Questions théoriques et applications pratiques en apprentissage statistique et statistique non paramétrique

Aurélie Fischer

Université Paris Cité

LPSM
LABORATOIRE DE PROBABILITÉS
STATISTIQUE & MODÉLISATION

Mémoire d'Habilitation à Diriger des Recherches

Spécialité : Mathématiques

Questions théoriques et applications pratiques en apprentissage statistique et statistique non paramétrique

Courbes principales, classification non supervisée, modèles prédictifs, agrégation d'estimateurs

Présenté par

Aurélie FISCHER

Le 21 juin 2022, devant le jury composé de :

Gérard BIAU	Sorbonne Université
Gilles BLANCHARD	Université Paris-Saclay
Stéphane BOUCHERON	Université Paris Cité
Jérôme DEDECKER	Université Paris Cité
Mathilde MOUGEOT	ENS Paris-Saclay
Anne PHILIPPE	Nantes Université
Wolfgang POLONIK	UC Davis, USA

Questions théoriques et applications pratiques en apprentissage statistique et statistique non paramétrique

Courbes principales, classification non supervisée,
modèles prédictifs, agrégation d'estimateurs

Theoretical issues and practical applications in statistical learning and nonparametric statistics

Principal curves, clustering, predictive models,
aggregation of estimators

Remerciements

En premier lieu, je tiens à remercier ceux qui ont consenti à écrire un rapport sur ce manuscrit. Tout d'abord, je suis reconnaissante à Stéphane Boucheron, mon "rapporteur local", à qui j'avais essayé en vain d'éviter cette corvée, et qui s'en est acquitté avec générosité, en rédigeant un document conséquent. Merci également, Stéphane, pour ta bienveillance constante, pour ton avis et ton aide concernant certains points d'organisation. Je suis très honorée d'avoir Gilles Blanchard et Wolfgang Polonik comme rapporteurs. Un très grand merci, Gilles, d'avoir accepté en toute simplicité cette tâche chronophage malgré un emploi du temps très chargé, et de l'avoir accomplie avec diligence en respectant strictement le délai annoncé. Herzlichen Dank dafür, Wolfgang, dass du einverstanden warst, diese Gutachter Arbeit zu übernehmen ; es ist mir eine Ehre.

Ensuite, je voudrais remercier très sincèrement Gérard Biau, Jérôme Dedecker, Mathilde Mougeot et Anne Philipe qui ont bien voulu participer au jury.

Gérard, merci d'avoir accepté de me faire une petite place dans ton agenda bien rempli. Malgré tes hautes responsabilités, tu as su rester ouvert et accessible, comme tu l'étais durant ma thèse.

Jérôme, je suis heureuse que tu présides ce jury. Maintenant que certaines charges prennent fin, j'espère que nous parviendrons à dégager du temps pour travailler ensemble.

Mathilde, mon arrivée ici il y a 10 ans, c'est un peu grâce à toi... C'est une grande joie de te compter parmi les membres de ce jury !

Anne, merci de ton intérêt pour mon travail, je me réjouis de t'en présenter une partie à l'occasion de la soutenance. Si l'opportunité d'entreprendre une collaboration se présente, ce sera avec plaisir.

Je remercie de tout coeur l'ensemble de mes collaborateurs, à qui les différents travaux présentés ici doivent tant. Réfléchir ensemble devant un tableau, une feuille

blanche, rebondir sur une interrogation, susciter de nouvelles questions, et faire ainsi avancer la réflexion, voilà une grande partie du charme de la recherche ! Merci Benjamin, Bertrand, Claire, Clément, Dominique, Gérard, Ilaria, Jérôme, Lucie, Pamela. Merci Sylvain pour les solénoïdes au tableau et "les dessins" sur mon bureau. Merci Mathilde pour tes questions et les idées qu'elles font éclore. Merci Riwal pour nos échanges interdisciplinaires.

Le laboratoire et l'UFR constituent une communauté accueillante et sympathique : merci à tous les collègues !

Merci aux directeurs et directeurs adjoints du laboratoire : Bastien, mon voisin à Sophie Germain, toujours de bon conseil, Francis, qui vient de nous quitter, trop tôt, Gilles et Lorenzo. Je remercie toute l'équipe administrative du laboratoire, qui nous facilite le quotidien, avec beaucoup de patience et une grande disponibilité. Merci Nathalie, Valérie, fidèles au poste depuis le début.

Des équipes de choc à la tête de l'UFR, ça donne la pêche ! Merci Georges pour ton accueil chaleureux. Merci à vous, Isabelle et Giambattista, puis Arnaud et Mathieu : c'était un plaisir de faire partie du Conseil d'UFR avec vous à la direction. Merci aux élus du Conseil qui ont partagé avec moi ce mandat qui ne voulait plus finir, entre fusion et crise sanitaire, et à notre nouveau tandem de directeurs, David et Xavier.

Je n'oublie pas les responsables administratifs successifs : merci pour votre compétence dévouée.

Merci Xavier et Olivier, vos conseils et votre soutien dans l'ensemble du processus administratif relatif à l'HDR ont été précieux !

Merci à tous les collègues côtoyés au cours des diverses aventures passées, présentes ou à venir que représentent le Comité National, la préparation du projet ANR GeoDSIC, l'organisation du Trimestre Thématique IHP "Geometry and Statistics in Data Science 2022", l'élaboration d'une plateforme Math & IA, le Conseil de la SFdS... Merci Bertrand, on ne peut rêver meilleur co-équipier ! J'en profite pour remercier également Frédéric Chazal et Pascal Massart, toujours prêts à transmettre leur expérience avec bonne humeur et bienveillance.

Je remercie tous les étudiants grâce à qui j'ai connu mes premières expériences d'encadrement, notamment Adam, Clarisse, Maya, Maëlle, Marie, Raphaël, et bien sûr Sothea, qui soutiendra sa thèse le mois prochain.

Enfin, une pensée à mes amis, à ma famille, tous ceux qui m'ont entendu répéter que j'aurais dû terminer la rédaction de ce manuscrit un peu plus rapidement... et qui ensuite ont continué à en entendre parler plusieurs mois durant, jusqu'à ce jour !

Liste des symboles

$\mathrm{Im} f$	Image de f.
$\|x\|$	Norme euclidienne de x.
$\langle x, y \rangle$	Produit scalaire euclidien entre x et y.
$d(x, y)$	Distance euclidienne entre x et y.
$d(x, A)$	Distance euclidienne du point x à l'ensemble A : $d(x, A) = \min_{y \in A} \|x - y\|$.
diam(A)	Diamètre de l'ensemble A : $\mathrm{diam}(A) = \max_{x,y \in A} \|x - y\|$.
$\mathrm{Card}(A)$	Cardinal de l'ensemble A.
$d_H(A, B)$	Distance de Hausdorff entre les ensembles A et B : $d_H(A, B) = \sup_{a \in A} d(a, B) \vee \sup_{b \in B} d(b, A)$.
$\mathscr{L}(f)$	Longueur d'une courbe f.
$\lambda(\cdot)$	Mesure de Lebesgue sur $\mathbb{R}$.
δ_x	Mesure de Dirac en x.
$\mathcal{H}^1$	Mesure de Hausdorff de dimension 1 sur $\mathbb{R}^d$.
$\mathcal{B}(E)$	Tribu borélienne sur l'espace E.
$C([0, 1])$	Espace métrique des fonctions continues de $[0, 1]$ dans $\mathbb{R}^d$, muni de la topologie de la convergence uniforme.
$\mathcal{N}(m, \sigma^2)$	Loi normale de moyenne μ et de variance σ^2.

List of symbols

$\mathrm{Im} f$	Image of f.
$\lvert x \rvert$	Euclidean norm of x.
$\langle x, y \rangle$	Euclidean inner product between x and y.
$d(x, y)$	Euclidean distance between x and y.
$d(x, A)$	Euclidean distance from point x to the set A : $d(x, A) = \min_{y \in A} \lvert x - y \rvert$.
diam(A)	Diameter of the set A : $\mathrm{diam}(A) = \max_{x,y \in A} \lvert x - y \rvert$.
$\mathrm{Card}(A)$	Cardinality of the set A.
$d_H(A, B)$	Hausdorff distance between sets A and B : $d_H(A, B) = \sup_{a \in A} d(a, B) \vee \sup_{b \in B} d(b, A)$.
$\mathscr{L}(f)$	Length of a curve f.
$\lambda(\cdot)$	Lebesgue measure on $\mathbb{R}$.
δ_x	Dirac measure at x.
$\mathcal{H}^1$	1-dimensional Hausdorff measure on $\mathbb{R}^d$.
$\mathcal{B}(E)$	Borel sigma-algebra of the space E.
$C([0, 1])$	Metric space of continuous functions from $[0, 1]$ to $\mathbb{R}^d$, equipped with the topology of uniform convergence.
$\mathcal{N}(m, \sigma^2)$	Normal distribution with mean μ and variance σ^2.

Avant-propos / Foreword

Ce document est construit de la manière suivante. La première partie est une introduction générale, comprenant deux chapitres. Le premier chapitre est une synthèse de mes travaux de recherche, rédigée en français. Le second chapitre en est une version en anglais. Les chapitres suivants, tous rédigés en anglais, regroupés en 4 parties thématiques, détaillent chacun des travaux présentés dans la partie introductive. Ils correspondent en général à des articles de recherche, qui ont été publiés dans une revue ou sont actuellement soumis. Il s'agit la plupart du temps d'une version abrégée de l'article, ne contenant pas les démonstrations complètes. Souvent, le texte a été simplifié, allégé ou réorganisé. Le dernier chapitre présente, de manière plus succincte, quelques collaborations interdisciplinaires supplémentaires. Le manuscrit se clôt par une partie conclusive dédiée aux perspectives : un premier chapitre en français puis un second chapitre en anglais mentionnent quelques travaux actuellement en cours ainsi que mes principaux projets de recherche.

This document is constructed as follows. The first part is a general introduction, consisting of two chapters. The first chapter is a synthesis of my research work, written in French. The second chapter is the corresponding English version. Then, the next chapters, all written in English, grouped in 4 thematic parts, detail each of the works presented in the introductory part. Most of them correspond to research articles, which have been published in a journal or are currently submitted. Usually, a chapter is an abbreviated version of an article, without complete proofs. Often, the text has been simplified, lightened or reorganized. The last chapter presents, more succinctly, some additional interdisciplinary collaborations. The manuscript ends with a concluding part dedicated to perspectives : a first chapter in French and a second chapter in English mention some work in progress and future research directions.

Contents

Part I

Introduction

Chapitre I.1

Présentation synthétique des travaux

I.1.1 Introduction

Mes thèmes de recherche concernent l'apprentissage statistique, la statistique non paramétrique, et la théorie des probabilités. Une partie de mes contributions est essentiellement théorique, tandis qu'une autre partie est consacrée à des applications pratiques, dans le cadre de collaborations industrielles ou interdisciplinaires.

Apprentissage supervisé et non supervisé. L'apprentissage statistique comporte deux branches, l'apprentissage supervisé et l'apprentissage non supervisé. En apprentissage non supervisé, on observe des données $X_1, \ldots, X_n$, supposées indépendantes, de même loi qu'une variable aléatoire générique X, et le but est d'apprendre certaines caractéristiques de la structure sous-jacente de X. En apprentissage supervisé, on dispose de $(X_1, Y_1), \ldots, (X_n, Y_n)$, copies indépendantes d'un couple (X, Y), et l'objectif est d'apprendre, grâce aux observations, la fonction reliant l'entrée X et la sortie Y, afin de pouvoir prédire la sortie associée à une nouvelle entrée.

Problèmes étudiés. Le cas non supervisé au sens large correspond ici à la classification non supervisée (*clustering*) ainsi qu'à certaines questions concernant les courbes principales. Un travail en déconvolution est aussi rattaché à ce cadre. Les problèmes supervisés abordés dans ce document, en classification ou en régression, sont tous liés à des stratégies d'agrégation et/ou à des problématiques réelles. Hormis l'apprentissage statistique, les courbes principales sont également considérées avec un point de vue probabiliste, puis dans un contexte d'estimation de courbe.

Optimalité minimax. Quelquefois, dans la suite, en discutant de la performance

d'un estimateur, nous évoquerons l'optimalité au sens minimax. Dans un cadre général où $s \in \mathcal{S}$ désigne une quantité inconnue à estimer et où la performance d'un estimateur est mesurée grâce à un risque R, un estimateur S^* est dit minimax s'il vérifie

$$R(S^*, s) = \inf_T \sup_{s \in \mathcal{S}} R(T, s),$$

où la borne inférieure est prise sur tous les estimateurs T de s possibles. Dans ce contexte, supposons qu'un estimateur S vérifie $R(S, s) \leq C\psi_n$ pour tout $s \in \mathcal{S}$, avec $(\psi_n)_{n \geq 1}$ une suite convergeant vers 0 et $C > 0$ une constante. Si, par ailleurs, on peut trouver s tel que le meilleur estimateur possible ne converge pas à une vitesse plus rapide que ψ_n, c'est-à-dire

$$\inf_T \sup_{s \in \mathcal{S}} R(T, s) \geq c\psi_n,$$

où $c > 0$ est une constante, cette vitesse ψ_n est appelée vitesse de convergence minimax, et l'estimateur S est donc optimal au sens minimax.

Dans tout le document, sauf mention contraire, $\mathbb{R}^d$ est muni de la norme euclidienne $|\cdot|$.

I.1.2 Courbes principales

Nous nous intéressons aux propriétés de régularité, ainsi qu'à la performance d'estimation des courbes principales contraintes, telles qu'introduites dans Kégl et al. (2000). Selon la définition introduite par ces auteurs, une courbe principale f pour une variable aléatoire X de carré intégrable est une fonction à valeurs dans $\mathbb{R}^d$ minimisant un critère $E\left[d^2(X, \mathrm{Im}f)\right]$ sous une certaine contrainte, typiquement une contrainte sur la longueur de la courbe. Ici, $d(\cdot, \cdot)$ représente la distance euclidienne dans $\mathbb{R}^d$ et $\mathrm{Im}f \subset \mathbb{R}^d$ désigne l'image de f. Cette notion de courbe principale correspond également à une version du « problème de distance moyenne » étudié dans la communauté de calcul des variations et d'optimisation de formes (Buttazzo et Stepanov, 2003; Buttazzo et al., 2002). La version empirique du critère, pour un ensemble d'observations $X_1, \ldots, X_n$, est $\frac{1}{n}\sum_{i=1}^n d^2(X_i, \mathrm{Im}f)$. Dans la suite, nous considérons aussi des mesures de distance plus générales.

I.1.2.1 Sélection de modèle

Une première direction de recherche sur les courbes principales, concernant la sélection de modèle, développée dans Biau et Fischer (2012) et Fischer (2013), est présentée dans le Chapitre II.2. La construction d'une courbe principale empirique suppose en effet le choix de certains paramètres de régularité (longueur, courbure, nombre de segments pour une ligne polygonale...) qui déterminent la forme exacte de la courbe. Intuitivement, l'objectif est de trouver la structure sous-jacente des données, en évitant l'interpolation. Considérant deux modèles différents, en estimation et en apprentissage statistique respectivement, nous effectuons une sélection de paramètres basée sur la théorie de sélection de modèle par pénalisation (Massart, 2007). Une garantie théorique sur l'estimateur sélectionné est fournie sous la forme d'une inégalité de type oracle, exprimant que sa performance est proche de la meilleure possible sur la collection de modèles considérée.

Dans Fischer (2013), nous étudions la sélection de la longueur dans l'estimation par courbes principales dans un modèle gaussien. Plus précisément, nous observons des vecteurs aléatoires $X_1, \dots, X_n$ à valeurs dans $\mathbb{R}^d$ tels que

$$X_i = x_i^* + \sigma\xi_i, \quad i = 1, \dots, n,$$

où les x_i^* sont inconnus, les ξ_i sont des vecteurs gaussiens standards de $\mathbb{R}^d$ indépendants, et $\sigma > 0$ représente le niveau de bruit, qui est supposé connu.

Soit $\{w_\ell\}_{\ell \in \mathcal{L}}$ une famille sommable de poids. Supposons que $(\hat{x}_{1,\ell}, \dots, \hat{x}_{n,\ell})$ minimise $\frac{1}{n}\sum_{i=1}^n |X_i - x_i|^2$ en $(x_1, \dots, x_n) \in (\mathrm{Im} f)^n$ où f appartient à une classe $\mathcal{F}_\ell$ de courbes de longueur ℓ avec les extrémités fixées. Alors, pour σ en-dessous d'un certain seuil, si

$$\mathrm{pen}(\ell) \geq \eta\sigma^2 \left[c_1 \left(\ln\left(\frac{\ell^{1/d}\lambda^{1-1/d}}{\sigma} \right) + c_2 \right) + \frac{4w_\ell}{nd} \right],$$

où λ dépend de ℓ et de la distance entre les extrémités, presque sûrement, il existe un minimiseur $\hat{\ell}$ du critère pénalisé

$$\mathrm{crit}(\ell) = \frac{1}{nd}\sum_{i=1}^n |X_i - \hat{x}_{i\ell}|^2 + \mathrm{pen}(\ell).$$

De plus, si $\tilde{x}_i = \hat{x}_{i\hat{\ell}}$ pour tout $i = 1, \dots, n$, on a

$$\frac{1}{nd}\sum_{i=1}^n E|\tilde{x}_i - x_i^*|^2 \leq c(\eta)\Big[\inf_\ell \Big\{ \inf_{f \in \mathcal{F}_\ell} \frac{1}{nd}\sum_{i=1}^n d^2(x_i^*, \mathrm{Im} f) + \mathrm{pen}(\ell) \Big\} + \frac{\sigma^2}{nd} \Big].$$

Dans Biau et Fischer (2012), la question de la sélection de paramètres est abordée dans le contexte de l'apprentissage statistique. Les estimateurs sont des lignes polygonales indexées par le nombre d'arêtes et la longueur ou la courbure. Nous présentons ici le résultat de sélection de modèle pour la longueur.

Soit $X_1, \ldots, X_n$ un échantillon d'un vecteur aléatoire générique X tel que $P(X \in K) = 1$, où K est un sous-ensemble convexe compact de $\mathbb{R}^d$, de diamètre δ. On pose

$$\Delta(f) = E\left[d^2(X, \mathrm{Im} f)\right], \quad \Delta_n(f) = \frac{1}{n}\sum_{i=1}^{n} d^2(X_i, \mathrm{Im} f),$$

et $\mathcal{D}(f, g) = \Delta(g) - \Delta(f)$. Soit f^* un minimiseur de $\Delta(f)$.

On suppose que $\hat{f}_{k,\ell}$ minimise le critère empirique $\Delta_n(f)$ sur une classe $\mathcal{F}_{k,\ell}$ de lignes polygonales de longueur au plus ℓ, à k segments. Soit $\{x_{k,\ell}\}_{k,\ell}$ une famille sommable de poids. Il existe $c, c_0, \ldots, c_2$, tels que, si

$$\mathrm{pen}(k, \ell) \geq \frac{1}{\sqrt{n}}\left[c_1\sqrt{k} + c_2\ell + c_0\right] + \delta^2\sqrt{\frac{x_{k,\ell}}{2n}},$$

alors, si $\tilde{f}$ désigne la ligne polygonale sélectionnée en minimisant $\mathrm{crit}(k, \ell) = \Delta_n(\hat{f}_{k,\ell}) + \mathrm{pen}(k, \ell)$, on a

$$E[\mathcal{D}(f^*, \tilde{f})] \leq \inf_{k,\ell}\left[\inf_{f \in \mathcal{F}_{k,\ell}} \mathcal{D}(f^*, f) + \mathrm{pen}(k, \ell)\right] + \frac{\delta^2 c}{\sqrt{n}}.$$

I.1.2.2 Régularité

Dans l'article Delattre et Fischer (2020), exposé dans le Chapitre II.3, nous étudions les propriétés théoriques satisfaites par une courbe principale $f^* : [0, 1] \to \mathbb{R}^d$ de longueur au plus L, associée à une loi de probabilité ayant un moment d'ordre 2. On suppose que le support de la loi n'est pas l'image d'une courbe de longueur L. Considérant des courbes optimales aussi bien ouvertes (ayant des extrémités) que fermées (sans extrémités), nous montrons qu'elles ont une courbure finie et établissons une équation d'Euler-Lagrange.

Plus précisément, notant

$$\Delta(f) = E\left[d^2(X, \mathrm{Im} f)\right],$$

on définit, pour $L \geq 0$,

$$G(L) = \min\{\Delta(f), f \in \mathcal{C}_L\},$$

où $\mathcal{C}_L$ désigne l'un des deux ensembles de courbes suivants :

$$\{f \in [0,1] \to \mathbb{R}^d, \mathscr{L}(f) \leq L\},$$
$$\{f \in [0,1] \to \mathbb{R}^d, \mathscr{L}(f) \leq L, f(0) = f(1)\}.$$

Soit $L > 0$ tel que $G(L) > 0$ et soit $f^* \in \mathcal{C}_L$ telle que $\Delta(f^*) = G(L)$. Alors, la longueur de f^* est égale à L. En supposant f^* L-lipschitzienne, nous obtenons les propriétés suivantes :

— f^* est dérivable à droite sur $[0,1)$, $|f_r^{*\prime}(t)| = L$ pour tout $t \in [0,1)$,
— f^* est dérivable à gauche sur $(0,1]$, $|f_\ell^{*\prime}(t)| = L$ pour tout $t \in (0,1]$,

et il existe une unique mesure signée $f^{*\prime\prime}$ sur $[0,1]$ (à valeurs dans $\mathbb{R}^d$) telle que
— $f^{*\prime\prime}((s,t]) = f_r^{*\prime}(t) - f_r^{*\prime}(s)$ pour tout $0 \leq s \leq t < 1$,
— $f^{*\prime\prime}([0,1]) = 0$.

En outre, si $\mathcal{C}_L = \{f : [0,1] \to \mathbb{R}^d, \mathscr{L}(f) \leq L\}$, on a
— $f^{*\prime\prime}(\{0\}) = f_r^{*\prime}(0)$,
— $f^{*\prime\prime}(\{1\}) = -f_\ell^{*\prime}(1)$.

De plus, il existe un unique $\lambda > 0$ et une variable aléatoire $\hat{t}$ à valeurs dans $[0,1]$, définie sur une extension $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{P})$ de l'espace de probabilité $(\Omega, \mathcal{F}, P)$, tels que
— $|X - f^*(\hat{t})| = d(X, \mathrm{Im} f^*)$ p.s.,
— pour toute fonction borélienne bornée $g : [0,1] \to \mathbb{R}^d$,

$$E\left[\langle X - f^*(\hat{t}), g(\hat{t})\rangle\right] = -\lambda \int_{[0,1]} \langle g(t), f^{*\prime\prime}(dt)\rangle.$$

En utilisant cette équation d'Euler-Lagrange, nous montrons ensuite qu'une courbe principale avec contrainte de longueur en dimension 2 n'a pas de point multiple.

Enfin, nous présentons deux classes d'exemples de courbes optimales. Nous étudions tout d'abord le problème de courbe principale pour la loi uniforme sur un épaississement d'une certaine courbe générative.

Pour un ensemble A, on note

$$A \oplus r = \left\{x \in \mathbb{R}^d \mid d(x, A) \leq r\right\}$$

l'épaississement de taille r de A. On considère une courbe générative $g : [0,1] \to \mathbb{R}^d$ et son épaississement de taille r, où r est inférieur au *reach* de $\mathrm{Im} g$. Le *reach* d'un

ensemble $A \subset \mathbb{R}^d$ est la borne supérieure des rayons ρ tels que tout point à distance au plus ρ de A a une unique projection sur A. Alors, l'image d'une courbe optimale de longueur $\mathscr{L}(g)$ pour la loi uniforme sur un épaississement de taille r de Img est Img elle-même.

Un autre exemple de courbe principale concerne la loi uniforme sur un cercle. Considérons le cercle unité centré en l'origine paramétré par

$$g(t) = (\cos(2\pi t), \sin(2\pi t))$$

pour $t \in [0,1]$. Soit U une variable aléatoire uniforme sur $[0,1]$ et soit $X = g(U)$. Alors, pour tout $L < 2\pi$, le cercle centré en l'origine de rayon $\dfrac{L}{2\pi}$ est l'unique courbe principale fermée de longueur L pour X.

I.1.2.3 Estimation

Soit $g : [0,1] \to \mathbb{R}^d$ une courbe rectifiable, telle que $\mathscr{L}(g) \leq \Lambda < \infty$, $|g'(t)| = \mathscr{L}(g)$ $dt-$a.e., et $\mathscr{L}(g) = \mathcal{H}^1(\text{Im}g)$. Pour $n \geq 1$, soient U_i^n, $i = 1, \ldots, n$, des variable aléatoires indépendantes à valeurs dans $[0,1]$, de support plein. On considère le modèle

$$X_i^n = g(U_i^n) + \varepsilon_i^n, \quad i = 1, \ldots, n,$$

où g est inconnue et on s'intéresse à l'estimation de l'image de g, en distance de Hausdorff. On suppose que le bruit est tel que $\frac{1}{n}\sum_{i=1}^n V(|\varepsilon_i^n|)$ converge vers 0 en probabilité lorsque n tend vers l'infini. Ainsi, dans ce modèle, le bruit n'est pas supposé borné.

Pour un vecteur aléatoire X tel que $E[V(|X|)] < \infty$, soit $f : [0,1] \to \mathbb{R}^d$ un minimiseur de

$$\Delta(f) = E\big[V(d(X, \text{Im}f))\big]$$

sur toutes les courbes de longueur au plus $L > 0$. Ici, $V : \mathbb{R}_+ \to \mathbb{R}_+$ est une fonction strictement croissante semi-continue inférieurement. Par exemple, on peut avoir $V(x) = x^p$, où $p > 0$, ou $V(x) = \frac{x}{1+x}$. Cette définition est une généralisation de la notion de courbe principale de Kégl et al. (2000). Une courbe principale empirique $\hat{f}_{n,L}$ associée à $X_1^n, \ldots, X_n^n$ peut être définie comme un minimiseur, de longueur au plus L, du critère

$$\Delta_n(f) = \frac{1}{n}\sum_{i=1}^n V(d(X_i^n, \text{Im}f)).$$

On choisit $\hat{f}_{n,L}$ L-lipschitzienne.

Nous proposons dans Delattre et Fischer (2021), présenté dans le Chapitre II.4, une méthode pour construire une suite de courbes principales empiriques généralisées, avec sélection de la longueur, telles qu'en distance de Hausdorff, leurs images convergent en probabilité vers l'image de g.

Plus précisément, soit $\hat{L}_n$ définie par

$$\hat{L}_n \in \underset{L \in a_n \mathbb{N} \cap [0, \Lambda_n \wedge \Lambda]}{\arg\min} \left[V(L) \mathcal{D} \left(\frac{1}{n} \sum_{i=1}^{n} \delta_{T(\hat{f}_{n,L}, X_i^n)}, \mathcal{M}_c \right) + \Delta_n(\hat{f}_{n,L}) \right].$$

Dans cette expression, $a_n > 0$ pour tout $n \geq 1$, $a_n \to 0$ lorsque $n \to +\infty$,

$$\Lambda_n = \inf\{L \geq 0, G_n(L) = 0\}, \text{ où } G_n(L) = \min_{\mathscr{L}(f) \leq L} \Delta_n(f),$$

et

$$T(f, x) = \max \underset{t \in [0,1]}{\arg\min} |x - f(t)|.$$

Alors, la distance de Hausdorff

$$d_H(\mathrm{Im} \hat{f}_{n, \hat{L}_n}, \mathrm{Im} g)$$

converge en probabilité vers 0 lorsque n tend vers l'infini.

I.1.2.4 Vitesse de convergence en apprentissage statistique

Soit X un vecteur aléatoire tel que $P(X \in K) = 1$, où K est un sous-ensemble compact de $\mathbb{R}^d$. On considère un échantillon $X_1, \ldots, X_n$ de X. On définit

$$\Delta(f) = E\left[d^2(X, \mathrm{Im} f)\right], \quad \Delta_n(f) = \frac{1}{n} \sum_{i=1}^{n} d^2(X_i, \mathrm{Im} f).$$

On note f une courbe optimale de longueur au plus L pour X et $\hat{f}_n$ une courbe optimale empirique, construite sur l'échantillon $X_1, \ldots, X_n$:

$$f^* \in \underset{g, \mathscr{L}(f) \leq L}{\arg\min} \Delta(f), \quad \hat{f}_n \in \underset{g, \mathscr{L}(f) \leq L}{\arg\min} \Delta_n(f).$$

L'objectif du Chapitre II.5 est d'étudier la vitesse de convergence vers 0 de $\Delta\big(\hat{f}_n\big) - \Delta(f^*)$. Jusqu'ici, la meilleure vitesse de convergence avait été obtenue par Kégl et al. (2000), qui ont construit une suite de courbes $\hat{f}_n$ telle que

$$\Delta(\hat{f}_n) - \Delta(f^*) = \mathcal{O}(n^{-1/3}).$$

Grâce à une approximation par des lignes polygonales plus fine, nous obtenons une vitesse de convergence améliorée $\mathcal{O}(n^{-2/5})$.

I.1.3 Classification non supervisée, segmentation, et déconvolution

I.1.3.1 Classification non supervisée avec des divergences de Bregman

Une partie de mon travail de thèse portait sur le problème de quantification d'une variable aléatoire à valeurs dans un espace de Banach séparable et réflexif, ainsi que sur la question connexe de classification non supervisée par la méthode des centres mobiles, en utilisant les divergences de Bregman comme mesures de proximité (Fischer, 2010). Les résultats sont exposés dans le Chapitre III.1 en dimension finie. C'est en effet le cadre dans lequel s'inscrit une extension présentée ci-dessous.

Si X désigne une variable aléatoire de loi μ, à valeurs dans un espace métrique $(\mathcal{E}, d)$, et $X_1, \ldots, X_n$ sont des variables aléatoires indépendantes de même loi que X, la quantification consiste à résumer X via un nombre fini k d'éléments de $(\mathcal{E}, d)$, les centres $\{c_1, \ldots, c_k\}$, en minimisant la distorsion

$$E\big[\min_{j=1,\ldots,k} d(X, c_j)\big].$$

Dans le cas où μ est inconnue, on peut effectuer une classification non supervisée des données $X_1, \ldots, X_n$ en minimisant la distorsion empirique

$$\frac{1}{n}\sum_{i=1}^{n} \min_{j=1,\ldots,k} d(X_i, c_j).$$

Une divergence de Bregman associée à la fonction strictement convexe ϕ est définie par

$$d_\phi(x, y) = \phi(x) - \phi(y) - \langle \nabla\phi(y), x - y\rangle,$$

où ∇ représente l'opérateur gradient. Cette classe de divergences, indexée par des fonctions strictement convexes, englobe un large éventail de fonctions distance, qui sont adaptées au partitionnement de mélanges de lois de la famille exponentielle. Nous établissons des conditions pour l'existence d'un ensemble de centres optimaux et démontrons la convergence presque sûre, pour la distorsion, d'un ensemble de centres empiriques optimaux vers l'optimum théorique. De plus, nous obtenons une vitesse de convergence $\frac{1}{\sqrt{n}}$.

Dans Brécheteau et al. (2021), résumé dans le Chapitre III.2, nous étudions une version plus sophistiquée de la classification non supervisée avec les divergences de Bregman, qui est robuste au bruit : étant donné un niveau de troncature $h \in (0, 1]$, nous recherchons à la fois des centres de classe optimaux et un sous-ensemble de données de masse au moins h pour μ. Nous étudions les propriétés théoriques de cette procédure avec troncature. Nous prouvons l'existence d'un ensemble de centres optimaux et la convergence presque sûre d'un ensemble de centres empiriques optimaux, et obtenons également une vitesse de convergence $\frac{1}{\sqrt{n}}$.

D'un point de vue pratique, nous développons une version de l'algorithme de Lloyd avec un paramètre de troncature, ainsi qu'une heuristique pour sélectionner ce paramètre et le nombre de classes à partir des données.

I.1.3.2 Une étude de détection de rupture

Dans Fischer et Picard (2020), sujet du Chapitre III.3, nous considérons l'estimation d'un point de rupture dans un modèle gaussien, pour des observations éventuellement de grande dimension, en utilisant une méthode de maximum de vraisemblance. Plus précisément, on observe des vecteurs aléatoires indépendants $Y_1, \dots, Y_n$ à valeurs dans $\mathbb{R}^d$, tels que

$$\begin{aligned} &Y_i = \theta_i + \eta_i, \quad \eta_i \sim \mathcal{N}(0, \sigma^2 I_d), \text{ indépendantes}, \quad 1 \leq i \leq n, \\ &\forall i \leq n\tau, \quad \theta_i = \theta^-, \\ &\forall i > n\tau, \quad \theta_i = \theta^+. \end{aligned}$$

Le but est d'estimer τ. L'estimateur du maximum de vraisemblance, également appelé dans ce cas estimateur CUSUM, est obtenu en minimisant en k

$$\sum_{i=1}^{k} \sum_{j=1}^{d} \left(Y_{i,j} - \frac{1}{k} \sum_{\ell=1}^{k} Y_{\ell,j} \right)^2 + \sum_{i=k+1}^{n} \sum_{j=1}^{d} \left(Y_{i,j} - \frac{1}{n-k} \sum_{\ell=k+1}^{n} Y_{\ell,j} \right)^2.$$

Nous nous intéressons à la manière dont une réduction de dimension peut affecter les performances de la méthode. Ainsi, nous considérons également le critère pour une dimension de projection $p \leq d$ au lieu de d. L'estimateur correspondant est noté $\hat{\tau}(p)$. Soient $\Delta^2 = \sum_{j=1}^{d} (\theta_j^- - \theta_j^+)^2$, $\Delta_p^2 = \sum_{j=1}^{p} (\theta_j^- - \theta_j^+)^2$ et $\Psi_n(p, \Delta_p) = \frac{\sigma^2}{n\Delta_p^2} \left(1 \vee \frac{\sigma^2 p}{n\Delta_p^2}\right)$. Alors, pour tout $\gamma > 0$, il existe des constantes $\kappa(\gamma, \varepsilon)$ et $c(\gamma, \varepsilon)$ telles que, si

$$\Delta_p^2 \geq c(\gamma, \varepsilon) \frac{\sigma^2 \ln(n)}{n},$$

alors

$$P\Big(|\hat{\tau}(p) - \tau| \geq \kappa(\gamma, \varepsilon) \ln(n) \Psi_n(p, \Delta_p)\Big) \leq c n^{-\gamma},$$

où c est une constante absolue. D'après Korostelev et Lepski (2008), la vitesse de convergence obtenue est minimax à un facteur logarithmique près. Elle est composée d'une vitesse rapide, qui ne dépend pas de la dimension, et d'une vitesse lente qui se détériore avec la dimension.

Ensuite, nous considérons le cas d'observations parcimonieuses, avec une régularité de type Sobolev. Pour $s > 0$, on définit

$$\Theta(s, L) = \left\{ \theta \in \mathbb{R}^d, \ \sup_{K \in \mathbb{N}^*} K^{2s} \sum_{k \geq K} (\theta_k)^2 \leq L^2 \right\},$$

et on suppose que les moyennes θ^- et θ^+ appartiennent à l'ensemble $\Theta(s, L)$.

Dans ce contexte, pour tout $\gamma > 0$, il existe des constantes $\kappa(\gamma, \varepsilon)$ et $c(\gamma, \varepsilon)$ telles que, si

$$\Delta^2 \geq \left[2c(\gamma, \varepsilon) \frac{\sigma^2 \ln(n)}{n} \vee 8L^2 p^{-2s} \right],$$

alors

$$P\Big(|\hat{\tau}(p) - \tau| \geq \kappa(\gamma, \varepsilon) \ln(n) \Psi_n(p, \Delta)\Big) \leq c n^{-\gamma}.$$

De plus, si $p_s = \left(\frac{8L^2 n}{\sigma^2}\right)^{\frac{1}{1+2s}}$, pour tout $\gamma > 0$, il existe des constantes $\kappa(\gamma, \varepsilon)$ et $c(\gamma, \varepsilon)$ telles que, si

$$\Delta^2 \geq \left[2c(\gamma, \varepsilon) \frac{\sigma^2 \ln(n)}{n} \vee \left(\frac{\sigma^2}{n}\right)^{\frac{2s}{1+2s}} \left(8L^2\right)^{\frac{1}{1+2s}} \right],$$

$$P\Big(|\hat{\tau}(p_s) - \tau| \geq \kappa(\gamma, \varepsilon) \frac{\sigma^2 \ln(n)}{n\Delta^2}\Big) \leq c n^{-\gamma}.$$

Cela montre qu'il existe un choix optimal de la dimension, conduisant à la vitesse rapide.

En pratique, le p_s optimal est inconnu. Nous proposons ainsi une procédure adaptative de réduction de la dimension basée sur la méthode de Lepski. Nous montrons que l'estimateur obtenu atteint la vitesse de convergence rapide. Enfin, nous proposons des stratégies pour effectuer la réduction de dimension en pratique.

I.1.3.3 Déconvolution Wasserstein en dimension 1

Lors de l'estimation d'une mesure par la mesure empirique associée, ou, dans un contexte bruité, par un estimateur par déconvolution, il peut être utile de contrôler la distance de Wasserstein entre les deux mesures, en particulier dans le contexte de l'inférence géométrique, en relation avec un outil appelé « distance à la mesure » (Chazal et al., 2011).

L'article Dedecker et al. (2015), repris dans le Chapitre III.4, traite de l'estimation d'une mesure de probabilité μ sur la droite réelle à partir de données observées avec un bruit additif. On observe $Y_i = X_i + \varepsilon_i$, où les X_i sont indépendantes, de loi de probabilité inconnue μ. Les variables aléatoires ε_i, $i = 1, \dots, n$, sont indépendantes et distribuées selon une mesure de probabilité connue μ_ε, non nécessairement symétrique. On suppose que $(X_1, \dots, X_n)$ est indépendant de $(\varepsilon_1, \dots, \varepsilon_n)$.

Nous nous intéressons à la vitesse de convergence pour la distance de Wasserstein d'ordre $p \geq 1$. La loi connue des erreurs est supposée appartenir à une classe de lois de probabilité ordinairement ou super régulières.

On note μ^* (respectivement f^*) la transformée de Fourier de la mesure de probabilité μ (respectivement de la fonction intégrable f), c'est-à-dire

$$\mu^*(x) = \int_{\mathbb{R}} e^{iux} \mu(du) \quad \text{et} \quad f^*(x) = \int_{\mathbb{R}} e^{iux} f(u) du.$$

Soit F la fonction de répartition de μ. On définit l'estimateur $\tilde{\mu}_n$ de la mesure μ comme la mesure de probabilité avec la fonction de répartition $\tilde{F}_n$, un estimateur de F construit en deux étapes. On construit tout d'abord un estimateur préliminaire $\hat{F}_n$ de F, en utilisant un noyau symétrique positif k avec une régularité appropriée :

$$\hat{F}_n(t) = \frac{1}{nh} \int_{-\infty}^{t} \sum_{k=1}^{n} \tilde{k}_h \left(\frac{u - Y_k}{h} \right) du,$$

où

$$\tilde{k}_h(x) = \frac{1}{2\pi} \int \frac{e^{iux} k^*(u)}{\mu_\varepsilon^*(-u/h)} du.$$

Observons que cet estimateur $\hat{F}_n$, basé sur l'estimateur de déconvolution standard à noyau de la densité $\tilde{k}_h$, introduit pour la première fois par Carroll et Hall (1988), n'est pas une fonction de répartition puisqu'il n'est pas nécessairement croissant. Par conséquent, on choisit pour l'estimateur $\tilde{F}_n$ un minimiseur approché, sur toutes les fonctions de répartition, de la quantité $\int_{\mathbb{R}} |x|^{p-1} |\hat{F}_n - G|(x)dx$: étant donné $\rho > 0$, $\tilde{F}_n$ est telle que, pour toute fonction de répartition G,

$$\int |x|^{p-1} |\hat{F}_n - \tilde{F}_n|(x)dx \leq \int |x|^{p-1} |\hat{F}_n - G|(x)dx + \rho\,.$$

Pour cet estimateur, nous obtenons une borne supérieure améliorée dans le cas ordinairement régulier et des conditions moins restrictives pour la borne existante dans le cas super régulier.

Soit $\rho \leq n^{-1/2}$. On pose $r_\varepsilon = 1/\mu_\varepsilon^*$. On fait les hypothèses suivantes :

$$\int_0^\infty |x|^{p-1} \sqrt{P(|Y| \geq x)} dx < \infty \text{ et } \sup_{t \in [-2,2]} |r_\varepsilon^{(m_0)}(t)| < \infty.$$

1. On suppose qu'il existe $\beta > 0$, $\tilde{\beta} \geq 0$, $\gamma > 0$ et $c > 0$, tels que pour tout $\ell \in \{0, 1, \ldots, m_1\}$ et tout $t \in \mathbb{R}$,

   $$|r_\varepsilon^{(\ell)}(t)| \leq c(1 + |t|)^{\tilde{\beta}} \exp(|t|^\beta / \gamma).$$

   Alors, si $h = (4/(\gamma \log n))^{1/\beta}$, il existe une constante $C > 0$ telle que

   $$E[W_p^p(\tilde{\mu}_n, \mu)] \leq C(\log n)^{-p/\beta}.$$

2. On suppose qu'il existe $\beta > 0$ et $c > 0$, tels que pour tout $\ell \in \{0, 1, \ldots, m_1\}$ et tout $t \in \mathbb{R}$,

   $$|r_\varepsilon^{(\ell)}(t)| \leq c(1 + |t|)^{\beta}.$$

   Alors, si $h = n^{-\frac{1}{2p + (2\beta - 1)_+}}$, il existe une constante $C > 0$ telle que

   $$E[W_p^p(\tilde{\mu}_n, \mu)] \leq C\psi_n\,,$$

   où

   $$\psi_n = \begin{cases} n^{-\frac{p}{2p+2\beta-1}} & \text{si } \beta > \frac{1}{2} \\ \sqrt{\frac{\log n}{n}} & \text{si } \beta = \frac{1}{2} \\ \frac{1}{\sqrt{n}} & \text{si } \beta < \frac{1}{2}. \end{cases}$$

Dans le cas ordinairement régulier, nous établissons également une borne inférieure. Soient $M > 0$, $q \geq 1$. On définit $\mathcal{D}(M, q)$ comme l'ensemble des mesures μ sur $\mathbb{R}$ telles que $\int |x|^q d\mu(x) \leq M$. Supposons qu'il existe $\beta > 0$ et $c > 0$, tels que pour tout $\ell \in \{0, 1, 2\}$ et tout $t \in \mathbb{R}$,

$$|\mu_\varepsilon^{*(\ell)}(t)| \leq c(1 + |t|)^{-\beta}.$$

Alors, il existe une constante $C > 0$ telle que, pour tout estimateur $\hat{\mu}$,

$$\liminf_{n \to +\infty} n^{\frac{p}{2\beta+1}} \sup_{\mu \in \mathcal{D}(M,q)} E[W_p^p(\hat{\mu}, \mu)] > C.$$

I.1.4 Agrégation d'estimateurs

L'intérêt pour l'agrégation de différentes procédures statistiques n'a cessé de croître ces dernières années. En effet, avec le nombre croissant de méthodes d'estimation disponibles, il semble naturel d'essayer de combiner plusieurs procédures, en recherchant la garantie d'une certaine optimalité. Dans de nombreuses méthodes d'agrégation, la prédiction pour une nouvelle observation x est calculée à l'aide d'une combinaison linéaire ou convexe d'une collection d'estimateurs. Les stratégies abordées dans ce document sont basées sur l'idée introduite en classification par Mojirsheibani (1999, 2000, 2002a,b). Il s'agit de combiner plusieurs classifieurs, en s'appuyant sur une notion de consensus. La méthode consiste à calculer la prédiction associée à une nouvelle observation en combinant les vraies étiquettes de certaines des données d'apprentissage, une donnée étant sélectionnée si les prédictions calculées pour ce point avec les différents estimateurs initiaux sont les mêmes que la prédiction pour la nouvelle observation à classer.

I.1.4.1 Agrégation consensuelle pour la régression

En partant de l'idée de Mojirsheibani (1999) en classification, nous proposons dans Biau et al. (2016), présenté dans le Chapitre IV.1, une approche connexe dans le contexte de la régression. Au lieu de demander des prédictions égales pour sélectionner une donnée, ce qui n'a pas de sens pour des sorties continues, on met la condition qu'elles soient proches, que l'écart ne dépasse pas un certain seuil.

Plus précisément, si $\mathbf{r} = (r_1, \ldots, r_M)$ désigne la collection d'estimateurs indivi-

duels, l'estimateur combiné est donné par

$$T_n\left(\mathbf{r}(x)\right) = \sum_{i=1}^{n} W_{n,i}(x) Y_i, \quad x \in \mathbb{R}^d,$$

où

$$W_{n,i}(x) = \frac{\mathbf{1}_{\bigcap_{m=1}^{M}\{|r_m(x)-r_m(X_i)|\leq\varepsilon\}}}{\sum_{j=1}^{n}\mathbf{1}_{\bigcap_{m=1}^{M}\{|r_m(x)-r_m(X_j)|\leq\varepsilon\}}}.$$

On montre que l'estimateur est asymptotiquement au moins aussi performant, au sens L^2, que la meilleure combinaison des estimateurs de la liste initiale, et donc, en particulier, que le meilleur parmi ces estimateurs.

Plus formellement, pour toute loi de (X, Y) avec $E[Y^2] < \infty$,

$$\begin{aligned} &E[|T_n(\mathbf{r}(X)) - r^\star(X)|]^2 \\ &\quad \leq E[|T_n(\mathbf{r}(X)) - T(\mathbf{r}(X))|]^2 + \inf_f E[|f(\mathbf{r}(X)) - r^\star(X)|]^2, \end{aligned}$$

où la borne inférieure est prise sur toutes les fonctions de carré intégrable de $\mathbf{r}(X)$. En particulier,

$$\begin{aligned} &E[|T_n(\mathbf{r}(X)) - r^\star(X)|^2] \\ &\quad \leq \min_{m=1,\ldots,M} E[|r_m(X) - r^\star(X)|]^2 + E|T_n(\mathbf{r}(X)) - T(\mathbf{r}(X))|^2. \end{aligned}$$

De plus, on obtient pour le terme $E[|T_n(\mathbf{r}(X)) - T(\mathbf{r}(X))|]^2$, qui représente le prix à payer pour l'agrégation, une vitesse de convergence de $n^{-\frac{2}{M+2}}$. Il s'agit de la vitesse non paramétrique habituelle, le nombre d'estimateurs initiaux jouant ici le rôle de la dimension.

I.1.4.2 Prise en compte des distances entre observations

Dans Fischer et Mougeot (2019), exposé dans le Chapitre IV.2, nous introduisons une nouvelle stratégie d'apprentissage, basée sur la même idée de consensus, impliquant également des informations de distance entre les entrées.

Dans le schéma original introduit dans Mojirsheibani (1999) et Biau et al. (2016), la condition d'égalité ou de proximité est en fait requise pour tous les estimateurs individuels, ce qui ne paraît pas très opportun s'il existe un mauvais estimateur initial. Mojirsheibani (2002a) note que cette règle peut sembler trop restrictive et

propose donc d'autoriser quelques désaccords (typiquement, un seul). Le classifieur obtenu demeure convergent à condition que le nombre de classifieurs ne présentant pas de désaccord tende vers l'infini. De même, dans Biau et al. (2016), nous notons que la contrainte d'unanimité peut être assouplie en pratique en demandant que la condition de distance pour conserver une observation soit vérifiée pour une certaine proportion γ des estimateurs au moins (par exemple, $\gamma = 80\%$). Ici, nous proposons une nouvelle approche, basée sur les distances entre les observations, qui vise également à réduire l'effet d'un éventuel mauvais estimateur initial. Adoptant un point de vue « noyau », nous proposons un estimateur combiné avec des poids construits en associant les distances entre les entrées avec les distances entre les prédictions provenant des estimateurs individuels.

Soit $K : \mathbb{R}^{d+p} \mapsto \mathbb{R}_+$ un noyau régulier, et soit $g : \mathbb{R}^d \times \mathbb{R}^p \to \mathbb{R}_+$ telle que $g(v_1, v_2) = K(v)$, où $v \in \mathbb{R}^{d+p}$ est la concaténation de $v_1 \in \mathbb{R}^d$ et $v_2 \in \mathbb{R}^p$.

On note $\mathbf{f} = (f_1, \ldots, f_p)$ la collection des estimateurs initiaux. En régression, l'estimateur combiné est défini par

$$T_n(x) = \frac{\sum_{i=1}^n Y_i g\left(\frac{X_i - x}{\alpha}, \frac{\mathbf{f}(X_i) - \mathbf{f}(x)}{\beta}\right)}{\sum_{i=1}^n g\left(\frac{X_i - x}{\alpha}, \frac{\mathbf{f}(X_i) - \mathbf{f}(x)}{\beta}\right)},$$

Pour la classification, le classifieur combiné est donné par

$$C_n(x) = \begin{cases} 0 & \text{si } \sum_{i=1}^n Y_i g\left(\frac{X_i - x}{\alpha}, \frac{\mathbf{f}(X_i) - \mathbf{f}(x)}{\beta}\right) \leq \sum_{i=1}^n (1 - Y_i) g\left(\frac{X_i - x}{\alpha}, \frac{\mathbf{f}(X_i) - \mathbf{f}(x)}{\beta}\right) \\ 1 & \text{sinon.} \end{cases}$$

Notre motivation pour introduire une telle stratégie est l'intuition que profiter de l'efficacité de l'idée de consensus de Mojirsheibani (1999) et Biau et al. (2016) sans pour autant laisser de côté les informations sur la proximité entre les entrées devrait permettre d'améliorer la prédiction, surtout en présence d'un estimateur initial peu performant. Nous prouvons la convergence de la procédure en classification et en régression. En pratique, pour appliquer cette méthode, il faut trouver les fenêtres α et β optimales. En particulier, grâce à un équilibre adéquat entre les deux termes intervenant dans les poids, la procédure reste relativement robuste lorsque les entrées sont de grande dimension : un plus grand poids est mis sur les sorties dans ce cas.

I.1.4.3 Classification non supervisée pour les modèles prédictifs

Dans Fischer et al. (2021), qui fait l'objet du Chapitre IV.3, nous nous intéressons aux problèmes d'apprentissage où les données d'entrée sont constituées de plusieurs groupes inconnus, liés à différents modèles prédictifs sous-jacents. Un tel cadre est particulièrement pertinent pour modéliser certains phénomènes physiques avec des transitions de phase, ou dans les situations où certaines informations sont manquantes en raison des lois sur la protection de la vie privée.

Nous proposons une procédure en trois étapes pour résoudre ce type de problème. La première étape consiste à appréhender la structure de groupe des données d'entrée, qui peut être caractérisée par plusieurs lois de probabilité. En utilisant plusieurs divergences de Bregman comme mesures de proximité pour le partitionnement, nous espérons récupérer précisément la structure sous-jacente pour une grande variété de lois possibles. Pour chaque partition obtenue, la deuxième étape consiste à ajuster un modèle prédictif spécifique basé sur les données contenues dans chaque classe. Ainsi, à ce stade, nous disposons d'un certain nombre de modèles, correspondant à différentes partitions de Bregman, chacun de ces modèles étant constitué de plusieurs sous-modèles simples ajustés sur un certain groupe de la partition. Enfin, le modèle global est calculé en agrégeant les modèles correspondant aux différentes partitions. Cette étape de combinaison est basée sur les méthodes présentées dans les Sections I.1.4.1 et I.1.4.2 ci-dessus.

Nous réalisons des expériences numériques sur différents jeux de données simulées et réelles, mettant en évidence la bonne performance de cette méthode en trois étapes dans différents problèmes prédictifs.

I.1.5 Collaborations interdisciplinaires

Au cours de ces dernières années, j'ai eu l'occasion de m'investir dans plusieurs collaborations interdisciplinaires. La majeure partie d'entre elles porte sur les thèmes de l'énergie ou du climat.

I.1.5.1 Modélisation de la production d'énergie électrique

Dans l'article Fischer et al. (2017), détaillé dans le Chapitre V.1, nous nous intéressons à la modélisation de l'énergie éolienne en utilisant des techniques d'apprentissage statistique. Nous disposons de données réelles mesurées sur des éoliennes, fournies par la société d'énergie éolienne Maïa Eolis (aujourd'hui Engie Green). Les observations proviennent de 3 parcs différents, de 4 à 6 turbines, situés dans le Nord et l'Est de la France, pour la période de 2011 à 2014. Dans un parc, chaque éolienne fournit, avec un pas de temps de 10 minutes, des mesures de la puissance électrique, de la vitesse et de la direction du vent, de la température, ainsi qu'un indicateur de l'état de fonctionnement de l'éolienne. La puissance électrique de l'ensemble du parc est également fournie avec le même pas de temps.

