidth selection. Then, we derive from our results some adaptive estimators on Sobolev classes and sets of functions with nonincreasing weights.

NOTA: Ce chapitre est une version légèrement modifiée d'un travail en collaboration avec Sylvain Arlot[(1)] et Matthieu Lerasle[(2)].

Contents

[(1)]École Normale Supérieure, Paris
[(2)]Université Nice Sophia Antipolis

3.1 Introduction

Having at hand a random sample from a distribution with some density function, one is often interested in estimating it. The most commonly used nonparametric estimator to solve this problem is the kernel density estimator (k.d.e.) which was introduced fifty years ago by Rosenblatt (1956) and Parzen (1962). Since then, its properties (stability of the kernel estimate, consistency, rates of convergence,..) were studied and commented in depth by many statisticians considering different frameworks and different losses (mainly the $\mathbb{L}_1$, $\mathbb{L}_2$ and Kullback loss).

The major issue when one handles with a family of k.d.e. is to select the bandwidth in an optimal way from the data. This problem was attacked by many authors from both theoretical (see among other references the book of Devroye & Lugosi (2001) for the $\mathbb{L}_1$-loss, and Wegkamp (1999); Tsybakov (2009) for the $\mathbb{L}_2$-loss) and practical point of view (we refer to Silverman (1986); Wand & Jones (1995); Jones *et al.* (1996); Chiu (1996) for an overview of practical data-driven methods to select the bandwidth). Dealing with the $\mathbb{L}_2$-loss, the two most widely used procedures are least squares cross-validation (LSCV), proposed independently by Rudemo (1982) and Bowman (1984), and the plug-in method (Sheather & Jones, 1991). Even if plug-in procedures seem more popular from a practical point of view, CV requires fewer assumptions and works well when the density is difficult to estimate (Loader, 1999).

The quality of CV to select a bandwidth quickly became an important subject. The consistency of the minimizer of the CV criterion as an estimate of the optimal bandwidth was first proved (Hall, 1983; Stone, 1984), followed by the asymptotic normality of the bandwidth estimate (Hall & Marron, 1987) and some rates of convergence (Scott & Terrell, 1987). Despite the large number of papers, all these techniques rely on asymptotic considerations and up to our knowledge no nonasymptotic results (such as oracle inequalities) were ever proved for these procedures. Several modifications of LSCV have been proposed in an attempt to improve its performance. Among others, these include the biased cross-validation method (Scott & Terrell, 1987), the corrected V-fold cross-validation (Burman, 1989), the trimmed cross-validation (Feluch & Koronacki, 1992), the modified cross-validation (Stute, 1992), the indirect cross-validation (Savchuk *et al.*, 2010), etc.

The development of k.d.e. have rapidly led to many extensions of their definition. A classical generalization is to consider a collection of linear estimator (sometimes called "delta-sequence estimator" (Walter & Blum, 1979) or "additive estimator" (Devroye & Lugosi, 2001)) which include projection and weighted projection estimators leading in particular to histogram and Pinsker estimators. The problem is then to select an hyperparameter (other than the bandwidth one can cite the choice of the partition, the weight function, etc.) in an optimal way from the data. The universality of CV permits one to deal with this problem (see Rudemo (1982) and Marron (1987)) whereas plug-in methods fail in such a general treatment.

The main goal of this work is to provide nonasymptotic analysis and first-order optimal results for different widely used data-driven procedures: the V-fold cross-validation (VFCV), the corrected VFCV, the penalized V-fold (penVF) and the leave-p-out (LPO) procedure. In particular, this extends the results obtained by Arlot & Lerasle (2014) for projection estimators. To do so, we take advantage of V-fold penalties (Arlot, 2008; Arlot & Lerasle, 2014) and a concentration

inequality for U-statistics of order two (Lerasle *et al.*, 2015). It is indeed possible to prove that V-fold penalization is related to these procedures through some U-statistics.

In addition, we show that CV can be a powerful tool to get an estimator which is adaptive to the unknown regularity of the density. Indeed, we deduce from our asymptotically optimal oracle inequalities some sharp adaptive results on Sobolev classes and functions with nonincreasing weights. Up to our knowledge, apart from Dalelane (2005a,b), such results have never been reached using CV procedures.

We finally provide simulation experiments on two different families of estimators (Parzen estimators, to deal with the bandwidth selection problem, and a mix of these together with histograms) to illustrate our oracle inequalities and to study the behavior of V-fold penalization with respect to V.

Notations

For any finite set A, $|A|$ denote its cardinal. For any integer n, let $[n] = \{1, \dots, n\}$ and for any $a < b \in \mathbb{N}^*$, let $[\![a, b]\!] = \{a, \dots, b\}$. For any sequence $(x_i)_{i\in\mathcal{I}}$ and any subset $A \subset \mathcal{I}$, x_A denotes $(x_i)_{i\in A}$. For any real numbers x and y, $x \vee y = \max(x, y)$, $x \wedge y = \min(x, y)$, $x_+ = x \vee 0$ and $x_- = (-x) \vee 0$.

3.2 Kernel selection for density estimation

3.2.1 Setup

Let X be a random variable taking values in a Polish space Ξ endowed with its Borel σ-algebra $\mathcal{Z}$ with distribution P. We assume that P has density s with respect to some known measure μ on $(\Xi, \mathcal{Z})$, and aim at estimating s from a sample $\mathbf{X} = (X_1, \dots, X_n)$ of independent copies of X. For any $p \in \mathbb{R}_+$, let $\mathbb{L}_p$ denote the space of measurable functions $f : \Xi \to \mathbb{R}$ such that $\|f\|_p^p = \int_\Xi |f|^p \, d\mu < \infty$. Let $\|\cdot\| = \|\cdot\|_2$ and $\langle \cdot, \cdot \rangle$ denote respectively the $\mathbb{L}_2$-norm and the associated inner product. Let $\mathbb{L}_\infty$ denote the space of bounded functions and let $\|\cdot\|_\infty$ denote the associated sup-norm. All along the paper, the unknown density $s \in \mathbb{L}_\infty$, which implies that $s \in \mathbb{L}_2$. The quadratic loss of any estimator t is equal to $\|s - t\|^2$. Let $(\mathcal{K}_m)_{m\in\mathcal{M}}$ denote a finite collection of bounded functions $\mathcal{K}_m : \Xi^2 \to \mathbb{R}$ such that $\mathcal{K}_m(x, y) = \mathcal{K}_m(y, x)$ and, such that for some constant $\Gamma \geq 1$

$$\forall m \in \mathcal{M}, \qquad \sup_{x\in\Xi} \int_\Xi \mathcal{K}_m(x, y)^2 d\mu(y) \ \vee \sup_{(x,y)\in\Xi^2} |\mathcal{K}_m(x, y)| \leqslant \Gamma n \ . \tag{3.1}$$

We also define $\chi_m(x) := \mathcal{K}_m(x, x)$ for all $x \in \Xi$. Hereafter $\mathcal{K}_m$ is called a *kernel*. To any $\mathcal{K}_m$, we associate the linear estimator $\widehat{s}_m$ and the kernel regularization function s_m, respectively defined for any $x \in \Xi$ by

$$\widehat{s}_m(x) := \frac{1}{n} \sum_{i=1}^n \mathcal{K}_m(x, X_i) \ , \tag{3.2}$$

$$s_m(x) := \int_\Xi \mathcal{K}_m(x,y)s(y)d\mu(y) = \mathbb{E}\left[\mathcal{K}_m(x,X)\right] \ .$$

3.2.2 Example of kernels

Let us present three examples of estimators that have been widely studied in the literature on density estimation. Each one can be naturally associated to a kernel.

Example 1 (Projection estimators)**.** For any $m \in \mathcal{M}$, let S_m denote a linear subspace of $\mathbb{L}_2$. The projection estimator on S_m is defined as

$$\widehat{s}_m = \operatorname*{argmin}_{t \in S_m} \left\{ \|t\|^2 - \frac{2}{n}\sum_{i=1}^n t(X_i) \right\} \ .$$

Classical algebra shows that, given any orthonormal basis $(\psi_\lambda)_{\lambda\in\Lambda_m}$ of S_m, $\widehat{s}_m$ is equal to

$$\forall x \in \Xi, \qquad \widehat{s}_m(x) = \sum_{\lambda\in\Lambda_m} \left(\frac{1}{n}\sum_{i=1}^n \psi_\lambda(X_i) \right) \psi_\lambda(x) \ .$$

Projection estimators can therefore be associated to the *projection kernel* $\mathcal{K}_m$ defined for the basis $(\psi_\lambda)_{\lambda\in\Lambda_m}$ by

$$\forall (x,y) \in \Xi^2, \qquad \mathcal{K}_m(x,y) = \sum_{\lambda\in\Lambda_m} \psi_\lambda(x)\psi_\lambda(y) \ .$$

The kernel $\mathcal{K}_m$ actually depends on the orthonormal basis $(\psi_\lambda)_{\lambda\in\Lambda_m}$ even if $\widehat{s}_m$ does not. This is why, in the sequel, an orthonormal basis is always assumed to be given with S_m. Notice that the kernel regularization

$$s_m = \sum_{\lambda\in\Lambda_m} \mathbb{E}\left[\psi_\lambda(X)\right]\psi_\lambda = \sum_{\lambda\in\Lambda_m} \langle s, \psi_\lambda\rangle \psi_\lambda \ ,$$

is the orthogonal projection of s onto S_m. Hence, s_m does not depend on $(\psi_\lambda)_{\lambda\in\Lambda_m}$ either. For projection kernels, we can choose, see Lerasle *et al.* (2015)

$$\Gamma = 1 \vee \frac{1}{n} \sup_{m\in\mathcal{M}} \sup_{x\in\Xi} |\chi_m(x)| \ .$$

Example 2 (Weighted projection estimators)**.** Projection estimators have been generalized in the following way. Let $p \in \mathbb{N}^* \cup \{\infty\}$, let $m = m_{[p]} \in [0,1]^p$ with $\sum_{i=1}^p m_i > 0$, let $(\varphi_j)_{j\in[p]}$ be an orthonormal system in $\mathbb{L}_2$. The weighted projection estimator is defined by

$$\widehat{s}_m = \sum_{i=1}^p m_i \left(\frac{1}{n}\sum_{j=1}^n \varphi_i(X_j) \right) \varphi_i \ .$$

It is the linear estimator associated to the weighted projection kernel $\mathcal{K}_m$ defined, for any $(x,y) \in \Xi^2$, by

$$\mathcal{K}_m(x,y) = \sum_{i=1}^{p} m_i \varphi_i(x)\varphi_i(y) \ .$$

For these kernels, one can choose $\Gamma = 1 \vee n^{-1} \left\| \sum_{i=1}^{p} \varphi_i^2 \right\|_\infty$, see Lerasle *et al.* (2015). It is clear that projection estimators correspond to the particular choices $m_i \in \{0,1\}$ for any $i = 1,\dots,p$. Another popular example of weighted estimators is given by Pinsker's estimators. Pinsker (1980) worked on Gaussian sequences but Efroimovich (Efroĭmovich, 1985; Efromovich, 2000, 2005) adapted Pinsker's construction to density estimation and derive estimators adaptive to Sobolev classes up to the constant. They are defined for the Fourier basis $(\varphi_k)_{k\geq 0}$ where $\varphi_0 = \mathbf{1}_{[0,1]}$ and, for any $k \geq 1$,

$$\forall x \in [0,1], \qquad \varphi_{2k-1}(x) = \sqrt{2}\sin(2k\pi x), \qquad \varphi_{2k}(x) = \sqrt{2}\cos(2k\pi x) \ . \tag{3.3}$$

Pinsker's weights $m^{\beta,Q} = (m_k^{\beta,Q})_{k\geq 0}$ are defined for fixed parameters $\beta > 0$ and $Q > 0$ by $m_0^{\beta,Q} = 1$ and for any $k \geq 1$, $m_{2k-1}^{\beta,Q} = m_{2k}^{\beta,Q} = \tau_k^{\beta,Q}$, where

$$\tau_k^{\beta,Q} = \left(1 - r^{\beta,Q}\frac{k^\beta}{n^{\frac{\beta}{2\beta+1}}}\right)_+ , \qquad r^{\beta,Q} = \left(\frac{\beta}{(\beta+1)(2\beta+1)Q}\right)^{\frac{\beta}{2\beta+1}} \ , \tag{3.4}$$

and for any $u \in \mathbb{R}$, $u_+ = u \vee 0$.

Example 3 (Parzen's estimators)**.** Let $\Xi \subset \mathbb{R}^d$ and let μ denote the Lebesgue measure on Ξ. Let $m \in \mathcal{M}$ denote a collection of couples $m = (k,h)$ where k is a bounded function in $\mathbb{L}_1$, such that $k(0) > 0$ and $k(x) = k(-x)$, and $h = (h_1,\dots,h_d) \in (\mathbb{R}_+^*)^d$ is vector of regularization parameters. Parzen's estimators are defined for any $m = (k,h)$ by

$$\widehat{s}_m(x) = \frac{1}{n\prod_{i=1}^d h_i} \sum_{i=1}^{n} k\left(\frac{x - X_i}{h}\right) \ ,$$

where by definition $\frac{x-X_i}{h}$ is the vector of $\mathbb{R}^d$ with coordinates $(x_j - (X_i)_j)/h_j$. Parzen's estimators are associated to the approximation kernels

$$\mathcal{K}_m(x,y) = \frac{1}{\prod_{i=1}^d h_i} k\left(\frac{x-y}{h}\right) \ .$$

For these kernels, one can choose $\Gamma = 1$ if $h \geq (\|k\|_\infty \vee \|k\|^2)/n$ for all $m = (k,h) \in \mathcal{M}$, see Lerasle *et al.* (2015). As examples of functions k, let us mention the *Epanechnikov* k_E and the *Gaussian* k_G kernels respectively defined (when $d = 1$) for any $x \in \mathbb{R}$ by

$$k_E(x) = \frac{3}{4}\left(1 - x^2\right)_+ , \qquad k_G(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \ .$$

Remark. When $\beta \in \mathbb{N}^*$, Pinsker estimators can also be defined as Parzen's estimators. Actually, Dalelane (2005b) shows that, given β and Q, one can define Pinsker's kernel

$$k_\beta(x) := \frac{\beta!}{\pi} \sum_{j=1}^{\beta} \frac{\sin^{(j)}(x)}{(\beta - j)! x^{j+1}} \ ,$$

and Pinsker's estimator is the associated Parzen's estimator defined for the bandwidth

$$h_{\beta,Q} = n^{-\frac{1}{2\beta+1}} \left(\frac{2\beta}{Q^2(\beta+1)(2\beta+1)} \right) \ .$$

3.2.3 Estimator selection

Let $(\mathcal{K}_m)_{m\in\mathcal{M}}$ denote a collection of kernels, and $(\widehat{s}_m)_{m\in\mathcal{M}}$ be the corresponding collection of linear estimators. Our purpose is to select from data some $\widehat{m} = \widehat{m}(\mathbf{X})$ such that the risk of $\widetilde{s} = \widehat{s}_{\widehat{m}}$ is as small as possible.

More precisely, the goal is to prove $\widetilde{s}$ satisfies with a large probability an oracle inequality, that is, an inequality of the form

$$\|\widetilde{s} - s\|^2 \leqslant C_n \inf_{m\in\mathcal{M}} \|\widehat{s}_m - s\|^2 + R_n \, . \tag{3.5}$$

The constant C_n above is called the leading constant of the oracle inequality, and R_n is a remainder term, assumed to be negligible in front of the oracle loss $\inf_{m\in\mathcal{M}} \|\widehat{s}_m - s\|^2$. Clearly, $C_n \geq 1$ and the best one can expect is to get C_n as close to 1 as possible. This is why, when $C_n \to 1$ as n goes to infinity, the oracle inequality is called first-order optimal. As the goal is to find a minimizer of the $\mathbb{L}_2$-loss, an ideal way to proceed would be to minimize the ideal criterion

$$\mathcal{C}_{\mathrm{id}}(m) := \|\widehat{s}_m - s\|^2 - \|s\|^2 = \|\widehat{s}_m\|^2 - 2\mathbb{E}\left[\widehat{s}_m(X)|X_1, \ldots, X_n\right] \ .$$

As the measure P is unknown, a natural idea is to estimate $\mathcal{C}_{\mathrm{id}}(m)$ by a data-driven procedure $\mathcal{C}(m)$, for example using cross-validation, see Section 3.3, or *penalization*

$$\mathcal{C}(m) = \|\widehat{s}_m\|^2 - \frac{2}{n} \sum_{i=1}^{n} \widehat{s}_m(X_i) + \mathrm{pen}(m) \ .$$

Concerning the latter one, one notices that the ideal criterion can be rewritten as a penalized criterion

$$\mathcal{C}_{\mathrm{id}}(m) = \|\widehat{s}_m\|^2 - \frac{2}{n} \sum_{i=1}^{n} \widehat{s}_m(X_i) + \mathrm{pen}_{\mathrm{id}}(m) \ ,$$

where the ideal penalty is defined as

$$\mathrm{pen}_{\mathrm{id}}(m) := 2 \left(\frac{1}{n} \sum_{i=1}^{n} \widehat{s}_m(X_i) - \mathbb{E}\left[\widehat{s}_m(X)|X_1, \ldots, X_n\right] \right) \ .$$

In order to calibrate the penalty pen, the idea is then to estimate the ideal one.

3.3 Cross-validation and resampling methods

The purpose of this paper is to study existing model selection procedures of the form

$$\widehat{m}_{\mathcal{C}} \in \operatorname*{argmin}_{m \in \mathcal{M}} \{ \mathcal{C}(m) \} \quad ,$$

where $\mathcal{C}$ is some data-dependent criterion based on cross-validation or resampling ideas. The goal is to prove that these procedures lead to estimators that satisfy first-order optimal oracle inequality. This section describes more precisely the criteria $\mathcal{C}$ considered in the following.

3.3.1 Cross-validation

For any function $t : \Xi \to \mathbb{R}$, let

$$P_n t = \frac{1}{n} \sum_{i=1}^{n} t(X_i)$$

and for any nonempty subset $A \subset [n]$, let $A^c := [n] \backslash A$,

$$P_n^{(A)} t = \frac{1}{|A|} \sum_{i \in A} t(X_i) \quad .$$

For any $t \in \mathbb{L}_2(\mu)$, let

$$Pt = \mathbb{E}\left[t(X) \right] \quad .$$

Let also γ denote the *least-squares contrast* defined, for any function $t \in \mathbb{L}_2(\mu)$ and any $x \in \Xi$, by

$$\gamma(t, x) = \|t\|^2 - 2t(x) \quad .$$

The main idea of cross-validation is data splitting. Some $T \subset [n]$ is chosen to *train* the estimators $\widehat{s}_m^{(T)}(x) = \frac{1}{|T|} \sum_{i \in T} \mathcal{K}_m(X_i, x)$, the remaining data are used to estimate the ideal criterion and build the *hold-out* criterion

$$\mathcal{C}_T^{\mathrm{HO}}(m) := \frac{1}{|T^c|} \sum_{i \in T^c} \gamma\left(\widehat{s}_m^{(T)} \right) \quad . \tag{3.6}$$

These criteria depend on a particular choice for T that makes them unstable. To reduce this variability, a collection $\mathcal{E}$ of training sets is chosen to build a cross-validation criterion

$$\mathcal{C}_{\mathcal{E}}(m) = \frac{1}{|\mathcal{E}|} \sum_{T \in \mathcal{E}} \mathcal{C}_T^{\mathrm{HO}}(m) \quad .$$

We are interested in two particular collections of training sets. The first one is the collection $\mathcal{E}_p$ of all subsets of $[n]$ with cardinality $n - p$, for some $p \in [n-1]$. The associated cross-validation criterion is called *leave-p-out*

$$\mathcal{C}_p^{\mathrm{LPO}}(m) := \frac{1}{\binom{n}{p}} \sum_{T \in \mathcal{E}_p} \mathcal{C}_T^{\mathrm{HO}}(m) \quad . \tag{3.7}$$

To define the second one, let $V \in [n]$ and let $B_{[V]} = (B_1, \ldots, B_V)$ denote a regular partition of $[n]$, that is a partition satisfying the following assumption

$$V \text{ divides } n \quad \text{and} \quad \forall k \in [V]\,,\ |B_k| = \frac{n}{V} \ . \tag{\textbf{Reg}}$$

The second collection of training sets that we consider is the collection $\mathcal{E}_{B_{[V]}} = \{ B_1^c, \ldots, B_V^c \}$. The associated cross-validation criterion is the V-fold cross-validation (VFCV) criterion of Breiman *et al.* (1984). In our setting, it is equal to

$$\mathcal{C}_V^{\mathrm{VFCV}}(m) := \frac{1}{V} \sum_{k=1}^{V} P_n^{(B_k)} \gamma \left(\widehat{s}_m^{(B_k^c)} \right) = \frac{1}{V} \sum_{k=1}^{V} \left\| \widehat{s}_m^{(B_k^c)} \right\|^2 - \frac{2}{V} \sum_{k=1}^{V} P_n^{(B_k)} \widehat{s}_m^{(B_k^c)} \ .$$

Since the VFCV criterion is known to be a biased estimator of $\mathcal{C}_{\mathrm{id}}$, a bias-corrected VFCV criterion was introduced by Burman (1989) in the regression setting. In our framework, it can be written as

$$\mathcal{C}_V^{\mathrm{corr,VFCV}}(m) := \mathcal{C}_V^{\mathrm{VFCV}}(m) + P_n \gamma \left(\widehat{s}_m \right) - \frac{1}{V} \sum_{k=1}^{V} P_n \gamma \left(\widehat{s}_m^{(B_k^c)} \right) \ . \tag{3.8}$$

3.3.2 Resampling penalties

A general alternative to cross-validation is resampling penalization (Efron, 1983; Arlot, 2009). Its principle is to select $\widehat{m}$ by minimizing a penalized criterion

$$\mathcal{C}(m) = P_n \gamma \left(\widehat{s}_m \right) + \mathrm{pen}(m)$$

where $\mathrm{pen}(m)$ estimates the expectation of the ideal penalty $\mathrm{pen}_{\mathrm{id}}(m)$ thanks to Efron's resampling heuristic. More precisely, let $W_{[n]} = (W_1, \ldots, W_n)$ be a collection of nonnegative random variables independent of $\mathbf{X}$ such that $\sum_{i=1}^n W_i = n$ and all the marginals W_i have the same distribution. The distribution of $W_{[n]}$ is called a resampling scheme and the corresponding estimator is denoted $\widehat{s}_m^W = n^{-1} \sum_{i=1}^n W_i \mathcal{K}_m(X_i, x)$. Let $P_n^W = n^{-1} \sum_{i=1}^n W_i \delta_{X_i}$ denote the resampling empirical distribution, the associated resampling penalty is defined by

$$\mathrm{pen}_W(m) = \frac{2}{\mathbb{E}\left[(W_1 - 1)^2 \right]} \mathbb{E}_W \left[\left(P_n^W - P_n \right) \widehat{s}_m^W \right] \ , \tag{3.9}$$

where $\mathbb{E}_W [\cdot]$ denotes the expectation with respect to the resampling randomness. The corresponding penalized criterion is written, for any constant $C > 0$,

$$\mathcal{C}_{C,W}(m) := P_n \gamma \left(\widehat{s}_m \right) + C \, \mathrm{pen}_W(m) \ .$$

If we note for all $i, j \in [n]$

$$\rho_{i,j}^W := \frac{\mathbb{E}\left[(W_i - 1)(W_j - 1) \right]}{\mathbb{E}\left[(W_1 - 1)^2 \right]} \ ,$$

elementary computations give

$$\mathrm{pen}_W(m) = \frac{2}{n^2}\left(\sum_{i=1}^{n}\chi_m(X_i) + \sum_{1\leqslant i\neq j\leqslant n}\rho_{i,j}^W \mathcal{K}_m(X_i,X_j)\right) \ . \tag{3.10}$$

A classical assumption on the weights $W_{[n]}$ is their exchangeability, i.e. the fact that the distribution of $W_{[n]}$ is the same as the one of $(W_{\pi(1)},\dots,W_{\pi(n)})$ for any permutation π. Under this extra assumption, it comes from $\mathbb{E}\left[(\sum_{i=1}^n (W_i-1))^2\right] = 0$ that $\rho_{i,j}^W = -1/(n-1)$ for any $i \neq j$, which implies

$$\mathrm{pen}_W(m) = \frac{2}{n^2}\left(\sum_{i=1}^{n}\chi_m(X_i) - \frac{1}{n-1}\sum_{1\leqslant i\neq j\leqslant n}\mathcal{K}_m(X_i,X_j)\right) \ . \tag{3.11}$$

In particular, all resampling penalties built with an exchangeable resampling scheme are equal, and they can be computed efficiently according to (3.11). Famous exchangeable weights include *leave-p-out* weights defined by $W_i^{\mathrm{LPO}} = \frac{n}{n-p}\mathbf{1}_{i\in B}$ where p is an integer $1 \leq p \leq n-1$ and B is a random set uniformly chosen among $\mathcal{E}_p$, the collection subsets of $[n]$ with cardinality $n-p$. The associated penalty is sometimes called *leave-p-out* penalty $\mathrm{pen}_p^{\mathrm{LPO}}(m)$ (Celisse, 2014).

3.3.3 V-fold penalization

Resampling penalties can also be defined with non-exchangeable weights $W_{[n]}$. For instance, the V-fold penalty (Arlot, 2008; Arlot & Lerasle, 2014) has been introduced as a resampling penalty using some V-fold subsampling weights, and it was shown to include Burman's criterion as a particular case (Arlot, 2008) (for $F = 1$). More precisely, let $B_{[V]}$ be a regular partition of $[n]$ into $V \in [n]$ subsets, satisfying (**Reg**), and let I denote a random variable independent of $\mathbf{X}$ uniformly distributed over $[V]$. Then, the V-fold penalty associated with $B_{[V]}$ is defined by (3.9) where for any $i \in [n]$, $W_i^{(\mathrm{VF})} = \frac{n}{|B_I^c|}\mathbf{1}_{i\in B_I^c} = \frac{V}{V-1}\mathbf{1}_{i\in B_I^c}$. These weights are identically distributed, they satisfy $\sum_{i=1}^n W_i^{(\mathrm{VF})} = n$, $\mathbb{E}\left[(W_1^{(\mathrm{VF})}-1)^2\right] = 1/(V-1)$ and if $i \in B_k$ and $j \in B_{k'}$, $\rho_{i,j}^{(\mathrm{VF})} := \rho_{i,j}^{W^{(\mathrm{VF})}} = 1 - (V/(V-1))\mathbf{1}_{k\neq k'}$. Hence, it comes from (3.10)

$$\mathrm{pen}_{\mathrm{VF}}(m,V) := \mathrm{pen}_{W^{(\mathrm{VF})}}(m) = \frac{2}{n^2}\left(\sum_{i=1}^{n}\chi_m(X_i) + \sum_{1\leqslant i\neq j\leqslant n}\rho_{i,j}^{(\mathrm{VF})}\mathcal{K}_m(X_i,X_j)\right) \ . \tag{3.12}$$

Notice that $W_{[n]}^{(\mathrm{VF})}$ are exchangeable if and only if $V = n$. In particular, as already discussed, any resampling penalty defined with an exchangeable resampling scheme is equal to the n-fold penalty. The V-fold penalized criteria are defined for any $F > 0$ by

$$\mathcal{C}_{V,F}^{\mathrm{pen}}(m) := P_n\gamma\left(\widehat{s}_m\right) + F\,\mathrm{pen}_{\mathrm{VF}}(m,V) \ . \tag{3.13}$$

In this paper, we shall study the quality of the estimator $\widetilde{s}$ of s defined by $\widetilde{s} = \widehat{s}_{\widehat{m}^{\rm pen}_{V,F}}$, where

$$\widehat{m}^{\rm pen}_{V,F} \in \operatorname*{argmin}_{m\in\mathcal{M}} \left\{ \mathcal{C}^{\rm pen}_{V,F}(m) \right\} \ , \tag{3.14}$$

for any divisor $V \geq 2$ of n and $F > 0$. As already discussed this allows to study simultaneously any corrected V-fold criteria and any empirical criterion penalized by a resampling penalty with exchangeable weights.

3.3.4 Links

The following lemma makes the link between the criteria defined in the previous subsections. The proof is left to Section 3.7.1.

Lemma 3.1. *In the setting of Section 3.2, if $B_{[V]}$ satisfies* (**Reg**)*, we have*

$$\mathcal{C}_V^{\rm corr,VFCV}(m) = \mathcal{C}^{\rm pen}_{V,1}(m)$$

and

$$\begin{aligned}\mathcal{C}_V^{\rm VFCV}(m) - \mathcal{C}^{\rm pen}_{V,1}(m) \\ = \frac{1}{n^2(V-1)} \left(\sum_{i=1}^n A_m(X_i,X_i) + \sum_{1\leqslant i\neq j\leqslant n} \rho^{\rm (VF)}_{i,j} A_m(X_i,X_j) \right)\end{aligned} \tag{3.15}$$

where

$$\forall (x,x') \in \Xi^2, \quad A_m(x,x') := \int_\Xi \mathcal{K}_m(x,y)\mathcal{K}_m(x',y)d\mu(y) \ .$$

Finally,

$$\begin{aligned}\mathcal{C}_p^{\rm LPO}(m) - \mathcal{C}^{\rm pen}_{n,1}(m) \\ = \frac{p}{(n-p)n^2} \left(\sum_{i=1}^n A_m(X_i,X_i) - \frac{1}{n-1} \sum_{1\leqslant i\neq j\leqslant n} A_m(X_i,X_j) \right) \ .\end{aligned} \tag{3.16}$$

Remark that, for the V-fold weights $W_i^{\rm (VF)} = (V/(V-1))\mathbf{1}_{i\notin B_I}$,

$$\begin{aligned}\rho^{\rm (VF)}_{i,j} &= \rho^{W^{\rm (VF)}}_{i,j} \\ &= \mathbf{1}_{i,j \text{ belong to the same block}} - \frac{1}{V-1}\mathbf{1}_{i,j \text{ belong to different blocks}} \ ,\end{aligned}$$

therefore, from (3.12), the differences $\mathcal{C}_V^{\rm VFCV}(m) - \mathcal{C}^{\rm pen}_{V,1}(m)$ and $\mathcal{C}_p^{\rm LPO}(m) - \mathcal{C}^{\rm pen}_{n,1}(m)$ have the same structure as $\mathrm{pen}_{\rm VF}(m,V)$, but $\mathcal{K}_m$ is replaced by A_m. Moreover, when considering projection kernels (Example 1), it holds,

$$\forall (x,x') \in \Xi^2, \qquad A_m(x,x') = \mathcal{K}_m(x,x') \ ,$$

so that (3.15) becomes

$$\mathcal{C}_V^{\mathrm{VFCV}}(m) - \mathcal{C}_V^{\mathrm{corr,VFCV}}(m) = \frac{1}{2(V-1)}\,\mathrm{pen}_{\mathrm{VF}}(m,V) \ ,$$

and we recover Equation (7) in Lemma 1 of Arlot & Lerasle (2014).

3.4 Oracle inequalities

3.4.1 Notation and assumption

An important quantity in our study is

$$\mathcal{D}_m := \int_{\Xi^2} \mathcal{K}_m(x,y)^2 s(x) d\mu(x) d\mu(y) = \mathbb{E}\left[A_m(X,X)\right] \ , \tag{3.17}$$

which is always assumed to be positive. We also define $\gamma_m := \mathbb{E}\left[\chi_m(X)\right]/\mathcal{D}_m$. Remark that, by definition of Γ,

$$\mathcal{D}_m \leq \sup_{x\in\Xi} \int \mathcal{K}_m(x,y)^2 d\mu(y) \int s(x) d\mu(x) \leq \Gamma n \ . \tag{3.18}$$

We shall work with similar assumptions as in Lerasle *et al.* (2015). We assume that Υ is a constant such that the following hypotheses hold:

$$\sup_{m\in\mathcal{M}} \|s_m\|^2 \vee \left[\Gamma(1+\|s\|_\infty)\right] \leq \Upsilon \ , \tag{H1}$$

$$\forall m \in \mathcal{M}, \qquad \mathbb{E}\left[\chi_m(X)^2\right] \leqslant \Upsilon n \mathcal{D}_m \ , \tag{H2}$$

$$\forall (m,m') \in \mathcal{M}^2, \qquad \|s_m - s_{m'}\|_\infty \leqslant \Upsilon \vee \left(\sqrt{\Upsilon n}\,\|s_m - s_{m'}\|\right) \ , \tag{H3}$$

$$\forall m \in \mathcal{M}, \qquad \mathbb{E}\left[A_m(X,Y)^2\right] \leq \Upsilon \mathcal{D}_m \ , \tag{H4}$$

$$\forall m \in \mathcal{M}, \qquad \sup_{x\in\Xi} \mathbb{E}\left[A_m(X,x)^2\right] \leq \Upsilon n \ , \tag{H5}$$

$$\forall m \in \mathcal{M}, \qquad v_m^2 \leqslant \Upsilon \vee \sqrt{\Upsilon \mathcal{D}_m} \ , \tag{H6}$$

where $v_m^2 = \sup_{t\in\mathbb{B}_m} \mathrm{Var}\left[t(X)\right]$ and $\mathbb{B}_m$ is the set of functions t such that there exists $a \in \mathbb{L}_2$ with $\|a\| \leq 1$ and $t(x) = \int_\Xi a(y)\mathcal{K}_m(y,x)d\mu(y)$. These assumptions are discussed in Lerasle *et al.* (2015) and recalled below for the examples of kernels given in Section 3.2.2.

3.4.2 Oracle inequality for V-fold penalization and resampling penalties

The following theorem describes the oracle properties of the estimators selected by a V-fold penalty. As the parameter V and the leading constant F are left free in the following theorem, it is easy to deduce the oracle inequalities satisfied by the estimators minimizing the corrected V-fold cross-validation criterion or an empirical contrast penalized using an exchangeable resampling scheme.

Theorem 3.1. *Let $(\widehat{s}_m)_{m\in\mathcal{M}}$ denote a collection of linear estimators. Let*

$$\widehat{s}^{\mathrm{pen}}_{V,F} = \widehat{s}_{\widehat{m}^{\mathrm{pen}}_{V,F}} \ ,$$

where $\widehat{m}^{\mathrm{pen}}_{V,F}$ is defined in (3.14) *with $0 < F \leq 1 + n/4$, assuming* (**Reg**) *holds true. Let Υ be any constant such that Assumptions* (**H1**)*,* (**H2**)*,* (**H3**)*,* (**H4**)*,* (**H5**) *and* (**H6**) *hold. Let $(w_m)_{m\in\mathcal{M}}$ be a collection of positive real numbers such that $\sum_{m\in\mathcal{M}} e^{-w_m} \leqslant 1$. There exists an absolute constant κ such that, for any $\varepsilon > 0$ and any $x \geq \log(19.8)$, with probability larger than $1 - e^{-x}$, for any $m \in \mathcal{M}$,*

$$\frac{1 - 2(F-1)_-\gamma_{\widehat{m}^{\mathrm{pen}}_{V,F}} - \varepsilon}{1 + 2(F-1)_+\gamma_m + \varepsilon} \left\| \widehat{s}^{\mathrm{pen}}_{V,F} - s \right\|^2 \leq \|\widehat{s}_m - s\|^2 + \kappa\Upsilon(F^2 \vee 1)\frac{(w_m \vee w_{\widehat{m}^{\mathrm{pen}}_{V,F}} + x)^2}{n\varepsilon^3} \ . \quad (3.19)$$

Taking $\varepsilon > 0$ small enough in Theorem 3.1 proves V-fold penalized procedures satisfy an oracle inequality with large probability. The remainder term can be asymptotically bounded under the following classical assumption:

$$\exists a' > 0\,,\ \forall n \in \mathbb{N}^\star,\ |\mathcal{M}| = |\mathcal{M}_n| \leqslant n^{a'} \ . \quad (\mathbf{A3})$$

Under (**A3**) and choosing F smaller than some absolute constant, the remainder term in Eq. (3.19) is bounded by $L(\log n + x)^2/(\varepsilon^3 n)$ for some $L > 0$.
The leading constant in the oracle inequality (3.19) is $(1 + (\delta_{\widehat{m}})_+)/(1 - (\delta_m)_-) + \mathrm{o}(1)$, where $\delta_m = 2(F-1)\gamma_m$, by choosing $\varepsilon = \mathrm{o}(1)$, so the first-order behavior of the upper bound on the loss is driven by the δ_m. These quantities depend on m in general. However, for any projection kernel (Example 1), $\gamma_m = 1$, hence, δ_m is independent of m. Actually, this result extends the oracle inequalities in Arlot & Lerasle (2014). For Parzen's kernels (Example 3),

$$\gamma_m = \frac{k(0)}{\|k\|^2} \ ,$$

for any $m = (k, h)$. Hence, if we are interested in the selection of h, γ_m does not depend on m either. An asymptotic optimality result can be derived from Eq. (3.19) only if $\delta_m = \mathrm{o}(1)$. The meaning of $2(F-1)$ is the amount of bias of the V-fold penalization criterion, as shown in Arlot & Lerasle (2014). Given this interpretation, the model selection literature suggests no asymptotic optimality result can be obtained in general when $\delta_m \neq \mathrm{o}(1)$, see for instance Shao (1997). Therefore, even if the leading constant $(1 + (\delta_{\widehat{m}})_+)/(1 - (\delta_m)_-)$ is only an upper bound,

we conjecture it cannot be taken as small as $1+\mathrm{o}(1)$ unless $\delta_m = \mathrm{o}(1)$; such a result can be proved in our setting using similar arguments and assumptions as in Arlot (2008) for instance.
For bias-corrected V-fold cross-validation is obtained when $F = 1$, which implies $\delta_m = 0$. Theorem 3.1 shows a first-order optimal nonasymptotic oracle inequality, since the leading constant $(1+\varepsilon)/(1-\varepsilon)$ can be taken equal to $1+\mathrm{o}(1)$, and the remainder term is small enough under assumption (**A3**), for instance. Such a result valid with no upper bound on V was only obtained for projection estimators in Arlot & Lerasle (2014).
Assuming $F \leqslant 1 + n/4$ is necessary in the proof but is not constraining in practice since we usually use a constant $F \leq 2$.
The sup-norm of s appears in the oracle inequality since $\Upsilon \geq \|s\|_\infty$. This is not a fundamental problem since, as discussed in Birgé (2014), this sup-norm cannot in general be avoided to control the $\mathbb{L}_2$-risks of the $\widehat{s}_m$ by a deterministic universal bound.

3.4.3 Oracle inequality for cross-validation criteria

The result is the following.

Theorem 3.2. *Let $(\widehat{s}_m)_{m\in\mathcal{M}}$ denote a collection of linear estimators. Let*

$$\widetilde{s}^{\mathrm{VFCV}} = \widehat{s}_{\widehat{m}^{\mathrm{VFCV}}}, \quad \widetilde{s}^{\mathrm{LPO}} = \widehat{s}_{\widehat{m}^{\mathrm{LPO}}} \ ,$$

where $\widehat{m}^{\mathrm{VFCV}} = \operatorname{argmin}_{m\in\mathcal{M}} \mathcal{C}_V^{\mathrm{VFCV}}(m)$, $\widehat{m}^{\mathrm{LPO}} = \operatorname{argmin}_{m\in\mathcal{M}} \mathcal{C}_p^{\mathrm{LPO}}(m)$. Let Υ be the smallest constant such that Assumptions (**H1**)*,* (**H2**)*,* (**H3**)*,* (**H4**)*,* (**H5**) *and* (**H6**) *hold. Let $(w_m)_{m\in\mathcal{M}}$ be a collection of positive real numbers such that $\sum_{m\in\mathcal{M}} e^{-w_m} \leqslant 1$. There exists an absolute constant κ such that for any $x \geq \log(19.8)$, with probability larger than $1 - 19.8e^{-x}$, for any $\varepsilon \geq 1/n$*

$$\frac{1 - \frac{2\gamma_{\widehat{m}^{\mathrm{VFCV}}}}{V-1} - \varepsilon}{1 + \frac{2\gamma_m}{V-1} + \varepsilon} \left\| \widetilde{s}^{\mathrm{VFCV}} - s \right\|^2 \leq \|\widehat{s}_m - s\|^2 + \frac{\kappa\Upsilon}{n} \frac{\left(w_m \vee w_{\widehat{m}^{\mathrm{VFCV}}} + x \right)^2}{(\varepsilon \wedge 1)^3} \ .$$

$$\frac{1 - \frac{2p\gamma_{\widehat{m}^{\mathrm{LPO}}}}{n-p} - \varepsilon}{1 + \frac{2p\gamma_m}{n-p} + \varepsilon} \left\| \widetilde{s}^{\mathrm{LPO}} - s \right\|^2 \leq \|\widehat{s}_m - s\|^2 + \frac{\kappa\Upsilon}{n} \frac{\left(w_m \vee w_{\widehat{m}^{\mathrm{LPO}}} + x \right)^2}{(\varepsilon \wedge 1)^3} \ .$$

Remark. Theorem 3.2 analyzes the performances of the estimators obtained by minimization of classical cross-validation criteria. The discussion following Theorem 3.1 applies for a large part here also, and the performances are essentially similar. The most important difference is that these estimators are always biased, hence, the terms γ_m have to be controlled to derive meaningful results from the oracle inequalities. As already discussed, these γ_m are for example controlled in two important examples. For projection kernels, $\gamma_m = 1$ and, for Parzen's kernels, $\gamma_m = \frac{k(0)}{\|k\|^2}$. When a bound on $\max_{m\in\mathcal{M}} \gamma_m$ is available, we can derive an asymptotically optimal oracle inequality from Theorem 3.2 when $V = V_n \to \infty$ for $\widetilde{s}^{\mathrm{VFCV}}$ and when $p = p_n$ satisfies $p_n/(n - p_n) \to 0$ for $\widetilde{s}^{\mathrm{LPO}}$. This is consistent with the results of Arlot & Lerasle (2014) for $\widetilde{s}^{\mathrm{VFCV}}$ and those of Celisse (2014) for $\widetilde{s}^{\mathrm{LPO}}$. On the other hand, as in the discussion of Theorem 3.1, this suggests no asymptotically optimal oracle inequalities can be obtained if these assumptions are not satisfied.

Discussion of the assumptions. The examples of kernels given in Section 3.2.2 satisfy these assumptions as shown in Lerasle *et al.* (2015) and recalled by the following proposition.

Proposition 3.1. *Assumptions* (**H1**)*,* (**H2**)*,* (**H3**)*,* (**H4**)*,* (**H5**) *and* (**H6**) *hold in the examples of Section 3.2.2 under the following lower bound on* Υ*.*

- *For projection kernels, if for all* $m \in \mathcal{M}$, $\|s_m\|_\infty \leqslant \Upsilon/2$ *or*

$$\forall (m, m') \in \mathcal{M}^2, \qquad S_m + S_{m'} \subset \{ S_m, S_{m'} \} \quad ,$$

  (see (Massart, 2007, Chapter 7)) then all assumptions are satisfied once $\Upsilon \geq \Gamma(1 + \|s\|_\infty)$*.*

- *For weighted projection kernels, is it needed to suppose* $\Upsilon \geq \Gamma(1 + \|s\|_\infty)$*.*

- *For approximation kernels, if* $h \geq \|k\|_\infty \|k\|_1 / n$ *for any* $m = (k, h) \in \mathcal{M}$*, one requires*

$$\Upsilon \geq \max_k \left\{ \frac{k(0)}{\|k\|^2} \vee \left(1 + 2 \|s\|_\infty \|k\|_1^2 \right) \right\} \quad .$$

3.5 Sharp minimax adaptivity

It is well-known that oracle inequalities are powerful tools to prove adaptivity in the minimax sense (see among others Barron *et al.* (1999); Massart (2007); Goldenshluger & Lepski (2014)). Consider some class of functions $\mathcal{F} = \bigcup_{\sigma \in \Sigma} \mathcal{F}_\sigma$. The minimax risk over each $\mathcal{F}_\sigma$ is defined by

$$\mathcal{R}_{\text{minimax}} (\mathcal{F}_\sigma) := \inf_{\widehat{s}} \sup_{s \in \mathcal{F}_\sigma} \mathbb{E} \left[\|\widehat{s} - s\|^2 \right] \quad ,$$

the infimum being taken over the set of all estimators $\widehat{s}$. A minimax estimator $\widetilde{s}$ over $\mathcal{F}_\sigma$ is such that for some bounded sequence K_n,

$$\sup_{s \in \mathcal{F}_\sigma} \mathbb{E} \left[\|\widetilde{s} - s\|^2 \right] \leq K_n \mathcal{R}_{\text{minimax}} (\mathcal{F}_\sigma) \quad .$$

The estimator $\widetilde{s}$ is said to be adaptive to σ if it is simultaneously minimax over all classes $\mathcal{F}_\sigma$, that is, if, for every $\sigma \in \Sigma$,

$$\sup_{s \in \mathcal{F}_\sigma} \mathbb{E} \left[\|\widetilde{s} - s\|^2 \right] \leq K_n \mathcal{R}_{\text{minimax}} (\mathcal{F}_\sigma) \quad .$$

It is clear that each $K_n \geq 1$. When $K_n \to 1$, the estimator $\widetilde{s}$ is said to be sharp adaptive up to the constant to σ. In this section, we show that the oracle inequality obtained for corrected V-fold criteria can be used to derive some estimators adaptive up to the constant. In the sequel, $\Xi = [0, 1]$, μ is the Lebesgue measure on Ξ and $(\varphi_j)_{j \geq 0}$ is the Fourier basis defined by (3.3). The estimators $(\widehat{s}_m)_{m \in \mathcal{M}}$ are weighted projection estimators as in Example 2. The weights $m = (m_j)_{j \geq 0}$ are sequences of real numbers such that $m_0 = 1$ and $m_{2j-1} = m_{2j} =: \tau_j \in [0, 1]$ for all $j \geq 1$. Any

estimator in this collection is thus written

$$\forall x \in [0,1], \quad \widehat{s}_m(x) = 1 + \sum_{j=1}^{+\infty} \tau_j \left((P_n\varphi_{2j-1})\varphi_{2j-1}(x) + (P_n\varphi_{2j})\varphi_{2j}(x) \right) \ . \tag{3.20}$$

Hereafter, we emphasize the dependence to τ and denote the estimators $\widehat{s}_\tau$. Since the unknown density s belongs to $\mathbb{L}_2$, it can be developed on the Fourier basis. For any $j \geq 0$, $P\varphi_j$ is the Fourier coefficient of s, hence

$$s = 1 + \sum_{j=1}^{\infty} (P\varphi_j)\varphi_j \ .$$

The risk of any estimator $\widehat{s}_\tau$ defined in (3.20) is equal to

$$R(\tau) := \mathbb{E}\left[\|\widehat{s}_\tau - s\|^2 \right] = \sum_{j=1}^{\infty} \left(\frac{\tau_j^2}{n} + \left((1-\tau_j)^2 - \frac{\tau_j^2}{n} \right) \theta_j^2 \right) \ , \tag{3.21}$$

where, from now on, $\theta_j^2 = (P\varphi_{2j-1})^2 + (P\varphi_{2j})^2$ for all $j \geq 1$.

We provide two different applications based on this collection. First we consider Pinsker estimators and suppose that s belongs to some Sobolev class with parameters β and Q. Since Pinsker's estimators are minimax over Sobolev classes, we deduce from our oracle inequality the sharp minimax adaptivity to the unknown parameters β and Q of the estimator selected by the corrected V-fold criterion. Then we look at the more general class of estimators with nonincreasing weights in order to get the optimal minimax rate for all smooth densities (until some degree). The proofs of both results are left to the Appendix.

3.5.1 Adaptivity over Sobolev ellipsoids and Pinsker's estimators

Consider the collection of Pinsker's estimators, defined by (3.20) with weights $\tau^{\beta,Q} = (\tau_j^{\beta,Q})_{j\geq 1}$ given by (3.4). The goal is to derive from Theorem 3.1 that the selected estimator by Burman's criterion (3.8) is sharp adaptive over Sobolev ellipsoids. Recall that the class of Sobolev functions $\mathcal{S}(\beta, L)$ is defined for any $L > 0$, $\beta \in \mathbb{N}^*$ by

$$\mathcal{S}(\beta, L) := \left\{ f : [0,1] \mapsto \mathbb{R}_+ \text{ s.t. } f \in \mathcal{C}^\beta \text{ and } \left\| f^{(\beta)} \right\|^2 \leq L^2 \right\} \ .$$

Now, let $Q > 0$ and $a = (a_n)_{n\geq 1}$ be a nondecreasing sequence of real numbers such that $a_n \to \infty$ as $n \to \infty$. The ellipsoid $\mathcal{E}_{a,Q}$ is defined as the set

$$\mathcal{E}_{a,Q} = \left\{ (u_n)_{n\geq 1} \in \ell^2(\mathbb{N}) \text{ s.t. } \sum_{n\geq 1} a_n^2 u_n^2 \leq Q \right\} \ .$$

It is well-known (see (Tsybakov, 2009, Proposition 1.14)) that for any function $f \in \mathcal{S}(\beta, L)$ the sequence $(\langle f, \varphi_j \rangle)_{j\geq 1}$ belongs to the *Sobolev ellipsoid* $\mathcal{E}_{a,Q}$ defined for $Q = L^2/(2\pi)^{2\beta}$ and

$a = (a_n)_{n\geq 1}$ such that

$$\forall n \in \mathbb{N}^*, \qquad a_{2n-1} = a_{2n} = n^\beta \ .$$

The definition of Sobolev classes is then extended to any $\beta > 0$. We note $\mathcal{S}'(\beta, Q)$ the space of functions f such that $(\langle f, \varphi_j \rangle)_{j\geq 1}$ belongs to the Sobolev ellipsoid $\mathcal{E}_{a,Q}$. For all $\beta > 1/2$ the functions belonging to $\mathcal{S}'(\beta, Q)$ are continuous. Moreover, we have monotonicity with respect to inclusion, that is for $0 < \beta < \beta'$, $\mathcal{S}'(\beta', Q) \subset \mathcal{S}'(\beta, Q)$.

The first step to prove minimax adaptivity is to show that Pinsker's estimators are minimax over Sobolev classes. The following classical result is proved in the Appendix for sake of completeness.

Proposition 3.2. *For each $\beta > 1/2, Q > 0$ Pinsker's estimator $\widehat{s}_{\tau^{\beta,Q}}$ is sharp minimax over $\mathcal{S}'(\beta, Q)$, more precisely*

$$\sup_{s\in\mathcal{S}'(\beta,Q)} R\left(\tau^{\beta,Q}\right) = \left(1 + \mathrm{O}\left(n^{-1/(2\beta+1)}\right)\right) \mathcal{R}_{\text{minimax}}\left(\mathcal{S}'(\beta, Q)\right) \qquad n \to \infty \ .$$

Unfortunately, the optimal value of the parameters β and Q is unknown in practice. We propose to use the corrected V-fold criterion to select these parameters. To apply our method, we have to discretize the set of parameters. First, we assume that $\beta \in (1/2, n)$ and $Q \in (0, (\log n)^2)$. Then, we discretize these sets using grids with steps equal to $1/n$. Let us call $\widetilde{s}$ the estimator selected by Burman's criterion over this set of linear estimators. The following corollary shows that it is asymptotically minimax adaptive (up to the constant) over the Sobolev classes $\mathcal{S}'(\beta, Q)$.

Corollary 3.1. *For any $n \geq 6(\log n)^2$, any $\beta \in (1/2, n)$ and any $Q \in (0, (\log n)^2)$,*

$$\sup_{s\in\mathcal{S}'(\beta,Q)} \mathbb{E}\left[\|s - \widetilde{s}\|^2\right] \leq \widetilde{C}_{\beta,Q}\, n^{\frac{-2\beta}{2\beta+1}} \left(1 + C_Q \frac{\log n}{n^{1/(8\beta+4)}}\right) \ ,$$

where $\widetilde{C}_{\beta,Q}$ is Pinsker's constant

$$\widetilde{C}_{\beta,Q} := Q^{\frac{1}{2\beta+1}} (2\beta + 1)^{\frac{1}{2\beta+1}} \left(\frac{\beta}{\beta+1}\right)^{\frac{2\beta}{2\beta+1}} \ .$$

Remark. We could have obtained the bound for any $(\beta, Q) \in (1/2, n^p) \times (0, (\log n)^q)$ changing in a straightforward way the collection of models, replacing the constant C_Q by a constant $C_{Q,p,q}$ and the exponent of the logarithm by some function $\kappa_{p,q}$.

To the best of our knowledge, all adaptive results over Sobolev classes were obtained using blockwise Stein construction, see for example Cavalier & Tsybakov (2002) in the inverse problem or Rigollet (2006a) for density estimation or some aggregation procedures, see Rigollet & Tsybakov (2007). We are not aware of other model selection procedures achieving this goal. In these papers, the authors considered the strongest problem of adaptivity over the class of weighted projection kernels with nonincreasing weights. Actually, Pinsker's estimators belong to this class of estimators. Other famous examples are discussed in Cavalier & Tsybakov (2002) for example.

3.5.2 Adaptivity over nonincreasing weights

In this section, we derive adaptive estimators over a large class of nonincreasing weights via corrected V-fold criteria. For this purpose let $\mathcal{S}_{ni}$ denote the set of nonincreasing admissible weight sequences.

$$\mathcal{S}_{ni} = \left\{ \tau \in \ell^2 \left(\mathbb{N}^* \right) \text{ s.t. } 1 \geq \tau_1 \geq \tau_2 \geq \ldots \geq 0 \right\} \ .$$

In addition, for any closed convex subset of sequences $\mathcal{S} \subset \ell^2$, let

$$\tau^{\mathcal{S}} := \operatorname*{argmin}_{\tau \in \mathcal{S}} \ R(\tau) \ ,$$

which is always well-defined since R is continuous and strictly convex. The following corollary shows the existence of a family of kernels such that the estimator selected by V-fold criteria is adaptive to the nonincreasing weights. This family of kernels is built using blockwise Stein's construction of Cavalier & Tsybakov (2002) or Rigollet (2006a).

Corollary 3.2. *Let s be such that $\|s\|_\infty \leq \left((1 + \sqrt{\log n})/3 \right)^2$. Then there exist some subset $\mathcal{S}^d_{B,I} \subset \mathcal{S}_{ni}$ and a collection of weighted projection estimators $\{\widehat{s}_\tau, \tau \in \mathcal{S}^d_{B,I}\}$ such that the $\mathbb{L}_2$-risk of $\widetilde{s} = \widehat{s}_{\tau^{\mathrm{pen}}_{V,1}}$, where*

$$\tau^{\mathrm{pen}}_{V,1} \in \operatorname*{argmin}_{\tau \in \mathcal{S}^d_{B,I}} \left\{ P_n \gamma \left(\widehat{s}_\tau \right) + \mathrm{pen}_{\mathrm{VF}}(\tau, V) \right\} \ ,$$

satisfies

$$R\left(\tau^{\mathrm{pen}}_{V,1} \right) \leq \left(1 + \frac{\kappa}{\sqrt{\log n}} \right) R\left(\tau^{\mathcal{S}_{ni}} \right) + \kappa \frac{(\log n)^{11}}{n} \ ,$$

for some absolute constant $\kappa > 0$ and any n larger than some absolute constant n_0.

This result is not as strong as the one in Rigollet (2006a) where the remainder term has the form κ/n. In particular, Corollary 3.2 does not show optimal adaptivity of the selected estimator over classes of super-smooth functions, where the rate $R(\tau^{\mathcal{S}_{ni}})$ is $\log n/n$. Nevertheless, it shows that V-fold criteria select adaptive over classes of polynomially decreasing rates.

3.6 Simulation experiments

3.6.1 Simulation protocol

In this section $\Xi = \mathbb{R}$. Since simulations were already made for histograms in Arlot & Lerasle (2014), we decided to focus on two different families of estimators. First, Parzen's kernel density estimators in order to deal with the famous bandwidth selection problem with a model selection strategy. Second, a mix of those estimators together with regular histograms to see if the penalized criterion is adaptive to the best collection, i.e. if it selects the best candidate between regular or Parzen estimator when one collection is worse than the other one.

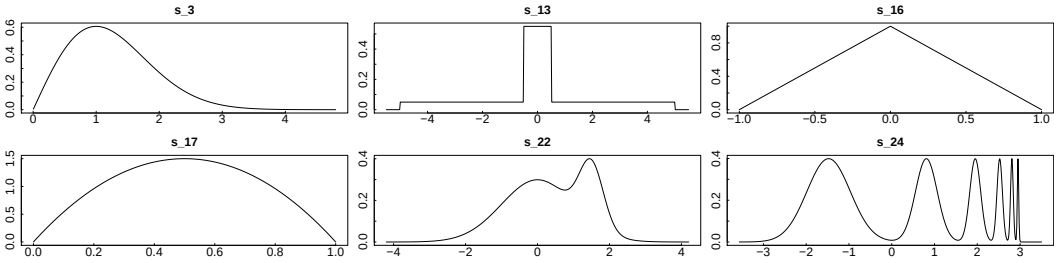

Figure 3.1: All the densities that we consider in this section.

- "Par" corresponds to the collection $(\widehat{s}_m)_{m\in\mathcal{M}_\mathcal{P}}$ of approximation Gaussian kernels with a "geometrical" grid of bandwidths, that is with

$$\mathcal{K}_m(x,y) = \frac{1}{\sqrt{2\pi}h_m}\exp\left(-\frac{(x-y)^2}{2h_m^2}\right) \quad , \tag{3.22}$$

$$\mathcal{M}_\mathcal{P} = \left\{ h_m = \frac{1}{n\log n}\left(1+\frac{3}{2\log n}\right)^m,\ m=1,\dots,\left\lfloor(\log n)^2\right\rfloor\right\} \quad . \tag{3.23}$$

- "Mix" corresponds to $(\widehat{s}_m)_{m\in\mathcal{M}} = (\widehat{s}_m)_{m\in\mathcal{M}_\mathcal{P}} \cup (\widehat{s}_m)_{m\in\mathcal{M}_\mathcal{R}}$, where $(\widehat{s}_m)_{m\in\mathcal{M}_\mathcal{R}}$ is a family of regular histogram estimators with

$$\mathcal{M}_\mathcal{R} = \left\{ m=1,\dots,\left\lfloor\frac{n}{\log n}\right\rfloor\right\} \quad . \tag{3.24}$$

  More precisely, for each $m\in\mathcal{M}_\mathcal{R}$, let $(I_\lambda)_{\lambda\in\Lambda_m}$ be a regular partition of Ξ with m bins. Then the corresponding projection kernel comes from the orthonormal family $(\psi_\lambda)_{\lambda\in\Lambda_m}$ with $\psi_\lambda(x) = |I_\lambda|^{-1/2}\mathbf{1}_{I_\lambda}(x)$.

We considered in our study a total of 14 densities[(3)]. For sake of clarity, we only show the results for some of them that are plotted in Figure 3.1. Note that s_{11} is the standard Gaussian density $\mathcal{N}(0,1)$.

All the simulations were carried out with $n = 100, 500, 1000$. We computed the classical least-squares V-fold criteria together with some penalized V-fold criteria with

- $V \in \{2,5,10,n\}$, the choice $V=n$ representing the leave-one-out;
- $\mathit{F} \in \{1, 5/4, 3/2, 7/4, 2, 5/2, 3\}$, the choice $\mathit{F}=1$ representing Burman's criterion.

For the LOO and the penalized LOO criteria, we used closed formulas which allow to speed up the computations. Concerning the classical LOO criterion, they appear explicitly in Celisse (2014) for histograms and Parzen estimators with Gaussian kernel. For the penalized LOO, they can be deduced from (3.12). Indeed, when $V=n$ we get $\rho_{i,j}^{(\mathrm{VF})} = -1/(n-1)$ for all $i\neq j$.

[(3)]from the R-package 'benchden', that implements the benchmark distributions of Berlinet & Devroye (1994) in density estimation.

Simple computations lead to the following penalty

$$\mathrm{pen}_{\mathrm{VF}}(m,n) = \sqrt{\frac{2}{\pi}}\frac{1}{n^2 h_m}\left(n - \frac{1}{n-1}\sum_{i\neq j=1}^{n}\exp\left(\frac{-(X_i-X_j)^2}{2h_m^2}\right)\right) \ ,$$

for Parzen estimators with kernel Gaussian and bandwidth $h_m > 0$. For regular histogram estimators with bin size $|I_\lambda| = |I_m|$ for all $\lambda \in \Lambda_m$

$$\mathrm{pen}_{\mathrm{VF}}(m,n) = \frac{2}{n^2(n-1)\,|I_m|}\left(n^2 - \sum_{\lambda\in\Lambda_m} N_\lambda^2\right), \quad \text{where } N_\lambda := \sum_{i=1}^{n}\mathbf{1}_{X_i\in I_\lambda} \ .$$

All procedures are compared on $N = 10000$ independent synthetic data sets of size n. To measure the quality of some procedure $\mathcal{C}$, we estimate, as in (Arlot & Lerasle, 2014),

$$C_{or}(\mathcal{C}) := \mathbb{E}\left[\frac{\left\|\widehat{s}_{\widehat{m}_\mathcal{C}} - s\right\|^2}{\inf_{m\in\mathcal{M}}\left\|\widehat{s}_m - s\right\|^2}\right] \tag{3.25}$$

which represents the constant that would appear in front of an oracle inequality.

3.6.2 Simulation results

For sake of clarity we only show here the results for $n = 500$ and $\digamma = 1$. The concerned reader should go to Section 3.11 for a more complete picture of our empirical study. The main question when considering VF type procedures is maybe "which V is optimal?" or, more generally, "what is the influence of V on the quality of the VF procedure?". In the tables below we provide the value of the following quantity

$$\xi(\mathcal{C}) = \mathbb{E}\left[C_{or}(\mathcal{C})\right] \pm \sqrt{\frac{\mathrm{Var}\left[C_{or}(\mathcal{C})\right]}{N}}$$

for different procedures $\mathcal{C}$.

The conclusion of this empirical study about the influence of V seems to be similar as the one drawn for the penalized V-fold criterion for projection estimators (Arlot & Lerasle, 2014, Section 5). We notice indeed from Table 3.1 and Table 3.2 that in every setting the quality of the estimation increases with V. The most significant gain appears between $V = 2$ and $V = 5$, then the quality keeps improving but with very little difference between $V = 10$ and $V = n$. This table suggests that the larger the value of V, the better the accuracy of the $\mathrm{pen}_{\mathrm{VF}}$ procedure.

The purpose of Table 3.3 is to show that at least the best VF procedure selects the optimal estimator when one collection is well-designed to a particular density. For instance for s_{24} it chooses a Parzen-Rosenblatt estimator since these are better than histograms, whereas it selects an histogram for s_{13} for the opposite reason.

procedure $\mathcal{C}$	s_3	s_{11}	s_{16}	s_{17}
pen2F	2.29 ± 0.04	3.08 ± 0.08	2.59 ± 0.04	2.34 ± 0.03
pen5F	1.75 ± 0.02	2.27 ± 0.05	1.96 ± 0.02	1.78 ± 0.02
pen10F	1.66 ± 0.02	2.14 ± 0.06	1.82 ± 0.02	1.67 ± 0.02
penLOO	1.54 ± 0.01	1.82 ± 0.02	1.68 ± 0.01	1.57 ± 0.01
LS2F	1.62 ± 0.01	1.87 ± 0.02	1.76 ± 0.01	1.70 ± 0.01
LS5F	1.52 ± 0.01	1.79 ± 0.02	1.69 ± 0.01	1.56 ± 0.01
LS10F	1.52 ± 0.01	1.81 ± 0.02	1.69 ± 0.01	1.56 ± 0.01
LOO	1.51 ± 0.01	1.81 ± 0.02	1.67 ± 0.01	1.53 ± 0.01
Epenid	1.62 ± 0.03	2.16 ± 0.07	1.69 ± 0.01	1.58 ± 0.01

Table 3.1: Values of $\xi(\mathcal{C})$ for several procedures for collection $\mathcal{F}_{\mathrm{K}}$ with $n = 500$. See the text.

procedure $\mathcal{C}$	s_3	s_{11}	s_{13}	s_{16}	s_{17}	s_{22}
LS2F	3.66 ± 0.04	4.81 ± 0.08	1.54 ± 0.009	3.95 ± 0.05	3.40 ± 0.035	2.82 ± 0.03
LS5F	3.40 ± 0.04	4.38 ± 0.09	1.49 ± 0.009	3.73 ± 0.05	3.10 ± 0.037	2.54 ± 0.03
LS10F	3.36 ± 0.04	4.30 ± 0.09	1.49 ± 0.010	3.72 ± 0.05	3.06 ± 0.038	2.49 ± 0.03
LOO	3.39 ± 0.04	4.37 ± 0.09	1.50 ± 0.010	3.78 ± 0.05	3.23 ± 0.040	2.46 ± 0.03
pen2F	6.87 ± 0.07	9.57 ± 0.15	1.91 ± 0.017	7.77 ± 0.09	6.88 ± 0.076	4.72 ± 0.04
pen5F	4.47 ± 0.06	6.17 ± 0.12	1.69 ± 0.014	4.88 ± 0.07	4.13 ± 0.053	3.20 ± 0.04
pen10F	3.88 ± 0.05	5.28 ± 0.11	1.63 ± 0.014	4.26 ± 0.06	3.57 ± 0.046	2.81 ± 0.03
penLOO	3.36 ± 0.04	4.39 ± 0.09	1.50 ± 0.010	3.68 ± 0.05	3.07 ± 0.039	2.47 ± 0.03
Epenid	3.26 ± 0.06	4.93 ± 0.12	1.59 ± 0.014	3.28 ± 0.05	2.52 ± 0.037	2.64 ± 0.04

Table 3.2: Values of $\xi(\mathcal{C})$ for several procedures for collection $\mathcal{F}_{\mathrm{KR}}$ with $n = 500$. See the text.

density	Best($\mathcal{F}_{\mathrm{K}}$)	Oracle($\mathcal{F}_{\mathrm{K}}$)	Best($\mathcal{F}_{\mathrm{R}}$)	Oracle($\mathcal{F}_{\mathrm{R}}$)	Best($\mathcal{F}_{\mathrm{KR}}$)	Oracle($\mathcal{F}_{\mathrm{KR}}$)
s_3	3.84±0.02	3±0.02	13.8±0.05	7.67±0.02	4.58±0.04	3±0.02
s_{11}	2.1±0.01	1.6±0.01	8.36±0.03	4.61±0.01	2.72±0.03	1.6±0.01
s_{13}	16.6±0.03	15.5±0.03	14.1±0.09	11±0.05	13.8±0.08	10.3±0.04
s_{16}	5.3±0.03	4.03±0.02	19.5±0.06	11.3±0.03	6.31±0.04	4.03±0.02
s_{17}	9.24±0.05	7.24±0.04	35.2±0.1	20.3±0.06	10.1±0.06	7.24±0.04
s_{22}	3.72±0.02	3.08±0.02	10±0.03	5.96±0.02	5.18±0.03	3.08±0.01
s_{24}	16.2±0.03	15.2±0.03	26.8±0.05	20±0.03	18.3±0.05	15.2±0.03

Table 3.3: $\mathbb{L}_2$-risk multiplied by 1000 $n = 500$. See the text.

3.7 Main proofs

3.7.1 Proof of Lemma 3.1

The first relation has already been proved in Arlot (2008) and can be found again in Lemma 1 of Arlot & Lerasle (2014). It does not depend on the particular estimators but relies only on the fact that, for regular partitions, $(V-1)(P_n - P_n^{(B_k^c)}) = P_n^{(B_k)} - P_n$.
Concerning the second relation, let us write

$$\|\widehat{s}_m\|^2 = \frac{1}{n^2}\left(\sum_{i=1}^{n} A_m(X_i,X_i) + \sum_{1\leqslant i\neq j\leqslant n} A_m(X_i,X_j)\right) \ ,$$

$$\left\|\widehat{s}_m^{(B_k^c)}\right\|^2 = \frac{V^2}{n^2(V-1)^2}\left(\sum_{i\in B_k^c} A_m(X_i,X_i) + \sum_{i\neq j\in B_k^c} A_m(X_i,X_j)\right) \ .$$

Since B is regular, $V^{-1}\sum_{k=1}^{V} P_n^{(B_k^c)} = P_n$ implies $V^{-1}\sum_{k=1}^{V} \widehat{s}_m^{(B_k^c)} = \widehat{s}_m$ and

$$\frac{1}{V}\sum_{k=1}^{V}\frac{V}{n(V-1)}\sum_{i\in B_k^c} A_m(X_i,X_i) = \frac{1}{n}\sum_{i=1}^{n} A_m(X_i,X_i) \ .$$

Moreover,

$$\begin{aligned}\sum_{K=1}^{V}\sum_{i\neq j\in B_K^c} A_m(X_i,X_j) &= \sum_{1\leqslant i\neq j\leqslant n}\left((V-1)\mathbf{1}_{k=k'} + (V-2)\mathbf{1}_{k\neq k'}\right)A_m(X_i,X_j)\\ &= \sum_{1\leqslant i\neq j\leqslant n}\left((V-1) - \mathbf{1}_{k\neq k'}\right)A_m(X_i,X_j) \ ,\end{aligned}$$

which allows to conclude, since

$$\begin{aligned}&\mathcal{C}_V^{\mathrm{VFCV}}(m) - \mathcal{C}_V^{\mathrm{corr,VFCV}}(m)\\ &= \frac{1}{V}\sum_{k=1}^{V}\left\|\widehat{s}_m^{(B_k^c)}\right\|^2 - \|\widehat{s}_m\|^2 + 2P_n\widehat{s}_m - \frac{2}{V}\sum_{k=1}^{V} P_n\widehat{s}_m^{(B_k^c)}\\ &= \frac{1}{n^2(V-1)}\sum_{i=1}^{n} A_m(X_i,X_i) + \frac{V}{n^2(V-1)^2}\sum_{k=1}^{V}\sum_{i\neq j\in B_k^c} A_m(X_i,X_j)\\ &\quad - \frac{1}{n^2}\sum_{1\leqslant i\neq j\leqslant n} A_m(X_i,X_j)\\ &= \frac{1}{n^2(V-1)}\left(\sum_{i=1}^{n} A_m(X_i,X_i) + \sum_{1\leqslant i\neq j\leqslant n}\rho_{i,j}^{(\mathrm{VF})} A_m(X_i,X_j)\right) \ .\end{aligned}$$

Let us turn to the leave-p-out criterion. Recall that $\mathcal{E}_p$ denotes the set of all subsets $T \subset [n]$ with same cardinality $|T| = n - p$. It is well-known that $P_n = |\mathcal{E}_p|^{-1} \sum_{T \in \mathcal{E}_p} P_n^{(T)}$, and for any $T \in \mathcal{E}_p$

$$P_n^{(T^c)} = \frac{n}{p} P_n - \frac{q}{p} P_n^{(T)}, \quad \text{with} \quad q = n - p \ .$$

Using these relations and denoting for any $i, j \in [n]$ and $T \subset [n]$ $E_{i,j}(T) = \mathbf{1}_{i,j \in T}$, one obtains

$$\begin{aligned}
\mathcal{C}_p^{\mathrm{LPO}}(m) &= \frac{1}{\binom{n}{p}} \sum_{T \in \mathcal{E}_p} \left(\left\| \widehat{s}_m^{(T)} \right\|^2 - 2 P_n^{(T^c)} \left(\widehat{s}_m^{(T)} \right) \right) \\
&= \frac{1}{\binom{n}{p}} \sum_{T \in \mathcal{E}_p} \left(\frac{1}{q^2} \sum_{i,j=1}^{n} E_{i,j}(T) A_m(X_i, X_j) - 2 \left(\frac{n}{p} P_n - \frac{q}{p} P_n^{(T)} \right) \left(\widehat{s}_m^{(T)} \right) \right) \\
&= \frac{1}{\binom{n}{p}} \sum_{T \in \mathcal{E}_p} \frac{1}{q^2} \sum_{i,j=1}^{n} E_{i,j}(T) \left(A_m(X_i, X_j) + \frac{2q}{p} \mathcal{K}_m(X_i, X_j) \right) - 2 \frac{n}{p} P_n \widehat{s}_m \ .
\end{aligned}$$

By definition of $\mathcal{E}_p$,

$$\frac{1}{\binom{n}{p}} \sum_{T \in \mathcal{E}_p} E_{i,i}(T) = \frac{\binom{n-1}{n-p-1}}{\binom{n}{p}} = \frac{q}{n} \ ,$$

and, for any $i \neq j$, if B denotes some element of $\mathcal{E}_p$ uniformly chosen,

$$\frac{1}{\binom{n}{p}} \sum_{T \in \mathcal{E}_p} E_{i,j}(T) = \mathbb{P}\left((i,j) \in B \right) = \frac{\binom{n-2}{n-p-2}}{\binom{n}{p}} = \frac{(n-p)(n-p-1)}{n(n-1)} \ .$$

Hence,

$$\begin{aligned}
\frac{1}{\binom{n}{p}} \sum_{T \in \mathcal{E}_p} \frac{1}{q^2} \sum_{i,j \in T} g(X_i, X_j) &= \frac{1}{q^2} \sum_{i,j=1}^{n} \left(\frac{1}{\binom{n}{p}} \sum_{T \in \mathcal{E}_p} E_{i,j}(T) \right) g(X_i, X_j) \\
&= \frac{1}{nq} \left(\sum_{i,j=1}^{n} g(X_i, X_j) - \frac{p}{n-1} \sum_{1 \leqslant i \neq j \leqslant n} g(X_i, X_j) \right) \ .
\end{aligned}$$

Using this expression with $g(x, y) = A_m(x, y) + \mathcal{K}_m(x, y) 2qp^{-1}$, one finds

$$\mathcal{C}_p^{\mathrm{LPO}}(m) = \frac{n}{q} \left\| \widehat{s}_m \right\|^2 - \frac{1}{n(n-1)} \sum_{1 \leqslant i \neq j \leqslant n} \left(\frac{p}{q} A_m(X_i, X_j) + 2 \mathcal{K}_m(X_i, X_j) \right) \ .$$

Therefore,

$$\mathcal{C}_p^{\mathrm{LPO}}(m) - P_n \gamma(\widehat{s}_m) = \frac{p}{qn^2} \left(\sum_{i=1}^{n} A_m(X_i, X_i) - \frac{1}{n-1} \sum_{1 \leqslant i \neq j \leqslant n} A_m(X_i, X_j) \right)$$

$$+\frac{2}{n^2}\left(\sum_{i=1}^{n}\chi_m(X_i)-\frac{1}{n-1}\sum_{1\leqslant i\neq j\leqslant n}\mathcal{K}_m(X_i,X_j)\right)$$

that is, by (3.12) with $V=n$,

$$\mathcal{C}_p^{\mathrm{LPO}}(m)-\mathcal{C}_{n,1}^{\mathrm{pen}}(m)=\frac{p}{qn^2}\left(\sum_{i=1}^{n}A_m(X_i,X_i)-\frac{1}{n-1}\sum_{1\leqslant i\neq j\leqslant n}A_m(X_i,X_j)\right)\ .$$

3.7.2 Concentration inequalities

A common technique for proving oracle inequalities for penalized criteria (see for instance Arlot & Lerasle (2014)) requires concentration inequalities for the loss $\|\widehat{s}_m-s\|^2$ and the difference $\mathcal{F}\,\mathrm{pen}_{\mathrm{VF}}(m,V)-\mathrm{pen}_{\mathrm{id}}(m)$, uniformly with respect to $m\in\mathcal{M}$. It follows from (3.12) and the definition of $\mathrm{pen}_{\mathrm{id}}$ that

$$\begin{aligned}\mathcal{F}\,\mathrm{pen}_{\mathrm{VF}}(m,V)-\mathrm{pen}_{\mathrm{id}}(m)=\frac{2(\mathcal{F}-1)}{n^2}\sum_{i=1}^{n}\chi_m(X_i)+\frac{2}{n}\sum_{i=1}^{n}s_m(X_i)\\+\frac{2}{n^2}\sum_{1\leqslant i\neq j\leqslant n}(\mathcal{F}\rho_{i,j}^{(\mathrm{VF})}-1)\mathcal{K}_m(X_i,X_j)\ .\end{aligned}$$

The two means $P_n\chi_m$ and P_ns_m will concentrate around their corresponding expectations thanks to Bernstein's inequality. The last term appearing in this expression is somewhat more complicated to handle. For this purpose, we introduce the totally degenerate U-statistic of order 2

$$U_m^{(\mathcal{K})}=\sum_{1\leqslant i\neq j\leqslant n}(\mathcal{F}\rho_{i,j}^{(\mathrm{VF})}-1)(\mathcal{K}_m(X_i,X_j)-s_m(X_i)-s_m(X_j)+Ps_m)\ .$$

Elementary algebra shows $\sum_{j=1,\,j\neq i}^{n}(\mathcal{F}\rho_{i,j}^{(\mathrm{VF})}-1)=-(n+\mathcal{F}-1)$, therefore the Hoeffding's decomposition of the U-statistics of interest is given by

$$\begin{aligned}\frac{2}{n^2}\sum_{1\leqslant i\neq j\leqslant n}(\mathcal{F}\rho_{i,j}^{(\mathrm{VF})}-1)\mathcal{K}_m(X_i,X_j)=\frac{2}{n^2}U_m^{(\mathcal{K})}\\-\frac{4}{n}(n+\mathcal{F}-1)(P_n-P)(s_m)-\frac{2(n+\mathcal{F}-1)}{n}Ps_m\ .\end{aligned}$$

Thus,

$$\begin{aligned}\mathcal{F}\,\mathrm{pen}_{\mathrm{VF}}(m,V)-\mathrm{pen}_{\mathrm{id}}(m)=\frac{2}{n^2}U_m^{(\mathcal{K})}+\frac{2(\mathcal{F}-1)}{n}\,(P_n-P)\,(\chi_m)\\+\frac{2(\mathcal{F}-1)}{n}P\,(\chi_m-s_m)-2\left(1+\frac{2(\mathcal{F}-1)}{n}\right)(P_n-P)(s_m)\ .\end{aligned}\quad(3.26)$$

Since we are actually interested in the difference $F\,\mathrm{pen}_{\mathrm{VF}}(m,V)-\mathrm{pen}_{\mathrm{id}}(m)-(F\,\mathrm{pen}_{\mathrm{VF}}(m',V)-\mathrm{pen}_{\mathrm{id}}(m'))$ for all $m,m'\in\mathcal{M}$, let us introduce $\delta(m,m')=(P_n-P)(\chi_m-\chi_{m'})$ and $\delta^*(m,m')=(P_n-P)(s_m-s_{m'})$.

Concentration of $U_m^{(\mathcal{K})}$

Proposition 3.3. *There exists an absolute constant $\kappa>0$ such that for any $x>\log(5.4)$, for any $m\in\mathcal{M}$, with probability larger than $1-5.4e^{-x}$, we have for any $\varepsilon>0$*

$$\left|\frac{2U_m^{(\mathcal{K})}}{n^2}\right|\leq\varepsilon\frac{\mathcal{D}_m}{n}+\frac{\kappa\Upsilon(F+1)^2x^2}{n(\varepsilon\wedge 1)}\ .$$

Proof: Let g be the function defined for all $(x,y)\in\Xi^2$ by

$$g(x,y)=\mathcal{K}_m(x,y)-\mathbb{E}\left[\mathcal{K}_m(x,X)\right]-\mathbb{E}\left[\mathcal{K}_m(y,X)\right]+\mathbb{E}\left[\mathcal{K}_m(X,Y)\right]\ .$$

Let $f_{i,j}(x,y)=(F\rho_{i,j}^{(\mathrm{VF})}-1)g(x,y)$ so that $U_m^{(\mathcal{K})}=\sum_{1\leqslant i\neq j\leqslant n}f_{i,j}(X_i,X_j)$. For any $x\in\Xi$,

$$\forall 1\leqslant i\neq j\leq n,\qquad \mathbb{E}\left[f_{i,j}(x,X)\right]=0\qquad \text{a.s. .}$$

Hence, Theorem 3.4 applies to the U-statistic $U_m^{(\mathcal{K})}$. By definition,

$$\max_{1\leq i,j\leq n}\left|F\rho_{i,j}^{(\mathrm{VF})}-1\right|=(F-1)\mathbf{1}_{\left\{F\geq\frac{2(V-1)}{V-2}\right\}}+\left(1+\frac{F}{V-1}\right)\mathbf{1}_{\left\{F\leq\frac{2(V-1)}{V-2}\right\}}\ ,$$

hence $\max_{1\leq i,j\leq n}\left|F\rho_{i,j}^{(\mathrm{VF})}-1\right|\leq F+1$. Moreover, for any $i\in[n]$,

$$\sum_{j=1}^{n}\left(F\rho_{i,j}^{(\mathrm{VF})}-1\right)^2\leq n(F+1)^2,\qquad \sum_{1\leq i\neq j\leq n}\left(F\rho_{i,j}^{(\mathrm{VF})}-1\right)^2<n^2(F+1)^2\ .$$

Following the notation of Theorem 3.4, we need to bound the terms A,B,C and D. First, we have

$$A=\max_{1\leq i,j\leq n}\sup_{(x,y)\in\Xi^2}|f_{i,j}(x,y)|\leq 4(F+1)\sup_{(x,y)\in\Xi^2}|\mathcal{K}_m(x,y)|\ ,$$

Then, we have $B^2\leq n(F+1)^2\sup_{x\in\Xi}\mathbb{E}\left[g(x,Z)^2\right]$, with

$$\begin{aligned}\mathbb{E}\left[g(x,Z)^2\right]&=\mathbb{E}\left[\left(\mathcal{K}_m(x,Z)-\mathbb{E}\left[\mathcal{K}_m(x,X)\right]-\mathbb{E}\left[\mathcal{K}_m(Z,X)\,|\,Z\right]+\mathbb{E}\left[\mathcal{K}_m(X,Y)\right]\right)^2\right]\\&=\mathrm{Var}\left[\mathcal{K}_m(x,Z)-\mathbb{E}\left[\mathcal{K}_m(Z,X)\,|\,Z\right]\right]\\&\leq\mathbb{E}\left[\left(\mathcal{K}_m(x,Z)-\mathbb{E}\left[\mathcal{K}_m(Z,X)\,|\,Z\right]\right)^2\right]\\&\leq 2\left(\mathbb{E}\left[\mathcal{K}_m(x,Z)^2\right]+\mathbb{E}\left[\left(\mathbb{E}\left[\mathcal{K}_m(Z,X)\,|\,Z\right]\right)^2\right]\right)\ .\end{aligned}$$

Now, Jensen's inequality implies

$$\mathbb{E}\left[\left(\mathbb{E}\left[\mathcal{K}_m(Z,X)\,|\,Z\right]\right)^2\right] \leq \mathbb{E}\left[\mathbb{E}\left[\mathcal{K}_m(Z,X)^2\,\middle|\,Z\right]\right] \leq \sup_{x\in\Xi}\mathbb{E}\left[\mathcal{K}_m(x,X)^2\right] \ .$$

Hence, $B^2 \leq 4n(\digamma+1)^2 \sup_{x\in\Xi}\mathbb{E}\left[\mathcal{K}_m(x,X)^2\right]$.

Concerning the term C, we obtain from the definition $C^2 \leq n^2(\digamma+1)^2\mathbb{E}\left[g(X,Y)^2\right]$. Setting $W = \mathcal{K}_m(X,Y) - \mathbb{E}\left[\mathcal{K}_m(Y,Z)\,|\,Y\right]$, we have

$$\mathbb{E}\left[g(X,Y)^2\right] = \mathbb{E}\left[\left(W - \mathbb{E}\left[W\,|\,X\right]\right)^2\right] = \mathbb{E}\left[\mathrm{Var}\left[W\,|\,X\right]\right] \ .$$

Then, by the definition of $\mathrm{Var}\left[W\right] = \mathrm{Var}\left[\mathbb{E}\left[W\,|\,Y\right]\right] + \mathbb{E}\left[\mathrm{Var}\left[W\,|\,Y\right]\right] = \mathrm{Var}\left[\mathbb{E}\left[W\,|\,X\right]\right] + \mathbb{E}\left[\mathrm{Var}\left[W\,|\,X\right]\right]$, and the fact $\mathbb{E}\left[W\,|\,Y\right] = 0$, we can write

$$\begin{aligned}\mathbb{E}\left[g(X,Y)^2\right] &= \mathbb{E}\left[\mathrm{Var}\left[W\,|\,Y\right]\right] - \mathrm{Var}\left[\mathbb{E}\left[W\,|\,X\right]\right] \leq \mathbb{E}\left[\mathrm{Var}\left[W\,|\,Y\right]\right] \\ &\leq \mathbb{E}\left[\mathbb{E}\left[\left(W + \mathbb{E}\left[\mathcal{K}_m(Y,Z)\,|\,Y\right]\right)^2\,\middle|\,Y\right]\right] = \mathbb{E}\left[\mathcal{K}_m(X,Y)^2\right] \ .\end{aligned}$$

Hence, $C^2 \leq n^2(\digamma+1)^2\mathbb{E}\left[\mathcal{K}_m(X,Y)^2\right]$.

Finally, let us bound

$$D = \sup_{(a,b)\in\mathcal{A}} \sum_{1\leq i<j\leq n} \left(\digamma\rho_{i,j}^{(\mathrm{VF})} - 1\right)\mathbb{E}\left[g(X,Y)a_i(X)b_j(Y)\right] \ ,$$

where $\mathcal{A} = \left\{(a,b),\ \text{s.t. } \mathbb{E}\left[\sum_{i=1}^{n-1} a_i(X_i)^2\right] \leq 1,\ \mathbb{E}\left[\sum_{j=2}^{n} b_j(X_j)^2\right] \leq 1\right\}$. Let $(a,b) \in \mathcal{A}$, since $\mathbb{E}\left[\mathcal{K}_m(X,Y)^2\right] \leq \|s\|_\infty \mathcal{D}_m \leq \Upsilon\mathcal{D}_m$, from Cauchy-Schwarz inequality,

$$\begin{aligned}\mathbb{E}\left[a_i(X)b_j(Y)g(X,Y)\right] &\leq \sqrt{\mathbb{E}\left[a_i(X)^2\right]\mathbb{E}\left[b_j(Y)^2\right]\mathbb{E}\left[\mathcal{K}_m(X,Y)^2\right]} \\ &\leq \sqrt{\mathbb{E}\left[a_i(X)^2\right]\mathbb{E}\left[b_j(Y)^2\right]\Upsilon\mathcal{D}_m} \ .\end{aligned}$$

Therefore,

$$\begin{aligned}&\mathbb{E}\left[\sum_{1\leq i<j\leq n} f_{i,j}(X_i,X_j)a_i(X_i)b_j(X_j)\right] \\ &\qquad\leq (\digamma+1)\sqrt{\Upsilon\mathcal{D}_m}\sum_{1\leq i\leq n}\sqrt{\mathbb{E}\left[a_i(X)^2\right]}\sum_{1\leq j\leq n}\sqrt{\mathbb{E}\left[b_j(Y)^2\right]} \ .\end{aligned}$$

Remark by (3.1)

$$\mathbb{E}\left[\mathcal{K}_m(X,x)^2\right] \leq \|s\|_\infty \sup_{z\in\Xi}\int_\Xi \mathcal{K}_m(z,y)^2 d\mu(y) \leq \|s\|_\infty \Gamma n \leq \Upsilon n \ . \tag{3.27}$$

Hence, by (3.18) and (**H1**),

$$A \leq 4(\digamma+1)\Upsilon n, \quad B^2 \leq n^2(\digamma+1)^2\Upsilon, \quad C^2 \leq n^2(\digamma+1)^2\Upsilon\mathcal{D}_m, \quad D \leq n(\digamma+1)\sqrt{\Upsilon\mathcal{D}_m} \ .$$

Thus, Theorem 3.4 gives, for any $x > \log(5.4)$, with probability larger than $1 - 5.4e^{-x}$,

$$\left|\frac{U_m^{(\mathcal{K})}}{2}\right| \leq (2)^{5/2} n(F+1)\sqrt{\Upsilon \mathcal{D}_m x} + 8n(F+1)\sqrt{\Upsilon \mathcal{D}_m} x + 108(\sqrt{\Upsilon} n(F+1)x^{3/2} + 4\Upsilon n(F+1)x^2) \ .$$

Since $x^2 \geq x^{3/2}$ and $x \geq \sqrt{x}$, we conclude the proof observing that $\forall \varepsilon > 0$

$$14n(F+1)\sqrt{\Upsilon \mathcal{D}_m} x \leq n\varepsilon \mathcal{D}_m + \frac{49n(F+1)^2 \Upsilon x^2}{\varepsilon} \ .$$

□

Additional Lemmas

We prove here the concentration of $\delta(m, m') = (P_n - P)(\chi_m - \chi_{m'})$, and $\delta^*(m, m') = (P_n - P)(s_m - s_{m'})$ for all $m, m' \in \mathcal{M}$.

Lemma 3.2. *Suppose that Assumption* (**H3**) *holds and* $\Upsilon \geq \|s\|_\infty$. *There exists an absolute constant* κ *such that for any* $x > \log(2)$, *for any* $m, m' \in \mathcal{M}$, *we have* $\forall \varepsilon > 0$,

$$\mathbb{P}\left(\left|\delta^*(m, m')\right| > \varepsilon \|s_m - s_{m'}\|^2 + \frac{\kappa \Upsilon x^2}{n(\varepsilon \wedge 1)} \right) \leqslant 2e^{-x} \ .$$

Proof: We have

$$\mathrm{Var}\left[(s_m - s_{m'})(X) \right] \leqslant \|s\|_\infty \|s_m - s_{m'}\|^2 \ .$$

Hence, we deduce from Bernstein's inequality that for any $x > \log(2)$, we have

$$\mathbb{P}\left(\left|\delta^*(m, m')\right| > \|s_m - s_{m'}\| \sqrt{\frac{2\|s\|_\infty x}{n}} + \frac{\|s_m - s_{m'}\|_\infty x}{3n} \right) \leq 2e^{-x} \ .$$

Using Assumption (**H3**), we have for any $\varepsilon > 0$

$$\begin{aligned} \frac{\|s_m - s_{m'}\|_\infty x}{3n} &\leq \frac{x}{3n}\left(\Upsilon + \sqrt{\Upsilon n}\, \|s_m - s_{m'}\| \right) \\ &\leq \frac{x\Upsilon}{3n} + \frac{\varepsilon}{2}\|s_m - s_{m'}\|^2 + \frac{x^2 \Upsilon}{18n\varepsilon} \ . \end{aligned}$$

Since $\|s\|_\infty \leq \Upsilon$, we also have for any $\varepsilon > 0$

$$\|s_m - s_{m'}\| \sqrt{\frac{2\|s\|_\infty x}{n}} \leq \frac{\varepsilon}{2}\|s_m - s_{m'}\|^2 + \frac{\Upsilon x}{n\varepsilon} \ .$$

The proof follows plugging these bounds together and from $x^2 \geq x$, since $x \geq \log(2)$. □

Lemma 3.3. *Assume that $\mathcal{K}_m$ satisfies* (3.1) *and suppose that Assumption* (**H2**) *holds. There exists an absolute constant κ such that for any $x > 0$, for any $m, m' \in \mathcal{M}$, we have $\forall \varepsilon > 0$,*

$$\mathbb{P}\left(\delta(m,m') > \varepsilon \left(\mathcal{D}_m + \mathcal{D}_{m'} \right) + \frac{\kappa \Upsilon x}{\varepsilon \wedge 1} \right) \leqslant e^{-x} \ .$$

Proof: Bernstein's inequality provides that, for any $x > 0$, we have

$$\mathbb{P}\left(\delta(m,m') > \sqrt{\frac{2\mathrm{Var}\left[\chi_m(X) - \chi_{m'}(X) \right] x}{n}} + \frac{\left\| \chi_m - \chi_{m'} \right\|_\infty x}{3n} \right) \leqslant e^{-x} \ .$$

Now, from (3.1)

$$\left\| \chi_m - \chi_{m'} \right\|_\infty \leq \left\| \chi_m \right\|_\infty + \left\| \chi_{m'} \right\|_\infty \leq 2\Upsilon n \ ,$$

and using Assumption (**H2**)

$$\begin{aligned} \mathrm{Var}\left[\chi_m(X) - \chi_{m'}(X) \right] &\leq \mathbb{E}\left[\left(\chi_m(X) - \chi_{m'}(X) \right)^2 \right] \\ &\leq 2\left(\mathbb{E}\left[\chi_m(X)^2 \right] + \mathbb{E}\left[\chi_{m'}(X)^2 \right] \right) \\ &\leq 2\Upsilon n \left(\mathcal{D}_m + \mathcal{D}_{m'} \right) \ . \end{aligned}$$

We finally notice that for every $\varepsilon > 0$

$$2\sqrt{\Upsilon \left(\mathcal{D}_m + \mathcal{D}_{m'} \right) x} + \frac{2\Upsilon x}{3} \leq \varepsilon \left(\mathcal{D}_m + \mathcal{D}_{m'} \right) + \frac{5\Upsilon x}{3(\varepsilon \wedge 1)} \ .$$

□

Lemma 3.4. *Let us define*

$$S_m^{(A)} := \sum_{i=1}^n \left(\mathbb{E}\left[A_m(X_i, X) | X_i \right] - \mathbb{E}\left[A_m(X,Y) \right] \right) \ .$$

Assume that $\mathcal{K}_m$ satisfies (3.1) *and suppose that Assumption* (**H4**) *holds. There exists an absolute constant κ such that for any $x > \log(2)$, for any $m \in \mathcal{M}$, we have*

$$\mathbb{P}\left(\left| S_m^{(A)} \right| > \kappa \Upsilon n x \right) \leqslant 2e^{-x} \ .$$

Proof: From Cauchy-Schwarz inequality,

$$\forall (x,x') \in \Xi^2, \qquad A_m(x,x')^2 \leq A_m(x,x) A_m(x',x') \ .$$

In particular, using (3.1)

$$\sup_{x \in \Xi} \mathbb{E}\left[A_m(x,X) \right] \leq \sqrt{\mathcal{D}_m} \sup_{x \in \Xi} \sqrt{A_m(x,x)} \leq \sqrt{\Gamma n \mathcal{D}_m}$$

and, by (**H4**)

$$\mathrm{Var}\left[A_m(X,Y) \right] \leq \mathbb{E}\left[A_m(X,Y)^2 \right] \leq \Upsilon \mathcal{D}_m \ .$$

Since $\mathcal{K}_m$ satisfies (3.1), we can bound $\mathcal{D}_m \leq \Upsilon n$. Therefore, from Bernstein's bound, we have with probability larger than $1 - 2e^{-x}$,

$$\left| S_m^{(A)} \right| \leq \sqrt{2n\Upsilon\mathcal{D}_m x} + \sqrt{\Gamma n \mathcal{D}_m}\frac{x}{3} \leq \left(\sqrt{2} + \frac{1}{3}\right)\Upsilon n x \ .$$

□

Concentration properties of the loss

In this section, we are interested in the $\mathbb{L}_2$-loss of the estimator $\widehat{s}_m$. The goal is to prove the following result.

Proposition 3.4. *Assume that $\mathcal{K}_m$ satisfies* (3.1) *and Assumptions* (**H4**)*,* (**H5**) *and* (**H6**) *hold. Then, there exists an absolute constant κ such that, for any $x > \log(11.4)$, for any $m \in \mathcal{M}$, and $\forall \varepsilon \in (0,1]$,*

$$\mathbb{P}\left(\left| \|\widehat{s}_m - s\|^2 - E_m \right| > \varepsilon E_m + \frac{\kappa \Upsilon x^2}{n\varepsilon^3} \right) \leqslant 11.4 e^{-x} \ ,$$

where $E_m := \|s_m - s\|^2 + \mathcal{D}_m/n = \mathbb{E}\left[\|\widehat{s}_m - s\|^2\right] + \|s_m\|^2/n$.

Proof: We have

$$\|\widehat{s}_m - s\|^2 = \|\widehat{s}_m - s_m\|^2 + \|s_m - s\|^2 + 2\langle \widehat{s}_m - s_m, s_m - s\rangle \ .$$

The expectation of the risk is easily obtained, actually,

$$\begin{aligned}
\mathbb{E}\left[\|\widehat{s}_m - s\|^2\right] &= \mathbb{E}\left[\|\widehat{s}_m - s_m\|^2\right] + \|s_m - s\|^2 \\
&= \|s_m - s\|^2 + \frac{1}{n}\int_\Xi \mathrm{Var}\left[\mathcal{K}_m(X,x)\right] d\mu(x) \\
&= \|s_m - s\|^2 + \frac{\mathcal{D}_m}{n} - \frac{\|s_m\|^2}{n} = E_m - \frac{\|s_m\|^2}{n} \ .
\end{aligned}$$

It follows that we are interested in

$$\begin{aligned}
\|\widehat{s}_m - s\|^2 - E_m = \|\widehat{s}_m - s_m\|^2 - \mathbb{E}\left[\|\widehat{s}_m - s_m\|^2\right] \\
+ 2\langle \widehat{s}_m - s_m, s_m - s\rangle - \frac{\|s_m\|^2}{n} \ . \quad (3.28)
\end{aligned}$$

By Fubbini,

$$\begin{aligned}
\|\widehat{s}_m - s_m\|^2 &= \int_\Xi \left(\frac{1}{n}\sum_{i=1}^n \mathcal{K}_m(X_i,x) - \mathbb{E}\left[\mathcal{K}_m(X,x)\right] \right)^2 d\mu(x) \\
&= \frac{\mathcal{D}_m - \|s_m\|^2}{n} + \frac{1}{n^2}\sum_{i=1}^n \left(A_m(X_i,X_i) - \mathcal{D}_m \right) + \frac{U_m^{(A)}}{n^2} - \frac{2S_m^{(A)}}{n^2} \ ,
\end{aligned}$$

where

$$U_m^{(A)} = \sum_{i\neq j=1}^{n} \Big(A_m(X_i,X_j) - \mathbb{E}\left[A_m(X_i,X_j)|X_i\right] \\ - \mathbb{E}\left[A_m(X_i,X_j)|X_j\right] + \mathbb{E}\left[A_m(X,Y)\right] \Big) \ .$$

By Lemma 3.4, there exists an absolute constant κ such that for any $x > \log(2)$, for any $m \in \mathcal{M}$, we have

$$\mathbb{P}\left(\frac{1}{n^2} \left| S_m^{(A)} \right| > \kappa \frac{\Upsilon x}{n} \right) \leqslant 2e^{-x} \ .$$

By Proposition 3.5, for any $x > \log(2)$,

$$\mathbb{P}\left(\exists \varepsilon > 0, \qquad \left| \frac{1}{n^2} \sum_{i=1}^{n} \left(A_m(X_i,X_i) - \mathcal{D}_m \right) \right| > \varepsilon \frac{\mathcal{D}_m}{n} + \frac{\kappa \Upsilon x}{n(\varepsilon \wedge 1)} \right) \leq 2e^{-x} \ .$$

By Proposition 3.6, for any $x > \log(5.4)$,

$$\mathbb{P}\left(\exists \varepsilon > 0, \qquad \frac{1}{n^2} \left| U_m^{(A)} \right| > \varepsilon \frac{\mathcal{D}_m}{n} + \frac{\kappa \Upsilon x^2}{n(\varepsilon \wedge 1)} \right) \leq 5.4e^{-x} \ .$$

Hence, for any $x > \log(9.4)$, for any $m \in \mathcal{M}$, with probability larger than $1 - 9.4e^{-x}$, for any $\varepsilon \in (0,1]$,

$$\left| \left\| \widehat{s}_m - s_m \right\|^2 - \mathbb{E}\left[\left\| \widehat{s}_m - s_m \right\|^2 \right] \right| \leq \varepsilon \frac{\mathcal{D}_m}{n} + \frac{\kappa \Upsilon x^2}{n\varepsilon} \ . \tag{3.29}$$

In order to complete the concentration of the risk, from (3.28), it remains to concentrate the term $\langle \widehat{s}_m - s_m, s_m - s \rangle$. Setting $\xi_m(y) = \int_\Xi \mathcal{K}_m(y,x)(s_m(x) - s(x))d\mu(x)$, this term can be written

$$\langle \widehat{s}_m - s_m, s_m - s \rangle = (P_n - P)\xi_m \ .$$

By Assumption (**H6**), since $\xi_m / \left\| s_m - s \right\| \in \mathbb{B}_m$,

$$\mathrm{Var}\left[\xi_m(X)\right] \leqslant \left\| s_m - s \right\|^2 v_m^2 \leqslant \left\| s_m - s \right\|^2 \left(\Upsilon \vee \sqrt{\Upsilon \mathcal{D}_m} \right) \ .$$

Moreover from (3.1)

$$\sup_{y\in\Xi} |\xi_m(y)| \leqslant \left\| s_m - s \right\| \sup_{y\in\Xi} \int_\Xi \mathcal{K}_m(y,x) \frac{|s_m(x) - s(x)|}{\left\| s_m - s \right\|} d\mu(x) \\ \leqslant \left\| s_m - s \right\| \sqrt{\sup_{y\in\mathbb{X}} \int_\Xi \mathcal{K}_m(y,x)^2 d\mu(x)} \leq \left\| s_m - s \right\| \sqrt{\Upsilon n} \ .$$

In addition, we have for any $\theta > 0$

$$2\frac{\left\| s_m - s \right\| x}{3} \sqrt{\frac{\Upsilon}{n}} \leq \theta \left\| s_m - s \right\|^2 + \frac{x^2 \Upsilon}{9n\theta} \ ,$$

and for any $\theta_1, \theta_2 > 0$

$$2\left\|s_m - s\right\| \sqrt{\frac{2x\left(\Upsilon + \sqrt{\Upsilon \mathcal{D}_m}\right)}{n}} \leq \theta_1 \left\|s_m - s\right\|^2 + \frac{2x\Upsilon}{n\theta_1} + \frac{2x\sqrt{\Upsilon \mathcal{D}_m}}{n\theta_1}$$
$$\leq \theta_1 \left\|s_m - s\right\|^2 + \frac{2x\Upsilon}{n\theta_1} + \theta_2 \frac{\mathcal{D}_m}{n} + \frac{x^2\Upsilon}{n\theta_1^2\theta_2} \ .$$

Hence, it comes from Bernstein's inequality that, for some constant $\kappa > 0$,

$$\mathbb{P}\left(2\left|\langle \widehat{s}_m - s_m, s_m - s\rangle\right| > \varepsilon \left\|s_m - s\right\|^2 + \varepsilon \frac{\mathcal{D}_m}{n} + \frac{\kappa\Upsilon x^2}{n(\varepsilon \wedge 1)^3}\right) \leqslant 2e^{-x} \ .$$

Combining this inequality with (3.29) concludes the proof. □

Concentration results for the risk and the classical V-fold criterion

From Lemma 3.1, it is sufficient to get the concentration of

$$\frac{1}{n}\sum_{i=1}^{n} A_m(X_i, X_i) \quad \text{and} \quad \sum_{1 \leqslant i \neq j \leqslant n} \rho_{i,j}^{(\mathrm{VF})} A_m(X_i, X_j) \ .$$

We begin by the concentration of the empirical mean $n^{-1}\sum_{i=1}^{n} A_m(X_i, X_i)$.

Proposition 3.5. *For any $x > \log(2)$, for any $m \in \mathcal{M}$,*

$$\mathbb{P}\left(\exists \varepsilon > 0, \qquad \left|\frac{1}{n}\sum_{i=1}^{n} A_m(X_i, X_i) - \mathcal{D}_m\right| > \varepsilon \mathcal{D}_m + \frac{\Gamma x}{\varepsilon \wedge 1}\right) \leq 2e^{-x} \ .$$

Proof: Remark that $\mathcal{D}_m = \mathbb{E}\left[A_m(X, X)\right]$, hence, Bernstein's bound gives $\forall x > \log(2)$, with probability larger than $1 - 2e^{-x}$,

$$\left|\frac{1}{n}\sum_{i=1}^{n} A_m(X_i, X_i) - \mathcal{D}_m\right| \leq \sqrt{\frac{2}{n}\mathrm{Var}\left[A_m(X, X)\right] x} + \frac{\sup_{y \in \Xi} A_m(y, y) x}{3n} \ .$$

Now, by definition of Γ, $0 \leq A_m(y, y) = \int_{\Xi} \mathcal{K}_m(y, z)^2 d\mu(z) \leq \Gamma n$ and

$$\mathrm{Var}\left[A_m(X, X)\right] \leq \mathbb{E}\left[A_m(X, X)^2\right] \leq \Gamma n \mathbb{E}\left[A_m(X, X)\right] = \Gamma n \mathcal{D}_m \ .$$

Hence,

$$\forall \varepsilon > 0, \qquad \sqrt{\frac{2}{n}\mathrm{Var}\left[A_m(X, X)\right] x} \leq \varepsilon \mathcal{D}_m + \frac{\Gamma x}{2\varepsilon} \ .$$

□

Proposition 3.6. *Assume that $\mathcal{K}_m$ satisfies* (3.1) *and Assumptions* (**H4**) *and* (**H5**) *hold. Then, there exists an absolute constant κ such that, for any $x > \log(7.4)$, for any $m \in \mathcal{M}$, with probability*

larger than $1 - 7.4e^{-x}$ *for any* $\varepsilon > 0$*,*

$$\left| \sum_{1 \leqslant i \neq j \leqslant n} \rho_{i,j}^{(\mathrm{VF})} A_m(X_i, X_j) \right| \leq n \left(\varepsilon \mathcal{D}_m + \frac{\kappa \Upsilon x^2}{\varepsilon \wedge 1} \right) \ .$$

Moreover,

$$\mathbb{P}\left(\exists \varepsilon > 0, \qquad \left| U_m^{(A)} \right| > n \left(\varepsilon \mathcal{D}_m + \frac{\kappa \Upsilon x^2}{\varepsilon \wedge 1} \right) \right) \leq 5.4 e^{-x} \ .$$

Proof: The proof follows the same lines as the one of Proposition 3.3, with $\mathcal{K}_m$ being replaced by A_m. We introduce the totally degenerate U-statistic

$$U_m^{(A,\rho)} := \sum_{1 \leqslant i \neq j \leqslant n} \rho_{i,j}^{(\mathrm{VF})} (A_m(X_i, X_j) - \mathbb{E}\left[A_m(X_i, X) | X_i \right] \\ - \mathbb{E}\left[A_m(X, X_j) | X_j \right] + \mathbb{E}\left[A_m(X, Y) \right]) \ . \quad (3.30)$$

Now, since $\mathbb{E}\left[A_m(X, Y) \right] = \|s_m\|^2$, the following Hoeffding's decomposition holds

$$\sum_{1 \leqslant i \neq j \leqslant n} \rho_{i,j}^{(\mathrm{VF})} A_m(X_i, X_j) = U_m^{(A,\rho)} - 2 S_m^{(A)} - n \|s_m\|^2 \ .$$

From Lemma 3.4 we have with probability larger than $1 - 2e^{-x}$, $\left| S_m^{(A)} \right| \leq \kappa \Upsilon n x$.

Concerning the U-statistics, we first write $\forall 1 \leqslant i \neq j \leq n$, $\forall (x, y) \in \Xi^2$, let $f_{i,j}(x, y) = \rho_{i,j}^{(\mathrm{VF})} g(x, y)$, where

$$g(x, y) = A_m(x, y) - \mathbb{E}\left[A_m(x, X) \right] - \mathbb{E}\left[A_m(y, X) \right] + \mathbb{E}\left[A_m(X, Y) \right] \ .$$

Let us find upper bounds for the different terms involved in Theorem 3.4. Using the notations A, B, C, D (respectively $A^\rho, B^\rho, C^\rho, D^\rho$) of this theorem, for $U^{(A)}$ (respectively $U^{(A,\rho)}$),

$$A \vee A^\rho \leq 4 \sup_{(x,y) \in \Xi^2} |A_m(x, y)| = 4 \sup_{x \in \Xi} |A_m(x, x)| \leq 4 \Gamma n \ . \quad (3.31)$$

The last equality comes from $A_m(x, y) \leq \sqrt{A_m(x, x) A_m(y, y)}$. Next, since, for any $i \in [n]$, $\sum_{j=1}^n (\rho_{i,j}^{(\mathrm{VF})})^2 = n/(V-1)$, by (**H5**),

$$B^2 \leq n \sup_{x \in \Xi} \mathbb{E}\left[A_m(x, X)^2 \right] \leq \Upsilon n, \ (B^\rho)^2 \leq \frac{n}{V-1} \sup_{x \in \Xi} \mathbb{E}\left[A_m(x, X)^2 \right] \leq \frac{\Upsilon n^2}{V-1} \ . \quad (3.32)$$

By (**H4**), $\mathbb{E}\left[g(X, Y)^2 \right] \leq \mathbb{E}\left[A_m(X, Y)^2 \right] \leq \Upsilon \mathcal{D}_m$, hence $C^2 \leq \Upsilon n^2 \mathcal{D}_m$. It follows from easy computations that

$$\sum_{1 \leq i \neq j \leq n} \left(\rho_{i,j}^{(\mathrm{VF})} \right)^2 = \frac{n(n - V + 1)}{V - 1} \ .$$

Hence,

$$(C^\rho)^2 = \mathbb{E}\left[g(X,Y)^2\right] \sum_{1\le i\neq j\le n} \left(\rho_{i,j}^{(\mathrm{VF})}\right)^2 \le \frac{n(n-V+1)}{V-1}\Upsilon\mathcal{D}_m \ . \tag{3.33}$$

Let a and b be sequences of real valued functions such that

$$\sum_{i=1}^n \mathbb{E}\left[a_i(X)^2\right] \le 1 \quad \text{and} \quad \sum_{j=1}^n \mathbb{E}\left[b_j(X)^2\right] \le 1 \ .$$

From Cauchy-Schwarz inequality, Assumption (**H4**) and using in addition that $\left|\rho_{i,j}^{(\mathrm{VF})}\right| \leqslant 1$, we obtain

$$\mathbb{E}\left[\sum_{1\le i<j\le n} f_{i,j}(X_i,X_j)a_i(X_i)b_j(X_j)\right] \vee \mathbb{E}\left[\sum_{1\le i<j\le n} g(X_i,X_j)a_i(X_i)b_j(X_j)\right]$$
$$\le \sqrt{\Upsilon\mathcal{D}_m}\sum_{1\le i\le n}\sqrt{\mathbb{E}\left[a_i(X)^2\right]}\sum_{1\le j\le n}\sqrt{\mathbb{E}\left[b_j(Y)^2\right]} \le n\sqrt{\Upsilon\mathcal{D}_m} \ .$$

Thus,

$$D \vee D^\rho \le n\sqrt{\Upsilon\mathcal{D}_m} \ . \tag{3.34}$$

Plugging (3.31), (3.32), (3.33) and (3.34) in Theorem 3.4 implies that one can find an absolute constant $\kappa > 0$ such that, for any $x > 1$, with probability larger than $1 - 5.4e^{-x}$, each of these inequalities holds : for any $\varepsilon > 0$

$$\left|U_m^{(A)}\right| \le n\left(\varepsilon\mathcal{D}_m + \frac{\kappa\Upsilon x^2}{\varepsilon \wedge 1}\right), \qquad \left|U_m^{(A,\rho)}\right| \le n\left(\varepsilon\mathcal{D}_m + \frac{\kappa\Upsilon x^2}{\varepsilon \wedge 1}\right) \ .$$

□

3.7.3 Oracle inequalities

Proof of Theorem 3.1

To simplify the notation we note $\widetilde{s} = \widetilde{s}_{V,F}^{\mathrm{pen}}$ and $\widehat{m} = \widehat{m}_{V,F}^{\mathrm{pen}}$ in this proof. It follows from (3.13) and classical computations that

$$\|\widetilde{s} - s\|^2 \le \|\widehat{s}_m - s\|^2 + \left(F\,\mathrm{pen}_{\mathrm{VF}}(m,V) - \mathrm{pen}_{\mathrm{id}}(m)\right)$$
$$- \left(F\,\mathrm{pen}_{\mathrm{VF}}(\widehat{m},V) - \mathrm{pen}_{\mathrm{id}}(\widehat{m})\right) \ . \tag{3.35}$$

Now, remark that (**H3**) ensures

$$\left|\mathbb{E}\left[\mathcal{K}_m(X,Y)\right] - \mathbb{E}\left[\mathcal{K}_{m'}(X,Y)\right]\right| \le \Upsilon \ .$$

Thus, we obtain from (3.35) and (3.26), for any $m \in \mathcal{M}$,

$$\|\widetilde{s}-s\|^2 \leq \|\widehat{s}_m - s\|^2 + \frac{2\left(U_m^{(\mathcal{K})} - U_{\widehat{m}}^{(\mathcal{K})}\right)}{n^2} + \frac{2(\digamma-1)}{n}\left(\gamma_m \mathcal{D}_m - \gamma_{\widehat{m}}\mathcal{D}_{\widehat{m}}\right) - \left(2 + \frac{4(\digamma-1)}{n}\right)\delta^*(m,\widehat{m}) + \frac{2(\digamma-1)}{n}\delta(m,\widehat{m}) + \frac{2(\digamma-1)\Upsilon}{n} \quad . \quad (3.36)$$

Let $(w_m)_{m\in\mathcal{M}}$ denote a set of real numbers larger than 1 such that $\sum_{m\in\mathcal{M}} e^{-w_m} \leqslant 1$. Let $\varepsilon, x > 0$ and

$$\Omega_{U_m^{(\mathcal{K})},\varepsilon}(x) := \bigcap_{m\in\mathcal{M}} \left\{ \left| \frac{2U_m^{(\mathcal{K})}}{n^2} \right| \leq \varepsilon \frac{\mathcal{D}_m}{n} + \frac{\kappa\Upsilon(\digamma+1)^2(w_m+x)^2}{n(\varepsilon\wedge 1)} \right\} \quad ,$$

$$\Omega_{\delta,\varepsilon}(x) := \bigcap_{m,m'\in\mathcal{M}} \left\{ \delta\left(m,m'\right) \leqslant \varepsilon\left(\mathcal{D}_m + \mathcal{D}_{m'}\right) + \frac{\kappa\Upsilon(w_m \vee w_{m'} + x)^2}{\varepsilon\wedge 1} \right\} \quad ,$$

$$\Omega_{\delta^*,\varepsilon}(x) := \bigcap_{m,m'\in\mathcal{M}} \left\{ \left|\delta^*\left(m,m'\right)\right| \leqslant \varepsilon \left\|s_m - s_{m'}\right\|^2 + \frac{\kappa\Upsilon(w_m \vee w_{m'} + x)^2}{n(\varepsilon\wedge 1)} \right\} \quad ,$$

$$\Omega_{r,\varepsilon}(x) := \bigcap_{m\in\mathcal{M}} \left\{ \left| \|\widehat{s}_m - s_m\|^2 - E_m \right| \leqslant \varepsilon E_m + \frac{\kappa\Upsilon(w_m+x)^2}{n\varepsilon^3} \right\} \quad .$$

A union bound in Lemma 3.2, Lemma 3.3, Proposition 3.3 and Proposition 3.4 gives that there exists an absolute constant κ such that, for any $\varepsilon > 0$ and $x > \log(19.8)$, $\mathbb{P}\left(\Omega_{U_m^{(\mathcal{K})},\varepsilon}(x)\right) \geqslant 1 - e^{-x}$, $\mathbb{P}\left(\Omega_{\delta,\varepsilon}(x)\right) \geqslant 1 - e^{-x}$, $\mathbb{P}\left(\Omega_{\delta^*,\varepsilon}(x)\right) \geq 1 - e^{-x}$ and $\mathbb{P}\left(\Omega_{r,\varepsilon}(x)\right) \geq 1 - e^{-x}$. Let us denote by

$$\Omega_{good,\varepsilon}(x) = \Omega_{\delta,\varepsilon}(x) \cap \Omega_{\delta^*,\varepsilon}(x) \cap \Omega_{U_m^{(\mathcal{K})},\varepsilon}(x) \cap \Omega_{r,\varepsilon}(x) \quad .$$

A union bound gives that there exists an absolute constant κ such that for any $\varepsilon > 0$ and $x \geq \log(19.8)$, $\mathbb{P}\left(\Omega_{good,\varepsilon}(x)\right) \geq 1 - e^{-x}$. Moreover, on $\Omega_{good,\varepsilon}(x)$ it comes from (3.36) that, for any $m \in \mathcal{M}$,

$$\begin{aligned}\|\widetilde{s}-s\|^2 \leq \|\widehat{s}_m - s\|^2 &+ \left(2 + \frac{4(\digamma-1)_+}{n}\right)\varepsilon\left\|s_m - s_{\widehat{m}}\right\|^2 \\ &+ \frac{\mathcal{D}_m}{n}\left(2(\digamma-1)_+\gamma_m + (2\digamma-1)_+\varepsilon\right) + \frac{\mathcal{D}_{\widehat{m}}}{n}\left(2(1-\digamma)_+\gamma_{\widehat{m}} + (2\digamma-1)_+\varepsilon\right) \\ &+ \left(2 + (\digamma+1)^2 + 4(\digamma-1)_+ + \frac{4(\digamma-1)_+}{n}\right)\frac{\kappa\Upsilon(w_m \vee w_{\widehat{m}} + x)^2}{n(\varepsilon\wedge 1)} \quad .\end{aligned}$$

Now, we notice that

$$\left\|s_m - s_{\widehat{m}}\right\|^2 \leq 2\left(\|s_m - s\|^2 + \|s - s_{\widehat{m}}\|^2\right) = 2\left(E_m - \frac{\mathcal{D}_m}{n} + E_{\widehat{m}} - \frac{\mathcal{D}_{\widehat{m}}}{n}\right) \quad .$$

Hence, we deduce from the inequality above and from Proposition 3.4 that there exists an absolute constante $\kappa > 0$ such that for any $x \geq \log(19.8)$, with probability larger than $1 - 19.8e^{-x}$, for any $\varepsilon > 0$

$$\left(1 - 2(\digamma-1)_-\gamma_{\widehat{m}} - \varepsilon\right)\|\widetilde{s}-s\|^2$$

$$\leqslant (1 + 2(F-1)_+\gamma_m + \varepsilon)\,\|\widehat{s}_m - s\|^2 + \kappa(1 \vee F^2)\Upsilon\frac{(w_m \vee w_{\widehat{m}} + x)^2}{n\varepsilon^3} \ .$$

Proof of Theorem 3.2

From Lemma 3.1,

$$\mathcal{C}_V^{\mathrm{VFCV}}(m) - \mathcal{C}_V^{\mathrm{corr,VFCV}}(m) - \frac{\mathcal{D}_m}{n(V-1)}$$
$$= \frac{1}{n^2(V-1)}\left(\sum_{i=1}^{n}(\,A_m(X_i,X_i) - \mathcal{D}_m\,) + \sum_{1\leqslant i\neq j\leqslant n}\rho_{i,j}^{(\mathrm{VF})}A_m(X_i,X_j)\right) \quad (3.37)$$

We deduce that, on the event $\Omega_{good,\varepsilon}(x)$ considered in the proof of Theorem 3.1,

$$\left|\mathcal{C}_V^{\mathrm{VFCV}}(m) - \mathcal{C}_{1,V}(m) - \frac{\mathcal{D}_m}{n(V-1)}\right| \leq \varepsilon\frac{\mathcal{D}_m}{n} + \left(1+\frac{1}{\varepsilon}\right)\frac{\kappa\Upsilon x^2}{n} \ .$$

The proof then follows the same lines as the one of Theorem 3.1. We proceed in the same way to get the result on the leave-p-out estimator.

3.8 Adaptation over Sobolev ellipsoids

In this section, we remind that $\Xi = [0,1]$ and μ is the Lesbegue measure.

3.8.1 Proof of Proposition 3.2

To prove this result we first remind the lower bound from Golubev (1992) proved in Dalelane (2005a).

Proposition 3.7. *For any $\beta > 1/2$, $Q > 0$,*

$$\liminf_{n\to\infty}\ \inf_{\widehat{s}_n}\ \sup_{s\in\mathcal{S}'(\beta,Q)}\ n^{\frac{2\beta}{2\beta+1}}\,\mathbb{E}\left[\,\|s-\widehat{s}_n\|^2\right] \geq \widetilde{C}_{\beta,Q} \ ,$$

where $\widetilde{C}_{\beta,Q}$ is Pinsker's constant and the infimum is taken over all estimators of s.

We show here the upper bound, that essentially prove Pinsker estimators are sharp minimax over Sobolev classes.