Nous modélisons la puissance électrique en fonction des données de vent et de température, en testant plusieurs algorithmes, paramétriques ou non paramétriques. Nous obtenons des résultats particulièrement stables et satisfaisants en agrégeant plusieurs arbres de décision au moyen d'une procédure *bagging*. Comme première étape vers la prévision, nous quantifions également l'impact de l'utilisation de données moyennées au lieu de mesures locales : l'objectif est d'imiter les données plus globales fournies par un météorologue.

I.1.5.2 Réduction d'échelle pour la vitesse du vent

Les Chapitres V.2 et V.3, correspondant respectivement aux articles Alonzo et al. (2018) et Goutham et al. (2021), sont le résultat d'une collaboration avec Riwal Plougonven, professeur au Laboratoire de Météorologie Dynamique (LMD) à l'École Polytechnique. Nous avons eu l'occasion d'encadrer ensemble à deux reprises des projets de troisième année d'étudiants de l'École Polytechnique. Les recherches initiées lors de ces projets ont ensuite été poursuivies dans le cadre de la thèse de Bastien Alonzo (actuellement chercheur post-doctoral chez Météo France), et du stage de Master 1 de Naveen Goutham (actuellement doctorant au LMD sous la direction de Riwal Plougonven). Nos collaborations portent sur une question de réduction d'échelle, c'est-à-dire l'estimation d'une quantité locale en se basant sur des observations réelles à l'endroit considéré et sur des sorties de modèle numérique de prévision météorologique à plus grande échelle. La variable d'intérêt est la vitesse du vent, à 10m et 100m.

Dans Alonzo et al. (2018), nous comparons les performances de deux méthodes statistiques de réduction d'échelle, respectivement basées sur la régression linéaire et

les forêts aléatoires, pour la reconstruction et la prévision de la vitesse du vent dans un lieu précis, à savoir la plateforme d'observation du SIRTA (Site Instrumental de Recherche par Télédétection Atmosphérique) située sur le plateau de Saclay. On utilise des sorties de modèles du Centre Européen pour les Prévisions Météorologiques à Moyen Terme (CEPMMT). Les données provenant du CEPMMT pour la vitesse du vent à 10m montrent un biais systématique, alors qu'à 100m, la vitesse du vent est mieux représentée. Notre étude montre que les méthodes d'apprentissage statistique paramétriques et non paramétriques conduisent à des résultats comparables. Un modèle linéaire associé à un pré-traitement et à une sélection de variables judicieuse montre des performances légèrement supérieures à celles des forêts aléatoires. Néanmoins, ces dernières constituent une option intéressante minimisant le temps nécessaire au pré-traitement et au calibrage des modèles.

Dans Goutham et al. (2021), cette méthodologie est étendue à plus de 150 stations réparties sur les territoires de France métropolitaine et de Corse, correspondant à plusieurs contextes géographiques. Différentes méthodes d'apprentissage statistique sont testées : régression linéaire, k plus proches voisins, forêts aléatoires, *boosting*. Le modèle basé sur les forêts aléatoires est ensuite exploré plus avant afin de réduire la liste des variables explicatives. La valeur ajoutée de l'utilisation des méthodes statistiques est indéniable, et on constate par exemple que l'amélioration est plus significative pour les stations côtières, où les erreurs des données fournies par le CEPMMT sont les plus importantes.

I.1.6 Autres collaborations appliquées

J'ai également eu l'opportunité de participer à d'autres collaborations interdisciplinaires, résumées dans le Chapitre V.4, notamment en m'impliquant dans des activités d'encadrement doctoral. En particulier, j'ai pris part à l'encadrement d'une thèse en informatique, appliquée à la biologie pour la cancérologie, ainsi qu'à une thèse en linguistique, plus précisément en phonétique. Par ailleurs, j'ai participé à des travaux de recherche en astrophysique nécessitant des outils d'apprentissage statistique, dans le cadre d'une collaboration avec le CEA.

Chapter I.2

Summary of research directions

I.2.1 Introduction

My research interests focus on statistical learning, nonparametric statistics, and probability theory. One part of my contributions is essentially theoretical, whereas another part concerns applications to practical problems, in the frame of industrial or interdisciplinary collaborations.

Supervised and unsupervised learning. Statistical learning may be divided in two branches, supervised learning and unsupervised learning. In unsupervised learning, we observe $X_1, \ldots, X_n$, supposed to be independent, with the same distribution as a generic random variable X, and the goal is to learn some features of the underlying structure of X. In supervised learning, however, we have at hand $(X_1, Y_1), \ldots, (X_n, Y_n)$, independent copies of a generic pair (X, Y), and the aim is to learn the relationship between the input X and the output Y, based on the observations, to be able to predict the output associated with a new input.

Problems considered. Here, the unsupervised scheme in the broad sense corresponds to cluster analysis as well as some questions related to principal curves. A deconvolution study is also attached to this framework. Besides, the supervised problems of classification and regression investigated in the present document are all linked to aggregation strategies, and / or to real-life questions. Apart from statistical learning, principal curves are also considered from a probabilistic point of view, and then in a curve estimation context.

Minimax optimality. Sometimes, in the sequel, when discussing the performance

of an estimator, we will refer to optimality in the minimax sense. In a general framework where $s \in \mathcal{S}$ denotes an unknown quantity to be estimated and the performance of an estimator is measured via a risk R, an estimator S^* is said to be minimax if it satisfies

$$R(S^*, s) = \inf_T \sup_{s \in \mathcal{S}} R(T, s),$$

where the lower bound is taken over all possible estimators T of s. In this context, suppose that an estimator S is such that $R(S, s) \leq C\psi_n$ for every $s \in \mathcal{S}$, with $(\psi_n)_{n \geq 1}$ a sequence converging to 0 and $C > 0$ a constant. If, on the other hand, we can find s such that the best possible estimator does not converge at a faster rate than ψ_n, that is

$$\inf_T \sup_{s \in \mathcal{S}} R(T, s) \geq c\psi_n,$$

where $c > 0$ is a constant, this rate ψ_n is called the minimax rate of convergence, and the estimator S is thus optimal in the minimax sense.

Throughout the document, unless otherwise stated, $\mathbb{R}^d$ is equipped with the Euclidean norm $|\cdot|$.

I.2.2 Principal curves

We are interested in regularity properties as well as estimation capabilities of constrained principal curves, as introduced in Kégl et al. (2000). According to their definition, a principal curve f for a square integrable random variable X is an $\mathbb{R}^d$-valued function minimizing a criterion $E\left[d^2(X, \mathrm{Im}f)\right]$ under some constraint, typically on the length of the curve. Here, $d(\cdot, \cdot)$ stand for the Euclidean distance in $\mathbb{R}^d$ and $\mathrm{Im}f \subset \mathbb{R}^d$ is the range of f. Principal curves optimization also corresponds to a version of the "average-distance problem" studied in the calculus of variation and shape optimization community (Buttazzo and Stepanov, 2003; Buttazzo et al., 2002). The empirical version of the criterion, for a set of observations $X_1, \ldots, X_n$, is $\frac{1}{n}\sum_{i=1}^n d^2(X_i, \mathrm{Im}f)$. In the sequel, we also consider more general distance measures.

I.2.2.1 Model selection

A first research direction on principal curves, pertaining to model selection tasks, developed in Biau and Fischer (2012) and Fischer (2013), is presented in Chapter II.2.

The construction of an empirical principal curve relies indeed on some regularity parameters (number of segments, length, turn...) which drive the exact shape of the curve. Intuitively, the aim is to recover the underlying structure of the data, without interpolating. Considering two different models, in estimation and in statistical learning respectively, we perform parameter selection based on model selection via penalization (Massart, 2007). A theoretical guarantee on the selected estimator is provided by means of a so-called oracle inequality, expressing that its performance is close to the best possible over the considered collection of models.

In Fischer (2013), we focus on length selection in estimation via principal curves in a Gaussian model. More specifically, we observe random vectors $X_1, \dots, X_n$ with values in $\mathbb{R}^d$ such that

$$X_i = x_i^* + \sigma \xi_i, \quad i = 1, \dots, n,$$

where the x_i^* are unknown, the ξ_i are independent standard Gaussian vectors of $\mathbb{R}^d$ and $\sigma > 0$ stands for the noise level, which is supposed known.

Let $\{w_\ell\}_{\ell \in \mathcal{L}}$ be a summable family of weights. Assume that $(\hat{x}_{1,\ell}, \dots, \hat{x}_{n,\ell})$ minimizes $\frac{1}{n}\sum_{i=1}^n |X_i - x_i|^2$ among all $(x_1, \dots, x_n) \in (\mathrm{Im} f)^n$ where f belongs to some class $\mathcal{F}_\ell$ of curves with length ℓ and fixed endpoints. Then, for σ below a certain threshold, if

$$\mathrm{pen}(\ell) \geq \eta \sigma^2 \left[c_1 \left(\ln \left(\frac{\ell^{1/d} \lambda^{1-1/d}}{\sigma} \right) + c_2 \right) + \frac{4 w_\ell}{nd} \right],$$

where λ depends on ℓ and the distance between the endpoints, then, almost surely, there exists a minimizer $\hat{\ell}$ of the penalized criterion

$$\mathrm{crit}(\ell) = \frac{1}{nd} \sum_{i=1}^n |X_i - \hat{x}_{i\ell}|^2 + \mathrm{pen}(\ell).$$

Moreover, if $\tilde{x}_i = \hat{x}_{i\hat{\ell}}$ for all $i = 1, \dots, n$, we have

$$\frac{1}{nd} \sum_{i=1}^n E|\tilde{x}_i - x_i^*|^2 \leq c(\eta) \Big[\inf_\ell \Big\{ \inf_{f \in \mathcal{F}_\ell} \frac{1}{nd} \sum_{i=1}^n d^2(x_i^*, \mathrm{Im} f) + \mathrm{pen}(\ell) \Big\} + \frac{\sigma^2}{nd} \Big].$$

In Biau and Fischer (2012), the parameter selection issue is addressed in a statistical learning framework. The estimators are polygonal lines indexed by their number of edges and their length or turn. We present here the model selection result for the length.

Let $X_1, \ldots, X_n$ denote a sample from a generic random vector X satisfying $P(X \in K) = 1$, where K is a convex compact subset of $\mathbb{R}^d$, with diameter δ. We set

$$\Delta(f) = E\left[d^2(X, \mathrm{Im}f)\right], \quad \Delta_n(f) = \frac{1}{n}\sum_{i=1}^{n} d^2(X_i, \mathrm{Im}f),$$

and $\mathcal{D}(f, g) = \Delta(g) - \Delta(f)$. Let f^* denote a minimizer of $\Delta(f)$.

We suppose that $\hat{f}_{k,\ell}$ minimizes the empirical criterion $\Delta_n(f)$ over some class $\mathcal{F}_{k,\ell}$ of polygonal lines with k segments and length at most ℓ. Let $\{x_{k,\ell}\}_{k,\ell}$ denote a summable family of weights. There exist $c, c_0, \ldots, c_2$, such that, if

$$\mathrm{pen}(k, \ell) \geq \frac{1}{\sqrt{n}}\left[c_1\sqrt{k} + c_2\ell + c_0\right] + \delta^2\sqrt{\frac{x_{k,\ell}}{2n}},$$

then letting $\tilde{f}$ denote the polygonal line selected by minimizing $crit(k, \ell) = \Delta_n(\hat{f}_{k,\ell}) + \mathrm{pen}(k, \ell)$, we have

$$E[\mathcal{D}(f^*, \tilde{f})] \leq \inf_{k,\ell}\left[\inf_{f \in \mathcal{F}_{k,\ell}} \mathcal{D}(f^*, f) + \mathrm{pen}(k, \ell)\right] + \frac{\delta^2 c}{\sqrt{n}}.$$

I.2.2.2 Regularity

In Delattre and Fischer (2020), exposed in Chapter II.3, we investigate the theoretical properties satisfied by a principal curve $f^* : [0, 1] \to \mathbb{R}^d$ with length at most L, associated to a probability distribution with second-order moment. We suppose that the probability distribution is not supported on the image of a curve with length L. Studying open (with endpoints) as well as closed (without endpoints) optimal curves, we show that they have finite curvature and derive an Euler-Lagrange equation.

More specifically, setting

$$\Delta(f) = E\left[d^2(X, \mathrm{Im}f)\right],$$

we define, for $L \geq 0$,

$$G(L) = \min\{\Delta(f), f \in \mathcal{C}_L\},$$

where $\mathcal{C}_L$ will denote either one of the following sets of curves:

$$\{f \in [0, 1] \to \mathbb{R}^d, \mathscr{L}(f) \leq L\},$$
$$\{f \in [0, 1] \to \mathbb{R}^d, \mathscr{L}(f) \leq L, f(0) = f(1)\}.$$

Let $L > 0$ such that $G(L) > 0$ and let $f^* \in \mathcal{C}_L$ such that $\Delta(f^*) = G(L)$. Then, the length of f^* equals L. Assuming that f^* is L-Lipschitz, we obtain that

— f^* is right-differentiable on $[0,1)$, $|f_r^{*\prime}(t)| = L$ for all $t \in [0,1)$,
— f^* is left-differentiable on $(0,1]$, $|f_\ell^{*\prime}(t)| = L$ for all $t \in (0,1]$,

and there exists a unique signed measure $f^{*\prime\prime}$ on $[0,1]$ (with values in $\mathbb{R}^d$) such that
— $f^{*\prime\prime}((s,t]) = f_r^{*\prime}(t) - f_r^{*\prime}(s)$ for all $0 \leq s \leq t < 1$,
— $f^{*\prime\prime}([0,1]) = 0$.
In the case $\mathcal{C}_L = \{f : [0,1] \to \mathbb{R}^d, \mathscr{L}(f) \leq L\}$, we also have
— $f^{*\prime\prime}(\{0\}) = f_r^{*\prime}(0)$,
— $f^{*\prime\prime}(\{1\}) = -f_\ell^{*\prime}(1)$.

Moreover, there exists a unique $\lambda > 0$ and there exists a random variable $\hat{t}$ with values in $[0,1]$, defined on an extension $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{P})$ of the probability space $(\Omega, \mathcal{F}, P)$, such that
— $|X - f^*(\hat{t})| = d(X, \mathrm{Im} f^*)$ a.s.,
— for every bounded Borel function $g : [0,1] \to \mathbb{R}^d$,

$$E\left[\langle X - f^*(\hat{t}), g(\hat{t})\rangle\right] = -\lambda \int_{[0,1]} \langle g(t), f^{*\prime\prime}(dt)\rangle.$$

Using this Euler-Lagrange equation, we then show that a length-constrained principal curve in two dimensions has no multiple point.

Finally, two classes of examples of optimal curves are presented. First, we study the principal curve problem for the uniform distribution on an enlargement of some generative curve. For a set A, we denote by

$$A \oplus r = \left\{x \in \mathbb{R}^d \mid d(x, A) \leq r\right\}$$

the r-enlargement of A. We consider a generative curve $g : [0,1] \to \mathbb{R}^d$ and its r-enlargement , where r does not exceed the reach of Img. Here, the reach of a set $A \subset \mathbb{R}^d$ is the supremum of the radii ρ such that every point at distance at most ρ of A has a unique projection on A. The image of an optimal curve with length $\mathscr{L}(g)$ for the uniform distribution on an r-enlargement of Img is Img itself.

Another example is about the uniform distribution on a circle. Consider the unit circle centered at the origin with parameterization given by

$$g(t) = (\cos(2\pi t), \sin(2\pi t))$$

for $t \in [0,1]$. Let U be a uniform random variable on $[0,1]$ and let $X = g(U)$. Then, for every $L < 2\pi$, the circle centered at the origin with radius $\dfrac{L}{2\pi}$ is the unique closed principal curve with length L for X.

I.2.2.3 Estimation

Let $g : [0,1] \to \mathbb{R}^d$ be a rectifiable curve, satisfying $\mathscr{L}(g) \leq \Lambda < \infty$, $|g'(t)| = \mathscr{L}(g)\ dt-$a.e., and $\mathscr{L}(g) = \mathcal{H}^1(\mathrm{Im}g)$. For $n \geq 1$, let U_i^n, $i = 1, \ldots, n$, denote independent random variables taking their values in $[0,1]$, with full support. Consider the model

$$X_i^n = g(U_i^n) + \varepsilon_i^n, \quad i = 1, \ldots, n,$$

where g is unknown and we are interested in the estimation of the image of g, in Hausdorff distance. The noise is supposed to be such that $\frac{1}{n}\sum_{i=1}^n V(|\varepsilon_i^n|)$ tends to 0 in probability as n tends to infinity. Let us stress that the noise is not assumed to be bounded.

Given a random vector X such that $E[V(|X|)] < \infty$, let $f : [0,1] \to \mathbb{R}^d$ be a minimizer of

$$\Delta(f) = E\big[V(d(X, \mathrm{Im}f))\big]$$

over all curves with length not greater than a certain threshold. Here, $V : \mathbb{R}_+ \to \mathbb{R}_+$ is a lower semi-continuous strictly increasing function. For instance, $V(x) = x^p$, where $p > 0$, or $V(x) = \frac{x}{1+x}$. This definition is a generalization of the principal curve notion of Kégl et al. (2000). An empirical principal curve $\hat{f}_{n,L}$ associated to $X_1^n, \ldots, X_n^n$ may be defined as a minimizer, with length at most L, of the criterion

$$\Delta_n(f) = \frac{1}{n}\sum_{i=1}^n V(d(X_i^n, \mathrm{Im}f)).$$

We choose $\hat{f}_{n,L}$ L-Lipschitz.

In Delattre and Fischer (2021), presented in Chapter II.4, we propose a method to build a sequence of generalized empirical principal curves, with selected length, so that, in Hausdorff distance, the images of the estimating principal curves converge in probability to the image of g.

More specifically, let $\hat{L}_n$ be defined by

$$\hat{L}_n \in \underset{L \in a_n\mathbb{N} \cap [0, \Lambda_n \wedge \Lambda]}{\arg\min} \left[V(L)\mathcal{D}\left(\frac{1}{n}\sum_{i=1}^n \delta_{T(\hat{f}_{n,L}, X_i^n)}, \mathcal{M}_c\right) + \Delta_n(\hat{f}_{n,L}) \right].$$

In this expression, $a_n > 0$ for every $n \geq 1$, $a_n \to 0$ as $n \to \infty$,

$$\Lambda_n = \inf\{L \geq 0, G_n(L) = 0\}, \text{ where } G_n(L) = \min_{\mathscr{L}(f) \leq L} \Delta_n(f),$$

and

$$T(f, x) = \max \arg\min_{t \in [0,1]} |x - f(t)|.$$

Then, the Hausdorff distance

$$d_H(\mathrm{Im}\hat{f}_{n,\hat{L}_n}, \mathrm{Im}g)$$

converges in probability to 0 as n tends to infinity.

I.2.2.4 Rate of convergence in statistical learning

Let X be a random vector such that $P(X \in K) = 1$, where K is a compact subset of $\mathbb{R}^d$, and let $X_1, \ldots, X_n$ denote independent realizations of X. We define

$$\Delta(f) = E\left[d^2(X, \mathrm{Im}f)\right], \quad \Delta_n(f) = \frac{1}{n}\sum_{i=1}^{n} d^2(X_i, \mathrm{Im}f),$$

We denote by f an optimal curve with length at most L for X and by $\hat{f}_n$ an empirical counterpart, built on the sample $X_1, \ldots, X_n$:

$$f^* \in \arg\min_{g, \mathscr{L}(f) \leq L} \Delta(f), \quad \hat{f}_n \in \arg\min_{g, \mathscr{L}(f) \leq L} \Delta_n(f).$$

The purpose of Chapter II.5 is to study the rate of convergence to 0 of $\Delta\big(\hat{f}_n\big) - \Delta(f^*)$. So far, the best rate of convergence has been obtained by Kégl et al. (2000), who constructed a sequence of curves $\hat{f}_n$, such that

$$\Delta(\hat{f}_n) - \Delta(f^*) = \mathcal{O}(n^{-1/3}).$$

Thanks to a refined approximation by piecewise linear curves, we obtain an improved rate of convergence, that is $\mathcal{O}(n^{-2/5})$.

I.2.3 Cluster analysis, segmentation, and deconvolution

I.2.3.1 Bregman clustering

A part of my PhD research work dealt with the quantization problem of a random variable with values in a separable and reflexive Banach space, and with the related k-means clustering issue, using Bregman divergences as proximity measures (Fischer, 2010). The results in finite dimension are reported in Chapter III.1. Indeed, the extension presented below pertains to the finite-dimensional framework. If X denotes a random variable with distribution μ, taking its values in a metric space $(\mathcal{E}, d)$, and $X_1, \ldots, X_n$ are independent realizations of X, quantization consists in summarizing X through a finite number k of elements in $(\mathcal{E}, d)$, a so-called codebook $\{c_1, \ldots, c_k\}$, by minimizing the distortion

$$E\big[\min_{j=1,\ldots,k} d(X, c_j)\big].$$

The associated clustering task for the data $X_1, \ldots, X_n$, in the case where μ is unknown, relies on the minimization of the empirical distortion

$$\frac{1}{n}\sum_{i=1}^{n} \min_{j=1,\ldots,k} d(X_i, c_j).$$

A Bregman divergence associated to the strictly convex function ϕ is defined by

$$d_\phi(x, y) = \phi(x) - \phi(y) - \langle \nabla\phi(y), x - y\rangle,$$

where ∇ stands for the gradient operator. This class of divergences, indexed by strictly convex functions, encompasses a wide range of distance-like functions, which are well-suited to perform clustering of exponential families mixtures. We provide conditions for the existence of an optimal codebook and show almost sure convergence, in terms of distortion, of an empirically optimal codebook to the theoretical optimum. Moreover, we obtain a rate of convergence of $\frac{1}{\sqrt{n}}$. In Brécheteau et al. (2021), summarized in Chapter III.2, we study a more sophisticated version of clustering with Bregman divergences, which is robust to noise thanks to a trimming approach: given a trim level $h \in (0, 1]$, we search for optimal cluster centers along with a subset of the data with mass at least h for μ. We investigate the theoretical properties of this trimmed k-means procedure. We prove the existence of an optimal codebook and almost sure convergence of an empirically optimal codebook. We obtain again a rate of convergence of $\frac{1}{\sqrt{n}}$.

From a practical point of view, we derive an associated Lloyd-type algorithm with a trimming parameter, along with a heuristic to select this parameter and the number of clusters from the sample.

I.2.3.2 A change-point study

In Fischer and Picard (2020), subject of Chapter III.3, we consider the estimation of a change-point for possibly high-dimensional data in a Gaussian model, using a maximum likelihood method. More precisely, we observe independent random vectors $Y_1, \dots, Y_n$ with values in $\mathbb{R}^d$, such that

$$\begin{aligned} &Y_i = \theta_i + \eta_i, \quad \eta_i \sim \mathcal{N}(0, \sigma^2 I_d), \text{ independent}, \quad 1 \le i \le n, \\ &\forall i \le n\tau, \quad \theta_i = \theta^-, \\ &\forall i > n\tau, \quad \theta_i = \theta^+. \end{aligned}$$

The goal is to estimate τ. The maximum likelihood estimator, also called in this case CUSUM estimator, is obtained by minimizing in k

$$\sum_{i=1}^{k}\sum_{j=1}^{d}\Big(Y_{i,j} - \frac{1}{k}\sum_{\ell=1}^{k} Y_{\ell,j}\Big)^2 + \sum_{i=k+1}^{n}\sum_{j=1}^{d}\Big(Y_{i,j} - \frac{1}{n-k}\sum_{\ell=k+1}^{n} Y_{\ell,j}\Big)^2.$$

We are interested in how dimension reduction can affect the performance of the method, and thus, we also consider this criterion for a projection dimension $p \le d$ instead of d. The corresponding estimator is called $\hat{\tau}(p)$. Let

$$\Delta^2 = \sum_{j=1}^{d}(\theta_j^- - \theta_j^+)^2, \qquad \Delta_p^2 = \sum_{j=1}^{p}(\theta_j^- - \theta_j^+)^2,$$

and

$$\Psi_n(p, \Delta_p) = \frac{\sigma^2}{n\Delta_p^2}\left(1 \vee \frac{\sigma^2 p}{n\Delta_p^2}\right).$$

Then, for any $\gamma > 0$, there exist constants $\kappa(\gamma, \varepsilon)$ and $c(\gamma, \varepsilon)$ such that, if

$$\Delta_p^2 \ge c(\gamma, \varepsilon)\frac{\sigma^2 \ln(n)}{n},$$

then

$$P\Big(|\hat{\tau}(p) - \tau| \ge \kappa(\gamma, \varepsilon)\ln(n)\Psi_n(p, \Delta_p)\Big) \le cn^{-\gamma},$$

where c is an absolute constant. The rate of convergence is minimax up to a logarithmic factor, according to Korostelev and Lepski (2008). It is composed of a fast rate, dimension-invariant, and a slow rate deteriorating with the dimension.

Then, we consider the case of sparse data, with a Sobolev regularity. For $s > 0$, we define

$$\Theta(s, L) = \left\{ \theta \in \mathbb{R}^d, \ \sup_{K \in \mathbb{N}^*} K^{2s} \sum_{k \geq K} (\theta_k)^2 \leq L^2 \right\},$$

and we suppose that the means θ^- and θ^+ are in $\Theta(s, L)$.

In this context, for any $\gamma > 0$, there exist constants $\kappa(\gamma, \varepsilon)$ and $c(\gamma, \varepsilon)$ such that, if

$$\Delta^2 \geq \left[2c(\gamma, \varepsilon) \frac{\sigma^2 \ln(n)}{n} \vee 8L^2 p^{-2s} \right],$$

then

$$P\Big(|\hat{\tau}(p) - \tau| \geq \kappa(\gamma, \varepsilon) \ln(n) \Psi_n(p, \Delta) \Big) \leq cn^{-\gamma}.$$

Moreover, setting $p_s = \left(\frac{8L^2 n}{\sigma^2} \right)^{\frac{1}{1+2s}}$, for any $\gamma > 0$, there exist constants $\kappa(\gamma, \varepsilon)$ and $c(\gamma, \varepsilon)$ such that, if

$$\Delta^2 \geq \left[2c(\gamma, \varepsilon) \frac{\sigma^2 \ln(n)}{n} \vee \left(\frac{\sigma^2}{n} \right)^{\frac{2s}{1+2s}} \left(8L^2 \right)^{\frac{1}{1+2s}} \right],$$

$$P\Big(|\hat{\tau}(p_s) - \tau| \geq \kappa(\gamma, \varepsilon) \frac{\sigma^2 \ln(n)}{n\Delta^2} \Big) \leq cn^{-\gamma}.$$

This shows that there exists an optimal choice of the dimension, leading to the fast rate of estimation.

In practice, the optimal p_s is unknown. We thus propose an adaptive dimension reduction procedure based on Lepski's method. It is shown that the resulting estimator attains the fast rate of convergence. Finally, practical strategies to perform dimension reduction are suggested.

I.2.3.3 Wasserstein deconvolution in dimension 1

When estimating a measure by the associated empirical measure, or, in a noisy context, by a deconvolution estimator, it may be useful to control the Wasserstein

distance between the two measures, particularly in the context of geometric inference, in connection with a tool called "distance to measure" (Chazal et al., 2011).

In Dedecker et al. (2015), discussed in Chapter III.4, we deal with the estimation of a probability measure μ on the real line from data observed with an additive noise. We observe $Y_i = X_i + \varepsilon_i$, where the X_i's are independent and identically distributed according to an unknown probability measure μ. The random variables ε_i, $i = 1, \ldots, n$, are independent and identically distributed according to a known probability measure μ_ε, not necessarily symmetric. We assume that $(X_1, \ldots, X_n)$ is independent of $(\varepsilon_1, \ldots, \varepsilon_n)$.

We are interested in rates of convergence for the Wasserstein metric of order $p \geq 1$. The known distribution of the errors is assumed to belong to a class of supersmooth or ordinary smooth distributions.

We denote by μ^* (respectively f^*) the Fourier transform of the probability measure μ (respectively of the integrable function f), that is

$$\mu^*(x) = \int_{\mathbb{R}} e^{iux} \mu(du) \quad \text{and} \quad f^*(x) = \int_{\mathbb{R}} e^{iux} f(u) du.$$

Let F be the cumulative distribution function of μ. We define the estimator $\tilde{\mu}_n$ of the measure μ as the probability measure with distribution function $\tilde{F}_n$, which is an estimator of F built in two steps. We first build a preliminary estimator $\hat{F}_n$ of F, using a symmetric nonnegative kernel k with appropriate smoothness:

$$\hat{F}_n(t) = \frac{1}{nh} \int_{-\infty}^{t} \sum_{k=1}^{n} \tilde{k}_h \left(\frac{u - Y_k}{h} \right) du,$$

where

$$\tilde{k}_h(x) = \frac{1}{2\pi} \int \frac{e^{iux} k^*(u)}{\mu_\varepsilon^*(-u/h)} du.$$

Observe that this estimator $\hat{F}_n$, based on the standard deconvolution kernel density estimator $\tilde{k}_h$ first introduced by Carroll and Hall (1988), is not a cumulative distribution function since it is not necessarily non-decreasing. Therefore, we choose the estimator $\tilde{F}_n$ as an approximate minimizer over all distribution functions of the quantity $\int_{\mathbb{R}} |x|^{p-1} |\hat{F}_n - G|(x) dx$: given $\rho > 0$, let $\tilde{F}_n$ be such that, for every distribution function G,

$$\int |x|^{p-1} |\hat{F}_n - \tilde{F}_n|(x) dx \leq \int |x|^{p-1} |\hat{F}_n - G|(x) dx + \rho \,.$$

For this estimator, we obtain an improved upper bound in the ordinary smooth case and less restrictive conditions for the existing bound in the supersmooth one. Let $\rho \leq n^{-1/2}$. We set $r_\varepsilon = 1/\mu_\varepsilon^*$. Assume that

$$\int_0^\infty |x|^{p-1}\sqrt{P(|Y| \geq x)}dx < \infty \text{ and } \sup_{t\in[-2,2]} |r_\varepsilon^{(m_0)}(t)| < \infty.$$

1. Assume that there exist $\beta > 0$, $\tilde{\beta} \geq 0$, $\gamma > 0$ and $c > 0$, such that for every $\ell \in \{0, 1, \ldots, m_1\}$ and every $t \in \mathbb{R}$,

   $$|r_\varepsilon^{(\ell)}(t)| \leq c(1 + |t|)^{\tilde{\beta}} \exp(|t|^\beta/\gamma).$$

   Then, taking $h = (4/(\gamma \log n))^{1/\beta}$, there exists $C > 0$ such that

   $$E[W_p^p(\tilde{\mu}_n, \mu)] \leq C(\log n)^{-p/\beta}.$$

2. Assume that there exist $\beta > 0$ and $c > 0$, such that for every $\ell \in \{0, 1, \ldots, m_1\}$ and every $t \in \mathbb{R}$,

   $$|r_\varepsilon^{(\ell)}(t)| \leq c(1 + |t|)^{\beta}.$$

   Then, taking $h = n^{-\frac{1}{2p+(2\beta-1)_+}}$, there exists $C > 0$ such that

   $$E[W_p^p(\tilde{\mu}_n, \mu)] \leq C\psi_n\,,$$

   where

   $$\psi_n = \begin{cases} n^{-\frac{p}{2p+2\beta-1}} & \text{if } \beta > \frac{1}{2} \\ \sqrt{\frac{\log n}{n}} & \text{if } \beta = \frac{1}{2} \\ \frac{1}{\sqrt{n}} & \text{if } \beta < \frac{1}{2}. \end{cases}$$

In the ordinary smooth case, a lower bound is also provided. Let $M > 0$, $q \geq 1$, and define $\mathcal{D}(M, q)$ as the set of measures μ on $\mathbb{R}$ such that $\int |x|^q d\mu(x) \leq M$. Assume that there exist $\beta > 0$ and $c > 0$, such that for every $\ell \in \{0, 1, 2\}$ and every $t \in \mathbb{R}$,

$$|\mu_\varepsilon^{*(\ell)}(t)| \leq c(1 + |t|)^{-\beta}.$$

Then, there exists a constant $C > 0$ such that, for any estimator $\hat{\mu}$,

$$\liminf_{n\to\infty} n^{\frac{p}{2\beta+1}} \sup_{\mu\in\mathcal{D}(M,q)} E[W_p^p(\hat{\mu}, \mu)] > C.$$

I.2.4 Aggregation of estimators

Recent years have witnessed a growing interest in the aggregation of different statistical procedures. Indeed, the increasing number of available estimation methods naturally suggests to try to combine several procedures, looking for a certain guarantee of optimality. In many aggregation methods, the prediction for a new observation x is computed by building a linear or convex combination over a collection of estimators. The strategies discussed here are based on the idea introduced in classification by Mojirsheibani (1999, 2000, 2002a,b), who proposed a smart method for combining several classifiers, relying on a consensus notion. The procedure consists in computing the prediction associated to a new observation by combining selected true labels of the training data. An data point is selected if the predictions computed for this point with the different initial estimators are the same as the prediction for the new observation to be classified.

I.2.4.1 Consensual aggregation for regression

Starting from the idea of Mojirsheibani (1999) in classification, we design in Biau et al. (2016), presented in Chapter IV.1, a related approach in the context of regression. Roughly, instead of requiring equal predictions to select a data point, which does not make sense for continuous output values, they are required to be close, not more distant than some threshold.

More specifically, letting $\mathbf{r} = (r_1, \ldots, r_M)$ denote the collection of individual estimators, the combined estimator is given by

$$T_n\left(\mathbf{r}(x)\right) = \sum_{i=1}^{n} W_{n,i}(x) Y_i, \quad x \in \mathbb{R}^d,$$

where

$$W_{n,i}(x) = \frac{\mathbf{1}_{\bigcap_{m=1}^{M}\{|r_m(x)-r_m(X_i)|\leq\varepsilon\}}}{\sum_{j=1}^{n}\mathbf{1}_{\bigcap_{m=1}^{M}\{|r_m(x)-r_m(X_j)|\leq\varepsilon\}}}.$$

The estimator is shown to perform asymptotically at least as well in the L^2 sense as the best combination of the basic estimators in the initial list.

More formally, for all distributions of (X, Y) with $E[Y^2] < \infty$,

$$\begin{aligned} &E[|T_n(\mathbf{r}(X)) - r^\star(X)|]^2 \\ &\quad \leq E[|T_n(\mathbf{r}(X)) - T(\mathbf{r}(X))|]^2 + \inf_f E[|f(\mathbf{r}(X)) - r^\star(X)|]^2, \end{aligned}$$

where the infimum is taken over all square integrable functions of $\mathbf{r}(X)$. In particular,

$$E[|T_n(\mathbf{r}(X)) - r^\star(X)|^2] \\ \leq \min_{m=1,\dots,M} E[|r_m(X) - r^\star(X)|]^2 + E|T_n(\mathbf{r}(X)) - T(\mathbf{r}(X))|^2.$$

Moreover, we obtain for the term $E[|T_n(\mathbf{r}(X)) - T(\mathbf{r}(X))|]^2$, which represents the price to pay for combining, a rate of convergence of $n^{-\frac{2}{M+2}}$. This is the usual nonparametric rate, with the number of initial estimators playing the role of the dimension.

I.2.4.2 Adding distance information

In Fischer and Mougeot (2019), exposed in Chapter IV.2, we introduce a new learning strategy, which is based on the same consensus idea, but also involves distance information between inputs.

In the original scheme investigated in Mojirsheibani (1999) and Biau et al. (2016), the agreement condition is actually required to hold for all individual estimators, which appears inadequate if there is one bad initial estimator. Mojirsheibani (2002a) notes that this rule may seem too restrictive and proposes to allow a few disagreements (typically, a single one). The resulting classifier is still consistent provided that the number of initial classifiers keeps tending to infinity after removing those with disagreement. Similarly, in Biau et al. (2016), we note that the unanimity constraint may be relaxed in practice by demanding that the distance condition for keeping an observation is true at least for a certain proportion γ of the estimators (for example, $\gamma = 80\%$). Here, we propose a new approach, based on distances between observations, which also aims at reducing the effect of a possibly bad initial estimator. Roughly, choosing a kernel point of view, we propose a combined estimator with weights constructed by mixing distances between entries with distances between predictions coming from the individual estimators.

Let $K : \mathbb{R}^{d+p} \mapsto \mathbb{R}_+$ be a regular kernel, and let the function $g : \mathbb{R}^d \times \mathbb{R}^p \to \mathbb{R}_+$ be such that $g(v_1, v_2) = K(v)$, where $v \in \mathbb{R}^{d+p}$ is the concatenation of $v_1 \in \mathbb{R}^d$ and $v_2 \in \mathbb{R}^p$.

We denote by $\mathbf{f} = (f_1, \dots, f_p)$ the collection of initial estimators. In a regression

context, the combined estimator is defined by

$$T_n(x) = \frac{\sum_{i=1}^n Y_i g\left(\frac{X_i - x}{\alpha}, \frac{\mathbf{f}(X_i) - \mathbf{f}(x)}{\beta}\right)}{\sum_{i=1}^n g\left(\frac{X_i - x}{\alpha}, \frac{\mathbf{f}(X_i) - \mathbf{f}(x)}{\beta}\right)},$$

For classification, the combined classifier $\mathcal{C}_n$ is defined by

$$C_n(x) = \begin{cases} 0 & \text{if } \sum_{i=1}^n Y_i g\left(\frac{X_i - x}{\alpha}, \frac{\mathbf{f}(X_i) - \mathbf{f}(x)}{\beta}\right) \leq \sum_{i=1}^n (1 - Y_i) g\left(\frac{X_i - x}{\alpha}, \frac{\mathbf{f}(X_i) - \mathbf{f}(x)}{\beta}\right) \\ 1 & \text{otherwise.} \end{cases}$$

Our motivation for introducing such a strategy is the intuition that taking advantage of the efficiency of the consensus idea of Mojirsheibani (1999) and Biau et al. (2016) without for all that forgetting the information related to the proximity between entries shall help improving the prediction, especially in the presence of an initial estimator that does not perform very well. We prove the consistency of the procedure in classification and in regression. Applying the method in practice requires finding optimal bandwidths α and β. In particular, through an adequate balance between both terms in the weights, the procedure remains relatively robust when the inputs are high dimensional : more weight is put on the output combination in this case.

I.2.4.3 Clustering for predictive models

In Fischer et al. (2021), which is the subject of Chapter IV.3, we focus on learning problems where the input data consists of more than one unknown clusters, linked to different underlying predictive models. Such a framework is particularly relevant for modeling some physical phenomena with phase transitions, or in situations when some information is missing due to privacy laws.

We propose a three-step procedure to automatically solve this problem. The first step is to catch the clustering structure of the input data, which may be characterized by several statistical distributions. By using several Bregman divergences as proximity measures for the partitioning, we hope to recover precisely the underlying structure for a large variety of possible distributions. For each obtained partition, the second step fits a specific predictive model based on the data in each cluster. Hence, at this stage, we have at hand a certain number of models, corresponding to different Bregman partitions, each of these models consisting of several simple submodels fitted on every cluster of the partition. Finally, the overall model is computed

by aggregating the models corresponding to the different partitions. This combining step is based on the methods presented in Sections I.2.4.1 and I.2.4.2 above.

We conduct numerical experiments on different simulated and real data sets, showing the fine performance of our three-step method in a broad range of prediction problems.

I.2.5 Interdisciplinary collaborations

During the last few years, I have been involved in several interdisciplinary collaborations. Most of them are related to energy or climate issues.

I.2.5.1 Modeling wind energy production

In Fischer et al. (2017), detailed in Chapter V.1, we focus on wind power modeling using machine learning techniques. We deal with real data measured on wind turbines, provided by the wind energy company Maïa Eolis (now Engie Green). The observations come from 3 different farms with 4 to 6 turbines, in the North and East of France, for the period 2011 to 2014. In a farm, each wind turbine provides 10 minute measurements of electrical power, wind speed, wind direction, temperature, as well as an indicator of the working state of the turbine. The electrical power output of the whole farm is also provided on a 10 minute basis.

We model the electrical power as a function of the wind and temperature data, trying several algorithms, parametric or nonparametric. We obtain particularly stable and satisfactory results by aggregating several decision trees using a bagging procedure. As a step toward forecast, we also quantify the impact of using averaged data instead of local measures : the goal is to mimic more global inputs provided by a weather forecaster.

I.2.5.2 Downscaling wind speed

Chapters V.2 and V.3, corresponding to the articles Alonzo et al. (2018) and Goutham et al. (2021) respectively, are the result of a collaboration with Riwal Plougonven, professor at the Laboratoire de Météorologie Dynamique (LMD), at École Polytechnique. We had twice the opportunity to supervise together third-year projects of students of École Polytechnique. The research initiated during the

time of these projects was then pursued in the frame of the PhD dissertation of Bastien Alonzo (currently postdoctoral fellow at Météo France), and the Master 1 internship of Naveen Goutham (currently PhD student at LMD under the supervision of Riwal Plougonven). Our collaborations focused on a downscaling issue, that is the estimation of a local quantity based on past observations at the given location and on outputs of a Numerical Weather Prediction model. The variable of interest is the wind speed, at 10m and 100m.

In Alonzo et al. (2018), we compare the performances of two downscaling statistical methods, respectively based on linear regression and random forests, for reconstructing and forecasting wind speed at a specific location, namely the SIRTA observation platform (an atmospheric research observatory) on the Saclay plateau. We use model outputs from the European Center of Medium-range Weather Forecasts (ECMWF). The assessment of ECMWF for 10m wind speed displays a systematic bias, while at 100m, the wind speed is better represented. Our study shows that both parametric and nonparametric statistical learning methods lead to comparable results: A linear model associated with a wise preprocessing and variable selection shows performances that are slightly better than random forests. Nevertheless, the latter are a valuable option, as they minimize the time required for model preprocessing and calibration.

In Goutham et al. (2021), this methodology is extended to more than 150 stations over mainland France and Corsica, corresponding to several geographical contexts. Various statistical learning methods are tried : linear regression, k-nearest neighbors, random forests, gradient boosting. The random forest model is further explored to reduce the list of explanatory variables. The added-value of using statistical methods is undeniable, and we see, for example, that the improvement is more significant for coastal stations, where the errors of ECMWF data are the largest.

I.2.6 Other applied collaborations

I also had the opportunity to participate in other interdisciplinary collaborations, summarized in Chapter V.4, especially through an involvement in doctoral supervision activities. In particular, I took part in the supervision of a Computer Science PhD, applied to biology for cancer study, as well as a PhD in linguistics, more precisely phonetics. In addition, I participated in research work in astrophysics requiring machine learning tools, in the frame of a collaboration with CEA.

Part II

Principal curves

Chapter II.1

Introduction

Statisticians use various methods in order to sum up information and represent the data by simpler quantities. Among these methods, Principal Component Analysis (PCA) aims at determining the maximal variance axes of a data cloud, as a means to represent the observations in a compact manner revealing as well as possible their variability (see, e.g., Mardia et al. (1979)). This technique, initiated at the beginning of the last century by Pearson (1901) and Spearman (1904), and further developed by Hotelling (1933), is certainly one of the most famous and most widely used procedure of multivariate analysis. Whether in the context of dimension reduction or feature extraction, PCA often provides a first important insight in the data structure.

However, in a number of situations, it may be of interest to summarize information in a nonlinear manner. This approach leads to the notion of principal curve, which can be thought of as a nonlinear generalization of the first principal component. Roughly, the purpose is to search for a curve passing through the middle of a probability distribution or a data cloud, as illustrated in Figure II.1.1. Principal curves have a broad range of applications in many different areas, such as physics (Hastie and Stuetzle (1989), Friedsam and Oren (1989)), character and speech recognition (Kégl and Krzyżak (2002), Reinhard and Niranjan (1999)), mapping and geology (Brunsdon (2007), Stanford and Raftery (2000), Banfield and Raftery (1992), Einbeck et al. (2005a,b)), natural sciences (De'ath (1999), Corkeron et al. (2004), Einbeck et al. (2005b)) and medicine (Wong and Chung (2008), Caffo et al. (2008)).

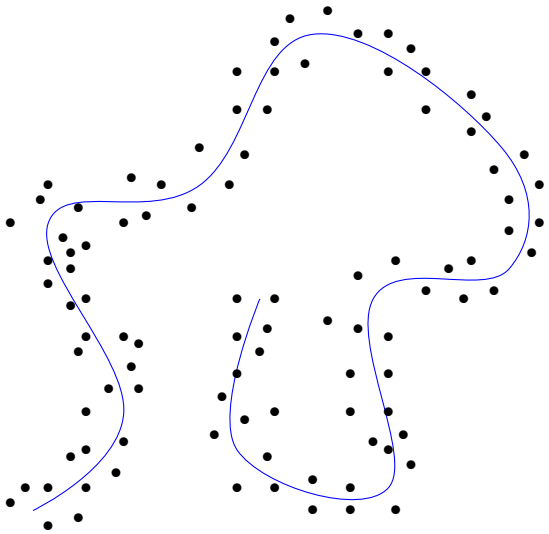


Figure II.1.1 – Example of principal curve for a data cloud

II.1.1 Principal curve definition

The definition of a principal curve typically depends of the principal component property of interest. In the sequel, we consider mainly principal curves with length constraint as introduced by Kégl et al. (2000), or extensions of them. Throughout this part, what we call a curve is a continuous function taking its values in $\mathbb{R}^d$. In the most general form, given a random vector $X \in \mathbb{R}^d$, we aim to study curves $f : [0,1] \to \mathbb{R}^d$ with length less than $L > 0$, minimizing a criterion of the form

$$\Delta(f) = E\left[V\big(d(X, \mathrm{Im} f)\big)\right],$$

with $V : \mathbb{R}_+ \to \mathbb{R}_+$ differentiable and increasing, and d a distance. The definition of Kégl et al. (2000), which is related to standard vector quantization and k-means clustering, consists in setting $V(x) = x^2$, and taking the Euclidean distance for d.

The length of a curve f is here defined by

$$\mathscr{L}(f) = \sup \sum_{j=1}^{m} |f(t_j) - f(t_{j-1})|,$$

where the supremum is taken over all subdivisions $a = t_0 < t_1 < \cdots < t_m = b$, $m \geq 1$.

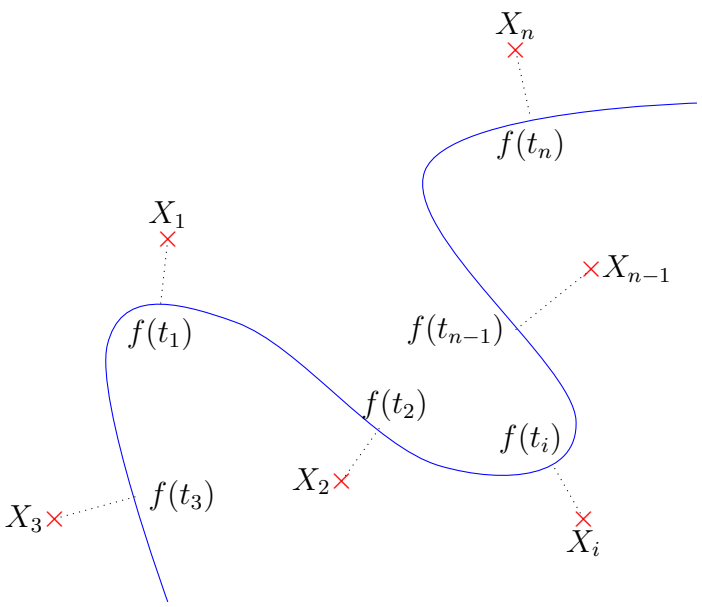


Figure II.1.2 – The projection index t_f. For all i, t_i stands for $t_f(X_i)$

The original definition of a principal curve goes back to Hastie and Stuetzle (1989) and relies on the self-consistency property of principal components. In words, a smooth (infinitely differentiable) curve $f(t) = (f_1(t), \ldots, f_d(t))$ is a principal curve for X if f does not intersect itself, if it has finite length inside any bounded subset of $\mathbb{R}^d$, and if it is self-consistent. This last requirement means that

$$f(t) = E[X|t_f(X) = t],$$

where the so-called projection index $t_f(x)$ is the largest real number t minimizing the squared Euclidean distance between x and $f(t)$, as depicted in Figure II.1.2. More formally,

$$t_f(x) = \sup\left\{t \,:\, |x - f(t)| = \inf_{t'} |x - f(t')|\right\}.$$

The self-consistency property may be interpreted by saying that each point $f(t)$ of the curve is the mean of the observations projecting on Imf near this point. Hastie and Stuetzle (1989) discuss an iterative algorithm, alternating between a projection and a conditional expectation step, which yields an approximate principal curve. Modifications of the algorithm are proposed by Banfield and Raftery (1992) and Chang and Ghosh (1998), whereas Tibshirani (1992) adopts a semiparametric strategy, defining principal curves in terms of a mixture model. Other points of

view are considered for instance in Sandilya and Kulkarni (2002) (principal curves with bounded turn), Verbeek et al. (2001) (k-segments algorithm), Delicado (2001) (principal curves of oriented points), Einbeck et al. (2005a) (local principal curves), Ozertem and Erdogmus (2011) (ridge of a density function), Gerber and Whitaker (2013) (self-consistency and optimization). The reader shall find a detailed presentation of different approaches in Fischer (2014).

A definition relying on the minimization of a criterion like Δ, without any smoothness condition other than continuity for the considered functions, is obviously more amenable to mathematical analysis than the original one, based on an implicit formulation. In particular, it is possible to show, under very general assumptions, that Δ admits a minimizer, that is there exists an optimal curve. On the contrary, the existence of principal curves defined in terms of self-consistency was only proved for a few particular examples (see Duchamp and Stuetzle, 1996a,b). The existence result is stated in the next lemma.

Lemma II.1.1.1. *Let $V : [0,\infty) \to [0,\infty)$ be a lower semi-continuous, strictly increasing function, continuous at 0, and such that $V(0) = 0$. Let X denote a random vector such that $E[V(|X|)] < \infty$. Then, for any finite length L, there exists a curve $f_L^* : [0,1] \to \mathbb{R}^d$ with length $\mathscr{L}(f_L^*) \leq L$ minimizing over all curves with length at most L the criterion*

$$\Delta(f) = E\big[V(d(X, \mathrm{Im} f))\big].$$

II.1.2 A related problem

Observe that principal curves are related to the constrained problem:

minimize $\displaystyle\int_{\mathbb{R}^d} d(x,\Sigma)^p d\mu(x)$ over compact connected sets Σ such that $\mathcal{H}^1(\Sigma) \leq L$.

Here, $\mathcal{H}^1$ denotes the 1-dimensional Hausdorff measure. A connected question is the minimization of the penalized version of the criterion:

$$\int_{\mathbb{R}^d} d(x,\Sigma)^p d\mu(x) + \lambda \mathcal{H}^1(\Sigma).$$

This issue, called in the calculus of variations and shape optimization community "average-distance problem" or, for $p = 1$, "irrigation problem", has been introduced

in Buttazzo et al. (2002) and Buttazzo and Stepanov (2003) (see also the survey Lemenant, 2012, and references therein).

II.1.3 Statistical learning and estimation with principal curves

From a statistical point of view, principal curves may be used both in a statistical learning context and in an estimation context.

In statistical learning, given a sample $X_1, \dots, X_n$, for a length $L > 0$, we consider a minimizer $\hat{f}_{n,L}$ of the empirical criterion $\Delta_n(f) = \frac{1}{n}\sum_{i=1}^{n} V\big(d(X_i, \mathrm{Im}f)\big)$, over all curves with length at most L. Note that, in this case, existence of a minimizer is more straightforward since the empirical measure has finite support. Letting Δ^* denote the minimum of Δ, the purpose is then to study the convergence of $\Delta\big(\hat{f}_{n,L}\big)$ to Δ^*.

Besides, empirical principal curves also provide estimation tools. Considering the model $X_i = g(U_i) + \varepsilon_i$, $i = 1, \dots, n$, where $g : [0,1] \to \mathbb{R}^d$ is an unknown curve and the U_i are independent and uniformly distributed, the aim is to study the convergence of $\mathrm{Im}\hat{f}_n$ to $\mathrm{Im}g$ for some empirically optimal curve $\hat{f}_n$ with suitable length.