Proposition 3.8. *For any $\beta > 1/2$, $Q > 0$,*

$$\sup_{s\in\mathcal{S}'(\beta,Q)} R\left(\tau^{\beta,Q}\right) \leq n^{\frac{-2\beta}{2\beta+1}}\widetilde{C}_{\beta,Q}(1+\mathrm{o}(1)) \quad n\to\infty \ .$$

Proof: First notice that for generic Pinsker's weights $\tau_j = \left(1 - rj^\beta n^{-\beta/(2\beta+1)}\right)_+$,

$$\tau_j = 0 \quad \text{for any} \quad j > N_{r,\beta} = \left\lceil\frac{n^{1/(2\beta+1)}}{r^{1/\beta}}\right\rceil, \quad \text{and} \quad N_{r,\beta}\to\infty \ .$$

It follows from (3.21) that the risk of such an estimator can be written

$$R(\tau) = \sum_{j=1}^{N_{r,\beta}} \left(\frac{\tau_j^2}{n} + \left((1-\tau_j)^2 - \frac{\tau_j^2}{n} \right) \theta_j^2 \right) + \sum_{j=N_{r,\beta}+1}^{\infty} \theta_j^2 \ .$$

Moreover, from the definition of the Sobolev ellipsoid, we have

$$\sum_{j=1}^{N_{r,\beta}} \left(\frac{rj^\beta}{n^{\beta/(2\beta+1)}} \right)^2 \theta_j^2 + \sum_{j=N_{r,\beta}+1}^{\infty} \theta_j^2 \leq \frac{r^2}{n^{2\beta/(2\beta+1)}} \sum_{j=1}^{N_{r,\beta}} a_j^2 \theta_j^2 + a_{N_{r,\beta}+1}^{-2} \sum_{j=N_{r,\beta}+1}^{\infty} a_j^2 \theta_j^2 \leq r^2 \frac{Q}{n^{2\beta/(2\beta+1)}} \ .$$

Since for any $a > 0$

$$\sum_{j=1}^{M} j^a = \frac{M^{a+1}}{a+1} \left(1 + O_a(1/M)\right) \quad \text{as } M \to \infty \ ,$$

we get

$$\begin{aligned} R(\tau) &\leq \frac{1}{n} \sum_{j=1}^{N_{r,\beta}} \left(1 - r \frac{j^\beta}{n^{\beta/(2\beta+1)}} \right)^2 + r^2 \frac{Q}{n^{2\beta/(2\beta+1)}} \\ &= \frac{1}{n^{2\beta/(2\beta+1)}} \left(\frac{1}{r^{1/\beta}} \left(1 - \frac{2}{\beta+1} + \frac{1}{2\beta+1} \right) + r^2 Q + O\left(\frac{1}{n^{1/(2\beta+1)}} \right) \right) \\ &= \frac{1}{n^{2\beta/(2\beta+1)}} \left(\frac{2\beta^2}{(\beta+1)(2\beta+1) r^{1/\beta}} + r^2 Q + O\left(\frac{1}{n^{1/(2\beta+1)}} \right) \right) \ . \end{aligned}$$

This last bound is optimized for $r = r^{\beta,Q}$ given in (3.4) and yields the following risk bound for Pinsker's estimators.

$$R\left(\tau^{\beta,Q}\right) \leqslant \widetilde{C}_{\beta,Q} \left(1 + \mathrm{O}(n^{-1/(1+2\beta)})\right) n^{\frac{-2\beta}{2\beta+1}} \ .$$

□

3.8.2 Proof of Corollary 3.1

Let us assume $n \geq 6(\log n)^2$ and $s \in \mathcal{S}'(\beta^*, Q^*)$ for some $\beta^* \in (1/2, n)$ and $Q^* \in (0, (\log n)^2)$. Let us consider the following set of weights

$$\mathcal{M} = \left\{ \tau^{\beta_j, Q_k},\ (j,k) \in \mathcal{A} \right\} \quad \text{with} \quad \mathcal{A} = (1, \ldots, n^2) \times (1, \ldots, n(\log n)^2) \ ,$$

with $\beta_j = 1/2 + j/n$ and $Q_k = k/n$. First let us notice since $\beta > 1/2$ and $Q \leq (\log n)^2$

$$N_{\beta,Q} = N_{r^{\beta,Q},\beta} = \left\lceil \left(\frac{n(\beta+1)(2\beta+1)Q}{\beta} \right)^{\frac{1}{2\beta+1}} \right\rceil$$
$$\leq \left(3n(\log n)^2(2\beta+1)\right)^{\frac{1}{2\beta+1}} \leq \sqrt{6n(\log n)^2} \leq n \ ,$$

which implies that

$$\Theta(x) = \sum_{j=1}^{2\max_{(j,k)\in\mathcal{A}} N_{\beta_j,Q_k}} \varphi_j(x)^2 \leq n, \qquad \text{hence} \quad \Gamma = 1 \ .$$

We want to verify all the assumptions of Section 3.4.1, so we need to compute a constant Υ such that

$$\Upsilon \geq 1 + \|s\|_\infty \ .$$

Let us bound the sup-norm of s. For any $x \in [0,1]$,

$$s(x) = 1 + \sum_{j=1}^{\infty} \left\lceil \frac{j}{2} \right\rceil^{\beta} (P\varphi_j) \frac{\varphi_j(x)}{\left\lceil \frac{j}{2} \right\rceil^{\beta}}$$
$$\leq 1 + \sqrt{\sum_{j=1}^{\infty} \left\lceil \frac{j}{2} \right\rceil^{2\beta} (P\varphi_j)^2} \sqrt{\sum_{j=1}^{\infty} \frac{\varphi_j(x)^2}{\left\lceil \frac{j}{2} \right\rceil^{2\beta}}}$$
$$\leq 1 + \sqrt{Q\left(1 + \frac{1}{2\beta+1}\right)} \leq 2\log n \ .$$

The last inequality holds since $n \geq 4$ implies

$$1 + \sqrt{3/2}\log n \leq 2\log n \ .$$

Hence, $\Upsilon = 3\log n$ is large enough. For any $\tau \in \mathcal{M}$, choose $w_\tau = \log(|\mathcal{A}|) \leq 4\log n$ and let $F = 1$. From Theorem 3.1, there exists an absolute constant κ such that, for any $\varepsilon > 0$ and any $x \geq \log(17.8)$, with probability larger than $1 - e^{-x}$, for all $\tau \in \mathcal{M}$

$$\frac{1-\varepsilon}{1+\varepsilon} \|\widetilde{s} - s\|^2 \leq \|\widehat{s}_\tau - s\|^2 + \kappa \log n \frac{(\log n + x)^3}{n(\varepsilon \wedge 1)^3} \ .$$

This bound can be integrated and yields for all $\tau \in \mathcal{M}$

$$\mathbb{E}\left[\|\widetilde{s} - s\|^2\right] \leq \frac{1+\varepsilon}{1-\varepsilon} R(\tau) + \frac{\kappa(\log n)^4}{n(\varepsilon \wedge 1)^3} \ .$$

In particular, taking $\tau = \tau^{\beta,Q}$ we have as a consequence of Proposition 3.8

$$\sup_{s\in\mathcal{S}'(\beta,Q)} \mathbb{E}\left[\|\widetilde{s} - s\|^2\right]$$

$$\leq \frac{1+\varepsilon}{1-\varepsilon}\, n^{\frac{-2\beta}{2\beta+1}} \widetilde{C}_{\beta,Q}(1+\mathrm{O}(1/n^{1/(2\beta+1)})) + \frac{\kappa(\log n)^4}{n(\varepsilon\wedge 1)^3} \ .$$

Optimizing the right-hand side in ε, we get

$$\sup_{s\in\mathcal{S}'(\beta,Q)} \mathbb{E}\left[\|s-\widetilde{s}\|^2\right] \leq \widetilde{C}_{\beta,Q}\, n^{-2\beta/(2\beta+1)} \left(1 + C_Q \frac{\log n}{n^{1/(8\beta+4)}}\right) \ .$$

Since $s \in \mathcal{S}'(\beta^*, Q^*)$, there exists a pair $(j,k) \in \mathcal{A}$ such that $\beta_j \leq \beta^* < \beta_{j+1}$ and $Q_k \leq Q^* < Q_{k+1}$. Therefore $\mathcal{S}'(\beta_{j+1}, Q_k) \subset \mathcal{S}'(\beta^*, Q^*) \subset \mathcal{S}'(\beta_j, Q_{k+1})$. In particular, s also belongs to $\mathcal{S}'(\beta_j, Q_{k+1})$. Thus

$$\begin{aligned}
\mathbb{E}\left[\|s-\widetilde{s}\|^2\right] &\leq \widetilde{C}_{\beta_j,Q_{k+1}}\, n^{-2\beta_j/(2\beta_j+1)} \left(1 + C_Q \frac{\log n}{n^{1/(8\beta+4)}}\right) \\
&= \mathcal{R}_{\mathrm{minmax}}(\beta^*, Q^*)\, C_n(\beta^*, \beta_j) \frac{\widetilde{C}_{\beta_j,Q_{k+1}}}{\widetilde{C}_{\beta^*,Q^*}} \left(1 + C_Q \frac{\log n}{n^{1/(8\beta+4)}}\right) \ ,
\end{aligned}$$

where $C_n(\beta^*, \beta_j) = (n)^{\frac{2\beta^*}{2\beta^*+1} - \frac{2\beta_j}{2\beta_j+1}}$ and

$$\begin{aligned}
&\frac{\widetilde{C}_{\beta_j,Q_{k+1}}}{\widetilde{C}_{\beta^*,Q^*}} \\
&\qquad = \frac{(2\beta_j+1)^{1/(2\beta_j+1)}}{(2\beta^*+1)^{1/(2\beta^*+1)}} \left(\frac{\beta_j}{\beta_j+1}\right)^{2\beta_j/(2\beta_j+1)} \left(\frac{\beta^*+1}{\beta^*}\right)^{2\beta^*/(2\beta^*+1)} \frac{Q_{k+1}^{\frac{1}{2\beta_j+1}}}{(Q^*)^{\frac{1}{2\beta^*+1}}} \ .
\end{aligned}$$

Since $0 \leq Q_{k+1} - Q^* \leq 1/n$

$$\frac{Q_{k+1}^{\frac{1}{2\beta_j+1}}}{(Q^*)^{\frac{1}{2\beta^*+1}}} \leq \frac{\left(Q^*+n^{-1}\right)^{\frac{1}{2\beta_j+1}}}{(Q^*)^{\frac{1}{2\beta_{j+1}+1}}} = \frac{\left(Q^*+n^{-1}\right)^{\frac{1}{2\beta_{j+1}+1-2/n}}}{(Q^*)^{\frac{1}{2\beta_{j+1}+1}}} = 1+\mathrm{O}(1/n) \ .$$

Moreover, observing that $x \mapsto (2x+1)^{1/(2x+1)} \left(\frac{x}{x+1}\right)^{2x/(2x+1)}$ is nondecreasing, we conclude $\widetilde{C}_{\beta_j,Q_{k+1}} \leq \widetilde{C}_{\beta^*,Q^*}(1+\mathrm{O}(1/n))$. Finally, since $0 \leq \beta^* - \beta_j \leq 1/n$, we also have

$$C_n(\beta^*, \beta_j) \leq e^{\log n/(2n)} = 1 + \mathrm{O}(\log n/n) \ .$$

Therefore, we have obtained that, for any $(\beta, Q) \in (1/2, n) \times (0, (\log n)^2)$,

$$\sup_{s\in\mathcal{S}'(\beta,Q)} \mathbb{E}\left[\|s-\widetilde{s}\|^2\right] \leq \widetilde{C}_{\beta,Q}\, n^{-2\beta/(2\beta+1)} \left(1 + C_Q \frac{\log n}{n^{1/(8\beta+4)}}\right) \ .$$

3.9 Proof of Corollary 3.2

The link between estimator (or model) selection and adaptive estimation is made through approximation theory. We shall therefore approximate the risk $R(\tau^{\mathcal{S}_{ni}})$ thanks to two successive subsets of $\mathcal{S}_{ni}$. Let $B = (B_j)_{j\geq 0}$ be the sequence of intervals $B_j = [u_j, v_j]$ with $u_j, v_j \in \mathbb{N}^*$, $u_0 = 1$ and $u_{j+1} = v_j + 1$. Moreover for $n \geq 10$ and $\eta = (\log n)^{-1/2}$, we assume that the lengths of the B_j satisfy $|B_0| = 1 + \lceil 2 \log n \rceil$ and for any $j \geq 0$

$$|B_{j+1}| = \lfloor (1+\eta)|B_j| \rfloor \quad .$$

Furthermore, let $I \in \mathbb{N}^*$ denote the smallest integer such that $v_{I-1} \geq n(\log n)^2$. Let $\mathcal{S}_{B,I} \subset \mathcal{S}_{ni}$ denote the subset of sequences constant on each B_j, i.e.

$$\mathcal{S}_{B,I} := \left\{ \tau \in \mathcal{S}_{ni} \text{ s.t. } \forall j \in [0, I-1],\ \forall k \neq k' \in B_j,\ \tau_k = \tau_{k'}, \text{ and } \forall k > v_{I-1},\ \tau_k = 0 \right\} \quad .$$

The following lemma gathers useful results borrowed from Rigollet (2006a).

Lemma 3.5. *Let $\eta = (\log n)^{-1/2}$ and $n \geq 10$. For any $\tau \in \mathcal{S}_{ni}$, there exists $\overline{\tau} \in \mathcal{S}_{B,I}$ such that*

$$R\left(\overline{\tau}\right) \leq (1+\eta) \sum_{j=1}^{v_{I-1}} \left(\frac{\tau_j^2(1-\theta_j^2)}{n} + (1-\tau_j)^2\theta_j^2 \right) + \sum_{j > v_{I-1}} \theta_j^2 + \frac{|B_0| + \eta \left\|s\right\|^2}{n} \quad .$$

In particular, for any s such that $\tau_{u_I}^{\mathcal{S}_{ni}} < 1 - 1/\sqrt{1+\eta}$,

$$R\left(\tau^{\mathcal{S}_{B,I}}\right) \leq (1+\eta) R\left(\tau^{\mathcal{S}_{ni}}\right) + \frac{|B_0| + \eta \left\|s\right\|^2}{n} \quad .$$

This last condition holds for example if $n \geq n_0$ for some absolute constant n_0, for any s such that

$$\left\|s\right\| \leq \frac{1}{3}\left(1 + \sqrt{\log n}\right) \quad .$$

Proof: Let $\tau \in \mathcal{S}_{ni}$, for any $i \in [\![0, I-1]\!]$ and $k \in B_i$, let $\overline{\tau_k} = \tau_{u_i}$ and for any $k > v_{I-1}$, let $\overline{\tau_k} = 0$. Then by definition $\overline{\tau} \in \mathcal{S}_{B,I}$ and $\overline{\tau_j} \geq \tau_j$ for any $j \in \cup_{i=0}^{I-1} B_i$. By (3.21) we have

$$\begin{aligned} R(\overline{\tau}) &= \sum_{i=0}^{I-1} \sum_{j \in B_i} \left(\frac{\overline{\tau_j}^2}{n} + \left((1-\overline{\tau_j})^2 - \frac{\overline{\tau_j}^2}{n} \right) \theta_j^2 \right) + \sum_{j > v_{I-1}} \theta_j^2 \\ &\leq \sum_{i=0}^{I-1} \sum_{j \in B_i} \frac{\overline{\tau_j}^2}{n} + \sum_{i=0}^{I-1} \sum_{j \in B_i} \left((1-\tau_j)^2 - \frac{\tau_j^2}{n} \right) \theta_j^2 + \sum_{j > v_{I-1}} \theta_j^2 \quad . \end{aligned}$$

Now, since $\overline{\tau_j} \in [0,1]$, we have

$$\sum_{i=0}^{I-1} \sum_{j \in B_i} \frac{\overline{\tau_j}^2}{n} \leq \frac{|B_0|}{n} + \sum_{i=1}^{I-1} \sum_{j \in B_i} \frac{\overline{\tau_j}^2}{n} \quad .$$

Moreover, for any $i \in [\![0, I-2]\!]$, $j \in B_i$ and $k \in B_{i+1}$,

$$\tau_j^2 \geq \tau_{v_i}^2 \geq \tau_{v_i+1}^2 = \tau_{u_{i+1}}^2 = \overline{\tau_k}^2 \ ,$$

hence for any $i \in [\![0, I-2]\!]$

$$\sum_{k \in B_{i+1}} \overline{\tau_k}^2 \leq |B_{i+1}| \tau_{v_i}^2 \leq (1+\eta)|B_i|\tau_{v_i}^2 \leq (1+\eta) \sum_{j \in B_i} \tau_j^2 \ .$$

Therefore,

$$\sum_{i=0}^{I-1} \sum_{j \in B_i} \frac{\overline{\tau_j}^2}{n} \leq \frac{|B_0|}{n} + (1+\eta) \sum_{i=0}^{I-2} \sum_{j \in B_i} \frac{\tau_j^2}{n} \leq \frac{|B_0|}{n} + (1+\eta) \sum_{i=0}^{I-1} \sum_{j \in B_i} \frac{\tau_j^2}{n} \ .$$

Overall, we get that

$$\begin{aligned} R(\overline{\tau}) &\leq (1+\eta) \sum_{i=0}^{I-1} \sum_{j \in B_i} \frac{\tau_j^2}{n} + \sum_{i=0}^{I-1} \sum_{j \in B_i} \left((1-\tau_j)^2 - \frac{\tau_j^2}{n} \right) \theta_j^2 + \sum_{j > v_{I-1}} \theta_j^2 + \frac{|B_0|}{n} \\ &\leq (1+\eta) \sum_{j=1}^{v_{I-1}} \left(\frac{\tau_j^2(1-\theta_j^2)}{n} + (1-\tau_j)^2 \theta_j^2 \right) + \eta \sum_{j=1}^{v_{I-1}} \frac{\tau_j^2 \theta_j^2}{n} + \sum_{j > v_{I-1}} \theta_j^2 + \frac{|B_0|}{n} \\ &\leq (1+\eta) \sum_{j=1}^{v_{I-1}} \left(\frac{\tau_j^2(1-\theta_j^2)}{n} + (1-\tau_j)^2 \theta_j^2 \right) + \sum_{j > v_{I-1}} \theta_j^2 + \frac{|B_0| + \eta \left\| s \right\|^2}{n} \ . \end{aligned}$$

In particular, for $\tau = \tau^{\mathcal{S}_{ni}}$, there exists some $\overline{\tau} \in \mathcal{S}_{B,I}$ such that

$$\begin{aligned} R(\overline{\tau}) \leq (1+\eta) \sum_{j=1}^{v_{I-1}} \left(\frac{\left(\tau_j^{\mathcal{S}_{ni}} \right)^2 (1-\theta_j^2)}{n} + \left(1 - \tau_j^{\mathcal{S}_{ni}} \right)^2 \theta_j^2 \right) \\ + \sum_{j > v_{I-1}} \theta_j^2 + \frac{|B_0| + \eta \left\| s \right\|^2}{n} \ . \end{aligned}$$

Since $R\left(\tau_{B,I}^{\mathcal{S}} \right) \leq R(\overline{\tau})$, $u_I = v_{I-1} + 1$ and

$$\begin{aligned} R\left(\tau^{\mathcal{S}_{ni}} \right) = \sum_{j=1}^{v_{I-1}} \left(\frac{\left(\tau_j^{\mathcal{S}_{ni}} \right)^2 \left(1 - \theta_j^2 \right)}{n} + \left(1 - \tau_j^{\mathcal{S}_{ni}} \right)^2 \theta_j^2 \right) \\ + \sum_{j=u_I}^{\infty} \left(\frac{\left(\tau_j^{\mathcal{S}_{ni}} \right)^2 \left(1 - \theta_j^2 \right)}{n} + \left(1 - \tau_j^{\mathcal{S}_{ni}} \right)^2 \theta_j^2 \right) \ , \end{aligned}$$

it suffices to prove

$$\sum_{j=u_I}^{\infty} \theta_j^2 \leq (1+\eta) \sum_{j=u_I}^{\infty} \left(\frac{\left(\tau_j^{\mathcal{S}_{ni}}\right)^2 \left(1-\theta_j^2\right)}{n} + \left(1-\tau_j^{\mathcal{S}_{ni}}\right)^2 \theta_j^2 \right) \ .$$

If $\tau_{u_I}^{\mathcal{S}_{ni}} \leq 1 - 1/\sqrt{1+\eta}$, then by definition $\tau_j^{\mathcal{S}_{ni}} \leq 1 - 1/\sqrt{1+\eta}$ for all $j \geq u_I$, hence, for $j \geq u_I$, that is for $j > v_{I-1}$, $1/(1+\eta) \leq \left(1-\tau_j^{\mathcal{S}_{ni}}\right)^2$, thus

$$(1+\eta) \sum_{j=u_I}^{\infty} \left(\frac{\left(\tau_j^{\mathcal{S}_{ni}}\right)^2 \left(1-\theta_j^2\right)}{n} + \left(1-\tau_j^{\mathcal{S}_{ni}}\right)^2 \theta_j^2 \right)$$
$$\geq (1+\eta) \sum_{j=u_I}^{\infty} \left(\left(1-\tau_j^{\mathcal{S}_{ni}}\right)^2 \theta_j^2 \right) \geq \sum_{j=u_I}^{\infty} \theta_j^2 \ .$$

Now, assume that $\|s\|^2 = \sum_{j\geq 1} \theta_j^2 \leq L^2$ and, for any $u \in (0,1]$, denote by $H_u^{\mathcal{S}_{ni}} = \{j \geq 1 \text{ s.t. } \tau_j^{\mathcal{S}_{ni}} \geq u\}$. Clearly $\left|H_u^{\mathcal{S}_{ni}}\right|$ is nonincreasing when u increases. Introducing the sequence τ^0 identically equal to zero, Rigollet (2006a) remarked that for any $u \in (0,1]$

$$L^2 \geq R(\tau^0) \geq R\left(\tau^{\mathcal{S}_{ni}}\right) \geq \frac{1}{n} \sum_{j \in H_u^{\mathcal{S}_{ni}}} \left(\tau_j^{\mathcal{S}_{ni}}\right)^2 - \frac{L^2}{n} \geq \frac{u^2 \left|H_u^{\mathcal{S}_{ni}}\right|}{n} - \frac{L^2}{n} \ .$$

This proves $\left|H_u^{\mathcal{S}_{ni}}\right| \leq (n+1)(L/u)^2$ for any $u \in (0,1]$. In particular, for

$$u = 1 - \frac{1}{\sqrt{1+\eta}} = \frac{\sqrt{1+\sqrt{\log n}} - \sqrt{\sqrt{\log n}}}{\sqrt{1+\sqrt{\log n}}} \geq \frac{1}{2(1+\sqrt{\log n})} \ ,$$

and $L = (1+\sqrt{\log n})/3$, we get

$$\left|H_u^{\mathcal{S}_{ni}}\right| \leq \frac{4}{9}(n+1)(1+\sqrt{\log n})^4 \leq n(\log n)^2 \leq v_{I-1} \ ,$$

where the inequality before the last one holds for $n \geq 4 * 10^8$ for example.

□

Now, for any $j \in [0, I-1]$ let $\varepsilon_j = (I|B_j|)^{-1/2}$. Let $\mathcal{S}_{B,I}^d \subset \mathcal{S}_{B,I}$ denote the subset of sequences $\tau \in \mathcal{S}_{B,I}$ such that, for any $i \in [0, I-1]$,

$$\forall j \in B_i, \qquad \tau_j = \overline{\tau_i}, \qquad \text{where} \quad \overline{\tau_i} \in \{k\varepsilon_i,\ 0 \leq k \leq 1/\varepsilon_i\} \ .$$

We need another approximation result.

Lemma 3.6. *Let $\varepsilon = (\varepsilon_0, \dots, \varepsilon_{I-1}) \in [0,1]^I$ be such that $\varepsilon_i = 1/\sqrt{I|B_i|}$. Let $\tau, \tau^\varepsilon \in \mathcal{S}_{B,I}$ be such that there exist $\overline{\tau}, \overline{\tau}^\varepsilon \in [0,1]^I$ such that for all $i \in [0, I-1]$*

$$0 \leq \overline{\tau}_i^\varepsilon - \overline{\tau}_i \leq \varepsilon_i, \qquad \textit{and} \quad \forall k \in B_i \quad \tau_k = \overline{\tau}_i \quad \textit{and} \quad \tau_k^\varepsilon = \overline{\tau}_i^\varepsilon \ .$$

Then, for any $u > 0$,

$$R(\tau^\varepsilon) \leq (1+u)R(\tau) + \frac{2 + u^{-1}}{n} \ .$$

Proof: By (3.21) and using the inequality $2ab \leq ua^2 + u^{-1}b^2$,

$$R(\tau^\varepsilon) - R(\tau) \leq \sum_{i=0}^{I-1} \frac{2\varepsilon_i |B_i| (\overline{\tau}_i + \varepsilon_i)}{n} \leq uR(\tau) + \left(2 + \frac{1}{u}\right) \sum_{i=0}^{I-1} \frac{\varepsilon_i^2 |B_i|}{n} \ .$$

□

Since $|B_0| \geq 2/\eta$, $|B_1| \geq (1+\eta/2)|B_0|$ and recursively, we have $|B_k| \geq (1+\eta/2)^k |B_0|$, therefore,

$$v_I \geq |B_0| \sum_{k=0}^{I} \left(1 + \frac{\eta}{2}\right)^k \geq \left(1 + \frac{\eta}{2}\right)^{I+1} - 1 \ .$$

Thus, $v_k \geq n(\log n)^2$ for any $k \geq \frac{\log(1+n(\log n)^2)}{\log(1+\eta/2)}$, hence, for example,

$$I \leq 20(\log n)^2 \ .$$

The cardinal of $\mathcal{S}_{B,I}^d$ is therefore upper bounded by

$$|\mathcal{S}_{B,I}^d| \leq \prod_{i=0}^{I-1} \left(\sqrt{I} \, (1+\eta)^{i/2} \right) \leq e^{\frac{I}{2}\log(I) + \frac{I^2}{4}\log(1+\eta)} \leq e^{120(\log n)^3} \ .$$

Hence we observe that the collection $(\widehat{s}_\tau)_{\tau \in \mathcal{S}_{B,I}^d}$ is a collection of weighted projection estimators defined for the Fourier basis $(\varphi_i)_{i=1,\dots 2p}$, with $p = v_I \leq 2n(\log n)^2$ so

$$\sum_{i=1}^{2p} \varphi_i^2(x) = p \leq 2n(\log n)^2$$

and this collection of kernels satisfies (3.1) with $\Gamma = 2(\log n)^2$. Let $w_\tau = \log |\mathcal{S}_{B,I}^d|$ for any $\tau \in \mathcal{S}_{B,I}^d$. Applying Theorem 3.1 with $F = 1$ yields the following result

$$\frac{1-\varepsilon}{1+\varepsilon} R\left(\tau_{V,1}^{\text{pen}}\right) \leq \inf_{\tau \in \mathcal{S}_{B,I}^d} R(\tau) + \kappa \Upsilon \frac{(\log n)^6}{n\varepsilon^3} \ .$$

In order to bound Υ, we will require furthermore that $\|s\|_\infty \leq \left(\frac{1}{3}(1+\sqrt{\log n})\right)^2$, which implies $\|s\| \leq \frac{1}{3}(1+\sqrt{\log n})$. We have then $\Upsilon \leq (\log n)^3$ and therefore, for any s such that $\|s\|_\infty \leq$

$\left((1+\sqrt{\log n})/3\right)^2$,

$$R\left(\tau_{V,1}^{\text{pen}}\right) \leq \left(1+\frac{\kappa}{\sqrt{\log n}}\right) R\left(\tau^{\mathcal{S}_{ni}}\right) + \kappa\frac{(\log n)^{11}}{n} \ .$$

3.10 Concentration tools

Theorem 3.3 (Bernstein's inequality)**.** *Let* $\mathbf{X} = (X_1, \ldots, X_n)$ *be an i.i.d. sample of random variables with values in a Polish space* $(\Xi, \mathcal{Z})$*, and let* u *be a measurable real valued function. Then for all* $x > \log(2)$

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\left(u(X_i) - \mathbb{E}\left[u(X)\right]\right)\right| > \sqrt{\frac{2\text{Var}\left[u(X)\right]x}{n}} + \frac{\|u\|_\infty x}{3n}\right) \leq 2e^{-x} \ . \tag{3.38}$$

The following updated version of the concentration inequality of Houdré & Reynaud-Bouret (2003) is proved in Lerasle *et al.* (2015).

Theorem 3.4. *Let* $X, X_1, \ldots X_n$ *be i.i.d. random variables defined on a Polish space* Ξ *and let* $(f_{i,j})_{1\leqslant i\neq j\leq n}$ *denote bounded real valued symmetric measurable functions defined on* Ξ^2*, such that for any* $i \neq j$*,* $f_{i,j} = f_{j,i}$ *and*

$$\forall 1 \leqslant i \neq j \leq n, \qquad \mathbb{E}\left[f_{i,j}(x, X)\right] = 0 \qquad \textit{for a.e. } x \in \Xi \ .$$

Let U *be the following totally degenerate* U*-statistic of order* 2*,*

$$U = \sum_{1\leqslant i\neq j\leq n} f_{i,j}(X_i, X_j) \ .$$

Let A *be an upper bound of* $|f_{i,j}(x,y)|$ *for any* i, j, x, y *and*

$$B^2 = \max\left(\sup_{i,x\in\Xi}\sum_{j=1}^{i}\mathbb{E}\left[f_{i,j}(x,X_j)^2\right], \sup_{j,t\in\Xi}\sum_{i=j+1}^{n}\mathbb{E}\left[f_{i,j}(X_i,t)^2\right]\right)$$

$$C^2 = \sum_{1\leq i\neq j\leq n}\mathbb{E}\left[f_{i,j}(X_i,X_j)^2\right]$$

$$D = \sup_{(a,b)\in\mathcal{A}}\mathbb{E}\left[\sum_{1\leq i<j\leq n} f_{i,j}(X_i,X_j)a_i(X_i)b_j(X_j)\right] \ ,$$

where $\mathcal{A} = \left\{(a,b), \ \textit{s.t.} \ \mathbb{E}\left[\sum_{i=1}^{n-1} a_i(X_i)^2\right] \leq 1, \ \mathbb{E}\left[\sum_{j=2}^{n} b_j(X_j)^2\right] \leq 1\right\}$*. Then for any* $u > 0$ *and* $\varepsilon > 0$*,*

$$\mathbb{P}\left(U \geq 2\sqrt{2}(1+\varepsilon)^{\frac{3}{2}}C\sqrt{u} + 8\left(1+\frac{1}{\varepsilon}\right)^{\frac{1}{2}}Du + 24\left(1+\varepsilon+\frac{1}{\varepsilon}\right)^2\left(Bu^{\frac{3}{2}} + Au^2\right)\right)$$
$$\leqslant 2.7e^{-u} \ .$$

3.11 Additional Simulations

procedure $\mathcal{C}$	s_3	s_{11}	s_{16}	s_{17}
LS2F	2.42 ± 0.035	2.65 ± 0.05	2.73 ± 0.04	2.74 ± 0.03
LS5F	2.27 ± 0.037	2.59 ± 0.06	2.58 ± 0.05	2.51 ± 0.04
LS10F	2.30 ± 0.040	2.63 ± 0.06	2.60 ± 0.05	2.53 ± 0.04
LOO	2.27 ± 0.041	2.67 ± 0.07	2.57 ± 0.05	2.54 ± 0.04
pen2F_{C1}	4.15 ± 0.106	4.81 ± 0.13	4.55 ± 0.11	4.70 ± 0.11
$\text{pen2F}_{\text{C1.25}}$	2.48 ± 0.040	2.92 ± 0.07	2.84 ± 0.05	2.86 ± 0.05
$\text{pen2F}_{\text{C1.5}}$	2.14 ± 0.028	2.32 ± 0.04	2.39 ± 0.03	2.35 ± 0.03
$\text{pen2F}_{\text{C1.75}}$	2.06 ± 0.021	2.16 ± 0.03	2.26 ± 0.03	2.24 ± 0.02
pen2F_{C2}	2.06 ± 0.018	2.09 ± 0.02	2.23 ± 0.02	2.20 ± 0.02
$\text{pen2F}_{\text{C2.5}}$	2.20 ± 0.017	2.12 ± 0.02	2.29 ± 0.02	2.29 ± 0.02
pen2F_{C3}	2.38 ± 0.019	2.20 ± 0.02	2.47 ± 0.02	2.44 ± 0.02
pen5F_{C1}	2.71 ± 0.052	3.17 ± 0.09	3.09 ± 0.06	3.02 ± 0.05
$\text{pen5F}_{\text{C1.25}}$	1.90 ± 0.022	2.14 ± 0.04	2.12 ± 0.03	2.10 ± 0.03
$\text{pen5F}_{\text{C1.5}}$	1.78 ± 0.014	1.88 ± 0.02	1.92 ± 0.02	1.94 ± 0.02
$\text{pen5F}_{\text{C1.75}}$	1.80 ± 0.012	1.85 ± 0.02	1.90 ± 0.01	1.95 ± 0.01
pen5F_{C2}	1.89 ± 0.011	1.90 ± 0.01	1.96 ± 0.01	2.02 ± 0.01
$\text{pen5F}_{\text{C2.5}}$	2.11 ± 0.013	2.10 ± 0.01	2.14 ± 0.01	2.23 ± 0.01
pen5F_{C3}	2.35 ± 0.015	2.30 ± 0.02	2.37 ± 0.02	2.44 ± 0.02
$\text{pen10F}_{\text{C1}}$	2.50 ± 0.047	2.90 ± 0.07	2.87 ± 0.06	2.75 ± 0.05
$\text{pen10F}_{\text{C1.25}}$	1.80 ± 0.020	2.02 ± 0.04	1.97 ± 0.02	1.97 ± 0.02
$\text{pen10F}_{\text{C1.5}}$	1.72 ± 0.012	1.82 ± 0.02	1.84 ± 0.02	$\mathbf{1.85 \pm 0.01}$
$\text{pen10F}_{\text{C1.75}}$	1.77 ± 0.011	1.81 ± 0.01	1.85 ± 0.01	1.89 ± 0.01
$\text{pen10F}_{\text{C2}}$	1.85 ± 0.010	1.88 ± 0.01	1.92 ± 0.01	1.98 ± 0.01
$\text{pen10F}_{\text{C2.5}}$	2.09 ± 0.012	2.10 ± 0.01	2.12 ± 0.01	2.22 ± 0.01
$\text{pen10F}_{\text{C3}}$	2.34 ± 0.014	2.32 ± 0.02	2.36 ± 0.02	2.45 ± 0.02
$\text{penLOO}_{\text{C1}}$	2.29 ± 0.041	2.70 ± 0.07	2.59 ± 0.05	2.50 ± 0.04
$\text{penLOO}_{\text{C1.25}}$	1.74 ± 0.018	1.93 ± 0.03	1.90 ± 0.02	1.88 ± 0.02
$\text{penLOO}_{\text{C1.5}}$	$\mathbf{1.71 \pm 0.011}$	$\mathbf{1.80 \pm 0.02}$	$\mathbf{1.82 \pm 0.01}$	$\mathbf{1.85 \pm 0.01}$
$\text{penLOO}_{\text{C1.75}}$	1.80 ± 0.010	1.84 ± 0.01	1.88 ± 0.01	1.93 ± 0.01
$\text{penLOO}_{\text{C2}}$	1.91 ± 0.010	1.94 ± 0.01	1.99 ± 0.01	2.05 ± 0.01
$\text{penLOO}_{\text{C2.5}}$	2.16 ± 0.011	2.28 ± 0.02	2.25 ± 0.01	2.34 ± 0.01
$\text{penLOO}_{\text{C3}}$	2.48 ± 0.015	2.62 ± 0.02	2.54 ± 0.01	2.65 ± 0.02
$\text{Epenid}_{\text{C1}}$	2.32 ± 0.042	2.72 ± 0.07	2.59 ± 0.05	2.54 ± 0.04
$\text{Epenid}_{\text{C1.25}}$	1.70 ± 0.018	1.91 ± 0.03	1.83 ± 0.02	1.84 ± 0.02
$\text{Epenid}_{\text{C1.5}}$	1.65 ± 0.010	1.74 ± 0.02	1.73 ± 0.01	1.78 ± 0.01
$\text{Epenid}_{\text{C1.75}}$	1.72 ± 0.009	1.75 ± 0.01	1.76 ± 0.01	1.82 ± 0.01
$\text{Epenid}_{\text{C2}}$	1.82 ± 0.009	1.81 ± 0.01	1.84 ± 0.01	1.91 ± 0.01
$\text{Epenid}_{\text{C2.5}}$	2.02 ± 0.010	2.07 ± 0.01	2.06 ± 0.01	2.14 ± 0.01
$\text{Epenid}_{\text{C3}}$	2.27 ± 0.012	2.42 ± 0.02	2.26 ± 0.01	2.41 ± 0.02

Table 3.4: Values of $\xi(\mathcal{C})$ for several procedures for collection $\mathcal{F}_{\text{K}}$ with $n = 100$ (see the text Section 3.6). The best procedure are bolded for each density.

procedure $\mathcal{C}$	s_3	s_{11}	s_{16}	s_{17}
LS2F	1.62 ± 0.011	1.87 ± 0.026	1.76 ± 0.015	1.70 ± 0.011
LS5F	1.52 ± 0.012	1.79 ± 0.024	1.69 ± 0.016	1.56 ± 0.012
LS10F	1.52 ± 0.014	1.81 ± 0.025	1.69 ± 0.019	1.56 ± 0.013
LOO	1.51 ± 0.015	1.81 ± 0.026	1.67 ± 0.017	1.53 ± 0.014
pen2F_{C1}	2.29 ± 0.043	3.08 ± 0.088	2.59 ± 0.041	2.34 ± 0.038
$\text{pen2F}_{\text{C1.25}}$	1.66 ± 0.015	1.97 ± 0.030	1.80 ± 0.017	1.69 ± 0.013
$\text{pen2F}_{\text{C1.5}}$	1.56 ± 0.010	1.74 ± 0.019	1.66 ± 0.012	1.60 ± 0.009
$\text{pen2F}_{\text{C1.75}}$	1.58 ± 0.008	1.73 ± 0.014	1.66 ± 0.010	1.64 ± 0.009
pen2F_{C2}	1.66 ± 0.009	1.79 ± 0.013	1.72 ± 0.010	1.71 ± 0.010
$\text{pen2F}_{\text{C2.5}}$	1.86 ± 0.011	1.99 ± 0.014	1.91 ± 0.012	1.91 ± 0.012
pen2F_{C3}	2.07 ± 0.013	2.22 ± 0.017	2.12 ± 0.014	2.12 ± 0.014
pen5F_{C1}	1.75 ± 0.029	2.27 ± 0.058	1.96 ± 0.028	1.78 ± 0.027
$\text{pen5F}_{\text{C1.25}}$	1.41 ± 0.010	1.57 ± 0.017	1.51 ± 0.010	1.44 ± 0.008
$\text{pen5F}_{\text{C1.5}}$	1.41 ± 0.006	1.51 ± 0.013	1.48 ± 0.007	1.44 ± 0.006
$\text{pen5F}_{\text{C1.75}}$	1.48 ± 0.006	1.55 ± 0.009	1.54 ± 0.007	1.53 ± 0.007
pen5F_{C2}	1.57 ± 0.007	1.65 ± 0.009	1.63 ± 0.008	1.63 ± 0.008
$\text{pen5F}_{\text{C2.5}}$	1.80 ± 0.008	1.90 ± 0.010	1.84 ± 0.009	1.86 ± 0.009
pen5F_{C3}	2.03 ± 0.010	2.15 ± 0.012	2.09 ± 0.011	2.08 ± 0.011
$\text{pen10F}_{\text{C1}}$	1.66 ± 0.027	2.14 ± 0.062	1.82 ± 0.024	1.67 ± 0.020
$\text{pen10F}_{\text{C1.25}}$	1.38 ± 0.010	1.52 ± 0.015	1.45 ± 0.009	1.39 ± 0.007
$\text{pen10F}_{\text{C1.5}}$	1.38 ± 0.005	1.46 ± 0.011	1.45 ± 0.007	1.42 ± 0.006
$\text{pen10F}_{\text{C1.75}}$	1.45 ± 0.005	1.50 ± 0.008	1.52 ± 0.006	1.51 ± 0.006
$\text{pen10F}_{\text{C2}}$	1.55 ± 0.006	1.62 ± 0.008	1.61 ± 0.007	1.62 ± 0.007
$\text{pen10F}_{\text{C2.5}}$	1.80 ± 0.008	1.91 ± 0.010	1.83 ± 0.009	1.85 ± 0.009
$\text{pen10F}_{\text{C3}}$	2.03 ± 0.009	2.15 ± 0.011	2.08 ± 0.011	2.08 ± 0.010
$\text{penLOO}_{\text{C1}}$	1.54 ± 0.016	1.82 ± 0.026	1.68 ± 0.017	1.57 ± 0.014
$\text{penLOO}_{\text{C1.25}}$	$\mathbf{1.35 \pm 0.009}$	1.48 ± 0.013	$\mathbf{1.41 \pm 0.008}$	$\mathbf{1.36 \pm 0.006}$
$\text{penLOO}_{\text{C1.5}}$	1.38 ± 0.005	$\mathbf{1.42 \pm 0.008}$	1.44 ± 0.006	1.39 ± 0.005
$\text{penLOO}_{\text{C1.75}}$	1.44 ± 0.005	1.44 ± 0.006	1.53 ± 0.006	1.49 ± 0.006
$\text{penLOO}_{\text{C2}}$	1.53 ± 0.006	1.58 ± 0.008	1.61 ± 0.006	1.63 ± 0.007
$\text{penLOO}_{\text{C2.5}}$	1.84 ± 0.009	2.00 ± 0.011	1.80 ± 0.007	1.88 ± 0.008
$\text{penLOO}_{\text{C3}}$	2.09 ± 0.009	2.22 ± 0.010	2.07 ± 0.010	2.05 ± 0.009
$\text{Epenid}_{\text{C1}}$	1.62 ± 0.037	2.16 ± 0.075	1.69 ± 0.017	1.58 ± 0.017
$\text{Epenid}_{\text{C1.25}}$	1.36 ± 0.009	1.49 ± 0.015	1.41 ± 0.008	1.36 ± 0.006
$\text{Epenid}_{\text{C1.5}}$	1.38 ± 0.005	1.42 ± 0.011	1.43 ± 0.006	1.38 ± 0.005
$\text{Epenid}_{\text{C1.75}}$	1.44 ± 0.005	1.42 ± 0.006	1.52 ± 0.006	1.48 ± 0.006
$\text{Epenid}_{\text{C2}}$	1.52 ± 0.006	1.53 ± 0.007	1.60 ± 0.006	1.61 ± 0.007
$\text{Epenid}_{\text{C2.5}}$	1.84 ± 0.009	1.94 ± 0.010	1.77 ± 0.007	1.86 ± 0.008
$\text{Epenid}_{\text{C3}}$	2.09 ± 0.009	2.20 ± 0.010	2.04 ± 0.010	2.02 ± 0.009

Table 3.5: Values of $\xi(\mathcal{C})$ for several procedures for collection $\mathcal{F}_{\text{K}}$ with $n = 500$ (see the text Section 3.6). The best procedure are bolded for each density.

procedure $\mathcal{C}$	s_3	s_{11}	s_{13}	s_{16}	s_{17}	s_{22}
LS2F	3.66 ± 0.04	4.81 ± 0.08	1.54 ± 0.009	3.95 ± 0.05	3.40 ± 0.035	2.82 ± 0.03
LS5F	3.40 ± 0.04	4.38 ± 0.09	1.49 ± 0.009	3.73 ± 0.05	3.10 ± 0.037	2.54 ± 0.03
LS10F	3.36 ± 0.04	4.30 ± 0.09	1.49 ± 0.010	3.72 ± 0.05	3.06 ± 0.038	2.49 ± 0.03
LOO	3.39 ± 0.04	4.37 ± 0.09	1.50 ± 0.010	3.78 ± 0.05	3.23 ± 0.040	2.46 ± 0.03
pen2F_{C1}	6.87 ± 0.07	9.57 ± 0.15	1.91 ± 0.017	7.77 ± 0.09	6.88 ± 0.076	4.72 ± 0.04
$\text{pen2F}_{\text{C1.25}}$	5.10 ± 0.05	6.71 ± 0.10	1.64 ± 0.011	5.55 ± 0.06	4.75 ± 0.048	3.70 ± 0.03
$\text{pen2F}_{\text{C1.5}}$	4.21 ± 0.04	5.51 ± 0.09	1.55 ± 0.009	4.54 ± 0.05	3.87 ± 0.037	3.32 ± 0.03
$\text{pen2F}_{\text{C1.75}}$	3.77 ± 0.03	4.91 ± 0.08	1.53 ± 0.009	4.03 ± 0.05	3.44 ± 0.032	3.15 ± 0.02
pen2F_{C2}	3.53 ± 0.03	4.50 ± 0.07	1.53 ± 0.009	3.73 ± 0.04	3.19 ± 0.028	3.11 ± 0.02
$\text{pen2F}_{\text{C2.5}}$	3.36 ± 0.03	4.26 ± 0.06	1.59 ± 0.010	3.53 ± 0.04	3.02 ± 0.025	3.19 ± 0.02
pen2F_{C3}	3.37 ± 0.03	4.24 ± 0.06	1.68 ± 0.012	3.51 ± 0.04	3.02 ± 0.024	3.35 ± 0.02
pen5F_{C1}	4.47 ± 0.06	6.17 ± 0.12	1.69 ± 0.014	4.88 ± 0.07	4.13 ± 0.053	3.20 ± 0.04
$\text{pen5F}_{\text{C1.25}}$	3.10 ± 0.04	3.90 ± 0.08	1.45 ± 0.009	3.40 ± 0.05	2.78 ± 0.032	2.51 ± 0.02
$\text{pen5F}_{\text{C1.5}}$	2.57 ± 0.03	3.13 ± 0.05	1.39 ± 0.008	2.81 ± 0.04	2.28 ± 0.024	2.31 ± 0.02
$\text{pen5F}_{\text{C1.75}}$	2.36 ± 0.02	2.78 ± 0.04	1.39 ± 0.008	2.52 ± 0.03	2.11 ± 0.020	2.28 ± 0.02
pen5F_{C2}	2.25 ± 0.02	2.66 ± 0.04	1.41 ± 0.008	2.41 ± 0.03	2.06 ± 0.017	2.31 ± 0.02
$\text{pen5F}_{\text{C2.5}}$	2.26 ± 0.02	2.64 ± 0.04	1.50 ± 0.009	2.37 ± 0.02	2.09 ± 0.014	2.49 ± 0.02
pen5F_{C3}	2.36 ± 0.02	2.75 ± 0.03	1.63 ± 0.011	2.50 ± 0.02	2.24 ± 0.014	2.71 ± 0.02
$\text{pen10F}_{\text{C1}}$	3.88 ± 0.05	5.28 ± 0.11	1.63 ± 0.014	4.26 ± 0.06	3.57 ± 0.046	2.81 ± 0.03
$\text{pen10F}_{\text{C1.25}}$	2.67 ± 0.03	3.29 ± 0.07	1.41 ± 0.009	2.92 ± 0.04	2.38 ± 0.028	2.22 ± 0.02
$\text{pen10F}_{\text{C1.5}}$	2.24 ± 0.02	2.66 ± 0.04	1.36 ± 0.007	2.42 ± 0.03	1.97 ± 0.020	2.07 ± 0.02
$\text{pen10F}_{\text{C1.75}}$	2.03 ± 0.02	2.39 ± 0.04	1.36 ± 0.007	2.21 ± 0.03	1.86 ± 0.016	2.07 ± 0.02
$\text{pen10F}_{\text{C2}}$	1.99 ± 0.02	2.30 ± 0.03	1.38 ± 0.008	2.14 ± 0.02	1.87 ± 0.014	2.12 ± 0.02
$\text{pen10F}_{\text{C2.5}}$	2.05 ± 0.01	2.35 ± 0.03	1.46 ± 0.009	2.16 ± 0.02	1.97 ± 0.012	2.32 ± 0.02
$\text{pen10F}_{\text{C3}}$	2.20 ± 0.01	2.48 ± 0.03	1.60 ± 0.010	2.31 ± 0.02	2.15 ± 0.012	2.53 ± 0.02
$\text{penLOO}_{\text{C1}}$	3.36 ± 0.04	4.39 ± 0.09	1.50 ± 0.010	3.68 ± 0.05	3.07 ± 0.039	2.47 ± 0.03
$\text{penLOO}_{\text{C1.25}}$	2.31 ± 0.03	2.91 ± 0.07	1.36 ± 0.007	2.60 ± 0.04	2.05 ± 0.024	2.02 ± 0.02
$\text{penLOO}_{\text{C1.5}}$	1.98 ± 0.02	2.28 ± 0.04	$\mathbf{1.33 \pm 0.007}$	2.20 ± 0.03	1.76 ± 0.017	$\mathbf{1.90 \pm 0.02}$
$\text{penLOO}_{\text{C1.75}}$	1.82 ± 0.02	2.09 ± 0.03	$\mathbf{1.33 \pm 0.007}$	2.06 ± 0.02	$\mathbf{1.70 \pm 0.014}$	$\mathbf{1.90 \pm 0.02}$
$\text{penLOO}_{\text{C2}}$	$\mathbf{1.80 \pm 0.01}$	$\mathbf{2.05 \pm 0.03}$	1.36 ± 0.007	$\mathbf{1.97 \pm 0.02}$	1.75 ± 0.012	1.97 ± 0.02
$\text{penLOO}_{\text{C2.5}}$	1.98 ± 0.01	2.28 ± 0.02	1.44 ± 0.008	1.99 ± 0.02	1.92 ± 0.010	2.17 ± 0.02
$\text{penLOO}_{\text{C3}}$	2.17 ± 0.01	2.41 ± 0.02	1.56 ± 0.010	2.20 ± 0.02	2.07 ± 0.010	2.42 ± 0.02
$\text{Epenid}_{\text{C1}}$	3.26 ± 0.06	4.93 ± 0.12	1.59 ± 0.014	3.28 ± 0.05	2.52 ± 0.037	2.64 ± 0.04
$\text{Epenid}_{\text{C1.25}}$	2.05 ± 0.03	2.93 ± 0.07	1.36 ± 0.008	2.20 ± 0.03	1.66 ± 0.018	2.03 ± 0.02
$\text{Epenid}_{\text{C1.5}}$	1.75 ± 0.02	2.31 ± 0.05	1.32 ± 0.007	1.86 ± 0.02	1.50 ± 0.012	1.89 ± 0.02
$\text{Epenid}_{\text{C1.75}}$	1.64 ± 0.01	2.01 ± 0.03	1.32 ± 0.007	1.77 ± 0.02	1.52 ± 0.009	1.87 ± 0.02
$\text{Epenid}_{\text{C2}}$	1.64 ± 0.01	1.96 ± 0.03	1.34 ± 0.007	1.76 ± 0.01	1.62 ± 0.008	1.90 ± 0.02
$\text{Epenid}_{\text{C2.5}}$	1.88 ± 0.01	2.19 ± 0.02	1.42 ± 0.008	1.82 ± 0.01	1.86 ± 0.008	2.05 ± 0.01
$\text{Epenid}_{\text{C3}}$	2.11 ± 0.01	2.34 ± 0.02	1.53 ± 0.009	2.06 ± 0.01	2.02 ± 0.009	2.25 ± 0.01

Table 3.6: Values of $\xi(\mathcal{C})$ for several procedures for collection $\mathcal{F}_{\text{KR}}$ with $n = 500$ (see the text Section 3.6). The best procedure are bolded for each density.

Chapter 4

Variance computations for V-fold criteria

Abstract. We compute the variance of the resampling procedures of Chapter 3 together with the variance of key quantities for model selection performance. We show these variances are nonincreasing with V (at least in some particular cases), suggesting the performance increases with V. We recover all the results, and the subsequent discussions, of Arlot & Lerasle (2014) for projection estimators. Our computations are illustrated by numerical experiments for the problem of bandwidth selection and suggest, again, that the performance increases much from $V = 2$ to $V = 5$ or 10, and then is almost constant.

NOTA: Ce chapitre reflète les avancées d'une recherche en cours, menée en collaboration avec Sylvain Arlot[(1)] et Matthieu Lerasle[(2)].

Contents

[(1)]École Normale Supérieure, Paris
[(2)]Université Nice Sophia Antipolis

4.1 Introduction

The search of (first-order optimal) oracle inequalities for selection procedures has been one of the biggest challenge in modern nonparametric statistics. As explained in the previous chapter, many of them have been proved in density estimation for a large number of procedures. Among others references in the least-squares framework one can cite Rigollet (2006a), Massart (2007), Bunea *et al.* (2010), Goldenshluger & Lepski (2011) and Birgé (2014). In this context, we proved in Section 3.4 of Chapter 3 that the resampling procedures $\mathcal{C}_{V,F}^{\mathrm{pen}}$ (thus $\mathcal{C}_{V}^{\mathrm{corr,LSVF}}$), $\mathcal{C}_{V}^{\mathrm{LSVF}}$ and $\mathcal{C}_{p}^{\mathrm{LPO}}$ are also optimal (up to some multiplicative constant) for selecting a linear estimator.

Nevertheless, practitioners have pointed out an important gap between theory and practice which can be illustrated by the relation between hold-out (HO) and cross-validation (CV). Both methods are quite popular among statisticians since their "universality" allows to use them to attack several types of problem (Arlot & Celisse, 2010). There are two opposite conclusions concerning their practical and theoretical performances. On the one hand, while CV techniques present rather good performances and are widely used in practice, it is well-known that HO does not work well due to its large variability. On the other hand, there are some impressive theoretical results for HO –it can be shown that HO with a training set of size $n(1-1/\log n)$ is asymptotically optimal and satisfies oracle inequalities (van der Laan & Dudoit, 2003; Blanchard & Massart, 2006)– whereas CV procedures were not well studied from a nonasymptotic point of view (this was true at least before the work of Arlot (2008) and Arlot & Lerasle (2014)). Moreover, even if one finds very precise oracle inequalities for CV procedures, something we were able to do in Section 3.4, it will be not sufficient to prove properly that HO is worse than CV.

The main general question that arises from these observations is how to compare "effectively" two given procedures from a theoretical point of view. Ideally, one would like to quantify the difference in their performances in terms of risk, that is, in the example we have in mind, to be able to evaluate the improvement of CV regarding to HO. This led some authors to focus on variance computations of CV criteria (see Section 5.2 of Arlot & Celisse (2010)) hoping to prove the common intuition that the "variance" of HO is larger than the "variance" of a CV procedure. Unfortunately, most of these results were asymptotic and did not show clearly the influence of the parameter V which appears in second-order terms. These works were therefore useless to help effectively the practitioners to calibrate their procedures.

In a recent paper Arlot & Lerasle (2014) have made some remarkable progress in this direction. Starting from these empirical observations they first suggest that one should go beyond first-order comparisons based only on oracle inequalities. Second, they provide a general heuristic to quantify the quality of any given procedure (not only resampling procedures). It permits one to decide between two procedures having same bias by comparing the variance of their increments and not the variance of the criteria. In particular, they validate experimentally their heuristic with histogram estimators showing that the quality of $\mathcal{C}_{V,1}^{\mathrm{pen}}$ improves with V but the improvement mainly appears between $V=2$ and $V=5$ and tends to vanish when V goes from 10 to n. This empirical study justifies the classical advice for VFCV (at least when taking computational considerations into account), that is "$V=5$ is enough".

The goal of this chapter is twofold: first to extend the nonasymptotic variance computations for V-fold criteria, made by Arlot & Lerasle (2014) for projection estimators, to linear estimators

in general. Second, to proceed an empirical study on synthetic data that validates the heuristic and confirms that $V = 5$ is enough to guarantee the quality of $\mathcal{C}^{\mathrm{pen}}_{V,1}$. We believe these calculations will be an important tool to understand precisely how the model selection performances of VFCV and VF penalization depend on V. In particular this is a first step in order to help the practitioners to deal with the famous bandwidth selection problem (which is not covered by the paper of Arlot and Lerasle) using resampling methods. As explained in Chapter 3, this problem was massively treated (since the eighties) with a CV approach without strong nonasymptotic theoretical guarantees.

We remind in Section 4.2 the heuristic of Arlot & Lerasle (2014) in order to emphasize the quantity of interest. In Section 4.3 we provide the main theorems for linear estimators and give some interpretation for projection and approximation kernels. Section 4.4 is dedicated to a small empirical study that confirms the heuristic for the bandwidth selection problem. All the proofs are in Section 4.5 and some additional theorems and figures can be found in Section 4.6.

4.2 Why the variance?

This chapter is dependent from Chapter 3. In particular, we consider again the least-squares density estimation setting and we shall therefore use the same notations. Suppose we have at hand a family $(\mathcal{K}_m)_{m\in\mathcal{M}}$ of symmetric kernels from which one derives the associated family $(\widehat{s}_m = \mathcal{A}_m(\mathbf{X}))_{m\in\mathcal{M}}$ of linear estimators of the type (3.2). Moreover, we consider model selection procedures of the form $\mathcal{C} : \mathcal{M} \to \mathbb{R}$ and set $\widehat{m}(\mathcal{C}) \in \operatorname{argmin}_{m\in\mathcal{M}} \mathcal{C}(m)$.

The ideal procedure in this framework is thus given by $\mathcal{C}_{\mathrm{id}}(m) = P\gamma(\widehat{s}_m)$, where $\gamma(\widehat{s}_m) = \|\widehat{s}_m\|^2 - 2\,\widehat{s}_m$. When one considers penalized procedures of the type $\mathcal{C}_{\mathrm{pen}}(m) := P_n\gamma(\widehat{s}_m) + \mathrm{pen}(m)$, the ideal penalty is written $\mathrm{pen}_{\mathrm{id}}(m, \mathbf{X}) = 2(P_n - P)\mathcal{A}_m(\mathbf{X})$. A classical approach to choose the penalty is given by the so-called "unbiased risk estimation principle" which says that a good penalty should satisfy

$$\mathbb{E}\left[\mathrm{pen}(m)\right] = \mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m, \mathbf{X})\right] \quad \text{for all } m \in \mathcal{M} \ .$$

Interestingly, one can prove the following explicit expression for the previous expectation (it can be deduced from (2.9) in Section 2.2.2)

$$\mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m, \mathbf{X})\right] = \mathbb{E}\left[(P - P_n)\gamma\left(\mathcal{A}_m(\mathbf{X})\right)\right] = \frac{1}{|\mathbf{X}|} f(m) \ , \tag{4.1}$$

with $f(m) = 2(P\chi_m - Ps_m)$.

In Chapter 3 we mainly considered V-fold penalization (see the general definition in Section 3.3.3)

$$\mathcal{C}^{\mathrm{pen}}_{V,F}(m) = P_n\gamma\left(\widehat{s}_m\right) + F\ \mathrm{pen}_{\mathrm{VF}}(m, V) \ ,$$

where the V-fold penalty can be written, for a regular partition of $\mathbf{X}$ in V samples of same size,

$$\mathrm{pen}_{\mathrm{VF}}(m, V) = \frac{V-1}{V^2} \sum_{j=1}^{V} \left(\left(P_n^{(j)} - P_n^{(-j)} \right) \gamma\left(\widehat{s}_m^{(-j)} \right) \right) \ .$$

Hence, the penalty $\mathrm{pen}_{\mathrm{VF}}$ is satisfactory since we have for any $m \in \mathcal{M}$ and $V \in \{2, \dots, n\}$

$$\begin{aligned}
\mathbb{E}\left[\mathrm{pen}_{\mathrm{VF}}(m, V)\right] &= \frac{V-1}{V^2} \sum_{j=1}^{V} \mathbb{E}\left[\left(P_n^{(j)} - P_n^{(-j)}\right) \gamma\left(\widehat{s}_m^{(-j)}\right)\right] \\
&= \frac{V-1}{V} \mathbb{E}\left[\left(P_n^{(1)} - P_n^{(-1)}\right) \gamma\left(\widehat{s}_m^{(-1)}\right)\right] \\
&= \frac{V-1}{V} \mathbb{E}\left[\left(P - P_n^{(-1)}\right) \gamma\left(\widehat{s}_m^{(-1)}\right)\right] \\
&= \frac{V-1}{V} \mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}\left(m, \mathbf{X}^{(-B_1)}\right)\right] \\
&= \frac{V-1}{V} \frac{1}{\left|\mathbf{X}^{(-B_1)}\right|} f(m) = \mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m, \mathbf{X})\right] \quad , \qquad (4.2)
\end{aligned}$$

where we used the fact that the data are i.i.d. and $\mathbf{X}^{(-B_1)}$ is independent from $\mathbf{X}^{(B_1)}$. In particular $\mathbb{E}\left[\mathcal{C}_{V,1}^{\mathrm{pen}}(m)\right] = \mathbb{E}\left[\mathcal{C}_{\mathrm{id}}(m)\right]$ for all $m \in \mathcal{M}$, whatever the value of $V \in \{2, \dots, n\}$.

The main general question is the following. Having at hand a collection $(\mathcal{C}_\alpha)_{\alpha \in \Gamma}$ of procedures that are known to satisfy an oracle inequality such as (3.5), how can we compare them? Let us consider the simplest case and imagine there are only two procedures (say, $\mathcal{C}_1$ and $\mathcal{C}_2$) to compare. Since the goal is estimation, we should say that $\mathcal{C}_1$ is better than $\mathcal{C}_2$ if we have (with large probability) for some $\delta_n > 0$

$$\left\|\widehat{s}_{\widehat{m}(\mathcal{C}_1)} - s\right\|^2 \leq (1 - \delta_n) \left\|\widehat{s}_{\widehat{m}(\mathcal{C}_2)} - s\right\|^2 \quad .$$

Unfortunately proving this kind of inequality is very tricky since it requires a very precise analysis of both procedures, including lower bounds on their losses. Up to our knowledge, such a result was only made once by Arlot (2008) in a very particular case that seems difficult to generalize. Let us instead understand how one should proceed by summarizing the main steps that lead to the heuristic of Arlot & Lerasle (2014). To do so, let us define the deterministic oracle as

$$\overline{m} := \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \mathbb{E}\left[\left\|\widehat{s}_m - s\right\|^2\right] \right\} \quad .$$

- One must go beyond the first-order comparison. Take for instance $\mathcal{C}_1 = \mathcal{C}_{2,1}^{\mathrm{pen}}$ and $\mathcal{C}_2 = \mathcal{C}_{n,1}^{\mathrm{pen}}$. In this case the leave-one-out generally outperforms the 2-fold procedure in practice, apparently because it reduces drastically the variability. Nevertheless, these procedures are at first-order theoretically identical since from (4.2) one has $\mathbb{E}\left[\mathcal{C}_1(m)\right] = \mathbb{E}\left[\mathcal{C}_2(m)\right]$ for all m. In general, two procedures with same bias as an estimator of $\mathbb{E}\left[\left\|\widehat{s}_m - s\right\|^2\right]$ will be indistinguishable when analyzing the corresponding oracle inequalities which are only upper bounds, even if one is known to be much better than the other one in practice!

- A natural hypothesis would be then to study the variance of the criteria. But second-order analysis has also to be made carefully. Indeed, while the selected model $\widehat{m}(\mathcal{C})$ is unchanged when $\mathcal{C}(m)$ is translated by any random quantity, it is clear that such a translation does change the value of $\mathrm{Var}\left[\mathcal{C}(m)\right]$ and can make it as large as desired. This fact also underlines the difference between risk estimation and estimator selection (Breiman & Spector, 1992).

- Their idea is the following. They start by saying that for a procedure $\mathcal{C}$, the smaller is $\mathbb{P}\left(\widehat{m}(\mathcal{C}) = m\right)$ for all $m \neq \overline{m}$, the better should be the performance of $\widehat{s}_{\widehat{m}(\mathcal{C})}$. Hence, their goal is to find a "proxy" for $\mathbb{P}\left(\widehat{m}(\mathcal{C}) = m\right)$, that is a quantity that should behave similarly as a function of $\mathcal{C}$. Then defining for all $m, m' \in \mathcal{M}$, $\Delta_{\mathcal{C}}(m, m') := \mathcal{C}(m) - \mathcal{C}(m')$, and

$$\mathrm{SNR}_{\mathcal{C}}(m) := \max_{m' \neq m} \frac{\mathbb{E}\left[\Delta_{\mathcal{C}}(m, m')\right]}{\sqrt{\mathrm{Var}\left[\Delta_{\mathcal{C}}(m, m')\right]}} \quad \forall m \in \mathcal{M} \ ,$$

their heuristic says that for every $m \in \mathcal{M}$

$$\begin{aligned}
\mathbb{P}\left(\widehat{m}(\mathcal{C}) = m\right) &= \mathbb{P}\left(\forall m' \neq m,\ \mathcal{C}(m) - \mathcal{C}(m') \leq 0\right) \\
&\approx \min_{m' \neq m} \mathbb{P}\left(\Delta_{\mathcal{C}}(m, m') \leq 0\right) \\
&\approx \min_{m' \neq m} \mathbb{P}\left(\mathbb{E}\left[\Delta_{\mathcal{C}}(m, m')\right] + \xi \sqrt{\mathrm{Var}\left[\Delta_{\mathcal{C}}(m, m')\right]} \leq 0\right) \\
&= \min_{m' \neq m} \mathbb{P}\left(\xi \geq \frac{\mathbb{E}\left[\Delta_{\mathcal{C}}(m, m')\right]}{\sqrt{\mathrm{Var}\left[\Delta_{\mathcal{C}}(m, m')\right]}}\right) \\
&= \varphi\left(\mathrm{SNR}_{\mathcal{C}}(m)\right) \ ,
\end{aligned}$$

where for some standard Gaussian variable ξ, $\varphi(t) = \mathbb{P}(\xi > t)$. Hence

$$\mathcal{C}_1 \text{ is better than } \mathcal{C}_2 \quad \Longleftrightarrow \quad \forall m \neq \overline{m} \qquad \mathrm{SNR}_{\mathcal{C}_1}(m) > \mathrm{SNR}_{\mathcal{C}_2}(m) \ .$$

Now, assuming that

$$\mathbb{E}\left[\mathcal{C}_1(m)\right] = \mathbb{E}\left[\mathcal{C}_2(m)\right] \ \forall m \in \mathcal{M}, \quad \text{and } \overline{m} = \operatorname*{argmin}_{m \in \mathcal{M}} \mathbb{E}\left[\mathcal{C}_1(m)\right] = \operatorname*{argmin}_{m \in \mathcal{M}} \mathbb{E}\left[\mathcal{C}_2(m)\right] \ ,$$

we find Arlot and Lerasle's heuristic (see a discussion about the previous assumption in (Arlot & Lerasle, 2014, Section 4))

$$\text{" if } \quad \forall m \neq m' \quad \mathrm{Var}\left[\Delta_{\mathcal{C}_1}(m, m')\right] < \mathrm{Var}\left[\Delta_{\mathcal{C}_2}(m, m')\right], \quad \text{then } \mathcal{C}_1 \text{ is better than } \mathcal{C}_2 \text{ " } \ .$$

They also enlighten that all pairs (m, m') do not matter equally for explaining a quantitative difference in the performance of a procedure $\mathcal{C}$. In particular we can fix $m' = \overline{m}$ since the strongest candidate against any $m \neq \overline{m}$ is $\overline{m}$. This explains why we consider $\mathrm{Var}\left[\Delta_{\mathcal{C}}(m, \overline{m})\right]$ in Section 4.4.

Remark. In particular, one can apply this heuristic to resampling procedures. It allows for instance to compare HO with any CV procedure with same bias (such as the LPO with same training set size). In addition, having two VFCV procedures one can hope to have a precise idea of their quality with respect to V. Nevertheless, this is still insufficient to quantify the difference between two procedures. Imagine we "prove" with this heuristic that leave-one-out is better than 10-fold, how much gain do we have by taking $V = n$ instead of $V = 10$?

Remark. Finally, one can note that the central quantity in the above analysis, $\Delta_{\mathcal{C}}(m, m')$, has some link with relative bounds (Audibert, 2004; Catoni, 2007) which were mainly used in a statistical learning setting.

4.3 Variance computations

We provide here the main results of this chapter in the general setting of Chapter 3. The proofs are to be found in Section 4.5. Let us fix m and $m' \in \mathcal{M}$ and let X, Y, Z denote independent copies of X_1. For $x, y \in \Xi$ we note $\Delta^s(x) = s_m(x) - s_{m'}(x)$ and

$$\Delta^{\mathcal{K}}(x,y) = \mathcal{K}_m(x,y) - \mathcal{K}_{m'}(x,y), \qquad \Delta^{A}(x,y) = A_m(x,y) - A_{m'}(x,y) \ .$$

4.3.1 Results

Classical resampling procedures

Resampling procedures such as LPO and LSVF are widely used for risk estimation and model selection. We therefore provide the relevant computations for the variance of their increments in the following theorems and for the variance of these criteria in Section 4.6.2.

Theorem 4.1. *For any $V \in \{2, \dots, n\}$, we set $\Delta_V^{\mathrm{LSVF}}(m, m') = \Delta_{\mathcal{C}_V^{\mathrm{LSVF}}}(m, m')$. We then have*

$$\begin{aligned}
\mathrm{Var}\left[\Delta_V^{\mathrm{LSVF}}(m,m')\right] &= \frac{8V}{n^2(V-1)}\left(\mathrm{Var}\left[\Delta^{\mathcal{K}}(X,Y)\right] - 2\,\mathrm{Cov}\left[\Delta^{\mathcal{K}}(X,Y), \Delta^{\mathcal{K}}(Y,Z)\right]\right) \\
&+ \frac{2V}{n^2(V-1)^2}\left(1 + \frac{(V-2)^2}{V-1} - \frac{V}{n}\right)\mathrm{Var}\left[\Delta^{A}(X,Y)\right] \\
&+ \frac{V^2}{n^3(V-1)^2}\mathrm{Var}\left[\Delta^{A}(X,X)\right] + \frac{16}{n}\,\mathrm{Cov}\left[\Delta^{\mathcal{K}}(X,Y), \Delta^{\mathcal{K}}(Y,Z)\right] \\
&- \frac{8V}{n^2(V-1)^2}\,\mathrm{Cov}\left[\Delta^{\mathcal{K}}(X,Y), (V-1)\Delta^{A}(X,X) + (V-2)\Delta^{A}(X,Y)\right] \\
&+ \frac{4V}{n^2}\left(\frac{n}{V} - \frac{3V^2-7V+5}{(V-1)^3} + \frac{2V}{n(V-1)^2}\right)\mathrm{Cov}\left[\Delta^{A}(X,Y), \Delta^{A}(Y,Z)\right] \\
&+ \frac{16V}{n^2(V-1)}\left(\frac{(V-1)(V-n)}{V} - \frac{(V-2)^2}{V-1}\right)\mathrm{Cov}\left[\Delta^{A}(X,Y), \Delta^{\mathcal{K}}(Y,Z)\right] \\
&+ \frac{4V}{n^2(V-1)}\left(1 - \frac{V}{n(V-1)}\right)\mathrm{Cov}\left[\Delta^{A}(X,Y), \Delta^{A}(X,X)\right] \ .
\end{aligned}$$

Theorem 4.2. *For any $p \in [n-1]$, we set $q = n - p$ and $\Delta_p^{\mathrm{LPO}}(m, m') = \Delta_{\mathcal{C}_p^{\mathrm{LPO}}}(m, m')$, so that we get the following variance formula*

$$\begin{aligned}
\mathrm{Var}\left[\Delta_p^{\mathrm{LPO}}(m,m')\right] &= \frac{1}{n^3 q^2}\mathrm{Var}\left[\Delta^{A}(X,X)\right] \\
&+ \frac{2}{n(n-1)}\mathrm{Var}\left[\frac{q-1}{q}\Delta^{A}(X,Y) - 2\Delta^{\mathcal{K}}(X,Y)\right]
\end{aligned}$$

$$+\frac{4}{n^2 q}\operatorname{Cov}\left[\frac{q-1}{q}\Delta^A(X,Y)-2\Delta^{\mathcal{K}}(X,Y),\Delta^A(X,X)\right]$$
$$+\frac{4(n-2)}{n(n-1)}\operatorname{Cov}\left[\frac{q-1}{q}\Delta^A(X,Y)-2\Delta^{\mathcal{K}}(X,Y),\frac{q-1}{q}\Delta^A(Y,Z)-2\Delta^{\mathcal{K}}(Y,Z)\right] .$$

Penalized V-fold procedures

It is well-known that classical V-fold procedures are biased and therefore cannot achieve asymptotically optimal oracle inequalities in many frameworks (see the discussion of Theorem 3.2). Penalized V-fold procedures (Arlot, 2008) were introduced as an alternative to correct this drawback.

Theorem 4.3. *For any $V \in \{2,\dots,n\}$ and $F>0$ we set $\Delta_{V,F}(m,m') = \Delta_{\mathcal{C}^{\mathrm{pen}}_{V,F}}(m,m')$, so that we get the following variance formula*

$$\begin{aligned}
\operatorname{Var}\left[\Delta_{V,F}(m,m')\right] &= \frac{1}{n^3}\operatorname{Var}\left[\Delta^A(X,X)+2(F-1)\Delta^{\mathcal{K}}(X,X)\right]+\frac{2(n-1)}{n^3}\operatorname{Var}\left[\Delta^A(X,Y)\right]\\
&+\frac{8}{n^2}\left(1+\frac{F^2}{V-1}-\frac{(F-1)^2}{n}\right)\left(\operatorname{Var}\left[\Delta^{\mathcal{K}}(X,Y)\right]-2\operatorname{Cov}\left[\Delta^{\mathcal{K}}(X,Y),\Delta^{\mathcal{K}}(Y,Z)\right]\right)\\
&+\frac{16}{n^3}\left((F-1)(2n+F-1)+n^2\right)\operatorname{Cov}\left[\Delta^{\mathcal{K}}(X,Y),\Delta^{\mathcal{K}}(Y,Z)\right]\\
&+\frac{4(n-1)}{n^3}\operatorname{Cov}\left[\Delta^A(X,Y),(n-2)\Delta^A(Y,Z)+\Delta^A(X,X)+2(F-1)\Delta^{\mathcal{K}}(X,X)\right]\\
&-\frac{8}{n^2}\left(1+\frac{(F-1)}{n}\right)\operatorname{Cov}\left[\Delta^{\mathcal{K}}(X,Y),\Delta^A(X,X)+2(F-1)\Delta^{\mathcal{K}}(X,X)\right]\\
&-\frac{8}{n^2}\left(1+\frac{(F-1)}{n}\right)\operatorname{Cov}\left[\Delta^{\mathcal{K}}(X,Y),\Delta^A(X,Y)+2(n-2)\Delta^A(Y,Z)\right] .
\end{aligned}$$

In particular, for Burman's criterion given by (3.8), that is for $F=1$, we have

$$\begin{aligned}
\operatorname{Var}\left[\Delta_{V,1}(m,m')\right] &= \frac{1}{n^3}\operatorname{Var}\left[\Delta^A(X,X)\right]+\frac{2(n-1)}{n^3}\operatorname{Var}\left[\Delta^A(X,Y)\right]\\
&+\frac{8V}{n^2(V-1)}\left(\operatorname{Var}\left[\Delta^{\mathcal{K}}(X,Y)\right]-2\operatorname{Cov}\left[\Delta^{\mathcal{K}}(X,Y),\Delta^{\mathcal{K}}(Y,Z)\right]\right)\\
&+\frac{4(n-1)}{n^3}\operatorname{Cov}\left[\Delta^A(X,Y),(n-2)\Delta^A(Y,Z)+\Delta^A(X,X)\right]\\
&+\frac{8}{n^2}\operatorname{Cov}\left[\Delta^{\mathcal{K}}(X,Y),2n\Delta^{\mathcal{K}}(Y,Z)-\Delta^A(X,X)-\Delta^A(X,Y)-2(n-2)\Delta^A(Y,Z)\right] .
\end{aligned}$$

Ideal procedures

For sake of completeness, we provide the computations for the ideal procedure $\mathcal{C}_{\mathrm{id}}$ and the penalized procedure with the ideal deterministic penalty $\mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m)\right]$ which serves as benchmark for comparison.

Theorem 4.4. *Considering the ideal procedure $\mathcal{C} = \mathcal{C}_{\mathrm{id}}$ and setting $\Delta_{\mathrm{id}}(m, m') = \Delta_{\mathcal{C}_{\mathrm{id}}}(m, m')$, one has*

$$\begin{aligned}\mathrm{Var}\left[\Delta_{\mathrm{id}}(m,m')\right] &= \frac{1}{n^3}\mathrm{Var}\left[\Delta^A(X,X) - 2n\Delta^s(X)\right] + \frac{2(n-1)}{n^3}\mathrm{Var}\left[\Delta^A(X,Y)\right] \\ &+ \frac{4(n-1)}{n^3}\,\mathrm{Cov}\left[\Delta^A(X,Y), (n-2)\Delta^A(Y,Z) + \Delta^A(X,X) - 2n\Delta^s(X)\right] \quad .\end{aligned}$$

For $\mathcal{C}(m) = P_n\gamma(\widehat{s}_m) + \mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m)\right]$, setting $\Delta_{\mathrm{id,det}}(m, m') = \mathcal{C}(m) - \mathcal{C}(m')$, one has

$$\begin{aligned}\mathrm{Var}\left[\Delta_{\mathrm{id,det}}(m,m')\right] &= \frac{1}{n^3}\mathrm{Var}\left[\Delta^A(X,X) - 2\Delta^{\mathcal{K}}(X,X)\right] \\ &+ \frac{2(n-1)}{n^3}\mathrm{Var}\left[\Delta^A(X,Y) - 2\Delta^{\mathcal{K}}(X,Y)\right] \\ &+ \frac{4(n-1)}{n^3}\,\mathrm{Cov}\left[\Delta^A(X,Y) - 2\Delta^{\mathcal{K}}(X,Y), \Delta^A(X,X) - 2\Delta^{\mathcal{K}}(X,X)\right] \\ &+ \frac{4(n-1)(n-2)}{n^3}\,\mathrm{Cov}\left[\Delta^A(X,Y) - 2\Delta^{\mathcal{K}}(X,Y), \Delta^A(Y,Z) - 2\Delta^{\mathcal{K}}(Y,Z)\right] \quad .\end{aligned}$$

4.3.2 Interpretation

Albeit the theorems involve many terms and seem therefore very hard to interpret we can underline some remarkable facts in general. Concerning VF procedures, as we focus on the influence of V, we observe that both the first term in Theorem 4.1 and the only term in Theorem 4.3 with a factor that depends on V are controlled by the sign of

$$\mathrm{Var}\left[\Delta^{\mathcal{K}}(X,Y)\right] - 2\,\mathrm{Cov}\left[\Delta^{\mathcal{K}}(X,Y), \Delta^{\mathcal{K}}(Y,Z)\right] \quad .$$

Interestingly this quantity is nonnegative for symmetric kernels as it is shown by the following result.

Lemma 4.1. *Let X, Y, Z be independent copies of X_1. Then, for any symmetric application K*

$$\mathrm{Var}\left[K(X,Y)\right] \geq 2\,\mathrm{Cov}\left[K(X,Y), K(Y,Z)\right] \quad . \tag{4.3}$$

This implies that $\mathrm{Var}\left[\Delta_{V,1}(m,m')\right]$ depends on V as $8V/(V-1)$ (plus some additive terms which are independent of V), that is a nonincreasing function of V. Hence, the heuristic of Arlot and Lerasle suggests that the performance of the penalized V-fold criterion improves when V increases but the improvement is at most in a second order term when V is large (since $V/(V-1) = 1 + 1/(V-1)$), whatever the order of magnitude of the additive terms. Since the second term in Theorem 4.1 is also a nonincreasing function of V, the same commentary holds for Δ_V^{LSVF} if the first two terms are truly the dominant terms in the right-hand side of the theorem. In this situation, the performance of the classical LSVF procedure should also improve with V (at least asymptotically, that is with $n \to \infty$ and $V \to \infty$) since both the bias and the variance get better when V increases.

Projection kernels

Let us focus on the projection kernel $\mathcal{K}_m$ which was presented in Section 3.2.2 (see Example 1). Let S_m denote a linear subspace of $\mathbb{L}_2$. Given any orthonormal basis $(\psi_\lambda)_{\lambda\in\Lambda_m}$ of S_m, the projection kernel $\mathcal{K}_m$ is defined for all $(x,y)\in\Xi^2$ as $\mathcal{K}_m(x,y)=\sum_{\lambda\in\Lambda_m}\psi_\lambda(x)\psi_\lambda(y)$. In this case we have $A_m(x,y)=\mathcal{K}_m(x,y)$, hence $\Delta^A(x,y)=\Delta^{\mathcal{K}}(x,y)$. Using the notations of Arlot & Lerasle (2014), let us note

$$B(m,m')=\beta\left(\Lambda_m,\Lambda_m\right)+\beta\left(\Lambda_{m'},\Lambda_{m'}\right)-2\beta\left(\Lambda_m,\Lambda_{m'}\right)\quad,$$

with

$$\beta\left(\Lambda_m,\Lambda_{m'}\right)=\sum_{\lambda\in\Lambda_m,\lambda'\in\Lambda_{m'}}\left(\mathbb{E}\left[\left(\psi_\lambda(X)-P\psi_\lambda\right)\left(\psi_{\lambda'}(X)-P\psi_{\lambda'}\right)\right]\right)^2\quad.$$

One easily finds

$$\begin{aligned}
&\mathrm{Var}\left[\Delta^{\mathcal{K}}(X,Y)\right]=B(m,m')+2\mathrm{Var}\left[\Delta^s(X)\right]\\
&\mathrm{Cov}\left[\Delta^{\mathcal{K}}(X,Y),\Delta^{\mathcal{K}}(X,X)\right]=\mathrm{Cov}\left[\Delta^s(X),\Delta^{\mathcal{K}}(X,X)\right]\\
&\mathrm{Cov}\left[\Delta^{\mathcal{K}}(X,Y),\Delta^{\mathcal{K}}(Y,Z)\right]=\mathrm{Cov}\left[\Delta^{\mathcal{K}}(X,Y),\Delta^s(X)\right]=\mathrm{Var}\left[\Delta^s(X)\right]\quad.
\end{aligned}$$

Hence, from Theorem 4.3 and Theorem 4.4 one finds

$$\begin{aligned}
\mathrm{Var}\left[\Delta_{\mathrm{id}}(m,m')\right]&=\frac{2}{n^2}\left(1-\frac{1}{n}\right)B(m,m')+\frac{1}{n^3}\mathrm{Var}\left[\Delta^{\mathcal{K}}(X,X)-2\Delta^s(X)\right]\quad,\\
\mathrm{Var}\left[\Delta_{\mathrm{id,det}}(m,m')\right]&=\frac{2}{n^2}\left(1-\frac{1}{n}\right)B(m,m')+\frac{4}{n}\mathrm{Var}\left[\frac{n-1}{n}\Delta^s(X)+\frac{1}{2n}\Delta^{\mathcal{K}}(X,X)\right]\quad,
\end{aligned}$$

and

$$\begin{aligned}
&\mathrm{Var}\left[\Delta_{V,1}(m,m')\right]\\
&\qquad=\frac{2}{n^2}\left(1+\frac{4}{V-1}-\frac{1}{n}\right)B(m,m')+\frac{4}{n}\mathrm{Var}\left[\frac{n+1}{n}\Delta^s(X)-\frac{1}{2n}\Delta^{\mathcal{K}}(X,X)\right]\quad.
\end{aligned}$$

Thus, when using the expectation of the ideal penalty instead of a V-fold penalty the factor appearing in front of the first term becomes $1+4/(V-1)-1/n$. Assuming that the heuristics of Arlot & Lerasle (2014) is true, this proves that the $\mathrm{pen}_{\mathrm{LOO}}$ behaves like the expectation of the ideal penalty. Together with Theorem 4.7 that gives the variance computations of the criteria (see Section 4.6.2 below), we recover Theorem 2 of Arlot & Lerasle (2014). As a consequence all their remarks and specific computations (in particular their discussion on the case of regular histograms) can be deduced from our result.

Approximation kernels

Let $\Xi\subset\mathbb{R}$ and let μ denote the Lebesgue measure on Ξ. Let k denote a bounded symmetric function in $\mathbb{L}_1$, such that $k(0)>0$. We remind that for any $h_m>0$, the approximation kernel associated to k and h_m is defined for any $(x,y)\in\Xi^2$ by $\mathcal{K}_m(x,y)=h_m^{-1}k((x-y)/h_m)$. In this

case one can notice that some important terms have no randomness

$$\Delta^{\mathcal{K}}(X,X) = \mathcal{K}_m(X,X) - \mathcal{K}_{m'}(X,X) = k(0)\left(\frac{1}{h_m} - \frac{1}{h_{m'}}\right)$$
$$\Delta^{A}(X,X) = \int \left(\mathcal{K}_m(x,X)^2 - \mathcal{K}_{m'}(x,X)^2\right) d\mu(x) = \|k\|^2 \left(\frac{1}{h_m} - \frac{1}{h_{m'}}\right) \ ,$$

hence

$$\mathrm{Var}\left[\Delta_{\mathrm{id,det}}(m,m')\right] = \frac{2(n-1)}{n^3}\mathrm{Var}\left[\Delta^A(X,Y) - 2\Delta^{\mathcal{K}}(X,Y)\right]$$
$$+ \frac{4(n-1)(n-2)}{n^3}\,\mathrm{Cov}\left[\Delta^A(X,Y) - 2\Delta^{\mathcal{K}}(X,Y), \Delta^A(Y,Z) - 2\Delta^{\mathcal{K}}(Y,Z)\right] \ ,$$

and

$$\mathrm{Var}\left[\Delta_{V,1}(m,m')\right]$$
$$= \frac{2(n-1)}{n^3}\mathrm{Var}\left[\Delta^A(X,Y)\right] + \frac{4(n-1)(n-2)}{n^3}\,\mathrm{Cov}\left[\Delta^A(X,Y), \Delta^A(Y,Z)\right]$$
$$+ \frac{8V}{n^2(V-1)}\mathrm{Var}\left[\Delta^{\mathcal{K}}(X,Y)\right] + \frac{16(n(V-1)-V)}{n^2(V-1)}\,\mathrm{Cov}\left[\Delta^{\mathcal{K}}(X,Y), \Delta^{\mathcal{K}}(Y,Z)\right]$$
$$- \frac{8}{n^2}\,\mathrm{Cov}\left[\Delta^{\mathcal{K}}(X,Y), \Delta^A(X,Y) + 2(n-2)\Delta^A(Y,Z)\right] \ .$$

4.4 Simulation experiments

4.4.1 Simulation protocol

As some experiments were already made for projection estimators in Arlot & Lerasle (2014) we focus only on approximation kernels (or, equivalently, kernel density estimators) with fixed Gaussian function k defined on $\Xi = \mathbb{R}$. In the sequence $(\widehat{s}_m)_{m\in\mathcal{M}_{\mathcal{P}}}$ denote a collection of approximation Gaussian kernels with a "geometrical" grid of bandwidths, that is with

$$\mathcal{K}_m(x,y) = \frac{1}{\sqrt{2\pi}h_m}\exp\left(\frac{-(x-y)^2}{2h_m^2}\right) \ ,$$

and

$$\mathcal{M}_{\mathcal{P}} = \left\{ h_m = \max\left(\frac{1}{n}, \frac{1}{n\log n}\left(1 + \frac{3}{2\log n}\right)^m\right),\ m = 1,\dots,\left\lfloor(\log n)^2\right\rfloor\right\} \ .$$

We considered 10 densities coming from the R-package *benchden*[(3)] which are illustrated in the Figure 4.2 in the supplementary material below. We study hereafter the behavior of the variance term $\mathrm{Var}\left[\Delta_{\mathcal{C}}(m,\overline{m})\right]$ as a function of $1/h_m$ which measures the "complexity" of the kernel $\mathcal{K}_m$ exactly as the dimension D_m of S_m represents the complexity of a projection kernel (see Section 4.1 in Lerasle *et al.* (2015)). We consider unbiased procedures $\mathcal{C} = \mathcal{C}^{\mathrm{pen}}_{V,F}$ with $F = 1$ and

[(3)] that implements the benchmark distributions of Berlinet & Devroye (1994).

$V \in \{2, 5, 10, n\}$ ($V = n$ being the leave-one-out penalty) together with the ideal deterministic penalty that is $\mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m)\right]$. All procedures are compared on $N = 10000$ independent synthetic data sets of size $n = 100$.

4.4.2 Simulation results

Figure 4.1 shows the values of $\mathrm{Var}\left[\Delta_{\mathcal{C}}(m, \overline{m})\right]$ for two well-known densities, the standard Gaussian (on the left) and the isosceles triangular density[(4)] (on the right). It should be noticed that these densities are no exceptions, we obtain indeed the same kind of picture for other densities in Section 4.6.3. Similar results were obtained for $n = 500$ but we restrict the presentation to $n = 100$ for sake of clarity. We observe from Figure 4.1 the same behavior of $\mathrm{Var}\left[\Delta_{\mathcal{C}}(m, \overline{m})\right]$

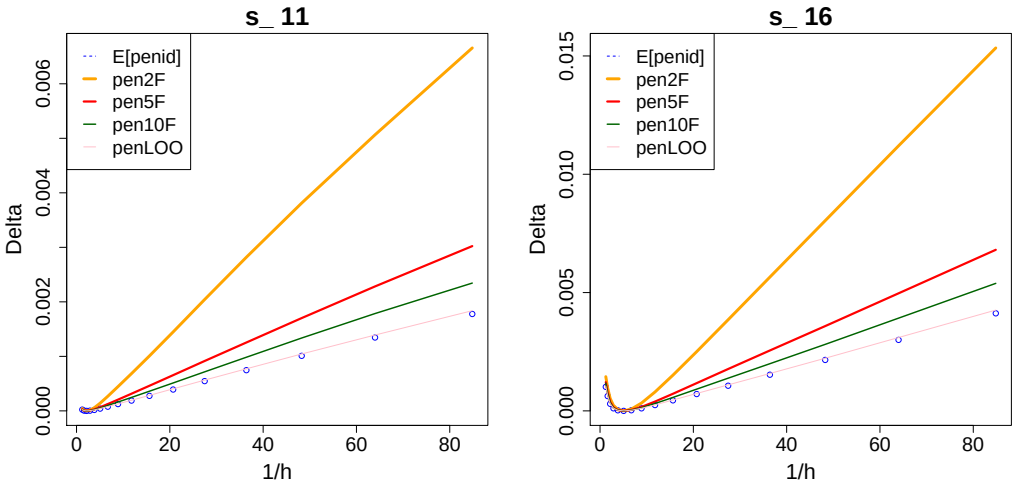

Figure 4.1: Illustration of the variance heuristic: $\mathrm{Var}\left[\Delta_{\mathcal{C}}(m, \overline{m})\right]$ as a function of $1/h_m$ for five different $\mathcal{C}$ with $n = 100$.

with respect to V as in Arlot & Lerasle (2014). First, the biggest difference appears between $V = 2$ and $V = 5$. Second, the improvement is very small for $V \geq 10$. Third, the leave-one-out penalty (denoted $\mathrm{pen}_{\mathrm{LOO}}$) is nearly indistinguishable from the choice $\mathrm{pen}(m) = \mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m)\right]$. Finally, let us mention that similar results were obtained for $n = 500$.

4.5 Proofs

Let us develop all the terms appearing in the different criteria of interest. As in Chapter 3 let us note for $i, j \in [n]$, $K_i, K_j \in [V]$ such that $i \in K_i, j \in K_j$, and define $\rho_{i,j}^{(\mathrm{VF})} = 1 - \frac{V}{V-1}\mathbf{1}_{K_i \neq K_j}$. Since in our setting there is no direct relation between penalized VF and VFCV as it is the case for projection estimators, we have two possible type of weight $\omega_{i,j}$ (instead of one in their case), defined for $i, j \in [n]$ by

$$\omega_{i,j}^{\mathrm{VF}} = 2(F\rho_{i,j}^{(\mathrm{VF})} - 1) \qquad \text{and} \qquad \omega_{i,j}^{\mathrm{VFCV}} = \left(1 - \frac{1}{V-1}\mathbf{1}_{K_i \neq K_j}\right) .$$

[(4)]The isosceles triangular density is given by $s(x) = (1 - |x|)_+$ for $x \in [-1, 1]$.

From (3.12), and

$$\|\widehat{s}_m\|^2 = \frac{1}{n^2}\sum_{i=1}^{n} A_m(X_i,X_i) + \frac{1}{n^2}\sum_{i\neq j=1}^{n} A_m(X_i,X_j)$$
$$P_n\widehat{s}_m = \frac{1}{n^2}\sum_{i,j=1}^{n} \mathcal{K}_m(X_i,X_j) = \frac{1}{n^2}\sum_{i=1}^{n} \mathcal{K}_m(X_i,X_i) + \frac{1}{n^2}\sum_{i\neq j=1}^{n} \mathcal{K}_m(X_i,X_j)$$

we have for any $m \in \mathcal{M}$,

$$\mathcal{C}_{V,F}^{\mathrm{pen}}(m) = \frac{1}{n^2}\sum_{i,j=1}^{n} \left(A_m(X_i,X_j) + \omega_{i,j}^{\mathrm{VF}} \mathcal{K}_m(X_i,X_j) \right) \tag{4.4}$$

$$\mathcal{C}_{\mathrm{id}}(m) = \frac{1}{n}\sum_{i=1}^{n} \left(\frac{1}{n} A_m(X_i,X_i) - 2s_m(X_i) \right) + \frac{1}{n^2}\sum_{i\neq j=1}^{n} A_m(X_i,X_j) \ . \tag{4.5}$$

From Lemma 3.1 in Chapter 3 one easily finds

$$\begin{aligned}\mathcal{C}_V^{\mathrm{LSVF}}(m) &= \mathcal{C}_{V,1}^{\mathrm{pen}}(m) + \frac{1}{n^2(V-1)} \left(\sum_{i=1}^{n} A_m(X_i,X_i) + \sum_{1\leqslant i\neq j\leqslant n} \rho_{i,j}^{(\mathrm{VF})} A_m(X_i,X_j) \right) \\ &= \frac{1}{n}\sum_{i=1}^{n} \frac{V}{n(V-1)} A_m(X_i,X_i) - \frac{2V}{n^2}\sum_{i\neq j=1}^{n} \mathcal{K}_m(X_i,X_j) \\ &\quad + \frac{V}{n^2(V-1)} \sum_{i\neq j} \omega_{i,j}^{\mathrm{VFCV}} \left(A_m(X_i,X_j) + 2(V-1)\mathcal{K}_m(X_i,X_j) \right) \ . \end{aligned} \tag{4.6}$$

and

$$\begin{aligned}\mathcal{C}_p^{\mathrm{LPO}}(m) &= \mathcal{C}_{n,1}^{\mathrm{pen}}(m) + \frac{p}{n^2(n-p)} \left(\sum_{i=1}^{n} A_m(X_i,X_i) - \frac{1}{n-1} \sum_{1\leqslant i\neq j\leqslant n} A_m(X_i,X_j) \right) \\ &= \frac{1}{n(n-p)}\sum_{i=1}^{n} A_m(X_i,X_i) \\ &\quad + \frac{1}{n(n-1)} \sum_{i\neq j} \left(\left(1 - \frac{1}{n-p} \right) A_m(X_i,X_j) - 2\mathcal{K}_m(X_i,X_j) \right) \ . \end{aligned} \tag{4.7}$$

4.5.1 Main Lemma

It appears that all expressions have the form

$$P_n f + \sum_{1\leqslant i\neq j\leqslant n} U_1(X_i,X_j) + \sum_{1\leqslant i\neq j\leqslant n} \omega_{i,j} U_2(X_i,X_j) \ ,$$

for some functions f, U_1 and U_2. The key tool to prove both theorems of Section 4.3 is the following lemma.

Lemma 4.2. *Let $(\omega_{i,j})_{1\leq i,j\leq n}$ be a symmetric matrix and let U_1, U_2 be symmetric functions. Let*

$$\mathcal{C} = P_n f + \sum_{1\leqslant i\neq j\leqslant n} U_1(X_i, X_j) + \sum_{1\leqslant i\neq j\leqslant n} \omega_{i,j} U_2(X_i, X_j) \ ,$$

$$\mathcal{C}' = P_n f' + \sum_{1\leqslant i\neq j\leqslant n} U_1'(X_i, X_j) + \sum_{1\leqslant i\neq j\leqslant n} \omega_{i,j} U_2'(X_i, X_j) \ .$$

We have

$$\begin{aligned}
&\mathrm{Cov}\left[\mathcal{C}, \mathcal{C}'\right] = \frac{1}{n}\,\mathrm{Cov}\left[f(X), f'(X)\right] + 2n(n-1)\,\mathrm{Cov}\left[U_1(X,Y), U_1'(X,Y)\right] \\
&+ 4n(n-1)(n-2)\,\mathrm{Cov}\left[U_1(X,Y), U_1'(Y,Z)\right] \\
&+ 2(n-1)\left(\mathrm{Cov}\left[f(X), U_1'(X,Y)\right] + \mathrm{Cov}\left[U_1(X,Y), f'(X)\right]\right) \\
&+ 2\left(\sum_{1\leqslant i\neq j\leqslant n} \omega_{i,j}^2\right)\mathrm{Cov}\left[U_2(X,Y), U_2'(X,Y)\right] \\
&+ 4\left(\sum_{1\leqslant i\neq j\leqslant n} \omega_{i,j} \sum_{k=1, k\neq(i,j)}^{n} \omega_{i,k}\right)\mathrm{Cov}\left[U_2(X,Y), U_2'(Y,Z)\right] \\
&+ 2\left(\sum_{1\leqslant i\neq j\leqslant n} \omega_{i,j}\right)\mathrm{Cov}\left[U_2(X,Y), \frac{1}{n}f'(X) + U_1'(X,Y) + 2(n-2)U_1'(Y,Z)\right] \\
&+ 2\left(\sum_{1\leqslant i\neq j\leqslant n} \omega_{i,j}\right)\mathrm{Cov}\left[U_2'(X,Y), \frac{1}{n}f(X) + U_1(X,Y) + 2(n-2)U_1(Y,Z)\right] \ .
\end{aligned}$$

In particular,

$$\begin{aligned}
&\mathrm{Var}\left[\mathcal{C}\right] = \frac{1}{n}\mathrm{Var}\left[f(X)\right] + 2n(n-1)\mathrm{Var}\left[U_1(X,Y)\right] + 2\left(\sum_{1\leqslant i\neq j\leqslant n} \omega_{i,j}^2\right)\mathrm{Var}\left[U_2(X,Y)\right] \\
&+ 4(n-1)\,\mathrm{Cov}\left[U_1(X,Y), f(X) + n(n-2)U_1(Y,Z)\right] \\
&+ 4\left(\sum_{1\leqslant i\neq j\leqslant n} \omega_{i,j} \sum_{k=1, k\neq(i,j)}^{n} \omega_{i,k}\right)\mathrm{Cov}\left[U_2(X,Y), U_2(Y,Z)\right] \\
&+ 4\left(\sum_{1\leqslant i\neq j\leqslant n} \omega_{i,j}\right)\mathrm{Cov}\left[U_2(X,Y), \frac{1}{n}f(X) + U_1(X,Y) + 2(n-2)U_1(Y,Z)\right] \ .
\end{aligned}$$

Proof: We develop the covariance to get, by independence and equidistribution of the variables,

$$\begin{aligned}
&\mathrm{Cov}\left[\mathcal{C}, \mathcal{C}'\right] = \sum_{i\neq j, k\neq l} \omega_{i,j}\omega_{k,l}\,\mathrm{Cov}\left[U_2(X_i, X_j), U_2'(X_k, X_l)\right] \\
&+ \sum_{i\neq j, k\neq l} \mathrm{Cov}\left[U_1(X_i, X_j), U_1'(X_k, X_l)\right] + \frac{1}{n^2}\sum_{1\leq i,j\leq n} \mathrm{Cov}\left[f(X_i), f'(X_j)\right]
\end{aligned}$$

$$
\begin{aligned}
&+ \sum_{1 \leqslant i \neq j \leqslant n} \omega_{i,j} \operatorname{Cov}\left[P_n f + \sum_{k \neq l} U_1(X_k, X_l), U_2'(X_i, X_j) \right] \\
&+ \sum_{1 \leqslant i \neq j \leqslant n} \omega_{i,j} \operatorname{Cov}\left[U_2(X_i, X_j), P_n f' + \sum_{k \neq l} U_1'(X_k, X_l) \right] \\
&+ \frac{1}{n} \sum_{1 \leqslant i, (k \neq l) \leqslant n} \left(\operatorname{Cov}\left[f(X_i), U_1'(X_k, X_l) \right] + \operatorname{Cov}\left[U_1(X_k, X_l), f'(X_i) \right] \right) \quad .
\end{aligned}
$$

We clearly have

$$
\begin{aligned}
&\frac{1}{n^2} \sum_{1 \leq i, j \leq n} \operatorname{Cov}\left[f(X_i), f'(X_j) \right] = \frac{1}{n} \operatorname{Cov}\left[f(X), f'(X) \right] \\
&\frac{1}{n} \sum_{1 \leqslant i, (k \neq l) \leqslant n} \operatorname{Cov}\left[f(X_i), U_1'(X_k, X_l) \right] = 2(n-1) \operatorname{Cov}\left[f(X), U_1'(X, Y) \right] \quad .
\end{aligned}
$$

Moreover, by symmetry of U_1 and U_2,

$$
\begin{aligned}
&\sum_{1 \leqslant i \neq j, k \neq l \leqslant n} \operatorname{Cov}\left[U_1(X_i, X_j), U_1'(X_k, X_l) \right] = \\
&\quad 2n(n-1) \operatorname{Cov}\left[U_1(X, Y), U_1'(X, Y) \right] + 4n(n-1)(n-2) \operatorname{Cov}\left[U_1(X, Y), U_1'(Y, Z) \right] \quad ,
\end{aligned}
$$

and

$$
\begin{aligned}
&\sum_{1 \leqslant i \neq j, k \neq l \leqslant n} \omega_{i,j} \omega_{k,l} \operatorname{Cov}\left[U_2(X_i, X_j), U_2'(X_k, X_l) \right] \\
&\qquad = 2 \operatorname{Cov}\left[U_2(X, Y), U_2'(X, Y) \right] \sum_{1 \leqslant i \neq j \leqslant n} \omega_{i,j}^2 \\
&\qquad\quad + 4 \operatorname{Cov}\left[U_2(X, Y), U_2'(Y, Z) \right] \sum_{1 \leqslant i \neq j \leqslant n} \omega_{i,j} \sum_{k=1, k \neq (i,j)}^{n} \omega_{i,k} \quad .
\end{aligned}
$$

We conclude the proof computing the last term

$$
\begin{aligned}
&\sum_{1 \leqslant i \neq j \leqslant n} \omega_{i,j} \operatorname{Cov}\left[U_2(X_i, X_j), P_n f' + \sum_{1 \leqslant k \neq l \leqslant n} U_1'(X_k, X_l) \right] \\
&= \frac{1}{n} \sum_{k, (i \neq j)} \omega_{i,j} \operatorname{Cov}\left[U_2(X_i, X_j), f'(X_k) \right] + \sum_{i \neq j} \omega_{i,j} \operatorname{Cov}\left[U_2(X_i, X_j), \sum_{k \neq l} U_1'(X_k, X_l) \right] \\
&= 2 \operatorname{Cov}\left[U_2(X, Y), \frac{1}{n} f'(X) + U_1'(X, Y) + 2(n-2) U_1'(Y, Z) \right] \sum_{1 \leqslant i \neq j \leqslant n} \omega_{i,j} \quad .
\end{aligned}
$$

□

4.5.2 Weights

Let us now compute the quantities appearing in Lemma 4.2 for the following weights

$$\omega_{i,j}^{\mathrm{VF}} = 2(F\rho_{i,j}^{(\mathrm{VF})} - 1) \qquad \text{and} \qquad \omega_{i,j}^{\mathrm{VFCV}} = \left(1 - \frac{1}{V-1}\mathbf{1}_{K_i \neq K_j}\right) \ .$$

Penalized VF weights

Lemma 4.3.

$$\sum_{1 \leqslant i \neq j \leqslant n} \omega_{i,j}^{\mathrm{VF}} = -2n^2 \left(1 + \frac{(F-1)}{n}\right) \ ,$$

$$\sum_{1 \leqslant i \neq j \leqslant n} \left(\omega_{i,j}^{\mathrm{VF}}\right)^2 = 4n^2 \left(1 + \frac{F^2}{V-1} - \frac{(F-1)^2}{n}\right) \ ,$$

$$\sum_{1 \leqslant i \neq j \leqslant n} \omega_{i,j}^{\mathrm{VF}} \sum_{k=1, k \neq (i,j)}^{n} \omega_{i,k}^{\mathrm{VF}} = 4n^2 \left(\frac{2(F-1)^2}{n} - \frac{F^2}{V-1} + 2(F-1) + n - 1\right) \ .$$

Proof: From Section A.4 in Arlot & Lerasle (2014), we know that

$$\sum_{i=1}^{n} \rho_{i,i}^{(\mathrm{VF})} = n, \qquad \sum_{i=1}^{n} \left(\rho_{i,i}^{(\mathrm{VF})}\right)^2 = n \ , \tag{4.8}$$

$$\sum_{1 \leqslant i \neq j \leqslant n} \rho_{i,j}^{(\mathrm{VF})} = -n, \qquad \sum_{1 \leqslant i \neq j \leqslant n} \left(\rho_{i,j}^{(\mathrm{VF})}\right)^2 = \frac{n^2}{V-1} - n \ . \tag{4.9}$$

It follows immediately that

$$\sum_{1 \leqslant i \neq j \leqslant n} \omega_{i,j}^{\mathrm{VF}} = -2n^2 \left(1 + \frac{1}{n}(F-1)\right) \ ,$$

$$\sum_{1 \leqslant i \neq j \leqslant n} \left(\omega_{i,j}^{\mathrm{VF}}\right)^2 = 4n^2 \left(1 + \frac{F^2}{V-1} - \frac{1}{n}(F-1)^2\right) \ .$$

Finally, let us consider the last expression in the lemma

$$\sum_{1 \leqslant i \neq j \leqslant n} \omega_{i,j}^{\mathrm{VF}} \sum_{k=1, k \neq (i,j)}^{n} \omega_{i,k}^{\mathrm{VF}} = 4 \sum_{i \neq j} (F\rho_{i,j}^{(\mathrm{VF})} - 1) \sum_{k=1, k \neq (i,j)}^{n} (F\rho_{i,k}^{(\mathrm{VF})} - 1)$$

$$= 4F^2 \sum_{i \neq j} \sum_{k=1, k \neq (i,j)}^{n} \rho_{i,j}^{(\mathrm{VF})} \rho_{i,k}^{(\mathrm{VF})} + 8Fn(n-2) + 4n(n-1)(n-2) \ .$$

We have thanks to (4.8) and (4.9)

$$\sum_{i \neq j} \sum_{k=1, k \neq (i,j)}^{n} \rho_{i,j}^{(\mathrm{VF})} \rho_{i,k}^{(\mathrm{VF})} = \sum_{j=1}^{n} \left(\sum_{1 \leqslant i \neq k \leqslant n:\ i,k \neq j} \rho_{i,j}^{(\mathrm{VF})} \rho_{i,k}^{(\mathrm{VF})} \right)$$

$$
\begin{aligned}
&= \sum_{j=1}^{n} \left(\sum_{1 \leqslant i \neq k \leqslant n} \rho_{i,j}^{(\mathrm{VF})} \rho_{i,k}^{(\mathrm{VF})} - \sum_{k=1,\ k \neq j}^{n} \rho_{j,j}^{(\mathrm{VF})} \rho_{j,k}^{(\mathrm{VF})} - \sum_{i=1,\ i \neq j}^{n} \rho_{i,j}^{(\mathrm{VF})} \rho_{i,j}^{(\mathrm{VF})} \right) \\
&= \sum_{j=1}^{n} \left(\sum_{1 \leqslant i \neq k \leqslant n} \rho_{i,j}^{(\mathrm{VF})} \rho_{i,k}^{(\mathrm{VF})} \right) - \sum_{j=1}^{n} \sum_{k=1,\ k \neq j}^{n} \rho_{j,k}^{(\mathrm{VF})} - \sum_{j=1}^{n} \sum_{i=1,\ i \neq j}^{n} \left(\rho_{i,j}^{(\mathrm{VF})} \right)^2 \\
&= \sum_{j=1}^{n} \left(\sum_{1 \leqslant i \neq k \leqslant n} \rho_{i,j}^{(\mathrm{VF})} \rho_{i,k}^{(\mathrm{VF})} \right) + n - \left(\frac{n^2}{V-1} - n \right) \\
&= \sum_{j=1}^{n} \left(\sum_{1 \leqslant i \neq k \leqslant n} \rho_{i,j}^{(\mathrm{VF})} \rho_{i,k}^{(\mathrm{VF})} \right) + 2n - \frac{n^2}{V-1} \ .
\end{aligned}
$$

We conclude with

$$
\sum_{1 \leqslant i,j,k \leqslant n} \rho_{i,j}^{(\mathrm{VF})} \left(1 - \frac{V}{V-1} \mathbf{1}_{K_i \neq K_k} \right) = -n \sum_{1 \leqslant i,j \leqslant n} \rho_{i,j}^{(\mathrm{VF})} = 0 \ .
$$

□

VFCV weights

Lemma 4.4.

$$
\begin{aligned}
\sum_{1 \leqslant i \neq j \leqslant n} \omega_{i,j}^{\mathrm{VFCV}} &= \frac{n^2(V-1)}{V} - n \ , \\
\sum_{1 \leqslant i \neq j \leqslant n} \left(\omega_{i,j}^{\mathrm{VFCV}} \right)^2 &= \frac{n^2}{V} \left(1 + \frac{(V-2)^2}{V-1} \right) - n \ , \\
\sum_{1 \leqslant i \neq j \leqslant n} \omega_{i,j}^{\mathrm{VFCV}} \sum_{k=1, k \neq (i,j)}^{n} \omega_{i,k}^{\mathrm{VFCV}} &= \frac{n^3(V-1)^2}{V^2} - \frac{n^2}{V} \left(2V - 1 + \frac{(V-2)^2}{V-1} \right) + 2n \ .
\end{aligned}
$$

Proof:

Since all blocks B_I have the same size n/V, we have

$$
\sum_{i=1}^{n} \omega_{i,i}^{\mathrm{VFCV}} = n, \qquad \sum_{i=1}^{n} \left(\omega_{i,i}^{\mathrm{VFCV}} \right)^2 = n \ .
$$

Using in addition that,

$$
\sum_{i=1}^{n} \omega_{i,i}^{\mathrm{VFCV}} + \sum_{i \neq j} \omega_{i,j}^{\mathrm{VFCV}} = \sum_{i,j} \omega_{i,j}^{\mathrm{VFCV}} = n \left(\frac{n}{V} + \frac{n(V-1)}{V} \frac{V-2}{V-1} \right) = \frac{n^2(V-1)}{V} \ ,
$$

we find, $\sum_{i\neq j}\omega_{i,j}^{\mathrm{VFCV}}=\frac{n^2(V-1)}{V}-n$, and

$$\sum_{i\neq j}\left(\omega_{i,j}^{\mathrm{VFCV}}\right)^2=n\left(\frac{n}{V}+\frac{n(V-1)}{V}\left(\frac{V-2}{V-1}\right)^2\right)-n=\frac{n^2}{V}\left(1+\frac{(V-2)^2}{V-1}\right)-n\ .$$

Let us now consider the last expression appearing in the lemma

$$\begin{aligned}&\sum_{1\leqslant i\neq j\leqslant n}\omega_{i,j}^{\mathrm{VFCV}}\sum_{k=1,k\neq(i,j)}^{n}\omega_{i,k}^{\mathrm{VFCV}}\\&=\sum_{j=1}^{n}\left(\sum_{i\neq k}\omega_{i,j}^{\mathrm{VFCV}}\omega_{i,k}^{\mathrm{VFCV}}-\sum_{1\leqslant k\leqslant n,\ k\neq j}\omega_{j,k}^{\mathrm{VFCV}}-\sum_{1\leqslant i\leqslant n,\ i\neq j}\left(\omega_{i,j}^{\mathrm{VFCV}}\right)^2\right)\\&=\sum_{i,j,k}\omega_{i,j}^{\mathrm{VFCV}}\omega_{i,k}^{\mathrm{VFCV}}-\sum_{i,j}\omega_{i,j}^{\mathrm{VFCV}}-\frac{n^2(V-1)}{V}-\frac{n^2}{V}\left(1+\frac{(V-2)^2}{V-1}\right)+2n\ .\end{aligned}$$

By definition $\omega_{i,k}^{\mathrm{VFCV}}=\left(1-\frac{1}{V-1}\mathbf{1}_{K_i\neq K_k}\right)$, hence

$$\begin{aligned}&\sum_{i,j,k}\omega_{i,j}^{\mathrm{VFCV}}\omega_{i,k}^{\mathrm{VFCV}}=\sum_{k=1}^{n}\left(\sum_{i,j}\omega_{i,j}^{\mathrm{VFCV}}\right)-\sum_{i,j}\frac{1}{V-1}\omega_{i,j}^{\mathrm{VFCV}}\left(\sum_{k=1}^{n}\mathbf{1}_{K_i\neq K_k}\right)\\&=\frac{n^3(V-1)}{V}-\frac{n^3(V-1)}{V^2}=\frac{n^3(V-1)^2}{V^2}\ .\end{aligned}$$

Finally we find

$$\begin{aligned}&\sum_{1\leqslant i\neq j\leqslant n}\omega_{i,j}^{\mathrm{VFCV}}\sum_{k=1,k\neq(i,j)}^{n}\omega_{i,k}^{\mathrm{VFCV}}\\&\qquad=\frac{n^3(V-1)^2}{V^2}-\frac{2n^2(V-1)}{V}-\frac{n^2}{V}\left(1+\frac{(V-2)^2}{V-1}\right)+2n\ .\end{aligned}$$

□

4.5.3 Proofs of the theorems

Proof of Theorem 4.3

From (4.4), we see that $\Delta_{V,F}(m,m')$ has the form of a function defined in Lemma 4.2 above with $\mathcal{C}=\mathcal{C}'$ and $\omega_{i,j}=\omega_{i,j}^{\mathrm{VF}}$,

$$U_1(x,y)=\frac{1}{n^2}\Delta^A(x,y),\qquad U_2(x,y)=\frac{1}{n^2}\Delta^{\mathcal{K}}(x,y)\ ,$$

$$\text{and}\qquad f(y)=\frac{1}{n}\left(\Delta^A(y,y)+2(F-1)\Delta^{\mathcal{K}}(y,y)\right)\ .$$

Hence from Lemma 4.2 we find,

$$\begin{aligned}
&\mathrm{Var}\left[\Delta_{V,F}(m,m')\right] = \frac{1}{n^3}\mathrm{Var}\left[\Delta^A(X,X) + 2(F-1)\Delta^{\mathcal{K}}(X,X)\right] \\
&+ \frac{2(n-1)}{n^3}\mathrm{Var}\left[\Delta^A(X,Y)\right] + \frac{2}{n^4}\sum_{1\leqslant i\neq j\leqslant n}(\omega_{i,j})^2\,\mathrm{Var}\left[\Delta^{\mathcal{K}}(X,Y)\right] \\
&+ \frac{4(n-1)}{n^3}\,\mathrm{Cov}\left[\Delta^A(X,Y), (n-2)\Delta^A(Y,Z) + \Delta^A(X,X) + 2(F-1)\Delta^{\mathcal{K}}(X,X)\right] \\
&+ \frac{4}{n^4}\sum_{1\leqslant i\neq j\leqslant n}\omega_{i,j}\sum_{k=1,k\neq(i,j)}^{n}\omega_{i,k}\,\mathrm{Cov}\left[\Delta^{\mathcal{K}}(X,Y), \Delta^{\mathcal{K}}(Y,Z)\right] \\
&+ \frac{4}{n^4}\sum_{1\leqslant i\neq j\leqslant n}\omega_{i,j}\,\mathrm{Cov}\left[\Delta^{\mathcal{K}}(X,Y), \Delta^A(X,X) + 2(F-1)\Delta^{\mathcal{K}}(X,X)\right] \\
&+ \frac{4}{n^4}\sum_{1\leqslant i\neq j\leqslant n}\omega_{i,j}\,\mathrm{Cov}\left[\Delta^{\mathcal{K}}(X,Y), \Delta^A(X,Y) + 2(n-2)\Delta^A(Y,Z)\right] \ .
\end{aligned}$$

Now we conclude using Lemma 4.3

$$\begin{aligned}
&\mathrm{Var}\left[\Delta_{V,F}(m,m')\right] = \frac{1}{n^3}\mathrm{Var}\left[\Delta^A(X,X) + 2(F-1)\Delta^{\mathcal{K}}(X,X)\right] \\
&+ \frac{2(n-1)}{n^3}\mathrm{Var}\left[\Delta^A(X,Y)\right] + \frac{8}{n^2}\left(1 + \frac{F^2}{V-1} - \frac{(F-1)^2}{n}\right)\mathrm{Var}\left[\Delta^{\mathcal{K}}(X,Y)\right] \\
&+ \frac{4(n-1)}{n^3}\,\mathrm{Cov}\left[\Delta^A(X,Y), (n-2)\Delta^A(Y,Z) + \Delta^A(X,X) + 2(F-1)\Delta^{\mathcal{K}}(X,X)\right] \\
&+ \frac{16}{n^2}\left(\frac{2(F-1)^2}{n} - \frac{F^2}{V-1} + 2(F-1) + n - 1\right)\mathrm{Cov}\left[\Delta^{\mathcal{K}}(X,Y), \Delta^{\mathcal{K}}(Y,Z)\right] \\
&- \frac{8}{n^2}\left(1 + \frac{(F-1)}{n}\right)\mathrm{Cov}\left[\Delta^{\mathcal{K}}(X,Y), \Delta^A(X,X) + 2(F-1)\Delta^{\mathcal{K}}(X,X)\right] \\
&- \frac{8}{n^2}\left(1 + \frac{(F-1)}{n}\right)\mathrm{Cov}\left[\Delta^{\mathcal{K}}(X,Y), \Delta^A(X,Y) + 2(n-2)\Delta^A(Y,Z)\right] \ .
\end{aligned}$$

Proof of Theorem 4.1

From (4.6), we see that $\Delta_V^{\mathrm{LSVF}}(m,m')$ has the form of a function defined in Lemma 4.2 above with $\mathcal{C} = \mathcal{C}'$ and $\omega_{i,j} = \omega_{i,j}^{\mathrm{VFCV}}$,

$$f(x) = \frac{V}{n(V-1)}\Delta^A(x,x), \qquad U_1(x,y) = -\frac{2V}{n^2}\Delta^{\mathcal{K}}(x,y) \ ,$$

$$\text{and} \qquad U_2(x,y) = \frac{V}{n^2(V-1)}\left(\Delta^A(x,y) + 2(V-1)\Delta^{\mathcal{K}}(x,y)\right) \ .$$

Hence from Lemma 4.2 we find,

$$\mathrm{Var}\left[\Delta_V^{\mathrm{LSVF}}(m,m')\right] = \frac{V^2}{n^3(V-1)^2}\mathrm{Var}\left[\Delta^A(X,X)\right]$$

$$
\begin{aligned}
&+ \frac{8V^2}{n^4}\left(n(n-1) - 2\sum_{1\leqslant i\neq j\leqslant n}\omega_{i,j}\right)\operatorname{Var}\left[\Delta^{\mathcal{K}}(X,Y)\right] \\
&+ \frac{2V^2}{n^4(V-1)^2}\left(\sum_{1\leqslant i\neq j\leqslant n}(\omega_{i,j})^2\right)\operatorname{Var}\left[\Delta^{A}(X,Y) + 2(V-1)\Delta^{\mathcal{K}}(X,Y)\right] \\
&+ \frac{8V^2}{n^4(V-1)}\left(\sum_{1\leqslant i\neq j\leqslant n}\omega_{i,j} - n(n-1)\right)\operatorname{Cov}\left[\Delta^{\mathcal{K}}(X,Y), \Delta^{A}(X,X)\right] \\
&+ \frac{16V^2}{n^4}\left((n-2)\left(n(n-1) - 2\sum_{1\leqslant i\neq j\leqslant n}\omega_{i,j}\right) + \sum_{1\leqslant i\neq j\leqslant n}\omega_{i,j}\sum_{k=1,k\neq(i,j)}^{n}\omega_{i,k}\right) \\
&\qquad\qquad \operatorname{Cov}\left[\Delta^{\mathcal{K}}(X,Y), \Delta^{\mathcal{K}}(Y,Z)\right] \\
&+ \frac{16V^2}{n^4(V-1)}\left(\sum_{1\leqslant i\neq j\leqslant n}\omega_{i,j}\sum_{k=1,k\neq(i,j)}^{n}\omega_{i,k} - (n-2)\sum_{1\leqslant i\neq j\leqslant n}\omega_{i,j}\right)\operatorname{Cov}\left[\Delta^{A}(X,Y), \Delta^{\mathcal{K}}(Y,Z)\right] \\
&+ \frac{4V^2}{n^4(V-1)^2}\left(\sum_{1\leqslant i\neq j\leqslant n}\omega_{i,j}\sum_{k=1,k\neq(i,j)}^{n}\omega_{i,k}\right)\operatorname{Cov}\left[\Delta^{A}(X,Y), \Delta^{A}(Y,Z)\right] \\
&+ \frac{4V^2}{n^4(V-1)^2}\left(\sum_{1\leqslant i\neq j\leqslant n}\omega_{i,j}\right)\operatorname{Cov}\left[\Delta^{A}(X,Y), \Delta^{A}(X,X) - 2(V-1)\Delta^{\mathcal{K}}(X,Y)\right] \quad .
\end{aligned}
$$

We conclude using Lemma 4.4 and some easy computations

$$
\begin{aligned}
\left(n(n-1) - 2\sum_{1\leqslant i\neq j\leqslant n}\omega_{i,j}^{\mathrm{VFCV}}\right) &= \frac{n(V-nV+2n)}{V} \quad , \\
\frac{4V^2}{n^4(V-1)^2}\left(\sum_{1\leqslant i\neq j\leqslant n}\omega_{i,j}^{\mathrm{VFCV}}\right) &= \frac{4V}{n^2(V-1)} - \frac{4V^2}{n^3(V-1)^2} \quad , \\
\frac{2V^2}{n^4(V-1)^2}\left(\sum_{1\leqslant i\neq j\leqslant n}\left(\omega_{i,j}^{\mathrm{VFCV}}\right)^2\right) &= \frac{2V}{n^2(V-1)^2}\left(1 + \frac{(V-2)^2}{V-1}\right) - \frac{2V^2}{n^3(V-1)^2} \quad , \\
\frac{8V^2}{n^4(V-1)}\left(\sum_{1\leqslant i\neq j\leqslant n}\omega_{i,j}^{\mathrm{VFCV}} - n(n-1)\right) &= \frac{-8V}{n^2(V-1)} \quad ,
\end{aligned}
$$

and,

$$
\begin{aligned}
&\frac{16V^2}{n^4(V-1)}\left(\sum_{1\leqslant i\neq j\leqslant n}\omega_{i,j}^{\mathrm{VFCV}}\sum_{k=1,k\neq(i,j)}^{n}\omega_{i,k}^{\mathrm{VFCV}} - (n-2)\sum_{1\leqslant i\neq j\leqslant n}\omega_{i,j}^{\mathrm{VFCV}}\right) \\
&= \frac{16V}{n^2(V-1)}\left(\frac{(V-1)(V-n)}{V} - \frac{(V-2)^2}{V-1}\right)
\end{aligned}
$$

$$\frac{V^2}{n^2}\left(n(n-1)(n-2) - 2(n-2)\sum_{1\leqslant i\neq j\leqslant n}\omega_{i,j}^{\mathrm{VFCV}} + \sum_{i\neq j}\sum_{k\neq(i,j)}\omega_{i,j}^{\mathrm{VFCV}}\omega_{i,k}^{\mathrm{VFCV}}\right) = n - \frac{V}{V-1} \ .$$

Proof of Theorem 4.2

The proof is straightforward. Indeed, from (4.7) one can use Lemma 4.2, with $U_2 \equiv 0$ and

$$f(y) = \frac{1}{n(n-p)}\Delta^A(y,y), \quad U_1(x,y) = \frac{1}{n(n-1)}\left(\left(1 - \frac{1}{n-p}\right)\Delta^A(x,y) - 2\Delta^{\mathcal{K}}(x,y)\right) \ .$$

Proof of Theorem 4.4

From (4.5), we see that $\Delta_{\mathrm{id}}(m,m')$ has the form of a function defined in Lemma 4.2 above with

$$f(y) = \frac{1}{n}\Delta^A(y,y) - 2\Delta^s(y), \qquad U_1(x,y) = \frac{1}{n^2}\Delta^A(x,y), \qquad \text{and} \qquad U_2 \equiv 0 \ .$$

Hence

$$\begin{aligned}\mathrm{Var}\left[\Delta_{\mathrm{id}}(m,m')\right] &= \frac{1}{n^3}\mathrm{Var}\left[\Delta^A(X,X) - 2n\Delta^s(X)\right] + \frac{2(n-1)}{n^3}\mathrm{Var}\left[\Delta^A(X,Y)\right] \\ &+ \frac{4(n-1)}{n^3}\,\mathrm{Cov}\left[\Delta^A(X,Y), (n-2)\Delta^A(Y,Z) + \Delta^A(X,X) - 2n\Delta^s(X)\right] \ .\end{aligned}$$

Moreover, up to an additive constant term $\Delta_{\mathrm{id,det}}(m,m')$ has the form of a function defined in Lemma 4.2 above with $U_2 \equiv 0$

$$U_1(x,y) = \frac{1}{n^2}\left(\Delta^A(x,y) - 2\Delta^{\mathcal{K}}(x,y)\right) \qquad \text{and} \qquad f(y) = \frac{1}{n}\left(\Delta^A(y,y) - 2\Delta^{\mathcal{K}}(y,y)\right) \ .$$

Hence,

$$\begin{aligned}&\mathrm{Var}\left[\Delta_{\mathrm{id,det}}(m,m')\right] \\ &= \frac{1}{n^3}\mathrm{Var}\left[\Delta^A(X,X) - 2\Delta^{\mathcal{K}}(X,X)\right] + \frac{2(n-1)}{n^3}\mathrm{Var}\left[\Delta^A(X,Y) - 2\Delta^{\mathcal{K}}(X,Y)\right] \\ &+ \frac{4(n-1)}{n^3}\,\mathrm{Cov}\left[\Delta^A(X,Y) - 2\Delta^{\mathcal{K}}(X,Y), \Delta^A(X,X) - 2\Delta^{\mathcal{K}}(X,X)\right] \\ &+ \frac{4(n-1)(n-2)}{n^3}\,\mathrm{Cov}\left[\Delta^A(X,Y) - 2\Delta^{\mathcal{K}}(X,Y), \Delta^A(Y,Z) - 2\Delta^{\mathcal{K}}(Y,Z)\right] \ .\end{aligned}$$

4.5.4 Proof of Lemma 4.1

Proof: We remember that $\mathrm{Var}\left[X \mid Y\right] := \mathbb{E}\left[(X - \mathbb{E}\left[X \mid Y\right])^2 \,\middle|\, Y\right]$.