Chapter II.2

Model selection

This chapter is related to articles written for my PhD thesis, Fischer (2013), published in the Electronic Journal of Statistics *and Biau and Fischer (2012), a collaboration with Gérard Biau (LPSM, Sorbonne Université) published in* IEEE Transactions on Information Theory*. The purpose is to help put into perspective the subsequent chapters of this part, some of which address similar considerations, with a quite different approach.*

Principal curves were studied from a model selection point of view in my doctoral research. The goal is to select regularity parameters of a principal curve, such as the length, or the number of segments for polygonal lines. Indeed, a curve passing through every point of a data cloud would probably not provide a very interesting summary of the data. This regularization issue is also addressed in Gerber and Whitaker (2013), based on a different principal curve definition.

Our approach relies on the model selection theory by penalization introduced by Birgé and Massart (1997) and Barron et al. (1999) (see also the monograph by Massart, 2007). Two frameworks have been investigated, an estimation issue within a Gaussian model, and a statistical learning framework for polygonal lines.

Some numerical illustrations are also presented, in Section II.2.3, based on heuristics inspired by the penalty calibration approach proposed by Birgé and Massart (2007) and Arlot and Massart (2009).

II.2.1 Length selection for estimation in a Gaussian model

We investigate a Gaussian model selection method in order to choose the length of a principal curve. Our context is similar to that of Caillerie and Michel (2011), who tackle model selection questions for graphs called simplicial complexes. In this section, the Euclidean space $\mathbb{R}^d$ is equipped with the inner product defined by

$$\langle u, v \rangle = \frac{1}{d} \sum_{j=1}^{d} u_j v_j, \tag{II.2.1}$$

and $|\cdot|$ denotes the associated Euclidean norm.

We assume that we observe random vectors $X_1, \ldots, X_n$ with values in $\mathbb{R}^d$ following the model

$$X_i = x_i^* + \sigma \xi_i, \quad i = 1, \ldots, n, \tag{II.2.2}$$

where the x_i^* are unknown, the ξ_i are independent standard Gaussian vectors of $\mathbb{R}^d$ and $\sigma > 0$ stands for the noise level, which is supposed known. Let us denote by $\mathbf{X}$ the column vector containing, in order, all components of the random vectors X_i, $i = 1, \ldots, n$. Defining $\mathbf{x}^*$ and $\boldsymbol{\xi}$ in the same way, the model (II.2.2) can be rewritten under the form

$$\mathbf{X} = \mathbf{x}^* + \sigma \boldsymbol{\xi}.$$

Let F and G be two fixed points of $\mathbb{R}^d$ and $\mathcal{L}$ a countable subset of $]0, +\infty[$. From a practical point of view, several methods can be employed to choose these two points from the observations, for example based on the minimum spanning tree of the data (or of some subset of the data). We introduce a countable collection $\{\mathcal{F}_\ell\}_{\ell \in \mathcal{L}}$, where each set $\mathcal{F}_\ell$ is a class of parameterized curves $f : I \to \mathbb{R}^d$ with length ℓ and endpoints F and G. Our aim is to select the length ℓ. To do this, we consider the criterion Δ_n' defined by

$$\begin{aligned} \Delta_n'(f) &= \frac{1}{n} \sum_{i=1}^{n} \inf_{t \in I} |X_i - f(t)|^2 \\ &= \frac{1}{n} \sum_{i=1}^{n} \inf_{X_i \in \mathrm{Im} f} |X_i - x_i|^2, \end{aligned}$$

where $\mathrm{Im} f$ denotes the range of the curve f. Due to the definition of the norm $|\cdot|$ chosen above (II.2.1), this is the empirical criterion $\Delta_n(f)$ normalized by the

dimension d. Suppose that, for all $\ell \in \mathcal{L}$, $\hat{\mathbf{x}}_\ell = (\hat{x}_{1\ell}, \dots, \hat{x}_{n\ell})$ minimizes

$$\frac{1}{n}\sum_{i=1}^{n} |X_i - x_i|^2$$

among all $\mathbf{x} = (x_1, \dots, x_n) \in \mathcal{C}_\ell = \bigcup_{f \in \mathcal{F}_\ell} (\mathrm{Im} f)^n$. In order to determine the length ℓ, our purpose is to minimize in ℓ a criterion of the type

$$\mathrm{crit}(\ell) = \frac{1}{n}\sum_{i=1}^{n} |X_i - \hat{x}_{i\ell}|^2 + \mathrm{pen}(\ell),$$

where pen : $\mathcal{L} \to \mathbb{R}^+$ is a penalty function, which should avoid the selection of a too large ℓ. Our goal is to design an appropriate penalty. Observe that the classical asymptotic model selection criteria AIC (Akaike, 1973), BIC (Schwarz, 1978) or Mallows'C_p (Mallows, 1973), which involve the "number of parameters" to be estimated, are not suitable in this framework. Therefore, our approach will rely on the non-asymptotic model selection theory developed by Birgé and Massart (2001) and Barron et al. (1999).

When the considered models are linear subspaces, the penalty can be chosen proportional to the dimension of the model, according to Birgé and Massart (2001). Here, the models $\mathcal{C}_\ell$ are not linear subspaces of $\mathbb{R}^{nd}$ and the dimension must be replaced by another quantity. In order to measure the complexity of these nonlinear models, we will use metric entropy. The metric entropy of a set S is given by

$$\mathcal{H}(S, |\cdot|, \varepsilon) = \ln \mathcal{N}(S, |\cdot|, \varepsilon),$$

were the covering number $\mathcal{N}(S, |\cdot|, \varepsilon)$ is the minimal number of balls with radius ε for the norm $|\cdot|$ needed to cover S.

Our approach is based on a general model selection theorem for nonlinear Gaussian models (Massart, 2007). Let us denote by $|\cdot|_{nd}$ the normalized norm of $\mathbb{R}^{nd}$, defined by the inner product $\langle \mathbf{u}, \mathbf{v} \rangle_{nd} = \frac{1}{nd}\sum_{i=1}^{nd} u_i v_i$. For every $\ell \in \mathcal{L}$, let φ_ℓ be a function such that $\varphi_\ell \geq \phi_\ell$, where ϕ_ℓ is given by

$$\phi_\ell(u) = \kappa \int_0^u \sqrt{\mathcal{H}(\mathcal{C}_\ell, |\cdot|_{nd}, \varepsilon)}\, d\varepsilon, \tag{II.2.3}$$

with κ an absolute constant. We define d_ℓ by the equation

$$\varphi_\ell\left(2\sigma\frac{\sqrt{d_\ell}}{\sqrt{nd}}\right) = \frac{\sigma d_\ell}{\sqrt{nd}}.$$

Assume that there exists a family of weights $\{w_\ell\}_{\ell\in\mathcal{L}}$ satisfying

$$\sum_{\ell\in\mathcal{L}} e^{-w_\ell} = \Sigma < \infty.$$

Under these assumptions and with this notation, Theorem 4.18 in Massart (2007) can be written in the following manner:

Theorem II.2.1.1. *Let $\eta > 1$ and*

$$\text{pen}(\ell) \geq \eta \frac{\sigma^2}{nd}\left(\sqrt{d_\ell} + \sqrt{2w_\ell}\right)^2.$$

Then, almost surely, there exists a minimizer $\hat{\ell}$ of the penalized criterion

$$\text{crit}(\ell) = \frac{1}{n}\sum_{i=1}^{n} |X_i - \hat{x}_{i\ell}|^2 + \text{pen}(\ell).$$

Moreover, writing $\tilde{x}_i = \hat{x}_{i\hat{\ell}}$ for all i=1,...,n, we have

$$\frac{1}{n}\sum_{i=1}^{n} E|\tilde{x}_i - x_i^*|^2 \leq c(\eta)\Big[\inf_{\ell\in\mathcal{L}}(d^2(\mathbf{x}^*, \mathcal{C}_\ell) + \text{pen}(\ell)) + \frac{\sigma^2}{nd}(\Sigma + 1)\Big],$$

where $d^2(\mathbf{x}^*, \mathcal{C}_\ell) = \inf_{\mathbf{y}\in\mathcal{C}_\ell} \frac{1}{n}\sum_{i=1}^{n} |y_i - x_i^*|^2$.

This result establishes, for a penalty $\text{pen}(\ell)$ which is large enough, an oracle-type inequality in expectation for the $\tilde{x}_i$, $i = 1, \ldots, n$. Provided a control of the Dudley integral (II.2.3) (Dudley, 1967), this theorem will apply in our context and allow us to select the length ℓ of the curve. To assess this integral, we will need the next lemmas, shown in Fischer (2013).

The first step consists in controlling the metric entropy of the classes $\mathcal{C}_\ell$, $\ell \in \mathcal{L}$. Note that, for all $\ell \in \mathcal{L}$, $\bigcup_{f\in f_\ell} \text{Im} f$ corresponds to an ellipsoid of $\mathbb{R}^d$, as stated in the next lemma. In the sequel, this ellipsoid will be denoted by $\mathcal{E}$.

Lemma II.2.1.1. *Every parameterized curve of $\mathbb{R}^d$ with endpoints F and G and length ℓ ($\ell > FG$), is included in an ellipsoid $\mathcal{E}$ with first principal axis of length ℓ, the other axes having length $\lambda = \sqrt{\ell^2 - FG^2}$.*

In particular, in $\mathbb{R}^2$, $\mathcal{E}$ is an ellipse with foci F and G (see Figure II.2.1), and in $\mathbb{R}^3$, it is a ellipsoid of revolution around the axis passing through these two points.

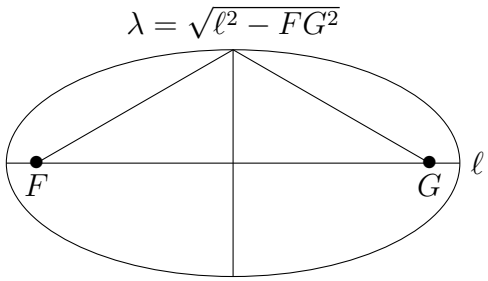


Figure II.2.1 – In the plane $\mathbb{R}^2$, ellipse $\mathcal{E}$ with foci F and G and axes ℓ and λ

We obtain then the following upper bound for $\mathcal{N}(\mathcal{C}_\ell, |\cdot|_{nd}, \varepsilon)$, $\ell \in \mathcal{L}$.

Lemma II.2.1.2. *Suppose that $\ell \geq \lambda \geq \varepsilon$. The covering number of $\mathcal{C}_\ell$ for the normalized norm $|\cdot|_{nd}$ of $\mathbb{R}^{nd}$ satisfies*

$$\mathcal{N}(\mathcal{C}_\ell, |\cdot|_{nd}, \varepsilon) \leq \left(\frac{2}{\varepsilon}\right)^{nd} (\ell\lambda^{d-1})^n.$$

Bounding the integral

$$\phi_\ell(u) = \kappa \int_0^u \sqrt{\mathcal{H}(\mathcal{C}_\ell, |\cdot|_{nd}, \varepsilon)} d\varepsilon$$

for all $\ell \in \mathcal{L}$, we can then define an adequate function φ_ℓ .

Lemma II.2.1.3. *The function φ_ℓ given by*

$$\varphi_\ell(r) = \begin{cases} \kappa r\sqrt{nd}\left(\sqrt{\ln\left(\dfrac{2\ell^{1/d}\lambda^{1-1/d}}{r}\right)} + \sqrt{\pi}\right) & \text{if} \quad r \leq \lambda \\ \varphi_\ell(\lambda) + (r-\lambda)\varphi'_\ell(\lambda) & \text{if} \quad r \geq \lambda \end{cases}$$

satisfies, for all r,

$$\varphi_\ell(r) \geq \phi_\ell(r).$$

Finally, in order to apply Theorem II.2.1.1, we have to assess d_ℓ, defined by the equation

$$\varphi_\ell\left(\frac{2\sigma\sqrt{d_\ell}}{\sqrt{nd}}\right) = \frac{\sigma d_\ell}{\sqrt{nd}},$$

which is the purpose of the next lemma.

Lemma II.2.1.4. *Let φ_ℓ be given by Lemma II.2.1.3. Suppose that*

$$\sigma \leq \frac{\lambda}{4\kappa}\left[\sqrt{\ln 2 + \frac{1}{d}\ln\left(\frac{\ell}{\lambda}\right)} + \sqrt{\pi}\right]^{-1}.$$

Then, equation

$$\varphi_\ell\left(\frac{2\sigma\sqrt{d_\ell}}{\sqrt{nd}}\right) = \frac{\sigma d_\ell}{\sqrt{nd}}$$

admits a solution d_ℓ satisfying

$$d_\ell \leq 8\kappa^2 nd\left(\ln\left(\frac{\ell^{1/d}\lambda^{1-1/d}}{2\sigma\kappa\sqrt{\pi}}\right) + \pi\right).$$

We are now in a position to state the main result of this section.

Theorem II.2.1.2. *Assume that there exists a family of weights $\{w_\ell\}_{\ell\in\mathcal{L}}$ such that*

$$\sum_{\ell\in\mathcal{L}} e^{-w_\ell} = \Sigma < \infty,$$

and that, for every $\ell \in \mathcal{L}$,

$$\sigma \leq \frac{\lambda}{4\kappa}\left[\sqrt{\ln 2 + \frac{1}{d}\ln\left(\frac{\ell}{\lambda}\right)} + \sqrt{\pi}\right]^{-1}. \tag{II.2.4}$$

Then, there exist constants c_1 and c_2 such that, for all $\eta > 1$, if

$$\mathrm{pen}(\ell) \geq \eta\sigma^2\left[c_1\left(\ln\left(\frac{\ell^{1/d}\lambda^{1-1/d}}{\sigma}\right) + c_2\right) + \frac{4w_\ell}{nd}\right], \tag{II.2.5}$$

then, almost surely, there exists a minimizer $\hat{\ell}$ of the penalized criterion

$$\mathrm{crit}(\ell) = \frac{1}{n}\sum_{i=1}^n |X_i - \hat{x}_{i\ell}|^2 + \mathrm{pen}(\ell).$$

Moreover, if $\tilde{x}_i = \hat{x}_{i\hat{\ell}}$ for all $i = 1, \ldots, n$, we have

$$\frac{1}{n}\sum_{i=1}^n E|\tilde{x}_i - x_i^*|^2 \leq c(\eta)\Big[\inf_{\ell\in\mathcal{L}}\{d^2(\mathbf{x}^*, \mathcal{C}_\ell) + \mathrm{pen}(\ell)\} + \frac{\sigma^2}{nd}(\Sigma + 1)\Big],$$

where $d^2(\mathbf{x}^*, \mathcal{C}_\ell) = \inf_{\mathbf{y}\in\mathcal{C}_\ell} \frac{1}{n}\sum_{i=1}^n |y_i - x_i^*|^2$.

Let us now comment on the theorem.

The first remark is about the fact that Theorem II.2.1.2 involves unknown constants. Lemma II.2.1.4 indicates that $c_1 = 16\kappa^2$ and $c_2 = \pi - \ln(2\kappa\sqrt{\pi})$ could be chosen. However, these values are (likely too large) upper bounds. Furthermore, the variance noise σ has been supposed known and is involved in the penalty. Nevertheless, the noise level is generally unknown in practice. In fact, the expression (II.2.5) does not provide a penalty function directly, but gives its shape instead. Note that it is possible to estimate σ separately and then proceed by plug-in. However, there is another solution to assess c_1, c_2 and σ, relying on the slope heuristics. This penalty calibration method introduced by Birgé and Massart (2007) (see also Arlot and Massart, 2009; Lerasle, 2012; Saumard, 2013) precisely allows to tune a penalty known up to a multiplicative constant.

According to the formula binding ℓ and λ, the quantity $\ln(\ell^{1/d}\lambda^{1-1/d})$ in the penalty characterizes each model of curves with length ℓ. The other elements varying over the collection of models are the weights $\{w_\ell\}_{\ell\in\mathcal{L}}$. For linear models $\mathcal{C}_\ell$ with dimension D_ℓ, a possible choice for w_ℓ is $w_\ell = w(D_\ell)$ where $w(D) = cD + \ln|\{k \in \mathcal{L}, D_k = D\}|$ and $c > 0$ (see Massart, 2007). If there is no redundancy in the models dimension, this strategy amounts to choosing w_ℓ proportional to D_ℓ. By analogy, w_ℓ may here be chosen proportional to $\ln(\ell^{1/d}\lambda^{1-1/d})$. More formally, we set $w_\ell = c\ln\ell^{1/d}\lambda^{1-1/d}$, where the constant $c > 0$ is such that $\sum_{\ell\in\mathcal{L}} \frac{1}{\ell^{c/d}\lambda^{c(1-1/d)}} = \Sigma < +\infty$. This choice finally yields a penalty proportional to $\ln(\ell^{1/d}\lambda^{1-1/d})$, which may be calibrated in practice thanks to the slope heuristics.

In addition, observe that condition (II.2.4) says that the noise level σ should not be too large with respect to λ. In other words, if $\lambda = \sqrt{\ell^2 - FG^2}$ is of the same order as σ, it is not possible to obtain a suitable principal curve with length ℓ.

Finally, let us point out that due to the exponent n in the covering number in Lemma II.2.1.2, the penalty shape obtained does not tend to 0 as n tends to infinity. This point is intrinsically related to the geometry of the problem. Indeed, its resolution is not made easier by increasing the size of the sample, since nothing has been specified about the repartition of the x_i^*'s.

II.2.2 Parameter selection in statistical learning

In this section, the considered estimators are polygonal lines, with the smoothness driven either by the length or by the turn. In both cases, we show that the polygonal

line obtained by minimizing some appropriate penalized criterion satisfies an oracle-type inequality.

Let X be a random vector such that $P(X \in K) = 1$, where K is a convex compact subset of $\mathbb{R}^d$, with diameter δ.

We let $\Delta(f) = E\left[\min_{t\in[0,1]} |X - f(t)|^2\right]$.

The empirical counterpart, based on a sample $X_1, \ldots, X_n$ of independent random variables distributed as X, is given by

$$\Delta_n(f) = \frac{1}{n}\sum_{i=1}^{n} \min_{t\in[0,1]} |X_i - f(t)|^2.$$

II.2.2.1 Principal curves with bounded length

For any given positive length L, there exists an optimal curve f^* for X with length at most L in K (Kégl, 1999, Lemma 1). We restrict ourselves to curves whose support is included in K and denote by $\mathcal{F}$ the set of such curves.

For $L > 0$, we set

$$f^* \in \underset{f\in\mathcal{F}, \mathscr{L}(f)\leq L}{\arg\min}\ \Delta(f).$$

Next, let $\mathcal{L}$ be a countable subset of $]0, L]$ and $\mathcal{Q}$ a grid over $\mathcal{C}$, that is $\mathcal{Q} = K \cap \Gamma$, where Γ is a lattice of $\mathbb{R}^d$. For every $k \geq 1$ and $\ell \in \mathcal{L}$, the model $\mathcal{F}_{k,\ell}$ is defined as the collection of polygonal lines with k segments, with length at most ℓ, and with vertices belonging to $\mathcal{Q}$. We note that each model $\mathcal{F}_{k,\ell}$ as well as the family of models $\{\mathcal{F}_{k,\ell}\}_{k\geq 1,\ell\in\mathcal{L}}$ are countable. For $k \geq 1$ and $\ell \in \mathcal{L}$, let

$$\hat{f}_{k,\ell} \in \underset{f\in\mathcal{F}_{k,\ell}}{\arg\min}\, \Delta_n(f)$$

be a curve achieving the minimum of the empirical criterion $\Delta_n(f)$ over the polygonal line class $\mathcal{F}_{k,\ell}$.

At this stage of the procedure, we have at hand a family of estimates $\{\hat{f}_{k,\ell}\}_{k\geq 1,\ell\in\mathcal{L}}$ and our goal is to select the best principal curve $\tilde{f}$ among this collection. We set

$$\mathcal{D}(f, g) = \Delta(g) - \Delta(f).$$

Let pen : $\mathbb{N}^* \times \mathcal{L} \to \mathbb{R}^+$ be some penalty function and denote by $(\hat{k}, \hat{\ell})$ a pair of minimizers of the criterion

$$\mathrm{crit}(k, \ell) = \Delta_n(\hat{f}_{k,\ell}) + \mathrm{pen}(k, \ell).$$

In order to obtain a suitable curve, we have to design an adequate penalty $\mathrm{pen}(k,\ell)$. This is done in the following theorem, which is an adaptation of a general model selection result for statistical learning (Massart, 2007, Theorem 8.1). Another example of using this result can be found in Fischer (2011), in the context of finding the number of groups in k-means clustering.

Theorem II.2.2.1. *Consider a family of nonnegative weights* $\{x_{k,\ell}\}_{k\geq 1,\ell\in\mathcal{L}}$ *such that*

$$\sum_{k\geq 1,\ell\in\mathcal{L}} e^{-x_{k,\ell}} = \Sigma < \infty,$$

and a penalty function $\mathrm{pen} : \mathbb{N}^* \times \mathcal{L} \to \mathbb{R}^+$. *Let* $\tilde{f} = \hat{f}_{\hat{k},\hat{\ell}}$. *If for all* $(k,\ell) \in \mathbb{N}^* \times \mathcal{L}$,

$$\mathrm{pen}(k,\ell) \geq E\left[\sup_{f\in\mathcal{F}_{k,\ell}} \Big(\Delta(f) - \Delta_n(f)\Big)\right] + \delta^2\sqrt{\frac{x_{k,\ell}}{2n}},$$

then

$$E[\mathcal{D}(f^*,\tilde{f})] \leq \inf_{k\geq 1,\ell\in\mathcal{L}} \Big[\mathcal{D}(f^*,\mathcal{F}_{k,\ell}) + \mathrm{pen}(k,\ell)\Big] + \frac{\delta^2\Sigma}{2^{3/2}}\sqrt{\frac{\pi}{n}},$$

where $\mathcal{D}(f^*,\mathcal{F}_{k,\ell}) = \inf_{f\in\mathcal{F}_{k,\ell}} \mathcal{D}(f^*,f)$.

Theorem II.2.2.1 offers a nonasymptotic bound, expressing the fact that the expected loss of the final estimate $\tilde{f}$ is close to the minimal loss over all $k \geq 1$ and $\ell \in \mathcal{L}$, up to a term tending to 0. Thus, in order to apply this theorem to the principal curve problem, we have to find an upper bound on the quantity

$$E\left[\sup_{f\in\mathcal{F}_{k,\ell}} \Big(E[\Delta(f,x)] - \Delta_n(f)\Big)\right]. \tag{II.2.6}$$

This is achieved by Proposition II.2.2.1 below, which is proved in Biau and Fischer (2012) by showing that the expected maximal deviation (II.2.6) may be bounded by a Rademacher average (see, e.g., Bartlett et al., 2001) and by resorting to a Dudley integral (Dudley, 1967).

Proposition II.2.2.1. *Let* $\mathcal{F}_{k,\ell}$ *be the set of all polygonal lines with* k *segments, length at most* ℓ, *and vertices in a grid* $\mathcal{Q} \subset K$. *Then there exist nonnegative constants* $a_0,\ldots,a_2$, *depending on the maximal length* L, *the dimension* d *and the diameter* δ *of the convex set* K, *such that*

$$E\left[\sup_{f\in\mathcal{F}_{k,\ell}} \Big(\Delta(f) - \Delta_n(f)\Big)\right] \leq \frac{1}{\sqrt{n}}\Big[a_1\sqrt{k} + a_2\ell + a_0\Big].$$

Finally, combining Theorem II.2.2.1 and Proposition II.2.2.1 leads to the following result.

Theorem II.2.2.2. *Consider a family of nonnegative weights* $\{x_{k,\ell}\}_{k\geq 1,\ell\in\mathcal{L}}$ *such that*

$$\sum_{k\geq 1,\ell\in\mathcal{L}} e^{-x_{k,\ell}} = \Sigma < \infty,$$

and a penalty function $\mathrm{pen} : \mathbb{N}^* \times \mathcal{L} \to \mathbb{R}^+$. *Let* $\tilde{f} = \hat{f}_{\hat{k},\hat{\ell}}$. *There exist nonnegative constants* $c_0, \ldots, c_2$, *depending on the maximal length* L, *the dimension* d *and the diameter* δ *of the convex set* K, *such that, if for all* $(k,\ell) \in \mathbb{N}^* \times \mathcal{L}$,

$$\mathrm{pen}(k,\ell) \geq \frac{1}{\sqrt{n}}\left[c_1\sqrt{k} + c_2\ell + c_0\right] + \delta^2\sqrt{\frac{x_{k,\ell}}{2n}},$$

then

$$E[\mathcal{D}(f^*,\tilde{f})] \leq \inf_{k\geq 1,\ell\in\mathcal{L}}\left[\mathcal{D}(f^*,\mathcal{F}_{k,\ell}) + \mathrm{pen}(k,\ell)\right] + \frac{\delta^2\Sigma}{2^{3/2}}\sqrt{\frac{\pi}{n}},$$

where $\mathcal{D}(f^*,\mathcal{F}_{k,\ell}) = \inf_{f\in\mathcal{F}_{k,\ell}} \mathcal{D}(f^*,f)$.

The penalty shape involves a term proportional to $\sqrt{k/n}$ and a term proportional to $\ell/\sqrt{n}$. This penalty form, which vanishes at the rate $1/\sqrt{n}$, seems relevant insofar as the number k of segments and the length ℓ of the curves measure the complexity of the models.

Note that the proof of Proposition II.2.2.1 provides possible values for the constants $c_0, \ldots, c_2$. However, these values are not very helpful since they are upper bounds which are probably far from being tight. Hence, again, a calibration heuristic is needed in practice. Nevertheless, the proof also reveals that $c_1 = c_1'\delta^2$, $c_2 = c_2'\delta$ and $c_0 = c_0'\delta^2$, where c_0', c_1' and c_2' are constants without dimension, so that the penalty is in fact homogeneous to a squared length.

Finally, an important practical issue is how to choose the weights $\{x_{k,\ell}\}_{k\geq 1,\ell\in\mathcal{L}}$. These weights should be large enough to ensure the finiteness of Σ, but not too large at the risk of overpenalizing. If the cardinality of the collection of models is not larger than n^2, we may set $x_{k,\ell} = 2\ln n$ for every (k,ℓ). This choice does not affect the penalty shape, though modifying the rate, and leads to $\Sigma = 1$ in the risk bound.

II.2.2.2 Principal curves with bounded turn

In this section, we considered principal curves indexed by the turn instead of the length, following Sandilya and Kulkarni (2002). The turn $\mathscr{K}(f)$ of a curve

$f : [0,1] \to \mathbb{R}^d$, is given by

$$\mathscr{K}(f) = \sup \sum_{j=1}^{m-1} \widehat{f(t_j)},$$

where $\widehat{f(t_j)}$ denotes the angle between the vectors $\overrightarrow{f(t_{j-1})f(t_j)}$ and $\overrightarrow{f(t_j)f(t_{j+1})}$, and the supremum is taken over all subdivisions $0 = t_0 < t_1 < \cdots < t_m = 1$, $m \geq 1$ (Alexandrov and Reshetnyak, 1989). Thus, the turn of a polygonal line with vertices $v_1, \ldots, v_{k+1}$ is the sum of the angles at $v_2, \ldots, v_k$ (see Figure II.2.2).

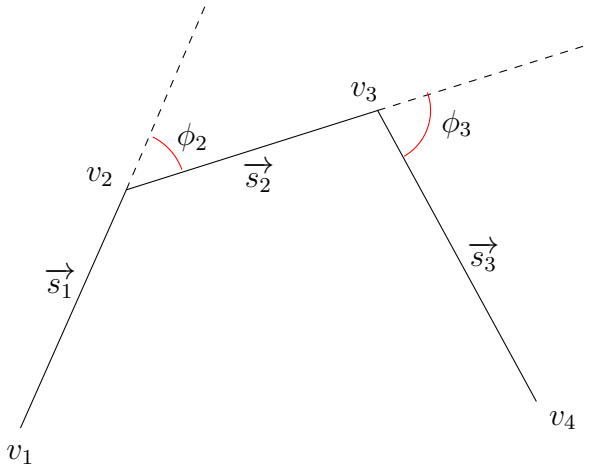


Figure II.2.2 – Denoting by $\overrightarrow{s_j}$ the vector $\overrightarrow{v_j v_{j+1}}$ for all $j = 1, \ldots, k$, the angles involved in the definition of the turn are defined by $\phi_{j+1} = (\overrightarrow{s_j}, \overrightarrow{s_{j+1}})$

For any turn T, there exists an optimal curve f^* for X with turn at most T in K (Sandilya and Kulkarni, 2002, Proposition 1). As above, we consider curves in $\mathcal{F}$.

For $T \geq 0$, we set

$$f^* \in \underset{f \in \mathcal{F}, \mathscr{T}(f) \leq T}{\arg\min} \ \Delta(f),$$

where $\mathscr{T}(f)$ denotes the turn of f. We let $\mathcal{T}$ be a countable subset of $[0,T]$ and define a countable collection of models $\{\mathcal{F}_{k,\kappa}\}_{k \geq 1, \kappa \in \mathcal{T}}$: each $\mathcal{F}_{k,\kappa}$ consists of polygonal lines with k segments, with turn at most κ, and with vertices belonging to some grid $\mathcal{Q}$ over K. For $k \geq 1$ and $\kappa \in \mathcal{T}$, define

$$\hat{f}_{k,\kappa} \in \underset{f \in \mathcal{F}_{k,\kappa}}{\arg\min} \, \Delta_n(f)$$

to be a polygonal line minimizing the empirical criterion $\Delta_n(f)$ over $\mathcal{F}_{k,\kappa}$. We wish to design an appropriate penalty function pen : $\mathbb{N}^* \times \mathcal{T} \to \mathbb{R}^+$ and minimize the criterion

$$\mathrm{crit}(k,\kappa) = \Delta_n(\hat{f}_{k,\kappa}) + \mathrm{pen}(k,\kappa).$$

We let $\tilde{f} = \hat{f}_{\hat{k},\hat{\kappa}}$, where $(\hat{k},\hat{\kappa})$ is a minimizer of the penalized criterion $\mathrm{crit}(k,\kappa)$.

An upper bound on the quantity

$$E\left[\sup_{f\in\mathcal{F}_{k,\kappa}}\Big(E[\Delta(f,x)] - \Delta_n(f)\Big)\right]$$

will lead to a result of the form of Theorem II.2.2.2. We will use the fact that a curve with bounded turn also has bounded length (Alexandrov and Reshetnyak, 1989, Chapter 5).

Lemma II.2.2.1. *Let f be a curve with turn κ and let δ be the diameter of K. Then $\mathscr{L}(f) \le \delta\zeta(\kappa)$, where the function ζ is defined by*

$$\zeta(x) = \begin{cases} \dfrac{1}{\cos(x/2)} & \text{if} \quad 0 \le x \le \dfrac{\pi}{2} \\ 2\sin(x/2) & \text{if} \quad \dfrac{\pi}{2} \le x \le \dfrac{2\pi}{3} \\ \dfrac{x}{2} - \dfrac{\pi}{3} + \sqrt{3} & \text{if} \quad x \ge \dfrac{2\pi}{3}. \end{cases}$$

The graph of the function ζ is shown in Figure II.2.3.

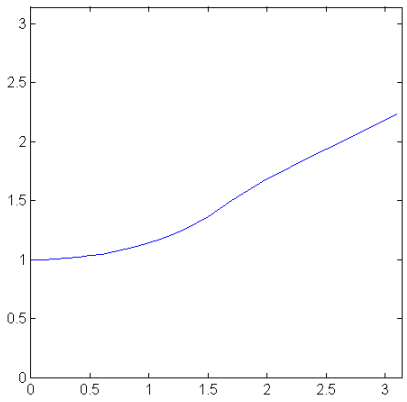


Figure II.2.3 – Graph of the function ζ

Thanks to this result, we obtain Proposition II.2.2.2 below, which is the counterpart of Proposition II.2.2.1 in the current setting.

Proposition II.2.2.2. *Let $\mathcal{F}_{k,\kappa}$ be the set of all polygonal lines with k segments, turn at most κ, and vertices in a grid $\mathcal{Q} \subset K$, and let δ be the diameter of the convex set K. Then there exist nonnegative constants $a_0, \ldots, a_4$, depending only on the dimension d, such that*

$$E\left[\sup_{f\in\mathcal{F}_{k,\kappa}}\Big(\Delta(f)-\Delta_n(f)\Big)\right] \leq \delta^2\left[a_1\sqrt{k}+a_2\sqrt{\zeta(\kappa)}+a_3\frac{\zeta(\kappa)}{\sqrt{k}}\mathbf{1}_{\{\frac{\zeta(\kappa)}{3k}<1\}}+a_4\sqrt{k\ln\frac{\zeta(\kappa)}{k}}\mathbf{1}_{\{\frac{\zeta(\kappa)}{3k}\geq 1\}}+a_0\right].$$

Putting finally Theorem II.2.2.1 and Proposition II.2.2.2 together, we obtain:

Theorem II.2.2.3. *Consider a family of nonnegative weights $\{x_{k,\kappa}\}_{k\geq 1,\kappa\in\mathcal{T}}$ such that*

$$\sum_{k\geq 1,\kappa\in\mathcal{T}} e^{-x_{k,\kappa}} = \Sigma < \infty,$$

and a penalty function $\mathrm{pen} : \mathbb{N}^* \times \mathcal{T} \to \mathbb{R}^+$. Let $\tilde{f} = \hat{f}_{\hat{k},\hat{\kappa}}$. There exist nonnegative constants $c_0, \ldots, c_2$, depending only on the dimension d, such that, if for all $(k,\kappa) \in \mathbb{N}^* \times \mathcal{T}$,

$$\mathrm{pen}(k,\kappa) \geq \frac{\delta^2}{\sqrt{n}}\left[c_1\sqrt{k}+c_2\sqrt{\zeta(\kappa)}+c_0+\sqrt{\frac{x_{k,\kappa}}{2}}\right],$$

then

$$E[\mathcal{D}(f^*,\tilde{f})] \leq \inf_{k\geq 1,\kappa\in\mathcal{T}}\Big[\mathcal{D}(f^*,\mathcal{F}_{k,\kappa})+\mathrm{pen}(k,\kappa)\Big]+\frac{\delta^2\Sigma}{2^{3/2}}\sqrt{\frac{\pi}{n}},$$

where $\mathcal{D}(f^*,\mathcal{F}_{k,\kappa}) = \inf_{f\in\mathcal{F}_{k,\kappa}}\mathcal{D}(f^*,f)$.

The expression of the penalty shape involves again a term of the order $\sqrt{k/n}$, whereas the length ℓ is replaced by $\sqrt{\zeta(\kappa)}$, which is an increasing function of the turn κ. This is relevant, since the number of segments k and the turn κ characterize the complexity of the models.

II.2.3 Experimental results

This section presents a few numerical experiments, carried out with the software MATLAB, to illustrate the model selection procedure suggested above. We consider

a setting where the parameters are the length and the number of segments. The minimization of the criterion $\Delta_n(f)$ is achieved through a MATLAB optimization routine.

As already mentioned, the penalty shapes involve constants which have to be practically calibrated. We use ideas inspired from the slope heuristics (Arlot and Massart, 2009; Birgé and Massart, 2007; Lerasle, 2012; Saumard, 2013). In short, the slope heuristics allows to tune a penalty known up to some multiplicative constant. The slope heuristics assumes that the empirical contrast decreases when the complexity of the models increases, which is clearly the case in our principal curve context. The procedure is based on the fact that the graph of the empirical contrast as a function of the penalty shape decreases strongly at the beginning and more slowly later, with a linear trend. The slope of this line is used to compute the constant.

In our framework, to a first approximation, we deal with a penalty of the form $c_1\sqrt{k} + c_2\ell$. Thus, we assume that, for large values of k and ℓ, $\Delta_n(\hat{f}_{k,\ell})$ behaves like $c_1\sqrt{k} + c_2\ell$. The constants $\hat{c}_1$ and $\hat{c}_2$ are chosen via an ordinary least square regression.

The algorithm may be described as follows.

Algorithm MS

---

1. For $k = 1, \ldots, k_{\max}$, $\ell \in \mathcal{L}$, compute $\hat{f}_{k,\ell}$ by minimizing the empirical criterion $\Delta_n(f)$ and record $\Delta_n(\hat{f}_{k,\ell})$.
2. Set $x_{k,\ell} = 2\ln n$ and consider a penalty of the form

$$\text{pen}(k,\ell) = c_1\sqrt{k} + c_2\ell.$$

3. Select the constants $\hat{c}_1$ and $\hat{c}_2$ using a bivariate version of the slope heuristics.
4. Retain the curve $\hat{f}_{\hat{k},\hat{\ell}}$ obtained by minimizing the penalized criterion

$$\text{crit}(k,\ell) = \Delta_n(\hat{f}_{k,\ell}) - 2(\hat{c}_1\sqrt{k} + \hat{c}_2\ell).$$

The results of the algorithm MS are compared to the outputs of the Polygonal Line Algorithm (PL hereafter) of Kégl et al. (2000), which is a more local procedure.

II.2.3.1 Simulated digit data

In this first series of experiments, we considered two-dimensional data distributed with bivariate Gaussian noise around a reference curve.

The generative curves are handwritten digits. The noise variance is set to 0.04, 150 observations were sampled around the digit 2 and the digit 3 and 250 observations around the digit 5. The observations and the results are depicted in Figure II.2.4.

We observe for the digit 2 data that the **MS** principal curve follows the observations more closely than what would be expected. On the other hand, the **PL** output looks smoother but a bit too short. Indeed, a comparison with the generative curve shows that the loop at the top and the angle at the bottom of the digit are not recovered precisely. For the digit 3, we note again that the algorithm **MS** slightly overfits the data, whereas the smoother curve **PL** misses the angle. The same comment holds for the digit 5, with however a smoother **MS** curve.

II.2.3.2 NIST database digits

In this second series of experiments, we use data from NIST Special Database 19 (`http://www.nist.gov/srd/nistsd19.cfm`), containing handwritten characters from 3600 writers. The data consists in binary images scanned at 11.8 dots per millimeter (300 dpi), which uniformly fill the area corresponding to the thickness of the pen stroke. Skeletonization, which consists in reducing foreground regions in such an image without affecting the general shape of the handwritten character, often constitutes a preliminary step in character recognition.

The results of both algorithms applied to three NIST database digits are visible in Figure II.2.5.

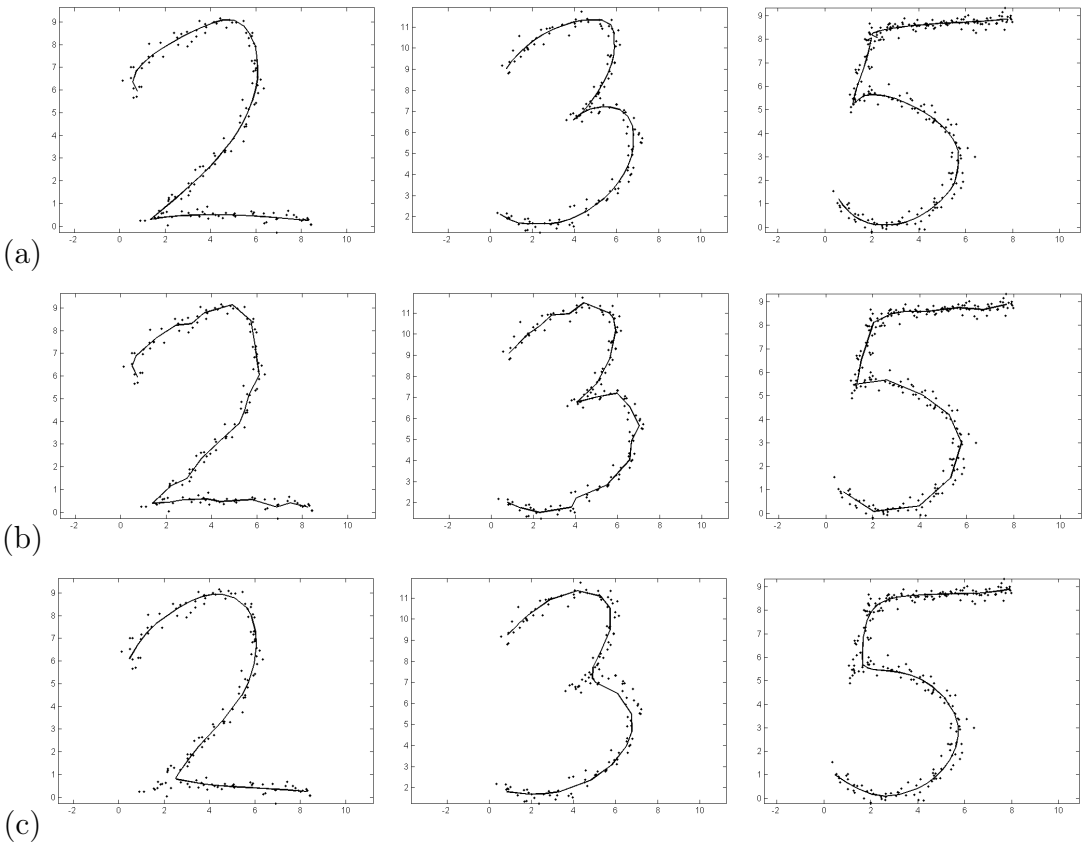


Figure II.2.4 – (a) Observations. (b) Principal curves selected by the method MS: $\hat{k} = 27$, $\hat{\ell} = 24$; $\hat{k} = 23$, $\hat{\ell} = 23$; $\hat{k} = 17$, $\hat{\ell} = 21$. (c) PL principal curves

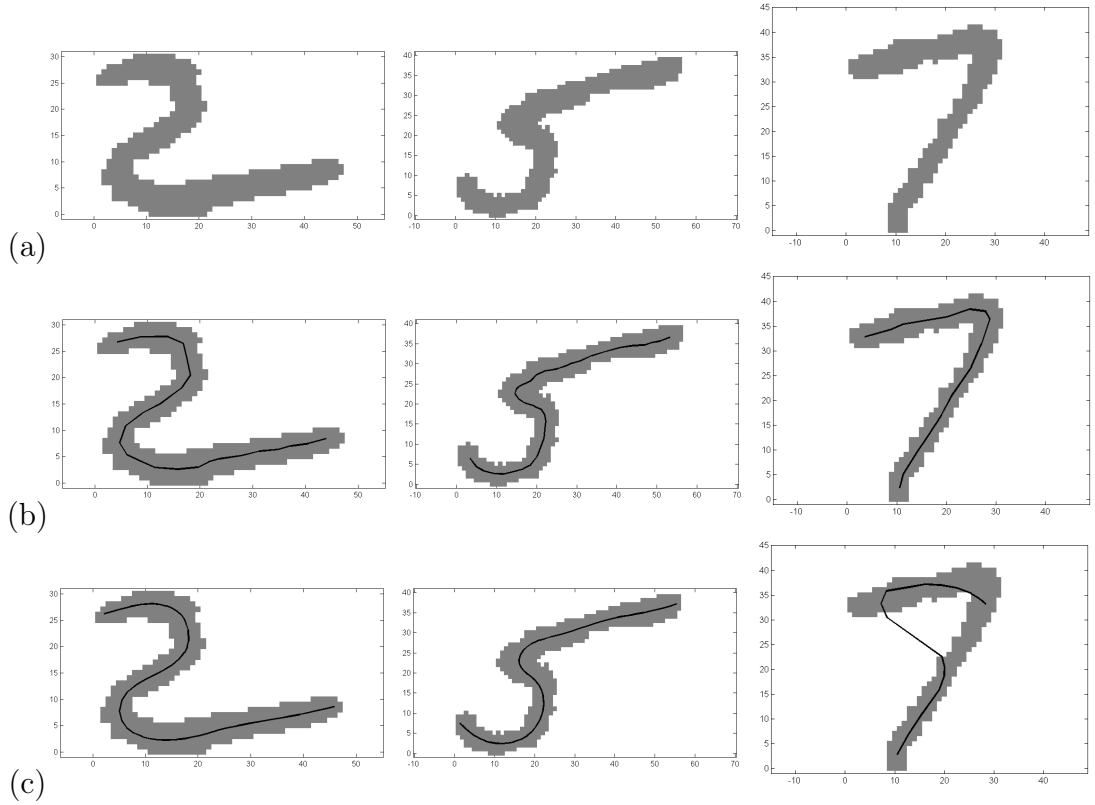


Figure II.2.5 – (a) Three NIST database handwritten digits. (b) Principal curves selected by the method MS: $\hat{k} = 23$, $\hat{\ell} = 80$; $\hat{k} = 38$, $\hat{\ell} = 82$; $\hat{k} = 15$, $\hat{\ell} = 66$. (c) PL principal curves

Chapter II.3

Some regularity properties of a principal curve

This chapter is the result of a collaborative work with Sylvain Delattre (LPSM, Université de Paris), published in the Probability and Statistics section of the Annales de l'Institut Henri Poincaré *(Delattre and Fischer, 2020).*

II.3.1 Introduction

II.3.1.1 Context of the problem and motivation

We focus on the problem:

find a curve $f : [0,1] \to \mathbb{R}^d$ minimizing the quantity

$$E\left[d^2(X, \mathrm{Im}f)\right] = \int d^2(x, \mathrm{Im}f) d\mu(x),$$

over all curves with length $\mathscr{L}(f)$, such that $\mathscr{L}(f) \leq L$. (II.3.1)

Here, X is some random vector with distribution μ, taking its values in $\mathbb{R}^d$. This corresponds to principal curves with length constraint, as described in Kégl et al. (2000). These authors show that there exists indeed a minimizer whenever X is square integrable.

Originally, principal curves were introduced in Hastie and Stuetzle (1989), with a different definition, based on the so-called self-consistency property. In this point of view, a curve f is said to be self-consistent for a random vector X with finite second moment if it satisfies:

$$f(t_f(X)) = E[X|t_f(X)] \quad \text{a.s.},$$

where the projection index t_f is given by

$$t_f(x) = \max \arg\min_t |x - f(t)|.$$

The self-consistency property may be interpreted as follows: each point on the curve is the average of the mass of the probability distribution projecting there (for more details about the notion of self-consistency, see Tarpey and Flury (1996)). Some regularity assumptions are made in addition: the principal curve is required to be smooth (C^∞), it does not intersect itself, and has finite length inside any ball in $\mathbb{R}^d$. The existence of principal curves designed according to this definition cannot be proved in general (see Duchamp and Stuetzle, 1996a,b, for results obtained in the case of some particular distributions in two dimensions), which is the main motivation for the least-square minimization definition proposed in Kégl et al. (2000).

Note that several other principal curve definitions, as well as algorithms, were proposed in the literature (Delicado, 2001; Einbeck et al., 2005a; Gerber and Whitaker, 2013; Ozertem and Erdogmus, 2011; Sandilya and Kulkarni, 2002; Tibshirani, 1992; Verbeek et al., 2001). Note also that principal curves, in their empirical version, have many applications in various areas (see for example Friedsam and Oren (1989); Hastie and Stuetzle (1989) for applications in physics, Kégl and Krzyżak (2002); Reinhard and Niranjan (1999) in character and speech recognition, Banfield and Raftery (1992); Brunsdon (2007); Einbeck et al. (2005a,b); Stanford and Raftery (2000) in mapping and geology, Corkeron et al. (2004); De'ath (1999); Einbeck et al. (2005a) in natural sciences, Caffo et al. (2008) in pharmacology, and Drier et al. (2013); Wong and Chung (2008) in medicine, for the study of cardiovascular disease or cancer).

II.3.1.2 Description of our results

We consider general distributions, assuming only that X has a second order moment, and search for a curve which is optimal for problem (II.3.1). We deal with open curves (with endpoints), as well as closed curves ($f(0) = f(1)$). Throughout, we will assume that the length-constraint is effective, that is the support of X is not

the image of a curve with length less than or equal to L. In this context, we prove that a minimizing curve cannot be self-consistent. We also show that, for an optimal curve, the set of points with several different projections of the curve, called ridge set in studies about the "average-distance problem" (see Section II.3.1.3), or ambiguity points in the principal curves literature, is negligible for the distribution of X. Then, we establish that an optimal curve is right- and left-differentiable everywhere and has bounded curvature. Moreover, we obtain a first order Euler-Lagrange equation: we show that there exist $\lambda > 0$ and a random variable $\hat{t}$ taking its values in $[0, 1]$ such that $|X - f(\hat{t})| = d(X, \mathrm{Im}f)$ a.s. and

$$E\left[X - f(\hat{t})|\hat{t} = t\right] m_{\hat{t}}(dt) = -\lambda f''(dt), \tag{II.3.2}$$

where $m_{\hat{t}}$ stands for the distribution of $\hat{t}$. To obtain that $\lambda \neq 0$, we use the fact that an optimal curve is not self-consistent. Equation (II.3.2) allows us to propose in dimension $d = 2$ a proof of the injectivity of an open principal curve as well as of a closed principal curve restricted to $[0, 1)$.

II.3.1.3 Comparison with previous results

Our framework is related to the constrained problem:

$$\text{minimize } \int_{\mathbb{R}^d} d(x, \Sigma)^p d\mu(x) \text{ over compact connected sets } \Sigma \text{ such that } \mathcal{H}^1(\Sigma) \leq L. \tag{II.3.3}$$

Here, $\mathcal{H}^\ell$ denotes ℓ-dimensional Hausdorff measure. A connected question is the minimization of the penalized version of the criterion:

$$\int_{\mathbb{R}^d} d(x, \Sigma)^p d\mu(x) + \lambda \mathcal{H}^1(\Sigma). \tag{II.3.4}$$

This issue, called in the calculus of variations and shape optimization community "average-distance problem" or, for $p = 1$, "irrigation problem", has been introduced in Buttazzo and Stepanov (2003); Buttazzo et al. (2002) (see also the survey Lemenant, 2012, and the references therein). Considering a compactly supported distribution, the penalized form is studied for connected sets, with $p = 1$, in Lu and Slepcev (2013), and for curves, with $p \geq 1$, in Lu and Slepcev (2016). In the first article, the authors prove that a minimizer is a tree made of a finite union of curves with

finite length, and they provide a bound on the total curvature of these curves. In the second one, they show existence of a curve minimizing the penalized criterion

$$\int_{\mathbb{R}^d} d(x, \mathrm{Im} f)^p d\mu(x) + \lambda \mathscr{L}(f). \tag{II.3.5}$$

They give a bound on the curvature of the minimizer, and prove that, in two dimensions, if $p \geq 2$ or the distribution μ has a bounded density with respect to Lebesgue measure, a minimizing curve is injective.

For the penalized irrigation problem (II.3.4), under the assumption that the distribution μ, with compact support, does not charge the sets that have finite $\mathcal{H}^{d-1}$ measure, which is true for instance if it has a density with respect to Lebesgue measure, an Euler-Lagrange equation is obtained for $p = 1$ in Buttazzo et al. (2009), whereas Lemenant (2011) uses arguments involving endpoints to derive one in the case of the constrained version (II.3.3), in $\mathbb{R}^2$, under the same assumption on μ. This assumption implies that X is almost surely different from its projection on the curve, which is required for differentiability when $p = 1$, and, moreover, it is used to ensure negligibility of the ridge set.

For the constrained problem (II.3.3), if Σ^* denotes a minimizer and

$$\int_{\mathbb{R}^d} d(x, \Sigma)^p d\mu(x) > 0,$$

it is shown in Paolini and Stepanov (2004) that $\mathcal{H}^1(\Sigma^*) = L$. A similar result in our context is stated in Corollary II.3.3.1 below.

Another related setting is the "lazy travelling salesman problem" of Polak and Wolansky (2007): in $\mathbb{R}^2$, taking for μ an empirical distribution and considering closed curves, the authors study the penalized problem (II.3.5) for $p = 2$ (with $\lambda\mathscr{L}(f)$ replaced by $\lambda\mathscr{L}^2(f)$). They show that for λ large enough, the problem is reduced to a convex optimization.

Recall that we study in this chapter the constrained problem (II.3.1), for open or closed curves. In our context, the distribution of X is not required to be compactly supported, and we do not need to assume that μ does not charge the sets with finite $\mathcal{H}^{d-1}$ measure to derive an Euler-Lagrange equation. Indeed, our proof does not rely on the fact that the ridge set is negligible. Besides, we prove that ambiguity points are actually negligible, which implies in particular that, for a given optimal curve, the Lagrange multiplier λ in (II.3.2) only depends on the curve f. We decided to focus on the case $p = 2$ for which we can state the more complete results. In particular, we

are only able to show the lack of self-consistency of an optimal curve when $p = 2$. As already mentioned, this is a key point to get the main result. Observe that it would be interesting to define a counterpart of the lack of self-consistency when considering other values of p.

The rest of the chapter is organized as follows. Section II.3.2 introduces relevant notation and recalls some basic facts about length-constrained principal curves. In Section II.3.3, negligibility of ambiguity points is given in Proposition II.3.3.1, and the main result is stated in his complete form in Theorem II.3.3.1.

Injectivity results are presented in Section II.3.4. Finally, we give in Section II.3.5 explicit examples of optimal curves.

II.3.2 Definitions and notation

Let $(\Omega, \mathcal{F}, P)$ be a probability space and X a random vector on $(\Omega, \mathcal{F}, P)$ with values in $\mathbb{R}^d$, such that $E[|X|^2] < \infty$. For $r > 0$, let $B(x, r)$ and $\bar{B}(x, r)$ denote, respectively, the open and the closed balls with center x and radius r. For a curve $f : [0, 1] \to \mathbb{R}^d$, let $\mathscr{L}(f) \in [0, \infty]$ denote its length.

We set

$$\Delta(f) = E\left[d^2(X, \mathrm{Im} f)\right],$$

and, for $L \geq 0$,

$$G(L) = \min\{\Delta(f), f \in \mathcal{C}_L\},$$

where, in the sequel, $\mathcal{C}_L$ will denote either one of the following sets of curves:

$$\{f \in [0, 1] \to \mathbb{R}^d, \mathscr{L}(f) \leq L\},$$
$$\{f \in [0, 1] \to \mathbb{R}^d, \mathscr{L}(f) \leq L, f(0) = f(1)\}.$$

Curves belonging to the latter set are closed curves. Note that G is well-defined. Indeed, the existence of an open curve f with $\mathscr{L}(f) \leq L$ achieving the infimum of the criterion $\Delta(f)$ is shown in Kégl et al. (2000), and the same proof applies for closed curves.

It will be useful to rewrite $G(L)$, for every $L \geq 0$, as the minimum of the quantity

$$E[|X - \hat{X}|^2]$$

over all possible random vectors $\hat{X}$ taking their values in the image $\mathrm{Im} f$ of a curve $f \in \mathcal{C}_L$.

Remark 1. If $f : [0,1] \to \mathbb{R}^d$ is Lipschitz with constant L, its length is at most L. This follows directly from the definition of the length. Conversely, if the curve $f : [0,1] \to \mathbb{R}^d$ has length $\mathscr{L}(f) \leq L$, then there exists a curve with the same image which is Lipschitz with constant L. Indeed, a curve with finite length may be parameterized by arc-length (1-Lipschitz) (see, e.g., Alexandrov and Reshetnyak (1989, Theorem 2.1.4)).

Remark 2. Let $L \geq 0$. Suppose that $\hat{X}$ satisfies $G(L) = E[|X - \hat{X}]|^2]$. Writing

$$E[|X - \hat{X}|^2] = E[|X - \hat{X} - E[X - \hat{X}]|^2] + |E[X] - E[\hat{X}]|^2,$$

we see that, necessarily,

$$E[X] = E[\hat{X}], \tag{II.3.6}$$

since, otherwise, the criterion could be made strictly smaller by replacing $\hat{X}$ by the translated variable $\hat{X} + E[X] - E[\hat{X}]$, which contradicts the optimality of $\hat{X}$.

Observe that (II.3.6) remains true in a more general setting, as soon as the constraint corresponds to a quantity invariant by translation.

II.3.3 Main results

II.3.3.1 Negligibility of the ridge set

Given a curve $f : [0,1] \to \mathbb{R}^d$, consider the set

$$\mathcal{P}_f(x) = \{y \in \mathrm{Im} f, |x - y| = d(x, \mathrm{Im} f)\} = \bar{B}(x, d(x, \mathrm{Im} f)) \cap \mathrm{Im} f.$$

If $\mathcal{P}_f(x)$ has cardinality at least 2, x is called an ambiguity point in the principal curves literature (see Hastie and Stuetzle (1989)). Properties of the set of such points, named ridge set in the shape optimization community, have been studied for instance in Mantegazza and Mennucci (2003). In particular, the ridge set is measurable. Using property (II.3.6), it may be shown that the ridge set of an optimal curve for X is negligible for the distribution of X.

Proposition II.3.3.1. *Let $f^* \in \mathcal{C}_L$ be an optimal curve for X, i.e. $\Delta(f^*) = G(L)$.*

1. *The set $\mathcal{A}_f^* = \{x \in \mathbb{R}^d, \mathrm{Card}(\mathcal{P}_f^*(x)) \geq 2\}$ of ambiguity points is measurable.*
2. *The set $\mathcal{A}_f^*$ is negligible for the distribution of X.*

Remark 3. The fact that the ridge set is negligible for the distribution of X may be extended to the context of computing optimal trees under $\mathcal{H}^1$ constraint. Indeed, the result relies on property (II.3.6), and $\mathcal{H}^1$ measure is translation invariant.

II.3.3.2 Main theorem and comments

Recall that a signed measure on $(\Omega, \mathcal{F})$ is a function $m : \mathcal{F} \to \mathbb{R}$ such that $m(\emptyset) = 0$ and m is σ-additive, that is $m\left(\bigcup_{k\geq 1} A_k\right) = \sum_{k\geq 1} m(A_k)$ for any sequence $(A_k)_{k\geq 1}$ of pairwise disjoint sets. For an $\mathbb{R}^d$-valued signed measure m on $[0,1]$, that is $m = (m^1, \dots, m^d)$, where each m^j is a signed measure, and for $g : [0,1] \to \mathbb{R}^d$ a measurable function, we will use the following notation: $\int \langle g(t), m(dt)\rangle = \sum_{j=1}^d \int g^j(t) m^j(dt)$.

A probability space $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{P})$ will be called an extension of $(\Omega, \mathcal{F}, P)$ if there exists a random vector $\tilde{X}$ defined on $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{P})$, with the same distribution μ as X. For simplicity, we still denote this random vector by X.

The next theorem is proved in Delattre and Fischer (2020).

Theorem II.3.3.1. *Let $L > 0$ such that $G(L) > 0$ and let $f^* \in \mathcal{C}_L$ such that $\Delta(f^*) = G(L)$. Then, $\mathscr{L}(f^*) = L$. Assuming that f^* is L-Lipschitz, we obtain that*

- *f^* is right-differentiable on $[0,1)$, $|f^{*\prime}_r(t)| = L$ for all $t \in [0,1)$,*
- *f^* is left-differentiable on $(0,1]$, $|f^{*\prime}_\ell(t)| = L$ for all $t \in (0,1]$,*

and there exists a unique signed measure $f^{*\prime\prime}$ on $[0,1]$ (with values in $\mathbb{R}^d$) such that

- *$f^{*\prime\prime}((s,t]) = f^{*\prime}_r(t) - f^{*\prime}_r(s)$ for all $0 \leq s \leq t < 1$,*
- *$f^{*\prime\prime}([0,1]) = 0$.*

In the case $\mathcal{C}_L = \{f : [0,1] \to \mathbb{R}^d, \mathscr{L}(f) \leq L\}$, we also have

- *$f^{*\prime\prime}(\{0\}) = f^{*\prime}_r(0)$,*
- *$f^{*\prime\prime}(\{1\}) = -f^{*\prime}_\ell(1)$.*

Moreover, there exists a unique $\lambda > 0$ and there exists a random variable $\hat{t}$ with values in $[0,1]$, defined on an extension $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{P})$ of the probability space $(\Omega, \mathcal{F}, P)$, such that

- *$|X - f^*(\hat{t})| = d(X, \mathrm{Im} f^*)$ a.s.,*
- *for every bounded Borel function $g : [0,1] \to \mathbb{R}^d$,*

$$E\left[\langle X - f^*(\hat{t}), g(\hat{t})\rangle\right] = -\lambda \int_{[0,1]} \langle g(t), f^{*\prime\prime}(dt)\rangle. \tag{II.3.7}$$

Remark 4. Let $m_{\hat{t}|X}$ denote the conditional distribution of $\hat{t}$ given X. Then, Equation (II.3.7) can be written in the following form:

$$\int_{\mathbb{R}^d} \int_{[0,1]} \langle x - f^*(t), g(t)\rangle m_{\hat{t}|X}(x, dt) d\mu(x) = -\lambda \int_{[0,1]} \langle g(t), f^{*\prime\prime}(dt)\rangle.$$

Remark 5. Whenever the function g is absolutely continuous, an integration by parts shows that Equation (II.3.7) may also be written

$$E\left[\langle X - f^*(\hat{t}), g(\hat{t})\rangle\right] = \lambda \int_0^1 \langle g'(t), {f^*}'_r(t)\rangle dt. \tag{II.3.8}$$

To see this, let us write

$$f^{*\prime\prime}([0,1])g(1) = f^{*\prime\prime}(\{0\})g(0) + \int_{(0,1]} \langle g(t), f^{*\prime\prime}(dt)\rangle + \int_{(0,1]} \langle g'(s), f^{*\prime\prime}([0,s])\rangle ds.$$

Since $f^{*\prime\prime}([0,1]) = 0$, we have

$$0 = \int_{[0,1]} \langle g(t), f^{*\prime\prime}(dt)\rangle + \int_{(0,1]} \langle g'(s), {f^*}'_r(s)\rangle ds,$$

which, combined with Equation (II.3.7), implies the announced formula Equation (II.3.8).

Remark 6. If the curve f^* has an angle at t, which means that ${f^*}'_r(t) \neq {f^*}'_\ell(t)$, we see that

$$E[(X - f^*(\hat{t}))\mathbf{1}_{\{\hat{t}=t\}}] = -\lambda f^{*\prime\prime}(\{t\}) = \lambda({f^*}'_\ell(t) - {f^*}'_r(t)) \neq 0.$$

So, at an angle, $P(\hat{t} = t) > 0$.

Besides, when $\mathcal{C}_L = \{f : [0,1] \to \mathbb{R}^d, \mathscr{L}(f) \leq L\}$, we have

$$E[(X - f^*(\hat{t}))\mathbf{1}_{\{\hat{t}=0\}}] = -\lambda f^{*\prime\prime}(\{0\}) = -\lambda {f^*}'_r(0),$$

which cannot be zero, since ${f^*}'_r(0)$ has norm $L > 0$. This implies that $P(\hat{t} = 0) > 0$.

Remark 7. Regarding the random variable $\hat{t}$, let us mention that $\hat{t}$ is unique almost surely whenever the curve is injective since $f^*(\hat{t})$ is unique almost surely (it is the case in dimension $d \leq 2$; see Section II.3.4). In general, it is worth pointing out that Theorem II.3.3.1 does not ensure that it is a function of X, as $(X, \hat{t})$ is, in fact, obtained as a limit in distribution of $(X, \hat{t}_n)$ for some sequence $(\hat{t}_n)_{n\geq 1}$. Besides, note that we do not know whether λ depends on the curve f^*.

Remark 8 (Principal curves in dimension 1). Let $\mathcal{C}_L = \{f : [0,1] \to \mathbb{R}^d, \mathscr{L}(f) \leq L\}$. It may be of interest to consider the simplest case of dimension 1, where the problem may be solved entirely and explicitly Assume that X is a real-valued random variable, and that, for some length $L > 0$, $G(L) > 0$. Consider an optimal curve f^* with length

$\mathscr{L}(f^*) \leq L$. Using Corollary II.3.3.1 below, we have that, in fact, $\mathscr{L}(f^*) = L$, so that the image of f^* is given by an interval $[a, a+L]$. In this context, solving directly the length-constrained principal curve problem in dimension 1 leads to minimizing in a the quantity

$$\Delta(a) = E\left[d^2(X, \mathrm{Im} f)\right] = E[(X-a)^2 \mathbf{1}_{\{X<a\}}] + E[(X-a-L)^2 \mathbf{1}_{\{X>a+L\}}].$$

The function Δ is differentiable in a, with derivative given by

$$\Delta'(a) = 2E[(a-X)\mathbf{1}_{\{X<a\}}] + 2E[(a+L-X)\mathbf{1}_{\{X>a+L\}}].$$

Moreover, Δ' admits a right-derivative $\Delta'_r(a) = 2(P(X<a) + P(X>a+L))$, which is positive since $G(L) > 0$ implies that we do not have $X \in [a, a+L]$ almost surely. Hence, Δ is strictly convex, which shows that the minimizing a is unique, so that the image of the principal curve f^* is also uniquely defined.

Besides, observe that Equation (II.3.7) from Theorem II.3.3.1 takes the following form in dimension 1: for every bounded Borel function $g : [0,1] \to \mathbb{R}^d$,

$$E[(X-a)\mathbf{1}_{\{X<a\}} g(0)] + E[(X-a-L)\mathbf{1}_{\{X>a+L\}} g(1)] = \lambda L(g(1) - g(0)).$$

In particular, we get

$$\begin{aligned} E[(X-a)\mathbf{1}_{\{X<a\}}] &= -\lambda L, \\ E[(X-a-L)\mathbf{1}_{\{X>a+L\}}] &= \lambda L, \end{aligned}$$

which characterizes λ. Let us stress that we directly see in this case that $\lambda > 0$, since, otherwise $X \in [a, a+L]$ almost surely, which contradicts the fact that $G(L) > 0$.

In the next sections, we present two lemmas, which are important both independently and for obtaining the main result Theorem II.3.3.1.

II.3.3.3 Properties of the function G

The first lemma is about the monotonicity and continuity properties of the function G. Observe that G is nonincreasing, since $\{f : [0,1] \to \mathbb{R}^d, \mathscr{L}(f) \leq L_1\} \subset \{f : [0,1] \to \mathbb{R}^d, \mathscr{L}(f) \leq L_2\}$ when $L_1 < L_2$, so that $G(L_2) \leq G(L_1)$.

Lemma II.3.3.1. *1. The function G is continuous.*

2. The function G is strictly decreasing over $[0, L_0)$, where

$$L_0 = \inf\{L \geq 0, G(L) = 0\} \in \mathbb{R}_+ \cup \{\infty\}.$$

In particular, Lemma II.3.3.1 admits the next useful corollary.

Corollary II.3.3.1. *For $L > 0$, if $G(L) > 0$ and $f \in \mathcal{C}_L$ is such that $\Delta(f) = G(L)$, then $\mathscr{L}(f) = L$.*

II.3.3.4 Lack of self-consistency

The next lemma states that a principal curve with length $\leq L$ does not satisfy the so-called self-consistency property, provided that the constraint is effective, that is $G(L) > 0$.

Lemma II.3.3.2. *Let $L > 0$ such that $G(L) > 0$, and let $f^* \in \mathcal{C}_L$ be such that $\Delta(f^*) = G(L)$. If $\hat{X}$ is a random vector with values in $\mathrm{Im} f^*$ such that $|X - \hat{X}| = d(X, \mathrm{Im} f^*)$ a.s., then $P(E[X|\hat{X}] \neq \hat{X}) > 0$.*

Equipped with Lemmas II.3.3.1 and II.3.3.2, we can present a sketch of the proof of the main result.

II.3.3.5 Sketch of proof of Theorem II.3.3.1

To obtain a length-constrained principal curve, we have to minimize a function which may not be differentiable. We propose to build a discrete approximation of the principal curve f^*, using a chain of points $v_1^n, \ldots, v_n^n$, $n \geq 1$, in $\mathbb{R}^d$. For every $n \geq 1$, linking the points yields a polygonal curve f_n. The properties of the principal curve f^* are shown by passing to the limit. The chain of points is obtained by minimizing a k-means-like criterion, which is differentiable, under a length constraint. This criterion is based on the distances from the random vector X to the n points and not to the corresponding segments of the polygonal line f_n, which allows to simplify the computation of the gradients.

Let us focus on open curves, that is in the case $\mathcal{C}_L = \{f : [0,1] \to \mathbb{R}^d, \mathscr{L}(f) \leq L\}$. The case of closed curves turns out to be even simpler since there are no endpoints and so all points of the curve play the same role. Note that the normalization factor "$n-1$" below becomes "n" in the closed curve context.

To facilitate understanding, we sketch the proof in a simpler case. Assume that X has a density with respect to Lebesgue measure, and consider a polygonal line f_n with vertices $v_1^n, \ldots, v_n^n$ obtained by minimizing under length constraint the criterion

$$F_n^0(x_1, \ldots, x_n) = E\left[\min_{1 \leq i \leq n} |X - x_i|^2\right].$$

For $h = (h_1, \dots, h_n) \in (\mathbb{R}^d)^n$, $\nabla F_n^0.h = \sum_{i=1}^n E\left[-2\langle X - \hat{X}_n, h_i\rangle \mathbf{1}_{\{\hat{X}=v_i^n\}}\right]$, where $\hat{X}$ is such that $|X - \hat{X}| = \min_{1\le j\le n} |X - v_i^n|$. For differentiability, it is convenient to write the length constraint as follows:

$$(n-1)\sum_{i=2}^{n} |x_i - x_{i-1}|^2 \le L^2.$$

Let $\hat{t}_n$ be defined by $\hat{t}_n = \frac{i-1}{n-1}$ on the event $\{\hat{X} = v_i^n\}$. For a test function g, set $h_i = g\left(\frac{i-1}{n-1}\right)$ for $i = 1, \dots, n$. Then, we obtain the Euler-Lagrange equation

$$E\left[\langle X - f_n(\hat{t}_n), g(\hat{t}_n)\rangle\right] = -\lambda_n \int_{[0,1]} \langle g(t), f_n''(dt)\rangle. \tag{II.3.9}$$

Up to an extraction, f_n converges uniformly to an optimal curve and $\hat{t}_n$ converges in distribution. Using the lack of self-consistency Lemma II.3.3.2, it may be shown that every limit point of the sequence $(\lambda_n)_{n\ge1}$ is positive. Together with the discrete Euler-Lagrange equation (II.3.9), this allows to prove that f_n'' converges weakly to a signed measure, which is $f^{*\prime\prime}$. Finally, the desired Euler-Lagrange equation is obtained as the limit of (II.3.9).

II.3.4 An application to injectivity

In this section, we present an application of formula (II.3.7) of Theorem II.3.3.1. We will use this first order condition to show in dimension $d = 2$ that an open optimal curve is injective, and a closed optimal curve restricted to $[0, 1)$ is injective, except in the case where its image is a segment. To obtain the result, we follow arguments exposed in Lu and Slepcev (2016) in the frame of the penalized problem, for open curves. The main difference is the fact that we have at hand the Euler-Lagrange equation, which allows to simplify the proof.

Again, we consider $L > 0$ such that $G(L) > 0$ and a curve $f^* \in \mathcal{C}_L$ such that $\Delta(f^*) = G(L)$, which is L-Lipschitz. We let $\hat{t}$ be defined as in Theorem II.3.3.1. The random vector $f^*(\hat{t})$ will sometimes be denoted by $\hat{X}$. Recall that $|X - \hat{X}| = d(X, \mathrm{Im} f^*)$ a.s. by Theorem II.3.3.1.

To prove the injectivity of f^*, we need several preliminary lemmas, proved in Delattre and Fischer (2020). Let us point out that Lemmas II.3.4.1 to II.3.4.5 below are valid for every $d \ge 1$.

First of all, we state the next lemma, which will be useful in the sequel, providing a lower bound on the curvature of any closed arc of f^*. Recall that the total variation of a signed measure ν is defined by

$$|\nu| = \Big(\sum_{j=1}^{d} |\nu^j|_{TV}^2\Big)^{1/2},$$

where $|\nu^j|_{TV}$ denotes the total variation norm of ν^j. For $0 \leq a < b \leq 1$, ${f^*_{(a,b]}}''$ denotes the vector-valued signed measure $f^{*\prime\prime}((a,b] \cap \cdot)$.

Lemma II.3.4.1. *If $0 \leq a < b \leq 1$ and $f^*(a) = f^*(b)$, then $|{f^*_{(a,b]}}''| \geq L$.*

As a first step toward injectivity, we then obtain that, if a point is multiple, it is only visited finitely many times.

Lemma II.3.4.2. *For every $t \in [0,1]$, the set $f^{*-1}(\{f^*(t)\})$ is finite.*

In the case $\mathcal{C}_L = \{f : [0,1] \to \mathbb{R}^d, \mathscr{L}(f) \leq L\}$, the endpoints of the curve f^* cannot be multiple points.

Lemma II.3.4.3. *Let $\mathcal{C}_L = \{f : [0,1] \to \mathbb{R}^d, \mathscr{L}(f) \leq L\}$. We have $f^{*-1}(\{f^*(0)\}) = \{0\}$ and $f^{*-1}(\{f^*(1)\}) = \{1\}$.*

For an open curve, there exists a multiple point which is the last multiple point.

Lemma II.3.4.4. *Let $\mathcal{C}_L = \{f : [0,1] \to \mathbb{R}^d, \mathscr{L}(f) \leq L\}$. There exists $\delta > 0$ such that for every $t \in [1-\delta, 1]$, $f^{*-1}(\{f^*(t)\}) = \{t\}$.*

We obtain that the two branches of the curve are necessarily tangent at a multiple point.

Lemma II.3.4.5. (*i*) *If there exist $0 < t_0 < t_1 < 1$ such that $f^*(t_0) = f^*(t_1)$, then $f^{*\prime}_\ell(t_0) = f^{*\prime}_r(t_0) = -f^{*\prime}_r(t_1) = -f^{*\prime}_\ell(t_1)$.*

(*ii*) *In the case $\mathcal{C}_L = \{f : [0,1] \to \mathbb{R}^d, \mathscr{L}(f) \leq L, f(0) = f(1)\}$, if there exists $0 < t < 1$ such that $f^*(t) = f^*(0)$, then $f^{*\prime}_\ell(t) = f^{*\prime}_r(t) = -f^{*\prime}_r(0) = -f^{*\prime}_\ell(1)$.*

We introduce the set

$$D = \Big\{t \in [0,1) \mid \mathrm{Card}\big(f^{*-1}(\{f^*(t)\}) \cap [0,1)\big) \geq 2\Big\}.$$

Lemma II.3.4.6. *If $f^*(t)$, $t \in (0,1)$, is a multiple point of $f^* : [0,1] \to \mathbb{R}^2$, then t cannot be right- or left-isolated:*
for all $t \in D \cap (0,1)$, for all $\varepsilon > 0$, $(t, t+\varepsilon) \cap D \neq \emptyset$ and $(t-\varepsilon, t) \cap D \neq \emptyset$.

Finally, we state the injectivity result in dimension 2, for open and closed curves.

Proposition II.3.4.1. *(i) If $\mathcal{C}_L = \{f \in [0,1] \to \mathbb{R}^2, \mathscr{L}(f) \leq L\}$, then f^* is injective.*

(ii) If $\mathcal{C}_L = \{f \in [0,1] \to \mathbb{R}^2, \mathscr{L}(f) \leq L, f^*(0) = f^*(1)\}$, then either f^* restricted to $[0,1)$ is injective or $\mathrm{Im} f^*$ is a segment.

II.3.5 Examples of principal curves

This section presents two examples of optimal curves. The proofs are available in Delattre and Fischer (2020).

II.3.5.1 Uniform distribution on an enlargement of a curve

The purpose of this section is to study the principal curve problem for the uniform distribution on an enlargement of some generative curve. For $A \subset \mathbb{R}^d$ and $r \geq 0$, we denote by

$$A \oplus r = \left\{x \in \mathbb{R}^d \mid d(x, A) \leq r\right\}$$

the r-enlargement of A. Under some conditions on the generative curve $g : [0,1] \to \mathbb{R}^d$, for r small enough, it turns out that the image of an optimal curve with length $\mathscr{L}(g)$ for the uniform distribution on an r-enlargement of $\mathrm{Im} g$ is necessarily $\mathrm{Im} g$. More specifically, the radius r must not exceed the reach of $\mathrm{Im} g$.

The reach of a set $A \subset \mathbb{R}^d$ is the supremum of the radii ρ such that every point at distance at most ρ of A has a unique projection on A. More formally, following Federer (1959), we define for $A \subset \mathbb{R}^d$

$$\mathrm{reach}(A) = \sup\left\{\rho \geq 0 \mid \forall x \in \mathbb{R}^d \quad d(x, A) \leq \rho \Rightarrow \exists! a \in A \quad d(x, a) = d(x, A)\right\} \in [0, +\infty].$$

The question of the optimality of the generative curve when considering the uniform distribution on an enlargement has been first addressed in dimension $d = 2$

in Mosconi and Tilli (2005). Observe that related ideas can be found in Genovese et al. (2012a). Our proof in arbitrary dimension $d \geq 1$ relies on arguments in Federer (1959), which moreover allow to show uniqueness.

Theorem II.3.5.1. *Let $g : [0,1] \to \mathbb{R}^d$ be a curve. Suppose that g is injective, differentiable, g' is Lipschitz, and there exists $c > 0$ such that $|g'(t)| \geq c$ for all $t \in [0,1]$. Then, the reach of $\mathrm{Im}g$ is positive. Let $r \in (0, \mathrm{reach}(\mathrm{Im}g)]$ and let X be a random vector uniformly distributed on $\mathrm{Im}g \oplus r$. Consider a function $V : [0,\infty) \to [0,\infty)$ continuous, increasing and such that $V(0) = 0$. Then, for every curve $f : [0,1] \to \mathbb{R}^d$ such that $\mathscr{L}(f) \leq \mathscr{L}(g)$ one has*

$$E\big[V\left(d(X, \mathrm{Im}g)\right)\big] \leq E\big[V\left(d(X, \mathrm{Im}f)\right)\big].$$

with equality if and only if $\mathrm{Im}f = \mathrm{Im}g$.

The proof of the theorem is based on the two next lemmas. For $k \geq 1$, λ_k denotes the Lebesgue measure on $\mathbb{R}^k$ and α_k the volume of the unit ball in $\mathbb{R}^k$. From (Mosconi and Tilli, 2005, Lemma 42), we have the next result.

Lemma II.3.5.1. *Let A be a compact connected subset of $\mathbb{R}^d$ with $\mathcal{H}^1(A) < \infty$. Then for all $r \geq 0$ one has*

$$\lambda_d(A \oplus r) \leq \mathcal{H}^1(A)\alpha_{d-1}r^{d-1} + \alpha_d r^d.$$

Lemma II.3.5.2. *Let $g : [0,1] \to \mathbb{R}^d$ be a curve. Suppose that g is injective, g is differentiable, g' is Lipschitz, and there exists $c > 0$ such that $|g'(t)| \geq c$ for all $t \in [0,1]$. Then, the reach of $A = \mathrm{Im}g$ is positive and for all $r \leq \mathrm{reach}(A)$ one has*

$$\lambda_d\left(A \oplus r\right) = \mathscr{L}(g)\alpha_{d-1}r^{d-1} + \alpha_d r^d \tag{II.3.10}$$

Moreover, one has

$$\{x \in A \oplus r \mid d\left(x, \partial(A \oplus r)\right) \geq r\} \subset A. \tag{II.3.11}$$

II.3.5.2 Uniform distribution on a circle

In this section, we investigate the principal curve problem for a particular distribution, the uniform distribution on a circle.

Proposition II.3.5.1. *Consider the unit circle centered at the origin with parameterization given by*

$$g(t) = (\cos(2\pi t), \sin(2\pi t))$$

for $t \in [0,1]$. *Let* U *be a uniform random variable on* $[0,1]$ *and let* $X = g(U)$. *Then, for every* $L < 2\pi$, *the circle centered at the origin with radius* $\dfrac{L}{2\pi}$ *is the unique closed principal curve with length* L *for* X.

Remark 9. Observe that radial symmetry of a distribution is not sufficient to guarantee that a given circle will be a constrained principal curve for this distribution. Let us exhibit two counterexamples.

— Let $p > 0$ and let $\mathcal{U}$ denote the uniform distribution on the unit circle. Consider a random variable X taking its values in $\mathbb{R}^2$, distributed according to the mixture distribution

$$p\delta_{(0,0)} + (1-p)\mathcal{U},$$

where $\delta_{(0,0)}$ stands for the Dirac mass at the origin $(0,0)$. Then, for every circle with center $(0,0)$ and radius $r \in (0,1]$, because of the atom at the origin, the projection of X on the circle is not unique almost surely, which implies, thanks to Proposition II.3.3.1, that none of these circles may be a constrained principal curve for X.

— We consider the case where X is a standard Gaussian random vector in $\mathbb{R}^2$. Lemma II.3.3.2 ensures that the circle with center $(0,0)$ and radius $E[|X|] = \sqrt{\pi/2}$ cannot be a constrained principal curve for X because it is self-consistent.

Chapter II.4

Estimation via length-constrained principal curves

This chapter is the result of a collaborative work with Sylvain Delattre, the complete version of which can be found in the article Delattre and Fischer (2021), currently submitted.

II.4.1 Introduction

II.4.1.1 Preliminary picture of the estimation result

Let $n \geq 1$. We observe random vectors X_i^n, given by

$$X_i^n = g(U_i^n) + \varepsilon_i^n, \quad i = 1, \ldots, n, \tag{II.4.1}$$

where the unknown function $g : [0,1] \to \mathbb{R}^d$ is continuous. Moreover, g is assumed to have finite length equal to its 1-dimensional Hausdorff measure and to have constant speed. Here, the random variables U_i^n, $i = 1, \ldots, n$, taking their values in $[0,1]$, are independent, and belong to a class of distributions with full support, enclosing for instance the uniform distribution as a particular case.

We study an asymptotic context, where the noise tends in probability to 0 (in a sense that will be specified below) when the number of observations n tends to infinity.

The main result in this chapter is the construction, relying on the principal curve notion, of an estimator $\hat{f}_n$, which converges to the unknown curve g in Hausdorff distance, in the sense that the Hausdorff distance between $\mathrm{Im}\hat{f}_n$ and $\mathrm{Im}g$ converges in probability to 0.

II.4.1.2 Related work

The problem of estimating the image of g may be cast into the general context of filament or manifold estimation from observations sampled on or near the unknown shape.

The literature mainly focuses on shapes with a reach bounded away from zero. The reach ρ, characterizing the regularity of the shape, is the maximal radius of a ball rolling on it (see Federer, 1959). In Genovese et al. (2012a), an additive noise model of the form (II.4.1) is studied. The curve g is parameterized by arc-length, normalized to $[0, 1]$. The authors assume that the U_i, $i = 1, \ldots, n$, have a common density with respect to the Lebesgue measure on $[0, 1]$, bounded and bounded away from zero. The noise has support in a ball $B(0, \sigma)$, with $\sigma < \rho(g)$, and admits a bounded density with respect to the Lebesgue measure, which is continuous on $\mathring{B}(0, \sigma)$, nondecreasing and symmetric, with a regularity condition on the boundary of the support. For an open curve (with endpoints), in addition, $|f(1) - f(0)|/2 > \sigma$. In the plane $\mathbb{R}^2$, the assumptions made allow to estimate the support S of the distribution of the observations, the boundary of this set S, in order to find its medial axis, which is the closure of the set of points in S that have at least two closest points in the boundary ∂S. In the same article, the authors also consider clutter noise, corresponding to the situation where one observes points sampled from a mixture density $(1 - \eta)u(x) + \eta h(x)$, where u is the uniform density over some compact set, and h is the density of points on the shape. Another additive model is investigated in Genovese et al. (2012b), for the estimation of manifolds without boundary, with dimension lower than the dimension of the ambient space, contained in a compact set. The model may be written

$$X_i = G_i + \varepsilon_i, \quad i = 1, \ldots, n,$$

where the random vectors G_i are drawn uniformly on the shape M, and the noise is drawn uniformly on the normal to the manifold, at distance at most $\sigma < \rho(M)$. The article Genovese et al. (2012c) is also dedicated to manifold estimation, under reach condition, first in a noiseless model, where the observations are exactly sampled on the manifold, according to some density with respect to the uniform distribution on

the manifold, and then in the presence of clutter noise. An additive noise model, with known Gaussian noise, is examined as well. This latter case is related to density deconvolution. Estimating manifolds without boundary, with low dimension and a lower bound on the reach, is also the purpose of Aamari and Levrard (2018, 2019). The points sampled on the manifold have a common density with respect to the d-dimensional Hausdorff measure of the manifold, which is bounded and bounded away from zero. In Aamari and Levrard (2018), estimation relies on Tangential Delaunay Complexes. It is performed in the noiseless case, with additive noise, bounded by σ, and under clutter noise. Aamari and Levrard (2019) deal with compact manifolds belonging to particular regularity classes. The authors examine the noiseless situation, as well as centered bounded noise perpendicular to the manifold. Estimators based on local polynomials are proposed.

To sum up, all these models involve strong conditions on the noise, which is either bounded, or of type clutter noise. Such assumptions allow the authors to derive rates of convergence. Here, we investigate a different situation, with a weak assumption on the noise. In particular, the noise does not need to be bounded. Regarding the regularity of the curve g, which has constant speed, there is no reach assumption, and g is not required to be injective. Although rates of convergence cannot be expected here, this weak framework is worth studying, since it is not obvious at first sight that it is even possible to build a convergent estimator without knowledge of either length or noise.

The estimation strategy relies on generalized empirical principal curves.

II.4.1.3 Extension of the notion of length-constrained principal curve

Recall that, according to Kégl et al. (2000), principal curves with length constraint are defined as follows. If X denotes a random vector with finite second moment, a principal curve is a continuous map $f^* : [0,1] \to \mathbb{R}^d$ minimizing under a length constraint the quantity

$$E\Big[\min_{t\in[0,1]} |X - f(t)|^2\Big] = E\Big[d^2(X, \mathrm{Im} f)\Big].$$

In order to allow for greater flexibility in the way we measure distances, we consider here the generalized principal curve notion introduced in Chapter II.1. Let $V : [0,\infty) \to [0,\infty)$ be a lower semi-continuous, strictly increasing function, continuous at 0, and such that $V(0) = 0$. For a random vector X such that $E[V(|X|)] < \infty$,

we are interested in the minimization over all curves with length at most L of the criterion

$$\Delta(f) = E\Big[V(d(X, \mathrm{Im} f))\Big].$$

This framework encloses for instance as particular cases the power functions $V(x) = x^p$, $p > 0$. An appropriate choice of V may enhance robustness. A typical example in this regard is the function defined by $V(x) = \frac{x}{1+x}$.

In a statistical context, one has at hand independent observations $X_1, \ldots, X_n$, and a generalized empirical principal curve is defined as a minimizer, under a length constraint, of the criterion

$$\frac{1}{n} \sum_{i=1}^{n} V(d\left(X_i, \mathrm{Im} f\right)).$$

We set up notation and introduce more formally the model in Section II.4.2. Then, in Section II.4.3, we describe the main result, that is the construction of a sequence of generalized empirical principal curves converging to the curve to be estimated in Hausdorff distance.

II.4.2 Definitions and notation

II.4.2.1 Notation

We denote by $d_H(A, B)$ the Hausdorff distance between two sets A and B.

Throughout, an interval (a, b) will denote an open interval of $[0, 1]$ equipped with the induced topology.

Let $\mathcal{D}$ denote a metric associated to weak convergence. For a probability measure μ and a closed set of probability measures $\mathcal{M}$, let $\mathcal{D}(\mu, \mathcal{M}) = \min_{\mu' \in \mathcal{M}} \mathcal{D}(\mu, \mu')$.

For two probability measures μ and μ', we define the bounded Lipschitz metric between μ and μ' by

$$|\mu - \mu'|_{BL} = \sup \left\{ |\mu(h) - \mu'(h)| : |h|_\infty \leq 1, \sup_{x \neq y} \frac{|h(x) - h(y)|}{|x - y|} \leq 1 \right\}.$$

II.4.2.2 Description of the model

Let $g : [0,1] \to \mathbb{R}^d$ be a curve with finite length and constant speed, such that the length equals the 1-dimensional Hausdorff measure.

Given $c > 0$, we define $\mathcal{M}_c$ as the closed family of probability distributions μ on $[0,1]$ satisfying $\mu \geq c\lambda$ on $[0,1]$.

For $n \geq 1$, we observe random vectors X_i^n, given by the model

$$X_i^n = g(U_i^n) + \varepsilon_i^n, \quad i = 1, \ldots, n, \tag{II.4.2}$$

where the U_i^n, $i = 1, \ldots, n$, are independent and for every $i = 1, \ldots, n$, the distribution μ_i^n of U_i^n belongs to $\mathcal{M}_c$.

Let $V : \mathbb{R}_+ \to \mathbb{R}_+$ be a lower semi-continuous, strictly increasing function, continuous at 0, and such that $V(0) = 0$. Moreover, we assume that V satisfies the following property: there exist a constant $C > 0$, such that, for every $(x, y) \in \mathbb{R}_+$

$$V(x + y) \leq C(V(x) + V(y)).$$

For a curve f, we define

$$\Delta_n(f) = \frac{1}{n} \sum_{i=1}^{n} V(d(X_i^n, \mathrm{Im} f)).$$

We also define a function $T(f, \cdot) : \mathbb{R}^d \to [0,1]$, by setting

$$T(f, x) = \max \arg\min_{t \in [0,1]} |x - f(t)|.$$

For every $L > 0$, let

$$G_n(L) = \min_{\mathscr{L}(f) \leq L} \Delta_n(f),$$

and let $\hat{f}_{n,L}$ denote an empirically optimal curve with length at most L, that is a random variable taking its values in $C([0,1])$ such that

$$\Delta_n(\hat{f}_{n,L}) = G_n(L).$$

Without loss of generality, we choose $\hat{f}_{n,L}$ L-Lipschitz. Let $\Lambda > 0$ and $\Lambda_n = \inf\{L \geq 0, G_n(L) = 0\}$.

II.4.3 Main result

We consider the estimation of the curve g in Model (II.4.2), using a sequence of generalized empirical principal curves. These optimal curves with respect to the criterion Δ_n are associated to a particular length selection described in the next statement, whose proof is to be found in Delattre and Fischer (2021).

Theorem II.4.3.1. *Let $g : [0,1] \to \mathbb{R}^d$ be a curve, such that $\mathscr{L}(g) \leq \Lambda < \infty$, and $|g'(t)| = \mathscr{L}(g)$ $dt-$a.e.. Assume that $\mathscr{L}(g) = \mathcal{H}^1(\mathrm{Im}g)$. We consider Model (II.4.2), with $\frac{1}{n}\sum_{i=1}^n V(|\varepsilon_i^n|)$ tending to 0 in probability as n tends to infinity. Let $\hat{L}_n$ be defined by*

$$\hat{L}_n \in \underset{L \in a_n\mathbb{N} \cap [0, \Lambda_n \wedge \Lambda]}{\arg\min} \left[V(L)\mathcal{D}\left(\frac{1}{n}\sum_{i=1}^n \delta_{T(\hat{f}_{n,L}, X_i^n)}, \mathcal{M}_c \right) + \Delta_n(\hat{f}_{n,L}) \right],$$

where $a_n > 0$ for every $n \geq 1$ and $a_n \to 0$ as $n \to \infty$. Then, $d_H(\mathrm{Im}\hat{f}_{n,\hat{L}_n}, \mathrm{Im}g)$ converges in probability to 0 as n tends to infinity.

Some comments are in order.

First, let us discuss the assumptions. The requirement $\mathscr{L}(g) \leq \Lambda < \infty$ is technical. It allows, in the proof, to consider limit points of the constructed sequence of empirical principal curves. From a applied point of view, this is not a limitation of the procedure. Indeed, in practice, we will always consider a finite grid for the length. Moreover, with a fixed number of observations, the minimal length needed to join all points is a finite upper bound for the length. The condition $\mathscr{L}(g) = \mathcal{H}^1(\mathrm{Im}g)$ ensures that the image of g is parameterized with minimal possible length. Indeed, there exist an infinite number of parameterizations, with infinite possibilities for the length. In words, generically, a portion of image of g cannot be traveled several times. The case were g is injective is a particular case. Nevertheless, here, an image with loops is allowed. We also require $|g'(t)| = \mathscr{L}(g)$ $dt-$a.e., which means that the image of g is parameterized with constant speed $\mathscr{L}(g)$. These assumptions about the parametrization allow to show a key relation between the distribution class $\mathcal{M}_c$ and its image by g (see Lemma II.4.4.3 below), the proof of which relies on the Cauchy-Crofton formula for the length of a rectifiable curve (Cauchy, 1850; Crofton, 1868).

Observe that the main strength of the result is that it provides a convergent estimator in a very general framework. Neither the length, nor the noise level, converging to 0 in a weak sense, is known. In particular, the noise is not supposed

bounded. Intuitively, considering a practical situation with a fixed number of observations, the same data cloud could arise from several different generative curves, longer or shorter, in a model with more or less noise. This illustrates the benefit of an estimator construction which does not require the knowledge of any of the two parameters. Apart from the upper bound Λ, which does not really need calibration in practice, as already mentioned, the procedure only depends on a single parameter, namely the constant c characterizing the class of possible sampling distributions $\mathcal{M}_c$.

It should be noticed that the theorem does not guarantee that the procedure allows to recover the true underlying length. Nevertheless, the proof below shows that the selected length cannot be too short: for all $\varepsilon > 0$ one has $P(\hat{L}_n \leq \mathscr{L}(g) - \varepsilon) \to 0$.

If g is a closed curve ($g(0) = g(1)$), then Theorem II.4.3.1 still holds when $\hat{f}_{n,L}$ is chosen as a closed empirically optimal curve with length at most L.

We sketch the proof of Theorem II.4.3.1 in the next section, divided into two parts, the first of which is dedicated to the equivalence linking $\mathcal{M}_c$ and its image by g, obtained thanks to results related to the Cauchy-Crofton formula.

II.4.4 Sketch of proof of the main result

II.4.4.1 Cauchy-Crofton formula and relation linking $\mathcal{M}_c$ to its image

As g is not supposed to be injective, but satisfies the condition $\mathscr{L}(g) = \mathcal{H}^1(\mathrm{Im}g)$, the next results, interesting in themselves, are crucial to enable a change of variables.

We first recall the Cauchy-Crofton formula.

Let $\mathcal{S}^{d-1} = \{z \in \mathbb{R}^d, |z| = 1\}$. For $\theta \in \mathcal{S}^{d-1}$ and $r \in [0, \infty)$, let

$$D_{\theta,r} = \{z \in \mathbb{R}^d \mid \langle \theta, z \rangle = r\}.$$

Lemma II.4.4.1 (Cauchy-Crofton formula)**.** *The length of a rectifiable curve $f : [0,1] \to \mathbb{R}^d$ is given by*

$$\mathscr{L}(f) = \frac{1}{c_d} \int_{\mathcal{S}^{d-1}} \int_0^\infty \mathrm{Card}(\{t \in [0,1], f(t) \in D_{\theta,r}\}) dr d\theta,$$

where $c_d > 0$ is a constant depending on the dimension d.

Equivalently, we may write:

$$\mathscr{L}(f) = \frac{1}{c_d} \int_{\mathcal{S}^{d-1}} \int_0^\infty \sum_{y \in \mathrm{Im} f \cap D_{\theta,r}} \mathrm{Card}(f^{-1}(\{y\})) dr d\theta.$$

The next equality corresponds to the Cauchy-Crofton formula applied to an open subset of the image of a curve.

Remark 10. Let $(a,b) \subset [0,1]$. Then,

$$\mathscr{L}(f|_{(a,b)}) = \frac{1}{c_d} \int_{\mathcal{S}^{d-1}} \int_0^\infty \mathrm{Card}(\{t \in (a,b), f(t) \in D_{\theta,r}\}) dr d\theta.$$

Since

$$\mathscr{L}(f|_{(a,b)}) = \int_0^1 \mathbf{1}_{(a,b)}(t) |f'(t)| dt,$$

we have

$$\int_0^1 \mathbf{1}_{(a,b)}(t) |f'(t)| dt = \frac{1}{c_d} \int_{\mathcal{S}^{d-1}} \int_0^\infty \sum_{t \in [0,1]} \mathbf{1}_{(a,b)}(t) \mathbf{1}_{\{f(t) \in D_{\theta,r}\}} dr d\theta.$$

Hence, by linearity, if (a_i, b_i), $i \geq 1$, are pairwise disjoint open intervals of $[0,1]$, we have

$$\int_0^1 \mathbf{1}_{\bigcup_{i\geq 1}(a_i,b_i)}(t) |f'(t)| dt = \frac{1}{c_d} \int_{\mathcal{S}^{d-1}} \int_0^\infty \sum_{t \in [0,1]} \mathbf{1}_{\bigcup_{i\geq 1}(a_i,b_i)}(t) \mathbf{1}_{\{f(t) \in D_{\theta,r}\}} dr d\theta.$$

We also need a Cauchy-Crofton-type formula for the curve g taking the form of an equality of measures. First, we may show an inequality, stated in the next lemma.

Lemma II.4.4.2. *Let $f : [0,1] \to \mathbb{R}^d$ be a rectifiable curve. Then, the trace of $\mathcal{H}^1$ on $\mathrm{Im} f$ satisfies $\mathcal{H}^1 \leq \gamma$, where γ is the measure defined on every Borel set $A \subset \mathrm{Im} f$ by*

$$\gamma(A) = \frac{1}{c_d} \int_{\mathcal{S}^{d-1}} \int_0^\infty \mathrm{Card}(A \cap D_{\theta,r}) dr d\theta.$$

In fact, for a curve g such that $\mathcal{H}^1(\mathrm{Im} g) = \mathscr{L}(g)$, we have $\mathcal{H}^1 = \gamma$. Indeed, on the one hand, $\mathcal{H}^1(\mathrm{Im} g) \leq \gamma(\mathrm{Im} g)$ by Lemma II.4.4.2, and on the other hand, $\gamma(\mathrm{Im} g) \leq \mathscr{L}(g)$ by the Cauchy-Crofton formula (Lemma II.4.4.1), so that the assumption $\mathcal{H}^1(\mathrm{Im} g) = \mathscr{L}(g)$ implies $\mathcal{H}^1(\mathrm{Im} g) = \gamma(\mathrm{Im} g)$. Thus, both measures have the same mass. Since $\mathcal{H}^1 \leq \gamma$ by Lemma II.4.4.2, they are equal.

Finally, there is an additional result useful for the change of variables. For g such that $\mathcal{H}^1(\mathrm{Im}g) = \mathscr{L}(g)$, since

$$\mathcal{H}^1(\mathrm{Im}g) = \gamma(\mathrm{Im}g) = \frac{1}{c_d}\int_{\mathcal{S}^{d-1}}\int_0^\infty \sum_{y\in \mathrm{Im}g\cap D_{\theta,r}} 1 dr d\theta,$$

and, by the Cauchy-Crofton formula for g,

$$\mathscr{L}(g) = \frac{1}{c_d}\int_{\mathcal{S}^{d-1}}\int_0^\infty \sum_{y\in \mathrm{Im}g\cap D_{\theta,r}} \mathrm{Card}(g^{-1}(\{y\})) dr d\theta,$$

we obtain that $\mathrm{Card}(g^{-1}(\{y\})) = 1$ for almost every y with respect to the trace of $\mathcal{H}^1$ on $\mathrm{Im}g$.

Thanks to all these properties around the Cauchy-Crofton formula, we are able to characterize the image by g of a distribution belonging to the class $\mathcal{M}_c$.

Lemma II.4.4.3. *Let $g : [0,1] \to \mathbb{R}^d$ be a curve such that $0 < \mathscr{L}(g) < \infty$, $|g'(t)| = \mathscr{L}(g)$ a.e., and $\mathcal{H}^1(\mathrm{Im}g) = \mathscr{L}(g)$. Let μ be a probability distribution supported in $[0,1]$, and let $c > 0$ denote a constant. Then,*

$$\mu \geq c\lambda \Leftrightarrow \forall A \subset \mathcal{B}(\mathbb{R}^d) \cap \mathrm{Im}g, \mu \circ g^{-1}(A) \geq c\frac{\mathcal{H}^1(A)}{\mathscr{L}(g)}. \tag{II.4.3}$$

Let us denote by $\mathcal{M}_c^g$ the family of probability distributions m on $\mathbb{R}^d$, with support $\mathrm{Im}g$, such that $\forall A \subset \mathcal{B}(\mathbb{R}^d) \cap \mathrm{Im}g, m(A) \geq c\frac{\mathcal{H}^1(A)}{\mathscr{L}(g)}$. Hence, the equivalence (II.4.3) means

$$\mu \in \mathcal{M}_c \Leftrightarrow \mu \circ g^{-1} \in \mathcal{M}_c^g.$$

Equipped with this key equivalence, we turn to the main part of the proof.

II.4.4.2 Overview of the proof of convergence

Note that $\mathscr{L}(g) - a_n < a_n \left\lfloor \frac{\mathscr{L}(g)}{a_n} \right\rfloor \leq \mathscr{L}(g)$. We set, for every $n \geq 1$,

$$f_n^* = \hat{f}_{n,a_n\left\lfloor \frac{\mathscr{L}(g)}{a_n}\right\rfloor},$$
$$\hat{f}_n = \hat{f}_{n,\hat{L}_n}.$$

There are two main steps, the first focusing on the f_n^* curve, before considering $\hat{f}_n$.

1. In the first step, the goal is to prove that

$$V(L)\mathcal{D}\left(\frac{1}{n}\sum_{i=1}^{n}\delta_{T(f_n^*,X_i^n)},\mathcal{M}_c\right)+\Delta_n(f_n^*)$$

converges in probability to 0 as n goes to infinity.

— For the term $\Delta_n(f_n^*)$, the convergence to 0 follows from the assumption on the noise.

— We set $\mathcal{D}_n^*(\mathcal{M}_c)=\mathcal{D}\left(\frac{1}{n}\sum_{i=1}^{n}\delta_{T(f_n^*,X_i^n)},\mathcal{M}_c\right)$. Let $\mathscr{L}(g)>0$. We show that the sequence $(\mathcal{D}_n^*(\mathcal{M}_c))_{n\geq 1}$ is tight, and that every limit point for the convergence in distribution is δ_0. Let $\nu_n^*=\frac{1}{n}\sum_{i=1}^{n}\delta_{\left(T(f_n^*,X_i^n),g(U_i^n)\right)}$. Since ν_n^* has compact support, f_n^* is $\mathscr{L}(g)$-Lipschitz, and $(f_n^*(0))_{n\geq 1}$ is tight, we obtain tightness of $(f_n^*,\nu_n^*)_{n\geq 1}$. Thus, $(\mathcal{D}_n^*(\mathcal{M}_c))_{n\geq 1}$ is tight. Considering a weakly convergent subsequence $(\mathcal{D}_{\sigma(n)}^*(\mathcal{M}_c))_{n\geq 1}$, we may assume by Prohorov's theorem, up to a further extraction, that $\left(f_{\sigma(n)}^*,\nu_{\sigma(n)}^*\right)_{n\geq 1}$ converges almost surely to a tuple (φ^*,ν^*), using Skorokhod's representation. Then, we show that the second marginal of ν^* belongs to $\mathcal{M}_c^g$. Moreover, $\nu^*\big(|Z-\varphi^*(T)|=\min_{t\in[0,1]}|Z-\varphi^*(t)|\big)=1$ almost surely, where (T,Z) denotes the identity on $[0,1]\times\mathbb{R}^d$, and $\int V(|z-\varphi^*(t)|)d\nu^*(t,z)=0$ almost surely, leading to $\mathrm{Im}g=\mathrm{Im}\varphi^*$ and $\mathscr{L}(\varphi^*)=\mathscr{L}(g)$ almost surely, by lower semi-continuity of the length. This allows to prove that φ^* satisfies the assumptions of Lemma II.4.4.3, and it follows that the first marginal of ν^* belongs to $\mathcal{M}_c$, meaning that $(\mathcal{D}_{\sigma(n)}^*(\mathcal{M}_c))_{n\geq 1}$ converges to 0.

2. In the second step, we first observe that, by definition of $\hat{L}_n$, we have the convergences in probability

$$\Delta_n(\hat{f}_n)\to 0 \tag{II.4.4}$$

$$V(\hat{L}_n)\mathcal{D}\left(\frac{1}{n}\sum_{i=1}^{n}\delta_{T(\hat{f}_n,X_i^n)},\mathcal{M}_c\right)\to 0. \tag{II.4.5}$$

Then, considering extractions of the sequence $(\hat{f}_n)_{n\geq 1}$ converging in distribution, we show that for every limit point φ of $(\hat{f}_n)_{n\geq 1}$, $d_H(\mathrm{Im}\varphi,\mathrm{Im}g)=0$ almost surely. To obtain that $\mathrm{Im}g=\mathrm{Im}\varphi$ almost surely, we exploit the convergence result (II.4.5), as well as a series of arguments similar to those in the first part, involving (II.4.4).

Chapter II.5

Improved rate of convergence in statistical learning

This chapter corresponds to a collaboration with Sylvain Delattre.

As the results have not been submitted to a journal yet, the complete proofs are collected in Section II.5.3.

II.5.1 Notation

We consider the space $(\mathbb{R}^d, |\cdot|)$. For an $\mathbb{R}^d$-valued signed measure $\nu = (\nu^1, \dots, \nu^d)$ on $[0,1]$, we set

$$|\nu| = \Big(\sum_{j=1}^{d} |\nu^j|_{TV}^2 \Big)^{1/2}$$

where $|\nu^j|_{TV}$ denotes the total variation norm of ν^j.