- Let us prove that for any symmetric function K

$$\begin{aligned}&\mathrm{Var}\left[K(X,Y)\right] - 2\,\mathrm{Cov}\left[K(X,Y), K(Y,Z)\right] \\ &\qquad = \mathbb{E}\left[\mathrm{Var}\left[K(X,Y) \mid Y\right]\right] - \mathrm{Var}\left[\mathbb{E}\left[K(X,Y) \mid Y\right]\right] \ .\end{aligned}$$

This equality comes from some easy computations. First, we remind the classic relation

$$\mathrm{Var}\left[K(X,Y)\right] = \mathbb{E}\left[\mathrm{Var}\left[K(X,Y)\,|\,Y\right]\right] + \mathrm{Var}\left[\mathbb{E}\left[K(X,Y)\,|\,Y\right]\right] \ ,$$

second, denoting $E = \mathbb{E}\left[K(X,Y)\right]$, we find that

$$\begin{aligned}
\mathrm{Cov}\left[K(X,Y),K(Y,Z)\right] &= \mathbb{E}\left[\left(K(X,Y)-\mathbb{E}\left[K(X,Y)\right]\right)\left(K(Z,Y)-\mathbb{E}\left[K(Z,Y)\right]\right)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\left(K(X,Y)-E\right)\left(K(Z,Y)-E\right)\,|\,Y\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[K(X,Y)-E\,|\,Y\right]^2\right] \\
&= \mathbb{E}\left[\left(\mathbb{E}\left[K(X,Y)\,|\,Y\right]-E\right)^2\right] \\
&= \mathbb{E}\left[\left(\mathbb{E}\left[K(X,Y)\,|\,Y\right]-\mathbb{E}\left[\mathbb{E}\left[K(X,Y)\,|\,Y\right]\right]\right)^2\right] \\
&= \mathrm{Var}\left[\mathbb{E}\left[K(X,Y)\,|\,Y\right]\right] \ .
\end{aligned}$$

- Now, since X and Y are independent, we can write

$$\mathbb{E}\left[K(X,Y)\right] = \mathbb{E}\left[\mathbb{E}\left[K(X,Y)\,|\,X\right]|\,Y\right] \ .$$

Thus using Jensen's inequality one gets

$$\begin{aligned}
\mathrm{Var}\left[\mathbb{E}\left[K(X,Y)\,|\,Y\right]\right] &= \mathbb{E}\left[\left(\mathbb{E}\left[K(X,Y)\,|\,Y\right]-\mathbb{E}\left[\mathbb{E}\left[K(X,Y)\,|\,X\right]|\,Y\right]\right)^2\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[K(X,Y)-\mathbb{E}\left[K(X,Y)\,|\,X\right]|\,Y\right]^2\right] \\
&\leq \mathbb{E}\left[\mathbb{E}\left[\left(K(X,Y)-\mathbb{E}\left[K(X,Y)\,|\,X\right]\right)^2\,\middle|\,Y\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\left(K(X,Y)-\mathbb{E}\left[K(X,Y)\,|\,X\right]\right)^2\,\middle|\,X\right]\right] \\
&= \mathbb{E}\left[\mathrm{Var}\left[K(X,Y)\,|\,X\right]\right] = \mathbb{E}\left[\mathrm{Var}\left[K(X,Y)\,|\,Y\right]\right] \ .
\end{aligned}$$

□

4.6 Supplementary material

4.6.1 Variance of normalized increments

Let X, Y, Z be independent copies of X_1. For a given criterion $\mathcal{C}$ we define for all $m, m' \in \mathcal{M}$

$$\frac{\Delta^{\mathcal{C}}_{\mathrm{id}}(m,m')}{\sqrt{n}} := \left(\mathcal{C}(m) - \mathcal{C}_{\mathrm{id}}(m) - \left(\mathcal{C}(m') - \mathcal{C}_{\mathrm{id}}(m')\right)\right) \ . \tag{4.10}$$

Theorem 4.5. *For $\mathcal{C} = \mathcal{C}^{\rm pen}_{V,F}$, setting $\Delta^{V,F}_{\rm id}(m,m') = \Delta^{\mathcal{C}^{\rm pen}_{V,F}}_{\rm id}(m,m')$, we get the following variance formula*

$$\begin{aligned}
&\mathrm{Var}\left[\Delta^{V,F}_{\rm id}(m,m')\right] = 4\mathrm{Var}\left[\frac{(F-1)}{n}\Delta^{\mathcal{K}}(X,X) + \Delta^{s}(X)\right] \\
&+ \frac{8}{n}\left(1 + \frac{F^2}{V-1} - \frac{(F-1)^2}{n}\right)\left(\mathrm{Var}\left[\Delta^{\mathcal{K}}(X,Y)\right] - 2\,\mathrm{Cov}\left[\Delta^{\mathcal{K}}(X,Y), \Delta^{\mathcal{K}}(Y,Z)\right]\right) \\
&+ 16\left(1 + \frac{(F-1)}{n^2}\left((F-1)+2n\right)\right)\mathrm{Cov}\left[\Delta^{\mathcal{K}}(X,Y), \Delta^{\mathcal{K}}(Y,Z)\right] \\
&- 16\left(1 + \frac{(F-1)}{n}\right)\mathrm{Cov}\left[\Delta^{\mathcal{K}}(X,Y), \frac{(F-1)}{n}\Delta^{\mathcal{K}}(X,X) + \Delta^{s}(X)\right] \ .
\end{aligned}$$

In particular, for Burman's criterion, that is for $F = 1$,

$$\begin{aligned}
\mathrm{Var}\left[\Delta^{V,1}_{\rm id}(m,m')\right] &= 4\mathrm{Var}\left[\Delta^{s}(X)\right] + 16\,\mathrm{Cov}\left[\Delta^{\mathcal{K}}(X,Y), \Delta^{\mathcal{K}}(Y,Z) - \Delta^{s}(X)\right] \\
&\quad + \frac{8V}{n(V-1)}\left(\mathrm{Var}\left[\Delta^{\mathcal{K}}(X,Y)\right] - 2\,\mathrm{Cov}\left[\Delta^{\mathcal{K}}(X,Y), \Delta^{\mathcal{K}}(Y,Z)\right]\right) \ .
\end{aligned}$$

Proof: Writing

$$\begin{aligned}
&\mathcal{C}^{\rm pen}_{V,F}(m) - \mathcal{C}_{\rm id}(m) = \\
&\qquad \frac{1}{n}\sum_{i=1}^{n}\left(\frac{2(F-1)}{n}\mathcal{K}_m(X_i,X_i) + 2s_m(X_i)\right) + \frac{1}{n^2}\sum_{i\neq j=1}^{n}\omega^{\rm VF}_{i,j}\mathcal{K}_m(X_i,X_j) \ ,
\end{aligned}$$

we see from (4.10) that $\Delta^{V,F}_{\rm id}(m,m')/\sqrt{n}$ has the form of a function defined in Lemma 4.2 above with $\mathcal{C} = \mathcal{C}'$, $\omega_{i,j} = \omega^{\rm VF}_{i,j}$, $U_1 \equiv 0$,

$$f(y) = \frac{2(F-1)}{n}\Delta^{\mathcal{K}}(y,y) + 2\Delta^{s}(y), \qquad \text{and} \qquad U_2(x,y) = \frac{1}{n^2}\Delta^{\mathcal{K}}(x,y) \ .$$

We then easily derive the result from Lemma 4.2 and Lemma 4.3. □

Theorem 4.6. *For $\mathcal{C} = \mathcal{C}^{\rm LSVF}_{V}$, setting $\Delta^{{\rm LSVF},V}_{\rm id}(m,m') = \Delta^{\mathcal{C}^{\rm LSVF}_{V}}_{\rm id}(m,m')$, we get the following variance formula*

$$\begin{aligned}
&\mathrm{Var}\left[\Delta^{{\rm LSVF},V}_{\rm id}(m,m')\right] = 4\mathrm{Var}\left[\Delta^{s}(X)\right] + 16\,\mathrm{Cov}\left[\Delta^{\mathcal{K}}(X,Y), \Delta^{\mathcal{K}}(Y,Z) - \Delta^{s}(X)\right] \\
&+ \frac{8V}{n(V-1)}\left(\mathrm{Var}\left[\Delta^{\mathcal{K}}(X,Y)\right] - 2\,\mathrm{Cov}\left[\Delta^{\mathcal{K}}(X,Y), \Delta^{\mathcal{K}}(Y,Z)\right]\right) \\
&+ \frac{8}{n(V-1)^2}\,\mathrm{Cov}\left[\Delta^{\mathcal{K}}(X,Y), \Delta^{A}(X,Y) + 2(V-2)\Delta^{A}(Y,Z)\right] \\
&+ \frac{4}{n(V-1)}\,\mathrm{Cov}\left[\Delta^{A}(X,X), \Delta^{s}(X) - 2\Delta^{\mathcal{K}}(X,Y)\right] - \frac{8}{n(V-1)}\,\mathrm{Cov}\left[\Delta^{s}(X), \Delta^{A}(X,Y)\right] \\
&+ \frac{1}{n^2(V-1)^2}\mathrm{Var}\left[\Delta^{A}(X,X)\right] - \frac{4}{n^2(V-1)^2}\,\mathrm{Cov}\left[\Delta^{A}(X,Y), \Delta^{A}(X,X)\right]
\end{aligned}$$

$$+\frac{2\,(n-(V-1))}{n^2(V-1)^3}\mathrm{Var}\left[\Delta^A(X,Y)\right]-\frac{4\,(n-2(V-1))}{n^2(V-1)^3}\,\mathrm{Cov}\left[\Delta^A(X,Y),\Delta^A(Y,Z)\right]\quad.$$

Proof: First, let us write

$$\begin{aligned}\mathcal{C}_V^{\mathrm{LSVF}}(m)-\mathcal{C}_{\mathrm{id}}(m)=&\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{n(V-1)}A_m(X_i,X_i)+2s_m(X_i)\right)\\&-\frac{1}{n^2}\sum_{i\neq j=1}^{n}\left(2V\mathcal{K}_m(X_i,X_j)+A_m(X_i,X_j)\right)\\&+\frac{V}{n^2(V-1)}\sum_{i\neq j}\omega_{i,j}^{\mathrm{VFCV}}\left(A_m(X_i,X_j)+2(V-1)\mathcal{K}_m(X_i,X_j)\right)\quad.\end{aligned}$$

Now, from (4.10) we notice that $\Delta_{\mathrm{id}}^{\mathrm{LSVF},V}(m,m')/\sqrt{n}$ has the form of a function defined in Lemma 4.2, with $\mathcal{C}=\mathcal{C}'$ and $\omega_{i,j}=\omega_{i,j}^{\mathrm{VFCV}}$, and for all $x,y\in\Xi$

$$f(y)=\frac{1}{n(V-1)}\Delta^A(y,y)+2\Delta^s(y),\qquad U_1(x,y)=-\frac{1}{n^2}\left(2V\Delta^{\mathcal{K}}(x,y)+\Delta^A(x,y)\right)\quad,$$

and

$$U_2(x,y)=\frac{V}{n^2(V-1)}\left(\Delta^A(x,y)+2(V-1)\Delta^{\mathcal{K}}(x,y)\right)\quad.$$

First, using Lemma 4.4 and plugging the weights into Lemma 4.2 one gets

$$\begin{aligned}&\mathrm{Var}\left[\Delta_{\mathrm{id}}^{\mathrm{LSVF},V}(m,m')\right]=\mathrm{Var}\left[f(X)\right]+2n^2(n-1)\mathrm{Var}\left[U_1(X,Y)\right]\\&+2n^2\left(\frac{n}{V}\left(1+\frac{(V-2)^2}{V-1}\right)-1\right)\mathrm{Var}\left[U_2(X,Y)\right]\\&+4n(n-1)\,\mathrm{Cov}\left[U_1(X,Y),n(n-2)U_1(Y,Z)+f(X)\right]\\&+4n\left(\frac{n^3(V-1)^2}{V^2}-\frac{n^2}{V}\left(2V-1+\frac{(V-2)^2}{V-1}\right)+2n\right)\mathrm{Cov}\left[U_2(X,Y),U_2(Y,Z)\right]\\&+4\left(\frac{n^2(V-1)}{V}-n\right)\mathrm{Cov}\left[U_2(X,Y),f(X)+nU_1(X,Y)+2n(n-2)U_1(Y,Z)\right]\quad.\end{aligned}$$

Then, we develop the three variance terms to obtain

$$\begin{aligned}\mathrm{Var}\left[f(X)\right]=&\frac{1}{n^2(V-1)^2}\mathrm{Var}\left[\Delta^A(X,X)\right]\\&+4\mathrm{Var}\left[\Delta^s(X)\right]+\frac{4}{n(V-1)}\,\mathrm{Cov}\left[\Delta^A(X,X),\Delta^s(X)\right]\quad,\end{aligned}$$

$$\begin{aligned}\mathrm{Var}\left[U_1(X,Y)\right]=&\frac{4V^2}{n^4}\mathrm{Var}\left[\Delta^{\mathcal{K}}(X,Y)\right]\\&+\frac{1}{n^4}\mathrm{Var}\left[\Delta^A(X,Y)\right]+\frac{4V}{n^4}\,\mathrm{Cov}\left[\Delta^{\mathcal{K}}(X,Y),\Delta^A(X,Y)\right]\quad,\end{aligned}$$

$$\frac{n^4(V-1)^2}{V^2}\mathrm{Var}\left[U_2(X,Y)\right] = \mathrm{Var}\left[\Delta^A(X,Y)\right] + 4(V-1)^2\mathrm{Var}\left[\Delta^{\mathcal{K}}(X,Y)\right] + 4(V-1)\,\mathrm{Cov}\left[\Delta^A(X,Y),\Delta^{\mathcal{K}}(X,Y)\right] \ .$$

Now, we develop the covariance terms

$$n^4\,\mathrm{Cov}\left[U_1(X,Y),U_1(Y,Z)\right] = 4V^2\,\mathrm{Cov}\left[\Delta^{\mathcal{K}}(X,Y),\Delta^{\mathcal{K}}(Y,Z)\right] + 4V\,\mathrm{Cov}\left[\Delta^{\mathcal{K}}(X,Y),\Delta^A(Y,Z)\right] + \mathrm{Cov}\left[\Delta^A(X,Y),\Delta^A(Y,Z)\right] \ ,$$

$$\frac{n^4(V-1)^2}{V^2}\,\mathrm{Cov}\left[U_2(X,Y),U_2(Y,Z)\right] = 4(V-1)^2\,\mathrm{Cov}\left[\Delta^{\mathcal{K}}(X,Y),\Delta^{\mathcal{K}}(Y,Z)\right] + 4(V-1)\,\mathrm{Cov}\left[\Delta^{\mathcal{K}}(X,Y),\Delta^A(Y,Z)\right] + \mathrm{Cov}\left[\Delta^A(X,Y),\Delta^A(Y,Z)\right] \ ,$$

$$-n^2\,\mathrm{Cov}\left[f(X),U_1(X,Y)\right] = \frac{2V}{n(V-1)}\,\mathrm{Cov}\left[\Delta^A(X,X),\Delta^{\mathcal{K}}(X,Y)\right] + 2\,\mathrm{Cov}\left[\Delta^s(X),2V\Delta^{\mathcal{K}}(X,Y)+\Delta^A(X,Y)\right] + \frac{1}{n(V-1)}\,\mathrm{Cov}\left[\Delta^A(X,X),\Delta^A(X,Y)\right] \ ,$$

$$\frac{n^2(V-1)}{V}\,\mathrm{Cov}\left[U_2(X,Y),f(X)\right] = 4(V-1)\,\mathrm{Cov}\left[\Delta^{\mathcal{K}}(X,Y),\Delta^s(X)\right] + 2\,\mathrm{Cov}\left[\Delta^A(X,Y),\Delta^s(X)\right] + \frac{1}{n}\,\mathrm{Cov}\left[2\Delta^{\mathcal{K}}(X,Y)+\frac{1}{V-1}\Delta^A(X,Y),\Delta^A(X,X)\right] \ ,$$

$$\frac{n^4(V-1)}{V}\,\mathrm{Cov}\left[U_2(X,Y),U_1(X,Y)\right] = -4V(V-1)\mathrm{Var}\left[\Delta^{\mathcal{K}}(X,Y)\right] - \mathrm{Var}\left[\Delta^A(X,Y)\right] - 2(2V-1)\,\mathrm{Cov}\left[\Delta^{\mathcal{K}}(X,Y),\Delta^A(X,Y)\right] \ ,$$

$$\frac{n^4(V-1)}{V}\,\mathrm{Cov}\left[U_2(X,Y),U_1(Y,Z)\right] = -4V(V-1)\,\mathrm{Cov}\left[\Delta^{\mathcal{K}}(X,Y),\Delta^{\mathcal{K}}(Y,Z)\right] - \mathrm{Cov}\left[\Delta^A(X,Y),\Delta^A(Y,Z)\right] - 2(2V-1)\,\mathrm{Cov}\left[\Delta^{\mathcal{K}}(X,Y),\Delta^A(Y,Z)\right] \ .$$

Thus, we find

- *Terms in front of* $\mathrm{Var}\left[\Delta^{\mathcal{K}}(X,Y)\right]$

$$\frac{8V^2}{n^2}\left(n-1+\frac{n}{V}\left(1+\frac{(V-2)^2}{V-1}\right)-1-2\left(\frac{n(V-1)}{V}-1\right)\right) = \frac{8V}{n(V-1)} \ .$$

- *Terms in front of* $\mathrm{Var}\left[\Delta^A(X,Y)\right]$

$$\frac{2}{n^2}\left(n-1+\frac{nV}{(V-1)^2}\left(1+\frac{(V-2)^2}{V-1}\right)-\frac{V^2}{(V-1)^2}-\frac{2V}{V-1}\left(\frac{n(V-1)}{V}-1\right)\right)$$
$$=\frac{2}{n^2}\left(\frac{nV(V^2+3-3V)-n(V-1)^3}{(V-1)^3}-\frac{1}{(V-1)^2}\right)=\frac{2\,(n-(V-1))}{n^2(V-1)^3}\ .$$

- *Terms in front of* $\mathrm{Cov}\left[\Delta^{\mathcal{K}}(X,Y),\Delta^{s}(X)\right]$

$$\frac{-16V(n-1)}{n}+\frac{16V}{n}\left(\frac{n(V-1)}{V}-1\right)=-16\ .$$

- *Terms in front of* $\mathrm{Cov}\left[\Delta^{\mathcal{K}}(X,Y),\Delta^{A}(X,X)\right]$

$$\frac{-8V(n-1)}{n^2(V-1)}+\frac{8V}{n^2(V-1)}\left(\frac{n(V-1)}{V}-1\right)=\frac{-8}{n(V-1)}\ .$$

- *Terms in front of* $\mathrm{Cov}\left[\Delta^{A}(X,Y),\Delta^{s}(X)\right]$

$$\frac{8}{n(V-1)}\left(n(V-1)-V-(n-1)(V-1)\right)=-\frac{8}{n(V-1)}\ .$$

- *Terms in front of* $\mathrm{Cov}\left[\Delta^{A}(X,X),\Delta^{A}(X,Y)\right]$

$$\frac{4}{n^2(V-1)^2}\left(n(V-1)-V-(n-1)(V-1)\right)=\frac{-4}{n^2(V-1)^2}\ .$$

- *Terms in front of* $\mathrm{Cov}\left[\Delta^{\mathcal{K}}(X,Y),\Delta^{\mathcal{K}}(Y,Z)\right]$

$$\frac{16V^2}{n^2}\left(\frac{n^2(V-1)^2}{V^2}-\frac{n}{V}\left(2V-1+\frac{(V-2)^2}{V-1}\right)\right.$$
$$\left.+2+(n-1)(n-2)-2(n-2)\left(\frac{n(V-1)}{V}-1\right)\right)=\frac{16\,(nV-n-V)\,)}{n(V-1)}\ .$$

- *Terms in front of* $\mathrm{Cov}\left[\Delta^{\mathcal{K}}(X,Y),\Delta^{A}(X,Y)\right]$

$$\frac{8V}{n^2(V-1)}\left(n+\frac{n(V-2)^2}{V-1}-V\right.$$
$$\left.+(n-1)(V-1)+(2V-1)\left(1-\frac{n(V-1)}{V}\right)\right)=\frac{8}{n(V-1)^2}\ .$$

- *Terms in front of* $\mathrm{Cov}\left[\Delta^{\mathcal{K}}(X,Y),\Delta^{A}(Y,Z)\right]$

$$\frac{16V}{n^2(V-1)}\left(2V+\frac{n^2(V-1)^2}{V}-n\left(2V-1+\frac{(V-2)^2}{V-1}\right)\right.$$

$$+ (n-1)(n-2)(V-1) + (n-2)(2V-1)\left(1 - \frac{n(V-1)}{V}\right)\Bigg) = \frac{16\,(V-2)}{n(V-1)^2} \ .$$

- *Terms in front of* $\mathrm{Cov}\left[\Delta^A(X,Y), \Delta^A(Y,Z)\right]$

$$\frac{4}{n^2(V-1)^2}\Bigg((n-1)(n-2)(V-1)^2 + n^2(V-1)^2 - 2V^2 n + nV - \frac{nV(V-2)^2}{V-1} + 2V^2 - 2n(n-2)(V-1)^2 + 2V(n-2)(V-1)\Bigg) = \frac{4\,(2(V-1)-n)}{n^2(V-1)^3} \ .$$

Putting all these terms in the calculation above provides the result.

□

4.6.2 Variance of the criteria

For sake of completeness we provide here the variance computations for all the previous criteria. We find the results by replacing Δ^A and $\Delta^{\mathcal{K}}$ respectively by A_m and $\mathcal{K}_m$ in all the theorems of Section 4.3.

Theorem 4.7. *Let X, Y, Z be independent copies of X_1. First,*

$$\begin{aligned}
\mathrm{Var}\left[\mathcal{C}^{\mathrm{pen}}_{V,F}(m)\right] &= \frac{1}{n^3}\mathrm{Var}\left[A_m(X,X) + 2(F-1)\mathcal{K}_m(X,X)\right] + \frac{2(n-1)}{n^3}\mathrm{Var}\left[A_m(X,Y)\right] \\
&+ \frac{8}{n^2}\left(1 + \frac{F^2}{V-1} - \frac{(F-1)^2}{n}\right)\left(\mathrm{Var}\left[\mathcal{K}_m(X,Y)\right] - 2\,\mathrm{Cov}\left[\mathcal{K}_m(X,Y), \mathcal{K}_m(Y,Z)\right]\right) \\
&+ \frac{16}{n^3}\left((F-1)(2n+F-1) + n^2\right)\mathrm{Cov}\left[\mathcal{K}_m(X,Y), \mathcal{K}_m(Y,Z)\right] \\
&+ \frac{4(n-1)(n-2)}{n^3}\,\mathrm{Cov}\left[A_m(X,Y), A_m(Y,Z)\right] \\
&+ \frac{4(n-1)}{n^3}\,\mathrm{Cov}\left[A_m(X,X) + 2(F-1)\mathcal{K}_m(X,X), A_m(X,Y)\right] \\
&- \frac{8}{n^2}\left(1 + \frac{1}{n}(F-1)\right)\mathrm{Cov}\left[\mathcal{K}_m(X,Y), A_m(X,X) + 2(F-1)\mathcal{K}_m(X,X)\right] \\
&- \frac{8}{n^2}\left(1 + \frac{1}{n}(F-1)\right)\mathrm{Cov}\left[\mathcal{K}_m(X,Y), A_m(X,Y)\right] \\
&- \frac{16(n-2)}{n^2}\left(1 + \frac{1}{n}(F-1)\right)\mathrm{Cov}\left[\mathcal{K}_m(X,Y), A_m(Y,Z)\right] \ .
\end{aligned}$$

Second,

$$\begin{aligned}
\mathrm{Var}\left[\mathcal{C}^{\mathrm{LSVF}}_{V}(m)\right] &= \frac{V^2}{n^3(V-1)^2}\mathrm{Var}\left[A_m(X,X)\right] \\
&+ \frac{8V}{n^2(V-1)}\mathrm{Var}\left[\mathcal{K}_m(X,Y)\right] + \frac{16}{n^2}\left(n - \frac{V}{V-1}\right)\mathrm{Cov}\left[\mathcal{K}_m(X,Y), \mathcal{K}_m(Y,Z)\right]
\end{aligned}$$

$$
\begin{aligned}
&- \frac{8V}{n^2(V-1)} \operatorname{Cov}\left[A_m(X,X), \mathcal{K}_m(X,Y)\right] - \frac{8V(V-2)}{n^2(V-1)^2} \operatorname{Cov}\left[A_m(X,Y), \mathcal{K}_m(X,Y)\right] \\
&+ \frac{2V}{n^2(V-1)^2}\left(1 + \frac{(V-2)^2}{V-1} - \frac{V}{n}\right) \operatorname{Var}\left[A_m(X,Y)\right] \\
&+ \frac{4V}{n^2}\left(\frac{n}{V} - \frac{3V^2-7V+5}{(V-1)^3} + \frac{2V}{n(V-1)^2}\right) \operatorname{Cov}\left[A_m(X,Y), A_m(Y,Z)\right] \\
&+ \frac{16V}{n^2(V-1)}\left(\frac{(V-1)(V-n)}{V} - \frac{(V-2)^2}{V-1}\right) \operatorname{Cov}\left[A_m(X,Y), \mathcal{K}_m(Y,Z)\right] \\
&+ \frac{4V}{n^2(V-1)}\left(1 - \frac{V}{n(V-1)}\right) \operatorname{Cov}\left[A_m(X,Y), A_m(X,X)\right] \quad .
\end{aligned}
$$

Finally, setting $q = n - p$,

$$
\begin{aligned}
\operatorname{Var}\left[\mathcal{C}_p^{\mathrm{LPO}}(m)\right] &= \frac{1}{n^3 q^2} \operatorname{Var}\left[A_m(X,X)\right] + \frac{2}{n(n-1)} \operatorname{Var}\left[\frac{q-1}{q} A_m(X,Y) - 2\mathcal{K}_m(X,Y)\right] \\
&+ \frac{4}{n^2 q} \operatorname{Cov}\left[\frac{q-1}{q} A_m(X,Y) - 2\mathcal{K}_m(X,Y), A_m(X,X)\right] \\
&+ \frac{4(n-2)}{n(n-1)} \operatorname{Cov}\left[\frac{q-1}{q} A_m(X,Y) - 2\mathcal{K}_m(X,Y), \frac{q-1}{q} A_m(Y,Z) - 2\mathcal{K}_m(Y,Z)\right] \quad .
\end{aligned}
$$

One can prove that the formulas obtained in Celisse & Robin (2008); Celisse (2014) for projection and approximation kernels are a particular result of the latter computation.

4.6.3 More figures

Figure 4.2 shows the 10 densities we considered in our empirical study.

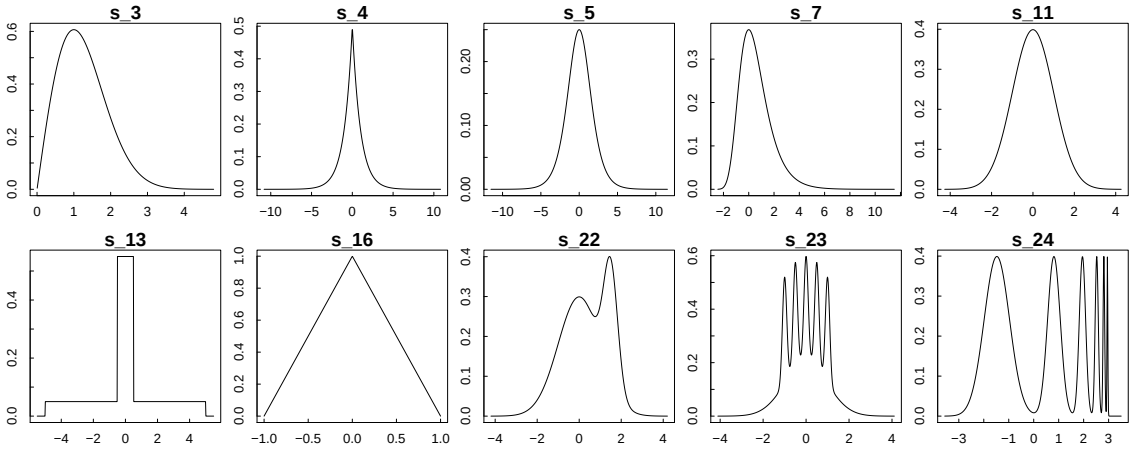

Figure 4.2: All densities considered in the paper.

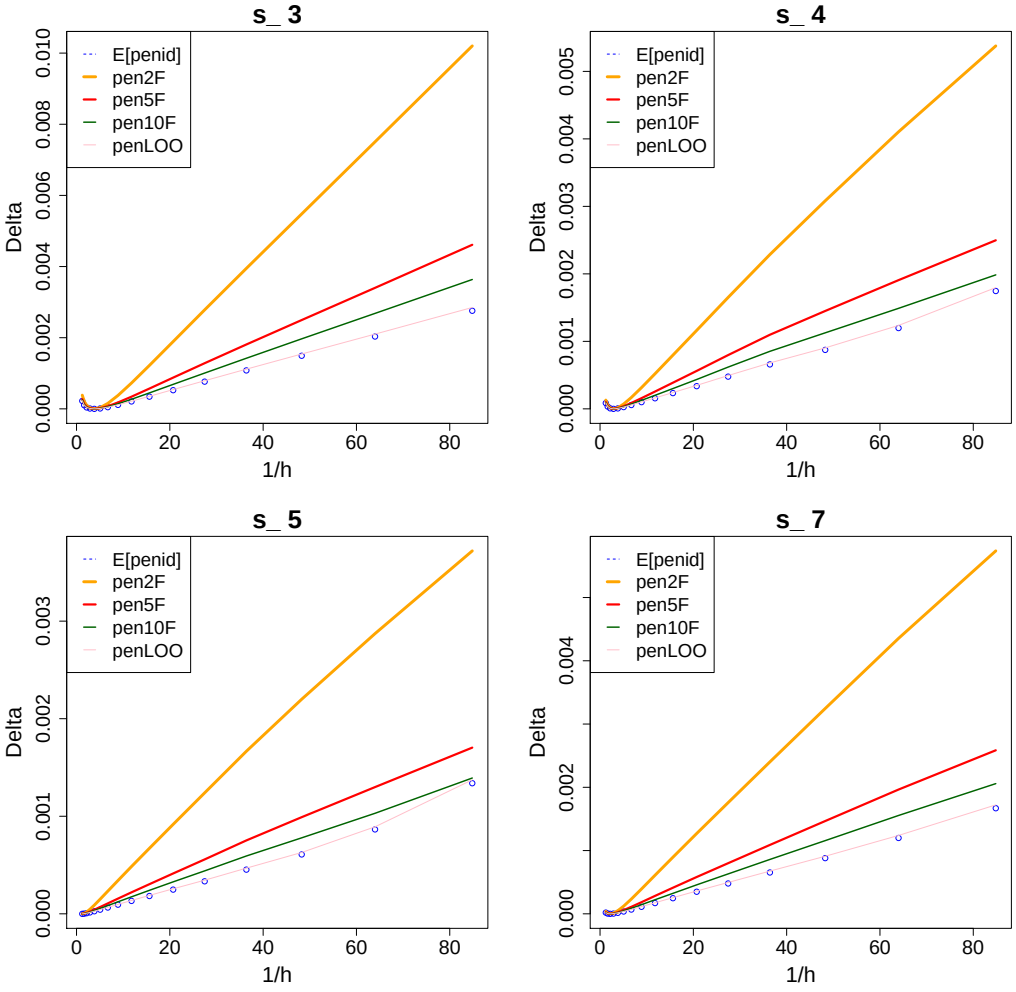

Figure 4.3: Illustration of the variance heuristic: $\mathrm{Var}\left[\Delta_{\mathcal{C}}(m, \overline{m})\right]$ as a function of $1/h_m$ for five different $\mathcal{C}$ with $n = 100$.

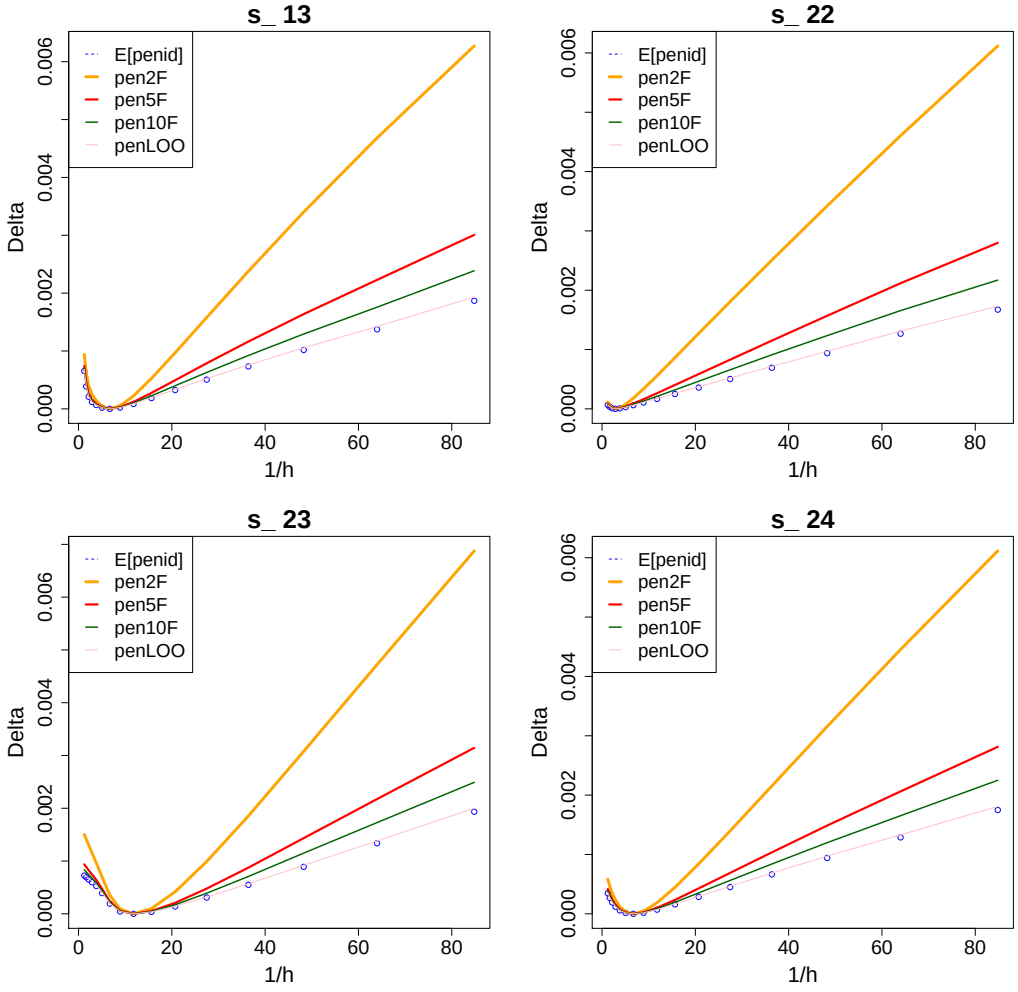

Figure 4.4: Illustration of the variance heuristic: $\mathrm{Var}\left[\Delta_{\mathcal{C}}(m, \overline{m})\right]$ as a function of $1/h_m$ for five different $\mathcal{C}$ with $n = 100$.

Part III

V-fold and robust tests

Chapter 5

Towards practical robust resampling procedures: the T-Hold-Out

Abstract. Cross-validation and data splitting techniques are widely used methods to proceed estimator selection while trying to avoid both underfitting and overfitting. The simplest case being the *hold-out*, which consists in splitting the sample into two subsamples: one is used to build a family of estimators, the other is dedicated to select among them. As an alternative to procedures based on the risk estimation principle, Birgé (2006a, Section 9) proposed to use robust tests between the preliminary estimators to select the final candidate, leading to an *apparent* quadratic cost. We introduce an efficient and exact algorithm, together with a faster but approximate version, which implements with a sub-quadratic complexity this robust hold-out. We study empirically their performance in the context of density estimation considering well-known competitors (hold-out derived from least-squares or Kullback-Leibler divergence, model selection procedures, etc.) and classical problems including histogram or bandwidth selection. We expect this algorithm to be the necessary key step in the construction of more general robust resampling methods. Our algorithms are available on the CRAN in a companion R-package in connexion with a companion website on RunMyCode to help transparency and reproducibility.

NOTA: Ce chapitre, fruit d'un travail effectué avec Yves Rozenholc[(1)], a été soumis au "Journal of the American Statistical Association" en février 2015.

Contents

[(1)]Université Paris Descartes

5.1 Introduction

Suppose we have at hand a sample of independent and identically distributed (i.i.d.) random variables from some unknown density s with respect to some dominating measure μ and that we want to estimate s from the sample.

Many papers have been published about the solution of this estimation problem with as little prior information on s as possible. A widely used strategy consists in starting from a family of preliminary estimators (for instance kernel or histogram estimators) with some varying smoothing parameter (the bandwidth or the partition) and selecting one candidate using the sample. Nevertheless, since the 30's (Larson, 1931) it has been known that building estimators and evaluating their quality with the same data yields an overoptimistic result. Many possibilities exist to treat this problem. Among others, cross-validation (and data splitting techniques in general) presents a popular and useful solution since it only requires an i.i.d. sample to be performed. The simplest procedure of this type - called *hold-out* or simple validation - consists in splitting the sample into two subsamples, building a family of estimators using the first subsample (which we shall call the training sample) and making the selection using the second subsample (which we shall call the validation sample).

Concerning the selection part, Birgé (2006a, Section 9) proposed a procedure - called *T-hold-out* hereafter - based on robust tests between the preliminary estimators. The procedure can be derived from Birgé's construction of T-estimators[(2)] oriented to model selection. The definition of these estimators is introduced in the same paper but relies on old ideas arising from Le Cam (1973); Birgé (1983, 1984b,a). Indeed, conditionally to the training sample, all the estimators are deterministic so that the models are reduced to points and the problem amounts to select one point from the validation sample. Whereas classical cross-validation procedures aim at estimating the risk of the preliminary estimators, this solution presents a rather different strategy since it associates to each estimator a plausibility index by doing a pairwise comparison between the candidates through robust tests. The –apparently– "too high computational complexity" of this construction led most authors –including its initiator– to consider it only as a theoretical tool, as pointed out in Birgé (2006a), Birgé (2007, p.45) and Baraud & Birgé (2009, p.241). However, through the simple T-hold-out procedure, we show hereafter that it is not true and that such approaches may indeed be implemented while controlling the computational cost, this being the necessary step for further extensions and generalization of the use of T-estimation based procedures.

The purpose of this paper is to provide an efficient algorithm that implements the T-hold-out, made available in our R-package called *Density.T.HoldOut* on the CRAN[(3)]. Our motivations are twofold. First, when we started this research in the summer of 2012 there was no practical application of T-estimation using the idea of data splitting[(4)]. We thought it would be of interest to compare empirically, in this simple setting, T-estimation with classical resampling and penalization procedures which are motivated by risk estimation. Second, since most cross-validation

[(2)]"T" refers to test.

[(3)]`http://cran.r-project.org/web/packages/Density.T.HoldOut/index.html`

[(4)]it should be noticed that Sart has recently applied a procedure based on robust tests in the special cases of dyadic partition selection (Sart, 2014) and parameter selection (Sart, 2013).

procedures, like V-fold cross-validation, consist in averaging Hold-Out criteria on different data splits, we believe this work to be the necessary first step towards a practical implementation of resampling procedures with robust tests.

In our empirical study, we considered several finite collections of preliminary estimators. These included histogram or kernel collections - leading to some well-known estimation problems: number of bin selection, partition selection, bandwidth selection, but also more complex collections mixing histograms and kernel estimators potentially completed with some parametric ones. The scripts, developed for this paper using our R-package, are available on the RunMyCode website[(5)] to increase transparency and reproducibility. These simulations show that T-Hold-Out represents a strong alternative to classical procedures in density estimation. Moreover, the encouraging results of this section are a source of motivation for future works on implementation of resampling procedures based on robust tests.

Hold-out is not specific to the density framework. Indeed, in all cases where we have at hand two independent random samples $\mathbf{X}_t$ and $\mathbf{X}_v$, one can build a collection of estimators using the training sample $\mathbf{X}_t$ and proceed to the selection with the validation sample $\mathbf{X}_v$. In density estimation, hold-out has been investigated theoretically for projection estimators (Arlot & Lerasle, 2014, Section 8.1) and kernel density estimates (Devroye & Lugosi, 2001) among other examples. Searching for the best linear (or convex) combination of the preliminary estimators in the validation step leads to the linear (or convex) aggregation problem (see Rigollet & Tsybakov (2007)). Moreover, theoretical properties of the hold-out have also been studied in classification (Bartlett *et al.*, 2002; Blanchard & Massart, 2006) and in regression -by Lugosi & Nobel (1999); Juditsky & Nemirovski (2000); Nemirovski (2000); Wegkamp (2003), among others.

5.1.1 Framework

Let us consider a sample $\mathbf{X} = \{X_1, \dots, X_n\}$ of i.i.d. random variables X_i with values in the measured space $(\Xi, \mathcal{Z}, \mu)$. We suppose that the distribution of X_i admits a density s with respect to μ and aim to estimate s. We turn the set $\mathcal{S}$ of all probability densities with respect to μ into a metric space using the Hellinger distance $h(t,u)$ where

$$h^2(t,u) = \frac{1}{2}\int \left(\sqrt{t(x)} - \sqrt{u(x)}\right)^2 d\mu(x) \ .$$

The quality of an approximation $t \in \mathcal{S}$ of the function s is measured by $\ell(s,t)$, where ℓ is a loss function (typically some power of a distance, as the $\mathbb{L}_q$-distances - derived from $\mathbb{L}_q$-norms denoted $\|\cdot\|_q$). The risk of an estimator $\widetilde{s} = \widetilde{s}(\mathbf{X})$ of the function s is defined through this loss function by $R_s(\widetilde{s}, \ell) := \mathbb{E}_s[\ell(s, \widetilde{s})]$, where $\mathbb{E}_s$ denotes the expectation when s obtains. The *Hellinger risk* $R_s(\widetilde{s}, h^2)$ comes from the loss $\ell = h^2$. The loss can also be defined as $\ell(s,t) = \mathbb{E}_s[\gamma(t,X) - \gamma(s,X)]$, where $\gamma : \mathcal{S} \times \Xi \mapsto [0,\infty)$ is a *contrast function* for which s appears as a minimizer of $\mathbb{E}_s[\gamma(t,X)]$ when $t \in \mathcal{S}$ (Birgé & Massart, 1993, Definition 1). In this context, the $\mathbb{L}_2$-loss (resp. the Kullback-Leibler loss) is defined via the contrast function $\gamma(t,x) = \|t\|_2^2 - 2t(x)$ (resp. $\gamma(t,x) = -\log(t(x))$) for any $t \in \mathcal{S}$, $x \in \Xi$.

[(5)] `http://www.runmycode.org/companion/view/589`

5.1.2 About the Hold-Out

Formally, the *hold-out* (HO) is a two-steps estimation procedure which relies on a split of $\mathbf{X}$ into two nonempty complementary subsamples, $\mathbf{X}_t$ and $\mathbf{X}_v$.

- **Step one**: Using the *training* sample $\mathbf{X}_t$, we build a finite set $S = \{\widehat{s}_m[\mathbf{X}_t], m \in \mathcal{M}\}$ of preliminary estimators.
- **Step two**: The *validation* sample $\mathbf{X}_v$ is dedicated to the selection of one point $\widehat{m}$ in $\mathcal{M}$.

The final estimator is either $\widehat{s}_{\widehat{m}}[\mathbf{X}_t]$ or $\widehat{s}_{\widehat{m}}[\mathbf{X}]$ depending on the authors. The goal is generally to select $\widehat{m} \in \mathcal{M}$ such that

$$R_s(\widehat{s}_{\widehat{m}}[\mathbf{X}_t], \ell) \sim \inf_{m \in \mathcal{M}} R_s(\widehat{s}_m[\mathbf{X}_t], \ell) \quad \text{or} \quad R_s(\widehat{s}_{\widehat{m}}[\mathbf{X}], \ell) \sim \inf_{m \in \mathcal{M}} R_s(\widehat{s}_m[\mathbf{X}], \ell) \ ,$$

where ℓ is the relevant loss function and the symbol $\sim$ means that quantities on both sides are of the same order.

Usually, after performing *Step one*, one defines some random criterion $\mathrm{crit}(m)$ for each m and selects the $\widehat{m} \in \mathcal{M}$ that minimizes crit. In the *classical* hold-out, when the loss ℓ is defined through a contrast function, this criterion is an estimation of the risk, made using the empirical contrast based on the validation sample:

$$\mathrm{crit}_{\mathrm{HO}}(m, \mathbf{X}_t, \mathbf{X}_v) = \frac{1}{|\mathbf{X}_v|} \sum_{X_i \in \mathbf{X}_v} \gamma(\widehat{s}_m[\mathbf{X}_t], X_i) \ ,$$

where $|A|$ denotes the cardinality of the set A. In this context one naturally selects the estimator with the smallest estimated risk,

$$\widehat{m} \in \operatorname*{argmin}_{m \in \mathcal{M}} \mathrm{crit}_{\mathrm{HO}}(m, \mathbf{X}_t, \mathbf{X}_v) \ .$$

We shall denote in what follows $\widehat{m}_{LS}$ and $\widehat{m}_{KL}$ for the estimators selected by the classical procedure using the contrast functions $\gamma(t, x) = \|t\|_2^2 - 2t(x)$ and $\gamma(t, x) = -\log(t(x))$ respectively. We call least-squares hold-out (LSHO) and Kullback-Leibler hold-out (KLHO) the corresponding HO procedures. Few theoretical results exist concerning this classical HO in the density framework. Nevertheless, considering projection estimators together with the least-squares contrast, Arlot & Lerasle (2014) have shown that the LSHO criterion can be written as a penalization criterion with some resampling-based penalty. They also proved an oracle inequality and provided variances computations for this criterion (see Theorem 3 and Section S.2. in the supplementary material in Arlot & Lerasle (2014)).

5.1.3 Overview of the paper

In practice the selection problem of *Step two* amounts to select one estimator in a given collection of $|\mathcal{M}|$ initial candidates. While the classical HO relies on the optimization of an empirical contrast

function and thus requires at most $|\mathcal{M}|$ computations, T-estimation involves pairwise comparisons based on robust tests leading to a quadratic number $\mathrm{O}(|\mathcal{M}|^2)$ of tests.

The first goal of this paper is to provide an algorithm in the general framework of T-estimation which allows an efficient and exact implementation of T-estimation in the HO context. This algorithm breaks this quadratic bound. The second goal is to compare the risk performance of this T-hold-out for three losses, a large set of densities and several sample sizes. We shall make a comparison against two types of procedures: those which select one point in a given family using the validation sample and those which estimate the density from the full sample.

Moreover, we provide a faster, albeit approximate, version of this exact algorithm. We shall study both algorithms from a computational complexity point of view as well as the risk performance of the resulting estimators.

The paper is organized as follows. In Section 2 we revisit the definition of the T-hold-out (THO) in a general framework. We introduce in Section 3 our exact and efficient algorithm which implements exact T-estimation and one approximate version derived from it. Section 4 presents the simulation protocol of our empirical study together with a short description of the main function of the companion R-package *Density.T.HoldOut*. Section 5 is dedicated to the study of the quality of the two possible T-hold-out in terms of risk. We also provide comparisons with other hold-out procedures, direct estimation procedures –penalized estimators or Lepski's method– and some bandwidth estimators obtained using asymptotic derivation of the risk. Section 6 is devoted to the empirical study of the complexity of the exact algorithm. Section 7 provides a comparison of exact and approximate algorithms both in terms of risk and complexity.

5.2 T-Hold-Out

Let us recall the T-hold-out procedure in a general framework where robust tests exist. We have at hand two independent samples, $\mathbf{X}_t$ and $\mathbf{X}_v$, and want to estimate some target s belonging to the metric space $(\mathcal{S}, d)$. Suppose that a family $S = \{\widehat{s}_m[\mathbf{X}_t], m \in \mathcal{M}\}$ of estimators of s has been built from $\mathbf{X}_t$, and we want to proceed to the selection step with $\mathbf{X}_v$. For $m_1, m_2 \in \mathcal{M}$, we write $d(m_1, m_2)$ instead of $d(\widehat{s}_{m_1}[\mathbf{X}_t], \widehat{s}_{m_2}[\mathbf{X}_t])$. Let us assume that ψ_{m_1,m_2} is a statistical test that decides between m_1 and m_2 which, conditionally to the knowledge of S, is based only on $\mathbf{X}_v$. The T-hold-out (THO) criterion is given by

$$\mathrm{crit}_{\mathrm{THO}}(m, \mathbf{X}_t, \mathbf{X}_v) := \sup_{j \in \mathcal{R}_m} d(j, m) \ , \tag{5.1}$$

with $\mathcal{R}_m$ the set of estimators preferred to m, namely $\{j \in \mathcal{M},\ j \neq m \mid \psi_{m,j} = j\}$. One finally chooses

$$\widehat{m}_{\mathrm{THO}} \in \operatorname*{argmin}_{m \in \mathcal{M}} \mathrm{crit}_{\mathrm{THO}}(m, \mathbf{X}_t, \mathbf{X}_v) \ . \tag{5.2}$$

To the best of our knowledge it is the first HO based on the Hellinger distance. There are several theoretical differences with classical HO methods. The criterion $\mathrm{crit}_{\mathrm{THO}}(m, \mathbf{X}_t, \mathbf{X}_v)$ does not estimate the risk but appears instead as a *plausibility index*. Its value is computed through robust tests between estimators, while the classical HO criterion is computed independently for each

estimator and thus does not take the geometrical structure of S into account. The key assumption in the construction is the existence of some test having the following robustness property.

Assumption A There exist two constants $a > 0, \theta \in (0, 1/2)$, such that, for any m_1 and $m_2 \in \mathcal{M}$, there exists a test $\psi_{m_1,m_2} = \psi_{m_2,m_1}$ which chooses between m_1 and m_2, and satisfies:

$$\sup_{\{\, s \in \mathcal{S} | d(s,m_1) \leqslant \theta d(m_1,m_2) \,\}} \mathbb{P}_s \left[\psi_{m_1,m_2} = m_2 \right] \leqslant \exp\left(-a |\mathbf{X}_v| d^2(m_1, m_2) \right) \quad , \tag{5.3}$$

$$\sup_{\{\, s \in \mathcal{S} | d(s,m_2) \leqslant \theta d(m_1,m_2) \,\}} \mathbb{P}_s \left[\psi_{m_1,m_2} = m_1 \right] \leqslant \exp\left(-a |\mathbf{X}_v| d^2(m_1, m_2) \right) \quad . \tag{5.4}$$

Considering two densities $\widehat{s}_i[\mathbf{X}_t]$ and $\widehat{s}_j[\mathbf{X}_t]$, the test is defined by

$$\psi_{i,j} = \begin{cases} i \quad \text{if} \quad T_{i,j} \leqslant 0 \\ \\ j \quad \text{otherwise} \ . \end{cases} \tag{5.5}$$

In the density framework, Assumption A is fulfilled with $d = h$ using one of the following test statistic $T_{i,j}$:

- setting $\omega = \arccos(1 - h^2(\widehat{s}_i[\mathbf{X}_t], \widehat{s}_j[\mathbf{X}_t]))$, Birgé (2013, Section 4) introduced

$$T_{i,j} = \sum_{X_k \in \mathbf{X}_v} \log \left(\frac{\sin(\theta\omega)\sqrt{\widehat{s}_i[\mathbf{X}_t]} + \sin(\omega(1-\theta))\sqrt{\widehat{s}_j[\mathbf{X}_t]}}{\sin(\theta\omega)\sqrt{\widehat{s}_j[\mathbf{X}_t]} + \sin(\omega(1-\theta))\sqrt{\widehat{s}_i[\mathbf{X}_t]}}(X_k) \right) \quad , \tag{5.6}$$

  and showed that $a = (1 - 2\theta)^2$.

- setting $\widehat{r}_{i,j}[\mathbf{X}_t] = \left(\widehat{s}_i[\mathbf{X}_t] + \widehat{s}_j[\mathbf{X}_t] \right)/2$, Baraud (2011, Section 2) considered

$$T_{i,j} = h^2\left(\widehat{s}_i[\mathbf{X}_t], \widehat{r}_{i,j}[\mathbf{X}_t] \right) - h^2\left(\widehat{s}_j[\mathbf{X}_t], \widehat{r}_{i,j}[\mathbf{X}_t] \right) + \frac{1}{|\mathbf{X}_v|} \sum_{X_k \in \mathbf{X}_v} \frac{\sqrt{\widehat{s}_j[\mathbf{X}_t]} - \sqrt{\widehat{s}_i[\mathbf{X}_t]}}{\sqrt{\widehat{r}_{i,j}[\mathbf{X}_t]}}(X_k) \quad , \tag{5.7}$$

  leading to a different value of a. This unpublished result of Sart (private communication) follows idea of Sart (2011, Section 6) developed in an other framework.

The Hellinger risk of $\widehat{s}_{\widehat{m}_{\mathrm{THO}}}[\mathbf{X}_t]$ is controlled by the following result.

Theorem 5.1. *Let $|\mathcal{M}| = M$, and let $T_{i,j}$ be the test statistic given by* (5.6)*. Then the estimator chosen by* (5.2) *satisfies the following oracle inequality*

$$\mathbb{E}_s \left[h\left(s, \widehat{s}_{\widehat{m}_{\mathrm{THO}}}[\mathbf{X}_t] \right) \right] \leq \left(1 + \theta + \frac{1}{2\log M} \right) \max \left(\frac{1}{\theta} \inf_{m \in \mathcal{M}} \mathbb{E}_s \left[h\left(s, \widehat{s}_m[\mathbf{X}_t] \right) \right], \frac{1}{1-2\theta} \sqrt{\frac{\log M}{|\mathbf{X}_v|}} \right) \quad .$$

The proof of Theorem 5.1 may be found in Section 5.9, it follows idea of Birgé (2006a, Corollary 9) tuned to our context to obtained a more precise bound. Similar bound holds for the $\mathbb{L}_2$-risk (see (Birgé, 2014, Corollary 1)) although under stronger assumptions.

5.3 Efficient algorithms for T-estimation

In this section, we describe our algorithms which are at the core of the *Density.T.HoldOut* package to implement THO. Both algorithms may be useful in a general framework of T-estimation as they allow one to reduce the combinatorial complexity. While our first algorithm computes the true T-estimator, the second implements a lossy approach which reduces the complexity further when the family S is very large, while maintaining good performance in terms of Hellinger risk. In both cases, we assume that *Step one* has already been performed, hence our aim is only to select $\widehat{m}$ among the finite collection S of preliminary estimators using $\mathbf{X}_v$, as described in Section 5.2. Since $\mathcal{M}$ is finite, we assume without loss of generality that $\mathcal{M} = [M]$. Since the estimators $\widehat{s}_m[\mathbf{X}_t]$ are built from a sample independent of $\mathbf{X}_v$, they are, conditionally to $\mathbf{X}_t$, deterministic points in $\mathcal{S}$. From now on we denote them s_m - or m when no confusion is possible - and the THO criterion $\mathrm{crit}_{\mathrm{THO}}(m, \mathbf{X}_t, \mathbf{X}_v)$ is denoted $\mathcal{D}(m) = \max_{i \in \mathcal{R}_m} d(i, m)$, where we recall that $\mathcal{R}_m$ consists of the $j \in [M] \setminus \{m\}$ which are chosen against m by the robust tests. Finally let us denote $\bar{\mathcal{B}}(m, r) = \{\, l \in [M] : d(m, l) \leqslant r \,\}$ the intersection of $\mathcal{M}$ with the closed ball with center m and radius $r > 0$. From a purely combinatorial point of view, the computation of $\widehat{m}$ minimizing the plausibility index $\mathcal{D}(m)$ requires the computation of $\mathrm{O}(M^2)$ tests with a "naive" algorithm, which is prohibitive as compared to the $\mathrm{O}(M)$ operations needed to compute the classical HO estimator.

5.3.1 Exact T-Hold-Out

The T-estimator search can be realized with a non-quadratic number of tests, thanks to a simple argument which is summarized by the following lemma and its corollary.

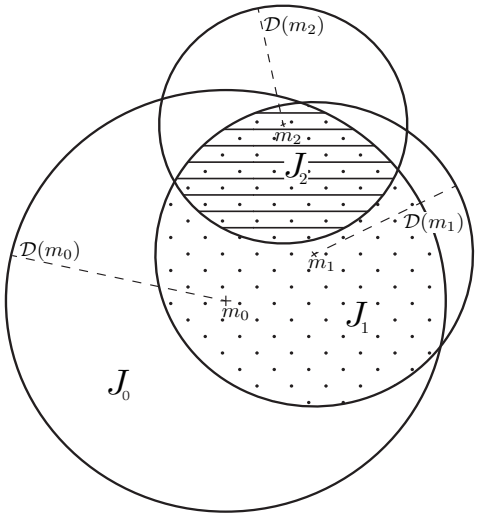

Figure 5.1: Illustration of our exact search for T-estimation. Along the three first iterations, the estimators m_i, $i = 0, 1, 2$ are considered with associated radii $\mathcal{D}(m_i)$ and the T-estimator belongs successively to J_i where J_0 is $\bar{\mathcal{B}}(m_0, \mathcal{D}(m_0))$, J_1 is the dotted and J_2 the hatched area.

Lemma 5.1. *For any point $m_0 \in [M]$, the T-estimator $\widehat{m}$ belongs to $\bar{\mathcal{B}}(m_0, \mathcal{D}(m_0))$.*

Proof: Suppose that there exists one point $m_0 \in [M]$ such that $\widehat{m}$ does not belong to the closed ball of radius $\mathcal{D}(m_0)$ centered at m_0. Then it does not belong to $\mathcal{R}_{m_0}$, and it follows that $\psi_{m_0,\widehat{m}} = m_0$. Hence m_0 belongs to $\mathcal{R}_{\widehat{m}}$ leading to $\mathcal{D}(\widehat{m}) \geqslant d(\widehat{m}, m_0) > \mathcal{D}(m_0)$ which provides a contradiction with $\mathcal{D}(\widehat{m}) = \min_{m\in[M]} \mathcal{D}(m)$. □

Corollary 5.1. *For any subset $J \subset [M]$, the T-estimator $\widehat{m}$ belongs to*

$$\bigcap_{m\in J} \bar{\mathcal{B}}(m, \mathcal{D}(m)) \ .$$

Proof: The proof, illustrated by 5.1, is straightforward using similar arguments as in Lemma 5.1. □

It follows that, starting from m_0, only a point inside $\bar{\mathcal{B}}(m_0, \mathcal{D}(m_0))$ may be the T-estimator. If any point m_1 in this first ball satisfies $\mathcal{D}(m_1) < \mathcal{D}(m_0)$, by Corollary 5.1, the T-estimator will belong to $\bar{\mathcal{B}}(m_0, \mathcal{D}(m_0)) \bigcap \bar{\mathcal{B}}(m_1, \mathcal{D}(m_1))$. Again, criterion $\mathcal{D}$ needs to be computed only for points inside this intersection. We keep intersecting balls $\bar{\mathcal{B}}(m, \mathcal{D}(m))$ until there are no more points with a value of $\mathcal{D}$ smaller than its running value. This approach provides an exact computation of the T-estimator.

At each step of the recursion, the current best point is denoted m with associated value $\mathcal{D}(m)$ denoted by $\mathcal{D}$. The running intersection which contains the potentially better points than m is denoted J (this set does not contain m). The recursion stops when J is empty. At a given step of the recursion, a point j in J is better than m - and thus replaces it - if $\mathcal{D}(j) < \mathcal{D}$. In all cases, j is removed from the set J. During the iteration, $|J|$ and $\mathcal{D}$ decrease ensuring that the algorithm

stops. The last running m is the T-estimator. The pseudo-code implementing the efficient and exact search of the T-estimator is provided by Algorithm 1.

Algorithm 1: Efficient and exact T-Hold-Out

Input: $m \in J = [M]$
1 **for** $(j \neq m)$ **do** compute $\psi_{m,j}(\mathbf{X}_v)$
2 Compute $\mathcal{D} = \mathcal{D}(m)$ and set $J = \bar{\mathcal{B}}(m, \mathcal{D}) \setminus \{m\}$
3 **while** $(|J| > 0)$ **do**
4 Set $\mathcal{D}_{tmp} = 0$, select $j \in J$ and set $J = J \setminus \{j\}$
5 **for** $(k \neq j)$ **do**
6 Compute $\psi_{k,j}(\mathbf{X}_v)$ `// if it has not been done yet`
7 **if** $(\psi_{k,j}(\mathbf{X}_v) == k)$ **then** `//` $k \in \mathcal{R}_j$
8 Set $\mathcal{D}_{tmp} = \max(\mathcal{D}_{tmp}, d(j,k))$
9 **if** $(\mathcal{D}_{tmp} > \mathcal{D})$ **then** break `// break the for loop`
10 Set $m = j$, $\mathcal{D} = \mathcal{D}_{tmp}$ and $J = J \bigcap \bar{\mathcal{B}}(m, \mathcal{D})$
Return: m `// the T-estimator`

Comments: This algorithm works for all the statistical frameworks of T-estimation, and does not depend on the considered robust test. The "for" loop is realized on all $k \neq j$, as $\mathcal{D}(k)$ depends on all points and not only on those in J. If there are N points in the first ball, the number of computed tests is at most $\mathrm{O}(N * M)$. Moreover, if the first ball is empty, i.e. if $\mathcal{D}(m) = 0$, the algorithm stops immediately, returning m for $\widehat{m}$. In this case, the complexity of our algorithm is $\mathrm{O}(M)$. Any preliminary estimator (maximum likelihood, least-squares, $\mathbb{L}_1$-minimizer, etc.) may be a starting point of our algorithm. We hope that by beginning from a good preliminary estimator, there will be only few points in the first ball, resulting in less computations. The computation requires $\mathrm{O}(M^2)$ operations if J decreases by only one point at each step of the recursion which happens only if the selected j satisfies

$$\max_{k \in J} d(j,k) = \max_{k \neq l \in J} d(k,l)$$

at each iteration.

5.3.2 Fast algorithm for approximate T-Hold-Out

Assumption A ensures that as soon as the Hellinger distance between two estimators of S is large enough, the probability that the robust test does not choose the best estimator is small. However, as shown in Lemma 1 of Le Cam (1973), when this distance is smaller than $c\, n^{-1/2}$, where c is a small positive constant, the two corresponding probabilities cannot be separated by a test built on n observations anymore. From this remark, we derive a lossy version from our efficient and exact algorithm. The main difference consists in ignoring points in S as soon as their Hellinger distance to a previously considered one is smaller than a given threshold $\delta_n > 0$.

We introduce this distance control at two steps of our efficient and exact algorithm. As the interior points of $\bar{\mathcal{B}}(m, \delta_n)$ cannot be properly distinguished from m by any test, the set J becomes, at lines 2 and 10 of Algorithm 1, the intersection of rings instead of balls, obtained by removing from the original ball $\bar{\mathcal{B}}(m, \mathcal{D}(m))$ the ball $\bar{\mathcal{B}}(m, \delta_n)$. In the same spirit, at line 5 of Algorithm 1, the current k, in the *for* loop, is considered if and only if its distance to $\mathcal{T}_j$ is larger than δ_n,

where $\mathcal{T}_j$ is made of the running j and the further points which have been tested against j. The pseudo-code of this lossy version is provided by Algorithm 2 and illustrated by 5.2.

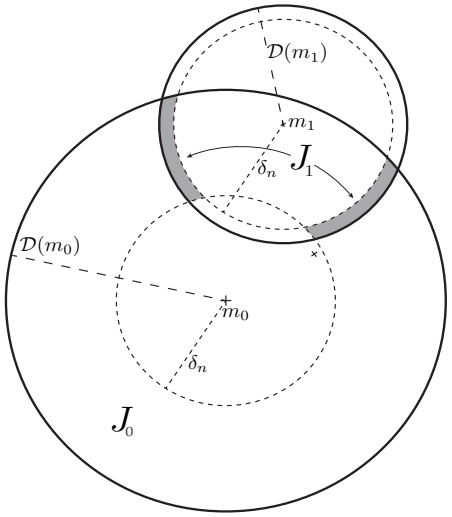

Figure 5.2: Illustration of the approximate T-estimation search: J_0 is a ring around m_0. The point following m_0 has changed with respect to 5.1 as the previously selected m_1 is now inside $\bar{\mathcal{B}}(m_0, \delta_n)$. J_1 (in grey) appears as the intersection of two rings.

Algorithm 2: Approximate T-Hold-Out

Input: $m \in J = [M]$; $\delta_n > 0$
for $(j \neq m)$ **do** compute $\psi_{m,j}(\mathbf{X}_v)$
Compute $\mathcal{D} = \mathcal{D}(m)$ and set $J = \bar{\mathcal{B}}(m, \mathcal{D}) \setminus \bar{\mathcal{B}}(m, \delta_n)$
while $(|J| > 0)$ **do**
  Set $\mathcal{D}_{tmp} = 0$, select $j \in J$ and set $J = J \setminus \{j\}$
  Define $\mathcal{T}_j = \{j\}$
  for $(k \neq j)$ **do**
    if $(d(j, \mathcal{T}_j) \leqslant \delta_n)$ **then** next k `// next k if distance is too small`
    Set $\mathcal{T}_j = \mathcal{T}_j \cup \{k\}$
    Compute $\psi_{k,j}(\mathbf{X}_v)$ `// if it has not been done yet`
    if $(\psi_{k,j}(\mathbf{X}_v) == k)$ **then** `// k ∈ R_j`
      Set $\mathcal{D}_{tmp} = \max(\mathcal{D}_{tmp}, d(j,k))$
      if $(\mathcal{D}_{tmp} > \mathcal{D})$ **then** break `// break the for loop`
  Set $m = j$, $\mathcal{D} = \mathcal{D}_{tmp}$ and $J = J \bigcap \left[\bar{\mathcal{B}}(m, \mathcal{D}) \setminus \bar{\mathcal{B}}(m, \delta_n)\right]$
Return: m `// the approximate T-estimator`

5.4 Simulation protocol

In our simulations, we consider only the density estimation framework. This is motivated by the fact that likelihood ratio tests are not robust in this context, and we hoped to observe differences in terms of risk.

We considered $\mathbf{X} = \{X_1, \ldots, X_n\}$ i.i.d. random variables from an unknown density s with respect to the Lebesgue measure on $\Xi = \mathbb{R}$ and, for a given proportion p in $(0, 1)$, we divide randomly $\mathbf{X}$ into $\mathbf{X}_t = \{X_1, \ldots, X_{n_1}\}$ and $\mathbf{X}_v = \{X_{n_1+1}, \ldots, X_n\}$, with $n_1 = \lfloor pn \rfloor$ where $\lfloor x \rfloor$ is the integer part of x. Simulations were carried out with four sample sizes ($n = 100, 250, 500, 1000$) and three different proportions ($p = 1/2, 2/3, 3/4$) using the two different robust tests (5.6) and (5.7). Our test functions s vary in a subset $\mathcal{L}$ made of the densities s_1,..., s_{28} of the R-package *benchden*[6] which are in $\mathbb{L}_1 \cap \mathbb{L}_2$ - to ensure that risks are computable. This set $\mathcal{L}$ is made of the densities s_i for

$$i \in \{1, \ldots, 5, 7, 11, 12, 13, 16, 17, 21, \ldots, 27\}.$$

We considered several estimator collections:

- S_R made of regular histograms with bin number varying from 1 to $\lceil n_1/\log(n_1) \rceil$ as described in Birgé & Rozenholc (2006);
- S_I made of the maximum likelihood irregular histograms when the bin number only varies from 1 to $\min(100, \lceil n_1/\log(n_1) \rceil$ as described in Rozenholc *et al.* (2010);
- S_K made of Gaussian kernel estimators with the varying bandwidths chosen as

$$(\max[\mathbf{X}_t] - \min[\mathbf{X}_t])/2j \quad \text{for} \quad j = 1, \ldots, \lceil n_1/\log(n_1) \rceil \ .$$

- S_P made of parametric estimates obtained by moment's method for the Gaussian, exponential, log-normal, chi-square, gamma and beta distributions together with a maximum likelihood estimate of the uniform distribution;
- $S_C = S_R \cup S_I$
- $S_1 = S_R \cup S_I \cup S_K$;
- $S_2 = S_R \cup S_I \cup S_K \cup S_P$.

The estimation accuracy of a given procedure $\widetilde{s}$ has been evaluated using an empirical version of the risk $R_s(\widetilde{s}, \ell) = \mathbb{E}_s[\ell(s, \widetilde{s})]$, obtained by generating 100 n-samples $\mathbf{X}^{(j)}$, $1 \leqslant j \leqslant 100$, of density s:

$$\bar{R}_s\left(\widetilde{s}, \ell\right) = \frac{1}{100} \sum_{j=1}^{100} \ell\left(s, \widetilde{s}[\mathbf{X}^{(j)}]\right) \ ,$$

where $\ell(t, u)$ is either $h^2(t, u)$ or $\|t - u\|_q^q$, for $q = 1, 2$.

[6] *Benchden* (see Mildenberger & Weinert (2012)) implements the benchmark distributions of Berlinet & Devroye (1994). Available on the CRAN `http://cran.r-project.org/web/packages/benchden/index.html`.

In order to compare two procedures $\widetilde{t}_1$ and $\widetilde{t}_2$, we introduce the normalized $\log_2$-ratio of their empirical risks, namely:

$$\overline{W_s}\left(\widetilde{t}_1, \widetilde{t}_2\right) = \frac{1}{r} \log_2 \frac{\bar{R}_s\left(\widetilde{t}_1, \ell\right)}{\bar{R}_s\left(\widetilde{t}_2, \ell\right)} = \log_2 \bar{R}_s^{1/r}\left(\widetilde{t}_1, \ell\right) - \log_2 \bar{R}_s^{1/r}\left(\widetilde{t}_2, \ell\right) \quad ,$$

where r is equal to q for $\mathbb{L}_q$ losses and 2 for the Hellinger loss. The aim of the normalization by r is to provide an easier comparison of $\overline{W_s}$ when the loss changes. In our empirical study, procedure $\widetilde{t}_2$ is thus considered better in terms of risk than $\widetilde{t}_1$ for a given loss function ℓ if the values of $\overline{W_s}(\widetilde{t}_1, \widetilde{t}_2)$ are positive when the density s varies.

We compared the four hold-out methods described above: T-estimation with the tests given by (5.6) and (5.7), LS and KL. We first computed $\widehat{s}_m[\mathbf{X}_t]$ for all $m \in \mathcal{M}$, and then selected $\widehat{m}$ minimizing the respective HO criterion resulting in $\widehat{m}_{T1}$, $\widehat{m}_{T2}$, $\widehat{m}_{LS}$ and $\widehat{m}_{KL}$, providing $\widetilde{s}$ as either $\widehat{s}_{\widehat{m}}[\mathbf{X}_t]$ or $\widehat{s}_{\widehat{m}}[\mathbf{X}]$. As $\widehat{m}$ depends on the chosen proportion p, in order to explicitly specify the dependency of $\widehat{m}$ with respect to this parameter, we will use the following notations $\widehat{s}_{\widehat{m}[p]}[\mathbf{X}_t]$ or $\widehat{s}_{\widehat{m}[p]}[\mathbf{X}]$ when needed. In Algorithms 1 and 2, the input m has been set to $\widehat{m}_{LS}$ and $j = \operatorname{argmax}_{k \in J} d(k, m)$, at line 4. In Algorithm 2, we fixed $\delta_n = 1/\sqrt{|\mathbf{X}_v|}$ as a lower bound for the Hellinger distance between distinguishable probabilities, following Le Cam (1973).

Moreover, we also considered some calibrated estimation procedures which choose m in some particular families. These are not direct competitors with the T-estimation as they cannot deal with general families S but provide a good benchmark in terms of risk:

- for S_R, S_I, S_C, the penalized maximum likelihood estimators, denoted $\widetilde{s}_{\text{pen}}$ introduced in Birgé & Rozenholc (2006); Rozenholc *et al.* (2010) and implemented in the R-package[(7)] *histogram*,
- for S_K, the L_1-version of the procedure introduced in Goldenshluger & Lepski (2011), denoted $\widetilde{s}_{GL}$.

For fairness, we applied these calibrated estimation procedures in their original setting which use the full sample replacing n_1 by n in the definition of S_R and S_K.

Finally, for the family S_K, we considered some bandwidth selectors (namely nrd, ucv, bcv, SJ) implemented in the *density* generic function available in R , providing some well-known estimators $\widetilde{s}_{nrd}$, $\widetilde{s}_{bcv}$, $\widetilde{s}_{ucv}$, $\widetilde{s}_{SJ}$ of the density which are not chosen in S (Silverman, 1986; Sheather & Jones, 1991; Scott, 1992).

The R-package[(8)] *Density.T.HoldOut* is a ready-to-use software that implements our algorithms in the density framework. The main function - called `DensityTestim` - receives as input a sample $\mathbf{X}$ and a family of estimators and returns the selected estimator. The previously described families are available and can be extended or adapted by the user (default family is S_2). Other important

[(7)] available on the CRAN `http://cran.r-project.org/web/packages/histogram/index.html`.

[(8)] available on the CRAN `http://cran.r-project.org/web/packages/Density.T.HoldOut/index.html`

input arguments are parameters p, θ and the starting point (default values are $p = 1/2$, $\theta = 1/4$ and $\widehat{m}_{LS}$). This function implements the exact and lossy algorithms, through the numeric `csqrt` (default value 1) which controls $\delta_n = \texttt{csqrt}/\sqrt{|\mathbf{X}_v|}$ in Algorithm 2. The robust test might be the one defined by (5.6) setting `test`='birge' (default), or by (5.7) setting `test`='baraud'. The resulting estimator is either built with $\mathbf{X}_t$ (`last`='training') or $\mathbf{X}$ (`last`='full', default).

5.5 Simulation results

This section, made using Algorithm 1, is devoted to the study of the quality of the T-hold-out. We illustrate our results showing boxplots of $\overline{W_s}(\widetilde{t}_1, \widetilde{t}_2)$ for all 18 densities $s \in \mathcal{L}$, various choices of estimators $\widetilde{t}_1$ and $\widetilde{t}_2$ and for different collections of estimators S, as described in the previous section. We begin by investigating how parameter θ influences the THO procedure deduced from (5.6). Then we show that the two robust procedures derived from (5.6) and (5.7) have similar behavior in terms of risk, and therefore pursue using the first one only. After studying how p influences the quality of estimation, we provide two main comparison types. First we look at HO methods which select among a family of points using the validation sample. Then we compare the THO against some density estimation methods, which are not necessarily selection procedures anymore. In this subsection, we divide the presentation between calibrated selection procedures build directly on the full sample and some selectors of the bandwidth obtained using asymptotic derivation of the risk for some specific loss.

5.5.1 Influence of θ

The robustness of the procedure build using (5.6) is controlled through the parameter $\theta < 1/2$ (see Eq. 5.6), the KLHO corresponding to $\theta = 0$ (no robustness). We computed the empirical risk using the THO procedure with $\theta = 1/16, 1/8, 1/4, 3/8, 7/16$, and $n = 100, 250, 500, 1000$. We observed that θ has little influence in terms of risk ($\theta = 1/16$ being slightly worse) and decided to pursue the empirical study with $\theta = 1/4$.

5.5.2 Influence of the robust test

As we dispose of two robust tests to proceed the THO, we compare the two corresponding strategies in 5.3 using $\widetilde{t}_1 = \widehat{s}_{\widehat{m}_{T1}[p]}[\mathbf{X}_t]$ and $\widetilde{t}_2 = \widehat{s}_{\widehat{m}_{T2}[p]}[\mathbf{X}_t]$ for $p = 1/2$, $2/3$ and $3/4$ (each value corresponding to one subfigure below). For a fixed n, there are 18×6 ratios obtained when both the density and the collection of estimators vary.

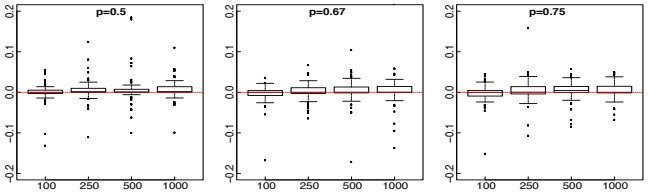

Figure 5.3: From left to right, normalized $\log_2$-ratio of the empirical risks $\overline{W_s}(\widehat{s}_{\widehat{m}_{T1}[p]}[\mathbf{X}_t], \widehat{s}_{\widehat{m}_{T2}[p]}[\mathbf{X}_t])$ for the Hellinger loss for $p = 1/2$, $2/3$ and $3/4$. Each subfigure shows the boxplot for n equals 100, 250, 500 and 1000. The horizontal red dotted line provides the reference value 0.

Surprisingly the two procedures behave very similarly in all settings, and only few differences can be observed in terms of Hellinger risk (generally less than 2%). We therefore pursue our empirical study with the procedure derived from (5.6), and from now on we denote $\widehat{m}_T$ instead of $\widehat{m}_{T1}$, when no confusion is possible.

5.5.3 Influence of p

We examine the dependence of the THO with respect to p, the proportion of the initial sample dedicated to building the estimators, using the Hellinger risk.

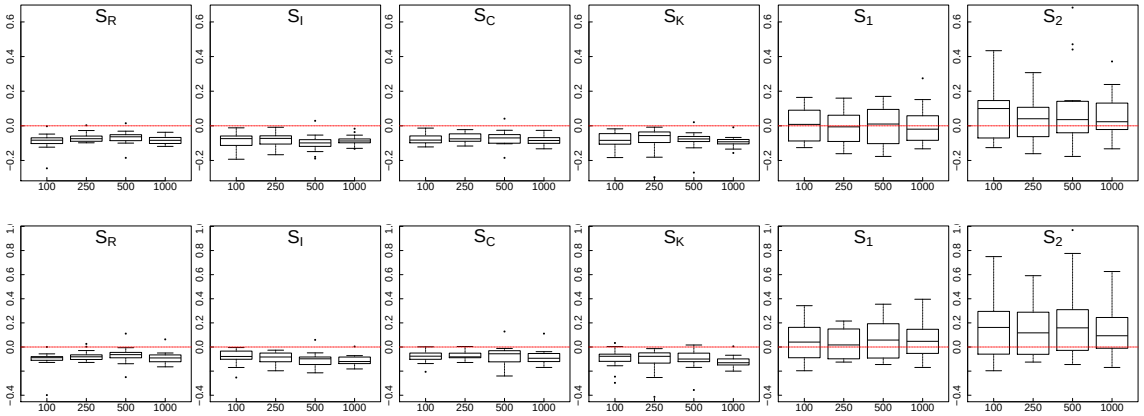

Figure 5.4: From left to right, normalized $\log_2$-ratio of the empirical risks $\overline{W_s}(\widehat{s}_{\widehat{m}_T[2/3]}[\mathbf{X}_t], \widehat{s}_{\widehat{m}_T[1/2]}[\mathbf{X}_t])$ (upper line) and $\overline{W_s}(\widehat{s}_{\widehat{m}_T[3/4]}[\mathbf{X}_t], \widehat{s}_{\widehat{m}_T[1/2]}[\mathbf{X}_t])$ (bottom line) for the Hellinger loss, using collections S_R, S_I, S_C, S_K, S_1 and S_2. Each subfigure shows the boxplot for n equals 100, 250, 500 and 1000. The horizontal red dotted line provides the reference value 0.

5.4 is built using $\widetilde{t}_1 = \widehat{s}_{\widehat{m}_T[p]}[\mathbf{X}_t]$ for p equals 2/3 (upper line), 3/4 (bottom line) and $\widetilde{t}_2 = \widehat{s}_{\widehat{m}_T[1/2]}[\mathbf{X}_t]$. We observe two different behaviors for families S_R, S_I, S_C and S_K on the one hand and for S_1 and S_2 on the other hand. For the first families $p = 2/3$ or $3/4$ is better than $p = 1/2$. For the second ones $p = 2/3$ seems equivalent to $p = 1/2$ but $p = 3/4$ is worst than $p = 1/2$. Hence we consider preferable to use $p = 2/3$, which makes the best compromise for all families.

5.5.4 Comparing Hold-Out methods

Hold-out procedures are universal since they do not depend on the choice of family S. They can be seen as methods that choose among some family of fixed points. Setting $p = 2/3$, we compare the THO to the KLHO and LSHO introduced in Section 5.1.2 using each of the 6 estimator collections described in Section 5.4.

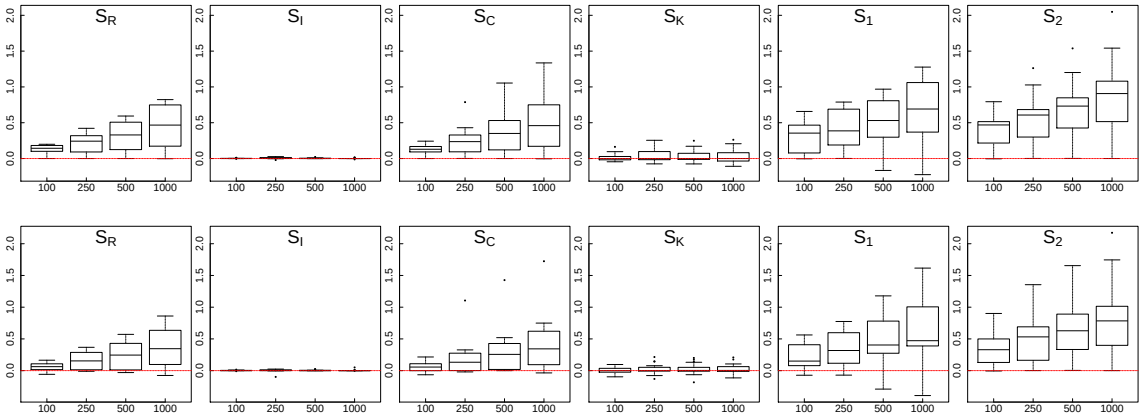

Figure 5.5: From left to right, normalized $\log_2$-ratio of the empirical risks $\overline{W_s}(\widehat{s}_{\widehat{m}_{KL}}[\mathbf{X}_t], \widehat{s}_{\widehat{m}_T}[\mathbf{X}_t])$ for $p = 2/3$, using collections S_R, S_I, S_C, S_K, S_1 and S_2. Upper line, using Hellinger loss, bottom line using $\mathbb{L}_1$ loss. See 5.4 for more details.

5.5 is built using $\widetilde{t}_1 = \widehat{s}_{\widehat{m}_{KL}}[\mathbf{X}_t]$ and $\widetilde{t}_2 = \widehat{s}_{\widehat{m}_T}[\mathbf{X}_t]$ considering Hellinger (upper line) and $\mathbb{L}_1$ (bottom line) losses. In all cases, the median and most of the distribution are positive, meaning that the THO outperforms the KLHO estimator. For collections S_I and S_K, empirical risks for both losses are similar, with $\overline{W_s}(\widehat{s}_{\widehat{m}_{KL}}[\mathbf{X}_t], \widehat{s}_{\widehat{m}_T}[\mathbf{X}_t])$ being respectively larger than -0.01 (except for the uniform density) for S_I, and -0.2 for S_K. When n grows, while for S_I and S_K the ratio remains stable, it increases for all other families in favor of the THO. Moreover when going from collection S_1 to S_2, that is adding the parametric collection S_P, we observe that the already good performance of the THO improves. We therefore suspect that the THO chooses the parametric estimator more often than KLHO when facing the corresponding densities.

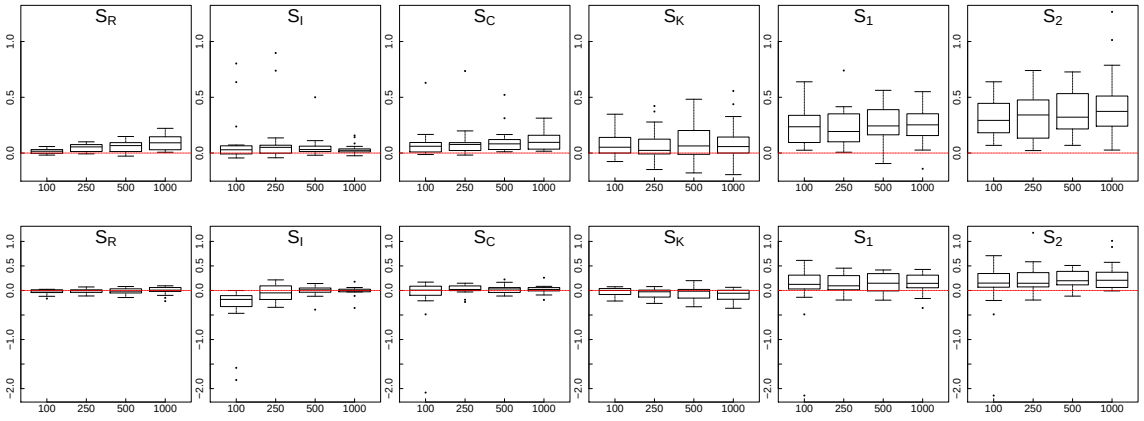

Figure 5.6: From left to right, normalized $\log_2$-ratio of the empirical risks $\overline{W_s}(\widehat{s}_{\widehat{m}_{LS}}[\mathbf{X}_t], \widehat{s}_{\widehat{m}_T}[\mathbf{X}_t])$ for $p = 2/3$, using collections S_R, S_I, S_C, S_K, S_1 and S_2. Upper line, using Hellinger loss, bottom line using $\mathbb{L}_2$ loss. See 5.4 for more details.

5.6 is built using $\widetilde{t}_1 = \widehat{s}_{\widehat{m}_{LS}}[\mathbf{X}_t]$ and $\widetilde{t}_2 = \widehat{s}_{\widehat{m}_T}[\mathbf{X}_t]$ considering Hellinger (upper line) and $\mathbb{L}_2$ (bottom line) losses. The THO performs better than the LSHO estimator for all collections except for the collection S_I when $n = 100$. For the larger collections S_1 and S_2, the THO outperforms the LSHO. However, as n grows, we observe that the relative quality of the two procedures remain stable.

5.5.5 Comparing final strategies for T-Hold-Out

Here, we investigate whether $\widehat{s}_{\widehat{m}_T}[\mathbf{X}_t]$ or $\widehat{s}_{\widehat{m}_T}[\mathbf{X}]$ performs better. For this purpose, we study the Hellinger risk of $\widehat{s}_{\widehat{m}_T}[\mathbf{X}]$ when p varies. 5.7 is built using $\widetilde{t}_1 = \widehat{s}_{\widehat{m}_T[p]}[\mathbf{X}]$ for p equals 2/3 (upper line), 3/4 (bottom line) and $\widetilde{t}_2 = \widehat{s}_{\widehat{m}_T[1/2]}[\mathbf{X}]$. We observe that against $p = 2/3$ or $p = 3/4$, the value $p = 1/2$ provides better results for the large families S_1 and S_2 while for the small families the results are more balanced. Hence we consider preferable to make use of this strategy with $p = 1/2$.

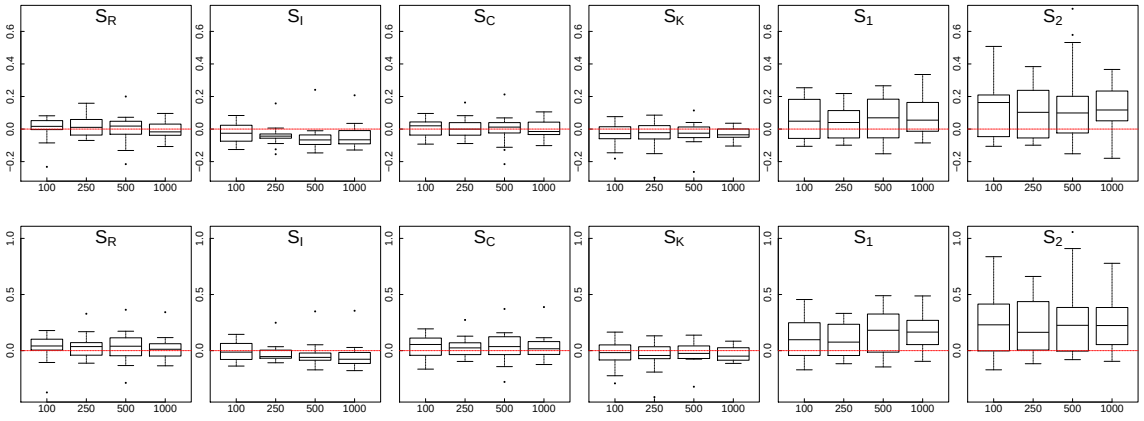

Figure 5.7: From left to right, normalized $\log_2$-ratio of the empirical risks $\overline{W_s}(\widehat{s}_{\widehat{m}_T[2/3]}[\mathbf{X}], \widehat{s}_{\widehat{m}_T[1/2]}[\mathbf{X}])$ (upper line) and $\overline{W_s}(\widehat{s}_{\widehat{m}_T[3/4]}[\mathbf{X}], \widehat{s}_{\widehat{m}_T[1/2]}[\mathbf{X}])$ (bottom line) for the Hellinger loss, using collections S_R, S_I, S_C, S_K, S_1 and S_2. See 5.4 for more details.

We now compare the Hellinger risks of $\widehat{s}_{\widehat{m}_T[2/3]}[\mathbf{X}_t]$ - which appeared as the best competitor in Section 5.5.3 - and $\widehat{s}_{\widehat{m}_T[1/2]}[\mathbf{X}]$. 5.8 is built using $\widetilde{t}_1 = \widehat{s}_{\widehat{m}_T[1/2]}[\mathbf{X}]$ and $\widetilde{t}_2 = \widehat{s}_{\widehat{m}_T[2/3]}[\mathbf{X}_t]$. We observe that the strategy $\widehat{s}_{\widehat{m}_T[1/2]}[\mathbf{X}]$ is preferable, since its median (and even most of its distribution) is negative in all considered settings. It should be noticed that our simulations show that, more than the value of p, it is the use of $\mathbf{X}$ instead of $\mathbf{X}_t$ which has the larger influence on the final risk.

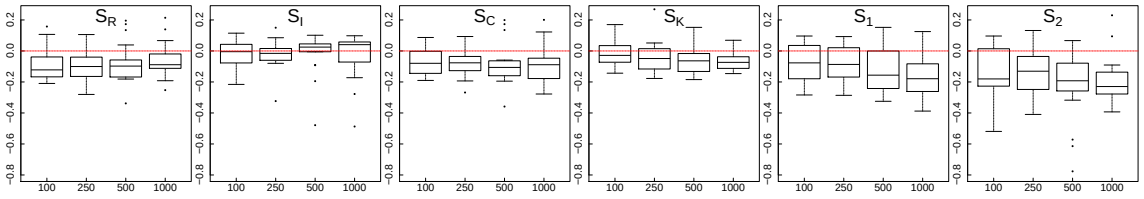

Figure 5.8: From left to right, normalized $\log_2$-ratio of the empirical risks $\overline{W_s}(\widehat{s}_{\widehat{m}_T[1/2]}[\mathbf{X}], \widehat{s}_{\widehat{m}_T[2/3]}[\mathbf{X}_t])$ for Hellinger loss, using collections S_R, S_I, S_C, S_K, S_1 and S_2. See 5.4 for more details.

5.5.6 T-Hold-Out against dedicated estimation procedures

We now compare the THO competitor $\widehat{s}_{\widehat{m}_T[1/2]}[\mathbf{X}]$ against the so-called dedicated methods. 5.9 is built using $\widetilde{t}_1 = \widetilde{s}[\mathbf{X}]$ ($\widetilde{s}$ being either $\widetilde{s}_{\text{pen}}$ or $\widetilde{s}_{GL}$) and $\widetilde{t}_2 = \widehat{s}_{\widehat{m}_T[1/2]}[\mathbf{X}]$ considering Hellinger (upper line) and $\mathbb{L}_1$ (bottom line) losses. We observe that the THO is slightly worse than a well-calibrated procedure for histograms but outperforms the $\mathbb{L}_1$-version of the Goldenshluger-Lepski procedure.

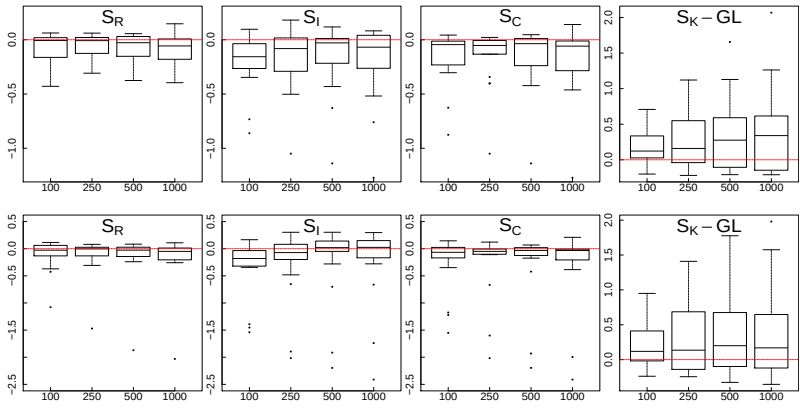

Figure 5.9: From left to right, normalized $\log_2$-ratio of the empirical risks $\overline{W_s}(\widetilde{s}[\mathbf{X}], \widehat{s}_{\widehat{m}_T[1/2]}[\mathbf{X}])$ using collections S_R, S_I, S_C and S_K with Hellinger (upper line) and $\mathbb{L}_1$ (bottom line) losses. For the 3 first collections $\widetilde{s}$ is $\widetilde{s}_{\text{pen}}$ and $\widetilde{s}_{GL}$ for S_K. Each subfigure shows the boxplot for n equals 100, 250, 500 and 1000. The horizontal red dotted line provides the reference value 0.

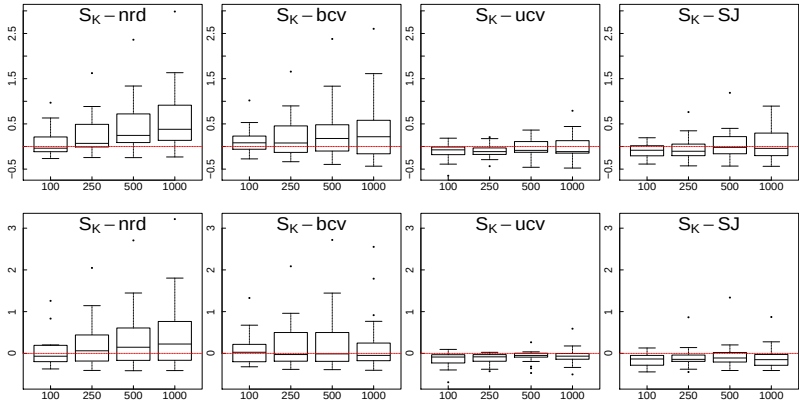

Figure 5.10: From left to right, normalized $\log_2$-ratio of the empirical risks $\overline{W_s}(\widetilde{s}[\mathbf{X}], \widehat{s}_{\widehat{m}_T[1/2]}[\mathbf{X}])$ for collection S_K. The 3 first competitors $\widetilde{s}$ are the kernel estimators with respective bandwidth provided by the bandwidth selectors nrd, bcv, ucv and SJ as defined in the function *density* of the *stats* package of *R*. Upper line, using Hellinger loss, bottom line using $\mathbb{L}_1$ loss. See 5.4 for more details.

For the sake of completeness, we also provide in 5.10 the comparison between the THO and well-known estimators of the density derived from bandwidth selectors available in the *density* generic function of R. We observe that $\widetilde{s}_{ucv}$ and $\widetilde{s}_{SJ}$ perform well (particularly for the $\mathbb{L}_1$-loss), whereas the THO outperforms $\widetilde{s}_{nrd}$ and $\widetilde{s}_{bcv}$.

5.6 Empirical complexity of the exact algorithm

To evaluate the complexity of our algorithms let us denote by N the number of tests needed in the computation of the THO for each generated sample of our simulations. As N is between $M-1$ and $M(M-1)/2$, we define the so-called "THO complexity" as the ratio of $N-M+1$ over its

maximal value, that is

$$\frac{2(N-M+1)}{(M-1)(M-2)} \quad . \tag{5.8}$$

For any run, this ratio belongs to $[0,1]$ by construction. For each fixed n, we get a global sample of size 10800 corresponding to "18 densities" times "6 families" times "100 simulations". 5.11 shows the empirical cumulative distribution function (CDF) of the latter sample with the quantiles 0.75, 0.9 and 0.95, for both tests (5.6) and (5.7). We observe from this figure that in both cases the complexity of our algorithm tends to improve with n. Moreover, 75% of the THO complexities are smaller than 0.1 for n equals 250, 500 and 1000 and 95% are smaller than 0.4 for all values of n. The THO complexity using (5.7) is slightly smaller. However the comparison of two estimators in (5.7) requires the computation of one integral to compute the difference of squared Hellinger distances involving the middle point. From a practical point of view, we indeed observed that using the test (5.7) is more CPU time-consuming. Since both strategies have similar THO complexity, we pursue our study again using the procedure derived from (5.6) only.

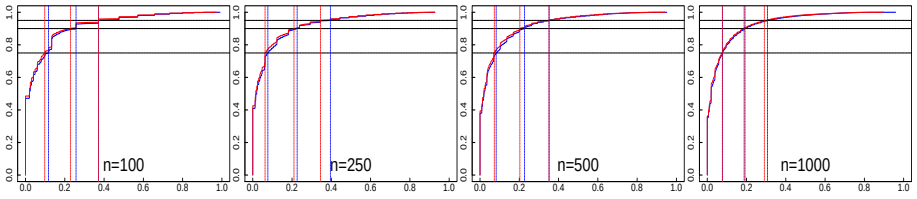

Figure 5.11: From left to right, the CDF for $n = 100$, 250, 500 and 1000 of the THO complexity using Algorithm 1 in plain line: procedure derived from (5.6) in blue and from (5.7) in red. The horizontal black dotted lines provide the values 0.75, 0.9 and 0.95 and the vertical dotted lines their respective quantiles using the respective colors.

In order to complete this study of the complexity we focused on the two collections S_R and S_K for which the number of estimators depends on n as $M = \lceil n_1/\log(n_1) \rceil$. Having in mind that N is not smaller than $M-1$ and not larger than $M(M-1)/2$, we assumed N to be of order $(M-1)^\beta$ with β in $[1,2]$. For each density and each value of n, we compute the average of $\log(N)$ over the 100 runs. In 5.12 these average values are drawn versus $\log(M-1)$ for the two collections and for each density.

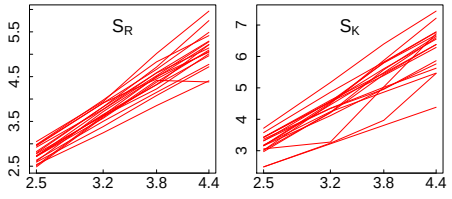

Figure 5.12: Graphs of $\log(N)$ versus $\log(M-1)$ for each density when using the collections S_R (left) and S_K (right).

As 5.12 exhibits mostly linear behaviors, we computed the slope in the linear model of $\log(N)$ versus $\log(M-1)$ as an estimator of β when n_1 varies. We observe that this estimator concentrates

around respectively 1.2 and 1.4 for the collections S_R and S_K providing a good indicator that our algorithm is typically sub-quadratic. The larger value of β for the collection S_K may be explained by the fact that, for our set of bandwidths, the kernel estimators may be very similar, inducing a slow decrease of the running intersection J in Algorithm 1.

5.7 Study of the approximate T-Hold-Out

We provide a comparison of the estimators selected using Algorithms 1 and 2 respectively, that is the exact T-estimator and its approximate version (denoted here by $\widehat{m}_T^g$) computed with $\delta_n = c/\sqrt{|\mathbf{X}_v|}$ for different values of c. We compare these estimators using the two strategies based on $\mathbf{X}_t$ and $\mathbf{X}$.

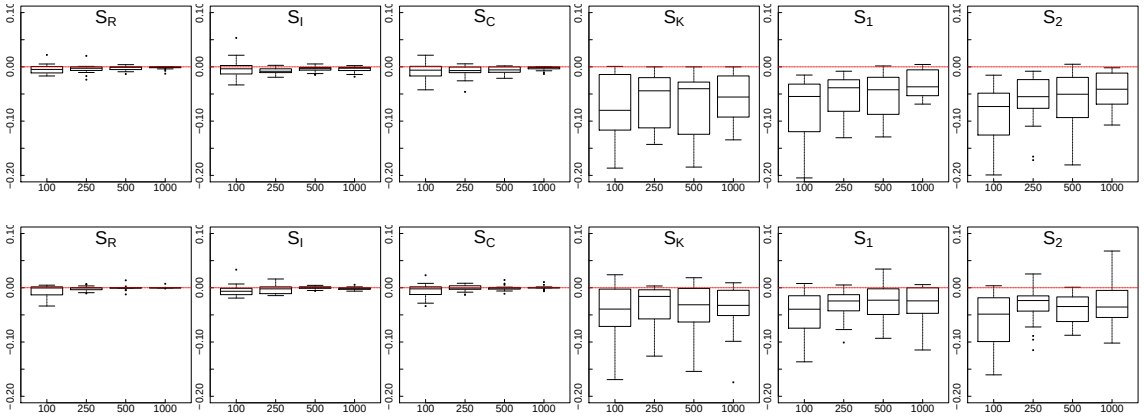

Figure 5.13: From left to right, normalized $\log_2$-ratio of the empirical risks $\overline{W_s}(\widehat{s}_{\widehat{m}_T}[\mathbf{X}_t], \widehat{s}_{\widehat{m}_T^g}[\mathbf{X}_t])$ (upper line) and $\overline{W_s}(\widehat{s}_{\widehat{m}_T}[\mathbf{X}], \widehat{s}_{\widehat{m}_T^g}[\mathbf{X}])$ using $c = 1$ (bottom line) for the Hellinger loss, using collections S_R, S_I, S_C, S_K, S_1 and S_2. See 5.4 for more details.

5.13 is built using $\widetilde{t}_1 = \widehat{s}_{\widehat{m}_T}[\mathbf{X}_t]$ and $\widetilde{t}_2 = \widehat{s}_{\widehat{m}_T^g}[\mathbf{X}_t]$ with $p = 2/3$ on the upper line and using $\widetilde{t}_1 = \widehat{s}_{\widehat{m}_T}[\mathbf{X}]$ and $\widetilde{t}_2 = \widehat{s}_{\widehat{m}_T^g}[\mathbf{X}]$ with $p = 1/2$ on the bottom line. As expected, the exact THO is better in terms of risk. For histogram families, the degradation of the Hellinger risk is negligible. For families S_K, S_1 and S_2, we observe that the risk increases not more than 20% in most of the cases (y-axis reference value equals to -0.13). The empirical cumulative distribution function (CDF) of the complexity ratio defined in (5.8) is shown in 5.14, for both tests, for comparison with 5.11. Clearly the CDFs of the lossy version are more concentrated around 0, showing a significant gain in terms of complexity when using Algorithm 2 (quantiles are divided by more than 2.5).

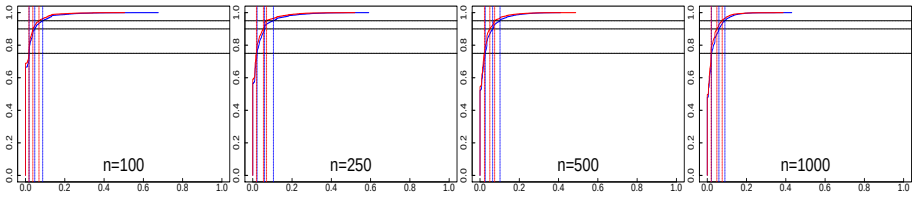

Figure 5.14: From left to right, the CDF for $n = 100$, 250, 500 and 1000 of the THO complexity in plain line using Algorithm 2 with $c = 1$. See 5.11 for more details.

A further study, using $c = 2$ in the approximate algorithm, shows that the risk increases up to 75% in most of the cases and does not offer a good trade-off between complexity and accuracy.

5.8 Conclusion

We introduce an efficient and exact algorithm, together with an approximate version, to implement T-estimation in the context of hold-out. We study the performances of this T-hold-out in the density framework using two different robust tests. Calibration study shows that, when building the final estimate only with the training sample, a good choice of the ratio between training and validation sample sizes is $p = 2/3$. However, risks can be improved using the full sample to build the final estimate when using $p = 1/2$. Our procedure is competitive compared to classical hold-out derived from Kullback-Leibler or least-squares contrasts. It still behaves well against model selection procedures derived from a calibrated penalized contrast for histogram selection, and against most of the bandwidth selectors for kernel estimators. Empirically, we observe that this algorithm improves clearly the combinatorial complexity. Moreover, it can be speeded up thanks to our proposed lossy version, which offers the expected trade-off between complexity and estimation quality. Finally, the two THO strategies are very similar in terms of Hellinger risk and THO complexity, but we recommend to proceed the THO procedure based on (5.6) since it is less time-consuming. This work is very encouraging in view of developing more complex resampling procedures based on robust tests. We believe that more complex data splitting scheme (such as V-fold cross-validation or leave-p-out) will provide even better results in terms of risk. However, the construction of fast algorithms together with theoretical guarantees will be unavoidable in order to provide a complete alternative to risk estimation procedures.

5.9 Proof of Theorem 5.1

Let us first set $\bar{m} = \text{argmin}_{m\in\mathcal{M}} h(s, \widehat{s}_m[\mathbf{X}_t])$, $\bar{h} = h(s, \widehat{s}_{\bar{m}}[\mathbf{X}_t])$ and $\mathcal{D}(m) = \text{crit}_{\text{THO}}(m, \mathbf{X}_t, \mathbf{X}_v)$ for all $m \in \mathcal{M}$. We have by definition of (5.1), for all $m, m' \in \mathcal{M}$,

$$h\left(\widehat{s}_{m'}[\mathbf{X}_t], \widehat{s}_m[\mathbf{X}_t]\right) \leqslant \max\left(\mathcal{D}(m), \mathcal{D}(m')\right) \quad .$$

Therefore, since $\widehat{m}_{\text{THO}}$ minimizes $\mathcal{D}$, and by the triangular inequality,

$$h\left(\widehat{s}_{\widehat{m}_{\text{THO}}}[\mathbf{X}_t], s\right) \leq h\left(\widehat{s}_{\bar{m}}[\mathbf{X}_t], s\right) + \mathcal{D}(\bar{m}) \quad . \tag{5.9}$$

Let us now focus on the random variable $\mathcal{D}(\bar{m})$. For $y \geq \theta^{-1} h(\widehat{s}_{\bar{m}}[\mathbf{X}_t], s)$

$$\begin{aligned}
\mathbb{P}_s\left[\mathcal{D}(\bar{m}) \geqslant y \,|\, \mathbf{X}_t\right] &= \mathbb{P}_s\left[\exists\, m \in \mathcal{M} \text{ s.t. } h(\widehat{s}_m[\mathbf{X}_t], \widehat{s}_{\bar{m}}[\mathbf{X}_t]) \geqslant y \text{ and } \psi_{m,\bar{m}}(\mathbf{X}_v) = m \,|\, \mathbf{X}_t\right] \\
&\leq \sum_{m \in \mathcal{M}:\ h(\widehat{s}_m[\mathbf{X}_t], \widehat{s}_{\bar{m}}[\mathbf{X}_t]) \geqslant y} \mathbb{P}_s\left[\psi_{m,\bar{m}}(\mathbf{X}_v) = m \,|\, \mathbf{X}_t\right] \\
&\leq \sum_{m \in \mathcal{M}:\ h(\widehat{s}_m[\mathbf{X}_t], \widehat{s}_{\bar{m}}[\mathbf{X}_t]) \geqslant y} \exp\left[-p(1-2\theta)^2 h^2(\widehat{s}_m[\mathbf{X}_t], \widehat{s}_{\bar{m}}[\mathbf{X}_t])\right] \\
&\leq M \exp\left[-p(1-2\theta)^2 y^2\right] \ ,
\end{aligned}$$

where we successively used the fact that $y \geq \theta^{-1} h(s, \widehat{s}_{\bar{m}}[\mathbf{X}_t])$, equation (5.3) and $|\mathbf{X}_v| = p$.

Since $\mathcal{D}(m) \leq 1$ for all $m \in \mathcal{M}$, we get

$$\begin{aligned}
\mathbb{E}_s\left[\mathcal{D}(\bar{m}) \,|\, \mathbf{X}_t\right] &= \int_0^1 \mathbb{P}_s\left[\mathcal{D}(\bar{m}) \geqslant y \,|\, \mathbf{X}_t\right] dy \\
&\leq \frac{\bar{h}}{\theta} + M \int_{\bar{h}/\theta}^1 \exp\left[-p(1-2\theta)^2 y^2\right] dy \ . \qquad (5.10)
\end{aligned}$$

Moreover, taking H which is, conditionally to $\mathbf{X}_t$, larger than $\bar{h}$ and using that

$$\int_z^\infty \exp\left(\frac{-x^2}{2}\right) dx \leq \frac{1}{z} \exp\left(\frac{-z^2}{2}\right) \text{ for } z > 0 \ ,$$

we obtain

$$\begin{aligned}
\int_{H/\theta}^1 \exp\left[-p(1-2\theta)^2 y^2\right] dy &= \frac{1}{\sqrt{2p}(1-2\theta)} \int_{\sqrt{2p}H(1-2\theta)/\theta}^1 \exp\left(\frac{-x^2}{2}\right) dx \qquad (5.11) \\
&\leq \frac{\theta}{2pH(1-2\theta)^2} \exp\left[\frac{-pH^2(1-2\theta)^2}{\theta^2}\right] \ . \qquad (5.12)
\end{aligned}$$

Hence we derive from (5.9), (5.10) and (5.11)

$$\mathbb{E}_s\left[h\left(\widehat{s}_{\widehat{m}_{\mathrm{THO}}}[\mathbf{X}_t], s\right) \,\middle|\, \mathbf{X}_t\right] \leq \frac{H}{\theta}\left(1 + \theta + \frac{M\theta^2}{2pH^2(1-2\theta)^2} \exp\left[\frac{-pH^2(1-2\theta)^2}{\theta^2}\right]\right) \ .$$

Finally, taking

$$H = \max\left(\mathbb{E}_s\left[\bar{h} \,|\, \mathbf{X}_t\right], \frac{\theta}{(1-2\theta)}\sqrt{\frac{\log M}{p}}\right) \ ,$$

implies

$$\mathbb{E}_s\left[h\left(\widehat{s}_{\widehat{m}_{\mathrm{THO}}}[\mathbf{X}_t], s\right) \,\middle|\, \mathbf{X}_t\right] \leq \frac{H}{\theta}\left(1 + \theta + \frac{1}{2\log M}\right) \ .$$

We conclude by taking the expectation in the above inequality

$$\mathbb{E}_s\left[h\left(\widehat{s}_{\widehat{m}_{\mathrm{THO}}}[\mathbf{X}_t], s\right)\right] \leq \frac{1}{\theta}\left(1 + \theta + \frac{1}{2\log M}\right) \max\left(\mathbb{E}_s\left[h(s, \widehat{s}_{\bar{m}}[\mathbf{X}_t])\right], \frac{\theta}{(1-2\theta)}\sqrt{\frac{\log M}{p}}\right) \ .$$

Chapter 6

A V-fold procedure based on robust tests

Abstract. We define a general V-fold cross-validation type procedure based on robust tests, which is an extension of the hold-out defined by Birgé (2006a). We give some theoretical results showing that, under some weak assumptions on the considered statistical methods, our selected estimator satisfies an oracle type inequality. We also introduce a fast algorithm that implements in practice our procedure. Moreover we show in our simulations that this V-fold performs generally well for estimating a density for different sample sizes, and can handle well-known problems, such as histogram or bandwidth selection. We finally provide a comparison with other classical V-fold procedures and study empirically the influence of the value of V in terms of risk.
NOTA: Ce chapitre est issu d'un travail en cours avec Lucien Birgé[(1)] et Pascal Massart[(2)].

Contents

[(1)]Université Pierre et Marie Curie
[(2)]Université Paris-Sud

6.1 Introduction

The purpose of this paper is to offer a new procedure to solve the following problem. Suppose we are given i.i.d. observations from an unknown distribution P_s to be estimated. This distribution is often assumed to have a density s with respect to some given measure μ, hence our notation, but we shall also consider the case where P_s is not absolutely continuous with respect to μ, keeping the same notation P_s for the true distribution, in which case the subscript s just indicates that P_s is the distribution of the observations.

We also have at hand a family of statistical methods or algorithms $\{\mathcal{A}_m, m \in \mathcal{M}\}$ that can be applied to the observations in order to derive estimators of P_s. How can we use our data in order to choose one potentially optimal algorithm in the family, provided that a criterion of quality for the estimators has been chosen? Let us now be somewhat more precise.

6.1.1 The problem of statistical method choice

We observe an n-sample $\mathbf{X} = \{X_1, \ldots, X_n\}$ of random variables X_i with values in the measured space $(\Xi, \mathcal{Z})$ and we assume (temporarily) that the distribution $P_s = s \cdot \mu$ of the X_i admits a density s with respect to some given positive measure μ on Ξ and that s belongs to some given subset $\mathcal{S}$ of $\mathbb{L}_1(\mu)$. The purpose here is to use the observations in order to design an estimator $\widehat{s} = \widehat{s}(\mathbf{X})$ of s.

There is a huge amount of strategies for solving this estimation problem, depending on the additional assumptions one makes about s. We shall use the notion of *statistical method* (method for short), also denoted *statistical algorithm* in what follows, in order to properly formalize these strategies. Following Arlot & Celisse (2010), we define a *statistical method or algorithm* as any measurable mapping $\mathcal{A}$ from $\bigcup_{q \geqslant 1} \Xi^q$ to $\mathcal{S}$. Such a statistical method associates to any random sample $\mathbf{Y}_q \in \Xi^q$ an estimator $\widehat{s}_q = \mathcal{A}(\mathbf{Y}_q) \in \mathbb{L}_1(\mu)$ of s. A classical criterion from decision theory used to measure the quality of a method $\mathcal{A}$ based on an i.i.d. sample of size q when s obtains is its *risk*: $\mathbb{E}_s\left[\ell(s, \mathcal{A}(\mathbf{Y}_q))\right]$, where ℓ is some given *loss function* and $\mathbb{E}_s$ denotes the expectation when s obtains, i.e. when the distribution of $\mathbf{Y}_q$ is $P_s^{\otimes q}$. The smaller the risk, the better the method $\mathcal{A}$.

To define the risk of an estimator one can consider various loss functions. Some popular ones are derived from a *contrast function* γ (Birgé & Massart, 1993, Definition 1) which is a mapping from $\mathcal{S} \times \Xi$ to $\mathbb{R}$ such that s minimizes over $\mathcal{S}$ the function $t \mapsto \mathbb{E}_s\left[\gamma(t, X)\right]$. The loss ℓ at t is then defined as

$$\ell(s, t) = \mathbb{E}_s\left[\gamma(t, X) - \gamma(s, X)\right] \geqslant 0 \quad \text{for all } t \in \mathcal{S} \ , \tag{6.1}$$

hence $\ell(s, s) = 0$. The $\mathbb{L}_2$-loss derives from the choice $\mathcal{S} = \mathbb{L}_2(\mu) \cap \mathbb{L}_1(\mu)$ and $\gamma(t, x) = \|t\|^2 - 2t(x)$, where $\|t\| = \left[\int_\Xi t^2 d\mu\right]^{1/2}$ denotes the $\mathbb{L}_2$-norm. The Kullback-Leibler loss corresponds to the contrast function $\gamma(t, x) = -\log(t(x))$ with $\mathcal{S}$ being the set of all probability densities with respect to μ.

In this paper, we consider the problem of *statistical method selection*. Let $(\mathcal{A}_m)_{m \in \mathcal{M}}$ denote a collection of candidate statistical methods. Our goal is to choose from the observations $\mathbf{X}$ one of these methods, that is some $\widehat{m}(\mathbf{X}) \in \mathcal{M}$, in order to have the most accurate estimation of s. If we apply all these methods to the sample $\mathbf{X}$ we get the corresponding collection of estimators

$\{\widehat{s}_m = \mathcal{A}_m(\mathbf{X}),\, m \in \mathcal{M}\}$. Given a loss ℓ, the best possible choice for m would be to select $m^* \in \mathcal{M}$ such that

$$\mathbb{E}_s\left[\ell\left(s, \widehat{s}_{m^*}(\mathbf{X})\right)\right] = \inf_{m\in\mathcal{M}} \mathbb{E}_s\left[\ell\left(s, \widehat{s}_m(\mathbf{X})\right)\right] \ .$$

Unfortunately, since s is unknown, all the risks $\mathbb{E}_s\left[\ell(s, \widehat{s}_m)\right]$ are unknown as well and we cannot select the so-called *oracle algorithm* $\mathcal{A}_{m^*}$. One can only hope to choose $\widehat{m} = \widehat{m}(\mathbf{X})$ in such a way that $\mathbb{E}_s\left[\ell(s, \widehat{s}_{\widehat{m}})\right]$ is close to $\mathbb{E}_s\left[\ell(s, \widehat{s}_{m^*})\right]$.

To make this presentation more explicit, let us mention some classical estimation problems that naturally fit into it:

- *Bandwidth selection* (see (Devroye & Lugosi, 2001, Chapter 11)). Let $\Xi = \mathbb{R}$, μ be the Lebesgue measure, $k : \mathbb{R} \to \mathbb{R}$ a given nonnegative function satisfying $\int_\Xi k(x)\,dx = 1$ and $\mathcal{H} = \{h_m, m \in \mathcal{M}\}$ be a finite or countable set of positive bandwidths. We define the *approximation statistical algorithm* $\mathcal{A}_m$ as the method that produces from any sample $\mathbf{Y}_q$ of size q a Parzen-Rosenblatt density estimator with bandwidth h_m, which means that

  $$\mathcal{A}_m(\mathbf{Y}_q)(x) = \frac{1}{qh_m} \sum_{Y_i\in\mathbf{Y}_q} k\left(\frac{x - Y_i}{h_m}\right) \quad \text{for all } x \in \mathbb{R} \ .$$

  The problem of choosing among $\{\widehat{s}_m,\, m \in \mathcal{M}\}$ amounts to select a "best" bandwidth in $\mathcal{H}$, that is the one that minimizes the risk $\mathbb{E}_s\left[\ell(s, \widehat{s}_m)\right]$ with respect to m.

- *Model selection* (see Massart (2007)). We recall that a *model* S for s is any subset of $\mathcal{S}$. It follows from (6.1) that minimizing, for t in S, the loss $\ell(s,t)$ derived from the contrast function γ amounts to minimizing $t \mapsto \mathbb{E}_s\left[\gamma(t, X)\right]$ over S. Since s is unknown, this is impossible but if we replace $\mathbb{E}_s\left[\gamma(t, X)\right]$ by its unbiased empirical version: $\gamma_n(t) = n^{-1}\sum_{i=1}^n \gamma(t, X_i)$ we can derive an estimator with values in S by minimizing $\gamma_n(t)$ instead. This method $\mathcal{A}_S$ is a *minimum contrast algorithm* that provides a *minimum contrast estimator* $\widehat{s}_S(\mathbf{X}) \in \operatorname{argmin}_{t\in S} \gamma_n(t)$ on S. Using for instance, the Kullback-Leibler contrast on a set S of densities leads to the so-called "maximum likelihood estimator" on S.

  If we have at hand some finite or countable collection of models $(S_m)_{m\in\mathcal{M}}$ and a suitable contrast function γ we may associate in this way to each model S_m a minimum contrast algorithm $\mathcal{A}_m$ and the corresponding minimum contrast estimator $\widehat{s}_m(\mathbf{X})$. The problem of "model selection" is to select from the data a "best model" (one with the minimal risk) in the family, leading to a "best" possible minimum contrast estimator.

An alternative choice for the loss function ℓ is the squared Hellinger distance. We recall that the *Hellinger distance* h and the *Hellinger affinity* ρ between two probabilities P and Q defined on Ξ are given respectively by

$$h^2(P,Q) = \frac{1}{2}\int \left(\sqrt{dP} - \sqrt{dQ}\right)^2 \quad \text{and} \quad \rho(P,Q) = \int \sqrt{dP dQ} = 1 - h^2(P,Q) \ , \quad (6.2)$$

where dP and dQ denote the densities of P and Q with respect to any dominating measure (the result being independent of this choice). One advantage of this loss function lies in the fact that h is a distance on the set $\mathcal{P}$ of *all* probabilities on Ξ and therefore does not require that P_s be absolutely continuous with respect to μ, which is one of the reasons why we shall use it in the sequel. In this case we take for $\mathcal{S}$ a set of probability densities with respect to μ and we set, for all t in $\mathcal{S}$ and $P_t = t \cdot \mu$, $\ell(s,t) = h^2(P_s, P_t)$ which we shall write $h^2(s,t)$ for simplicity. We shall also write $\rho(t,u)$ for $\rho(P_t, P_u)$. This loss then leads to the *quadratic Hellinger risk*.

6.1.2 Cross-validation

The biggest difficulty for selecting a method in a given family $\{\mathcal{A}_m,\, m \in \mathcal{M}\}$ comes from the fact that we use the same data $\mathbf{X}$ to build the estimators $\widehat{s}_m(\mathbf{X})$ and to evaluate their quality. It is indeed well-known that evaluating the statistical performance of a method with the same data that have been used for the construction of the corresponding estimator leads to an overoptimistic result. One solution to avoid this drawback is to save a fraction of the initial sample to test the output of the methods $\mathcal{A}_m$ on it. This is the basic idea behind *cross-validation* (CV) which relies on data splitting.

The simplest CV method is the *hold-out* (HO) which corresponds to a single split of the data. The set $\mathbf{X}$ is divided once and for all into two nonempty proper subsets $\mathbf{X}^t$ and $\mathbf{X}^v = \mathbf{X} \setminus \mathbf{X}^t$ to be called respectively the *training* and the *validation* sample. First, with the training sample $\mathbf{X}^t$, we construct a set $\{\mathcal{A}_m(\mathbf{X}^t),\, m \in \mathcal{M}\}$ of preliminary estimators. Then, using the validation sample $\mathbf{X}^v$, we choose a criterion in order to evaluate the quality of each method $\mathcal{A}_m$ from the observation of $\mathcal{A}_m(\mathbf{X}^t)$. Finally, we select $\widehat{m}(\mathbf{X}^v)$ minimizing this criterion over $\mathcal{M}$. Depending on the author, the final estimator might be either $\mathcal{A}_{\widehat{m}}(\mathbf{X}^t)$ (as in Devroye & Lugosi (2001)) or $\mathcal{A}_{\widehat{m}}(\mathbf{X})$ (as in Arlot & Lerasle (2014)). All CV methods are deduced from the HO: instead of using one single partition of our sample, we use different partitions, compute the HO criterion for each one and finally define the CV criterion by averaging all the HO criteria. The goal, by considering several partitions instead of one, is to reduce the variability with the hope that the CV criterion will lead to a more accurate evaluation of the quality of each method.

We shall focus here on V-fold cross-validation (VFCV) which corresponds to a particular set of data splits[(3)]. One divides the sample $\mathbf{X}$ into $V \geq 2$ disjointed and therefore independent subsamples $\mathbf{X}_j$, $j = 1, \ldots, V$, of the same size $p = n/V$ (assuming, for simplicity, that p is an integer) so that $\mathbf{X} = \bigcup_{j=1}^{V} \mathbf{X}_j$. For each split $j \in \{1, \ldots, V\}$, one uses $\mathbf{X}_j^c$ to build the family of "partial estimators" $\{\widehat{s}_m^{(-j)} = \mathcal{A}_m(\mathbf{X}_j^c),\, m \in \mathcal{M}\}$ and the corresponding validation sample $\mathbf{X}_j$ to define an evaluation criterion $\mathrm{crit}_{\mathrm{j}}(m) = \mathrm{crit}_{\mathrm{j}}(m)(\mathbf{X}_j)$ of the method $\mathcal{A}_m$ corresponding to the partition $(\mathbf{X}_j, \mathbf{X}_j^c)$ of the data. One finally selects a strategy $\widehat{m}_{\mathrm{VF}}$ minimizing the averaged criterion:

$$\widehat{m}_{\mathrm{VF}} \in \operatorname*{argmin}_{m \in \mathcal{M}} \mathrm{crit}(m) \qquad \text{with} \qquad \mathrm{crit}(m) = \frac{1}{V} \sum_{j=1}^{V} \mathrm{crit}_{\mathrm{j}}(m) \ .$$

[(3)]The concerned reader should have a look at the survey of Arlot & Celisse (2010) to get a complete overview of other CV methods.

There are as many V-fold procedures as there are different ways to define $\mathrm{crit_j}(m)$. If we work with a loss of the type (6.1), the best estimator in the family $\{\widehat{s}_m^{(-j)},\ m \in \mathcal{M}\}$ is the one minimizing the loss, i.e. the one minimizing $\mathbb{E}_s\left[\gamma(\widehat{s}_m^{(-j)}, X)\right]$ (with X being independent of $\mathbf{X}_j^c$). A natural idea for evaluating this quantity, that we cannot compute since we do not know s, is to estimate it by its empirical version based on the independent sample $\mathbf{X}_j$ of size p, which leads to the criterion

$$\mathrm{crit_j}(m) = \frac{1}{p}\sum_{X_i \in \mathbf{X}_j} \gamma\left(\widehat{s}_m^{(-j)}, X_i\right) \ .$$

In this classical context, we naturally select the statistical method with the lowest estimated average loss $\mathrm{crit}(m)$. The choice $\gamma(t,x) = -\log(t(x))$ leads to the Kullback-Leibler V-fold (KLVF) whereas $\gamma(t,x) = \|t\|^2 - 2t(x)$ provides the Least-Squares V-fold (LSVF). The chosen estimators will be respectively denoted $\widehat{m}_{\mathrm{KLVF}}$ and $\widehat{m}_{\mathrm{LSVF}}$, and the relevant classical criterion will be denoted $\mathrm{crit}_{\mathrm{VFCV}}$ in what follows.

6.1.3 An alternative criterion

When the chosen loss function that we use is the squared Hellinger distance, an alternative empirical criterion to evaluate the quality of an estimator has been proposed by Birgé (1983) following ideas of Le Cam (1973, 1975) then also used by Baraud (2011) to process estimator selection. An HO strategy based on this criterion was first proposed by Birgé (2006a), this latter procedure being recently implemented in Magalhães & Rozenholc (2014). The idea behind the construction is as follows. Suppose we have at hand a set $\mathcal{T}$ of densities with respect to μ and, for each pair (t,u), $t \neq u$, of points of $\mathcal{T}$, a test $\psi_{t,u}$ between t and u ($\psi_{t,u} = \psi_{u,t} = t$ meaning accepting t). Given a sample $\mathbf{X}$ we may perform all the tests $\psi_{t,u}(\mathbf{X})$ and consider the criterion $\mathcal{D}(t)$ defined on $\mathcal{T}$ by

$$\mathcal{D}(t) = \sup_{u \in \mathcal{T},\, u \neq t} h(t,u)\ \mathbf{1}_{\{\,\psi_{t,u}(\mathbf{X}) = u\,\}} \ . \tag{6.3}$$

It immediately follows from this definition that

$$h(t,u) \leq \max\{\mathcal{D}(t), \mathcal{D}(u)\} \quad \text{for all } t,u \in \mathcal{T} \ . \tag{6.4}$$

This definition means that $\mathcal{D}(t)$ is large when there exists some u which is far from t and which is preferred to t by the test $\psi_{t,u}(\mathbf{X})$, suggesting that t is likely to be far from s, at least if s does belong to $\mathcal{T}$. In order that this be actually true even if P_s does not belong to $\{P_t, t \in \mathcal{T}\}$, it is necessary to design suitable tests. It has been shown in Birgé (1983) that one can build a special test $\psi_{t,u}$ between the two Hellinger balls $\mathcal{B}(t,r)$ and $\mathcal{B}(u,r)$ with $r < h(t,u)/2$ (where $\mathcal{B}(t,r)$ denotes the closed ball of center t and radius r in the metric space $(\mathcal{P}, h)$) which posesses the required properties. With this special choice of tests $\psi_{t,u}$ for all pairs (t,u), $\mathcal{D}(t)$ becomes indeed a good indicator of the quality of t as an estimator of s (the smaller $\mathcal{D}(t)$, the better t) and, more generally, of P_t as an estimator of P_s even if P_s is not absolutely continuous with respect to μ. This property of $\mathcal{D}$ suggests to define the following criterion on which to base a new VFCV

procedure. Starting from the family of preliminary density estimators

$$\left\{ \widehat{s}_m^{(-j)} = \mathcal{A}_m(\mathbf{X}_j^c),\, m \in \mathcal{M},\; 1 \leq j \leq V \right\} \;\; ,$$

we build all the corresponding tests $\psi_{\widehat{s}_l^{(-j)}, \widehat{s}_m^{(-j)}}(\mathbf{X}_j)$, hereafter denoted for simplicity by $\psi_{l,m}(\mathbf{X}_j)$, between the densities $\widehat{s}_l^{(-j)}$ and $\widehat{s}_m^{(-j)}$ for $l, m \in \mathcal{M}$, $l \neq m$. Then, for each j and m, we define the criterion $\mathrm{crit_j}(m)$ by

$$\mathrm{crit_j}(m) = \mathcal{D}_j^2(m) \quad \text{with} \quad \mathcal{D}_j(m) = \sup_{l \in \mathcal{M},\, l \neq m} h\left(\widehat{s}_l^{(-j)}, \widehat{s}_m^{(-j)}\right) \mathbf{1}_{\left\{\psi_{l,m}(\mathbf{X}_j) = l\right\}} \;\; . \tag{6.5}$$

We then naturally define our test-based V-fold criterion as

$$\mathcal{C}_V^{\mathrm{TVF}}(m) := \overline{\mathcal{D}}(m) = \frac{1}{V} \sum_{j=1}^{V} \mathcal{D}_j^2(m) \quad \text{for all } m \in \mathcal{M} \;\; .$$

Up to our knowledge, this is the first V-fold type procedure based on the Hellinger distance. Note that this construction requires that the estimators $\widehat{s}_{m,j}$ be genuine probability densities with respect to μ which we shall assume from now on.