In the sequel, we consider a random vector X with values in $\mathbb{R}^d$, such that $P(X \in K) = 1$ for some compact set K. We let D denote the diameter of K. Let $X_1, \dots, X_n$ independent random vectors with the same distribution as X. For curves $f : [0,1] \to$

$\mathbb{R}^d$, we define

$$\Delta(f) = E[d^2(X, \mathrm{Im} f)],$$
$$\Delta_n(f) = \frac{1}{n}\sum_{i=1}^{n} d^2(X_i, \mathrm{Im} f).$$

Given a maximal length $L > 0$, we denote by f^* an optimal curve and by $\hat{f}_n$ an empirically optimal curve, that is

$$f^* \in \underset{f, \mathscr{L}(f) \leq L}{\arg\min}\, \Delta(f).$$
$$\hat{f}_n \in \underset{f, \mathscr{L}(f) \leq L}{\arg\min}\, \Delta_n(f).$$

The purpose is to study the rate of convergence to 0 for $\Delta\big(\hat{f}_n\big) - \Delta(f^*)$.

Note that the best rate of convergence so far has been obtained by Kégl et al. (2000), who constructed a sequence of curves $\hat{f}_n$, such that

$$\Delta(\hat{f}_n) - \Delta(f^*) = \mathcal{O}(n^{-1/3}).$$

II.5.2 Rate of convergence

Let $\mathcal{F}_n$ be the set of piecewise linear curves with n segments, with length at most L. We have

$$\begin{aligned}\Delta(\hat{f}_n) - \Delta(f^*) &= \Delta(\hat{f}_n) - \Delta_n(\hat{f}_n) + \Delta_n(\hat{f}_n) - \Delta_n(f^*) + \Delta_n(f^*) - \Delta(f^*) \\ &\leq \Delta(\hat{f}_n) - \Delta_n(\hat{f}_n) + \Delta_n(f^*) - \Delta(f^*) \\ &\leq \sup_{f \in \mathcal{F}_n} (\Delta(f) - \Delta_n(f)) + \Delta_n(f^*) - \Delta(f^*).\end{aligned}$$

Hoeffding's inequality applied to the second term $\Delta_n(f^*) - \Delta(f^*)$ shows that, for every $\varepsilon > 0$,

$$P\left(\Delta_n(f^*) - \Delta(f^*) \geq \varepsilon\right) \leq \exp\left(-\frac{2n\varepsilon^2}{D^4}\right).$$

The next proposition, proved in Section II.5.3, allows to control the first term $\sup_{f \in \mathcal{F}_n}(\Delta(f) - \Delta_n(f))$.

Proposition II.5.2.1. *For $k \geq 2$,*

$$P\left(\sup_{f\in\mathcal{F}_n}(\Delta(f)-\Delta_n(f)) \geq \frac{C_1}{k^2}\right) \leq (Dk^2)^{dk}\exp\left(-\frac{2nC_2}{k^4}\right),$$

where $C_1, C_2 \geq 0$ only depend on the dimension d and the diameter D.

This upper bound is based on the result stated in the next lemma, about approximation by piecewise linear curves.

Lemma II.5.2.1. *Let $f : [0,1] \to \mathbb{R}^d$ be such that $\mathscr{L}(f) \leq L$, f is right-differentiable on $[0,1)$ and left-differentiable on $(0,1]$ and there is a signed measure f'' such that*

- *$f''((s,t]) = f'_r(t) - f'_r(s)$ for $0 \leq s \leq t < 1$,*
- *$f''([0,1]) = 0$,*
- *$f''(\{0\}) = f'_r(0)$,*
- *$f''(\{1\}) = -f'_\ell(1)$,*

Then, for every $k \geq 2$, there exists a piecewise linear curve f_k with k segments (with breakpoints $f(s)$, $s \in \{t_0 = 0, \ldots, t_\ell, \ldots, t_k = 1\}$) such that

$$\max_{t\in[0,1]} |f(t) - f_k(t)| \leq \frac{A}{k^2},$$

for some constant $A \geq 0$.

To prove the lemma, in Section II.5.3, we first check that it is enough to consider smooth functions (C^2), and then, we show the piecewise linear approximation when f is C^2.

Finally, the main result may be stated as follows.

Proposition II.5.2.2. *The rate of convergence to 0 of $\Delta(\hat{f}_n) - \Delta(f^*)$ is $\mathcal{O}(n^{-2/5})$.*

Proof. To see this, let us optimize in k the quantity $(Dk^2)^{dk}\exp\left(-\frac{2nC_2}{k^4}\right)$. Taking the logarithm, we deal with $dk(\ln D + 2\ln k) - \frac{2nC_2}{k^4}$. Leaving out the constants, we obtain that k^5 should be of the order n, and, thus, k of the order $n^{1/5}$. Consequently, we end up with the rate of convergence $\mathcal{O}(n^{-2/5})$. □

II.5.3 Proofs

Proof of Lemma II.5.2.1. 1. Let us first check that it is enough to consider C^2 functions. We denote by f''_+ the positive part of the signed measure f'' and f''_- its negative part. For each $j = 1, \dots, d$, ${f^j}''_+$ and ${f^j}''_-$ are the positive and negative part of the j-th coordinate of f''.

Let Z^j denote a random variable with cumulative distribution function $t \mapsto {f^j}''_+(]0,t])$ and let h_n^j denote a density of the random variable $Z^j + \frac{1}{n}\xi$, where $\xi \sim \mathcal{N}(0,1)$. As $Z^j + \frac{1}{n}\xi$ converges in distribution to Z^j, we may write, for every $t \in [0,1]$ such that ${f^j}''_+(\{t\}) = 0$,

$$\int_{-\infty}^{t} h_n^j(s)ds \to P(Z^j \leq t) = {f^j}''_+(]0,t]).$$

Yet, $\int_{-\infty}^{0} h_n^j(s)ds = P(Z^j + \frac{1}{n}\xi < 0)$ tends to 0 as n tends to infinity since the random variable Z^j puts no mass at 0. Consequently, for a.e. $t \in [0,1]$,

$$\int_{0}^{t} h_n^j(s)ds \to {f^j}''_+(]0,t]).$$

By Lebesgue's dominated convergence theorem, we obtain

$$\int_0^1 \left| \int_0^t h_n^j(s)ds - {f^j}''_+(]0,t]) \right| dt \to 0.$$

Applying the same argument to the negative part ${f^j}''_-$, there also exist a function $\tilde{h}_n^j : [0,1] \to \mathbb{R}_+$ such that

$$\int_0^1 \left| \int_0^t \tilde{h}_n^j(s)ds - {f^j}''_-(]0,t]) \right| dt \to 0.$$

Hence,

$$\int_0^1 \left| \int_0^t (h_n^j(s) - \tilde{h}_n^j(s))ds - f^{j\prime\prime}(]0,t]) \right| dt \to 0.$$

Let us write

$$f^j(t) = f^j(0) + t{f^j}'_r(0) + \int_0^t f^{j\prime\prime}(]0,s])ds.$$

and set

$$f_n^j(t) = f(0) + tf^{j'}_r(0) + \int_0^t \int_0^s (h_n^j(u) - \tilde{h}_n^j(u))duds.$$

Then

$$\begin{aligned}\max_{t\in[0,1]} |f_n^j(t) - f^j(t)| &= \max_{t\in[0,1]} \left| \int_0^t \left(\int_0^s (h_n^j(u) - \tilde{h}_n^j(u))du - f^{j\prime\prime}(]0,s]) \right) ds \right| \\ &\leq \max_{t\in[0,1]} \int_0^t \left| \int_0^s (h_n^j(u) - \tilde{h}_n^j(u))du - f^{j\prime\prime}(]0,s]) \right| ds \\ &= \int_0^1 \left| \int_0^s (h_n^j(u) - \tilde{h}_n^j(u))du - f^{j\prime\prime}(]0,s]) \right| ds,\end{aligned}$$

which tends to 0. Consequently, $\max_{t\in[0,1]} |f_n(t) - f(t)| \to 0$. Moreover, for every $j = 1, \ldots, d$,

$$\int_0^1 |f_n^{j\prime\prime}(u)|du = \int_0^1 |h_n^j(u) - \tilde{h}_n^j(u)|du \leq |f^{j\prime\prime}|_{TV},$$

which is finite by assumption, and, hence,

$$|f_n''| = \left(\sum_{j=1}^d \left(\int_0^1 |f_n^{j\prime\prime}(u)|du \right)^2 \right)^{1/2} \leq \left(\sum_{j=1}^d |f^{j\prime\prime}|_{TV}^2 \right)^{1/2} < +\infty.$$

2. Assume that f is C^2. Let $k \geq 2$. For $t_0 = 0 < t_1 < \cdots < t_{k-1} < t_k = 1$, we consider the piecewise linear curve f_k, defined by the breakpoints $\{f(t_0), \ldots, f(t_k)\}$. We have

$$\max_{t\in[0,1]} |f(t) - f_k(t)| = \max_{\ell=1,\ldots,k} \max_{t\in[t_{\ell-1},t_\ell]} |f(t) - f_k(t)|.$$

For $t \in [t_{\ell-1}, t_\ell]$,

$$f_k(t) = f(t_{\ell-1}) + \frac{t - t_{\ell-1}}{t_\ell - t_{\ell-1}}(f(t_\ell) - f(t_{\ell-1})) = f(t_{\ell-1}) + \frac{t - t_{\ell-1}}{t_\ell - t_{\ell-1}} \int_{t_{\ell-1}}^{t_\ell} f'(u)du.$$

We write $f(t) = f(t_{\ell-1}) + \int_{t_{\ell-1}}^{t} f'(u)du$. Then,

$$\begin{aligned} f(t) - f_k(t) &= \int_{t_{\ell-1}}^{t} f'(u)du - \frac{t - t_{\ell-1}}{t_\ell - t_{\ell-1}} \int_{t_{\ell-1}}^{t_\ell} f'(u)du \\ &= \int_{t_{\ell-1}}^{t} \left(f'(s) - \frac{1}{t_\ell - t_{\ell-1}} \int_{t_{\ell-1}}^{t_\ell} f'(u)du \right) ds \\ &= \int_{t_{\ell-1}}^{t} \left(\frac{1}{t_\ell - t_{\ell-1}} \int_{t_{\ell-1}}^{t_\ell} (f'(s) - f'(u))du \right) dsg \end{aligned}$$

Thus,

$$\begin{aligned} |f(t) - f_k(t)| &\leq \int_{t_{\ell-1}}^{t} \left(\frac{1}{t_\ell - t_{\ell-1}} \int_{t_{\ell-1}}^{t_\ell} |f'(s) - f'(u)|du \right) ds \\ &\leq \int_{t_{\ell-1}}^{t} \left(\frac{1}{t_\ell - t_{\ell-1}} \int_{t_{\ell-1}}^{t_\ell} \int_{t_{\ell-1}}^{t_\ell} |f''(v)|dvdu \right) ds \\ &\leq (t_\ell - t_{\ell-1}) \int_{t_{\ell-1}}^{t_\ell} |f''(v)|dv, \end{aligned}$$

and we obtain

$$\max_{t \in [t_{\ell-1}, t_\ell]} |f(t) - f_k(t)| \leq (t_\ell - t_{\ell-1}) \int_{t_{\ell-1}}^{t_\ell} |f''(v)|dv.$$

The t_ℓ, $\ell = 0, \ldots, k$, have to be optimized to obtain the best approximating f_k. To minimize $\max_{\ell=1,\ldots,k}(t_\ell - t_{\ell-1}) \int_{t_{\ell-1}}^{t_\ell} |f''(v)|dv$, it is of interest to choose the t_ℓ so that almost all the products $(t_\ell - t_{\ell-1}) \int_{t_{\ell-1}}^{t_\ell} |f''(v)|dv$, $\ell = 1, \ldots, k$, are constant. Let $\varepsilon > 0$. We set $t_0 = 0$ and, for $\ell \in \mathbb{N}$,

$$t_\ell = \inf \left\{ t > t_{\ell-1} : (t - t_{\ell-1}) \int_{t_{\ell-1}}^{t} |f''(v)|dv = \varepsilon \right\} \wedge 1.$$

By assumption, there exists a constant $C \geq 0$, such that $\int_0^1 |f''(v)|dv \leq C$. Hence, for every $p \geq 1$ such that $t_p < 1$,

$$\sum_{\ell=1}^{p} \frac{1}{t_\ell - t_{\ell-1}} = \sum_{\ell=1}^{p} \frac{\int_{t_{\ell-1}}^{t} |f''(v)|dv}{\varepsilon} \leq \frac{C}{\varepsilon}.$$

Since $\frac{1}{p}\sum_{\ell=1}^{p}\frac{1}{t_\ell - t_{\ell-1}} \geq \frac{1}{\frac{1}{p}\sum_{\ell=1}^{p} t_\ell - t_{\ell-1}} = p$, we obtain $p^2 \leq \frac{C}{\varepsilon}$. This shows that 1 is attained. Now, let k be such that $t_{k-1} < 1$ and $t_k = 1$: we have $(k-1)^2 \leq \frac{C}{\varepsilon}$, that is $\varepsilon \leq \frac{C}{(k-1)^2}$. Finally, letting f_k be the piecewise linear curve associated to this choice of $t_0 = 0 < t_1 < \cdots < t_{k-1} < t_k = 1$, we get

$$\max_{t\in[0,1]} |f(t) - f_k(t)| \leq \frac{C}{(k-1)^2} \leq \frac{C+3}{k^2}.$$

□

Proof of Proposition II.5.2.1. Let $k \geq 2$. Let us build, for the sup-norm, a δ-net $\mathcal{F}_k^\delta$ of the set $\mathcal{F}_n$, with $\delta = 1/k^2$, made of piecewise linear curves with k segments. Lemma II.5.2.1 shows that every element f of $\mathcal{F}_n$ may be approximated in the sup-norm by a piecewise linear curve with k segments : there exists f_k such that

$$\max_{t\in[0,1]} |f(t) - f_k(t)| \leq \frac{A}{k^2},$$

where $A \geq 0$ is some constant. Moreover, noticing that, provided an adequate parametrization, the sup-norm between two piecewise linear curves determined by $k+1$ points, $\{x_0, \ldots, x_k\}$ and $\{y_0, \ldots, y_k\}$, is $\max_{\ell=0,\ldots,k} |x_\ell - y_\ell|$, we obtain that there exists $f_{k,\delta}$ belonging to the δ-net, such that

$$\max_{t\in[0,1]} |f_k(t) - f_{k,\delta}(t)| \leq \frac{\sqrt{d}}{k^2}.$$

Hence, there exists a constant $B \geq 0$, such that

$$\max_{t\in[0,1]} |f(t) - f_{k,\delta}(t)| \leq \frac{B}{k^2}.$$

Yet, for $f \in \mathcal{F}_n$,

$$\begin{aligned}
\Delta(f) - \Delta_n(f) &= \Delta(f) - \Delta_n(f) - \Delta(f_{k,\delta}) + \Delta_n(f_{k,\delta}) + \Delta(g_{k,\delta}) - \Delta_n(g_{k,\delta}) \\
&\leq E\left[\min_{t\in[0,1]} |X - g(t)|^2 - \min_{t\in[0,1]} |X - g_{k,\delta}(t)|^2\right] \\
&\quad + \frac{1}{n}\sum_{i=1}^{n}\left(\min_{t\in[0,1]} |X_i - g_{k,\delta}(t)|^2 - \min_{t\in[0,1]} |X_i - g(t)|^2\right) \\
&\quad + \Delta(g_{k,\delta}) - \Delta_n(g_{k,\delta}) \\
&\leq \frac{4DB}{k^2} + \Delta(g_{k,\delta}) - \Delta_n(g_{k,\delta}).
\end{aligned}$$

$$
\begin{aligned}
P\left(\sup_{g\in\mathcal{F}_n}(\Delta(g)-\Delta_n(g))\geq\frac{5DB}{k^2}\right) &\leq P\left(\sup_{g\in\mathcal{F}_k^\delta}(\Delta(g)-\Delta_n(g))+\frac{4DB}{k^2}\geq\frac{5DB}{k^2}\right)\\
&\leq P\left(\sup_{g\in\mathcal{F}_k^\delta}(\Delta(g)-\Delta_n(g))\geq\frac{DB}{k^2}\right)\\
&\leq |\mathcal{F}_k^\delta|\exp\left(-\frac{2nB^2}{k^4D^2}\right)\\
&= (Dk^2)^{dk}\exp\left(-\frac{2nB^2}{k^4D^2}\right).
\end{aligned}
$$

□

Part III

Cluster analysis, segmentation, and deconvolution

Chapter III.1

Clustering with Bregman divergences

The content of this chapter is extracted from an article published in the Journal of Multivariate Analysis *(Fischer, 2010) as part of my doctoral work. The goal is to provide the background for the next chapter.*

III.1.1 Introduction

Bregman divergences are a broad class of dissimilarity measures indexed by strictly convex functions. Introduced by Bregman (1967), these proximity functions are useful in a wide range of areas, among which statistical learning and data mining (Banerjee et al., 2005b; Cesa-Bianchi and Lugosi, 2006), computational geometry (Nielsen et al., 2007), natural sciences, speech processing and information theory (Gray et al., 1980). A lot of well-known proximity measures such as squared Euclidean, Mahalanobis, Kullback-Leibler and L^2 distances are particular cases of Bregman divergences. In $\mathbb{R}^d$, a Bregman divergence d_ϕ has the form

$$d_\phi(x, y) = \phi(x) - \phi(y) - \langle x - y, \nabla\phi(y)\rangle,$$

where $\langle\cdot,\cdot\rangle$ denotes the standard inner product, and $\nabla\phi(y)$ the gradient of ϕ at y. For example, taking for ϕ the squared Euclidean norm gives back the squared Euclidean distance. The same definition is valid in Hilbert spaces, and it even generalizes to Banach spaces (Alber and Butnariu, 1997; Csiszár, 1995; Frigyik et al., 2008b; Jones and Byrne, 1990).

Note that a Bregman divergence is not necessarily a true metric, since it may be asymmetric or fail to satisfy the triangle inequality. However, Bregman divergences fulfill an interesting projection property which generalizes the Hilbert projection on a closed convex set, as shown in Alber and Butnariu (1997); Bregman (1967). Banerjee et al. (2005b) have highlighted a relation between finite-dimensional Bregman divergences and exponential families, which will be discussed in the numerical section of Chapter III.2. These divergences are well-suited to measure proximity between observations arising from a mixture of such distributions. Moreover, the authors have shown that the standard k-means clustering algorithm (Lloyd, 1982) generalizes to these divergences. Consequently, clustering with Bregman divergences is particularly appropriate for mixtures from exponential families.

Following the approach of Banerjee et al. (2005b), we propose to use this class of proximity measures for quantization and clustering purposes. Quantization, also called lossy data compression in information theory, is the problem of replacing data by an efficient and compact representation which allows one to reconstruct the original observations with a certain accuracy. More formally, for a fixed integer $k \geq 1$, a random variable X with distribution μ, taking values in a set $\mathcal{X}$, will be represented by a so-called k-quantizer $q(X)$. Here q is a Borel measurable mapping from $\mathcal{X}$ to a finite subset of $\mathcal{X}$ with at most k elements. The error committed when representing X by $q(X)$ is given by the distortion

$$D(\mu, q) = E[d(X, q(X))],$$

where E denotes expectation with respect to the distribution μ and $d(\cdot, \cdot)$ is called the distortion measure. For more information on quantization, we refer the reader to Gersho and Gray (1992); Graf and Luschgy (2000); Linder (2002). In practice, the distribution μ is unknown, and $D(\mu, q)$ is replaced by the empirical criterion

$$D(\mu_n, q) = \frac{1}{n} \sum_{i=1}^{n} d(X_i, q(X_i)),$$

where $X_1, \ldots, X_n$ are independent random observations with distribution μ, and μ_n denotes the empirical measure associated with $X_1, \ldots, X_n$, i.e.,

$$\mu_n(A) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{X_i \in A\}}$$

for any Borel subset A of $\mathcal{X}$. The goal is to minimize $D(\mu_n, q)$ over all possible k-quantizers. This context corresponds to a clustering task, that is, the problem of

grouping data items in k meaningful classes, so that the classes are as homogeneous and at the same time as well separated as possible (Duda et al., 2000). The partitioning depends of course on the choice of the distance-like function $d(\cdot, \cdot)$ measuring the notion of proximity between points chosen. The constructed data-based quantizer q_n should have a clustering risk $D(\mu, q_n)$ "close" to the optimal risk $\inf_q D(\mu, q)$ as the size of the data set grows.

III.1.2 Context and assumptions

In Fischer (2010), we consider the problem of quantization and clustering when $d(\cdot, \cdot)$ is a general Bregman divergence $d_\phi(\cdot, \cdot)$ defined on a reflexive and separable Banach space. Our approach completes the more algorithmic-oriented results presented in Banerjee et al. (2005b). Here, we focus on data in the Euclidean space $(\mathbb{R}^d, |\cdot|)$, since this is the framework of the extension developed in the next chapter. Let $\langle \cdot, \cdot \rangle$ denote the associated inner product.

Definition III.1.2.1. *Let ϕ be a strictly convex $\mathcal{C}^1$ real-valued function defined on a convex set $\mathcal{C} \subset \mathbb{R}^d$. The Bregman divergence associated with ϕ is defined by*

$$d_\phi(x, y) = \phi(x) - \phi(y) - \langle x - y, \nabla\phi(y) \rangle,$$

where $\nabla f(u)$ denotes the gradient of f at u.

Although Bregman divergences are not true metrics, they satisfy some interesting properties, such as non-negativity and separation of points, convexity in the first argument and linearity (see Bregman, 1967; Frigyik et al., 2008a; Nielsen et al., 2007). Table III.1.1 collects some most common examples of Bregman divergences.

Sq	$\mathbb{R}$	x^2	$(x-y)^2$
Expo	$\mathbb{R}$	e^x	$e^x - e^y - (x-y)e^y$
Norm	$\mathbb{R}^+$	x^α	$x^\alpha + (\alpha-1)y^\alpha - \alpha x y^{\alpha-1}$
I-div uni	$\mathbb{R}^+$	$x \ln x$	$x \ln \frac{x}{y} - (x-y)$
Logistic	$[0,1]$	$x \ln x + (1-x)\ln(1-x)$	$x \ln \frac{x}{y} + (1-x) \ln \left(\frac{1-x}{1-y}\right)$
I-S uni	$(0,+\infty)$	$-\ln x$	$\frac{x}{y} - \ln \frac{x}{y} - 1$
Sq Eucl	$\mathbb{R}^d$	$\lvert x\rvert^2$	$\lvert x-y\rvert^2$
Mahal	$\mathbb{R}^d$	${}^t x A x$	${}^t(x-y)A(x-y)$
KL	$(d-1)$-simplex	$\sum_{\ell=1}^d x_\ell \ln x_\ell$	$\sum_{\ell=1}^d x_\ell \ln \frac{x_\ell}{y_\ell}$
I-div	$(\mathbb{R}^+)^d$	$\sum_{\ell=1}^d x_\ell \ln x_\ell$	$\sum_{\ell=1}^d x_\ell \ln \frac{x_\ell}{y_\ell} - \sum_{\ell=1}^d (x_\ell - y_\ell)$

Table III.1.1 – Some examples of Bregman divergences : squared loss, exponential loss, norm-like, I-divergence for $d = 1$, logistic loss, Itakura-Saito for $d = 1$, squared Euclidean distance, Mahalanobis distance, with the matrix A supposed to be positive definite, Kullback-Leibler, multivariate I-divergence.

Now, let X be a random variable with distribution μ, with values in $\mathcal{C}$. We make the following assumptions:

1. $E[|X|] < +\infty$.
2. $E[|\phi(X)|] < +\infty$ and, for all $c \in \mathcal{C}$, $E[|\langle \nabla\phi(c), X\rangle|] < +\infty$. This implies in particular that $E[d_\phi(X,c)] < +\infty$ for all c.

For $k \geq 1$, a k-quantizer is a Borel measurable mapping $q : \mathcal{C} \subset \mathbb{R}^d \to \mathbf{c}$, where $\mathbf{c} = \{c_1, \ldots, c_\ell\}$, $\ell \leq k$, is a subset of $\mathcal{C}$ called its codebook. In the sequel, the elements of $\mathbf{c}$ will also be named the centers associated to q. Every $x \in \mathcal{C}$ is represented by a unique $\hat{x} = q(x) \in \mathbf{c}$ and q induces a partition of $\mathcal{C}$ in cells $S_1, \ldots, S_\ell$. Each cell S_j is made of the elements of $\mathcal{C}$ whose image by q is c_j. Every k-quantizer is characterized by its codebook $\mathbf{c} = \{c_1, \ldots, c_\ell\}$ and its partition cells $S_1, \ldots, S_\ell$.

The error committed when representing X by $q(X)$ is assessed by the distortion

$$D(\mu, q) = E[d_\phi(X, q(X))] = \int_{\mathcal{C}} d_\phi(x, q(x)) d\mu(x). \tag{III.1.1}$$

Let

$$D^*(\mu) = \inf_{q \in \mathcal{Q}_k} D(\mu, q),$$

where $\mathcal{Q}_k$ is the set of all k-quantizers. To get a representation that is as accurate as possible, we look for an optimal quantizer, i.e., a quantizer q^* satisfying

$$D(\mu, q^*) = D^*(\mu).$$

In a statistical context, we only have at hand independent random observations $X_1, \ldots, X_n$ with distribution μ. The empirical distortion associated with $X_1, \cdots, X_n$ is given by

$$D(\mu_n, q) = \frac{1}{n} \sum_{i=1}^{n} d_\phi(X_i, q(X_i)), \tag{III.1.2}$$

where μ_n is the empirical measure. Observe that this is just the distortion (III.1.1) calculated with μ_n instead of μ. Clustering data into k groups means looking for an optimal quantizer q_n^* with respect to the empirical distortion (III.1.2).

Codebook and partition characterize a quantizer. As in the Euclidean case, it is easy to show that among all quantizers with same codebook, the best one (with respect to the distortion) is the nearest neighbor quantizer, whose partition $S_1, \ldots, S_\ell$ is the Voronoi partition, i.e.,

$$S_1 = \{x \in \mathcal{C}, d_\phi(x, c_1) \leq d_\phi(x, c_p), p = 1, \ldots, \ell\}$$

and for $j = 2, \ldots, \ell$,

$$S_j = \{x \in \mathcal{C}, d_\phi(x, c_j) \leq d_\phi(x, c_p), p = 1, \ldots, \ell\} \setminus \bigcup_{m=1}^{j-1} S_m$$

(see Linder, 2002). If an optimal quantizer exists, it is necessarily a nearest neighbor quantizer. Hence, in the sequel, we will always consider nearest neighbor quantizers. Conversely, given a partition $\{S_j\}_{j=1}^{\ell}$, with $\mu(S_j) > 0$ for $j = 1, \ldots, \ell$, the best quantizer is obtained by setting

$$c_j \in \arg\min_{c \in \mathcal{C}} E[d_\phi(X, c) | X \in S_j] \quad \text{for } j = 1, \ldots, \ell.$$

The next proposition is due to Banerjee et al. (2005a).

Proposition III.1.2.1. *Let d_ϕ be a Bregman divergence. If S is a Borel subset of $\mathcal{C}$ with $\mu(S) > 0$, the function*

$$c \mapsto E[d_\phi(X, c) | X \in S]$$

reaches its infimum at a unique point, $E[X|X \in S]$.

Thus, for every Bregman divergence, the minimizer is the conditional expectation, just as for the squared Euclidean distance. Observe that it is for instance the median instead of the expectation when the distortion measure is an L^1 norm.

Remark 11. The optimality of the Voronoi partition and Proposition III.1.2.1 are of computational interest (see Banerjee et al., 2005b). Indeed, even for squared Euclidean distance, minimizing the empirical distortion is generally a computationally hard problem, the complexity of an exact algorithm being exponential in the dimension of the space. In practice, a k-means type algorithm converging to local minima yields approximate solutions, and this adapts to general Bregman divergences. More precisely, given an initial codebook, which is made for instance of data items chosen at random, the algorithm proceeds by alternating between two steps. The first one consists in computing the Voronoi partition corresponding to the current centers. Then, during the second step, the new codebook is obtained by computing the mean of the data points falling in each cluster, according to Proposition III.1.2.1. Some numerical experiments about clustering with different Bregman divergences are available in Fischer (2015).

For ease of exposition, the results in the next section are stated in the case where $\mathcal{C}$ is closed and bounded.

III.1.3 Main results

III.1.3.1 Existence of an optimal quantizer

First, we focus on the question of the existence of an optimal quantizer q^*. Since a nearest neighbor quantizer is characterized by its codebook $\mathbf{c} = (c_1, \ldots, c_k)$, we may rewrite the distortion

$$D(\mu, \mathbf{c}) = E[\min_{j=1,\ldots,k} d_\phi(X, c_j)]$$

and look for an optimal codebook $\mathbf{c}^*$.

Thanks to the continuity of $y \mapsto d_\phi(x, y)$, the existence of a minimum may be proved via a compactness argument.

Theorem III.1.3.1. *There exists an optimal codebook* $\mathbf{c}^*$*, that is,*

$$D(\mu, \mathbf{c}^*) = D^*(\mu).$$

For the empirical measure μ_n, the existence of an optimal codebook $\mathbf{c}_n^*$ results anyway from the fact that the support of μ_n contain at most n points.

III.1.4 Convergence

III.1.4.1 Convergence of the distortion

Suppose that there exists an optimal codebook $\mathbf{c}_n^*$ that achieves the minimum of the empirical distortion $D(\mu_n, \mathbf{c})$. We study the asymptotic behavior of the "true" distortion $D(\mu, \mathbf{c})$ taken in $\mathbf{c} = \mathbf{c}_n^*$ with respect to the minimal distortion $D^*(\mu)$.

Assuming that $\mathbf{c}^*$ exists,

$$\begin{aligned} D(\mu, \mathbf{c}_n^*) - D^*(\mu) &= D(\mu, \mathbf{c}_n^*) - D(\mu, \mathbf{c}^*) \\ &= D(\mu, \mathbf{c}_n^*) - D(\mu_n, \mathbf{c}_n^*) + D(\mu_n, \mathbf{c}_n^*) - D(\mu, \mathbf{c}^*) \\ &\leq D(\mu, \mathbf{c}_n^*) - D(\mu_n, \mathbf{c}_n^*) + D(\mu_n, \mathbf{c}^*) - D(\mu, \mathbf{c}^*) \\ &\leq 2 \sup_{\mathbf{c} \in ri(\mathcal{C})^k} |D(\mu_n, \mathbf{c}) - D(\mu, \mathbf{c})|. \end{aligned}$$

Thus, to show that $D(\mu, \mathbf{c}_n^*)$ converges to $D^*(\mu)$ as n tends to infinity, it will be enough to prove that $\sup_{\mathbf{c} \in \mathcal{C}^k} |D(\mu_n, \mathbf{c}) - D(\mu, \mathbf{c})|$ tends to 0 as n tends to infinity.

Recall that we focus on the compact case.

Theorem III.1.4.1. *If, for all $c \in \mathcal{C}$, $|\nabla \phi(c)| \leq M$, then*

$$\lim_{n \to \infty} D(\mu, \mathbf{c}_n^*) = D^*(\mu) \quad a.s.$$

and

$$\lim_{n \to \infty} E[D(\mu, \mathbf{c}_n^*)] = D^*(\mu).$$

Remark 12. In the unbounded case, the existence of a minimizer $\mathbf{c}^*$ and the almost sure convergence of the distortion $\lim_{n \to \infty} D(\mu, \mathbf{c}_n^*) = D^*(\mu)$ may be proved by exploiting ideas in Sabin and Gray (1986) based on the Alexandroff one-point-compactification (see Fischer, 2010).

III.1.4.2 Rates of convergence

The previous section indicates that $D(\mu, \mathbf{c}_n^*)$ gets close to the minimal distortion when the sample size grows. However, it gives no information about the rates of convergence. To establish an non-asymptotic upper bound, we use the following inequality

$$E[D(\mu, \mathbf{c}_n^*)] - \inf_{\mathbf{c} \in \mathcal{C}^k} D(\mu, \mathbf{c}) \leq 2E \left[\sup_{\mathbf{c} \in \mathcal{C}^k} (D(\mu_n, \mathbf{c}) - D(\mu, \mathbf{c})) \right],$$

together with the next theorem, which is proved in Fischer (2010) by resorting to Rademacher averages as complexity measure for a function class (see, e.g., Bartlett et al., 2001).

Theorem III.1.4.2. *We have:*

$$E\left[\sup_{\mathbf{c}\in\mathcal{C}^k}\left(D(\mu_n,\mathbf{c})-D(\mu,\mathbf{c})\right)\right] \leq \frac{2k}{\sqrt{n}}\Big(\sup_{c\in\mathcal{C}}|\langle\nabla\phi(c),c\rangle-\phi(c)|+\sup_{c\in\mathcal{C}}|\nabla\phi(c)|(E|X|^2)^{1/2}\Big).$$

Corollary III.1.4.1. *If for all $c\in\mathcal{C}$, $|\langle\nabla\phi(c),c\rangle-\phi(c)|\leq M_1$ and $|\nabla\phi(c)|\leq M_2$, then*

$$E[D(\mu,\mathbf{c}_n^*)]-D^*(\mu)\leq\frac{4k}{\sqrt{n}}\left(M_1+M_2(E[|X|]^2)^{1/2}\right),$$

and thus

$$E[D(\mu,\mathbf{c}_n^*)]-D^*(\mu)\leq\frac{4k}{\sqrt{n}}\left(M_1+M_2\mathrm{diam}(\mathcal{C})\right).$$

Note that Corollary III.1.4.1 yields dimension-free upper bounds, which is a valuable feature when dealing with high-dimensional data.

Chapter III.2

Robust Bregman clustering

This chapter is the summary of a collaboration with Claire Brécheteau (IRMAR, Université Rennes 2) and Clément Levrard (LPSM, Université de Paris), published in The Annals of Statistics *(Brécheteau et al., 2021).*

III.2.1 Introduction

The theoretical performance of k-means has been studied mostly without particular considerations about the presence of outliers among the observations. As for classical mean estimation, the clustering procedure is actually quite sensitive to outliers. However, in real data sets, the source signal is often corrupted by noise.

To tackle this problem, a trimmed k-means heuristic is introduced by Cuesta-Albertos et al. (1997) (see also Garcí a Escudero et al. (2008) for trimmed clustering with Mahalanobis distances). Our purpose is to extend this trimming approach to the general framework of clustering with Bregman divergences.

Section III.2.2 describes a robust clustering technique, based on the computation of a trimmed empirically optimal codebook for some fixed trim level). Theoretical properties of the trimmed empirical codebook are then exposed. In Section III.2.3, a modified Lloyd's type algorithm is proposed, along with a heuristic to select both the trim level h and the number k of clusters from data. The numerical performances of our algorithm are then investigated on simulated and real data.

III.2.2 Trimming approach for Bregman clustering

Recall that a Bregman divergence is defined d_ϕ is defined for all $x, y \in \mathcal{C}$ by $d_\phi(x, y) = \phi(x) - \phi(y) - \langle \nabla\phi(y), x - y\rangle$, where ϕ is a strictly convex $\mathcal{C}^1$ real-valued function defined on a convex set $\mathcal{C} \subset \mathbb{R}^d$.

Let μ be a distribution on $\mathbb{R}^d$, and $\mathbf{c} = (c_1, \ldots, c_k)$ a codebook. The clustering performance of $\mathbf{c}$ is measured via its distortion, namely

$$D(\mu, \mathbf{c}) = \int \min_{j=1,\ldots,k} d_\phi(u, c_j) d\mu(u).$$

In the statistical context, when the distribution μ is unknown, but we have at hand a sample $X_1, \ldots, X_n$, only the empirical distortion may be computed, given by

$$D(\mu_n, \mathbf{c}) = \frac{1}{n} \sum_{i=1}^{n} \min_{j=1,\ldots,k} d_\phi(X_i, c_j)$$

where μ_n denotes the empirical measure.

To tackle the noise issue, the general idea proposed in Cuesta-Albertos et al. (1997); Gordaliza (1991) consists in searching, for a trim level $h \in (0, 1]$, both for a codebook and a subset of μ-mass not smaller than h (trimming set).

For a measure ν on $\mathbb{R}^d$, we write $\nu \leq \mu$ if $\nu(A) \leq \mu(A)$ for every Borel set A. Let μ_h denote the set $\mathcal{P}_h = \{\frac{1}{h}\nu \mid \nu \leq \mu, \nu(\mathbb{R}^d) = h\}$, and $\mathcal{P}_{+h} = \cup_{s \geq h} \mathcal{P}_s$. By analogy with Cuesta-Albertos et al. (1997), optimal trimming sets and codebooks are designed to achieve the optimal h-trimmed k-variation:

$$V_{k,h} = \inf_{\tilde{\mu} \in \mathcal{P}_{+h}} \inf_{\mathbf{c} \in \mathcal{C}^k} D(\tilde{\mu}, \mathbf{c}).$$

The h-trimmed k-variation may be thought of as the k-points optimal distortion of the best "denoised" version of μ, with denoising level $1 - h$. For instance, in a mixture setting, if $\mu = \gamma\mu_0 + (1 - \gamma)N$, where μ_0 is a signal supported by k points and N is a noise distribution, then, provided that $h \leq \gamma$, $V_{k,h} = 0$.

Now that the setting has been introduced, we present the main results on Bregman h-trimmed clustering obtained in Brécheteau et al. (2021). To avoid the use of cumbersome notation, for the sake of clarity, we decided to summarize them in a relatively informal way.

We assume that ϕ is $\mathcal{C}^2$ and that the closure of the convex hull of the support of μ is a subset of the interior of $\mathcal{C}$.

The first result concerns the existence of an optimal codebook.

Theorem III.2.2.1 (Existence of an optimal codebook)**.** *Let* $0 < h < 1$*, and assume that* $\int |u| d\mu(u) < +\infty$*. Then, there exists an optimal codebook for trimmed* k*-means.*

The next statement is about the convergence of empirically optimal trimmed codebooks.

Theorem III.2.2.2 (Convergence of the distortion)**.** *Assume that* μ *is absolutely continuous with respect to the Lebesgue measure and satisfies* $\int |u|^p d\mu(u) < \infty$ *for some* $p > 2$*. Then, almost sure convergence of a sequence of empirically optimal trimmed codebooks in terms of distortion is ensured.*

Moreover, provided that the optimal h*-trimmed distortion decreases when adding a codebook, the convergence rate is* $\frac{1}{\sqrt{n}}$*.*

More specifically, rate of convergence results are stated in Brécheteau et al. (2021) on the one hand in large probability, and on the other hand in expectation, provided additional technical assumptions on ϕ. Note that the moment condition required for the rate of convergence is $\int |u|^2 d\mu(u) < \infty$.

III.2.3 Numerical experiments

III.2.3.1 Description of the algorithm

The algorithm introduced in this section is inspired by the trimmed version of Lloyd's algorithm (Cuesta-Albertos et al., 1997), and is also a generalization of the Bregman clustering algorithm (Banerjee et al., 2005b, Algorithm 1). We assume that we observe $X_1, \ldots, X_n$, and that the mass parameter h equals $\frac{q}{n}$ for some positive integer q. We also let C_j denote the subset of $\{1, \ldots, n\}$ corresponding to the j-th cell S_j.

Bregman trimmed k-means algorithm

1. **Input:** $X_1, \ldots, X_n$, q, k.
2. **Initialization:** Sample c_1, c_2,..., c_k from the observations without replacement, $\mathbf{c}^{(0)} \leftarrow (c_1, \ldots, c_k)$.
3. **Iterations:** Repeat until stabilization of $\mathbf{c}^{(t)}$.
   - $N_q^{(t)} \leftarrow$ indices of the q smallest values of $d_\phi(x, \mathbf{c}^{(t-1)})$, $x \in \{X_1, \ldots, X_n\}$.
   - For $j = 1, \ldots, k$, $C_j^{(t)} \leftarrow S_j(\mathbf{c}^{(t-1)}) \cap N_q^{(t)}$.
   - For $j = 1, \ldots, k$, $c_j^{(t)} \leftarrow \frac{\sum_{x \in C_j^{(t)}} x}{|C_j^{(t)}|}$.
4. **Output:** $\mathbf{c}^{(t)}$, $C_1^{(t)}, \ldots, C_k^{(t)}$.

Since the algorithm may be quite sensitive to initialization, several random starts will be proceeded in practice. Note that, in full generality, the output of the algorithm is not a global minimizer.

III.2.3.2 Exponential Mixture Models

In this section we describe the generative models to which he algorithm will be applied. We consider mixtures of distributions belonging to some exponential family. As presented in Banerjee et al. (2005b), a distribution from an exponential family may be associated to a Bregman divergence via Legendre duality of convex functions. For a particular distribution, the corresponding Bregman divergence is more adapted for clustering than other divergences.

Recall that an exponential family associated to a proper closed convex function ψ defined on an open parameter space $\Theta \subset \mathbb{R}^d$ is a family of distributions $\mathcal{F}_\psi = \{P_{\psi,\theta} \mid \theta \in \Theta\}$, such that, for all $\theta \in \Theta$, $P_{\psi,\theta}$, defined on $\mathbb{R}^d$, is absolutely continuous with respect to some distribution P_0, with Radon-Nikodym density $p_{\psi,\theta}$ defined for all $x \in \mathcal{C}$ by

$$p_{\psi,\theta}(x) = \exp(\langle x, \theta \rangle - \psi(\theta)).$$

The function ψ is called the cumulant function and θ is the natural parameter. For this model, the expectation of $P_{\psi,\theta}$ may be expressed as $m(\theta) = \nabla\psi(\theta)$. We define

$$\phi(m) = \sup_{\theta \in \Theta} \{\langle m, \theta \rangle - \psi(\theta)\}.$$

By Legendre duality, for all m such that ϕ is defined, we get $\phi(m) = \langle \theta(m), m \rangle - \psi(\theta(m))$, with $\theta(m) = \nabla\phi(m)$. The density of $P_{\psi,\theta}$ with respect to P_0 can be rewritten using the Bregman divergence associated to ϕ as follows:

$$p_{\psi,\theta}(x) = \exp(-d_\phi(x, m) + \phi(x)).$$

In the next experiments, we use Gaussian, Poisson, Binomial and Gamma mixture distributions and the corresponding Bregman divergences. Table III.2.1 presents the 4 densities together with the functions ψ and ϕ, as well as the associated Bregman divergences d_ϕ.

Distribution	$p_{\psi,\theta}(x)$	θ	$\psi(\theta)$	m
Gaussian	$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-a)^2}{2\sigma^2}\right)$	$\frac{a}{\sigma^2}$	$\frac{\sigma^2}{2}\theta^2$	a
Poisson	$\frac{\lambda^x \exp(-\lambda)}{x!}$	$\log(\lambda)$	$\exp(\theta)$	λ
Binomial	$\frac{N!}{x!(N-x)!} q^x (1-q)^{N-x}$	$\log\left(\frac{q}{1-q}\right)$	$N \log\left(1 + \exp(\theta)\right)$	Nq
Gamma	$\frac{x^{k-1}\exp(-\frac{x}{b})}{\Gamma(k)b^k}$	$-\frac{k}{m}$	$k \log\left(-\frac{1}{\theta}\right)$	kb

	$\phi(m)$	$d_\phi(x, m)$
Gaussian	$\frac{1}{2\sigma^2}m^2$	$\frac{1}{2\sigma^2}(x-m)^2$
Poisson	$m\log(m) - m$	$x \log\left(\frac{x}{m}\right) - (x-m)$
Binomial	$m \log\left(\frac{m}{N}\right) + (N-m)\log\left(\frac{N-m}{N}\right)$	$x \log\left(\frac{x}{m}\right) + (N-x)\log\left(\frac{N-x}{N-m}\right)$
Gamma	$-k + k\log\left(\frac{k}{m}\right)$	$\frac{k}{m}\left(m\log\left(\frac{m}{x}\right) + x - m\right)$

Table III.2.1 – Exponential family distributions and associated Bregman divergences

The next remark gives an illustration of the connection between divergences and distributions in a simple case.

Remark 13. We let $k = 2$, $\theta_1 \neq \theta_2$, $z_1^*, \ldots, z_n^*$ be hidden labels in $\{1, 2\}$, and $X_1, \ldots, X_n$ be independent random variables with density

$$\mathbf{1}_{z_i^*=1} p_{\psi,\theta_1}(x) + \mathbf{1}_{z_i^*=2} p_{\psi,\theta_2}(x),$$

where $p_{\psi,\theta_j}(x) = \exp(-d_\phi(x, m_j) + \phi(x))$, for $j \in \{1, 2\}$. The parameters of this model are $(z_i^*)_{i\in\{1,\ldots,n\}}, \theta_1, \theta_2$.

Let $z_{i,j}$, $i \in \{1, \ldots, n\}$, $j \in \{1, 2\}$, be defined by $z_{i,j} = 1$ if X_i is assigned to class j and 0 otherwise. Also set

$$m = \sum_{i=1}^n z_{i,1},\ n - m = \sum_{i=1}^n z_{i,2},\ \bar{X}_1 = \sum_{i=1}^n X_i z_{i,1}/m,\ \bar{X}_2 = \sum_{i=1}^n X_i z_{i,2}/(n-m).$$

Maximizing the log-likelihood of the observations corresponds to maximizing in $(z_{i,j})_{i,j}$:

$$\ln \prod_{i=1}^{n} \exp\left[-z_{i,1} d_\phi\left(X_i, \bar{X}_1\right) - z_{i,2} d_\phi\left(X_i, \bar{X}_2\right) + \phi(X_i)\right] \\ = -\sum_{i=1}^{n} z_{i,1} d_\phi\left(X_i, \bar{X}_1\right) - \sum_{i=1}^{n} z_{i,2} d_\phi\left(X_i, \bar{X}_2\right) + \sum_{i=1}^{n} \phi(X_i).$$

On the other hand, since optimal codebooks are local means of their Bregman-Voronoi cells, performing Bregman k-means clustering is equivalent to minimizing

$$\sum_{i=1}^{n} z_{i,1} d_\phi\left(X_i, \bar{X}_1\right) + \sum_{i=1}^{n} z_{i,2} d_\phi\left(X_i, \bar{X}_2\right).$$

Thus, clustering with a Bregman divergence is the same as computing a maximum likelihood estimator for a distribution from the exponential family.

III.2.3.3 Calibration of trimming parameter and number of clusters

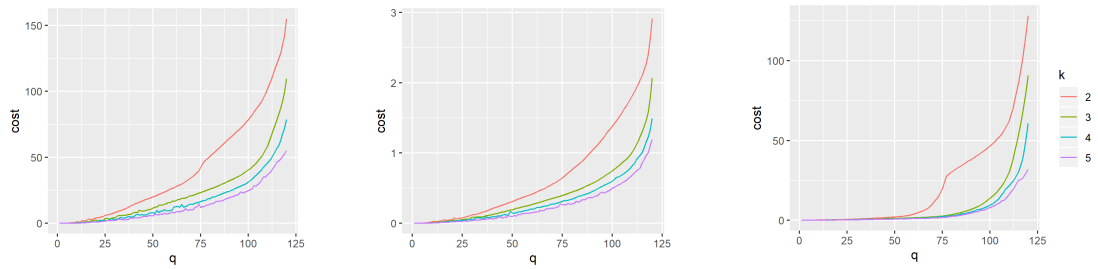
When the number of clusters k is known beforehand, we propose the following heuristic to select the trimming parameter q, that is, the number of points in the sample which are assigned to a cluster and not considered as noise. We let q vary from 1 to the sample size n, plot the curve $q \mapsto cost[q]$ where $cost[q]$ denotes the optimal empirical distortion at trimming level q, and choose $q^\star$ by seeking for a cut-point on the curve. Indeed, when the parameter q gets large enough, it is likely that the procedure starts assigning outliers to clusters

When both k and q are unknown, we propose to select these two parameters following the same principle as in the algorithm `tclust` (Fritz et al., 2012). We draw, for different values of k, the cost curves $q \mapsto cost_k[q]$, for $1 \leq q \leq n$. For each curve, the q's for which there is an abrupt slope increase can correspond to cases where outliers are assigned to clusters, or where some small clusters are included in the set of signal points (if k is chosen too small). In the sequel, we split $\{1, \ldots, n\}$ into several bins $\{q_j, \ldots, q_{j+1}\}$. On every such bin, we select a k that provides a significant cost decrease, as well as the q yielding a slope jump. Note that this heuristic may result in several possible pairs (k, q), corresponding to different point of views, depending on which data points are considered as outliers or not.

III.2.3.4 Simulated mixture distributions

We replicate some experiments in Banerjee et al. (2005b) and Fischer (2010), with additional noise. We consider mixture models of Gaussian, Poisson, Binomial, Cauchy and Gamma distributions in $\mathbb{R}^2$. Namely, we sample 100 points from $X = (X^1, X^2)$, where X^1 and X^2 are independent, distributed according to a mixture distribution with 3 components. In each case, the means of the components are set to $10, 20, 40$. The weights of the components are $(1/3, 1/3, 1/3)$. We also consider a mixture of 3 different components: Gamma, Gaussian and Binomial, with respective means 10, 20 and 40. In the Gaussian case, the standard deviation of the components is set to 5, in the Binomial case, the number of trials is set to 100 and in the Gamma case the shape parameter is set to 40. For Gaussian and Cauchy distributions, we discard negative realizations. 20 uniformly distributed outliers are added.

First, we run the clustering algorithm with 20 random starts for each of these noisy mixture distributions, using the "corresponding" divergence, and also make the same experiment for the Cauchy distribution with squared Euclidean distance. We select k and q following the heuristic exposed in Section III.2.3.3. According to Figure III.2.1, this leads to the choice $k = 3$, $q = 104$ for the Gaussian mixture and $q = 110$ for the other mixtures. The resulting partitions for the selected parameters are depicted in Figure III.2.2.



Gaussian mixture Poisson mixture Cauchy mixture

Figure III.2.1 – Cost curves for selection of k and q

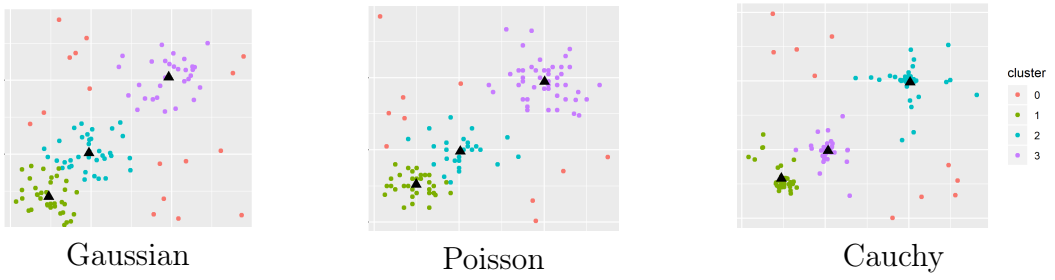


Figure III.2.2 – Clustering associated to the selected parameters k and q, where cluster 0 refers to noise

Then, we compare the proposed method, in every case, to clustering with other Bregman divergences (including trimmed k-means Cuesta-Albertos et al. (1997)), trimmed k-median (Cardot et al., 2013), `tclust` (Fritz et al., 2012), a single linkage procedure with ad hoc outlier removal, the ToMATo algorithm (Chazal et al., 2013) and `dbscan` (Hahsler et al., 2019). Quality of the partitions is assessed via the normalized mutual information (hereafter NMI, Strehl and Ghosh (2002)) with respect to the ground truth clustering, where the "noise" points are all considered as one single cluster. The results in terms of NMI for 1000 repetitions are exposed in Figure III.2.3: Algorithm 1 refers to our method with $q = 110$ and $k = 3$.

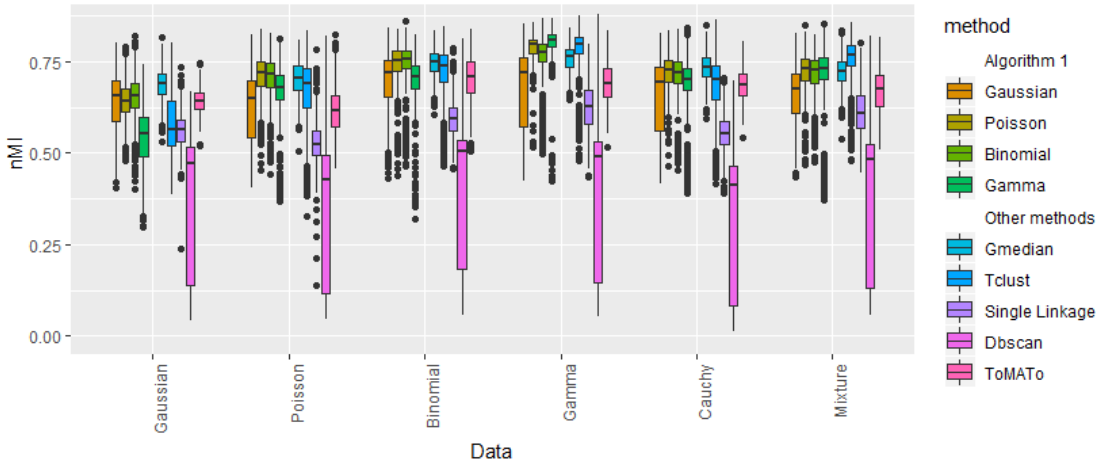


Figure III.2.3 – Comparison of robust clustering methods, for mixtures of Binomial, Gamma, Gaussian, Poisson, Cauchy, and the "heterogeneous" mixture

Note that our algorithm with the proper Bregman divergence (almost) systematically outperforms other clustering schemes.