6.1.4 Organization of the paper

Our goal is to study our new VFCV procedure from both a theoretical and a practical point of view. Section 6.2 is dedicated to the theoretical study. We prove some oracle type result thanks to concentration inequalities concerning the V-fold criterion that we use and discuss in details the implications of the resulting risk bounds to the case of histogram estimators. Section 6.3 contains an empirical study of the influence of the value of V on the performance of our procedure in terms of Hellinger risk and also comparisons with classical V-fold and calibrated procedures. Section 6.4 describes the fast algorithm that we have designed and implemented in order to compute the selected estimator efficiently. We also provide additional simulations in Section 6.5.

6.2 T-V-fold

As already mentioned, the method proposed in Birgé (2006a) is based on tests and it results in what Birgé called T-estimators (T for "test"). We shall therefore call our cross-validation method based on the same tests T-V-fold cross-validation (TVF for short).

6.2.1 Tests between Hellinger balls

The tests that are needed for our procedure have to satisfy the following assumption, which ensures their robustness. We recall that $\mathcal{S}$ is the set of all probability densities with respect to μ.

Assumption (TEST). *Let $\theta \in (0, 1/2)$ be given. For all t and u in $\mathcal{S}$, $z \in \mathbb{R}$ and $r = \theta h(t,u)$ there exists some test statistic $T_{t,u,\theta}(\mathbf{X})$ depending on t, u, θ and $\mathbf{X}$ with the following properties.*

The test $\psi_{t,u}$ between t and u defined by

$$\psi_{t,u}(\mathbf{X}) = \begin{cases} t & \text{if} \quad T_{t,u,\theta}(\mathbf{X}) > z; \\ u & \text{if} \quad T_{t,u,\theta}(\mathbf{X}) < z, \end{cases} \tag{6.6}$$

with an arbitrary choice when $t = u$, satisfies

$$\sup_{\{\, P_s \in \mathcal{P} \,|\, h(s,t) \leq r \,\}} \mathbb{P}_s \left[\psi_{t,u}(\mathbf{X}) = u \right] \leq \exp\left[-n(1-2\theta)^2 h^2(t,u) + z \right] \tag{6.7}$$

and

$$\sup_{\{\, P_s \in \mathcal{P} \,|\, h(s,u) \leq r \,\}} \mathbb{P}_s \left[\psi_{t,u}(\mathbf{X}) = t \right] \leq \exp\left[-n(1-2\theta)^2 h^2(t,u) - z \right] \quad , \tag{6.8}$$

where $\mathbb{P}_s$ denotes the probability that gives $\mathbf{X}$ the distribution $P_s^{\otimes n}$.

Any test satisfying (6.7) and (6.8) will be suitable for our needs.

Tests between balls In order to define tests between two Hellinger balls $\mathcal{B}(t,r)$ and $\mathcal{B}(u,r)$ with $r = \theta h(t,u)$, $0 < \theta < 1/2$, Birgé introduced the following test statistic

$$T_{t,u,\theta}(\mathbf{X}) = \sum_{i=1}^{n} \log\left(\frac{\sin(\omega(1-\theta))\sqrt{t}(X_i) + \sin(\omega\theta)\sqrt{u}(X_i)}{\sin(\omega(1-\theta))\sqrt{u}(X_i) + \sin(\omega\theta)\sqrt{t}(X_i)} \right) \quad \text{with } \omega = \arccos \rho(t,u) \ . \tag{6.9}$$

We should notice that for $\theta = 0$, the test given by (6.9) is exactly the likelihood ratio test between t and u. The fact that Assumption (TEST) holds for this test whatever $\theta \in (0, 1/2)$ has been proven in Birgé (1984a) and a more uptodate version is to be found in Birgé (2013) (see Corollary 1).

6.2.2 TVF estimators

Let $(\Delta_m)_{m \in \mathcal{M}}$ denote some collection of positive numbers satisfying

$$\Delta_m \geqslant 0 \quad \text{for all } m \in \mathcal{M}, \qquad \text{and} \qquad \frac{1}{2} \leq \Gamma = \sum_{m \in \mathcal{M}} \exp(-\Delta_m) < \infty \ . \tag{6.10}$$

Starting from the family of estimators $\widehat{s}_{m,j}$ defined in Section 6.1.3, we consider the corresponding tests $\psi_{l,m}(\mathbf{X}_j) = \psi_{\widehat{s}_l^{(-j)}, \widehat{s}_m^{(-j)}}(\mathbf{X}_j)$ with $t = \widehat{s}_l^{(-j)}$, $u = \widehat{s}_m^{(-j)}$ and $z = \Delta_l - \Delta_m$ in (6.6). This results in the estimator $\widehat{s}_{\widehat{m}_{\mathrm{TVF}}}$ derived from the method $\mathcal{A}_{\widehat{m}_{\mathrm{TVF}}}$ with

$$\widehat{m}_{\mathrm{TVF}} \in \operatorname*{argmin}_{m \in \mathcal{M}} \overline{\mathcal{D}}(m) = \operatorname*{argmin}_{m \in \mathcal{M}} \frac{1}{V} \sum_{j=1}^{V} \mathcal{D}_j^2(m) \ . \tag{6.11}$$

6.2.3 Assumption on the family of methods

The idea of V-fold relies on the heuristic that, for each method $\mathcal{A}_m$, the observation of V partial estimators $\widehat{s}_m^{(-j)}$, $1 \leq j \leq V$ based on samples of size $n - p$ with $p = n/V$ allows to predict the behavior of an estimator $\widehat{s}_m$ based on an n-sample. This requires that there exists a link between

the loss of $\widehat{s}_m$ and the losses of the $\widehat{s}_m^{(-j)}$. We shall need the following assumption on the collection of methods we consider.

Assumption (LOSS). *For all methods $\mathcal{A}_m$ with $m \in \mathcal{M}$, the loss at s satisfies*

$$h^2\left(s, \widehat{s}_m\right) \leqslant \frac{1}{V}\sum_{j=1}^{V} h^2\left(s, \widehat{s}_m^{(-j)}\right) \quad .$$

This implies in particular that $R(\mathcal{A}_m, n, s) \leq R(\mathcal{A}_m, n-p, s)$, where

$$R(\mathcal{A}, q, s) = \mathbb{E}_s\left[h^2\left(s, \mathcal{A}(\mathbf{Y}_q)\right)\right]$$

denotes the risk at s of the method $\mathcal{A}$ based on a sample of size q. Assumption (LOSS) is in particular satisfied by the "additive estimators" (Devroye & Lugosi, 2001, Chapter 10).

Definition 6.1. *An additive estimator $\widehat{s} = \widehat{s}(\mathbf{X})$ derived from a sample $\mathbf{X}$ of size n is an estimator that can be written in the form:*

$$\widehat{s}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{K}(x, X_i) \quad \textit{for all } x \in \Xi \ , \tag{6.12}$$

where $\mathcal{K}$ is a symmetric and real valued function from $\Xi \times \Xi$ to $\mathbb{R}$.

There is a huge amount of literature about these estimators which already appeared in an early version in Whittle (1958). The first results about their asymptotic properties in general were made by Watson & Leadbetter (1964b), followed by Winter (1975) and Walter & Blum (1979) who established rates (the latter authors called them *delta sequence density estimators*). They were introduced in the context of CV by Rudemo (1982) and used by Marron (1987) for comparison of CV techniques. As shown in Walter & Blum (1979) and Devroye & Lugosi (2001), additive estimators include in particular:

- *Histogram estimators*. Given a partition $\{I_\lambda,\ \lambda \in \Lambda\}$ of Ξ with $\mu(I_\lambda) > 0$ for all λ one defines the histogram estimator based on this partition as

$$\widehat{s}(x) = \sum_{\lambda\in\Lambda}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{I_\lambda}(X_i)\right)\frac{\mathbb{1}_{I_\lambda}(x)}{\mu(I_\lambda)} \quad . \tag{6.13}$$

  It corresponds to the case of $\mathcal{K}(x, X_i) = \sum_{\lambda\in\Lambda}[\mu(I_\lambda)]^{-1}\mathbb{1}_{I_\lambda}(X_i)\mathbb{1}_{I_\lambda}(x)$.

- *Parzen kernel estimators on the line*. Set $\mathcal{K}(x, X_i) = h^{-1}k\left(h^{-1}(X_i - x)\right)$ for a given nonnegative kernel k with $\int_{\mathbb{R}} k(x)\,dx = 1$ and a positive bandwidth h. This leads to a density estimator with respect to the Lebesgue measure on $\mathbb{R}$.

It is straightforward to check that if the method $\mathcal{A}_m$ results in additive estimators, the following relationship which says that the estimator built with the whole sample is exactly the convex

combination of the V partial estimators holds:

$$\widehat{s}_m = \frac{1}{V}\sum_{j=1}^{V} \widehat{s}_m^{(-j)} \quad . \tag{6.14}$$

As a consequence, we get the following elementary property:

Proposition 6.1. *Any statistical method $\mathcal{A}_m$ which results in additive estimators does satisfy Assumption (LOSS).*

Proof: It follows from (6.14) and the concavity of the square root function that

$$\rho(s,\widehat{s}_m) = \rho\left(s, \frac{1}{V}\sum_{j=1}^{V}\widehat{s}_m^{(-j)}\right) \geq \frac{1}{V}\sum_{j=1}^{V}\rho\left(s,\widehat{s}_m^{(-j)}\right) \quad ,$$

which is exactly Assumption (LOSS) in view of (6.2). □

6.2.4 The main result

Proposition 6.1 ensures that, for the methods that we consider, the loss of some estimator is bounded by the mean of the losses of the partial estimators. This motivates us to work separately on each split $j \in \{1,\dots,V\}$ and then to deduce a risk bound for the estimator built with the whole sample. It is therefore natural to study for each j the deviations of random variable $\mathcal{D}_j(\cdot)$. A deviation inequality for $\mathcal{D}$ has been proven in Theorem 9 in Birgé (2006a). Let us now recall it and provide a proof for the sake of completeness.

Proposition 6.2. *Let $(\Delta_m)_{m\in\mathcal{M}}$ be a collection of weights satisfying* (6.10) *and*

$$A = \frac{n(1-2\theta)^2}{2V}; \qquad y_{m,j} = \max\left(\frac{h\left(s,\widehat{s}_m^{(-j)}\right)}{\theta}, \sqrt{\frac{\Delta_m}{A}}\right) \quad .$$

Then, for all $m \in \mathcal{M}$, and $j \in \{1,\dots,V\}$,

$$\mathbb{P}_s\left[\mathcal{D}_j(m) \geqslant y \,\middle|\, \mathbf{X}_j^c\right] \leq \Gamma \exp\left[-2Ay^2 + \Delta_m\right] \qquad \textit{for } y \geqslant y_{m,j} \quad .$$

Proof: Let us fix some $m \in \mathcal{M}$ and $j \in \{1,\dots,V\}$ and work conditionally to the training sample $\mathbf{X}_j^c$ so that the collection of estimators $(\widehat{s}_l^{(-j)})_{l\in\mathcal{M}}$ is "frozen". We perform the robust test $\psi_{l,m}(\mathbf{X}_j)$ with $z = \Delta_l - \Delta_m$ in (6.7). Then

$$\begin{aligned}
\mathbb{P}_s\left[\mathcal{D}_j(m) \geqslant y \,\middle|\, \mathbf{X}_j^c\right] &= \mathbb{P}_s\left[\exists\, l \in \mathcal{M} \text{ such that } h(\widehat{s}_l^{(-j)}, \widehat{s}_m^{(-j)}) \geqslant y \text{ and } \psi_{l,m}(\mathbf{X}_j) = l \,\middle|\, \mathbf{X}_j^c\right] \\
&\leq \sum_{l\in\mathcal{M}:\ h(\widehat{s}_l^{(-j)},\widehat{s}_m^{(-j)})\geqslant y} \mathbb{P}_s\left[\psi_{l,m}(\mathbf{X}_j) = l \,\middle|\, \mathbf{X}_j^c\right] \\
&\leq \sum_{l\in\mathcal{M}:\ h(\widehat{s}_l^{(-j)},\widehat{s}_m^{(-j)})\geqslant y} \exp\left[-2Ah^2\left(\widehat{s}_l^{(-j)}, \widehat{s}_m^{(-j)}\right) - (\Delta_l - \Delta_m)\right]
\end{aligned}$$

$$\leq \exp\left[-2Ay^2 + \Delta_m\right] \sum_{l \in \mathcal{M}} \exp(-\Delta_l) \leq \Gamma \exp\left[-2Ay^2 + \Delta_m\right] \ ,$$

where we successively used the fact that $y \geq y_{m,j} \geq \theta^{-1} h(s, \widehat{s}_m^{(-j)})$ and (6.10). □

For each fixed j, that is conditionally to each $\mathbf{X}_j^c$, we deal with some "fixed geometrical configuration" since the points $(\widehat{s}_m^{(-j)})_{m \in \mathcal{M}}$ are "frozen". On this configuration, Proposition 6.2 controls the deviations of $\mathcal{D}_j^2(m)$ which allows us to bound the expectation of $\overline{\mathcal{D}}(m)$. This results in the following theorem.

Theorem 6.1. *Under Assumption (LOSS), the estimator $\widehat{s}_{\widehat{m}_{\mathrm{TVF}}} = \mathcal{A}_{\widehat{m}_{\mathrm{TVF}}}(\mathbf{X})$ with $\widehat{m}_{\mathrm{TVF}}$ minimizing the criterion $\overline{\mathcal{D}}(m)$ satisfies the following inequality:*

$$\mathbb{E}_s\left[h^2\left(s, \widehat{s}_{\widehat{m}_{\mathrm{TVF}}}\right)\right] \leq \inf_{m \in \mathcal{M}} \left\{ 2\left(\frac{\theta^2+2}{\theta^2}\right) R\left(\mathcal{A}_m, \frac{V-1}{V}n, s\right) + \frac{4V[\Delta_m + \log(2\Gamma) + 1]}{n(1-2\theta)^2} \right\} \ . \quad (6.15)$$

Proof: Let m' be any algorithm in $\mathcal{M}$. It follows from (6.4) that, for all $m \in \mathcal{M}$ and $1 \leq j \leq V$,

$$h\left(s, \widehat{s}_{m'}^{(-j)}\right) \leqslant h\left(s, \widehat{s}_m^{(-j)}\right) + h\left(\widehat{s}_{m'}^{(-j)}, \widehat{s}_m^{(-j)}\right) \leqslant h\left(s, \widehat{s}_m^{(-j)}\right) + \max\left(\mathcal{D}_j(m), \mathcal{D}_j(m')\right).$$

Setting $m' = \widehat{m}_{\mathrm{TVF}} = \widehat{m}$ for short, we derive that

$$\begin{aligned}
\frac{1}{V}\sum_{j=1}^{V} h^2\left(s, \widehat{s}_{\widehat{m}}^{(-j)}\right) &\leqslant 2\left\{ \frac{1}{V}\sum_{j=1}^{V} h^2\left(s, \widehat{s}_m^{(-j)}\right) + \frac{1}{V}\sum_{j=1}^{V} \max\left(\mathcal{D}_j^2(m), \mathcal{D}_j^2(\widehat{m})\right) \right\} \\
&\leqslant 2\left\{ \frac{1}{V}\sum_{j=1}^{V} h^2\left(s, \widehat{s}_m^{(-j)}\right) + \frac{1}{V}\sum_{j=1}^{V} (\mathcal{D}_j^2(m) + \mathcal{D}_j^2(\widehat{m})) \right\} \\
&\leqslant \frac{2}{V}\sum_{j=1}^{V} h^2\left(s, \widehat{s}_m^{(-j)}\right) + 4\overline{\mathcal{D}}(m) \ ,
\end{aligned}$$

for all $m \in \mathcal{M}$. Using Assumption (LOSS) and taking expectations, we derive that

$$\mathbb{E}_s\left[h^2\left(s, \widehat{s}_{\widehat{m}}\right)\right] \leq \frac{1}{V}\sum_{j=1}^{V} \mathbb{E}_s\left[h^2\left(s, \widehat{s}_{\widehat{m}}^{(-j)}\right)\right] \leq 2R(\mathcal{A}_m, n-p, s) + 4\mathbb{E}_s\left[\overline{\mathcal{D}}(m)\right], \quad (6.16)$$

since the risk of $\widehat{s}_m^{(-j)}$ is the same for all j and equal to $R(\mathcal{A}_m, n-p, s)$.

Let now m and j be fixed. Integrating the bound for $\mathbb{P}_s\left[\mathcal{D}_j^2(m) \geqslant y \,\middle|\, \mathbf{X}_j^c\right]$ provided by Proposition 6.2 with respect to y leads to

$$\mathbb{E}_s\left[\mathcal{D}_j^2(m) \,\middle|\, \mathbf{X}_j^c\right] \leq y_{m,j}^2 + \Gamma e^{\Delta_m} \int_{y_{m,j}^2}^{1} e^{-2Az}\, dz \leq y_{m,j}^2 + \frac{\Gamma e^{\Delta_m}}{A} \exp\left(-2Ay_{m,j}^2\right)$$

and, since $Ay_{m,j}^2 \geq \Delta_m$,

$$\mathbb{E}_s\left[\mathcal{D}_j^2(m)\right] \leq \mathbb{E}_s\left[y_{m,j}^2\right] + \Gamma A^{-1}\exp(-\Delta_m) \leq \frac{1}{\theta^2}\mathbb{E}_s\left[h^2\left(s, \widehat{s}_m^{(-j)}\right)\right] + \frac{\Delta_m + \Gamma e^{-\Delta_m}}{A} \ .$$

Finally

$$\mathbb{E}_s\left[\overline{\mathcal{D}}_m\right] \leq \frac{1}{\theta^2}R(\mathcal{A}_m, n-p, s) + \frac{\Delta_m + \Gamma e^{-\Delta_m}}{A} \ .$$

One should then observe that changing Δ_m into $\Delta_m + B$ with $B \geq 0$ does not change the procedure since the tests only depend on differences $\Delta_m - \Delta_l$. Since the new weights $\Delta_m + B$ also satisfy (6.10) with Γ changed to Γe^{-B}, the previous bound remains valid for the new weights leading to

$$\mathbb{E}_s\left[\overline{\mathcal{D}}_m\right] \leq \frac{1}{\theta^2}R(\mathcal{A}_m, n-p, s) + \frac{\Delta_m + B + \Gamma e^{-\Delta_m - 2B}}{A} \ .$$

An optimization with respect to B (taking into account the fact that $\Gamma \geq 1/2$) together with (6.16) leads to our conclusion. □

It is often the case that $\mathcal{M}$ is finite and that we use equal weights $\Delta_m = \Delta \leq \log(2|\mathcal{M}|)$ for all $m \in \mathcal{M}$, in which case $\Gamma = |\mathcal{M}|e^{-\Delta}$ which leads to the following risk bound which only depends on $|\mathcal{M}|$:

$$\mathbb{E}_s\left[h^2\left(s, \widehat{s}_{\widehat{m}_{\mathrm{TVF}}}\right)\right] \leq 2\left(\frac{\theta^2+2}{\theta^2}\right)\inf_{m\in\mathcal{M}} R\left(\mathcal{A}_m, \frac{V-1}{V}n, s\right) + \frac{4V\log(2e|\mathcal{M}|)}{n(1-2\theta)^2} \ .$$

If we assume, to be specific and for simplicity, that $\theta = 1/4$ and that $\log(2\Gamma) + 1 \leq 3\Delta_m$ for all m, (6.15) becomes

$$\mathbb{E}_s\left[h^2\left(s, \widehat{s}_{\widehat{m}_{\mathrm{TVF}}}\right)\right] \leq 66\inf_{m\in\mathcal{M}}\left\{R\left(\mathcal{A}_m, \frac{V-1}{V}n, s\right) + \frac{V\Delta_m}{n}\right\} \ . \tag{6.17}$$

Remark. It should be noted that the following analogue of (6.15) holds

$$\mathbb{E}_s\left[h^2\left(s, \widehat{s}_{\widehat{m}_{\mathrm{TVF}}}\right)\right] \leq \inf_{m\in\mathcal{M}}\left\{C_1(\theta,a)R\left(\mathcal{A}_m, \frac{V-1}{V}n, s\right) + C_2(\theta,a)\frac{V(\Delta_m + \log(2\Gamma)+1)}{n}\right\}$$

if we replace Assumption (TEST) by the following.

Assumption (TEST'). *Let $\theta \in (0, 1/2)$ and $a > 0$ be given. For all t and u in $\mathcal{S}$, $z \in \mathbb{R}$ and $r = \theta h(t,u)$ there exists some test statistic $T_{t,u,\theta}(\mathbf{X})$ depending on t, u, θ and $\mathbf{X}$ with the following properties. The test $\psi_{t,u}$ between t and u defined by*

$$\psi_{t,u}(\mathbf{X}) = \begin{cases} t & \text{if} \quad T_{t,u,\theta}(\mathbf{X}) > z; \\ u & \text{if} \quad T_{t,u,\theta}(\mathbf{X}) < z, \end{cases}$$

with an arbitrary choice when $t = u$, satisfies

$$\sup_{\{\, P_s \in \mathcal{P} \,|\, h(s,t) \leq r \,\}} \mathbb{P}_s\left[\psi_{t,u}(\mathbf{X}) = u\right] \leq \exp\left[-nah^2(t,u) + z\right]$$

and

$$\sup_{\{\,P_s\in\mathcal{P}\,|\,h(s,u)\le r\,\}} \mathbb{P}_s\left[\psi_{t,u}(\mathbf{X})=t\right] \le \exp\left[-nah^2(t,u)-z\right] \quad .$$

In particular Baraud (2011) introduced the following test statistic that relies on a variational formula for the Hellinger affinity. For $r=(t+u)/2$, let

$$T_{t,u}(\mathbf{X}) = \frac{1}{2}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\sqrt{t}(X_i)-\sqrt{u}(X_i)}{\sqrt{r}(X_i)} + \int\left(\sqrt{t(x)}-\sqrt{u(x)}\right)\sqrt{r(x)}\,d\mu(x)\right). \quad (6.18)$$

The corresponding test $\psi_{t,u}$ actually satisfies Assumption (TEST') for small enough constants θ and a. This follows from Baraud (2008, unpublished manuscript). Therefore the test $\psi(t,u)$ derived from Baraud's statistic could be used instead of the tests between balls. Some simulations based on this alternative test will be provided in Section 6.5.

6.2.5 The case of regular histograms

Although our risk bound (6.15) is certainly not optimal it is nevertheless already enlightening as shown by the following example. Let us consider the problem of estimating an unknown density with respect to the Lebesgue measure on $[0,1]$. We consider, for each positive integer m, the histogram estimator $\widehat{s}_m$ based on the partition $\mathcal{I}_m$ of $[0,1]$ into m intervals of length m^{-1}. It is known (Birgé & Rozenholc, 2006, Theorem 1) that the risk at s of an histogram estimator $\widehat{s}_m$ built on $\mathcal{I}_m$ from n i.i.d. observations is bounded by

$$\mathbb{E}_s\left[h^2\left(s,\widehat{s}_m\right)\right] \le h^2\left(s,\overline{s}_m\right) + \frac{m-1}{2n} \quad (6.19)$$

where $\overline{s}_m$ is the $\mathbb{L}_2$-projection of s onto the m-dimensional linear space of piecewise constant functions on the partition $\mathcal{I}_m$. It is also shown in this theorem that this bound is asymptotically optimal, up to a factor 4, since the asymptotic risk (when n tends to infinity) is of the form

$$\mathbb{E}_s\left[h^2\left(s,\widehat{s}_m\right)\right] = h^2\left(s,\overline{s}_m\right) + \frac{m-1}{8n}\left(1+\mathrm{o}(1)\right) \quad . \quad (6.20)$$

In view of (6.20), the bound in (6.19) can be considered as optimal, up to a constant factor.

It follows from (6.19) that

$$R\left(\mathcal{A}_m, \frac{V-1}{V}n, s\right) \leqslant h^2\left(s,\overline{s}_m\right) + \frac{(m-1)V}{2n(V-1)} = h^2\left(s,\overline{s}_m\right) + \frac{m-1}{2n} + \frac{m-1}{2n(V-1)} \quad (6.21)$$

and

$$\inf_{m\in\mathcal{M}} \mathbb{E}_s\left[h^2\left(s,\widehat{s}_m\right)\right] \le h^2\left(s,\overline{s}_{m^*}\right) + \frac{(m^*-1)}{2n} = \inf_{m\in\mathcal{M}}\left\{h^2\left(s,\overline{s}_m\right) + \frac{(m-1)}{2n}\right\} \quad , \quad (6.22)$$

where this last bound can be considered as a benchmark for the risk of any selection procedure applied to our family of histograms. Since the Hellinger distance is bounded by 1, it clearly appears that one should restrict to values of m that are smaller than $2n$.

Setting $\theta = 1/4$ and assuming that $\mathcal{M} = \{1, \ldots, n\}$ and

$$1 \leq 2\Gamma \leq \exp(3\Delta_m - 1) \quad \text{for all } m \in \mathcal{M} \ , \tag{6.23}$$

we derive from (6.17) and (6.21) that

$$\frac{1}{66}\mathbb{E}_s\left[h^2\left(s, \widehat{s}_{\widehat{m}_{\mathrm{TVF}}}\right)\right] \leq \inf_{m\in\mathcal{M}}\left\{R\left(\mathcal{A}_m, \frac{V-1}{V}n, s\right) + \frac{V\Delta_m}{n}\right\} \tag{6.24}$$

$$\leq \inf_{m\in\mathcal{M}}\left\{\left(h^2\left(s, \overline{s}_m\right) + \frac{m-1}{2n}\right) + \frac{m-1}{2n(V-1)} + \frac{V\Delta_m}{n}\right\} \tag{6.25}$$

$$\leq \left[h^2\left(s, \overline{s}_{m^*}\right) + \frac{(m^*-1)}{2n} + \frac{m^*-1}{2n(V-1)} + \frac{V\Delta_{m^*}}{n}\right] \ , \tag{6.26}$$

with m^* defined by (6.22). We see from (6.25) that, up to the multiplicative constant 66, we have to optimize with respect to m a bound for the risk of $\widehat{s}_m$ plus a residual term which depends in a non-monotonous way of V. The bound (6.26) shows that, up to a constant factor, we actually recover our benchmark (6.22) plus an error term which writes

$$g(V-1) \quad \text{with} \quad g(x) = \frac{1}{n}\left(\frac{m^*-1}{2x} + x\Delta_{m^*}\right) + \frac{\Delta_{m^*}}{n} \ .$$

Clearly, $g(x)$ is minimum for $x = x_0 = \sqrt{(m^*-1)/(2\Delta_{m^*})}$. It follows that the optimal value of V is two if $m^* - 1 \leq 2\Delta_{m^*}$. This occurs in particular if $m^* = 1$, for instance when P_s is the uniform distribution on $[0,1]$ or close enough to it. It also occurs if $\Delta_m \geq (m-1)/2$ for all $m \geq 2$. Let us now consider the situation for which $m^* - 1 > 2\Delta_{m^*}$ so that $x_0 > 1$ and the optimal value of V belongs to $(x_0 - 1, x_0 + 1)$. If $(m-1)/\Delta_m$ is an increasing function of m, the optimal value of V will be a nondecreasing function of m^* which depends on the true unknown value of s, large values of m^* leading to large values for V and vice-versa. For instance, the choice of equal weights, $\Delta_m = \log n$ for $m \in \mathcal{M}$ leads to $\Gamma = 1$ which satisfies (6.23) and to an optimal V of order $\sqrt{(m^*-1)/(2\log n)}$. But this choice of Δ_m is certainly not optimal in view of (6.24). A better one would be $\Delta_m = 1 + 2\log m$ which also satisfies (6.23) but improves (6.24) substantially. Then the optimal value of V is of order $\sqrt{(m^*-1)/(2+4\log m^*)}$, still depending on the true unknown s. Only larger values of Δ_m of the form $\Delta_m = a(m-1)$ for $m \geq 2$, that deteriorate the bound (6.24) and therefore should not be recommended, lead to an optimal value of V which is independent of m^*, hence of s.

This dependence of the optimal value of V with respect to the true density s will actually be confirmed by our simulations below. A density which is difficult to estimate by a histogram with a few bins will lead to a large value of m^* hence a large optimal V while a simple density, for which m^* is rather small is better estimated by a V-fold with a small V.

6.3 Empirical study

The theoretical bounds that we have derived, for instance (6.17), are quite pessimistic because of the large constants that are present in our risk bounds. It is therefore crucial to know whether

such large values are only artifacts or really enter the risk. In order to check the real quality of our selection procedure and evaluate the influence of the various parameters involved in it, we performed an extensive set of simulations the results of which are described below.

6.3.1 Simulation protocol

We have studied the performances of the TVF procedure on 18 out of the 28 densities described in the *benchden* [(4)] R-package (Mildenberger & Weinert, 2012) which provides a full implementation of the distributions introduced in Berlinet & Devroye (1994) as benchmarks for nonparametric density estimation. We only show our simulations for the eleven densities in the subset $\mathcal{L} = \{s_i,\ i = 1, 2, 3, 4, 5, 7, 12, 13, 22, 23, 24\}$ which can be viewed in Figure 6.1, except for the uniform density s_1 on $[0, 1]$.

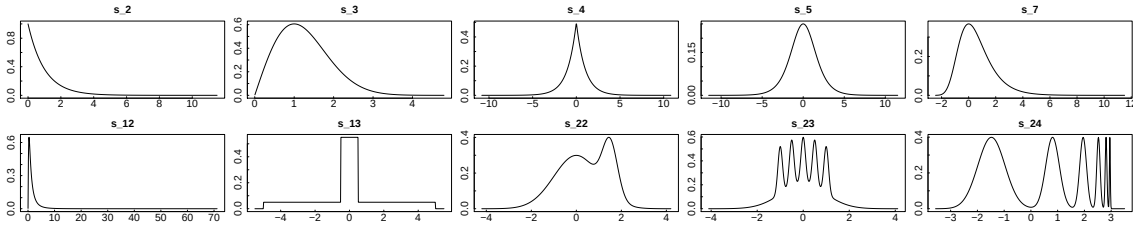

Figure 6.1: All densities mentioned in the paper.

For a given loss $\ell = h^2, d_1$ or d_2^2 (namely the Hellinger, $\mathbb{L}_1$ and $\mathbb{L}_2$-loss), we decide to judge the accuracy of some estimator $\widetilde{s} = \widehat{s}_{\widehat{m}}$ by estimating its risk $R(\widetilde{s}, s, \ell) = \mathbb{E}_s\left[\ell(s, \widetilde{s})\right]$. To do so, we generate 1000 pseudo-random samples $\mathbf{X}^i = \{X_1^i, \ldots, X_n^i\}$, $1 \leqslant i \leqslant 1000$, of size n and density s and we approximate $R(\widetilde{s}, s, \ell)$ by its empirical version:

$$\overline{R}_n\left(\widetilde{s}, s, \ell\right) = \frac{1}{1000}\sum_{i=1}^{1000} \ell\left(s, \widetilde{s}(\mathbf{X}^i)\right) \quad .$$

As in Magalhães & Rozenholc (2014), we have considered several families of estimators. In particular, we present here our simulations for the well-known problems of bandwidth selection for kernel estimators and number of bin selection for regular histograms. For this purpose, we used regular histograms and Gaussian kernel estimators:

- $\mathcal{F}_{\mathrm{R}}$ is the set of regular histograms with bin number varying from 1 to $\lceil n/\log(n)\rceil$ as described in Birgé & Rozenholc (2006),
- $\mathcal{F}_{\mathrm{K}}$ is the set of of Gaussian kernel estimators with bandwidths chosen as

$$h_m = \frac{1}{n\log(n)}\left(1 + \frac{1.5}{\log(n)}\right)^m, \qquad \text{for} \quad m = 1, \ldots, (\log(n))^2 \quad ,$$

[(4)] Available on the CRAN `http://cran.r-project.org`.

- $\mathcal{F}_{\mathrm{KR}} = \mathcal{F}_{\mathrm{K}} \cup \mathcal{F}_{\mathrm{R}}$.

Besides the classical VF methods, we have considered two alternative estimation methods that are known to perform well in practice in order to have an idea of the performance of the T-V-fold as compared to some especially calibrated procedures. When studying the problem of bandwidth selection, we compared the TVF with the unbiased cross-validation selector, implemented in the *density* generic function available in R, which provides an estimator which does not belong to the set $\{\widehat{s}_m,\, m \in \mathcal{M}\}$. When dealing with the partition problem we also implemented the penalization procedure of Birgé and Rozenholc (described in Birgé & Rozenholc (2006)) which selects a regular histogram in $\mathcal{F}_{\mathrm{R}}$. These two competitors will be denoted "UCV" and "BR" in the study below. To implement the TVF and process our simulations we used an algorithm which is described in Section 6.4 with the tests defined in (6.9) with constant weights $\Delta_m = \Delta = 0$ for all $m \in \mathcal{M}$.

We made thousands of simulations (varying the sample size n, the density, the family of estimators, the number of split V, etc.) but since the results were very similar, we only show the conclusion for $n = 500$, $V = 2, 5, 10$ and 20.

6.3.2 Influence of θ

As in Section 5.1 of Magalhães & Rozenholc (2014), we have studied the influence of the parameter θ, used in equation (6.9), on the TVF procedure. It influences the test theoretically and thus the entire procedure. Since on the one hand $\theta = 0$ corresponds to the KLVF and on the other hand θ must be less than 1/2, we have made comparisons between versions of $\widetilde{s}_{\mathrm{TVF}}$ deduced from the test with $\theta \in \Theta = \{1/16, 1/8, 1/4, 3/8, 7/16\}$. For the sake of clarity and to emphasize the stability of the behavior of the procedure in terms of risk, we present for each V the ratio

$$\inf_{s\in\mathcal{L}} \left\{ \inf_{\theta\in\Theta} \overline{R}_n\left(\widehat{s}_{\widehat{m}(\theta)}, s, h^2\right) \Big/ \sup_{\theta\in\Theta} \overline{R}_n\left(\widehat{s}_{\widehat{m}(\theta)}, s, h^2\right) \right\} \quad ,$$

which gives the largest difference in terms of risk among the densities in $\mathcal{L}$. The closer it is to 1, the more stable the procedure is when θ varies for all densities.

family	$V = 2$	$V = 5$	$V = 10$	$V = 20$
$\mathcal{F}_{\mathrm{R}}$	92,95	94,87	96,39	96,96
$\mathcal{F}_{\mathrm{K}}$	91,31	92,94	94,79	96,44
$\mathcal{F}_{\mathrm{KR}}$	87,81	94,36	97,48	95,15

Table 6.1: Ratios multiplied by 100 for $n = 500$ and families $\mathcal{F}_{\mathrm{R}}$ and $\mathcal{F}_{\mathrm{K}}$, see the text.

From this picture we conclude that θ has little influence on the quality of the resulting estimator for families $\mathcal{F}_{\mathrm{K}}$ and $\mathcal{F}_{\mathrm{R}}$, even if we did observe that $\theta = 1/16$ is in general slightly worse than the other values (in particular for the family $\mathcal{F}_{\mathrm{R}}$). The behaviour is different when considering the family $\mathcal{F}_{\mathrm{KR}}$. In this case the Hellinger risk decreases when θ increases so that $\theta = 1/16$ and

$\theta = 7/16$ are always respectively the worst and the best candidate. We also noticed some stability between $\theta = 3/8$ and $\theta = 7/16$.

6.3.3 Influence of V

The main question when considering VF type procedures is maybe "which V is optimal?" or, more generally, "what is the influence of V on the quality of the VF procedure?". According to our theoretical study in Section 6.2.5 the optimal value of V depends on the optimal value m^* of m. In the case of equal weights the best V seems to be an increasing function of the variance term in the risk of $\widehat{s}_{m^*}$. In the case of histograms, if the best one has many bins, one should take a large value of V and the same would hold for a kernel estimator with a small bandwidth. To understand what actually happens in practice, we study here how the risk of the chosen estimator behaves when V varies.

Since θ has little influence, we made the simulations with $\theta = 1/4$. We also implemented the calibrated procedures described in Section 6.3.1 to have a benchmark for the risk for families $\mathcal{F}_{\mathrm{R}}$ and $\mathcal{F}_{\mathrm{K}}$.

family	V	s_1	s_2	s_3	s_4	s_5	s_7	s_{12}	s_{13}	s_{22}	s_{23}	s_{24}
$\mathcal{F}_{\mathrm{R}}$	2	**2,9**	10,4	9,29	13,8	10,9	11,4	17,9	14,5	10,5	20,8	27,5
	5	4,31	9,9	8,75	12,7	10	10,6	17,3	**13,5**	9,56	18,4	25,2
	10	6,18	9,81	8,64	12,3	9,77	10,6	**17,2**	13,7	9,51	**17,8**	**24,8**
	20	9,39	**9,65**	**8,54**	**12,2**	**9,59**	**10,4**	17,3	14,1	**9,28**	17,9	**24,8**
	BR	2,20	9,94	9,27	12,98	10,53	11,14	17,85	14,63	10,37	17,98	25,15
$\mathcal{F}_{\mathrm{K}}$	2	15,4	29,9	5,67	5,1	**3,56**	4,26	28,5	20	3,96	10,6	18,1
	5	12,7	25,5	5,06	**4,95**	3,61	**3,98**	23,4	18,1	**3,86**	9,28	16,2
	10	12,4	24,3	**4,94**	5,01	3,96	4,04	21,8	17,7	3,91	9,08	15,8
	20	**12,2**	**23,5**	4,97	5,41	4,9	4,27	**20,9**	**17,6**	4,11	**9,05**	**15,7**
	UCV	15,86	22,20	5,57	6,16	3,74	4,10	18,80	17,16	3,88	9,52	15,91
$\mathcal{F}_{\mathrm{KR}}$	2	**2,88**	10,4	8,32	6,35	5,81	6,57	18,5	14,4	7,3	12,8	20
	5	4	9,91	7,86	**5,64**	**5,11**	**6,06**	17,7	**13,2**	**5,76**	9,66	16,7
	10	4,34	9,95	7,66	**5,64**	5,4	6,18	17,6	13,7	5,82	9,12	16
	20	4,34	**9,86**	**7,49**	5,91	5,81	6,5	**17,5**	14,5	5,88	**9,08**	**15,7**

Table 6.2: Hellinger risks multiplied by 1000 of the TVF procedure based on Birgé's test.

The empirical results summarized in Table 6.2 actually confirm what we derived above from our theoretical bound (6.26). The quality of the estimation increases with V when the true density is difficult to estimate which corresponds to an optimal estimator $\widehat{s}_{m^*}$ with a large value of m^* in (6.26). For a simple density like the uniform s_1 which is better estimated by an histogram with few bins, the best choice of V is 2 for the families $\mathcal{F}_R$ and $\mathcal{F}_{KR}$ which include histograms. On the contrary, when dealing with the family $\mathcal{F}_K$ for which s_1 is not easy to estimate, we need to use a larger value of V. A similar situation occurs with densities s_4, s_5, s_7 and s_{22} which appears to be

easily estimated by a kernel estimator with a large bandwidth but poorly by histograms. It seems that, apart from the exceptional situation of s_1, the best value of V is not 2 and the most significant gain appears between $V = 2$ and $V = 5$, then the quality sometimes keeps improving from $V = 5$ to $V = 20$, but with very little difference between $V = 10$ and $V = 20$.

The complexity of the TVF procedure is quite important in practice so that a large value of V should be avoided because of too large computation time. In particular the Leave-one-out ($V = n$) should be excluded since it is typically impossible to compute it in a reasonable amount of time. Of course, since the optimal value of V, as we have seen, depends of unknown properties of the methods with respect to the true density (like $D_{\mathcal{A}_{m^*}}$) it is impossible to define an optimal choice of V but, from our empirical study, we would rather recommend the user to process the TVF procedure with $V = 5$.

Interestingly, we also observe that when using the mixed collection $\mathcal{F}_{\mathrm{KR}}$ the TVF procedure shows a good adaptation behaviour since it selects the best family in all settings. For instance for s_5 it chooses a kernel estimator since these are better than histograms, whereas it selects an histogram for s_2 for the opposite reason.

6.3.4 Comparison with others VF

The goal of this section is to compare the TVF procedure to VF based one the unbiased risk estimation principle derived from a contrast function (that is LSVF and KLVF). It is hard to find a simple way to summarize this comparison since the best solution would be to present for each family, all risks for all densities and all procedures with different choices of V. In order to compare two VF procedures $\widetilde{t}_1$ and $\widetilde{t}_2$, we introduce the $\log_2$-ratio of their empirical risk, namely:

$$\overline{W}_s\left(\widetilde{t}_1, \widetilde{t}_2, \ell\right) = \log_2 \frac{\overline{R}_n\left(\widetilde{t}_1, s, \ell\right)}{\overline{R}_n\left(\widetilde{t}_2, s, \ell\right)} \ .$$

If one has $\overline{W}_s\left(\widetilde{t}_1, \widetilde{t}_2, \ell\right) = c$, for some constant c, it means that $\overline{R}_n\left(\widetilde{t}_1, s, \ell\right) = 2^c \times \overline{R}_n\left(\widetilde{t}_2, s, \ell\right)$. Hence, for a given density s, $\widetilde{t}_2$ is a better estimator than $\widetilde{t}_1$ if $c > 0$. In our empirical study, a selection procedure $\widetilde{t}_2$ is thus considered better in terms of risk than $\widetilde{t}_1$ for a given loss function ℓ if the values of $\overline{W}_s(\widetilde{t}_1, \widetilde{t}_2, \ell)$ are positive when the density s varies in $\mathcal{L}$. We illustrate our results showing boxplots of $\{\overline{W}_s(\widetilde{t}_1, \widetilde{t}_2, \ell),\ s \in \mathcal{L}\}$ with the discriminating value zero emphasized in red. We actually observed similar results and behaviours for all losses and all sample sizes but present here only the results for $\ell = h^2$ and $n = 500$ for the sake of simplicity. Figure 6.2 is built using $\widetilde{t}_1 = \widehat{s}_{\widehat{m}_{\mathrm{LSVF}}}$ (upper line) or $\widehat{s}_{\widehat{m}_{\mathrm{KLVF}}}$ (bottom line), and $\widetilde{t}_2 = \widehat{s}_{\widehat{m}_{\mathrm{TVF}}}$ with $\theta = 1/4$, $\ell = h^2$ and $n = 500$. Similar results and behaviors were observed for other sample sizes and losses.

In nearly all cases, the median and most of the distribution are positive, meaning that the TVF outperforms LSVF (average gain of about 20% for each family) and KLVF. For collection $\mathcal{F}_{\mathrm{K}}$ we observe that empirical risks are similar comparing to KLVF, the boxplot of $\overline{W}_s(\widehat{s}_{\widehat{m}_{\mathrm{KLVF}}}, \widehat{s}_{\widehat{m}_{\mathrm{TVF}}}, h^2)$ being concentrated around zero. There is a huge difference between TVF and KLVF procedures for families $\mathcal{F}_{\mathrm{R}}$ and $\mathcal{F}_{\mathrm{KR}}$ (average gain of respectively about 100% and 180%). For the uniform density estimated with regular histograms, the estimator derived from our procedure is worse since we found, for both classical VF, $\overline{W}_{s_1}(\widetilde{s}, \widehat{s}_{\widehat{m}_{\mathrm{TVF}}}, h^2) < 0$ (with an increasing difference with V for

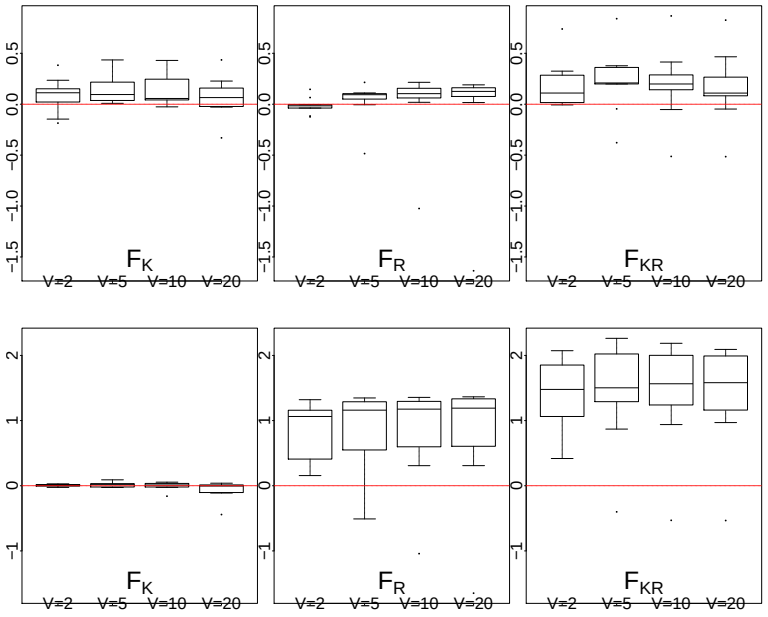

Figure 6.2: From left to right, the boxplot $\overline{W}_s(\widetilde{s}, \widehat{s}_{\widehat{m}_{\mathrm{TVF}}}, h^2)$, using families $\mathcal{F}_{\mathrm{K}}, \mathcal{F}_{\mathrm{R}}, \mathcal{F}_{\mathrm{KR}}$ (up for $\widetilde{s} = \widehat{s}_{\widehat{m}_{\mathrm{LSVF}}}$, down for $\widetilde{s} = \widehat{s}_{\widehat{m}_{\mathrm{KLVF}}}$). Each subfigure shows the boxplot for $V = 2, 5, 10$ and 20. The horizontal red dotted line provides the reference value 0.

$\mathcal{F}_{\mathrm{R}}$). Finally, let us notice that the difference between TVF and classical VF does not change much with V.

6.4 Our computational algorithm

For the practical computation of the TVF as well as any other VF procedure, we assume that $\mathcal{M}$ is finite with $|\mathcal{M}| = M$.

Let us compare the complexity of a classical V-fold against ours. Since for every VF method the construction of all partial estimators $(\widehat{s}_m^{(-j)})_{1 \leqslant j \leqslant V, 1 \leqslant m \leqslant M}$ is required, we only have to focus on the "validation part" which requires to compute all $\mathcal{D}_j^2(m)$ for $1 \leq j \leq V$ and $m \in \mathcal{M}$ and therefore perform all tests $\psi_{l,m}(\mathbf{X}_j)$ for $1 \leq j \leq V$ and $l, m \in \mathcal{M}$ with $l \neq m$. This means performing $V * M * (M-1)/2$ tests leading to a computational cost of order $O(V * M^2)$ that can then be prohibitive as compared to the one of either LSVF or KLVF which have a maximum complexity of order $O(V * M)$ (since in this case less than M calculations are needed for each split). For instance, a 10-fold with 100 different methods would require at most 1000 evaluations for a classical version of the VF whereas we would need the computation of 49500 tests for the TVF. It is already huge and does not even take into account the computation of the distances $h^2(\widehat{s}_l^{(-j)}, \widehat{s}_m^{(-j)})$, each one requiring the evaluation of an integral. Therefore a "naive" algorithm based on the computation of all the $V * M$ $\mathcal{D}_j^2(m)$ would be very slow.

Fortunately, there is a smarter way to determine which $\widehat{m}$ minimizes $\mathcal{D}'(\cdot)$ over $\mathcal{M}$. Our algorithm is inspired in some way by the one described in Section 3 of Magalhães & Rozenholc (2014). In order to explain how this "fast" algorithm works, it will be convenient to single an

element of $\mathcal{M}$, that we shall denote by "m_s", to serve as a starting point for our algorithm, therefore starting with the computation of $\mathcal{D}'(m_s)$. We shall store in R the minimal value of those $\mathcal{D}'(m)$ that have already been computed and in *opt* the corresponding optimal value of m starting with $opt = m_s$ and $R = \mathcal{D}'(m_s)$. We update them after each computation of a $\mathcal{D}'(m)$ such that $\mathcal{D}'(m) < R$, then setting $opt := m$ and $R := \mathcal{D}'(opt)$ so that R can only decrease during the computational procedure.

Let us first observe that minimizing our criterion $\overline{\mathcal{D}}(m)$ with respect to $m \in \mathcal{M}$ is actually equivalent to minimizing $\mathcal{D}'(m) = V\overline{\mathcal{D}}(m) = \sum_{j=1}^{V} \mathcal{D}_j^2(m)$. Since

$$\mathcal{D}_j^2(m) = \sup_{l \in \mathcal{M}_m} h^2\left(\widehat{s}_l^{(-j)}, \widehat{s}_m^{(-j)}\right) \mathbf{1}_{\left\{\psi_{l,m}(\mathbf{X})=l\right\}} \quad \text{with} \quad \mathcal{M}_m = \mathcal{M} \setminus \{m\} \ ,$$

one can compute it iteratively, starting with $\mathcal{L}_j(m) = 0$ and setting

$$\mathcal{L}_j(m) := \max\left(\mathcal{L}_j(m), h^2\left(\widehat{s}_l^{(-j)}, \widehat{s}_m^{(-j)}\right)\right) \quad \text{when} \quad \psi_{l,m}(\mathbf{X}_j) = l \quad \text{for} \quad l \in \mathcal{M}_m \ .$$

If $\psi_{l,m}(\mathbf{X}_j) = m$ we can instead update $\mathcal{L}_j(l)$ by $\mathcal{L}_j(l) := \max(\mathcal{L}_j(l), h^2(\widehat{s}_l^{(-j)}, \widehat{s}_m^{(-j)}))$ using the result of the test $\psi_{l,m}(\mathbf{X}_j)$ for the calculation of both $\mathcal{D}_j^2(m)$ and $\mathcal{D}_j^2(l)$. Our algorithm proceeds in this way, with a set of M V-dimensional vectors $\mathcal{L}_\cdot(m)$, $m \in \mathcal{M}$, initially set to zero. The updating procedure of $\mathcal{L}_j(m)$ stops when all updates, with $l \in \mathcal{M}_m$, have been done (which means that the present value of $\mathcal{L}_j(m)$ is $\mathcal{D}_j^2(m)$) and we finally set $\mathcal{D}'(m) = \sum_{j=1}^{V} \mathcal{L}_j(m)$.

We also use another trick in order to shorten our computations. Since $\mathcal{L}_j(m)$ can only increase during the updating procedure, $\sum_{j=1}^{V} \mathcal{L}_j(m)$ is, at any time, a lower bound for $\mathcal{D}'(m)$, whatever $m \in \mathcal{M}$. Therefore it is useless to go on with the computation of the vector $\mathcal{L}_\cdot(m)$ if $\sum_{j=1}^{V} \mathcal{L}_j(m) > R$ since then $\mathcal{D}'(m) \geq \sum_{j=1}^{V} \mathcal{L}_j(m)$ cannot minimize the function $\mathcal{D}'(\cdot)$ over $\mathcal{M}$. Taking this fact into account, we denote by $\mathcal{G} \subset \mathcal{M}$ the set of all methods which are potentially "better" than the current optimal one stored in *opt*. This means that we store in $\mathcal{G}$ all $m \in \mathcal{M}$ for which we do not yet know whether $\mathcal{D}'(m) < R$ or not and each time we find m such that $\sum_{j=1}^{V} \mathcal{L}_j(m) > R$, we remove it from $\mathcal{G}$. We also remove m from $\mathcal{G}$ once we have computed $\mathcal{D}'(m)$ with $m \in \mathcal{G}$ and then proceed with the computation of some new vector $\mathcal{L}_\cdot(l)$ for $l \in \mathcal{G}$ until $\mathcal{G}$ is empty and the algorithm stops with the final value $\widehat{m} = opt$.

Some important remarks

- The algorithm is designed to work with a test procedure ψ which satisfies Assumption (TEST) or, more generally Assumption (TEST'), like the procedures based on the statistics (6.9) or (6.18).

- It is important to notice that, at any step, we cannot "delete" once and for all the methods which do not belong to the set $\mathcal{G}$! Even if we do not compute the value of $\mathcal{D}'$ for these methods, we still need to test them against the remaining methods in $\mathcal{G}$.

- We hoped that by starting from a good estimator, only a few methods would be in the first set $\mathcal{G}$, resulting in just a few tests. In the simulations we always started from $m_s = \widehat{m}_{\mathrm{LSVF}}$. If $\mathcal{D}'(\widehat{m}_{\mathrm{LSVF}}) = 0$ at the first step the algorithm stops immediately and the chosen method

is $\widehat{m} = \widehat{m}_{\mathrm{LSVF}}$. In this special case, the complexity of our algorithm is the same as the one of the classical approach.

- Clearly, the choice of m at line 17 of the algorithm, as well as the choice of the starting method, have no influence on the final estimator. To avoid a quadratic complexity, we need to ensure that we don't "jump" to the worst method inside the set $\mathcal{G}$ at each iteration. In our simulations, we chose to jump to the statistical method $k \in \mathcal{G}$ with the lowest temporary criterion among the methods in $\mathcal{G}$, that is $k = \mathrm{argmin}_{l \in \mathcal{G}} \sum_{j=1}^{V} \mathcal{L}_j(l)$. We also tried two alternative options: jumping to $k = \mathrm{argmax}_{l \in \mathcal{G}} \sum_{j=1}^{V} \mathcal{L}_j(l)$ and to the most chosen statistical method k in $\mathcal{G}$ against m. Both options lead of course to the same final estimator but were definitely slower.

Algorithm 3: Selection of the TVF estimator

Initialization:
Set $\mathcal{G} = \mathcal{M}_{m_s}$ and $opt = m_s$
for $(l \in \mathcal{M})$ **do**
  for $(j = 1, \ldots, V)$ **do**
    $\mathcal{L}_j(l) = 0$
  end
end

1st step:
for $(l \in \mathcal{G})$ **do**
  Compute $\psi_{m_s,l}(\mathbf{X}_j)$
  if $(\psi_{m_s,l}(\mathbf{X}_j) = m_s)$ **then**
    $\mathcal{L}_j(l) = h^2(\widehat{s}_l^{(-j)}, \widehat{s}_{m_s}^{(-j)})$
  else
    $\mathcal{L}_j(m_s) = \max(\mathcal{L}_j(m_s), h^2(\widehat{s}_l^{(-j)}, \widehat{s}_{m_s}^{(-j)}))$
  end
end
Set $R = \sum_{j=1}^{V} \mathcal{L}_j(m_s)$ and $\mathcal{G} = \mathcal{G} \setminus \{l \in \mathcal{G} : \sum_{j=1}^{V} \mathcal{L}_j(l) > R\}$

Next steps:
while $(|\mathcal{G}| > 0)$ **do**
  Choose $m \in \mathcal{G}$ and set $\mathcal{G} = \mathcal{G} \setminus \{m\}$
  for $(j = 1, \ldots, V)$ **do**
    for $(l \in \mathcal{M}_m)$ **do**
      Compute $\psi_{m,l}(\mathbf{X}_j)$ `// if it has not been done yet`
      if $(\psi_{m,l}(\mathbf{X}_j) = m$ *and* $l \in \mathcal{G})$ **then**
        $\mathcal{L}_j(l) = \max(\mathcal{L}_j(l), h^2(\widehat{s}_l^{(-j)}, \widehat{s}_m^{(-j)}))$
        if $(\sum_{i=1}^{V} \mathcal{L}_i(l) > R)$ **then**
          $\mathcal{G} = \mathcal{G} \setminus \{l\}$
        end
      end
      if $(\psi_{m,l}(\mathbf{X}_j) = l)$ **then**
        $\mathcal{L}_j(m) = \max(\mathcal{L}_j(m), h^2(\widehat{s}_l^{(-j)}, \widehat{s}_m^{(-j)}))$
        if $(\sum_{i=1}^{V} \mathcal{L}_i(m) > R)$ **then**
          break `// quit the two ''for'' loops`
        end
      end
    end
  end
  if $(\sum_{j=1}^{V} \mathcal{L}_j(m) < R)$ **then**
    Set $opt = m$, $R = \sum_{j=1}^{V} \mathcal{L}_j(m)$ and $\mathcal{G} = \mathcal{G} \setminus \{l \in \mathcal{G} : \sum_{j=1}^{V} \mathcal{L}_j(l) > R\}$
  end
end
Return opt

6.5 Supplementary material

We provide here additional simulations concerning the test designed by Baraud (2011) which is given in (6.18). As in Section 6.3, we study the influence of V and we compare the TVF based

on this test against classical VF procedures. Moreover, we compare the squared Hellinger risk of both our TVF procedures.

Influence of V

family	V	s_1	s_2	s_3	s_4	s_5	s_7	s_{12}	s_{13}	s_{22}	s_{23}	s_{24}
$\mathcal{F}_{\mathrm{R}}$	2	**2,89**	9,97	9,07	13,2	10,5	11	17,5	14,7	10,3	19,9	26,9
	5	4,33	9,68	8,61	12,4	9,87	10,4	17,1	**13,4**	9,37	17,8	24,7
	10	6,13	9,65	8,56	12,1	9,65	10,4	17	13,7	9,36	17,5	**24,3**
	20	9,28	**9,47**	**8,4**	**12**	**9,36**	**10,3**	**16,9**	14,2	**9,17**	**17,4**	24,6
	BR	2,20	9,94	9,27	12,98	10,53	11,14	17,85	14,63	10,37	17,98	25,15
$\mathcal{F}_{\mathrm{K}}$	2	15,6	29,4	5,69	5,07	**3,55**	4,24	27,2	20	3,97	10,3	18
	5	13,2	25,7	5,1	**4,94**	3,58	**3,97**	23	18,1	**3,85**	9,18	16,2
	10	12,9	24,8	5	5,02	3,86	4,01	22,2	17,7	3,87	9,04	**15,8**
	20	**12,7**	**24,4**	**4,98**	5,28	4,54	4,1	**21,6**	**17,6**	3,98	**8,98**	**15,8**
	UCV	15,86	22,20	5,57	6,16	3,74	4,10	18,80	17,16	3,88	9,52	15,91
$\mathcal{F}_{\mathrm{KR}}$	2	**2,87**	10	7,47	5,88	5,04	5,6	18,9	14,7	6,38	11,6	19,1
	5	3,68	**9,77**	6,81	**5,48**	**4,64**	**5,19**	17,7	**13,3**	**5,01**	9,3	16,4
	10	3,58	9,84	6,71	5,53	4,99	5,26	**17,6**	13,7	5,11	9,04	15,9
	20	3,79	9,84	**6,45**	5,65	5,31	5,83	**17,6**	14,6	5,22	**9,01**	**15,7**

Table 6.3: Hellinger risks multiplied by 1000 for the TVF procedure based on Baraud's test.

Comparison with others VF

Influence of the test on the TVF

We compare here the performances of the best TVF procedure (among the five values of θ described above) derived from Birgé's test (6.9) against the one deduced from Baraud's test (6.18) (denoted $\widehat{s}_{\widehat{m}_{\mathrm{TVF}}}$). We show the conclusion of our study for the families $\mathcal{F}_{\mathrm{R}}$, $\mathcal{F}_{\mathrm{K}}$ and $\mathcal{F}_{\mathrm{KR}}$, $n = 500$, $V = 2, 5, 10$ and 20. The results were very similar for other values of n. For the sake of clarity and to emphasize the similarity of both procedures in terms of Hellinger risk, we present for each family, for each V, the supremum and the infimum over $\mathcal{L}$ of the ratio

$$\Upsilon(s) = \left\{ \inf_{\theta\in\Theta} \overline{R}_n\left(\widehat{s}_{\widehat{m}(\theta)}, s, h^2\right) \Big/ \overline{R}_n\left(\widehat{s}_{\widehat{m}_{\mathrm{TVF}}}, s, h^2\right) \right\} .$$

If $\inf_{s\in\mathcal{L}} \Upsilon(s) \geq 1$ the TVF using Baraud's test behaves in a better way than the one using Birgé's test for all densities in $\mathcal{L}$ while if $\sup_{s\in\mathcal{L}} \Upsilon(s) \leq 1$ the opposite holds. The closer the two values, the more similar the quality of both procedures.

We see from this table that Baraud's and Birgé's test are very similar to proceed the TVF procedure for families $\mathcal{F}_{\mathrm{R}}$ and $\mathcal{F}_{\mathrm{K}}$. There is indeed no noticeable difference for these families, the

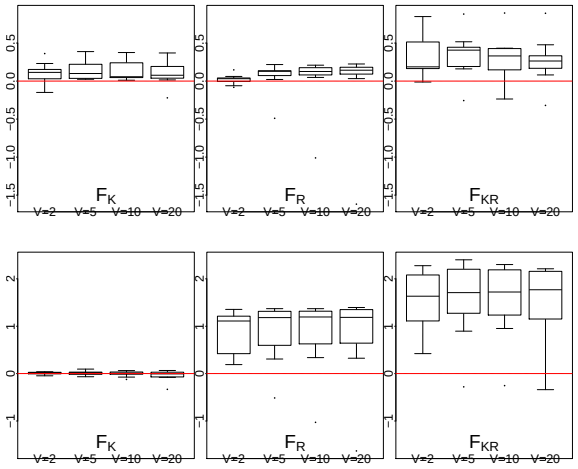

Figure 6.3: From left to right, the boxplot $\overline{W}_s\left(\widetilde{s}, \widetilde{s}_{\mathrm{TVF}}, h^2\right)$, using families $\mathcal{F}_{\mathrm{K}}, \mathcal{F}_{\mathrm{R}}, \mathcal{F}_{\mathrm{KR}}$ (up for $\widetilde{s} = \widehat{s}_{\widehat{m}_{\mathrm{LSVF}}}$, down for $\widetilde{s} = \widehat{s}_{\widehat{m}_{\mathrm{KLVF}}}$). Each subfigure shows the boxplot for $V = 2, 5, 10$ and 20. The horizontal red dotted line provides the reference value 0.

family	$\Upsilon(s)$	$V=2$	$V=5$	$V=10$	$V=20$
$\mathcal{F}_{\mathrm{R}}$	$\sup_s$	103,68	102,59	101,72	102,27
	$\inf_s$	98,16	100,07	99,59	99,13
$\mathcal{F}_{\mathrm{K}}$	$\sup_s$	102,78	100,80	100,92	105,10
	$\inf_s$	99,58	98,72	97,45	96,13
$\mathcal{F}_{\mathrm{KR}}$	$\sup_s$	116,71	115,80	116,79	116,73
	$\inf_s$	96,70	98,84	99,08	99,30

Table 6.4: Supremum and infimum of the ratio multiplied by 100, see the text.

largest gain (for a density in $\mathcal{L}$) being of 5% only. The procedure based on Baraud's test becomes much better for the family $\mathcal{F}_{\mathrm{KR}}$. We observe indeed that a potential gain of 15% appears (since the $\sup_s$ is close to 115%) while the loss is negligible (since the $\inf_s$ is close to 99%). Moreover, the ratios are quite similar when V increases. Finally, let us recall that the TVF procedure based on (6.9) is less time-consuming since it requires to compute only one integral instead of two for (6.18).

Ne pas conclure !

En guise de conclusion, nous proposons une nouvelle procédure de sélection d'estimateurs qui repose uniquement sur les idées et heuristiques présentées dans l'Introduction. Nous proposons quelques pistes pour prouver sa pertinence du point de vue théorique et étudions sa qualité sur des données simulées en la comparant aux procédures définies dans cette thèse. Nous formulons également quelques questions ouvertes qui forment un panorama non-exhaustif de perspectives de recherche qui peuvent être menées à partir des différents travaux que contiennent ce manuscrit. Aussi, ces deux sous-sections nous invitent à ne pas conclure cette recherche puisque les progrès à réaliser, notamment pour avoir une réelle compréhension théorique de ce qui se passe en pratique, semblent nettement plus importants que le travail qui a déjà pu être effectué.