III.2.3.5 Authors stylometric clustering

In this section, we perform clustering on texts based on stylometric descriptors (Arnold and Tilton, 2015, Section 10). To be more precise, raw data consist in 26 annotated texts from 4 authors (Mark Twain, Sir Arthur Conan Doyle, Nathaniel Hawthorne and Charles Dickens). These texts are available as supplementary material for Arnold and Tilton (2015), and are framed as a sequence of lemmatized string characters (for instance "be" and "is" are instances of the same lemma "be"). Following Arnold and Tilton (2015), we base our stylometric comparison on lemmas corresponding to nouns, verbs and adverbs, and split every original text in chunks of size 5000 of such lemmas that will be considered as data points. Then the 50 overall most frequent lemmas are chosen, and every chunk is described as the vector of counts of these lemmas within it. Thus, signal points consists of 189 count vectors with dimension 50, originating from 4 different authors.

Outliers are produced using the same process for the 8 State of the Union Addresses given by Barack Obama (available in `obama` dataset from package CleanNLP in R), resulting in 5 additional points, and for the King James Version of the Bible (available on Project Gutenberg) that we preliminary lemmatize using the CleanNLP package, resulting in 15 more additional points. Our final dataset consists of the 189 signal points and the 20 outlier points described above. Of course, these 20 outliers might also be thought of as two additional small clusters with size 5 and 15.

Since every individual lemma count can be modeled as a Poisson random variable in the random character sequence model (Evert, 2004), the appropriate Bregman divergence for this dataset is likely to be the associated divergence, that is relative entropy. In the sequel, we compare our method with relative entropy to trimmed k-means, trimmed k-medians, and t-clust.

In Figure III.2.4, we draw the cost of our method as a function of q, for different cluster numbers k. According to this figure, several choices of k and q are possible. For values of q up to 175, the significant jumps in the risk function are for $k = 3$ and $k = 6$. For $k = 3$, we obtain $q = 175$, whereas for $k = 6$ it seems that no data point might be considered as an outlier. When q ranges from 175 to 193, the significant jumps are for $k = 4$ and $k = 6$, and another possible choice is then $k = 4$ and $q = 188$. When q is larger than 193, the only significant jump is for $k = 6$. To

summarize, the pairs $(k = 3, q = 175)$, $(k = 4, q = 188)$, $(k = 6, q = n = 209)$ seem reasonable.

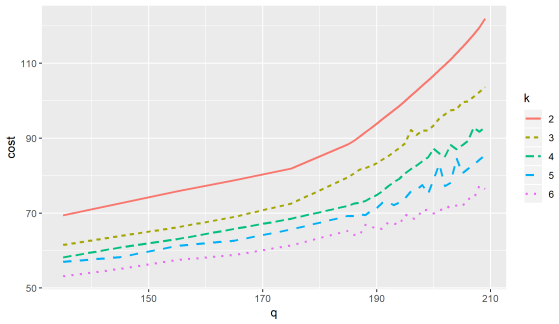


Figure III.2.4 – Cost curves for authors clustering with relative entropy.

These three solutions correspond to the 3 natural trimmed partitions: clustering only 3 authors writings (Twain writings being considered as outliers), clustering the 4 authors writings and removing the outliers from the Bible and B. Obama addresses, and at last, clustering the six sources of writings (none of them being considered as noise). The two latter situations are depicted in Figure III.2.5 (projection onto linear discriminant analysis factorial plane).

Figure III.2.5 – Author stylometric clustering with relative entropy.

For $k = 6$ and $q = 209$, our clustering globally retrieves the corresponding author. When $k = 4, q = 188$ is chosen, outliers are correctly identified and only one sample text from C. Dickens is labeled as outlier. The sample points seem on the whole well classified, which is assessed by an NMI of 0.7347. This performance is compared

with the other clustering algorithms in Table III.2.2. Note that q has been chosen to minimize the NMI, leading to $q = 190$ for trimmed k-means, $q = 202$ for trimmed k-medians, and $q = 184$ for `tclust`.

Method	trimmed 4-means	trimmed 4-medians	tclust	Poisson
NMI	0.5336	0.4334	0.4913	0.7347

Table III.2.2 – Comparison of robust clustering methods

The associated partitions for k-median and `tclust` are depicted in Figure III.2.6, where we see that these two methods fail in correctly identifying outliers.

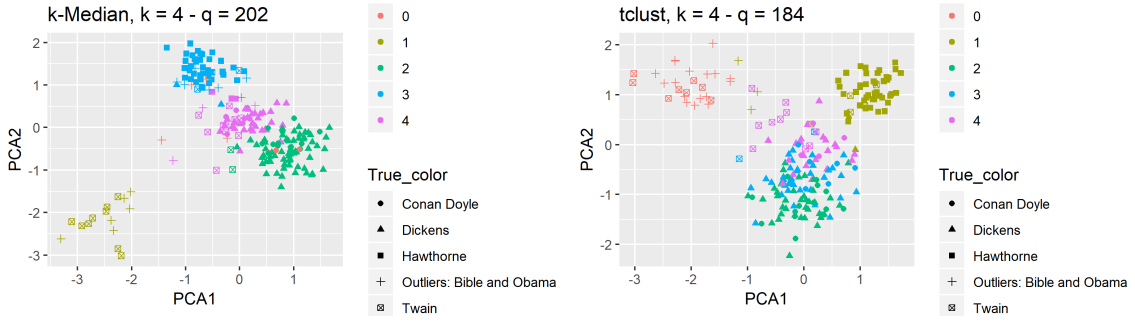


Figure III.2.6 – Author stylometric clustering with trimmed k-median and tclust

Chapter III.3

A change-point study

This chapter corresponds to an article written in collaboration with Dominique Picard (LPSM, Université de Paris), published in the Electronic Journal of Statistics *(Fischer and Picard, 2020).*

III.3.1 Introduction

An important problem in the vast domain of statistical learning is the question of unsupervised classification of high-dimensional data. Here, we address a framework where the change between classes occurs on a time scale, that is, a change-point setting. For the sake of simplicity, we consider the case of exactly two classes.

More specifically, we assume that there exists a change-point τ: before $n\tau$, the observations are in a certain state, after $n\tau$, they are in another state.

We observe independent random vectors $Y_1, \dots, Y_n$ with values in $\mathbb{R}^d$, such that

$$\begin{aligned} &Y_i = \theta_i + \eta_i, \quad \eta_i \sim \mathcal{N}(0, \sigma^2 I_d), \text{ independent}, \quad 1 \leq i \leq n, \\ &\forall i \leq n\tau, \quad \theta_i = \theta^-, \\ &\forall i > n\tau, \quad \theta_i = \theta^+. \end{aligned}$$

In practice, such a framework may model for instance the monitoring of patients, where the variables Y_i are a bunch of d biological, chemical or clinical observations collected each ten minutes (for example) on a patient, and $n\tau$ reflects a time of change in the patient's condition.

Our aim is to estimate the change-point τ. Using the maximum likelihood approach, also known in this context as CUSUM method, we derive rates of convergence for the estimation of τ, under conditions specified below.

For high-dimensional data, from a computational point of view, there is an obvious need for dimension reduction when estimating τ. Without such a step, the segmentation algorithm might be unstable or even not work at all. Here, we consider the dimension reduction problem from a theoretical point of view: one might suspect that it should always be better to keep the whole data, to get the best precision on the estimation of the change-point, but dimension reduction proves to be useful also on the theoretical side.

We will show that sparsity assumptions, as well as smoothing adaptive methods, can be directly borrowed from the nonparametric statistical inference and fruitfully applied in this context.

The change-point problem has a long history, going back at least to Page (1955). For an introduction to the field, the reader may refer for instance to the monographs and articles by Shiryaev (1978), Ritov (1990), Müller (1992), Basseville and Nikiforov (1993), Brodsky and Darkhovsky (1993), Carlstein et al. (1994), Csörgő and Horváth (1997) or Horváth and Rice (2014) (see also further references in Fischer and Picard, 2020).

Regarding the high-dimensional context, Jirak (2015) considers several dependent change-point tests and studies the behavior of the maximum over all test statistics as both the sample size and the number of tests tend to infinity. Cho and Fryzlewicz (2015) propose a sparse version of binary segmentation. The high dimension problem is addressed through changes in cross-covariance in Lavielle and Teyssière (2006), Aue et al. (2009), Bücher et al. (2014), Preuss et al. (2015), Cribben and Yu (2017). In Soh and Chandrasekaran (2017), convex optimization is used to perform regularization for solving the high dimensional change-point problem. In Chen and Zhang (2015); Shi et al. (2017), graph-based approaches which are efficient in high dimension are designed. Jin et al. (2016) proposes a method based on a statistics inspired by Hotelling's T^2 statistics. Enikeeva and Harchaoui (2019) consider high-dimensional change-point detection, from the testing point of view.

Here, we study the performance of the procedure from a minimax point of view, in a high-dimensional context. This approach provides an evaluation of the best expectable performances in a particular framework, and the aim is then to provide a procedure attaining these performances.

Minimax estimation is considered already in Korostelev (1987), in the Gaussian

white noise model. High-dimensional change-point problems are also studied in Korostelev and Lepski (2008), which proposes an asymptotically minimax estimator of the change-point location, when the Euclidean norm of the gap tends to infinity as the dimension d goes to infinity.

Our approach is deeply connected to this article and can be considered as a continuation of this project. The main difference is that in Korostelev and Lepski (2008), the authors do not question the dimension reduction problem and do not consider the same estimation method. Moreover, Korostelev and Lepski (2008) make the assumption that the change-point only occurs after a known number of observations and before another known number of observations. Another related reference is the article by Wang and Samworth (2018), who proposed a two-stage procedure based on a projection followed by a univariate change point estimation algorithm applied to the projected data, providing rates of convergence for the estimator of the change-point location.

There is a crucial difference between the present setting and the one in Korostelev and Lepski (2008) since knowing that some observations are in the first or last state allows to provide an efficient estimation of this state. Without this assumption, the problem gets more difficult. However, we prove that, for a fixed dimension, up to a logarithmic term, the maximum likelihood method, has a minimax rate of convergence. Moreover, we show that if the data is sparse, in a Sobolev sense, there exists an optimal dimension reduction, depending on the sparsity constants. Since these constants are not known in practice, we provide a procedure relying on the Lepski method, which behaves as well as if the sparsity constants were known. The proposed method has the advantage of being performed off-line, before the main segmentation step.

Numerical experiments are provided in Section III.3.4.

III.3.2 Main result: minimax convergence rate

III.3.2.1 Change-point model and assumptions

Let $n \geq 3$. We observe n independent signals $Y_1, \ldots, Y_n$. We assume that each signal Y_i, $i = 1, \ldots, n$, is a d dimensional vector: for every i, $Y_i = (Y_{i,1}, \ldots, Y_{i,d})$ is a random vector with values in $\mathbb{R}^d$.

We suppose that there exist a change-point $0 < \tau < 1$ and two vectors θ^- and

θ^+ of $\mathbb{R}^d$, such that the model is given by

$$\begin{aligned} &Y_i = \theta_i + \eta_i, \quad \eta_i \sim \mathcal{N}(0, \sigma^2 I_d), \text{ independent}, \quad 1 \leq i \leq n, \\ &\forall i \leq n\tau, \quad \theta_i = \theta^-, \\ &\forall i > n\tau, \quad \theta_i = \theta^+. \end{aligned} \tag{III.3.1}$$

III.3.2.2 Estimation method

We are interested in the behavior of the maximum likelihood estimator, also called in this case CUSUM estimator:

$$\hat{k}(d) = \underset{k \in \{2,\dots,n-2\}}{\arg\min} \left\{ \sum_{i=1}^{k} \sum_{j=1}^{d} \left(Y_{i,j} - \frac{1}{k} \sum_{\ell=1}^{k} Y_{\ell,j} \right)^2 + \sum_{i=k+1}^{n} \sum_{j=1}^{d} \left(Y_{i,j} - \frac{1}{n-k} \sum_{\ell=k+1}^{n} Y_{\ell,j} \right)^2 \right\},$$

To prove some of our results, we will need the following sparsity conditions.

Condition on the means

For $s > 0$, we define

$$\Theta(s, L) := \left\{ \theta \in \mathbb{R}^d, \ \sup_{K \in \mathbb{N}^*} K^{2s} \sum_{k \geq K} (\theta_k)^2 \leq L^2 \right\}.$$

We will suppose that θ^- and θ^+ are in $\Theta(s, L)$.

Remark 14. This assumption expresses a form of sparsity of the coefficients which is standard in nonparametric settings. It corresponds to conditions which are directly connected to the regularity of the function to be estimated in nonparametric estimation. In the more general setting of a high-dimensional physical observation, it is commonly accepted to solve learning problems by introducing sparsity constraints (see for instance Huang et al., 2011; Johnstone and Silverman, 2004, among many others). These constraints can take various forms: we choose here the Sobolev type, which is among the simplest forms to handle technically. Note that it reflects an

ordering: the first coefficients are supposed to be more important than the last ones. This is quite a reasonable assumption since, generally, modeling of high-dimensional or functional data via a basis expansion results in such a situation.

Example 1. Let $s = 1/2$ and $L = 1$. Assume that θ is defined by $\theta_k = \frac{1}{\sqrt{2k}}$ for $k = 1, \ldots, k_{\max}$ and $\theta_k = 0$ for $k > k_{\max}$. Then, for every $K = 1, \ldots, k_{\max}$,

$$K^{2s} \sum_{k \geq K} (\theta_k)^2 \leq L^2 = K \sum_{k=K}^{k_{\max}} \left(\frac{1}{\sqrt{2}k} \right)^2 \leq 1.$$

Hence, $\theta \in \Theta(1/2, 1)$.

To end up this section we introduce the following important parameter:

$$\varepsilon := \min\{\tau, 1 - \tau\}.$$

We consider a not degenerate case (there exists a change), ε is a strictly positive quantity, which measures the potential lack of information at the border of the interval $[0, 1]$. If the theoretical performances of the procedures depend on ε, the procedure is agnostic to ε, which therefore is not be supposed to have a known lower bound.

III.3.2.3 Dimension reduction for the estimation of τ

Our aim is to determine if it is efficient to perform a dimension reduction when estimating the change-point τ. More specifically, we will investigate the effect of replacing the vectors $Y_i = (Y_{i,1}, \ldots, Y_{i,d})$, $i \leq n$ (called "raw data"), by, for $p < d$, $Y_i(p) := (Y_{i,1}, \ldots, Y_{i,p})$, $i \leq n$, the vectors of $\mathbb{R}^p$ composed of the p first coordinates of Y_i.

For each projection dimension p, we may define:

$$\hat{k}(p) = \underset{k \in \{2, \ldots, n-2\}}{\arg\min} \left\{ \sum_{i=1}^{k} \sum_{j=1}^{p} \left(Y_{i,j} - \frac{1}{k} \sum_{\ell=1}^{k} Y_{\ell,j} \right)^2 + \sum_{i=k+1}^{n} \sum_{j=1}^{p} \left(Y_{i,j} - \frac{1}{n-k} \sum_{\ell=k+1}^{n} Y_{\ell,j} \right)^2 \right\}.$$

We set

$$\hat{\tau}(p) = \frac{\hat{k}(p)}{n}.$$

In the sequel, we will use the notation

$$\Delta^2 := \sum_{j=1}^{d} (\theta_j^- - \theta_j^+)^2 = |\theta^+ - \theta^-|^2.$$

We also define, for $p \leq d$,

$$\Delta_p^2 := \sum_{j=1}^{p} (\theta_j^- - \theta_j^+)^2, \quad \Psi_n(p, \Delta_p) = \frac{\sigma^2}{n\Delta_p^2} \left(1 \vee \frac{\sigma^2 p}{n\Delta_p^2}\right).$$

Example 2. Let θ^+ and θ^- be defined by $\theta_k^+ = \frac{1}{\sqrt{2k}}$ and $\theta_k^- = -\frac{1}{\sqrt{2k}}$. The rate Ψ_n is plotted as a function of n and p in Figure III.3.1.

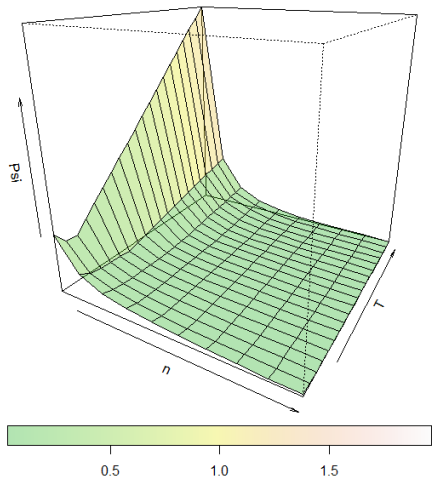


Figure III.3.1 – Example of plot of Ψ_n as a function of n and p ($n = 1, \ldots, 10$; $p = 1, \ldots, 20$)

The next result describes the behavior of the estimated change-point $\hat{\tau}(p)$.

Proposition III.3.2.1. *For any $\gamma > 0$, there exist constants $\kappa(\gamma, \varepsilon)$ and $c(\gamma, \varepsilon)$ such that, if*

$$\Delta_p^2 \geq c(\gamma, \varepsilon) \frac{\sigma^2 \ln(n)}{n},$$

then

$$P\Big(|\hat{\tau}(p) - \tau| \geq \kappa(\gamma, \varepsilon)\ln(n)\Psi_n(p, \Delta_p)\Big) \leq cn^{-\gamma},$$

where c is an absolute constant.

The proof is available in Fischer and Picard (2020). It is based on Lemma III.3.2.1 below, which shows that $\hat{\tau}$ may be written using Gaussian and chi-square random variables, and it makes extensive use of standard concentration inequalities for these distributions.

Lemma III.3.2.1. *Under Model (III.3.1), the estimator $\hat{\tau}(p)$ may be written:*

$$\hat{\tau}(p) = \underset{t\in\{\frac{2}{n},\ldots,\frac{n-2}{n}\}}{\arg\min} K^p(t),$$

where

$$K^p(t) = -\sum_{j=1}^{p}\sigma^2 V_j^2(t) - \sum_{j=1}^{p}\sigma^2 W_j^2(t) + \sum_{j=1}^{p}\delta_j^2\frac{(nt - n\tau)n\tau}{nt} + 2N_1(t) - 2N_2(t).$$

Here, $V_j^2(t)$ and $W_j^2(t)$, $j = 1, \ldots, p$, are independent $\chi^2(1)$ random variables,

$$N_1(\tau) = N_2(\tau) = 0, \text{ for every } t \neq \tau,$$

$$N_1(t) \sim \mathcal{N}\left(0, \sum_{j=1}^{p}\sigma^2(nt - n\tau)\delta_j^2\right) \text{ and } N_2(t) \sim \mathcal{N}\left(0, \sum_{j=1}^{p}\frac{\sigma^2(nt - n\tau)^2\delta_j^2}{nt}\right).$$

Using this expression $\hat{\tau}(p)$, we then build an upper bound for

$$P\left(|\hat{\tau}(p) - \tau| \geq \lambda\Psi_n\right),$$

where $\lambda > 0$. Since

$$\begin{aligned}
&P\left(|\hat{\tau}(p) - \tau| \geq \lambda\Psi_n\right) \\
&\leq P\left(\inf_{\frac{k}{n} - \tau \geq \lambda\Psi_n} K^p\left(\frac{k}{n}\right) < K^p(\tau)\right) + P\left(\inf_{\frac{k}{n} - \tau \leq -\lambda\Psi_n} K^p\left(\frac{k}{n}\right) < K^p\tau\right),
\end{aligned}$$

we focus on evaluating one term of the right-hand side, the second being treated in a symmetrical way.

Remark 15.

Minimax optimality and parameter ε. Thanks to Korostelev and Lepski (2008), we know that $\Psi_n(p, \Delta_p)$ is the minimax rate in this framework. Compared to their result, we are loosing a logarithmic factor. However, in Korostelev and Lepski (2008), a fixed lower bound on ε is supposed to be known, whereas our estimator $\hat{\tau}(p)$ is adaptive in ε. Looking carefully at the proofs, the constants $c(\varepsilon, \gamma)$ and $\kappa(\varepsilon, \gamma)$ can be taken proportional to $\frac{(\gamma+1)}{\varepsilon^2}$.

Note that $\ln(n)$ could be substituted by any sequence r_n, provided that the factor $n^{-\gamma}$ is simultaneously replaced by $\exp(-cr_n)$ in Proposition III.3.2.1. (Here, c is a constant which can be made explicit in the proof.)

Fast rate/slow rate. For $p = d$, $\Psi_n(d, \Delta_d) = \Psi_n(d, \Delta)$. The rate is composed of two different regimes: a "fast one" $\frac{\sigma^2 \ln(n)}{n\Delta^2}$, which does not depend on the dimension d and a "slow one" $\frac{\sigma^4 \ln(n) d}{(n\Delta^2)^2}$, which is rapidly deteriorating with the dimension. From the results above, we deduce that if $c(\gamma, \varepsilon)\frac{\sigma^2 \ln(n)}{n} \leq \Delta^2 < \frac{\sigma^2 d}{n}$, the rate of convergence is $\frac{\sigma^4 \ln(n) d}{(n\Delta^2)^2}$ (small gap, slow performances), whereas if $\Delta^2 \geq \frac{\sigma^2 d}{n} \vee c(\gamma, \varepsilon)\frac{\sigma^2 \ln(n)}{n}$, it is $\frac{\sigma^2 \ln(n)}{n\Delta^2}$. This last rate is obviously much better, and with this latter condition on Δ, taking $p = d$ (so raw data) allows to obtain the best rate $\frac{\sigma^2 \ln(n)}{n\Delta^2}$.

If we introduce sparsity assumptions, that is means θ^- and θ^+ belonging to $\Theta(s, L)$, then, for p such that $\Delta^2 \geq 8L^2 p^{-2s}$, Δ_p and Δ are comparable, in the sense that $\Delta_p^2 \geq \Delta^2/2$. Indeed, $\Delta^2 - \Delta_p^2 = \sum_{j=p+1}^{d} (\theta_j^- - \theta_j^+)^2 \leq 4p^{-2s}L^2$, so that

$$\frac{\Delta_p^2}{\Delta^2} \geq 1 - \frac{4p^{-2s}L^2}{\Delta^2} \geq 1/2.$$

Note that, if Δ_p and Δ are comparable, then $\Psi_n(p, \Delta_p) \sim \Psi_n(p, \Delta)$ becomes much easier to analyze. In particular, we see that, again, there are two regimes —a slow one and a fast one— and the dependence in p becomes easily understandable: $\frac{\sigma^2 \ln(n)}{n\Delta^2}$ for $p \leq \frac{n\Delta^2}{\sigma^2}$, and $\frac{\sigma^4 \ln(n) p}{(n\Delta^2)^2}$ for larger p's.

Note that two different convergence rates have also been highlighted in other change-point settings, for instance in Wang et al. (2021).

III.3.2.4 Minimax convergence rate under sparsity condition

The following theorem is an immediate consequence of Proposition III.3.2.1 in the case where one assumes the sparsity condition on the means.

Theorem III.3.2.1. *We consider Model (III.3.1), with the means θ^+ and θ^- in $\Theta(s, L)$. For any $\gamma > 0$, there exist constants $\kappa(\gamma, \varepsilon)$ and $c(\gamma, \varepsilon)$ such that, if*

$$\Delta^2 \geq \left[2c(\gamma, \varepsilon)\frac{\sigma^2 \ln(n)}{n} \vee 8L^2 p^{-2s}\right],$$

then

$$P\Big(|\hat{\tau}(p) - \tau| \geq \kappa(\gamma, \varepsilon) \ln(n) \Psi_n(p, \Delta)\Big) \leq cn^{-\gamma}.$$

If, now,

$$\Delta^2 \geq \left[2c(\gamma, \varepsilon)\frac{\sigma^2 \ln(n)}{n} \vee 8L^2 p^{-2s} \vee \frac{\sigma^2 p}{n}\right], \tag{III.3.2}$$

then

$$P\Big(|\hat{\tau}(p) - \tau| \geq \kappa(\gamma, \varepsilon)\frac{\sigma^2 \ln(n)}{n\Delta^2}\Big) \leq cn^{-\gamma}.$$

Here, c is an absolute constant.

Here, contrary to Proposition III.3.2.1, the rate is $\Psi_n(p, \Delta)$ instead of $\Psi_n(p, \Delta_p)$. The price to pay is then, as is intuitive, that p should be large enough. In the second statement, we look at the condition on Δ and p to obtain the fast rate. For Δ fixed, we see that p must not be too large or too small.

Condition (III.3.2) contains two terms: one is increasing in p, one decreasing. Hence it can be optimized leading to

$$p_{opt} \sim p_s := \left(\frac{8L^2 n}{\sigma^2}\right)^{\frac{1}{1+2s}}. \tag{III.3.3}$$

We obtain the next corollary, corresponding to this projection dimension p_s.

Corollary III.3.2.1. *Under the conditions above, for any $\gamma > 0$, there exist constants $\kappa(\gamma, \varepsilon)$ and $c(\gamma, \varepsilon)$ such that, if*

$$\Delta^2 \geq \left[2c(\gamma, \varepsilon)\frac{\sigma^2 \ln(n)}{n} \vee \left(\frac{\sigma^2}{n}\right)^{\frac{2s}{1+2s}} \left(8L^2\right)^{\frac{1}{1+2s}}\right],$$

$$P\Big(|\hat{\tau}(p_s) - \tau| \geq \kappa(\gamma, \varepsilon)\frac{\sigma^2 \ln(n)}{n\Delta^2}\Big) \leq cn^{-\gamma}.$$

Interpretation is that the quantity $\left[2c(\gamma,\varepsilon)\frac{\sigma^2\ln(n)}{n} \vee \left(\frac{\sigma^2}{n}\right)^{\frac{2s}{1+2s}} (8L^2)^{\frac{1}{1+2s}}\right]$ is the minimal gap between the two regimes ensuring that the "fast" rate $\frac{\sigma^2\ln(n)}{n\Delta^2}$ can be obtained, with an appropriate projection dimension.

Remark 16.

1. We see here that there is an obvious advantage in reducing the dimension, since it allows to obtain the best rate with less restricting conditions on the gap Δ.
2. Observe that σ^2 often has the form $\frac{\sigma_0^2}{d}$. In this case, the rate of convergence is of the order $\left(\frac{nd}{\sigma_0^2}\right)^{\frac{-2s}{1+2s}}\Delta^{-2}$.
3. Formula (III.3.3) indicates that the optimal p depends on the sparsity constant s, which is rarely known, which shows the interest of looking for an adaptive procedure selecting an optimal p (without knowing the regularity s). Since any adaptive smoothing performed individually on each signal Y_i (such as thresholding, Lasso...) would lead at best to a projection dimension of the form $p_{opt} = \left(\frac{d}{\sigma_0^2}\right)^{\frac{1}{1+2s}}$, inducing a lost of a polynomial factor in n in the rates, a procedure performing the smoothing globally (off-line) will be more efficient.

III.3.3 Fast convergence rate: adaptive choice of p

We propose a strategy based on Lepski's method to adaptively choose p, taking inspiration from nonparametric statistics.

Note that Wang and Samworth (2018) also prove adaptivity under slightly different conditions.

III.3.3.1 Lepski's procedure

The Lepski method (Lepski, 1991, 1992, 1993) is a strategy allowing to obtain adaptivity in various functional estimation settings such as white noise model, regression or density estimation. In these models, minimax optimality is linked with the regularity assumptions imposed on the functions which are estimated. In these nonparametric problems, there is a balance to obtain between a "variance" term, typically of the form $\frac{p}{n}$, and a "bias" term, typically of the form p^{-2s}. The Lepski procedure proposes to choose the minimal p among those such that an estimated version of the bias is below a bound.

For the sake of clarity, let us first recall the classical Lepski procedure in the standard Gaussian white noise model. We will not describe the procedure in the original form presented in the first articles, corresponding to kernel estimation methods, but a version adapted to orthogonal series estimation methods, which is more suitable for a transposition to our case.

Consider the following model:

$$Z_j = \beta_j + \varepsilon_j,\ j = 1, \dots, d, \tag{III.3.4}$$

where the ε_j's are independent random variables with distribution $\mathcal{N}(0, \nu^2)$. The Lepski procedure for choosing the optimal projection dimension p consists in defining $\hat{p}$ as follows:

$$\hat{p} := \min \left\{ k \geq 1 : \forall d \geq j \geq m \geq k, \sum_{\ell=m}^{j} (Z_\ell)^2 \leq C_{\mathcal{L}} j \nu^2 \ln(d) \right\},$$

where $C_{\mathcal{L}}$ is a tuning constant of the procedure.

In our change-point setting, a transformation of the data is necessary to fall into the frame of Model (III.3.4). We will apply the Lepski method to a surrogate data vector built on the whole observation.

III.3.3.2 Preprocessing

Using the complete data set (so off-line), we define a surrogate data vector, which will be used to find an optimal $\hat{p}$. We assume, for the sake of simplicity, that n is even; otherwise, the modifications are elementary.

We set:

$$Z_j = \frac{1}{n} \sum_{i=1}^{n} Y_{i,j} - \frac{2}{n} \sum_{i=1}^{n/2} Y_{i,j}, \quad j = 1, \dots, d.$$

This vector $Z = (Z_j)_{1 \leq j \leq d}$ is a special case of Model (III.3.4), where

$$\beta_j = (1-\tau)(\theta_j^+ - \theta_j^-)\mathbf{1}_{\{\tau \geq 1/2\}} + \tau(\theta_j^+ - \theta_j^-)\mathbf{1}_{\{\tau < 1/2\}}, \quad j = 1, \dots, d$$

$$\varepsilon_j = \sum_{i=1}^{n/2} \frac{-1}{n} \eta_{i,j} + \sum_{i=n/2+1}^{n} \frac{1}{n} \eta_{i,j}, \quad j = 1, \dots, d$$

$$\nu^2 = \frac{\sigma^2}{n}.$$

III.3.3.3 Adaptive convergence rate

We consider the Lepski procedure applied to the vector Z, producing a projection dimension $\hat{p}$. This parameter is then just plugged in the maximum likelihood procedure for estimating $\hat{\tau}$. Lepski's procedure provides a projection dimension $\hat{p}$ which, with overwhelming probability, is smaller than the optimal p_s (defined in (III.3.3) above) and such that the bias $\Delta^2 - \Delta_{\hat{p}}$ is controlled. The next theorem, proved in Fischer and Picard (2020), states that the method leads to an optimal selection, up to logarithmic terms: though the optimal p_s is unknown, we are able to achieve the same convergence rate as in Corollary III.3.2.1.

Theorem III.3.3.1. *We consider Model (III.3.1) and assume that θ^+ and θ^- belong to $\Theta(s, L)$. We suppose that there exists a constant $\alpha > 0$ such that*

$$\frac{n}{\sigma^2} \geq \alpha \ln d.$$

For any $\gamma > 0$, there exist constants $C_{\mathcal{L}}$, $\kappa(\gamma, \varepsilon)$, $c(\gamma, \varepsilon)$ and R such that, letting

$$\hat{p} := \min \left\{ k \geq 1 : \forall d \geq j \geq m \geq k, \sum_{\ell=m}^{j} (Z_\ell)^2 \leq C_{\mathcal{L}} j \frac{\sigma^2}{n} \ln(d \vee n) \right\},$$

if

$$\Delta^2 \geq 2c(\gamma, \varepsilon) \frac{\sigma^2 \ln(n)}{n} \vee R \left(\frac{\sigma^2 \ln(d \vee n)}{n} \right)^{\frac{2s}{1+2s}},$$

then

$$P\left(|\hat{\tau}(\hat{p}) - \tau| \geq \kappa(\gamma, \varepsilon) \frac{\sigma^2 \ln(n)}{n\Delta^2} \right) \leq cn^{-\gamma}.$$

Remark 17. The sense of this theorem is that, if $C_{\mathcal{L}}$, which should be considered as a tuning parameter of the method, is large enough, then an optimal result is obtained.

III.3.4 Numerical study

In this section, we provide some simulations illustrating our theoretical results.

III.3.4.1 Rate of convergence

In this experiment, we study the rate of convergence of the estimator $\hat{\tau}$. Let $d = 20$, $p = 10$, $\sigma = 1$, $\tau = 0.3$. Let us consider data generated from Model (III.3.1) with the means θ^- and θ^+ obtained from the following distribution: $\theta^- \sim \mathcal{N}(0, \frac{1}{20j^2}))$, $\theta^+ \sim \mathcal{N}(-\theta^-, 10^{-4})$.

To get a first insight about the rate of convergence, we simulate 1000 times a sample of length n, for n chosen between 20 and 4000, and plot in Figure III.3.2 the mean and median of the error $|\tau - \hat{\tau}|$ over the 1000 trials in function of n, together with the function $n \mapsto \ln(n)\Psi_n(p, \Delta_p)$ corresponding to the theoretical rate of convergence obtained in Proposition III.3.2.1. Note that the rate of convergence of $|\tau - \hat{\tau}|$ is given in the proposition up to a constant $\kappa(\gamma, \varepsilon)$. Nevertheless, the figure provides an appropriate illustration of the result as soon as n is large enough.

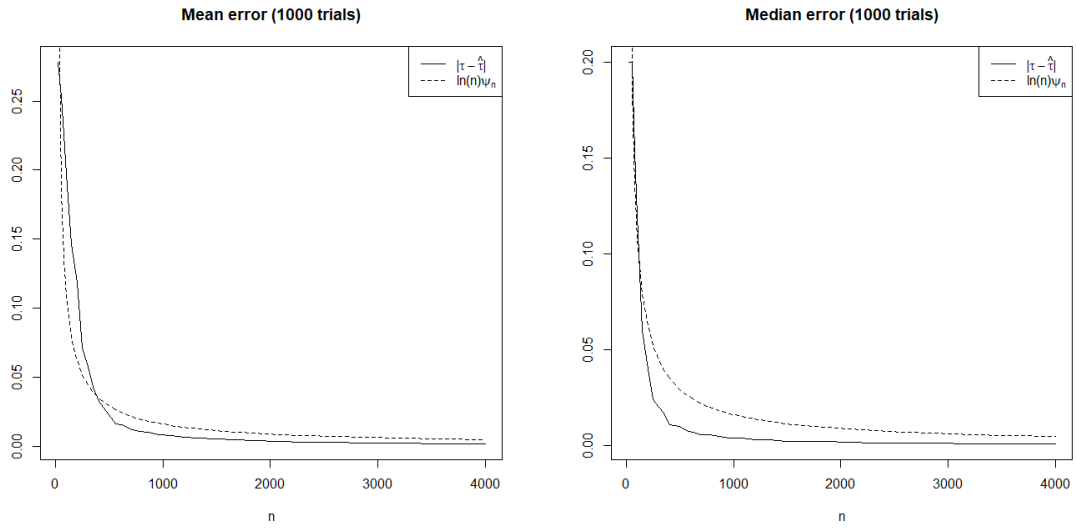


Figure III.3.2 – Plot of $|\hat{\tau} - \tau|$ as a function of n (mean and median over 1000 trials)

Then, simulating 1000 samples, for each value of the sample size n between 500 and 4000, we try to estimate of the rate of convergence by computing the linear regression of $|\tau - \hat{\tau}|$ by $\ln(n)$: omitting the logarithmic factor, an exponent -1 is to be found, corresponding to the rate of convergence $\frac{1}{n}$. Figure III.3.3 provides an illustration of this linear regression, considering again the mean and the median over the 1000 trials. On this example, the estimated slope of the regression line is -1.172

for the mean and -1.098 for the median.

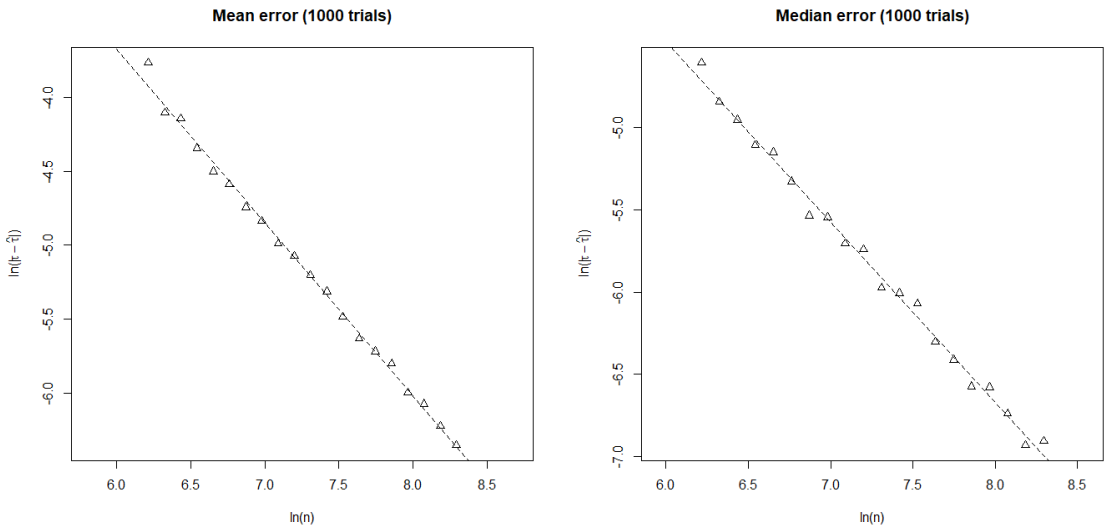


Figure III.3.3 – Plot of $\ln(|\hat{\tau} - \tau|)$ as a function of $\ln(n)$ (mean and median over 1000 trials)

III.3.4.2 Selection of p

In Theorem III.3.3.1, we suggest to select p using Lepski's method. Before introducing a practical procedure for the selection of p, let us illustrate the fact that the performance of the estimator $\hat{\tau}$ may indeed vary a lot as a function of p, so that selecting the right p is a crucial issue in the estimation of τ.

We set $d = 200$, $n = 100$, $\sigma = 1$, $\tau = 0.3$. We consider data generated from Model (III.3.1) with means θ^- and θ^+ built as follows:

— Case A: $\theta^- \sim \mathcal{N}(0, V)$, $\theta^+ \sim \mathcal{N}(0, V)$, $V = \mathrm{diag}(v_1, \ldots, v_d)$, $v_j = \frac{1}{2j^2}$ for $j = 1, \ldots, d$.

— Case B: θ^- is such that $\theta^-_j \sim \mathcal{N}(0, 1/2)$ for $j = 1, \ldots, 20$, $\theta^-_j \sim \mathcal{N}\big(0, \frac{1}{2(j-20)^2}\big)$ for $j = 21, \ldots, d$. θ^+ is such that $\theta^+_j \ \mathcal{N}(\theta^-_j, 10^{-2})$ for $j = 1, \ldots, 20$, $\theta^+_j \sim \mathcal{N}(0, \frac{1}{2(j-20)^2})$ for $j = 21, \ldots, d$.

We simulated 5000 data sets according to Model (III.3.1) in each of the two cases. Figures III.3.4 and III.3.5 show the mean and median error $|\hat{\tau} - \tau|$ over the 5000

trials as a function of p. In the first case, the best result is obtained already with $p = 1$, whereas for the second, taking p around 30 is a good choice.

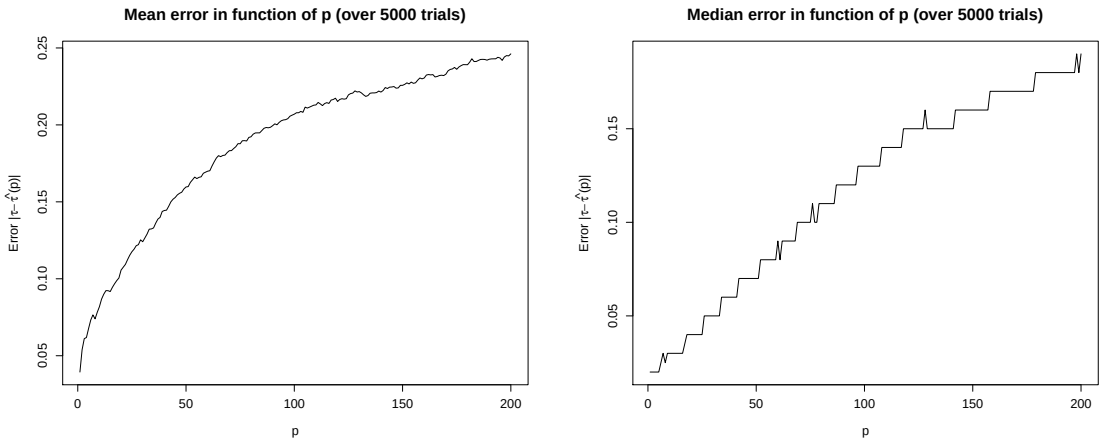


Figure III.3.4 – Mean and median of the error over 5000 trials for Model A

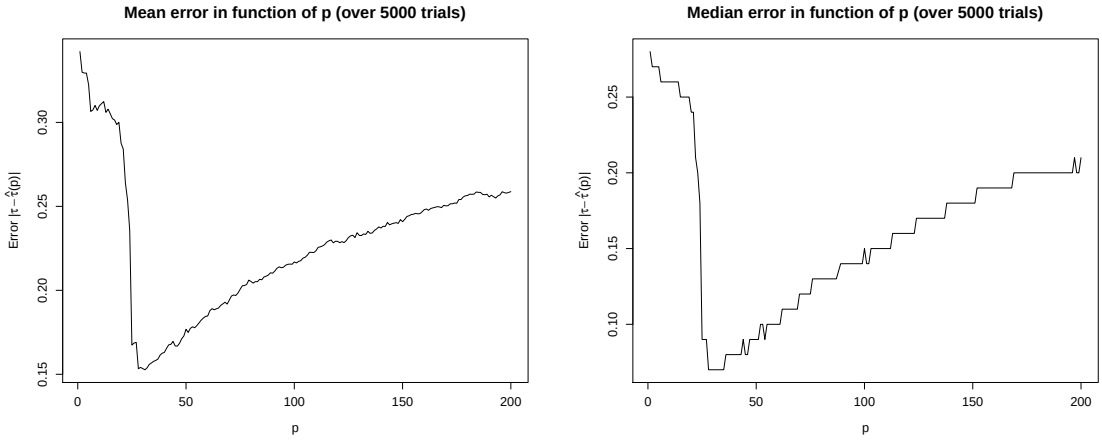


Figure III.3.5 – Mean and median of the error over 5000 trials for Model B

Theorem III.3.3.1 provides a theoretical way to select p. However, the statement depends on a tuning constant $C_{\mathcal{L}}$. In practice, it is simpler to try to select directly p. In the sequel, two procedures are investigated, yielding two estimators $\hat{p}_1$ and $\hat{p}_2$.
Method 1. This method is often used to search for tuning constants in adaptive methods. The idea is to find a division of the set $\{1, \dots, d\}$ into $\{1, \dots, \hat{p}_1\}$ and its

complementary, where the two subsets are corresponding to two "regimes" for the data, one with "big coefficients", one with small ones.
Let $\bar{Z}^{(p)} = \frac{1}{p}\sum_{j=1}^{p} Z_j$ and $\bar{Z}^{(-p)} = \frac{1}{d-p}\sum_{j=p+1}^{d} Z_j$, and consider

$$V(p) = \sum_{j=1}^{p}(Z_j - \bar{Z}^{(p)})^2 + \sum_{j=p+1}^{d}(Z_j - \bar{Z}^{(-p)})^2.$$

This quantity V is computed for every $p = 1, \dots, d$ and the value $\hat{p}_1$ is chosen such that

$$\hat{p}_1 \in \arg\min_{p=1,\dots,d} V(p).$$

Indeed, this procedure, by searching for a change-point along $Z_1, \dots, Z_d$, should separate the first most significative differences $\theta_j^- - \theta_j^+$, where $j = 1, \dots, \hat{p}_1$, from the remaining ones, expected to be less significative for estimating $\hat{\tau}$, in such a way that keeping for the estimation all components until $\hat{p}_1$ seems a reasonable choice.
Method 2. The second idea is more computationally involved and based on subsampling. When performing subsampling, the indices drawn at random are sorted, so that the parameter of interest τ remains indeed approximatively unchanged. For each $p = 1, \dots, d$, we compute $\hat{\tau}(p)$ for a collection of subsamples. Then, $\hat{p}_2$ is set to the value of p minimizing the variance of $\hat{\tau}$ over all subsamples. Here, 100 subsamples are built, each of them containing 80% of the initial sample.

Remark 18. Proportions of data from 50% to 90% have also been tried, with quite similar results. Observe that picking a quite small proportion of data for subsampling could be interesting since it provides more variability between the subsamples, but, at the same time, the fact that the ratio between the dimension d and the sample size is modified may be annoying when the aim is to select p. We also considered a version of subsampling where a different subsampling index is drawn for every $p = 1, \dots, d$: again, this provides more variability in the subsamples, but τ may also vary more than in the classical version. The results were not significantly different.

The performance of the two methods is compared with the result obtained using the value of p minimizing the average value of $|\tau - \hat{\tau}(p)|$ over a large number of trials, called hereafter oracle $p^\star$ (here, $p^\star = 30$ as obtained above for 5000 trials). Of course, $p^\star$ is not available in practice, since it depends on the true τ. However, it is introduced as a benchmark. The results, corresponding to 1000 trials, for Model B, are shown in Figure III.3.6 and Table III.3.1. The performances of the proposed methods could seem unsatisfactory in absolute terms. Nevertheless, the data has deliberately been chosen difficult to segment. Indeed, to illustrate the selection of

p, it seems more appropriate to consider a high-dimensional, hard situation, rather than an easy one where the true τ is always found exactly. Observe that the two methods perform very similarly, with a slight advantage of Method 2 over Method 1. However, Method 2 is based on subsampling, and, as such, is more CPU-time consuming.

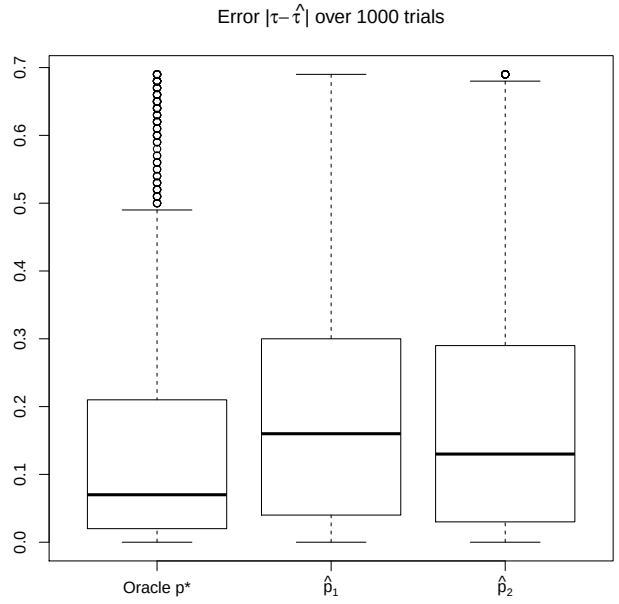


Figure III.3.6 – Error of the two selection procedures over 1000 trials, compared with the error obtained using the oracle $p^\star = 30$

Error over 1000 trials	Oracle $p^\star$	$\hat{p}_1$	$\hat{p}_2$
Mean	0.1524	0.2207	0.2047
(Standard deviation)	(0.18735)	(0.21329)	(0.20841)

Table III.3.1 – Mean and standard deviation over 1000 trials of the error obtained with the oracle $p^\star$ and the two selection methods

Chapter III.4

Wasserstein deconvolution in dimension 1

This chapter is the result of a collaboration with Jérôme Dedecker (MAP5, Université de Paris) and Bertrand Michel (LMJL, École Centrale de Nantes), published in the Electronic Journal of Statistics *(Dedecker et al., 2015).*

III.4.1 Introduction

Consider the following convolution model: we observe n real-valued random variables $Y_1, \dots, Y_n$ such that

$$Y_i = X_i + \varepsilon_i, \tag{III.4.1}$$

where the X_i's are independent and identically distributed according to an unknown probability μ, which we want to estimate. The random variables ε_i, $i = 1, \dots, n$, are independent and identically distributed according to a known probability measure μ_ε, not necessarily symmetric. Moreover we assume that $(X_1, \dots, X_n)$ is independent of $(\varepsilon_1, \dots, \varepsilon_n)$.

Our purpose is to investigate rates of convergence for the estimation of the measure μ under Wasserstein metrics. For $p \in [1, \infty)$, the Wasserstein distance W_p between μ and ν is given by

$$W_p(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \left(\int_{\mathbb{R}^2} |x - y|^p \pi(dx, dy) \right)^{\frac{1}{p}},$$

where $\Pi(\mu, \nu)$ is the set of probability measures on $\mathbb{R} \times \mathbb{R}$ with marginal distributions μ and ν (see Rachev and Rüschendorf, 1998; Villani, 2008,). The distances W_p are natural metrics for comparing measures. For instance they can compare two singular measures, which is of course impossible with the functional metrics commonly used in density estimation. Convergence of measure under Wasserstein distances is an active domain of research in probability and statistics. For instance, the rate of convergence of the empirical measure under these metrics has been obtained recently by both Dereich et al. (2013) and Fournier and Guillin (2015) in $\mathbb{R}^d$ and also by Bobkov and Ledoux (2019) in the one-dimensional framework. Moreover, Wasserstein metrics are involved in many fields of mathematics and computer sciences. For instance, in the field of Topological Data Analysis (TDA) (Carlsson, 2009), Wasserstein distances appear to be natural metrics for controlling the estimation of geometric and topological features of the sampling measure and its support. Indeed, in Chazal et al. (2011), a distance function to measures is introduced to solve geometric inference problems in a probabilistic setting: if a known measure ν is close enough with respect to W_2 to a measure μ concentrated on a given shape, then the topological properties of the shape can be recovered by using the distance to ν. More generally, the Wasserstein loss could be used as a guide for inferring the support. Other work in TDA with stability results involving the Wasserstein distances can be found in Chazal et al. (2014); Guibas et al. (2013). In practice, the data can be observed with noise, which motivates the study of the Wasserstein deconvolution problem (Caillerie et al., 2011), in particular if the deconvolved measure and the "true measure" are singular.

Rates of convergence in deconvolution have mostly been considered in density estimation, for pointwise or global convergence. Minimax rates can be found for instance in Butucea and Tsybakov (2008a,b); Fan (1991a) and in the monograph of Meister (2009). Here, however, we shall not assume that μ has a density with respect to the Lebesgue measure. In this context, rates of convergence for the W_2 Wasserstein distance have first been studied for several noise distributions by Caillerie et al. (2011). Then, Dedecker and Michel (2013) have obtained optimal rates of convergence in the minimax sense for a class of supersmooth error distributions, in any dimension, under any Wasserstein metric W_p. The result relies on the fact that lower bounds in any dimension can be deduced in this case from the lower bounds in dimension 1. Such a method cannot be used in the ordinary smooth case, where the rate of convergence depends on the dimension. As noticed by Fan (1991a), establishing optimal rates of convergence in the ordinary smooth case is more difficult than in the supersmooth one, even for pointwise estimation.

A key fact in the univariate context is that Wasserstein metrics are linked to integrated risks between cumulative distribution functions (cdf). In dimension 1, when estimating the density of μ, optimal rates of convergence for integrated risks can be found in Fan (1991b, 1993). When estimating the cdf F of μ, optimal rates for the pointwise and integrated quadratic risks are given in Hall and Lahiri (2008), where it is shown in particular that the rate $\sqrt{n}$ can be reached when the error distribution is ordinary smooth with a smoothness index less than $1/2$. Concerning the pointwise estimation of $F(x_0)$, optimal rates for the quadratic risk are also given in Dattner et al. (2011), when the density of μ belongs to a Sobolev class.

The case $\beta = 0$ in the upper bound (3.9) of Hall and Lahiri (2008) corresponds to the case where no assumption (except a moment assumption) is made on the measure μ (in particular μ is not assumed to be absolutely continuous with respect to the Lebesgue measure). This is precisely the case which we want to consider here. However the results by Hall and Lahiri (2008) cannot be applied to the Wasserstein deconvolution problems for two reasons: firstly, the integrated quadratic risk for estimating a cdf is not linked to Wasserstein distances, and secondly, the estimator of the cdf of μ proposed in Hall and Lahiri (2008) is the cdf of a signed measure, and is not well defined as an estimator of μ for the Wasserstein metric.

In this chapter, we propose in the univariate situation, when the error is ordinary smooth, an improved upper bound for deconvolving μ under W_p, as well as a lower bound. We also recover the optimal rate of convergence in the supersmooth case with slightly weaker regularity conditions than in Dedecker and Michel (2013). The estimator of the cdf F of μ is built in two steps: firstly, as in Hall and Lahiri (2008), we define a preliminary estimator through a classical kernel deconvolution method, and secondly we take an appropriate isotone approximation of this estimator. For controlling the random term, we use a moment inequality on the cdfs, which is due to Èbralidze (1971).

In Section III.4.2, some facts about the case without error are recalled and discussed. The upper bounds for Wasserstein deconvolution with supersmooth or ordinary smooth errors are given in Section III.4.3, and Section III.4.4 is about lower bounds. Section III.4.5 presents the implementation of the method and some experimental results. In particular, observed rates of convergence are compared with the theoretical bounds for the Wasserstein metrics W_1 and W_2, and we study as an illustrative example the deconvolution of the uniform measure on the Cantor set.

III.4.2 On the case without error

We begin by considering the simple case when one observes directly $X_1, \dots, X_n$ with values in $\mathbb{R}$ without error. Let us recall some results for the quantities $W_p(\mu_n, \mu)$, where μ_n is the empirical measure, given by

$$\mu_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}.$$

Let F be the cdf of X_1, F_n the cdf of μ_n, and let F^{-1} and F_n^{-1} be their usual cadlag inverses. Recall that, for any $p \geq 1$,

$$W_p^p(\mu_n, \mu) = \int_0^1 |F_n^{-1}(u) - F^{-1}(u)|^p du\,, \tag{III.4.2}$$

and, if $p = 1$

$$W_1(\mu_n, \mu) = \int |F_n^{-1}(u) - F^{-1}(u)| du = \int |F_n(t) - F(t)| dt\,.$$

The case $p = 1$ has been well understood since the article by del Barrio et al. (1999). The random variable $\sqrt{n} W_1(\mu_n, \mu)$ converges in distribution to $\int |B(F(t))| dt$, where B is a standard Brownian bridge, if and only if

$$\int_0^\infty \sqrt{P(|X| > x)} dx < \infty, \tag{III.4.3}$$

or equivalently if

$$\int_0^\infty \sqrt{F(x)(1 - F(x))} dx < \infty.$$

More recently, Bobkov and Ledoux (2019) have shown that the rate of $E[W_1(\mu_n, \mu)]$ can be characterized by the quantities

$$\int_{4nF(x)(1-F(x)) \leq 1} F(x)(1 - F(x)) dx$$

and

$$\int_{4nF(x)(1-F(x)) > 1} \sqrt{F(x)(1 - F(x))}\, dx.$$

More precisely, the rate $1/\sqrt{n}$ is achieved if and only if (III.4.3) is satisfied. When this is not the case, $E[W_1(\mu_n, \mu)]$ may decay at an arbitrary slow rate.

For $p > 1$, the situation is more complicated. Extra conditions are necessary to ensure that $W_p(\mu_n, \mu)$ is of order $1/\sqrt{n}$. If the random variables take their values in a compact interval $[a, b]$ and if the cdf F is continuously differentiable on $[a, b]$ with strictly positive derivative f, then $n^{p/2} W_p^p(\mu_n, \mu)$ converges in distribution to $\int_0^1 |B(u)|^p / |f \circ F^{-1}(u)|^p du$ (see Lemma 3.9.23 in van der Vaart and Wellner, 1996). But in general, the rate can be much slower. The convergence in distribution for the case $p = 2$ has been studied in detail by del Barrio et al. (2005). Under additional conditions on F, which must be twice differentiable, the rate of convergence depends on the behavior of F^{-1} in a neighborhood of 0 and 1. For instance, if

$$F(t) = \left(1 - \frac{1}{t^{\alpha-1}}\right) \mathbf{1}_{\{t>1\}},$$

where $\alpha > 3$, then

$$n^{(\alpha-3)/(\alpha-1)} W_2^2(\mu_n, \mu) \tag{III.4.4}$$

converges in distribution.

The rates of decay of $E[W_p(\mu_n, \mu)]$ and $\left(E[W_p^p(\mu_n, \mu)]\right)^{1/p}$ have been studied in Bobkov and Ledoux (2019). These quantities decay at the standard rate $1/\sqrt{n}$ if and only if

$$J_p(\mu) = \int_{\mathbb{R}} \frac{[F(x)(1 - F(x))]^{p/2}\, dx}{f(x)^{p-1}} < \infty\,,$$

where f is the density of the absolutely continuous component of μ. In particular,

$$\left(E[W_p^p(\mu_n, \mu)]\right)^{1/p} \leq \frac{5p}{\sqrt{n+2}} J_p^{1/p}(\mu).$$

However, this approach cannot be applied when the measure μ and the Lebesgue measure are singular. An alternative approach to obtain the rate of decay of $E[W_p^p(\mu_n, \mu)]$ is to use the following inequality, due to Èbralidze (1971) : for any $p \geq 1$,

$$W_p^p(\mu, \nu) \leq \kappa_p \int |x|^{p-1} |F_\mu - F_\nu|(x) dx\,, \tag{III.4.5}$$

where $\kappa_p = 2^{p-1} p$. Starting from (III.4.5), we get that

$$\begin{aligned} E[W_p^p(\mu_n, \mu)] &\leq \int |x|^{p-1} E|F_n(x) - F(x)| dx \\ &\leq \int |x|^{p-1} \sqrt{E|F_n(x) - F(x)|^2} dx \\ &\leq \frac{1}{\sqrt{n}} \int |x|^{p-1} \sqrt{F(x)(1 - F(x))} dx \end{aligned}$$

where F_n is the empirical cdf. This last integral is finite if and only if

$$\int_0^\infty |x|^{p-1}\sqrt{P(|X|>x)}dx < \infty\,. \tag{III.4.6}$$

For instance, taking $p=2$, a tail satisfying $P(|X|>x) = \mathcal{O}\Big(\frac{1}{x^4\log x^{2+\varepsilon}}\Big)$ gives the rate $1/\sqrt{n}$. Hence, we obtain the same rate as in (III.4.4) for $\alpha=5$, with a slightly stronger tail condition (due to the fact that we control the expectation), but without additional assumptions on the cdf F.

Since we want to estimate singular measures, we shall follow this approach in the sequel.

III.4.3 Upper bounds for W_p in deconvolution

III.4.3.1 Construction of the estimator

Let us start with some notations. For μ a probability measure and ν another probability measure, with density g, we denote by $\mu\star g$ the density of $\mu\star\nu$, given by

$$\mu\star g(x) = \int_{\mathbb{R}} g(x-y)\mu(dy).$$

We further denote by μ^* (respectively f^*) the Fourier transform of the probability measure μ (respectively of the integrable function f), that is

$$\mu^*(x) = \int_{\mathbb{R}} e^{iux}\mu(du) \quad \text{and} \quad f^*(x) = \int_{\mathbb{R}} e^{iux}f(u)du.$$

Finally, let F be the cdf of μ.

The estimator $\tilde{\mu}_n$ of the measure μ is built in two steps.

1. First, we build a preliminary estimator of F. Let $\lceil p\rceil$ be the least integer greater than or equal to p. We first introduce a symmetric nonnegative kernel k such that its Fourier transform k^* is $\lceil p\rceil$ times differentiable with Lipschitz $\lceil p\rceil$-th derivative and is supported on $[-1,1]$. An example of such a kernel is given by

$$k(x) = C_p\left[\frac{(2\lceil p/2\rceil+2)\sin\frac{x}{2\lceil p/2\rceil+2}}{x}\right]^{2\lceil p/2\rceil+2}, \tag{III.4.7}$$

where C_p is such that $\int k(x)dx = 1$.

We define now a preliminary estimator $\hat{F}_n$ of F:

$$\hat{F}_n(t) = \frac{1}{nh}\int_{-\infty}^{t}\sum_{k=1}^{n}\tilde{k}_h\left(\frac{u-Y_k}{h}\right)du,$$

where

$$\tilde{k}_h(x) = \frac{1}{2\pi}\int \frac{e^{iux}k^*(u)}{\mu_\varepsilon^*(-u/h)}du.$$

Let us first give some conditions under which these quantities are well defined. Clearly, $\tilde{k}_h(x)$ is well defined as soon as μ_ε^* does not vanish, since in that case it is the Fourier transform of a continuous and compactly supported function (it can be easily checked that $\tilde{k}_h(x)$ is a real function). In the sequel, we shall always assume that $r_\varepsilon = 1/\mu_\varepsilon^*$ is at least two times continuously differentiable. In that case, the function $w(u) = \frac{k^*(u)}{\mu_\varepsilon^*(-u/h)}$ is two times differentiable with bounded and compactly supported derivatives. An integration by parts yields

$$\tilde{k}_h(x) = -\frac{1}{2\pi ix}\int e^{iux}w'(u)du \quad \text{and} \quad \tilde{k}_h(x) = -\frac{1}{2\pi x^2}\int e^{iux}w''(u)du.$$

It follows that $\tilde{k}_h$ is a continuous function such that $\tilde{k}_h(x) = \mathcal{O}(1/(1+x^2))$. Hence $\tilde{k}_h$ belongs to $L^1(dx)$ and $\hat{F}_n$ is well defined. Now the inverse Fourier formula gives that $\tilde{k}_h^*(x) = \frac{k^*(u)}{\mu_\varepsilon^*(u/h)}$. Consequently $\tilde{k}_h^*(0) = 1$, proving that $\int \tilde{k}_h(x)dx = 1$ and that $\lim_{t\to\infty}\hat{F}_n(t) = 1$.

However, this estimator $\hat{F}_n$, based on the standard deconvolution kernel density estimator $\tilde{k}_h$ first introduced by Carroll and Hall (1988), is not a cumulative distribution function since it is not necessarily non-decreasing.

2. We need to define an estimator $\tilde{F}_n$ of F which is a cumulative distribution function. We choose the estimator $\tilde{F}_n$ as an approximate minimizer over all distribution functions of the quantity $\int_{\mathbb{R}}|x|^{p-1}|\hat{F}_n - G|(x)dx$. More precisely, given $\rho > 0$, let $\tilde{F}_n$ be such that, for every distribution function G,

$$\int |x|^{p-1}|\hat{F}_n - \tilde{F}_n|(x)dx \le \int |x|^{p-1}|\hat{F}_n - G|(x)dx + \rho\,.$$

Here, ρ may be chosen equal to 0 (best isotone approximation) but the condition $\rho = O(n^{-1/2})$ is the only condition required to get the rates of Section III.4.3.3 below.

Finally, the estimator $\tilde{\mu}_n$ is then defined as follows:

$$\tilde{\mu}_n \text{ is the probability measure with distribution function } \tilde{F}_n. \qquad \text{(III.4.8)}$$

Remark 19. In Dedecker and Michel (2013), $\tilde{\mu}_n$ is chosen as the (normalized) positive part of μ_n. We will see that the isotone approximation described here allows to get better rates of convergence in the ordinary smooth case. However, this approach works only in the one-dimensional case.

III.4.3.2 First upper bounds for $W_p^p(\tilde{\mu}_n, \mu)$

The control of $W_p^p(\tilde{\mu}_n, \mu)$ is done in three steps:

1. **A bias/random decomposition.** Denoting by K_h the function $h^{-1}k(\cdot/h)$, we have

$$W_p^p(\tilde{\mu}_n, \mu) \leq 2^{p-1} W_p^p(\mu \star K_h, \mu) + 2^{p-1} W_p^p(\tilde{\mu}_n, \mu \star K_h).$$

   The non-random quantity $W_p^p(\mu \star K_h, \mu)$ is the bias of the estimator $\tilde{\mu}_n$.

2. **Control of the bias.** Let V_h be a random variable with distribution K_h and independent of X_1, in such a way that the distribution of $X_1 + V_h$ is $\mu \star K_h$. By definition of W_p, we have

$$W_p^p(\mu \star K_h, \mu) \leq E[|X_1 + V_h - X_1|^p] = E[|V_h|^p] = h^p \int |x|^p k(x) dx.$$

3. **Control of the random term.** Note that

$$E[\hat{F}_n(t)] = \int_{-\infty}^{t} \mu \star K_h(x) dx$$

   is the cdf of $\mu \star K_h$. Applying Èbralidze's inequality (III.4.5), we obtain that

$$W_p^p(\tilde{\mu}_n, \mu \star K_h) \leq \kappa_p \int |x|^{p-1} |\tilde{F}_n - E[\hat{F}_n]|(x) dx \,.$$

   Now, by the triangle inequality and the definition of $\tilde{F}_n$,

$$\begin{aligned} W_p^p(\tilde{\mu}_n, \mu \star K_h) &\leq \kappa_p \left(\int |x|^{p-1} |\tilde{F}_n - \hat{F}_n|(x) dx + \int |x|^{p-1} |\hat{F}_n - E[\hat{F}_n]|(x) dx \right) \\ &\leq \rho + 2\kappa_p \int |x|^{p-1} |\hat{F}_n - E[\hat{F}_n]|(x) dx \,. \end{aligned}$$

Consequently, to get explicit rates of convergence for $E[W_p^p(\tilde{\mu}_n, \mu)]$, it remains to control the term

$$E\left(\int |x|^{p-1} |\hat{F}_n - E[\hat{F}_n]|(x) dx \right) .$$

III.4.3.3 Main results

Let $r_\varepsilon = 1/\mu_\varepsilon^*$, and let $r_\varepsilon^{(\ell)}$ be the ℓ-th derivative of r_ε. Let m_0 denote the least integer strictly greater than $p + \frac{1}{2}$, and m_1 be the least integer strictly greater than $p - \frac{1}{2}$.

First, we establish an upper bound for $E[W_p^p(\tilde{\mu}_n, \mu)]$ involving a tail condition on Y and the regularity of r_ε. The next proposition is proved in Dedecker et al. (2015).

Proposition III.4.3.1. *Let $\rho \leq n^{-1/2}$, and let $\tilde{\mu}_n$ be the estimator defined in (III.4.8). Assume that r_ε is m_0 times continuously differentiable. For any $h \leq 1$, we have*

$$E[W_p^p(\tilde{\mu}_n, \mu)] \leq \frac{1}{\sqrt{n}} + h^p 2^{p-1} \int |x|^p k(x)dx + \frac{C}{\sqrt{n}}(A_1 + A_2 + A_3 + A_4)$$

where

$$A_1 = \Big(\sup_{t\in[-2,2]} \sum_{\ell=0}^{1} |r_\varepsilon^{(\ell)}(t)| \Big) \int_0^\infty |x|^{p-1} \sqrt{P\left(|Y| \geq x\right)} dx$$

$$A_2 = \sup_{t\in[-2,2]} \sum_{\ell=0}^{m_0} |r_\varepsilon^{(\ell)}(t)|$$

$$A_3 = \left[E|Y|^{2p-\frac{1}{2}} \int_{-1/h}^{1/h} \frac{|r_\varepsilon(x)|^2}{|x|^2} \mathbf{1}_{[-1,1]^c}(x)dx \right]^{1/2}$$

$$A_4 = \left[\sum_{\ell=0}^{m_1} \int_{-1/h}^{1/h} \frac{|r_\varepsilon^{(\ell)}(x)|^2}{|x|^2} \mathbf{1}_{[-1,1]^c}(x)dx \right]^{1/2}.$$

Now, the rates of convergence obtained for the Wasserstein deconvolution, for a class of supersmooth error distributions and for a class of ordinary smooth error distributions, are stated in the next theorem.

Theorem III.4.3.1. *Let $\rho \leq n^{-1/2}$, and let $\tilde{\mu}_n$ be the estimator defined in (III.4.8). Assume that*

$$\int_0^\infty |x|^{p-1} \sqrt{P(|Y| \geq x)} dx < \infty \text{ and } \sup_{t\in[-2,2]} |r_\varepsilon^{(m_0)}(t)| < \infty.$$

1. *Assume that there exist* $\beta > 0$, $\tilde{\beta} \geq 0$, $\gamma > 0$ *and* $c > 0$, *such that for every* $\ell \in \{0, 1, \ldots, m_1\}$ *and every* $t \in \mathbb{R}$,

$$|r_\varepsilon^{(\ell)}(t)| \leq c(1 + |t|)^{\tilde{\beta}} \exp(|t|^\beta/\gamma). \tag{III.4.9}$$

Then, taking $h = (4/(\gamma \log n))^{1/\beta}$, *there exists* $C > 0$ *such that*

$$E[W_p^p(\tilde{\mu}_n, \mu)] \leq C(\log n)^{-p/\beta}.$$

2. *Assume that there exist* $\beta > 0$ *and* $c > 0$, *such that for every* $\ell \in \{0, 1, \ldots, m_1\}$ *and every* $t \in \mathbb{R}$,

$$|r_\varepsilon^{(\ell)}(t)| \leq c(1 + |t|)^{\beta}.$$

Then, taking $h = n^{-\frac{1}{2p+(2\beta-1)_+}}$, *there exists* $C > 0$ *such that*

$$E[W_p^p(\tilde{\mu}_n, \mu)] \leq C\psi_n\,,$$

where

$$\psi_n = \begin{cases} n^{-\frac{p}{2p+2\beta-1}} & \text{if } \beta > \frac{1}{2} \\ \sqrt{\frac{\log n}{n}} & \text{if } \beta = \frac{1}{2} \\ \frac{1}{\sqrt{n}} & \text{if } \beta < \frac{1}{2}. \end{cases}$$

Some comments are in order.

Remark 20. In the ordinary smooth case, when $\beta < 1/2$, any bandwidth $h = \mathcal{O}(n^{-1/2p})$ leads to the rate $n^{-1/2}$. The fact that there are three different situations according as $\beta > 1/2, \beta = 1/2$ or $\beta < 1/2$ has already been pointed out in Dattner et al. (2011); Hall and Lahiri (2008) for the estimation of the cdf F.

Remark 21. Since the function $H_Y(x) = P(|Y| \geq x)$ is non-increasing, the tail condition

$$\int_0^\infty |x|^{p-1}\sqrt{P(|Y| \geq x)}dx < \infty \tag{III.4.10}$$

implies that $H_Y(x) = \mathcal{O}(1/|x|^{2p})$. Hence $|Y|$ has a weak moment of order $2p$, which implies a strong moment of ordrer q for any $q < 2p$. Note that (III.4.10) is the same as the tail condition (III.4.6) obtained in Section III.4.2 to get the rate $E[W_p^p(\mu_n, \mu)] \leq Cn^{-1/2}$ in the case without noise. Recall that, in the case without noise when $p = 1$, this condition is necessary and sufficient for the weak convergence of $\sqrt{n}W_1(\mu_n, \mu)$. We have

$$\text{(III.4.10) holds iff (III.4.6) holds and } \int_0^\infty |x|^{p-1}\sqrt{P(|\varepsilon| \geq x)}dx < \infty.$$

Remark 22. The rate $E[W_p^p(\tilde{\mu}_n, \mu)] \leq C(\log n)^{-p/\beta}$ in the supersmooth case has already been given in Dedecker and Michel (2013) and is valid in any dimension. However, in Dedecker and Michel (2013) , the condition (III.4.9) on the regularity of r_ε is assumed to be true for $\ell \in \{0, 1, \ldots, \lceil p \rceil + 1\}$. Note that this rate is minimax.

III.4.4 Lower bound

For some $M > 0$ and $q \geq 1$, we denote by $\mathcal{D}(M, q)$ the set of measures μ on $\mathbb{R}$ such that $\int |x|^q d\mu(x) \leq M$.

Theorem III.4.4.1. *Let $M > 0$ and $q \geq 1$. Assume that there exist $\beta > 0$ and $c > 0$, such that for every $\ell \in \{0, 1, 2\}$ and every $t \in \mathbb{R}$,*

$$|\mu_\varepsilon^{*(\ell)}(t)| \leq c(1 + |t|)^{-\beta}.$$

Then, there exists a constant $C > 0$ such that, for any estimator $\hat{\mu}$,

$$\liminf_{n\to\infty} n^{\frac{p}{2\beta+1}} \sup_{\mu\in\mathcal{D}(M,q)} E[W_p^p(\hat{\mu}, \mu)] > C.$$

Remark 23. For W_1, this lower bound matches the upper bound given in Theorem III.4.3.1 for $\beta \geq 1/2$. For W_p, $p > 1$, we conjecture that the upper bounds given by Theorem III.4.3.1 are appropriate under the assumed tail conditions. Getting better rates of convergence for W_p ($p > 1$) is an open question. From Section III.4.2, it seems reasonable to think that better rates can be obtained when μ has an absolutely continuous component with respect to the Lebesgue measure which is strictly positive on the support of μ.

III.4.5 Numerical experiments

In this section, we study the W_1 and W_2 univariate deconvolution problems in the ordinary smooth case and we compare our numerical results with the upper and lower bounds given above. We also apply our procedure to the deconvolution of the uniform measure on the Cantor set.

The complete details about the implementation of the deconvolution estimators may be found in Dedecker et al. (2015).

In the experiments, we consider symmetric distributions for μ_ε. Thus, k^* and μ_ε^* are even functions, and the Fourier coefficients involved in the computation of the estimators are real. They are computed thanks to the Fast Fourier Transform.

We use the kernel

$$k(x) = \frac{3}{16\pi}\left(\frac{8\sin(x/8)}{x}\right)^4,$$

which corresponds to the kernel given by (III.4.7) with $p = 2$ and a Fourier support over $[-1/2, 1/2]$.

The isotone approximation is performed thanks to the function `gpava` from the R package `isotonic` (Mair et al., 2009).

In the sequel, the obtained estimator will be denoted by $\hat{\mu}_{n,h}^{\mathrm{isot},p}$, and $\hat{\mu}_n^{\mathrm{naive}}$ will stand for the "positive part" estimator studied in Dedecker and Michel (2013).

For fixed distributions μ and μ_ε, we simulate $Y_1, \ldots, Y_n$ according to the convolution model (III.4.1). For a given bandwidth h and $p \geq 1$, we can compute $W_p^p(\hat{\mu}_n^{\mathrm{naive}}, \mu)$ and $W_p^p(\hat{\mu}_{n,h}^{\mathrm{isot},p}, \mu)$ using the quantile functions of the measures, thanks to the relation (III.4.2). The Wasserstein risks $\mathcal{R}^{\mathrm{naive}}(n,h) := E[W_p^p(\hat{\mu}_{n,h}^{\mathrm{naive}}, \mu)]$ and $\mathcal{R}^{\mathrm{isot}}(n,h) := E[W_p^p(\hat{\mu}_{n,h}^{\mathrm{isot},p}, \mu)]$ can be estimated via a Monte-Carlo method by repeating the simulation of the Y_i's and averaging the Wasserstein distances. Let $\bar{r}_p^{\mathrm{isot}}(n,h)$ and $\bar{r}_p^{\mathrm{naive}}(n,h)$ be the estimated risks obtained this way, and $\bar{r}_{p,*}^{\mathrm{isot}}(n)$ and $\bar{r}_{p,*}^{\mathrm{naive}}(n)$ be the minimal risks over a bandwidth grid.

III.4.5.1 Estimation of the rates of convergence

We study the rates of convergence of the estimators for the deconvolution of three distributions:

— Dirac distribution at 0,

— Uniform distribution on $[-0.5, 0.5]$,

— Mixture of the Dirac distribution at 0 and the uniform distribution on $[-0.5, 0]$.

We take for μ_ε the ordinary smooth distributions summarized in Table III.4.1. Recall that the coefficient β of a symmetrized Gamma distribution is twice the shape parameter of the distribution.

Distribution	μ_ε^*	β
Symmetrized Gamma	$t \mapsto (1+t^2)^{-\beta/2}$	0.3, 0.5, 1.2, 2 ,3, 4
Laplace	$t \mapsto (1+t^2)^{-1}$	2
Symmetrized χ^2	$t \mapsto (1+4t^2)^{(-1/2)}$	1

Table III.4.1 – Ordinary smooth distributions used for the error

For each error distribution and for n chosen between 100 and 2000, we simulate 200 times a sample of length n from which we compute the estimated minimal risks $\bar{r}_{p,*}^{\text{isot}}(n)$ and $\bar{r}_{p,*}^{\text{naive}}(n)$. We study the Wasserstein risks W_1 and W_2. We obtain some estimation of the exponent of the rate of convergence for each deconvolution problem by computing the linear regression of $\log \bar{r}_{p,*}(n)$ by $\log n$. An example is depicted in Figure III.4.1. Observe that the risks are smaller for the isotone estimators than for the naive ones.

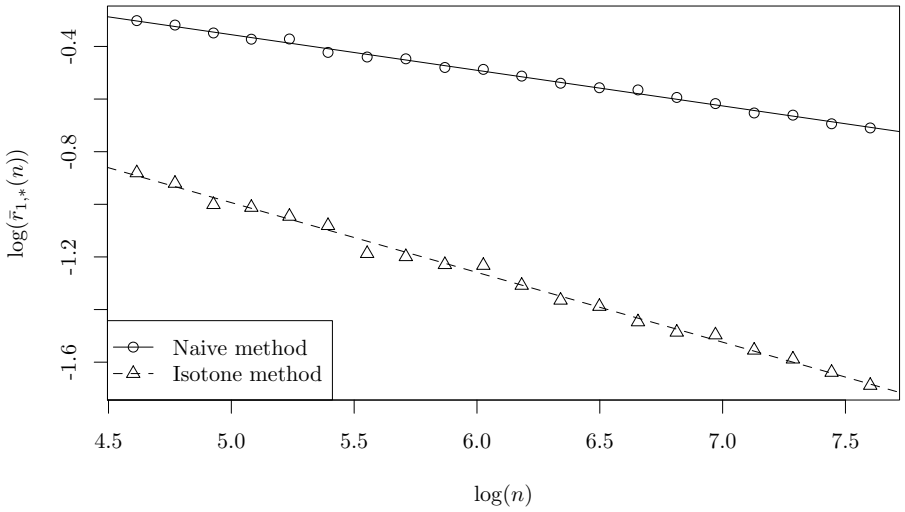


Figure III.4.1 – Estimated rates of convergence to zero of the W_1-risk for the naive method and the isotone method for μ being a Dirac distribution at 0, and the noise distribution the symmetrized Gamma distribution with $\beta = 2$

The estimated exponents of the convergence rates for W_1 and W_2 are plotted in Figure III.4.2 as functions of the ordinary smooth coefficient β, for the naive and the isotone deconvolution estimator. These estimated rates can be compared with the upper and lower bounds obtained stated above. Of course, the rates of convergence of the isotone estimator have no reason to match exactly the lower bounds. However it can be checked that the estimated rates are consistent with the theoretic bounds obtained. In particular, the parametric rate is reached for values of β close to 0, at least in the Dirac case.

III.4.5.2 Cantor set experiment

We now illustrate the deconvolution method by taking for μ the uniform distribution on the Cantor set $\mathfrak{C}$. Remember that the Cantor set can be defined by repeatedly deleting the open middle thirds of a set of line segments:

$$\mathfrak{C} = \bigcap_{m \geq 1} F_n$$

where $F_0 = [0, 1]$ and F_{m+1} is obtained by cutting out the middle thirds of all the intervals of F_m: $F_1 = [0, \frac{1}{3}] \cup [\frac{2}{3}, 1]$ and $F_2 = [0, \frac{1}{9}] \cup [\frac{2}{9}, \frac{1}{3}] \cup [\frac{2}{3}, \frac{7}{9}] \cup [\frac{8}{9}, 1]$, etc... The uniform measure $\mu_{\mathfrak{C}}$ on $\mathfrak{C}$ can be defined as the distribution of the random variable $X := 2\sum_{k\geq 1} 3^{-k} B_k$ where $(B_k)_{k\geq 1}$ is a sequence of independent random variables with Bernoulli distribution of parameter $1/2$. Note that the Lebesgue measure of $\mathfrak{C}$ is zero and thus the Lebesgue measure and $\mu_{\mathfrak{C}}$ are singular. The deconvolution estimators being densities for the Lebesgue measure, the Wasserstein distances are relevant metrics for comparing these with $\mu_{\mathfrak{C}}$.

Let $\mu_{\mathfrak{C},K}$ be the distribution of the random variable defined by the partial sum $\tilde{X} := 2\sum_{k=1}^{K} 3^{-k} B_k$ where the B_k's are defined as before. The distribution $\mu_{\mathfrak{C},K}$ is an approximation of $\mu_{\mathfrak{C}}$ which can be computed in practice. We simulate a sample of $n = 10^4$ observations from $\mu_{\mathfrak{C},K}$ with $K = 100$. These observations are contaminated by random variables with symmetrized Gamma distribution (the shape parameter is equal to $1/4$ (so that $\beta = 0.5$) and the scale parameter is equal to $1/2$).

In Figure III.4.3, the isotone estimators and naive estimator for W_1 and W_2 are plotted on the first four levels F_m of the Cantor set. We the W_1-isotone deconvolution estimator is able to detect the first three levels of the Cantor set and the three other deconvolution methods recover the first two levels. A kernel density estimator (with no deconvolution) only recovers the first level.

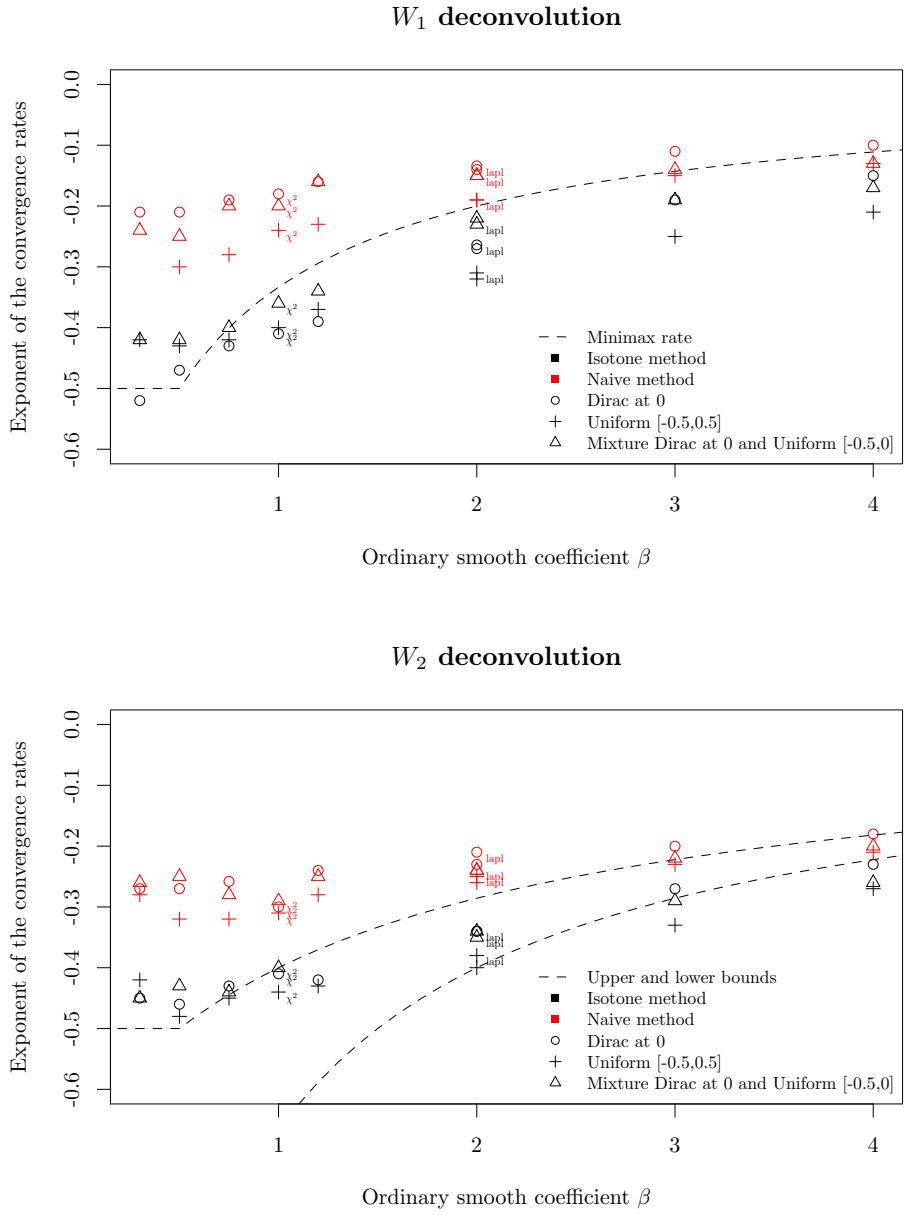


Figure III.4.2 – Estimated exponents of the convergence rates for W_1 and W_2 deconvolution, with the χ^2 and Laplace noise distributions indicated directly on the graph, the others experiments corresponding to the symmetrized Gamma distribution

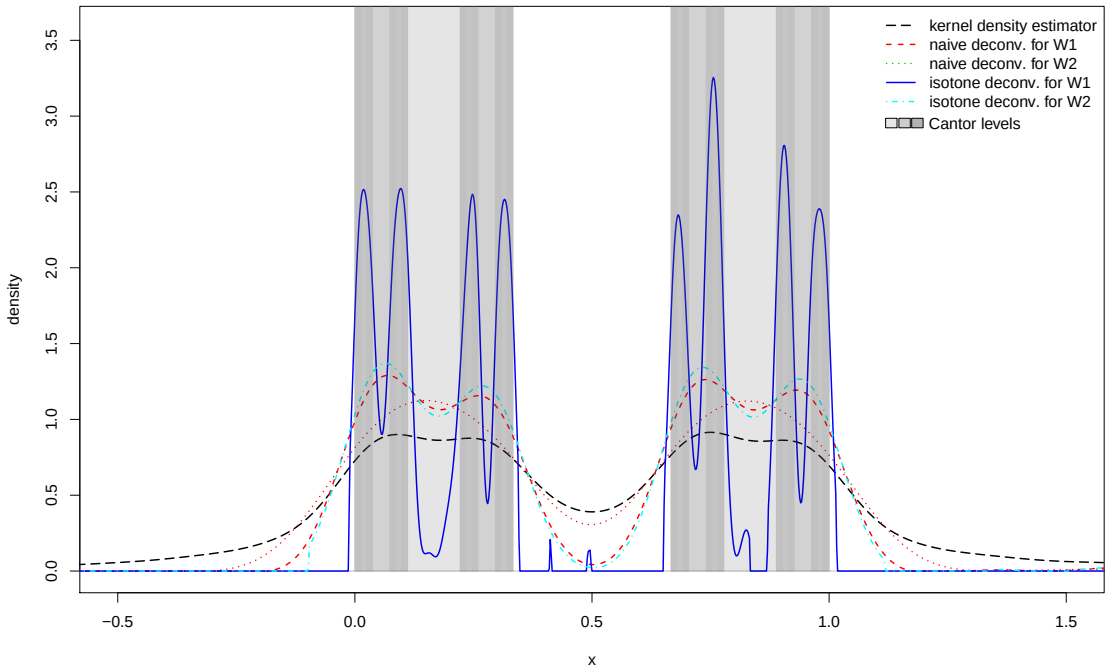


Figure III.4.3 – Deconvolution of the uniform measure on the Cantor set

Part IV

Aggregation

Chapter IV.1

Consensual aggregation for regression

This chapter corresponds to an article in collaboration with Gérard Biau, Benjamin Guedj (INRIA & University College London) and James Malley (National Institutes of Health, USA), published in the Journal of Multivariate Analysis *(Biau et al., 2016).*

Recent years have witnessed a growing interest in the aggregation of statistical procedures. Indeed, the increasing number of available estimation methods in a wide range of statistical problems naturally suggests to try to combine several procedures, all the more so since model and smoothness assumptions under which a particular method works well are usually unverifiable in practice. If the combined strategy is known to be optimal in some sense and relatively free of assumptions that are hard to evaluate, then, this strategy is a valuable and practical research tool.

In this regard, numerous contributions have enriched the aggregation literature with various approaches, such as model selection (select the optimal single estimator from a list of models), convex aggregation (searching for the optimal convex combination of given estimators, such as exponentially weighted aggregates) and linear aggregation (selecting the optimal linear combination).

Model selection, linear-type aggregation strategies and related problems have been studied by Catoni (2004); Györfi et al. (2002); Juditsky and Nemirovski (2000); Nemirovski (2000); Wegkamp (2003); Yang (2000, 2001, 2004). Minimax results have been derived by Nemirovski (2000) and Tsybakov (2003), leading to the notion of optimal rates of aggregation. Similar results can be found in Bunea et al. (2007). Further, upper bounds for the risk in model selection and convex aggregation have been established for instance by Audibert (2004); Birgé (2006); Dalalyan and Tsy-

bakov (2008).

Beside the usual linear aggregation and model selection methods, a quite different point of view has been introduced by Mojirsheibani (1999) for classification. The idea consists in an original combination method, which is non linear in the initial estimators and is based on a consensus concept between them. More precisely, an observation is considered to be reliable for contributing to the classification of a new query point if all initial classifiers predict the same label for both points. Then, the label for the query point is estimated thanks to a majority vote among the labels of the observations which have been retained this way. When the initial list contains a consistent estimator, it is shown that the combined estimator inherits this consistency property. Note that more regular versions, based on smoothing kernels, have also been developed (Mojirsheibani, 2000). A numerical comparison study of several combining schemes is available in Mojirsheibani (2002b), and recently, a variant of the method has been proposed in Balakrishnan and Mojirsheibani (2015).

In Biau et al. (2016), we have adapted this strategy in the regression framework. An observation is used in the combination step if all the initial estimators predict a similar value for the observation and the new point: the difference between both predictions is required to be less than some prespecified threshold. Then, the new prediction by the combined estimator is the average of the true outputs corresponding to the selected entries. Note that the functional data framework has also been considered, by Cholaquidis et al. (2016).

The resulting regression estimator is a nonlinear data-dependent function of the basic estimators. in fact, they are used to derive a local distance between a new test instance and the original training data.

In Section IV.1.1, we describe the combined estimator and derive a nonasymptotic risk bound. We present the main result, that is, the collective is asymptotically at least as good as any of the basic estimators. We also provide a rate of convergence for our procedure. Let us mention that the techniques of proof, and consequently, the assumptions are quite different in Mojirsheibani (1999) and Biau et al. (2016). For instance, the number of initial estimators is expected to tend to infinity with the sample size in Mojirsheibani (1999) whereas it is fixed here. Section IV.1.2 presents several numerical experiments, on simulated data sets, including high-dimensional models. Let us mention that the procedure is not too CPU-time consuming, in particular thanks to the possibility of parallelization. We compare our strategy with Super Learner (van der Laan et al., 2007) and exponentially weighted aggregation (see, among many other references see Dalalyan and Tsybakov, 2008).

IV.1.1 The combined estimator

We assume that we are given a training sample $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$. $\mathcal{D}_n$ of independent random variables, taking their values in $\mathbb{R}^d \times \mathbb{R}$, and distributed as a generic random pair (X, Y) satisfying $E\big[Y^2\big] < \infty$. $\mathbb{R}^d$ is equipped with the standard Euclidean metric. Our goal is to estimate the regression function $r^*(x) = E\big[Y|X = x\big]$, $x \in \mathbb{R}^d$, using the data $\mathcal{D}_n$.

The original data set $\mathcal{D}_n$ is split into two subsets $\mathcal{D}_k = \{(X_1, Y_1), \ldots, (X_k, Y_k)\}$ and $\mathcal{D}_\ell = \{(X_{k+1}, Y_{k+1}), \ldots, (X_n, Y_n)\}$, with $\ell = n - k \geq 1$. For simplicity, with a slight abuse of notation, the elements of $\mathcal{D}_\ell$ are renamed $\{(X_1, Y_1), \ldots, (X_\ell, Y_\ell)\}$.

We suppose that we are given a collection of $p \geq 1$ estimators $r_{k,1}, \ldots, r_{k,p}$, based on the first subsample $\mathcal{D}_k$.

These estimators can be any among the researcher's favorite toolkit, such as linear regression, kernel smoother, SVM, Lasso, neural networks, naive Bayes, or random forests. They could equally well be any ad hoc regression rule suggested by the experimental context. They can be parametric, nonparametric, or semi-parametric, with possible tuning rules. All that is asked for is that each of the $r_{k,m}(x)$, $m = 1, \ldots, p$, is able to provide an estimation of $r^*(x)$ on the basis of $\mathcal{D}_k$.

Here, the number of estimators p is fixed, it is not expected to grow. In Section IV.1.2, there are never more than 10 estimators in the list.

Given the collection of individual estimators $\mathbf{r}_k = (r_{k,1}, \ldots, r_{k,p})$, we define the combined estimator T_n to be

$$T_n\left(\mathbf{r}_k(x)\right) = \sum_{i=1}^{\ell} W_{n,i}(x) Y_i, \quad x \in \mathbb{R}^d,$$

where the random weights $W_{n,i}(x)$ take the form

$$W_{n,i}(x) = \frac{\mathbf{1}_{\bigcap_{m=1}^{p}\{|r_{k,m}(x) - r_{k,m}(X_i)| \leq \varepsilon_\ell\}}}{\sum_{j=1}^{\ell} \mathbf{1}_{\bigcap_{m=1}^{p}\{|r_{k,m}(x) - r_{k,m}(X_j)| \leq \varepsilon_\ell\}}}.$$

In this definition, ε_ℓ is some positive parameter and, by convention, $0/0 = 0$.

T_n is a local averaging estimator. For Y_i to be included in the average, it is required that all basic estimators predict a close value for X_i and for the query point x.

This approach is motivated by the fact that a major issue in learning problems consists in finding a metric that is suited to the data (see, e.g., the monograph by

Duin and Pekalska, 2005). In this context, ε_ℓ plays the role of a smoothing parameter, so that the practical performance of T_n strongly relies on an appropriate choice of ε_ℓ.

The condition for an observation to be reliable is here required to be satisfied for all estimators. In Mojirsheibani (2002a), the author notes that this rule may seem too restrictive and proposes to allow a few disagreements (typically, a single one). The resulting classifier is still consistent provided that the number of initial classifiers keeps tending to infinity after removing those with disagreement. Similarly, here, the unanimity constraint may be relaxed in practice by imposing, for example, that a fixed fraction $\gamma \in \{1/p, 2/p, \dots, 1\}$ of the estimators agree on the importance of X_i. In that case, the weights take the more sophisticated form

$$W_{n,i}(x) = \frac{\mathbf{1}_{\{\sum_{m=1}^{p} \mathbf{1}_{\{|r_{k,m}(x) - r_{k,m}(X_i)| \leq \varepsilon_\ell\}} \geq p\gamma\}}}{\sum_{j=1}^{\ell} \mathbf{1}_{\{\sum_{m=1}^{p} \mathbf{1}_{\{|r_{k,m}(x) - r_{k,m}(X_j)| \leq \varepsilon_\ell\}} \geq p\gamma\}}}.$$

It turns out that adding the parameter γ does not change the asymptotic properties of T_n, provided $\gamma \to 1$. For simplicity, we have decided to state the results in the case $\gamma = 1$. In Section IV.1.2, the parameters ε and γ are selected via cross-validation.

Let us turn to the theoretical study of the combined estimator T_n.

For ease of exposition, we assume that the estimators are bounded. Let

$$T\left(\mathbf{r}_k(X)\right) = E\left[Y | \mathbf{r}_k(X)\right].$$

By definition of the L^2 conditional expectation, we have

$$E\left[|T(\mathbf{r}_k(X)) - Y|^2\right] \leq \inf_f E\left[|f(\mathbf{r}_k(X)) - Y|\right]^2,$$

where the infimum is taken over all square integrable functions of $\mathbf{r}_k(X)$.

Our first result, proved in Biau et al. (2016), is a nonasymptotic inequality, which states that the combined estimator behaves as well as the best one in the original list, up to a term measuring how far T_n is from T.

Proposition IV.1.1.1. *For all distributions of (X, Y) with $E\left[Y^2\right] < \infty$,*

$$\begin{aligned} &E\left[|T_n(\mathbf{r}_k(X)) - r^*(X)|^2\right] \\ &\quad \leq E\left[|T_n(\mathbf{r}_k(X)) - T(\mathbf{r}_k(X))|^2\right] + \inf_f E\left[|f(\mathbf{r}_k(X)) - r^*(X)|^2\right], \end{aligned}$$

where the infimum is taken over all square integrable functions of $\mathbf{r}_k(X)$*. In particular,*

$$E\big[|T_n(\mathbf{r}_k(X)) - r^*(X)|^2\big] \\ \leq \min_{m=1,\dots,p} E\big[|r_{k,m}(X) - r^*(X)|^2\big] + E\big[|T_n(\mathbf{r}_k(X)) - T(\mathbf{r}_k(X))|^2\big].$$

Since Proposition IV.1.1.1 holds for any square integrable function of $\mathbf{r}_k(X)$, it allows to derive results involving any existing aggregation procedure, including for instance linear or convex aggregation.

The term $\min_{m=1,\dots,p} E\big[|r_{k,m}(X) - r^*(X)|^2\big]$ may be regarded as a bias term, whereas the term $E\big[|T_n(\mathbf{r}_k(X)) - T(\mathbf{r}_k(X))|^2\big]$ is a variance-type term, which can be asymptotically neglected, as shown in Biau et al. (2016).

Proposition IV.1.1.2. *Assume that* $\varepsilon_\ell \to 0$ *and* $\ell\varepsilon_\ell^p \to \infty$ *as* $\ell \to \infty$*. Then*

$$E\big[|T_n\left(\mathbf{r}_k(X)\right) - T\left(\mathbf{r}_k(X)\right)|^2\big] \to 0 \quad \text{as } \ell \to \infty,$$

for all distributions of (X, Y) *with* $E\big[Y^2\big] < \infty$*. Thus,*

$$\limsup_{\ell\to\infty} E\big[|T_n\left(\mathbf{r}_k(X)\right) - r^*(X)|^2\big] \leq \min_{m=1,\dots,p} E\big[|r_{k,m}(X) - r^*(X)|^2\big].$$

Thus, in terms of quadratic risk, the combined estimator T_n does asymptotically at least as well as the best primitive estimator, regardless of which initial estimator is actually the best. The result is universal, in the sense that it is true for all distributions of (X, Y).

The result does not require any regularity assumption on the basic estimators. However, this universality comes at a price since we have no guarantee on the rate of convergence of the variance term. Nevertheless, with light additional smoothness conditions, one has the following statement.

Theorem IV.1.1.1. *Assume that* Y *and the basic estimators* $\mathbf{r}_k$ *are bounded by some constant* R*, and that there exists* $L \geq 0$ *such that, for every* $k \geq 1$*,*

$$|T(\mathbf{r}_k(x)) - T(\mathbf{r}_k(y))| \leq L|\mathbf{r}_k(x) - \mathbf{r}_k(y)|, \quad x, y \in \mathbb{R}^d.$$

Then, with the choice $\varepsilon_\ell \propto \ell^{-\frac{1}{p+2}}$*, one has*

$$E\big[|T_n\left(\mathbf{r}_k(X)\right) - r^*(X)|^2\big] \leq \min_{m=1,\dots,p} E\big[|r_{k,m}(X) - r^*(X)|^2\big] + C\ell^{-\frac{2}{p+2}},$$

where C *depends on* R *and* L*.*

Theorem IV.1.1.1 offers an oracle-type inequality with leading constant 1, stating that the risk of the combined estimator is bounded by the lowest risk among those of the basic estimators, plus a remainder term of the order of $\ell^{-2/(p+2)}$, which is the price to pay for aggregating. In our setting, it is important to observe that this term has a limited impact. Indeed, since the number of basic estimators p is assumed to be fixed and not too large (the implementation presented in Section IV.1.2 considers p at most 5), the remainder term is negligible compared to the standard nonparametric rate $\ell^{-2/(d+2)}$ in dimension d. While the rate $\ell^{-2/(d+2)}$ is affected by the curse of dimensionality when d is large, this is not the case for the term $\ell^{-2/(p+2)}$. Obviously, if the distribution of (X, Y) can in fact be described parametrically, faster rates of the order of $1/\ell$ may appear in the bias term.

If one of the initial estimators is consistent for a given class $\mathcal{P}$ of distributions, then, under appropriate smoothness assumptions, T_n inherits the same property. To be more precise, assume that one of the original estimators, say r_{k,m_0}, satisfies

$$E\big[|r_{k,m_0}(X) - r^*(X)|\big]^2 \to 0 \quad \text{as } k \to \infty,$$

for all distribution of (X, Y) in some class $\mathcal{P}$. Then, under the assumptions of Theorem IV.1.1.1, with the choice $\varepsilon_\ell \propto \ell^{-\frac{1}{p+2}}$, one has

$$\lim_{k,\ell\to\infty} E\left|T_n\left(\mathbf{r}_k(X)\right) - r^*(X)\right|^2 = 0.$$

IV.1.2 Numerical study

This section is devoted to a series of numerical experiments, conducted with the R package `Cobra` (standing for COmBined Regression Alternative) implemented by Guedj (2013).

The performance of the method is measured using $R(\hat{Y}_1, \ldots, \hat{Y}_m) = \frac{1}{m}\sum_{j=1}^m (\hat{Y}_j - Y_j)^2$, where $\{(X_j, Y_j)\}_{j=1}^m$ is a testing sample and $\hat{Y}_i$ denotes the predicted value for Y_i.

As raised in the previous section, a precise calibration of the smoothing parameter ε_ℓ is crucial. Clearly, a too small value will discard many estimators, with most weights equal to zero. Conversely, that is too large just predicts the mean over the entire sample $\mathcal{D}_\ell$. We consider the relaxed version of the unanimity constraint. Instead of requiring global agreement over the implemented estimators, we keep observation Y_i in the construction of T_n if there is at least a proportion $\gamma \in (0, 1]$

of the estimators agreeing on the importance of X_i. This parameter, which may be seen as a measure of homogeneity of the estimators, also requires calibration. Both parameters ε_ℓ and γ are chosen by cross-validation, from a grid

$$\{\varepsilon_{\ell,\min}, \ldots, \varepsilon_{\ell,\max}\} \times \{1/p, \ldots, 1\},$$

where $\varepsilon_{\ell,\min} = 10^{-300}$ and $\varepsilon_{\ell,\max}$ is proportional to the largest absolute difference between two predictions of the pool of estimators.

We use the following individual estimators.

— Lasso (R package `lars`, see Hastie and Efron, 2012).

— Ridge regression (R package `ridge`, see Cule, 2012).

— k-nearest neighbors (R package `FNN`, see Li, 2013).

— CART algorithm (R package `tree`, see Ripley, 2012).

— Random Forest algorithm (R package `randomForest`, see Liaw and Wiener, 2002, `RF` hereafter).

The training data set was set to 80% of the whole sample, then split into two equal parts corresponding to $\mathcal{D}_k$ and $\mathcal{D}_\ell$. In the first experiments presented, we consider each time 2 different error distributions: Uniform over $(-1, 1)^d$ (referred to as "Uncorrelated" below), and Gaussian with mean 0 and covariance matrix Σ with $\Sigma_{ij} = 2^{-|i-j|}$ ("Correlated"). Let Z $\mathcal{N}(0, 0.5)$.

Model 1. $n = 800$, $d = 50$, $Y = X_1^2 + \exp(-X_2^2)$.

Model 2. $n = 600$, $d = 100$, $Y = X_1X_2 + X_3^2 - X_4X_7 + X_8X_{10} - X_6^2 + Z$.

Model 3. $n = 600$, $d = 100$, $Y = -\sin(2X_1) + X_2^2 + X_3 - \exp(-X_4) + Z$.

Model 4. $n = 600$, $d = 100$, $Y = X_1 + (2X_2 - 1)^2 + \sin(2\pi X_3)/(2 - \sin(2\pi X_3)) + \sin(2\pi X_4) + 2\cos(2\pi X_4) + 3\sin^2(2\pi X_4) + 4\cos^2(2\pi X_4) + Z$.

Model 5. $n = 700$, $d = 20$, $Y = \mathbf{1}_{\{X_1>0\}} + X_2^3 + \mathbf{1}_{\{X_4+X_6-X_8-X_9>1+X_{14}\}} + \exp(-X_2^2) + Z$.

Model 6. $n = 600$, $d = 300$, $Y = X_1^2 + X_2^2X_3\exp(-|X_4|) + X_6 - X_8 + Z$.