Sommaire

7.1 Sélection d'estimateurs par test rééchantillonné

Nous avons présenté dans ce travail différentes procédures VF reposant sur deux pertes distinctes qui ne peuvent être comparées qu'en pratique. La première remarque générale sur cette thèse est qu'elle met en lumière, si cela était encore nécessaire, un écart entre la théorie et la pratique qu'il nous semble important de combler. Il ressort en effet des différentes études empiriques que le critère pénalisé, théoriquement bien compris et pour lequel nous savons que la situation s'améliore quand V augmente, est sensiblement moins bon (du moins tant qu'on ne sur-pénalise pas) que la plupart des procédures de la Partie III. A l'inverse, la procédure $\mathcal{C}_V^{\mathrm{TVF}}$ introduite au Chapitre 6, pour laquelle nous n'avons pas pu prouver une inégalité oracle fine, semble être la meilleure procédure à condition de ne pas être regardant sur le coût algorithmique ! Effectivement, bien que nous ayons réduit le coût calculatoire grâce à notre algorithme, le TVF reste une procédure plus gourmande que les VF classiques[(5)]. Alors que la rédaction de ce manuscrit avançait, il est apparu qu'on pouvait combiner certaines idées phares pour construire une nouvelle procédure qui garderait la qualité du TVF tout en étant significativement plus rapide en temps de calcul. Celle-ci est une alternative au TVF et à la procédure de sélection d'estimateurs proposée dans Baraud (2011).

7.1.1 Combiner les avantages ?

Nous allons développer notre heuristique en utilisant successivement les points suivants présentés précédemment dans la thèse.

- Existence d'un test idéal dans la procédure de sélection d'estimateurs par T-estimation.
- Heuristique d'Efron et bootstrap avec poids V-fold.
- Algorithme pour le calcul d'un T-estimateur à partir d'une famille finie.

Nous nous plaçons dans le cadre du Chapitre 6, avec $\ell = h^2$ la perte de Hellinger et $\mathcal{S}$ l'ensemble des densités par rapport à la mesure μ sur Ξ. Démarrons de la collection de méthodes d'estimation à noyau $\{\mathcal{A}_m, m \in \mathcal{M}\}$ qui fournit la collection d'estimateurs linéaires $\{\widehat{s}_m, m \in \mathcal{M}\}$, où pour tout $m \in \mathcal{M}$,

$$\widehat{s}_m(x) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{K}_m(x, X_i) \qquad \text{pour tout } x \in \Xi$$

avec

$$\mathcal{K}_m(x,y) \geq 0 \quad \forall x, y \in \Xi \qquad \text{et} \qquad \int_\Xi \mathcal{K}_m(x,y) d\mu(x) = 1 \quad \forall y \in \Xi \ .$$

Nous avons vu à la Section 1.2.2 que le test idéal pour choisir entre deux estimateurs $\widehat{s}_l$ et $\widehat{s}_m$, pouvait s'écrire, en posant $\widehat{r} = (\widehat{s}_l + \widehat{s}_m)/2$,

$$T_{\widehat{s}_l, \widehat{s}_m}(P) = \frac{1}{2}\left(P\left(\frac{\sqrt{\widehat{s}_l} - \sqrt{\widehat{s}_m}}{\sqrt{\widehat{r}}} \right) + \int \left(\sqrt{\widehat{s}_l(x)} - \sqrt{\widehat{s}_m(x)} \right) \sqrt{\widehat{r}(x)} d\mu(x) \right) \ .$$

[(5)] Si on prend en compte cette contrainte, le conseil que l'on donnerait au praticien serait d'effectuer un 5-fold légèrement sur-pénalisé, voir Chapitre 3.

Baraud (2011, Corollary 1) montre en effet l'implication suivante

$$T_{\widehat{s}_l,\widehat{s}_m}(P) \geqslant 0 \qquad \Longrightarrow \qquad h^2(s,\widehat{s}_l) \leqslant \frac{\sqrt{2}+1}{\sqrt{2}-1}\, h^2(s,\widehat{s}_m) \ .$$

Le critère "idéal" donné par la T-estimation s'écrit alors comme une fonctionnelle $\mathcal{L}(P,P_n)$, qui dépend de P et P_n, donnée par

$$\sup_{l\in\mathcal{R}_m} h^2\left(\widehat{s}_l,\widehat{s}_m\right) \quad \text{où} \quad \mathcal{R}_m = \left\{ l\in\mathcal{M},\, l\neq m \mid T_{\widehat{s}_l,\widehat{s}_m}(P) \geqslant 0 \right\} \ .$$

Pour rappel, l'idée de la procédure $\mathcal{C}_V^{\mathrm{TVF}}$ est d'approcher ce critère $\mathcal{L}(P,P_n)$ en moyennant sur V découpages les critères $\mathcal{L}(P_n^{(j)},P_n^{(-j)})$ selon l'heuristique de la validation croisée V-fold présentée en (1.28). Comme expliqué au Chapitre 6, le coût algorithmique de cette moyenne est malheureusement important puisque nous sommes obligés de calculer tous les critères $\mathcal{D}_j$ (autrement dit de tester pour chaque nouveau j, tous les estimateurs entre eux) pour effectuer la sélection de l'estimateur final. Même si l'algorithme présenté en Section 6.4 échappe à ce coût exorbitant, on comprend qu'il serait préférable pour le praticien d'avoir une procédure où on évite cette *moyenne sur le critère*. Étant donné que le critère est "idéal" parce que le test sur lequel il repose l'est, nous allons estimer celui-ci uniquement. Ainsi, nous n'allons plus moyenner V critères pour approcher le critère idéal (ce qui nous oblige à tester V fois les méthodes entre elles) mais faire la *moyenne sur les tests* pour approcher le test idéal (ce qui permet de tester une seule fois les méthodes entre elles) !

De la même façon qu'il faut éviter de réutiliser les mêmes données pour construire et évaluer la qualité d'un estimateur, on peut penser qu'il peut être catastrophique d'estimer $T_{\widehat{s}_l,\widehat{s}_m}(P)$ par $T_{\widehat{s}_l,\widehat{s}_m}(P_n)$. Baraud (2011) propose ainsi d'introduire une pénalité pen, définie sur les modèles auxquels appartiennent les estimateurs, menant au test suivant :

$$\mathbb{T}_{\widehat{s}_l,\widehat{s}_m}(P_n) = T_{\widehat{s}_l,\widehat{s}_m}(P_n) + \mathrm{pen}\left(\widehat{s}_l\right) - \mathrm{pen}\left(\widehat{s}_m\right) \ . \tag{7.27}$$

Pour éviter ce problème, nous allons mettre à profit les heuristiques de rééchantillonnage et proposer deux manières d'estimer le test idéal. D'abord, en appliquant directement l'heuristique (1.28) sur le test, ensuite en utilisant l'heuristique d'Efron. Étant donné que celle-ci ne fournit pas de bonnes performances au premier ordre[6] nous allons l'utiliser au second ordre (Efron, 1983, Section 8).

7.1.2 Test rééchantillonné

Nous effectuons la partition de $\mathbf{X} = \{X_1,\ldots,X_n\}$ en V sous-échantillons de même taille, et utilisons les mêmes notations qu'à la Section 1.2.3. En reprenant l'idée du bootstrap à poids nous estimons

$$T_{\widehat{s}_l,\widehat{s}_m}(P) - T_{\widehat{s}_l,\widehat{s}_m}(P_n) = \frac{1}{2}\left((P-P_n)\left(\frac{\sqrt{\widehat{s}_l}-\sqrt{\widehat{s}_m}}{\sqrt{\widehat{r}}} \right) \right) \ ,$$

[6] C'est la raison pour laquelle nous estimons la pénalité idéale $(P-P_n)\gamma(\widehat{s}_m)$, plutôt que d'estimer le critère idéal $P\gamma(\widehat{s}_m)$ directement par $\mathbb{E}_W\left[P_n\gamma(\widehat{s}_m^W)\right]$.

par

$$\frac{1}{2\mathbb{E}\left[(W_1-1)^2\right]}\,\mathbb{E}_W\left[\left(P_n-P_n^W\right)\left(\frac{\sqrt{\widehat{s}_l^W}-\sqrt{\widehat{s}_m^W}}{\sqrt{r^W}}\right)\right] \quad \text{où} \quad r^W=\frac{\widehat{s}_l^W+\widehat{s}_m^W}{2}\ .$$

Nous considérons, comme précédemment, les poids V-fold $W^{(\mathrm{VF})(7)}$ qui mènent à l'estimateur suivant

$$T^W(l,m):=\frac{(V-1)}{2V}\sum_{j=1}^{V}\left(P_n-P_n^{(-j)}\right)\left(\frac{\sqrt{\widehat{s}_l^{(-j)}}-\sqrt{\widehat{s}_m^{(-j)}}}{\sqrt{\widehat{r}^{(-j)}}}\right) \quad \text{où} \quad \widehat{r}^{(-j)}=\frac{\widehat{s}_l^{(-j)}+\widehat{s}_m^{(-j)}}{2}\ .$$

En utilisant le fait que pour une partition régulière de $\mathbf{X}$ on a $(V-1)(P_n-P_n^{(-j)})=P_n^{(j)}-P_n$ pour tout $j\in[V]$, on trouve

$$T^W(l,m)=\frac{1}{2V}\sum_{j=1}^{V}\left(P_n^{(j)}-P_n\right)\left(\frac{\sqrt{\widehat{s}_l^{(-j)}}-\sqrt{\widehat{s}_m^{(-j)}}}{\sqrt{\widehat{r}^{(-j)}}}\right)\ .$$

Ainsi nous proposons les estimateurs suivants du test idéal $T_{\widehat{s}_l,\widehat{s}_m}(P)$.

- Le *test V-fold*
$$\mathcal{T}_{l,m}^{(VF)}(P_n):=\frac{1}{V}\sum_{j=1}^{V}T_{\widehat{s}_l^{(-j)},\widehat{s}_m^{(-j)}}\left(\mathbf{X}^{(B_j)}\right)\ .$$

- Pour une constante $C>0$, le *test Efron avec poids V-fold*

$$\begin{aligned}&\mathcal{T}_{l,m}^{V,C}(P_n):=T_{\widehat{s}_l,\widehat{s}_m}(P_n)+CT^W(l,m)\\&=\frac{1}{2}\left(\rho\left(\widehat{s}_l,\widehat{r}\right)-\rho\left(\widehat{s}_m,\widehat{r}\right)+P_n\left(\frac{\sqrt{\widehat{s}_l}-\sqrt{\widehat{s}_m}}{\sqrt{\widehat{r}}}\right)+\frac{C}{V}\sum_{j=1}^{V}\left(P_n^{(j)}-P_n\right)\frac{\sqrt{\widehat{s}_l^{(-j)}}-\sqrt{\widehat{s}_m^{(-j)}}}{\sqrt{\widehat{r}^{(-j)}}}\right)\ .\end{aligned}$$

  En considérant ce dernier test, on effectue la procédure de sélection d'estimateurs de Baraud (2011), en remplaçant la quantité $\mathrm{pen}(\widehat{s}_l)-\mathrm{pen}(\widehat{s}_m)$ dans (7.27) par $T^W(l,m)$.

La *procédure avec test rééchantillonné* se définit naturellement par

$$\mathcal{C}^{\mathrm{res.test}}(m):=\sup_{l\in\mathcal{R}_m}h^2(\widehat{s}_l,\widehat{s}_m) \qquad \text{avec} \quad \mathcal{R}_m=\left\{l\in\mathcal{M},\ l\neq m\mid\mathcal{T}_{l,m}^{\mathrm{res}}(P_n)\geqslant 0\right\}\ , \quad (7.28)$$

où le test $\mathcal{T}_{l,m}^{\mathrm{res}}$ est donné par $\mathcal{T}_{l,m}^{(VF)}$ ou $\mathcal{T}_{l,m}^{V,C}$. L'estimateur final est naturellement $\widehat{s}_{\widehat{m}(\mathcal{C}^{\mathrm{res.test}})}$. On peut noter que nous profitons à nouveau du fait que nous disposons de méthodes d'estimation et non pas d'estimateurs figés puisque pour tester deux estimateurs, nous utilisons les estimateurs

(7) défini pour tout i par $W_i^{(\mathrm{VF})}=\frac{V}{V-1}\mathbf{1}_{i\notin B_J}$, où $J\sim\mathcal{U}([V])$ est indépendant des données de sorte que $\mathbb{E}\left[(W_1-1)^2\right]=(V-1)^{-1}$ et $P_n^W=P_n^{(-J)}$.

construits successivement avec un nouveau sous-échantillon. Cette stratégie peut se voir comme une procédure de T-estimation avec "test V-fold", ou comme une nouvelle procédure de sélection d'estimateurs avec test rééchantillonné.

Premiers pas théoriques. Il nous semble que l'analyse des deux procédures commence de la même façon que pour Baraud. Pour $m \in \mathcal{M}$ et $\kappa > 0$ fixés, on a

$$\begin{aligned} h^2\left(s, \widehat{s}_{\widehat{m}}\right) &\leqslant (1+\kappa)\, h^2\left(s, \widehat{s}_m\right) + \left(1+\kappa^{-1}\right) h^2\left(\widehat{s}_m, \widehat{s}_{\widehat{m}}\right) \\ &\leqslant (1+\kappa)\, h^2\left(s, \widehat{s}_m\right) + \left(1+\kappa^{-1}\right)\left(\mathcal{C}^{\mathrm{res.test}}(\widehat{m}) \vee \mathcal{C}^{\mathrm{res.test}}(m)\right) \\ &\leqslant (1+\kappa)\, h^2\left(s, \widehat{s}_m\right) + \left(1+\kappa^{-1}\right) \mathcal{C}^{\mathrm{res.test}}(m) \ . \end{aligned}$$

Pour tout $l, m \in \mathcal{M}$, nous définissons le processus qui mesure l'erreur que nous faisons en utilisant un test rééchantillonné plutôt que le test idéal $\mathcal{Z}(\widehat{s}_l, \widehat{s}_m)$ par

$$\mathcal{Z}(\widehat{s}_l, \widehat{s}_m) = \mathcal{T}^{\mathrm{res}}_{l,m}(P_n) - T_{l,m}(P) \ .$$

On trouve alors pour tout $l \in \mathcal{R}_m$ (c'est-à-dire pour lequel $\mathcal{T}^{\mathrm{res}}_{l,m}(P_n) \geqslant 0$)

$$\begin{aligned} h^2(s, \widehat{s}_l) - h^2(s, \widehat{s}_m) &= \rho(s, \widehat{s}_m) - \rho(s, \widehat{s}_l) \\ &= T_{l,m}(P) + \rho(s, \widehat{s}_m) - \rho(s, \widehat{s}_l) - \mathcal{T}^{\mathrm{res}}_{l,m}(P_n) + \mathcal{Z}(\widehat{s}_l, \widehat{s}_m) \\ &\leqslant \frac{1}{\sqrt{2}}\left(h^2(s, \widehat{s}_l) + h^2(s, \widehat{s}_m)\right) + \mathcal{Z}(\widehat{s}_l, \widehat{s}_m) \ . \end{aligned}$$

Par conséquent

$$\sup_{l \in \mathcal{R}_m} h^2(\widehat{s}_l, s) \leqslant \left(\frac{\sqrt{2}+1}{\sqrt{2}-1}\right) h^2(s, \widehat{s}_m) + \left(\frac{\sqrt{2}}{\sqrt{2}-1}\right) \sup_{l \in \mathcal{R}_m} \mathcal{Z}(\widehat{s}_l, \widehat{s}_m) \ . \tag{7.29}$$

On en déduit pour tout $\alpha > 0$

$$\begin{aligned} \mathcal{C}^{\mathrm{res.test}}(m) = \sup_{l \in \mathcal{R}_m} h^2(\widehat{s}_l, \widehat{s}_m) &\leqslant (1+\alpha) h^2(s, \widehat{s}_m) + \left(1+\alpha^{-1}\right) \sup_{l \in \mathcal{R}_m} h^2(\widehat{s}_l, s) \\ &\leqslant (1+\alpha)\left(1 + \alpha^{-1}\frac{\sqrt{2}+1}{\sqrt{2}-1}\right) h^2(s, \widehat{s}_m) + \left(1+\alpha^{-1}\right)\frac{\sqrt{2}}{\sqrt{2}-1} \sup_{l \in \mathcal{R}_m} \mathcal{Z}(\widehat{s}_l, \widehat{s}_m) \ . \end{aligned}$$

A partir de ces calculs, il apparaît clairement qu'il est nécessaire de contrôler le processus $\mathcal{Z}(\widehat{s}_l, \widehat{s}_m)$ pour tout $l, m \in \mathcal{M}$. Il nous semble que les outils théoriques développés par Baraud (2011); Baraud *et al.* (2014) pourraient nous aider en ce sens et mener à un résultat similaire à celui obtenu pour la procédure $\mathcal{C}^{\mathrm{TVF}}_V$ (voir Théorème 6.1). Mais ceux-ci risquent de ne pas suffire pour obtenir une compréhension fine du rôle de V dans la procédure.

7.1.3 Comparaison empirique avec les autres VF

Nous proposons de l'étudier avec les deux tests et de la comparer empiriquement à celles étudiées dans cette thèse. Nous notons respectivement $\mathcal{C}^{\mathrm{res.test}}_V$ si $\mathcal{T}^{\mathrm{res}}_{l,m} = \mathcal{T}^{(VF)}_{l,m}$ et $\mathcal{C}^{\mathrm{res.test}}_{V,C}$ si $\mathcal{T}^{\mathrm{res}}_{l,m} = \mathcal{T}^{V,C}_{l,m}$.

Nous avons effectué le même type de simulations que dans les Sections 3.6 et 6.3 avec $N = 100$ échantillons de taille $n = 500$. Les deux familles[8] de méthode d'estimation considérées proviennent de noyaux d'approximation Gaussien (notée $\mathcal{F}_{\mathrm{K}}$) et d'histogrammes réguliers (notée $\mathcal{F}_{\mathrm{R}}$) de sorte que tous les estimateurs sont effectivement des densités. Nous avons calculé pour 12 densités le risque Hellinger, $\mathbb{L}_1$ et $\mathbb{L}_2$ des procédures $\mathcal{C}_V^{\mathrm{LSVF}}$, $\mathcal{C}_{V,F}^{\mathrm{pen}}$, $\mathcal{C}_V^{\mathrm{TVF},Bar}$ (avec le test de Baraud, donné par (1.43)), $\mathcal{C}_V^{\mathrm{TVF},Bir}$ (avec le test de Birgé, donné par (1.42)), $\mathcal{C}_V^{\mathrm{res.test}}$ et $\mathcal{C}_{V,C}^{\mathrm{res.test}}$ pour $V = 2, 5, 10, 20$ et $C = F = 1, 5/4, 3/2, 7/4, 2$. Nous présentons d'abord un aperçu du risque Hellinger pour $C = F = 1$ au travers de 5 densités.

		Chapitre 3 et Chapitre 4		Chapitre 6		Conclusion	
densité	V	$\mathcal{C}_V^{\mathrm{LSVF}}$	$\mathcal{C}_{V,1}^{\mathrm{pen}}$	$\mathcal{C}_V^{\mathrm{TVF},Bir}$	$\mathcal{C}_V^{\mathrm{TVF},Bar}$	$\mathcal{C}_V^{\mathrm{res.test}}$	$\mathcal{C}_{V,1}^{\mathrm{res.test}}$
s_3	2	9,37	10,73	9,21	9,10	**8,97**	9,77
	5	9,13	9,82	**8,75**	8,89	8,97	9,23
	10	9,27	9,33	8,82	**8,80**	9,10	9,10
	20	9,06	9,14	**8,48**	8,58	8,97	8,95
s_{11}	2	9,20	11,84	9,39	9,06	**8,94**	9,75
	5	9,17	9,71	8,35	**8,16**	8,74	9,02
	10	9,05	9,58	8,29	**8,27**	8,76	8,79
	20	9,28	9,30	8,33	**8,04**	8,79	8,87
s_{13}	2	15,36	18,71	14,44	15,15	14,90	**13,67**
	5	15,76	16,35	**13,18**	13,23	13,37	13,21
	10	15,87	16,42	13,42	13,58	**12,96**	13,04
	20	16,10	16,45	14,16	13,81	**12,85**	**12,85**
s_{16}	2	8,25	9,18	7,64	7,65	**7,25**	7,48
	5	7,63	7,63	7,33	**7,20**	7,36	7,50
	10	7,58	7,88	**7,20**	7,28	7,30	7,44
	20	7,51	7,97	7,43	**7,40**	7,41	7,50
s_{23}	2	19,19	21,21	21,07	19,67	**19,15**	19,27
	5	19,76	20,92	18,44	**18,01**	18,73	18,77
	10	19,99	20,52	18,01	**17,80**	18,57	18,87
	20	20,11	20,23	18,17	**17,57**	18,83	18,88

TABLE 7.5 – Risques Hellinger multipliés par 1000 pour la famille $\mathcal{F}_{\mathrm{R}}$.

Premièrement, on observe sur les Tables 7.5 et 7.6 que les quatre dernières procédures sont plus précises que les deux premières qui sont construites pour la perte des moindres carrés. Ensuite, on voit que les deux procédures introduites dans ce chapitre sont aussi bonnes voire meilleures que les procédures $\mathcal{C}_V^{\mathrm{TVF}}$. Pour celles-ci, le risque diminue dans presque toutes les configurations quand V augmente[9]. Concernant la procédure $\mathcal{C}_{V,C}^{\mathrm{res.test}}$, nous avons remarqué que le meilleur choix pour la constante C est 1, ce qui va à l'inverse de ce qu'on observe pour la procédure $\mathcal{C}_{V,F}^{\mathrm{pen}}$ (pour laquelle il est clairement préférable de sur-pénaliser avec une constante $F = 3/2$ ou 2 par exemple). Lorsqu'on s'autorise de prendre cette dernière procédure pour le meilleur choix de F, on se rapproche des risques des procédures qui reposent sur les tests robustes.

[8] Ces deux familles sont définies dans la Section 3.6.

[9] Il arrive que l'on observe une certaine stabilité ou une légère augmentation qui peut être due au faible nombre de pseudo-échantillons générés et à l'erreur de précision liée aux calculs d'intégrale.

		Chapitre 3 et Chapitre 4		Chapitre 6		Conclusion	
densité	V	$\mathcal{C}_V^{\mathrm{LSVF}}$	$\mathcal{C}_{V,1}^{\mathrm{pen}}$	$\mathcal{C}_V^{\mathrm{TVF},Bir}$	$\mathcal{C}_V^{\mathrm{TVF},Bar}$	$\mathcal{C}_V^{\mathrm{res.test}}$	$\mathcal{C}_{V,1}^{\mathrm{res.test}}$
s_3	2	6,86	6,69	5,65	5,75	5,64	**5,13**
	5	6,23	6,07	4,95	5,01	5,07	**4,93**
	10	5,90	5,84	5,00	5,02	**4,92**	**4,92**
	20	5,80	5,76	5,02	4,99	**4,86**	4,87
s_{11}	2	3,36	4,04	2,90	**2,89**	2,92	3,16
	5	3,07	3,24	**2,76**	2,81	2,77	2,89
	10	3,05	3,17	2,99	2,99	**2,80**	2,84
	20	3,04	3,05	3,41	3,22	**2,76**	2,77
s_{13}	2	18,06	21,00	20,11	20,07	19,89	**17,99**
	5	18,56	20,04	18,21	18,25	18,18	**17,69**
	10	18,59	19,73	17,74	17,86	17,85	**17,63**
	20	18,68	19,56	17,76	17,68	17,69	**17,62**
s_{16}	2	4,52	4,21	3,65	3,65	3,48	**3,30**
	5	3,84	3,71	3,16	3,18	3,14	**3,13**
	10	3,84	3,75	3,11	3,22	**3,07**	**3,07**
	20	3,79	3,78	3,21	3,27	**3,03**	3,04
s_{23}	2	9,41	10,47	10,35	10,06	10,18	**9,05**
	5	9,39	9,80	9,17	9,08	9,16	**8,99**
	10	9,41	9,61	9,02	**8,95**	9,06	9,03
	20	9,45	9,53	9,01	**8,80**	8,96	8,91

TABLE 7.6 – Risques Hellinger multipliés par 1000 pour la famille $\mathcal{F}_{\mathrm{K}}$.

7.1.4 Coût algorithmique

Du point de vue pratique les procédures $\mathcal{C}^{\mathrm{res.test}}$ diminuent sensiblement le coût algorithmique de $\mathcal{C}_V^{\mathrm{TVF}}$ et semble avoir un coût similaire à celui de $\mathcal{C}^{\mathrm{THO}}$, c'est-à-dire que nous atteignons à peine plus de $|\mathcal{M}|$ tests calculés. Pour illustrer cette faible complexité, nous avons calculé la moyenne, sur les 12 densités, des tests effectués pour les procédures $\mathcal{C}_V^{\mathrm{TVF},Bar}$, $\mathcal{C}_V^{\mathrm{TVF},Bir}$, $\mathcal{C}_V^{\mathrm{res.test}}$ et $\mathcal{C}_{V,1}^{\mathrm{res.test}}$. Dans la Table 7.7, nous considérons la famille $\mathcal{F}_{\mathrm{K}}$ qui contient $|\mathcal{M}| = 38$ méthodes d'estimation. Toutefois, si le nombre de tests calculés est proche de celui que l'on obtient avec la procédure $\mathcal{C}^{\mathrm{THO}}$, le nombre d'opérations des procédures introduites ici est supérieur et dépend, lui, du paramètre V. Ainsi les procédures $\mathcal{C}_V^{\mathrm{res.test}}$ et $\mathcal{C}_{V,1}^{\mathrm{res.test}}$ montrent une aussi bonne précision que $\mathcal{C}_V^{\mathrm{TVF}}$ mais avec un coût algorithmique nettement inférieur ce qui leur donne un avantage certain sur cette dernière.

Famille	V	$\mathcal{C}_V^{\mathrm{TVF},Bir}$	$\mathcal{C}_V^{\mathrm{TVF},Bar}$	$\mathcal{C}_V^{\mathrm{res.test}}$	$\mathcal{C}_{V,1}^{\mathrm{res.test}}$
$\mathcal{F}_{\mathrm{K}}$	2	216	214	104	101
	5	770	753	55	100
	10	2067	2025	44	102
	20	5573	5429	40	103

TABLE 7.7 – Moyenne sur 12 densités du nombre de tests effectués par les différentes procédures.

7.2 Questions ouvertes

Nous dressons ici une liste non-exhaustive de questions que soulèvent les différents chapitres de cette thèse.

7.2.1 Concernant la Partie I

Nous avons prouvé que l'heuristique de pente formulée par Birgé & Massart (2007) n'est plus vraie pour les estimateurs linéaires en général. Elle peut cependant être modifiée dans certains cas, comme pour les estimateurs provenant d'un noyau d'approximation de sorte que la pénalité optimale et minimale sont connues pour sélectionner la fenêtre d'un estimateur à noyau. De manière générale, la recherche d'une méthode automatique construite à partir des données, et qui soit fondée sur des preuves mathématiques, pour calibrer la pénalité est un objectif à garder en tête pour aider les praticiens.

- Existe-t-il un *algorithme de pente*, semblable à celui introduit par Arlot & Bach (2009) en régression, liant pénalité minimale et pénalité optimale dans le cadre général ?
- Comme expliqué au Chapitre 2, l'inégalité oracle du Theorem 2.1 reste optimale tant que $\log(|\mathcal{M}| \vee n)^2$ est négligeable par rapport à n. Ceci convient en particulier pour les familles polynomiales au sens où $|\mathcal{M}| \leq Ln^\alpha$ pour des constantes numériques $L, \alpha > 0$. Des résultats numériques (Lebarbier, 2005) semblent indiquer que l'heuristique de pente reste vraie pour des grandes collections d'estimateurs. Dès lors, peut-on prouver des résultats similaires aux nôtres pour des grandes familles de modèle[(10)] ?
- Que signifie la "complexité" $P\Theta_m$, mis en lumière dans la Section 2.4, pour un estimateur linéaire quelconque ? Existe-t-il une alternative pour désigner celle-ci en général ?

7.2.2 Concernant la Partie II

Nous avons obtenu une généralisation aux estimateurs linéaires de l'analyse minutieuse de Arlot & Lerasle (2014) (restreinte aux estimateurs par projection) notamment au travers d'inégalités oracles optimales au premier ordre. De plus, nous avons également étendu leurs calculs de variance sans toutefois pouvoir interpréter les nombreux termes en général.

- Pourquoi la performance des procédures de sélection semble meilleure lorsque $\mathcal{C}$ est un estimateur biaisé du risque ? En particulier, pourquoi est-il bon de sur-pénaliser par un facteur F égal à 3/2 ou 2 dans la procédure $\mathcal{C}^{\mathrm{pen}}_{V,F}$ (comme on l'observe, par exemple, dans le matériel supplémentaire du Chapitre 3) ? Peut-on déterminer à l'aide des données le *bon niveau* de sur-pénalisation ?
- Peut-on formaliser l'heuristique de Arlot & Lerasle (2014) sur laquelle repose le Chapitre 4 ? Cette heuristique peut-elle mener à une comparaison quantitative des procédures[(11)] ?

[(10)] Par exemple dans des contextes tels que la sélection complète de variables ou la détection de ruptures multiples.

[(11)] L'idéal serait de pouvoir quantifier précisément à *quel point* une procédure de sélection d'estimateurs VF est meilleure qu'une autre pour les estimateurs linéaires (par exemple le LOO est meilleur que le 5-fold).

- Un résultat important qui fait toujours défaut est la preuve théorique que la validation croisée V-fold est strictement meilleure que le hold-out pour la sélection d'estimateurs.

- Arlot & Lerasle (2014) ont réussi à interpréter avec précision les différents termes dans les calculs de variance pour les histogrammes. De la même façon, que pouvons-nous dire sur les quantités qui apparaissent pour les noyaux par approximation ? Un tel résultat, s'il est assez précis, aurait un impact important puisqu'il constituerait un pas important dans la compréhension d'un problème abondamment traité, la sélection d'une fenêtre par validation croisée V-fold.

- Peut-on utiliser les calculs de variance pour comprendre d'autres procédures de validation croisée utilisées en pratique, comme par exemple le "Repeated Learning-Testing" (Breiman *et al.*, 1984; Burman, 1989; Zhang, 1993) qui consiste à faire une validation croisée en choisissant, au hasard, sans remplacement et indépendamment des données, des échantillons de validation de taille fixée ?

- Peut-on généraliser l'analyse de la Partie II au contraste du maximum de vraisemblance (avec pour perte de référence $\ell = \mathrm{KL}$) ?

7.2.3 Concernant la Partie III

Concernant la procédure $\mathcal{C}_V^{\mathrm{TVF}}$, notre analyse théorique n'est pas satisfaisante puisque le principal résultat (le Theorem 6.1) est insuffisant pour expliquer les expériences empiriques convaincantes. De plus, nous ne sommes pas en mesure d'expliquer l'influence du paramètre V sur les performances de la procédure autrement que par nos simulations. Le plus gros défaut de cette procédure reste son coût algorithmique qui retire tout son attrait pour le praticien qui préférera sans doute une procédure légèrement moins précise mais nettement plus rapide.

- Peut-on améliorer l'inégalité oracle (6.17) en adoptant une autre stratégie de preuve que celle du Chapitre 6 ?

- Est-il possible de comprendre théoriquement le rôle de V au travers de l'heuristique de Arlot & Lerasle (2014) ?

- Existe-t-il un algorithme qui calcule très rapidement cette procédure, au moins dans des cas particuliers ?

- Peut-on obtenir une inégalité oracle pour la procédure $\mathcal{C}^{\mathrm{res.test}}$? Nous pensons qu'il est possible d'obtenir un résultat semblable au Theorem 6.1 sans trop de difficulté. L'étape suivante serait alors la même que pour la procédure $\mathcal{C}_V^{\mathrm{TVF}}$. Comprendre l'influence de V du point de vue théorique est un autre défi important.

Bibliographie

Adamczak, R. (2006). Moment inequalities for U-statistics. *Ann. Probab.* **34**, 2288–2314. ISSN 0091-1798.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory (Tsahkadsor, 1971)*, pp. 267–281. Akadémiai Kiadó, Budapest.

Arlot, S. (2007). Resampling and model selection. Ph.D. thesis, Université Paris-Sud 11. `http://tel.archives-ouvertes.fr/tel-00198803/`.

Arlot, S. (2008). V-fold cross-validation improved : V-fold penalization URL `http://hal.archives-ouvertes.fr/hal-00239182/en/`. ArXiv :0802.0566v2.

Arlot, S. (2009). Model selection by resampling penalization. *Electron. J. Stat.* **3**, 557–624 (electronic). ISSN 1935-7524.

Arlot, S. (2015). Minimal penalties and the slope heuristics : a survey. In preparation.

Arlot, S. & Bach, F. (2009). Data-driven calibration of linear estimators with minimal penalties. In Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C. K. I. & Culotta, A., eds., *Advances in Neural Information Processing Systems 22*, pp. 46–54.

Arlot, S. & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statist. Surv.* **4**, 40–79. ISSN 1935-7516.

Arlot, S. & Lerasle, M. (2014). Why $V = 5$ is enough in V-fold cross-validation. ArXiv :1210.5830.

Arlot, S. & Massart, P. (2009). Data-driven calibration of penalties for least-squares regression. *J. Mach. Learn. Res.* **10**, 245–279 (electronic).

Audibert, J.-Y. (2004). A better variance control for pac-bayesian classification. Tech. rep., Laboratoire de Probabilités et Modèles Aléatoires, Universités Paris 6 and Paris 7.

Baraud, Y. (2011). Estimator selection with respect to Hellinger-type risks. *Probab. Theory Related Fields* **151**, 353–401. ISSN 0178-8051.

Baraud, Y. & Birgé, L. (2009). Estimating the intensity of a random measure by histogram type estimators. *Probab. Theory Related Fields* **143**, 239–284. ISSN 0178-8051.

Baraud, Y., Birgé, L. & Sart, M. (2014). A new method for estimation and model selection : ρ-estimation. ArXiv :1403.6057v3.

Barron, A., Birgé, L. & Massart, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory Related Fields* **113**, 301–413. ISSN 0178-8051.

Bartlett, P., Boucheron, S. & Lugosi, G. (2002). Model selection and error estimation. *Machine Learning* **48**, 85–113.

Baudry, J.-P. (2009). Model Selection for Clustering. Choosing the Number of Classes. Theses, Université Paris Sud - Paris XI. URL `https://tel.archives-ouvertes.fr/tel-00461550`.

Baudry, J.-P., Maugis, C. & Michel, B. (2012). Slope heuristics : overview and implementation. *Stat. Comput.* **22**, 455–470. ISSN 0960-3174.

Beran, R. (1977). Robust location estimates. *Ann. Statist.* **5**, 431–444. ISSN 0090-5364.

Beran, R. (1978). An efficient and robust adaptive estimator of location. *Ann. Statist.* **6**, 292–313. ISSN 0090-5364.

Beran, R. (1980). Asymptotic lower bounds for risk in robust estimation. *Ann. Statist.* **8**, 1252–1264. ISSN 0090-5364.

Beran, R. (1981a). Efficient robust estimates in parametric models. *Z. Wahrsch. Verw. Gebiete* **55**, 91–108. ISSN 0044-3719.

Beran, R. (1981b). Efficient robust tests in parametric models. *Z. Wahrsch. Verw. Gebiete* **57**, 73–86. ISSN 0044-3719.

Berlinet, A. & Devroye, L. (1994). A comparison of kernel density estimates. *Publ. Inst. Statist. Univ. Paris* **38**, 3–59.

Bickel, P. J. (1981). Quelques aspects de la statistique robuste. In *Ninth Saint Flour Probability Summer School—1979 (Saint Flour, 1979)*, vol. 876 of *Lecture Notes in Math.*, pp. 1–72. Springer, Berlin-New York.

Billingsley, P. (1995). *Probability and measure*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York, third edn. ISBN 0-471-00710-2. A Wiley-Interscience Publication.

Birgé, L. (1983). Approximation dans les espaces métriques et théorie de l'estimation. *Z. Wahrsch. Verw. Gebiete* **65**, 181–237. ISSN 0044-3719.

Birgé, L. (1984a). Stabilité et instabilité du risque minimax pour des variables indépendantes équidistribuées. *Ann. Inst. H. Poincaré Probab. Statist.* **20**, 201–223. ISSN 0246-0203.

Birgé, L. (1984b). Sur un théorème de minimax et son application aux tests. *Probab. Math. Statist.* **3**, 259–282. ISSN 0208-4147.

Birgé, L. (2006a). Model selection via testing : an alternative to (penalized) maximum likelihood estimators. *Ann. Inst. H. Poincaré Probab. Statist.* **42**, 273–325. ISSN 0246-0203.

Birgé, L. (2006b). Statistical estimation with model selection. *Indag. Math. (N.S.)* **17**, 497–537. ISSN 0019-3577.

Birgé, L. (2007). Model selection for Poisson processes. In *Asymptotics : particles, processes and inverse problems*, vol. 55 of *IMS Lecture Notes Monogr. Ser.*, pp. 32–64. Inst. Math. Statist., Beachwood, OH.

Birgé, L. (2013). Robust tests for model selection. In *From probability to statistics and back : high-dimensional models and processes*, vol. 9 of *Inst. Math. Stat. (IMS) Collect.*, pp. 47–64. Inst. Math. Statist., Beachwood, OH.

Birgé, L. (2014). Model selection for density estimation with $\mathbb{L}_2$-loss. *Probab. Theory Related Fields* **158**, 533–574. ISSN 0178-8051.

Birgé, L. & Massart, P. (1993). Rates of convergence for minimum contrast estimators. *Probab. Theory Related Fields* **97**, 113–150. ISSN 0178-8051.

Birgé, L. & Massart, P. (1997). From model selection to adaptive estimation. In *Festschrift for Lucien Le Cam*, pp. 55–87. Springer, New York.

Birgé, L. & Massart, P. (2007). Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields* **138**, 33–73. ISSN 0178-8051.

Birgé, L. & Rozenholc, Y. (2006). How many bins should be put in a regular histogram. *ESAIM Probab. Stat.* **10**, 24–45 (electronic). ISSN 1292-8100.

Blanchard, G. & Massart, P. (2006). Discussion : "Local Rademacher complexities and oracle inequalities in risk minimization" [Ann. Statist. **34** (2006), no. 6, 2593–2656] by V. Koltchinskii. *Ann. Statist.* **34**, 2664–2671. ISSN 0090-5364.

Bogachev, V. I. (2007). *Measure theory. Vol. I, II*. Springer-Verlag, Berlin. ISBN 978-3-540-34513-8 ; 3-540-34513-2.

Boucheron, S., Lugosi, G. & Massart, P. (2013). *Concentration inequalities*. Oxford University Press, Oxford. ISBN 978-0-19-953525-5. A nonasymptotic theory of independence, With a foreword by Michel Ledoux.

Bousquet, O. (2002). A Bennett concentration inequality and its application to suprema of empirical processes. *C. R. Math. Acad. Sci. Paris* **334**, 495–500. ISSN 1631-073X.

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

Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. (1984). *Classification and regression trees*. Wadsworth Statistics/Probability Series. Wadsworth Advanced Books and Software, Belmont, CA. ISBN 0-534-98053-8 ; 0-534-98054-6.

Breiman, L. & Spector, P. (1992). Submodel selection and evaluation in regression. the x-random case. *International Statistical Review/Revue Internationale de Statistique* pp. 291–319.

Bunea, F., Tsybakov, A. B., Wegkamp, M. H. & Barbu, A. (2010). Spades and mixture models. *Ann. Statist.* **38**, 2525–2558. ISSN 0090-5364.

Burman, P. (1989). A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. *Biometrika* **76**, 503–514.

Castellan, G. (2000). Sélection d'histogrammes à l'aide d'un critère de type Akaike. *C. R. Acad. Sci. Paris Sér. I Math.* **330**, 729–732. ISSN 0764-4442.

Catoni, O. (1999). "universal" aggregation rules with exact bias bounds. *Preprint n.510, LPMA. Available at* `http://www.proba.jussieu.fr/mathdoc/preprints`.

Catoni, O. (2007). *Pac-Bayesian Supervised Classification : The Thermodynamics of Statistical Learning*.

Cavalier, L. & Tsybakov, A. B. (2002). Sharp adaptation for inverse problems with random noise. *Probab. Theory Related Fields* **123**, 323–354.

Celisse, A. (2014). Optimal cross-validation in density estimation with the l^2-loss. *Ann. Statist.* **42**, 1879–1910.

Celisse, A. & Robin, S. (2008). Nonparametric density estimation by exact leave-p-out cross-validation. *Comput. Statist. Data Anal.* **52**, 2350–2368. ISSN 0167-9473.

Chiu, S.-T. (1996). A comparative review of bandwidth selection for kernel density estimation. *Statist. Sinica* **6**, 129–145. ISSN 1017-0405.

Dalelane, C. (2005a). Exact minimax risk for density estimators in non-integer sobolev classes. Available at `http://hal.ccsd.cnrs.fr/ccsd-00004754`.

Dalelane, C. (2005b). Exact oracle inequality for a sharp adaptive kernel density estimator. Available at `http://hal.archives-ouvertes.fr/hal-00004753`.

Daly, J. E. (1988). The construction of optimal histograms. *Comm. Statist. Theory Methods* **17**, 2921–2931. ISSN 0361-0926.

Deheuvels, P. (1977). Estimation non paramétrique de la densité par histogrammes généralisés. II. *Publ. Inst. Statist. Univ. Paris* **22**, 1–23.

Deheuvels, P. & Ouadah, S. (2013). Uniform-in-bandwidth functional limit laws. *J. Theoret. Probab.* **26**, 697–721. ISSN 0894-9840.

DeVore, R. A. & Lorentz, G. G. (1993). *Constructive approximation*, vol. 303 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin. ISBN 3-540-50627-6.

Devroye, L. (1987). *A course in density estimation*, vol. 14 of *Progress in Probability and Statistics*. Birkhäuser Boston, Inc., Boston, MA. ISBN 0-8176-3365-0.

Devroye, L. & Györfi, L. (1985). *Nonparametric density estimation : The L_1 view*. Wiley Series in Probability and Mathematical Statistics : Tracts on Probability and Statistics. John Wiley & Sons, Inc., New York. ISBN 0-471-81646-9.

Devroye, L. & Lugosi, G. (2001). *Combinatorial methods in density estimation*. Springer Series in Statistics. Springer-Verlag, New York. ISBN 0-387-95117-2.

Donoho, D. L. & Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81**, 425–455. ISSN 0006-3444.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. & Picard, D. (1995). Wavelet shrinkage : asymptopia ? *J. Roy. Statist. Soc. Ser. B* **57**, 301–369. ISSN 0035-9246. With discussion and a reply by the authors.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. & Picard, D. (1996). Density estimation by wavelet thresholding. *Ann. Statist.* **24**, 508–539. ISSN 0090-5364.

Efroĭmovich, S. Y. (1985). Nonparametric estimation of a density of unknown smoothness. *Teor. Veroyatnost. i Primenen.* **30**, 524–534. ISSN 0040-361X.

Efroĭmovich, S. Y. & Pinsker, M. S. (1984). A self-training algorithm for nonparametric filtering. *Avtomat. i Telemekh.* pp. 58–65. ISSN 0005-2310.

Efromovich, S. (2000). On sharp adaptive estimation of multivariate curves. *Math. Methods Statist.* **9**, 117–139. ISSN 1066-5307.

Efromovich, S. (2005). Estimation of the density of regression errors. *Ann. Statist.* **33**, 2194–2227. ISSN 0090-5364.

Efromovich, S. (2008). Adaptive estimation of and oracle inequalities for probability densities and characteristic functions. *Ann. Statist.* **36**, 1127–1155. ISSN 0090-5364.

Efron, B. (1979). Bootstrap methods : another look at the jackknife. *Ann. Statist.* **7**, 1–26. ISSN 0090-5364.

Efron, B. (1983). Estimating the error rate of a prediction rule : improvement on cross-validation. *J. Amer. Statist. Assoc.* **78**, 316–331. ISSN 0162-1459.

Feluch, W. & Koronacki, J. (1992). A note on modified cross-validation in density estimation. *Comput. Statist. Data Anal.* **13**, 143–151. ISSN 0167-9473.

Fromont, M. & Tuleau, C. (2006). Functional classification with margin conditions. In *Learning theory*, vol. 4005 of *Lecture Notes in Comput. Sci.*, pp. 94–108. Springer, Berlin.

Gach, F., Nickl, R. & Spokoiny, V. (2013). Spatially adaptive density estimation by localised Haar projections. *Ann. Inst. Henri Poincaré Probab. Stat.* **49**, 900–914. ISSN 0246-0203.

Geisser, S. (1975). The predictive sample reuse method with applications. *J. Amer. Statist. Assoc.* **70**, 320–328.

Giné, E., Latała, R. & Zinn, J. (2000). Exponential and moment inequalities for U-statistics. In *High dimensional probability, II (Seattle, WA, 1999)*, vol. 47 of *Progr. Probab.*, pp. 13–38. Birkhäuser Boston, Boston, MA.

Giné, E. & Nickl, R. (2009). Uniform limit theorems for wavelet density estimators. *Ann. Probab.* **37**, 1605–1646. ISSN 0091-1798.

Giné, E. & Nickl, R. (2010). Adaptive estimation of a distribution function and its density in sup-norm loss by wavelet and spline projections. *Bernoulli* **16**, 1137–1163. ISSN 1350-7265.

Goldenshluger, A. & Lepski, O. (2011). Bandwidth selection in kernel density estimation : oracle inequalities and adaptive minimax optimality. *Ann. Statist.* **39**, 1608–1632. ISSN 0090-5364.

Goldenshluger, A. & Lepski, O. (2014). On adaptive minimax density estimation on R^d. *Probab. Theory Related Fields* **159**, 479–543. ISSN 0178-8051.

Golubev, G. K. (1992). Nonparametric estimation of smooth probability densities in L_2. *Probl. Inf. Transm.* **28**, 44–54. ISSN 0032-9460 ; 1608-3253/e.

Hall, P. (1983). Large sample optimality of least squares cross-validation in density estimation. *Ann. Statist.* **11**, 1156–1174. ISSN 0090-5364.

Hall, P. (1987). On Kullback-Leibler loss and density estimation. *Ann. Statist.* **15**, 1491–1519. ISSN 0090-5364.

Hall, P. (1990). Akaike's information criterion and Kullback-Leibler loss for histogram density estimation. *Probab. Theory Related Fields* **85**, 449–467. ISSN 0178-8051.

Hall, P. & Marron, J. S. (1987). Extent to which least-squares cross-validation minimises integrated square error in nonparametric density estimation. *Probab. Theory Related Fields* **74**, 567–581. ISSN 0178-8051.

Hellinger, E. (1909). Neue Begründung der Theorie quadratischer Formen von unendlichvielen Veränderlichen. *J. Reine Angew. Math.* **136**, 210–271.

Houdré, C. & Reynaud-Bouret, P. (2003). Exponential inequalities, with constants, for U-statistics of order two. In *Stochastic inequalities and applications*, vol. 56 of *Progr. Probab.*, pp. 55–69. Birkhäuser, Basel.

Huber, P. J. (1965). A robust version of the probability ratio test. *Ann. Math. Statist.* **36**, 1753–1758. ISSN 0003-4851.

Jones, M. C., Marron, J. S. & Sheather, S. J. (1996). A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association* **91**, 401–407.

Juditsky, A. & Nemirovski, A. (2000). Functional aggregation for nonparametric regression. *Ann. Statist.* **28**, 681–712. ISSN 0090-5364.

Kakutani, S. (1948). On equivalence of infinite product measures. *Annals of Mathematics* **49**, pp. 214–224. ISSN 0003486X.

Korostelev, A. & Nussbaum, M. (1999). The asymptotic minimax constant for sup-norm loss in nonparametric density estimation. *Bernoulli* **5**, 1099–1118. ISSN 1350-7265.

Kraft, C. (1955). Some conditions for consistency and uniform consistency of statistical procedures. *Univ. California Publ. Statist.* **2**, 125–141.

Larson, S. C. (1931). The shrinkage of the coefficient of multiple correlation. *J. Edic. Psychol.* **22**, 45–55.

Le Cam, L. (1973). Convergence of estimates under dimensionality restrictions. *Ann. Statist.* **1**, 38–53. ISSN 0090-5364.

Le Cam, L. (1975). On local and global properties in the theory of asymptotic normality of experiments. In *Stochastic processes and related topics (Proc. Summer Res. Inst. Statist. Inference for Stochastic Processes, Indiana Univ., Bloomington, Ind., 1974, Vol. 1 ; dedicated to Jerzy Neyman)*, pp. 13–54. Academic Press, New York.

Le Cam, L. (1986). *Asymptotic methods in statistical decision theory*. Springer Series in Statistics. Springer-Verlag, New York. ISBN 0-387-96307-3.

Le Cam, L. & Yang, G. L. (2000). *Asymptotics in statistics*. Springer Series in Statistics. Springer-Verlag, New York, second edn. ISBN 0-387-95036-2. Some basic concepts.

Lebarbier, E. (2005). Detecting multiple change-points in the mean of a Gaussian process by model selection. *Signal Proces.* **85**, 717–736.

Lecué, G. (2007). Méthodes d'agrégation : optimalité et vitesses rapides. Ph.D. thesis. URL `http://www.theses.fr/2007PA066687`. Thèse de doctorat dirigée par Tsybakov, Alexandre B. Mathématiques Paris 6 2007.

Ledoux, M. (2001). *The concentration of measure phenomenon*, vol. 89 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI. ISBN 0-8218-2864-9.

Ledoux, M. & Talagrand, M. (1991). *Probability in Banach spaces*, vol. 23 of *Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)]*. Springer-Verlag, Berlin. ISBN 3-540-52013-9. Isoperimetry and processes.

Lepskiĭ, O. V. (1991). Asymptotically minimax adaptive estimation. I. Upper bounds. Optimally adaptive estimates. *Teor. Veroyatnost. i Primenen.* **36**, 645–659. ISSN 0040-361X.

Lepskiĭ, O. V. (1992a). Asymptotically minimax adaptive estimation. II. Schemes without optimal adaptation. Adaptive estimates. *Teor. Veroyatnost. i Primenen.* **37**, 468–481. ISSN 0040-361X.

Lepskiĭ, O. V. (1992b). On problems of adaptive estimation in white Gaussian noise. In *Topics in nonparametric estimation*, vol. 12 of *Adv. Soviet Math.*, pp. 87–106. Amer. Math. Soc., Providence, RI.

Lerasle, M. (2012). Optimal model selection in density estimation. *Ann. Inst. Henri Poincaré Probab. Stat.* **48**, 884–908. ISSN 0246-0203.

Lerasle, M., Magalhães, N. & Reynaud-Bouret, P. (2015). Optimal kernel selection for density estimation. *Preprint* .

Loader, C. (1999). *Local regression and likelihood*. Statistics and Computing. Springer-Verlag, New York. ISBN 0-387-98775-4.

Lugosi, G. & Nobel, A. B. (1999). Adaptive model selection using empirical complexities. *Ann. Statist.* **27**, 1830–1864. ISSN 0090-5364.

Magalhães, N. & Rozenholc, Y. (2014). An efficient algorithm for T-estimation. *http ://hal.archives-ouvertes.fr/hal-00986229* .

Mallows, C. L. (1973). Comments on C_p. *Technometrics* **15**, 661–675.

Marron, J. S. (1987). A comparison of cross-validation techniques in density estimation. *The Annals of Statistics* **15**, 152–162.

Mason, D. M. & Newton, M. A. (1992). A rank statistics approach to the consistency of a general bootstrap. *Ann. Statist.* **20**, 1611–1624. ISSN 0090-5364.

Mason, D. M. & Swanepoel, J. W. H. (2011). A general result on the uniform in bandwidth consistency of kernel-type function estimators. *TEST* **20**, 72–94. ISSN 1133-0686.

Massart, P. (2007). *Concentration inequalities and model selection*, vol. 1896 of *Lecture Notes in Mathematics*. Springer, Berlin. ISBN 978-3-540-48497-4 ; 3-540-48497-3. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

Maugis, C. & Michel, B. (2011a). Data-driven penalty calibration : a case study for Gaussian mixture model selection. *ESAIM Probab. Stat.* **15**, 320–339. ISSN 1292-8100.

Maugis, C. & Michel, B. (2011b). A non asymptotic penalized criterion for Gaussian mixture model selection. *ESAIM Probab. Stat.* **15**, 41–68. ISSN 1292-8100.

Mildenberger, T. & Weinert, H. (2012). The benchden package : Benchmark densities for nonparametric density estimation. *Journal of Statistical Software* **46**, 1–14.

Nemirovski, A. (2000). Topics in non-parametric statistics. In *Lectures on probability theory and statistics (Saint-Flour, 1998)*, vol. 1738 of *Lecture Notes in Math.*, pp. 85–277. Springer, Berlin.

Parzen, E. (1962). On estimation of a probability density function and mode. *Ann. Math. Statist.* **33**, 1065–1076. ISSN 0003-4851.

Pinelis, I. (1994). Optimum bounds for the distributions of martingales in banach spaces. *Ann. Probab.* **22**, 1679–1706.

Pinsker, M. S. (1980). Optimal filtration of square-integrable signals in Gaussian noise. *Problems Inform. Transmission* **16**, 52–68.

Præstgaard, J. & Wellner, J. A. (1993). Exchangeably weighted bootstraps of the general empirical process. *Ann. Probab.* **21**, 2053–2086. ISSN 0091-1798.

Reynaud-Bouret, P., Rivoirard, V. & Tuleau-Malot, C. (2011). Adaptive density estimation : A curse of support ? *Journal of Statistical Planning and Inference* **141**, 115 – 139. ISSN 0378-3758.

Rigollet, P. (2006a). Adaptive density estimation using the blockwise Stein method. *Bernoulli* **12**, 351–370. ISSN 1350-7265.

Rigollet, P. (2006b). Oracle inequalities, aggregation and adaptation. Theses, Université Pierre et Marie Curie - Paris VI. URL `https://tel.archives-ouvertes.fr/tel-00115494`.

Rigollet, P. & Tsybakov, A. B. (2007). Linear and convex aggregation of density estimators. *Math. Methods Statist.* **16**, 260–280. ISSN 1066-5307.

Rio, E. (2012). Sur la fonction de taux dans les inégalités de Talagrand pour les processus empiriques. *C. R. Math. Acad. Sci. Paris* **350**, 303–305. ISSN 1631-073X.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.* **27**, 832–837. ISSN 0003-4851.

Rozenholc, Y., Mildenberger, T. & Gather, U. (2010). Combining regular and irregular histograms by penalized likelihood. *Comput. Statist. Data Anal.* **54**, 3313–3323. ISSN 0167-9473.

Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scand. J. Statist.* **9**, 65–78. ISSN 0303-6898.

Samarov, A. & Tsybakov, A. (2004). Nonparametric independent component analysis. *Bernoulli* **10**, 565–582. ISSN 1350-7265.

Sart, M. (2011). Model selection for Poisson processes with covariates. ArXiv :1112.5634.

Sart, M. (2013). Robust estimation on a parametric model with tests. ArXiv :1308.2927v2.

Sart, M. (2014). Estimation of the transition density of a markov chain. *Ann. Inst. H. Poincaré Probab. Statist.* **50**, 1028–1068.

Saumard, A. (2010). Regular Contrast Estimation and the Slope Heuristics. Ph.D. thesis, Université Rennes 1. URL `https://tel.archives-ouvertes.fr/tel-00569372`.

Savchuk, O. Y., Hart, J. D. & Sheather, S. J. (2010). Indirect cross-validation for density estimation. *J. Amer. Statist. Assoc.* **105**, 415–423. ISSN 0162-1459. With supplementary material available online.

Scott, D. W. (1992). *Multivariate density estimation*. Wiley Series in Probability and Mathematical Statistics : Applied Probability and Statistics. John Wiley & Sons, Inc., New York. ISBN 0-471-54770-0. Theory, practice, and visualization, A Wiley-Interscience Publication.

Scott, D. W. & Terrell, G. R. (1987). Biased and unbiased cross-validation in density estimation. *J. Amer. Statist. Assoc.* **82**, 1131–1146. ISSN 0162-1459.

Shao, J. (1997). An asymptotic theory for linear model selection. *Statist. Sinica* **7**, 221–264.

Sheather, S. J. & Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *J. Roy. Statist. Soc. Ser. B* **53**, 683–690. ISSN 0035-9246.

Shibata, R. (1981). An optimal selection of regression variables. *Biometrika* **68**, 45–54. ISSN 0006-3444.

Silverman, B. W. (1986). *Density Estimation.* London : Chapman and Hall.

Stone, C. J. (1984). An asymptotically optimal window selection rule for kernel density estimates. *The Annals of Statistics* **12**, 1285–1297.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *J. Roy. Statist. Soc. Ser. B* **36**, 111–147. ISSN 0035-9246. With discussion by G. A. Barnard, A. C. Atkinson, L. K. Chan, A. P. Dawid, F. Downton, J. Dickey, A. G. Baker, O. Barndorff-Nielsen, D. R. Cox, S. Giesser, D. Hinkley, R. R. Hocking, and A. S. Young, and with a reply by the authors.

Stute, W. (1992). Modified cross-validation in density estimation. *J. Statist. Plann. Inference* **30**, 293–305. ISSN 0378-3758.

Talagrand, M. (1996). New concentration inequalities in product spaces. *Invent. Math.* **126**, 505–563. ISSN 0020-9910.

Tsybakov, A. B. (2003). Optimal rates of aggregation. In Schölkopf, B. & Warmuth, M. K., eds., *Learning Theory and Kernel Machines*, vol. 2777 of *Lecture Notes in Computer Science*, pp. 303–313. Springer Berlin Heidelberg. ISBN 978-3-540-40720-1.

Tsybakov, A. B. (2009). *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York. ISBN 978-0-387-79051-0. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats.

van der Laan, M. J. & Dudoit, S. (2003). "unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator : Finite sample oracle inequalities and examples". *U.C. Berkeley Division of Biostatistics Working Paper Series. Working Paper 130.* .

Verzelen, N. (2010). High-dimensional Gaussian model selection on a Gaussian design. *Ann. Inst. Henri Poincaré Probab. Stat.* **46**, 480–524. ISSN 0246-0203.

Villers, F. (2007). Tests et sélection de modèles pour l'analyse de données protéomiques et transcriptomiques. Ph.D. thesis. URL `http://www.theses.fr/2007PA112198`. Thèse de doctorat dirigée par Massart, Pascal et Huet, Sylvie Mathématiques Paris 11 2007.

Walter, G. & Blum, J. (1979). Probability density estimation using delta sequences. *Ann. Statist.* **7**, 328–340. ISSN 0090-5364.

Wand, M. P. (1997). Data-based choice of histogram bin width. *The American Statistician* **51**, pp. 59–64. ISSN 00031305.

Wand, M. P. & Jones, M. C. (1995). *Kernel smoothing*, vol. 60 of *Monographs on Statistics and Applied Probability*. Chapman and Hall, Ltd., London. ISBN 0-412-55270-1.

Watson, G. S. & Leadbetter, M. R. (1963). On the estimation of the probability density. I. *Ann. Math. Statist.* **34**, 480–491. ISSN 0003-4851.

Watson, G. S. & Leadbetter, M. R. (1964a). Hazard analysis. I. *Biometrika* **51**, 175–184. ISSN 0006-3444.

Watson, G. S. & Leadbetter, M. R. (1964b). Hazard analysis. II. *Sankhyā Ser. A* **26**, 101–116. ISSN 0581-572X.

Wegkamp, M. (2003). Model selection in nonparametric regression. *Ann. Statist.* **31**, 252–273. ISSN 0090-5364.

Wegkamp, M. H. (1999). Quasi-universal bandwidth selection for kernel density estimators. *Canad. J. Statist.* **27**, 409–420. ISSN 0319-5724.

Whittle, P. (1958). On the smoothing of probability density functions. *J. Roy. Statist. Soc. Ser. B* **20**, 334–343. ISSN 0035-9246.

Winter, B. B. (1975). Rate of strong consistency of two nonparametric density estimators. *Ann. Statist.* **3**, 759–766. ISSN 0090-5364.

Yang, Y. (2000). Mixing strategies for density estimation. *Ann. Statist.* **28**, 75–87. ISSN 0090-5364.

Zhang, P. (1993). Model selection via multifold cross validation. *Ann. Statist.* **21**, 299–313. ISSN 0090-5364.

Validation croisée et pénalisation pour l'estimation de densité

Résumé. Cette thèse s'inscrit dans le cadre de l'estimation d'une densité, considéré du point de vue non-paramétrique et non-asymptotique. Elle traite du problème de la sélection d'une méthode d'estimation à noyau. Celui-ci est une généralisation, entre autre, du problème de la sélection de modèle et de la sélection d'une fenêtre. Nous étudions des procédures classiques, par pénalisation et par rééchantillonnage (en particulier la validation croisée V-fold), qui évaluent la qualité d'une méthode en estimant son risque. Nous proposons, grâce à des inégalités de concentration, une méthode pour calibrer la pénalité de façon optimale pour sélectionner un estimateur linéaire et prouvons des inégalités d'oracle et des propriétés d'adaptation pour ces procédures. De plus, une nouvelle procédure rééchantillonnée, reposant sur la comparaison entre estimateurs par des tests robustes, est proposée comme alternative aux procédures basées sur le principe d'estimation sans biais du risque.

Un second objectif est la comparaison de toutes ces procédures du point de vue théorique et l'analyse du rôle du paramètre V pour les pénalités V-fold. Nous validons les résultats théoriques par des études de simulations.

Mots-clés : Statistiques non-paramétriques, estimation de densité, sélection d'estimateur, sélection d'une méthode d'estimation, validation croisée V-fold, pénalisation, T-estimation, inégalités d'oracle, heuristique de pente, estimation adaptative, perte Hellinger.

Cross-validation and penalization in density estimation

Abstract. This thesis takes place in the density estimation setting from a nonparametric and nonasymptotic point of view. It concerns the statistical algorithm selection problem which generalizes, among others, the problem of model and bandwidth selection. We study classical procedures, such as penalization or resampling procedures (in particular V-fold cross-validation), which evaluate an algorithm by estimating its risk. We provide, thanks to concentration inequalities, an optimal penalty for selecting a linear estimator and we prove oracle inequalities and adaptative properties for resampling procedures. Moreover, new resampling procedure, based on estimator comparison by the mean of robust tests, is introduced as an alternative to procedures relying on the unbiased risk estimation principle.

A second goal of this work is to compare these procedures from a theoretical point of view and to understand the role of V for V-fold penalization. We validate these theoretical results on empirical studies.

Keywords : Non-parametric statistics, density estimation, estimator selection, statistical algorithm selection, linear estimators, V-fold cross-validation, penalization, T-estimation, oracle inequalities, slope heuristics, adaptive estimation, Hellinger loss.

AMS Classification : 62G07, 62G09, 62G10, 62G35, 62C20