Some models are borrowed form the literature: Mod. 2 comes from van der Laan et al. (2007), Mod. 3 and 4 appear in Meier et al. (2009). Table IV.1.1 presents the mean squared error and standard deviation over 100 independent replications, for each model and design. Bold numbers identify the lowest error. Corresponding boxplots are presented in Figures IV.1.1 and IV.1.2. Further, Figures IV.1.3 and IV.1.4 shows predicted versus true output values. We observe that the procedure performs at least as well as the best estimator, significantly so in Mod. 3 and 5 for instance.

Uncorr.		`lars`	`ridge`	`FNN`	`tree`	`RF`	`Cobra`
Mod. 1	m.	0.1561	0.1324	0.1585	0.0281	0.0330	**0.0259**
	sd.	0.0123	0.0094	0.0123	0.0043	0.0033	0.0036
Mod. 2	m.	0.4880	0.4562	0.3070	0.1746	**0.1366**	0.1645
	sd.	0.0676	0.0233	0.0303	0.0270	0.0161	0.0207
Mod. 3	m.	0.2536	0.5347	1.1603	0.4954	0.4027	**0.2332**
	sd.	0.0271	0.4469	0.1227	0.0772	0.0558	0.0272
Mod. 4	m.	7.6056	6.3271	10.5890	3.7358	3.5262	**3.3640**
	sd.	0.9419	1.0800	0.9404	0.8067	0.3223	0.5178
Mod. 5	m.	0.2943	0.3311	0.5169	0.2918	0.2234	**0.2060**
	sd.	0.0214	0.1012	0.0439	0.0279	0.0216	0.0210
Mod. 6	m.	1.0920	0.5452	0.9459	0.3638	0.3110	**0.3052**
	sd.	0.2265	0.0920	0.0833	0.0456	0.0325	0.0298

Corr.		`lars`	`ridge`	`FNN`	`tree`	`RF`	`Cobra`
Mod. 1	m.	2.3736	1.9785	2.0958	0.3312	0.5766	**0.3301**
	sd.	0.4108	0.3538	0.3414	0.1285	0.1914	0.1239
Mod. 2	m.	8.1710	4.0071	4.3892	**1.3609**	1.4768	1.3612
	sd.	1.5532	0.6840	0.7190	0.4647	0.4415	0.4654
Mod. 3	m.	6.1448	6.0185	8.2154	4.3175	4.0177	**3.7917**
	sd.	11.9450	12.0861	13.3121	11.7386	12.4160	11.1806
Mod. 4	m.	60.5795	42.2117	51.7293	**9.6810**	14.7731	9.6906
	sd.	11.1303	9.8207	10.9351	3.9807	5.9508	3.9872
Mod. 5	m.	6.2325	7.1762	10.1254	3.1525	4.2289	**2.1743**
	sd.	2.4320	3.5448	3.1190	2.1468	2.4826	1.6640
Mod. 6	m.	20.8575	4.4367	5.8893	3.6865	**2.7318**	2.9127
	sd.	7.1821	1.0770	1.2226	1.0139	0.8945	0.9072

Table IV.1.1 – Errors of the implemented estimators and combination: means and standard deviations over 100 independent replications

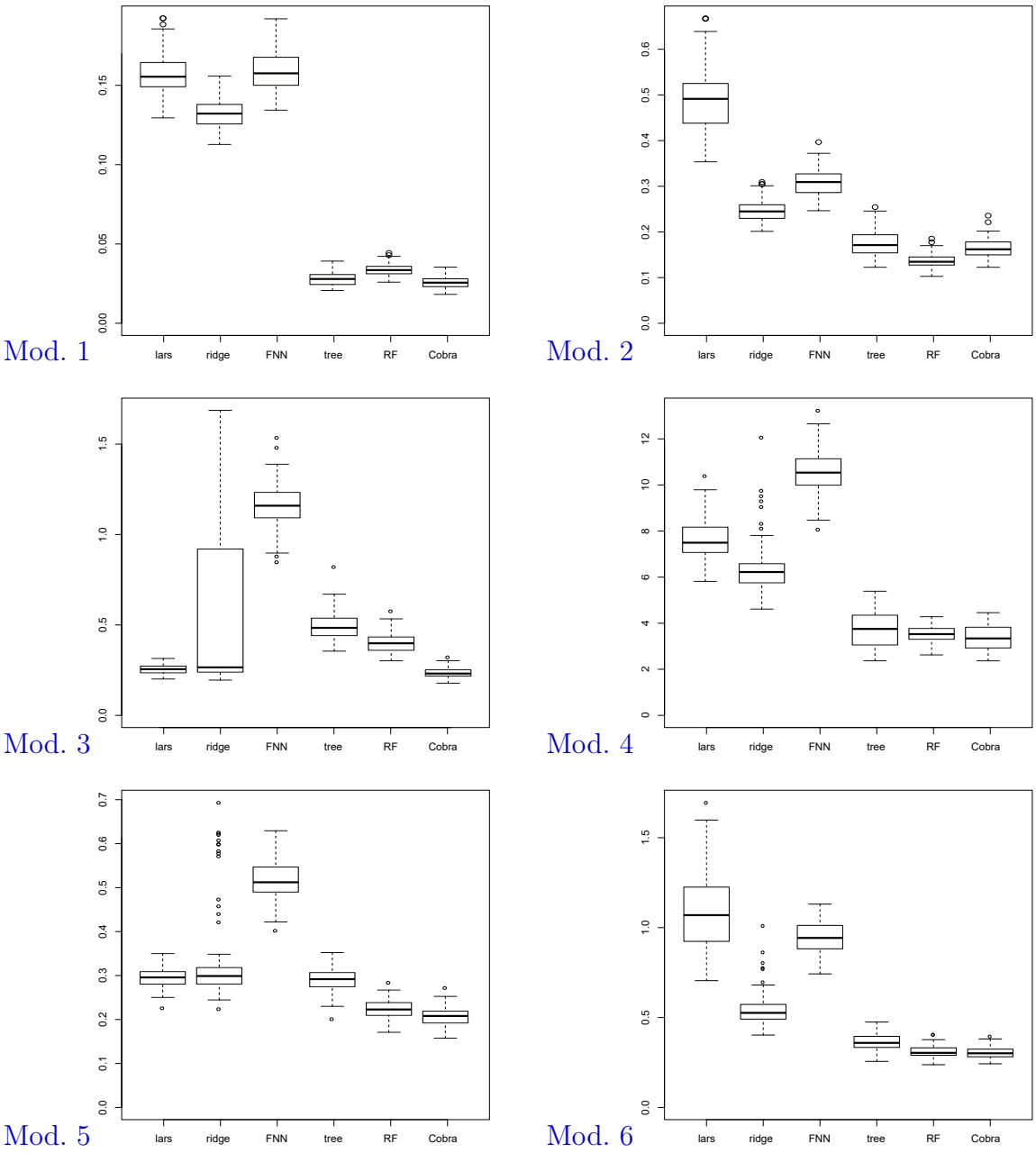


Figure IV.1.1 – Boxplots of errors, uncorrelated design.

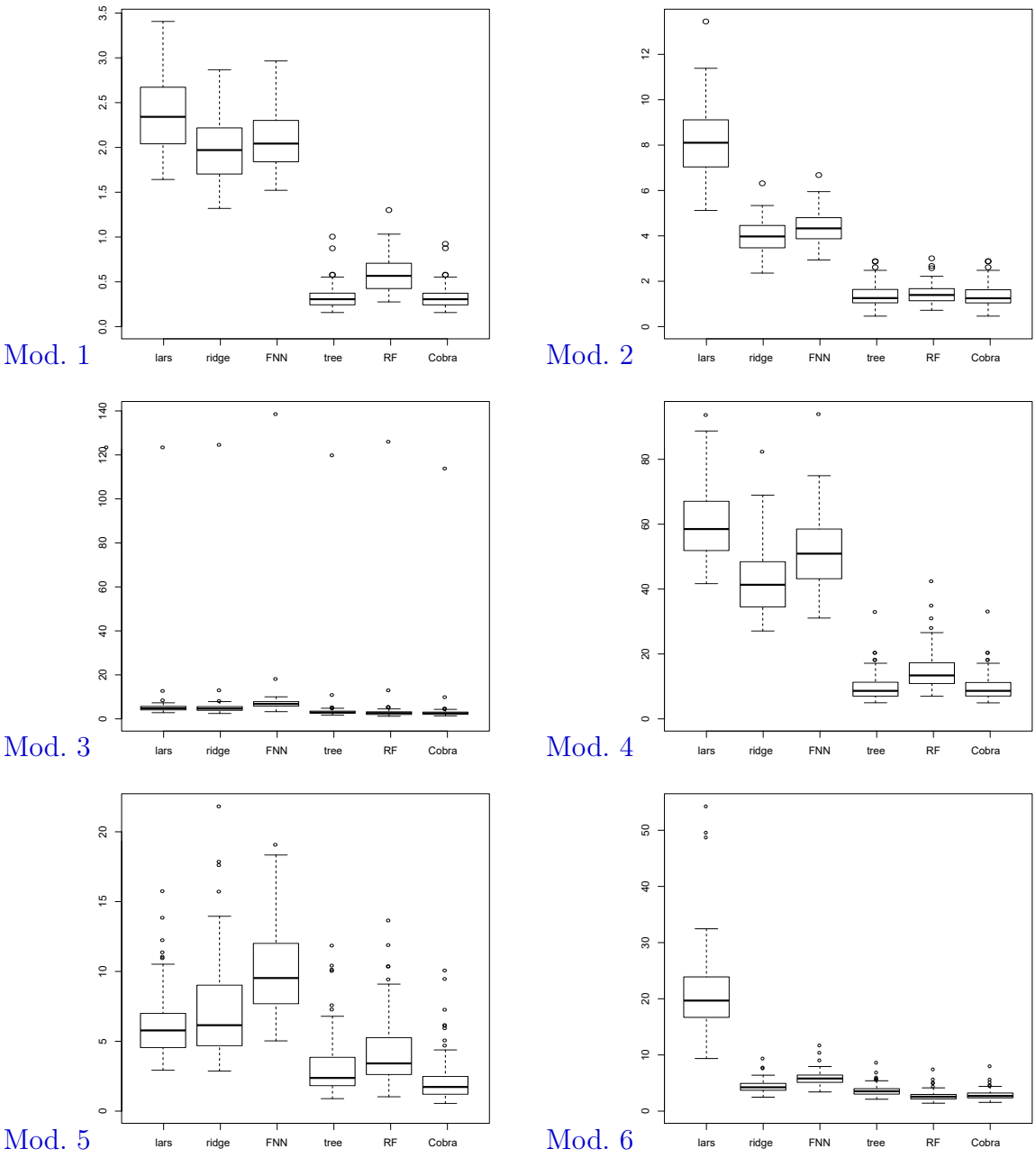
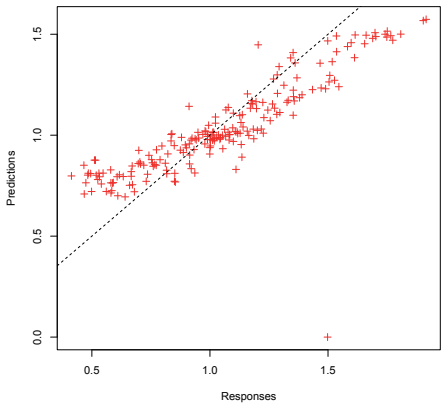
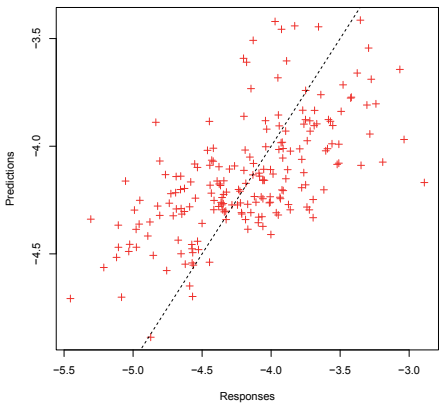


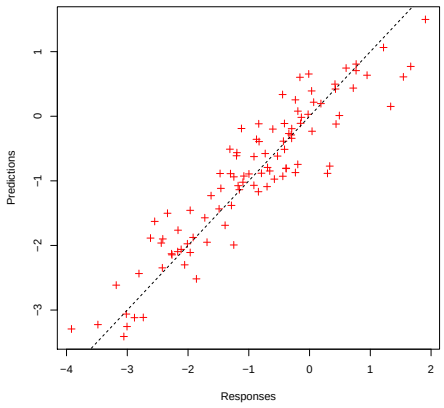
Figure IV.1.2 – Boxplots of errors, correlated design.



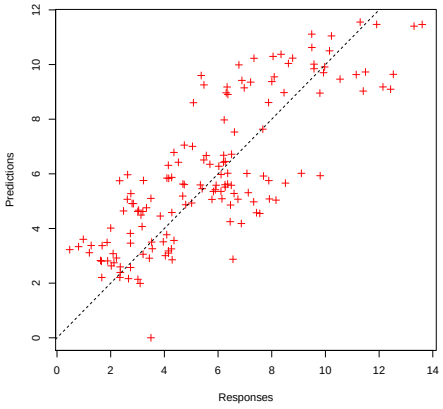
Mod. 1



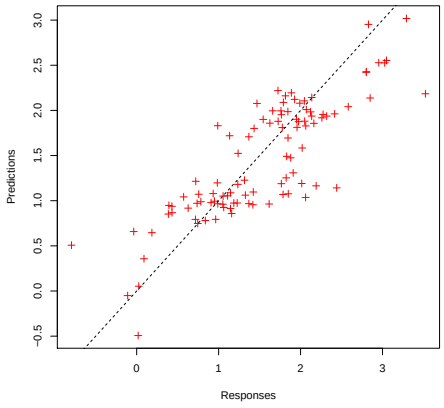
Mod. 2



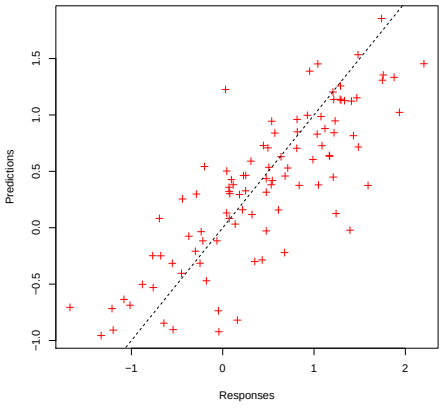
Mod. 3



Mod. 4

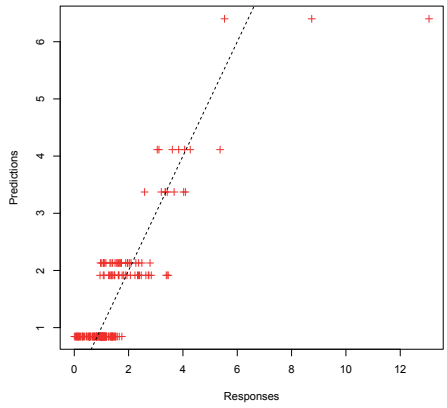


Mod. 5

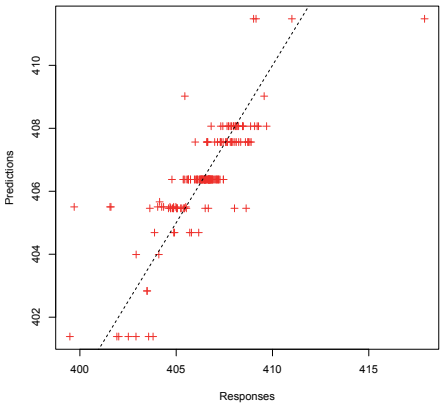


Mod. 6

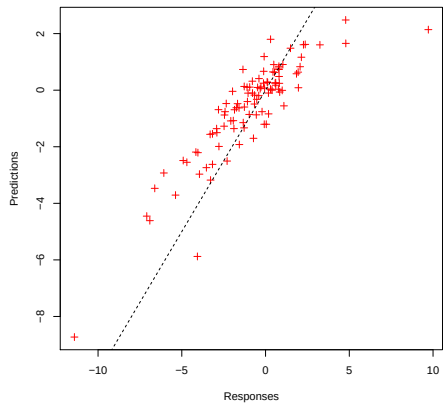
Figure IV.1.3 – Predicted versus true outputs values, uncorrelated design



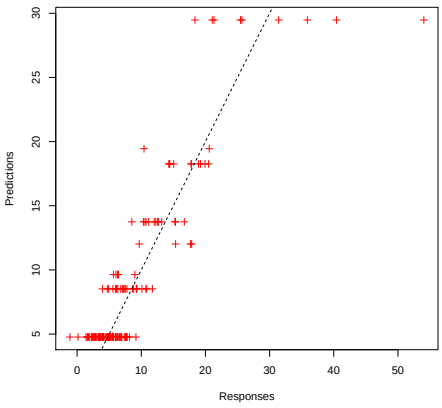
Mod. 1



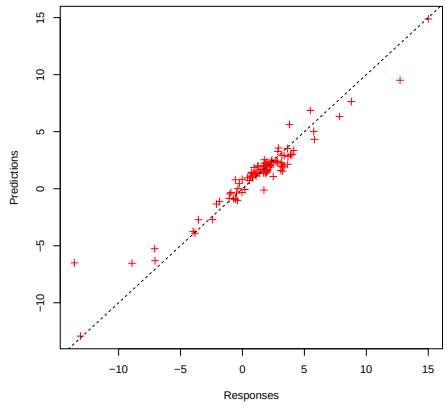
Mod. 2



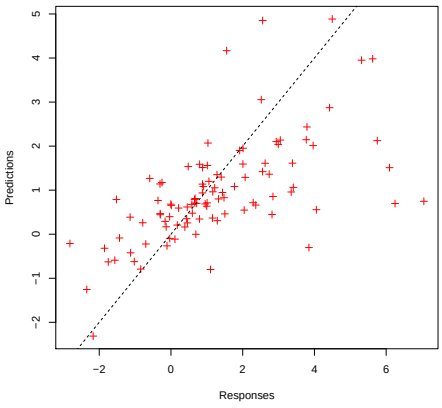
Mod. 3



Mod. 4



Mod. 5



Mod. 6

Figure IV.1.4 – Predicted versus true outputs values, correlated design

Since many problems in contemporary statistics involve high-dimensional data, we have also tested the abilities of our combination method in this context. To this aim, we simulated 200 independent replications following the next models:

Model 7. $n = 500$, $d = 1000$, $Y = X_1 + 3X_3^2 - 2\exp(-X_5) + X_6$. Uncorrelated design.

Model 8. $n = 500$, $d = 1000$, $Y = X_1 + 3X_3^2 - 2\exp(-X_5) + X_6$. Correlated design.

Model 9. $n = 500$, $d = 1500$, $Y = \exp(-X_1) + \exp(X_1) + \sum_{j=2}^{d} X_j^{j/100}$. Uncorrelated design.

As highlighted by Figure IV.1.5 and Table IV.1.2, the procedure is perfectly able to deal with high-dimensional data, provided that it is generated over estimators, at least some of which are known to perform well in such situations, possibly at the price of a sparsity assumption.

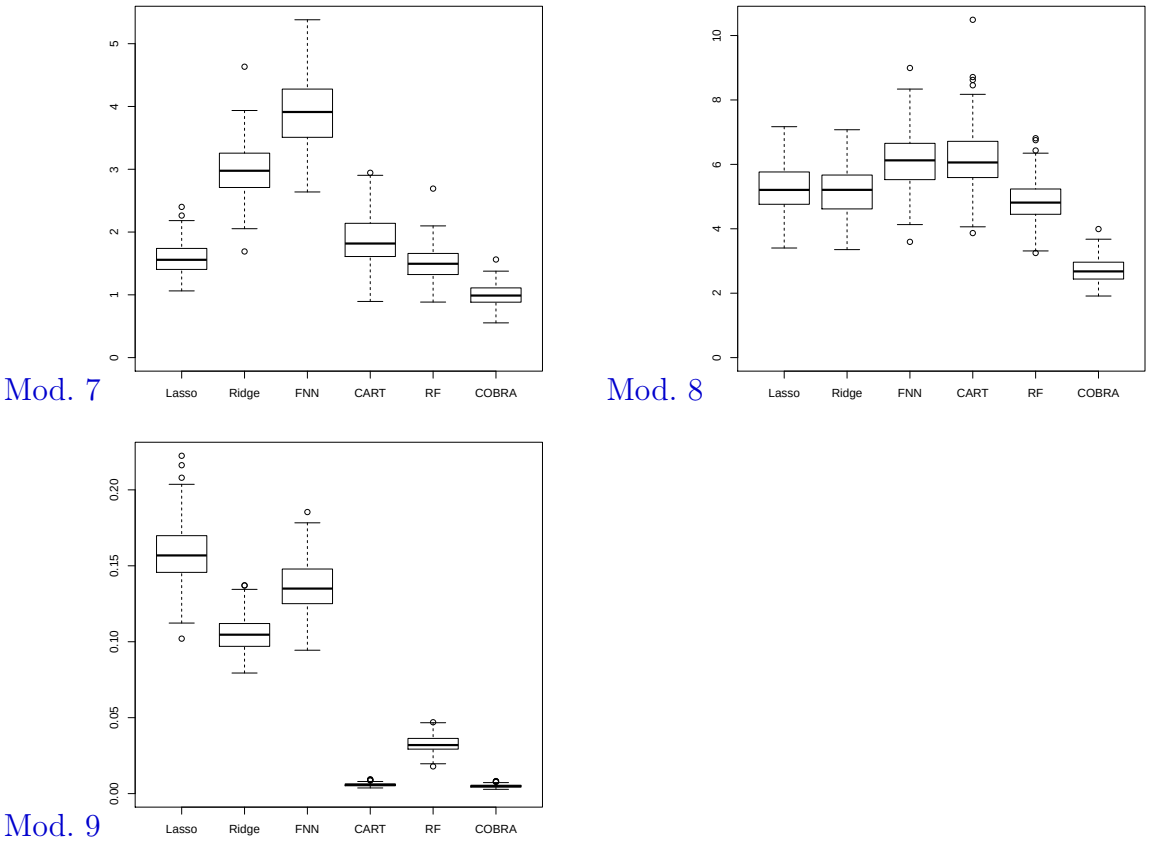


Figure IV.1.5 – Boxplot of errors, high-dimensional models

		`lars`	`ridge`	`FNN`	`tree`	`RF`	`Cobra`
Mod. 7	m.	1.5698	2.9752	3.9285	1.8646	1.5001	**0.9996**
	sd.	0.2357	0.4171	0.5356	0.3751	0.4591	0.1733
Mod. 8	m.	5.2356	5.1748	6.1395	6.1585	4.8667	**2.7076**
	sd.	0.6885	0.7139	0.9192	0.9298	0.6634	0.3810
Mod. 9	m.	0.1584	0.1055	0.1363	0.0058	0.0327	**0.0049**
	sd.	0.0199	0.0119	0.0176	0.0010	0.0052	0.0009

Table IV.1.2 – Errors of the implemented estimators and combination in high-dimensional situations: means and standard deviations over 200 independent replications

Finally, we compare our procedure with the Super Learner algorithm (Polley and van der Laan, 2012) and the exponentially weighted aggregation method (EWA).

The Super Learner algorithm was first described in van der Laan et al. (2007) and extended in Polley and van der Laan (2010). In short, this ensemble method, also called stacking, computes a weighted combination of the basic estimators, with optimization based on V-fold cross-validation.

Table IV.1.3 summarizes the performances. Both methods, Super Learner and our strategy, use `SL.randomForest`, `SL.ridge` and `SL.glmnet` as individual estimation procedures, for the fairness of the comparison. On the whole, they show similar results. However, our procedure turns out to be much more efficient, for instance, on correlated design in Mod. 2 and Mod. 4. Observe, in Table IV.1.4, that even without parallelization, `Cobra` is about five times faster.

Exponentially weighted aggregation method (EWA) for preliminary estimators $r_{k,1}, \ldots, r_{k,p}$ was implemented as follows: for $x \in \mathbb{R}^d$,

$$\mathrm{EWA}_\beta(x) = \sum_{j=1}^{p} \hat{w}_j r_{k,j}(x), \qquad \hat{w}_j = \frac{\exp(-\beta R_j)}{\sum_{i=1}^{p} \exp(-\beta R_i)}, \quad j = 1, \ldots, p,$$

where $R_1, \ldots, R_p$ are the individual mean squared errors computed on a subsample of $\mathcal{D}_\ell$, and $\beta > 0$ is the so-called temperature parameter, selected over a grid. We conducted 200 independent replications, on Mod. 7 to Mod. 9. We observe that `Cobra` leads to results that are roughly comparable to those of the EWA estimator, as shown in Table IV.1.5 and Figure IV.1.6.

Uncorr.		SL	Cobra
Mod. 1	m.	0.0541	**0.0320**
	sd.	0.0053	0.0104
Mod. 2	m.	**0.1765**	0.3569
	sd.	0.0167	0.8797
Mod. 3	m.	**0.2081**	0.2573
	sd.	0.0282	0.0699
Mod. 4	m.	4.3114	**3.7464**
	sd.	0.4138	0.8746
Mod. 5	m.	**0.2119**	0.2187
	sd.	0.0317	0.0427
Mod. 6	m.	**0.1705**	0.3103
	sd.	0.0260	0.0490

Corr.		SL	Cobra
Mod. 1	m.	0.8733	**0.3262**
	sd.	0.2740	0.1242
Mod. 2	m.	2.3391	**1.3984**
	sd.	0.4958	0.3804
Mod. 3	m.	**3.1885**	3.3201
	sd.	1.5101	1.8056
Mod. 4	m.	25.1073	**9.3964**
	sd.	7.3179	2.8953
Mod. 5	m.	5.6478	**4.9990**
	sd.	7.7271	9.3103
Mod. 6	m.	**3.0367**	3.1401
	sd.	1.6225	1.6097

Table IV.1.3 – Errors of `Cobra` and `SuperLearner` : means and standard deviations over 100 independent replications

Uncorr.		SL	Cobra
Mod. 1	m.	53.92	**10.92**
	sd.	1.42	0.29
Mod. 2	m.	57.96	**11.90**
	sd.	0.95	0.31
Mod. 3	m.	53.70	**10.66**
	sd.	0.55	0.11
Mod. 4	m.	55.00	**11.15**
	sd.	0.74	0.18
Mod. 5	m.	28.46	**5.01**
	sd.	0.73	0.06
Mod. 6	m.	127.80	**35.67**
	sd.	5.69	1.91

Corr.		SL	Cobra
Mod. 1	m.	61.92	**11.96**
	sd.	1.85	0.27
Mod. 2	m.	70.90	**14.16**
	sd.	2.47	0.57
Mod. 3	m.	59.91	**11.92**
	sd.	2.06	0.41
Mod. 4	m.	63.58	**13.11**
	sd.	1.21	0.34
Mod. 5	m.	31.45	**5.02**
	sd.	0.86	0.07
Mod. 6	m.	145.18	**41.28**
	sd.	8.97	2.84

Table IV.1.4 – Average CPU-times in seconds (no parallelization): means and standard deviations over 10 independent replications

		EWA	Cobra
Mod. 7	m.	1.1712	**1.1360**
	sd.	0.2090	0.4568
Mod. 8	m.	**9.4789**	12.4353
	sd.	5.6275	9.1267
Mod. 9	m.	0.0244	**0.0128**
	sd.	0.0042	0.0237

Table IV.1.5 – Errors of EWA and `Cobra`: means and standard deviations over 200 independent replications

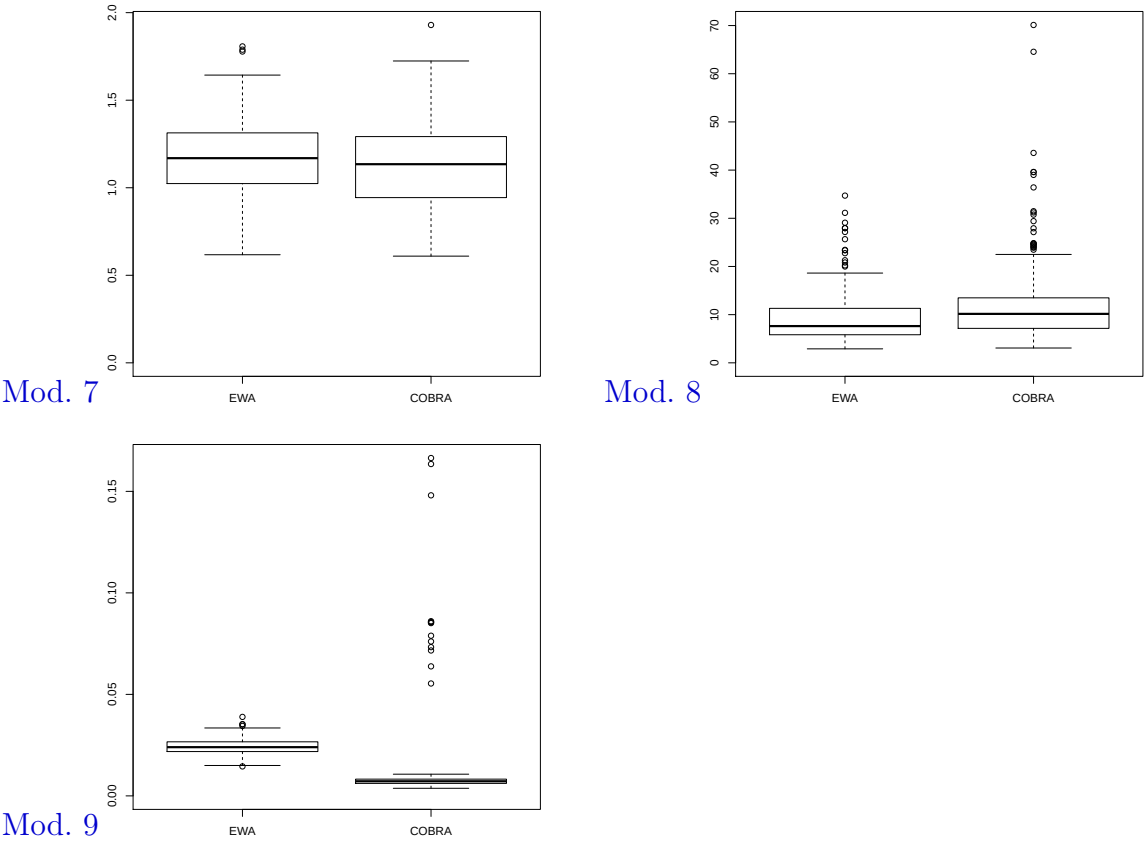


Figure IV.1.6 – Boxplot of errors: `EWA` vs `Cobra`

Chapter IV.2

Adding distance information

This chapter is the result of a collaboration with Mathilde Mougeot (Centre Borelli, ENS Paris-Saclay), published in the Journal of Statistical Planning and Inference *(Fischer and Mougeot, 2019).*

IV.2.1 Introduction

We focus on the question of combining estimators in classification and regression, using the same consensus idea as in Chapter IV.1.

Recall that, in Mojirsheibani (1999) (for classification), as well as in Chapter IV.1 (for regression), the condition for an observation to be reliable is in principle required to be satisfied for all estimators. In a further article, Mojirsheibani (2002a) notes that this rule may seem too restrictive and proposes to allow a few disagreements (typically, a single one). The resulting classifier is still consistent provided that the number of initial classifiers keeps tending to infinity after removing those with disagreement. Similarly, in Chapter IV.1, we have seen that this unanimity constraint may be relaxed in practice by demanding that the distance condition for keeping an observation is true at least for a certain proportion α of the estimators (for example, $\alpha = 80\%$).

Here, our purpose is to investigate a new approach, based on distances between observations, which also aims to reduce the effect of a possibly bad initial estimator. Roughly, choosing a kernel point of view, we will propose a combined estimator with weights constructed by mixing distances between entries with distances between pre-

dictions coming from the individual estimators. Our motivation for introducing such a strategy is the intuition that taking advantage of the efficiency of the consensus idea of Mojirsheibani (1999) and Chapter IV.1 without for all that forgetting the information related to the proximity between entries shall help improving the prediction, especially in the presence of an initial estimator that does not perform very well.

Our modified rule is consistent under general assumptions (Section IV.2.3). In particular, the combined estimator may perfectly be consistent even if the list of initial estimators does not contain any consistent estimator. We also conduct numerical experiments, both on simulated and real data, which demonstrate the benefits of our strategy, with respect to the original combining method and the individual estimators (Section IV.2.4).

IV.2.2 Notation and definition of the estimator

Let (X, Y) denote a random pair taking its values in $\mathbb{R}^d \times \mathcal{Y}$. The variable X has distribution μ. We are interested in two different situations: $\mathcal{Y} = \{0, 1\}$, which corresponds to the binary classification problem, and $\mathcal{Y} = [0, 1]$, that is bounded regression. Let η stand for the regression function $\eta(x) = E[Y|X = x]$. Note that $\eta(x) = P(Y = 1|X = x)$ in the classification context.

Let ψ^* denote the Bayes classifier, given by

$$\psi^*(x) = \begin{cases} 1 & \text{if } \eta(x) > 1/2 \\ 0 & \text{otherwise.} \end{cases}$$

It is well-known that ψ^* minimizes over all possible classifiers ψ the missclassification error $L(\psi) = P(\psi(X) \neq Y)$.

We assume that we are given a sample $\mathcal{D}_n = \{(X_1, Y_1) \ldots, (X_n, Y_n)\}$ of the random pair (X, Y).

Our goal in regression is to estimate the function η using $\mathcal{D}_n$. In classification, we aim at building a classifier based on $\mathcal{D}_n$ whose error mimics the Bayes classifier error.

Let $K : \mathbb{R}^{d+p} \mapsto \mathbb{R}_+$ be a kernel, that is a function which is nonnegative and decreasing along rays *ting from the origin. The next assumption will be made on the kernel K (see Devroye et al., 1996).

Assumption 1. We suppose that the kernel K is regular, that is, there exist $c > 0$ and $\rho > 0$ such that

— For all z, $K(z) \geq c\mathbf{1}_{B_{d+p}(0,\rho)}(z)$.

— $\int \sup_{t \in B_{d+p}(z,\rho)} K(t)dz < \infty$.

Here $B_k(x, r)$ denotes the closed ball of $\mathbb{R}^d$, with center x and radius r.

We propose to combine the predictions of p initial estimators $f_1, \ldots, f_p$. For $x \in \mathbb{R}^d$, let $\mathbf{f}(x) = (f_1(x), \ldots, f_p(x))$. We suppose that the estimators $f_1, \ldots, f_p$ take their values in $[0, 1]$ (regression) or $\{0, 1\}$ (classification). For ease of exposition, we assume that $f_1, \ldots, f_p$ do not depend on the sample $\mathcal{D}_n$. Using a simple sample-splitting device as in Chapter IV.1, the results extend to the case where the individual estimators depend on the sample.

The definition of our combined estimator is first introduced in the regression framework.

Let the function $g : \mathbb{R}^d \times \mathbb{R}^p \to \mathbb{R}_+$ be such that $g(v_1, v_2) = K(v)$, where $v \in \mathbb{R}^{d+p}$ is the concatenation of $v_1 \in \mathbb{R}^d$ and $v_2 \in \mathbb{R}^p$.

Definition IV.2.2.1. *Suppose that we are given a set of initial regression estimators $f_1, \ldots, f_p$. The regression combined estimator T_n is defined by*

$$T_n(x) = \frac{\sum_{i=1}^n Y_i g\left(\frac{X_i - x}{\alpha}, \frac{\mathbf{f}(X_i) - \mathbf{f}(x)}{\beta}\right)}{\sum_{i=1}^n g\left(\frac{X_i - x}{\alpha}, \frac{\mathbf{f}(X_i) - \mathbf{f}(x)}{\beta}\right)},$$

We now state the definition of the estimator in the context of classification.

Definition IV.2.2.2. *Suppose that we are given a set of initial classifiers $f_1, \ldots, f_p$. The combined classifier C_n is defined by*

$$C_n(x) = \begin{cases} 0 & \text{if } \sum_{i=1}^n Y_i g\left(\frac{X_i - x}{\alpha}, \frac{\mathbf{c}(X_i) - \mathbf{c}(x)}{\beta}\right) \leq \sum_{i=1}^n (1 - Y_i) g\left(\frac{X_i - x}{\alpha}, \frac{\mathbf{c}(X_i) - \mathbf{c}(x)}{\beta}\right) \\ 1 & \text{otherwise,} \end{cases}$$

Remark 24. In the particular case where K is such that $K(x) = H(|x|^2)$ for $x \in \mathbb{R}^{d+p}$, then the quantity $g\left(\frac{X-x}{\alpha}, \frac{\mathbf{f}(X) - \mathbf{f}(x)}{\beta}\right)$ takes the somewhat more explicit form $H\left(\left|\frac{X-x}{\alpha}\right|^2 + \left|\frac{\mathbf{f}(X) - \mathbf{f}(x)}{\beta}\right|^2\right)$. This is the case for a Gaussian kernel for instance.

Remark 25. It is worth pointing out that further definitions of combining estimate linking a consensus aggregation part and distances between entries part could be investigated. A point of view that seem very appealing is to employ general multivariate kernels and allow the bandwidth to be different depending on the direction, that is the machine and also the direction in $\mathbb{R}^d$. The advantage of this method would be a greater flexibility through his data-driven adaptive nature, whereas the increase of the number of parameters necessary for the estimation of a bandwidth matrix would represent a challenge both for estimation and computational aspects.

IV.2.3 Main results

The convergence results in regression and classification, proved in Fischer and Mougeot (2019), are stated below.

Theorem IV.2.3.1 (Regression case)**.** *If $\alpha \to 0$ and $n\alpha^d\beta^p \to \infty$ as $n \to \infty$, then, for every $\varepsilon > 0$, there exists n_0 such that for $n \geq n_0$, the following exponential bound holds:*

$$P\left(\int |\eta(x) - T_n(x)|\mu(dx) > \varepsilon\right) \leq 2\exp\left(-\frac{n\varepsilon^2}{32R^2}\right),$$

where $R \geq 0$ depends on K and $d+p$.

Let L^* denote the Bayes error and L_n the missclassification error of C_n.

Theorem IV.2.3.2 (Classification case)**.** *If $\alpha \to 0$ and $n\alpha^d\beta^p \to \infty$ as $n \to \infty$, then, for every $\varepsilon > 0$, there exists n_0 such that for $n \geq n_0$, the following exponential bound holds:*

$$P(L_n - L^* > \varepsilon) \leq 2\exp\left(-\frac{n\varepsilon^2}{32R^2}\right),$$

where $R \geq 0$ depends on K and $d+p$.

These results may be seen as "combining" versions of the strong consistency results for kernel regression and the kernel classification rule, described in Devroye and Krzyżak (1989) (see also the monographs of Devroye et al. (1996) and Györfi et al. (2002)).

Let us provide some insight into the proof. We introduce the notation T_n^*, for the quantity defined by

$$\frac{\sum_{i=1}^n Y_i g\left(\frac{X_i - x}{\alpha}, \frac{\mathbf{f}(X_i) - \mathbf{f}(x)}{\beta}\right)}{nE\left[g\left(\frac{X-x}{\alpha}, \frac{\mathbf{f}(X) - \mathbf{f}(x)}{\beta}\right)\right]}.$$

For Theorem IV.2.3.1 (regression), let us write

$$\int |\eta(x) - T_n(x)|\mu(dx) \leq \int |\eta(x) - T_n^*(x)|\mu(dx) + \int |T_n^*(x) - T_n(x)|\mu(dx).$$

Thus, the result will be obtained by replacing, on the one hand, T_n by T_n^*, and by controlling, on the other hand, the error due to the difference between the two terms. This is done respectively in Lemmas IV.2.3.1 and IV.2.3.3 below.

Lemma IV.2.3.1. *If $\alpha \to 0$ and $n\alpha^d\beta^p \to \infty$ as $n \to \infty$, then there exists $R \geq 0$, depending on K and $d+p$, such that for every $\varepsilon > 0$, if n is large enough,*

$$P\left(\int |\eta(x) - T_n^*(x)|\mu(dx) > \varepsilon/2\right) \leq \exp\left(-\frac{n\varepsilon^2}{32R^2}\right).$$

To control the quantity $\int |\eta(x) - T_n^*(x)|\mu(dx)$ in order to prove Lemma IV.2.3.1, we may use the following decomposition, for $x \in \mathbb{R}^d$:

$$|\eta(x) - T_n^*(x)| = E[|\eta(x) - T_n^*(x)|] + (|\eta(x) - T_n^*(x)| - E[|\eta(x) - T_n^*(x)|]).$$

The term $\int E[|\eta(x) - T_n^*(x)|]\mu(dx)$ is studied first and then McDiarmid's inequality is employed to handle the deviation $\int |\eta(x) - T_n^*(x)|\mu(dx) - \int E[|\eta(x) - T_n^*(x)|]\mu(dx)$.

The proof uses the following important result, which extends the covering lemma of Devroye and Krzyżak (1989) to our context.

Lemma IV.2.3.2 (Covering lemma)**.** *1. There exists $R \geq 0$, depending on K and $d+p$, such that*

$$\sup_u \int \frac{g\left(\frac{u-x}{\alpha}, \frac{\mathbf{f}(u)-\mathbf{f}(x)}{\beta}\right)}{E\left[g\left(\frac{X-x}{\alpha}, \frac{\mathbf{f}(X)-\mathbf{f}(x)}{\beta}\right)\right]}\mu(dx) \leq R < +\infty.$$

2. $\forall \delta, \varepsilon$, there exists α_0 such that

$$\sup_{u, \alpha \leq \alpha_0} \int \frac{g\left(\frac{u-x}{\alpha}, \frac{\mathbf{f}(u)-\mathbf{f}(x)}{\beta}\right)\mathbf{1}_{\{|x-u| \geq \delta\}}}{E\left[g\left(\frac{X-x}{\alpha}, \frac{\mathbf{f}(X)-\mathbf{f}(x)}{\beta}\right)\right]}\mu(dx) \leq \varepsilon.$$

The next lemma is devoted to the control of the difference between T_n and T_n^*.

Lemma IV.2.3.3. *If $\alpha \to 0$ and $n\alpha^d\beta^p \to \infty$ as $n \to \infty$, then there exists $R \geq 0$, depending on K and $d+p$, such that for every $\varepsilon > 0$, if n is large enough,*

$$P\left(\int |T_n^*(x) - T_n(x)|\mu(dx) > \varepsilon/2\right) \leq \exp\left(-\frac{n\varepsilon^2}{8R^2}\right).$$

For Theorem IV.2.3.2 (classification), we use the following version of Theorem 2.3 in Devroye et al. (1996) adapted to our context.

Lemma IV.2.3.4. *The following upper bound holds:*

$$L_n - L^* \leq \int \left|1 - \eta(x) - \frac{\sum_{i=1}^n g\left(\frac{X_i - x}{\alpha}, \frac{\mathbf{f}(X_i) - \mathbf{f}(x)}{\beta}\right)}{nE\left[g\left(\frac{X-x}{\alpha}, \frac{\mathbf{f}(X) - \mathbf{f}(x)}{\beta}\right)\right]} + T_n^*(x)\right| \mu(dx) + \int |\eta(x) - T_n^*(x)|\mu(dx).$$

Consequently, Theorem IV.2.3.2 follows from Lemma IV.2.3.1, applied to Y and to $1 - Y$.

IV.2.4 Numerical Experiments

The section presents numerical experiments to illustrate the benefits of using the new combining approach. The classification case is illustrated with numerical simulations and the regression case with real operational data recorded from two applications: modeling of the electrical power consumption of an air compressor and modeling of the electricity production of different wind turbines in a wind farm.

For comparison purposes, two aggregation strategies are run, the original method developed in Mojirsheibani (1999) and Chapter IV.1, which only combines output predictions (`Cobra` hereafter, from the name of the R package Guedj, 2013), and the new strategy, which combines input distances and output predictions (called `MixCobra`).

IV.2.4.1 Classification

We consider 6 different binary classification models, based on uniform and Gaussian distributions, called "gauss", "comete", "nuclear", "spot", "circles" and "spirals" in the sequel. Each time, $n = 200$ observations (100 per class) were simulated.

The data shapes are depicted in Figure IV.2.1, with $n = 1000$ observations in order to better visualize the distributions.

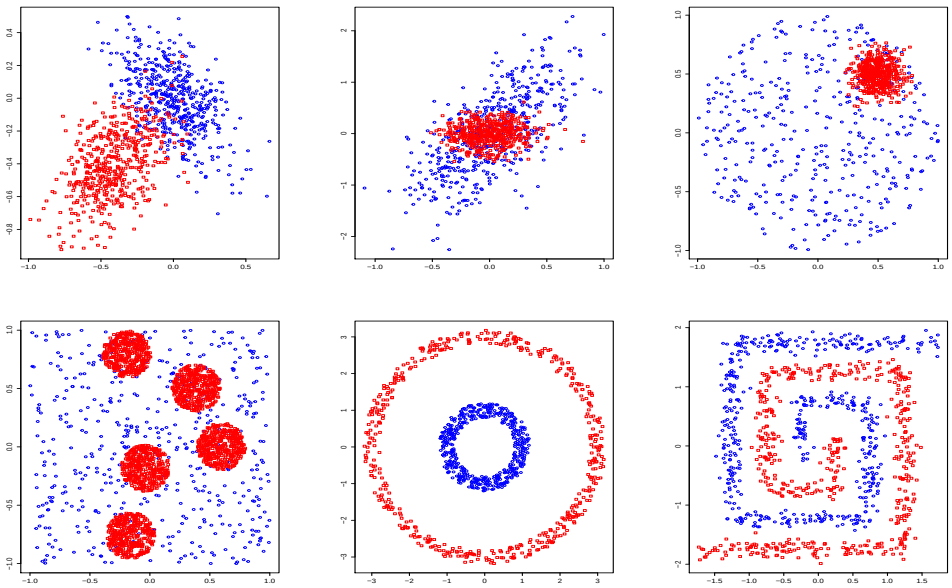


Figure IV.2.1 – Simulated data for classification. From left to right, from up to bottom: "gauss", "comete", "nuclear", "spot", "circles" and "spirals" examples.

Parametric and nonparametric methods are considered in the combination. For parametric methods, linear discriminant analysis (`lda`) and logistic regression (`logit`) are used. For nonparametric methods, k-nearest neighbors (`knn`) (with $k = 5$), support vector machines (`svm`), classification and regression trees (`cart`) and random forest (`rf`) are employed. In the *Cobra* aggregation procedure, the "unanimity" definition of the weights is used. For each simulated example, the performances of each estimator are computed using $N = 100$ repetitions. From the $n = 200$ observations generated, 75% randomly chosen observations are used as training set. The parameters α and β are selected via 5-fold cross-validation, using values on a grid. The 25% remaining observations are used to compute the test performances.

One first interesting question is whether a specific classifier always shows the best performance over all the repetitions. In an offline study, if a practitioner observes that a given classifier always gives the best performance, then aggregation is of course not necessary.

Table IV.2.1 shows, for each classifier, the number of runs where this classifier provides the best model. For each column, corresponding to a simulated example, the sum of the rows corresponding to the different methods equals the total number of repetitions ($N = 100$). We observe that, except for the "circles" example for which the knn classifier outperforms all the others, almost every classifier wins at least once over all the runs. Hence, for all simulated examples except the "circles" example, an aggregation procedure may be interesting.

	gauss	comete	nuclear	spot	circles	spirals
`lda`	75	9	35	4	0	2
`logit`	13	0	13	2	0	0
`knn`	5	46	27	44	100	76
`svm`	5	20	13	16	0	9
`cart`	0	18	8	19	0	2
`bag`	1	5	4	12	0	11
`rf`	1	2	0	3	0	0

Table IV.2.1 – Number of runs where every classifier provides the best model for the simulated classification examples (smallest test error).

Table IV.2.2 presents the average performances for all the classification examples. As expected for the "gauss" example corresponding to the rather well-separated mixture of two Gaussian distributions, both parametric and nonparametric classifiers perform well; otherwise, the nonparametric methods perform in general better.

The aggregation procedures succeed in yielding very satisfactory performances. When local behavior provides crucial information about the class label, as in the "circles" and "spirals" examples, `MixCobra` outperforms `Cobra`. This may be explained by the fact that the input part of `MixCobra` weights behaves like a nonparametric kernel-like method, performing well in such cases, as shown by the low error of the `knn` classifier.

IV.2.4.2 Regression

For the regression framework, we use the R package `Cobra` (Guedj, 2013). Recall that two parameters may drive the behavior of this aggregation strategy: ε sets a threshold for the distance between the prediction $f_m(x)$ given by an estimator f_m for a new observation x and the prediction $f_m(X_i)$ given for an observation X_i in

	gauss	comete	nuclear	spot	circles	spirals
`lda`	7.0	51.0	12.9	17.6	52.5	34.3
	(8.7)	(15.6)	(10.6)	(7.2)	(16.5)	(15.8)
`logit`	7.5	51.1	9.6	17.8	52.4	34.4
	(8.2)	(15.6)	(9.0)	(7.3)	(16.3)	(16.0)
`knn`	7.0	25.0	6.0	9.2	0.0	4.1
	(8.1)	(14.5)	(7.1)	(6.6)	(0.0)	(6.4)
`svm`	8.7	25.9	5.9	10.3	1.8	12.2
	(9.0)	(14.6)	(8.0)	(6.7)	(4.1)	(10.8)
`cart`	7.2	23.0	4.0	7.2	0.0	17.8
	(8.3)	(14.8)	(6.0)	(5.8)	(0.0)	(13.7)
`bag`	8.7	28.3	5.1	7.5	0.5	3.0
	(8.8)	(15.8)	(7.5)	(5.5)	(2.2)	(5.4)
`rf`	8.9	27.5	4.7	7.4	0.3	2.2
	(9.1)	(15.3)	(6.8)	(5.5)	(1.7)	(5.0)
`Cobra`	8.6	25.8	8.2	8.5	1.1	6.2
	(8.6)	(15.2)	(8.9)	(5.9)	(3.5)	(8.3)
`Mixcobra`	7.7	28.2	5.4	7.2	0.3	2.8
	(9.2)	(15.4)	(6.7)	(5.8)	(1.7)	(6.2)

Table IV.2.2 – Average classification error (and standard deviation into brackets) for the different classifiers and both aggregation methods `Cobra` and `MixCobra`, computed for the simulated classification examples (1 unit $= 10^{-2}$).

the training set, whereas γ is the proportion of estimators for which the consensus condition should be satisfied. In the sequel, we consider both the case $\gamma = 1$ (referred to as `CobraF`, for "fixed", in the sequel) and the case where γ is selected from the data (`CobraA` for "adaptive"). `Mixcobra` also involves two parameters, playing more symmetric roles in the weights: α and β control respectively the importance given to the inputs and to the combination of predictions. All parameters are chosen via cross-validation, as in the classification context.

The next paragraphs illustrate the benefit of combining several regression estimators in two industrial applications.

Modeling Air Compressors The goal is to model the power consumption of an air compressor, based on the flow, input pressure, output pressure, cooling water temperature and air temperature (Cadet et al., 2005). We consider $p = 8$ regression

estimators: linear regression model (`lm`), regression tree (`cart`), bagging regression trees (`bag`), random forest (`rf`), Support Vector Machines (`svm`) and k-nearest neighbors (`knn`) with different values of k ($k = 2, 5, 10$). The data set contains $n = 1000$ hourly observations of a working air compressor. Each variable is first transformed to have zero mean and unit variance.

The performances are computed using $N = 100$ replications, each consisting of randomly selecting $2/3$ of the observations for the training set, with the remaining $1/3$ used to estimate the performances of the different estimators. The performances on the test sets are presented in Table IV.2.3: for each case, the mean squared error is computed. We observed that the `bagging` and the `svm` estimators provide in average the smallest error. Moreover, we observe that the best estimator changes from run to run. For $N = 100$ replications, the best estimator is alternatively the `bagging` estimator (49 runs), the `svm` estimator (49 runs) or the `lm` estimator (2 runs). Figure IV.2.2 shows the boxplots computed with the different estimators and the 3 aggregation algorithms: `Cobra` with unanimity condition, `Cobra` with an adaptive number of estimators and `MixCobra`. The `Cobra` algorithm with all estimators yields in average the worst performance among the three aggregation techniques. Choosing adaptively the number of estimators allows to discard a possibly bad estimator, which improves the performance. We observe that `MixCobra` provides, in average, the best aggregation performance, associated with a low standard deviation.

	mean	sd
`lm`	11.64	0.00
`cart`	26.43	0.02
`bag`	10.75	0.01
`rf`	20.05	0.02
`svm`	10.77	0.01
`knn` 2	18.35	0.01
`knn` 5	18.79	0.01
`knn` 10	20.49	0.01
`CobraF`	21.63	0.06
`CobraA`	14.81	0.05
`MixCobra`	10.85	0.01

Table IV.2.3 – Average test performance and standard deviation for the different estimators and the aggregation methods, for the air compressor equipment (1 unit $= 10^{-2}$).

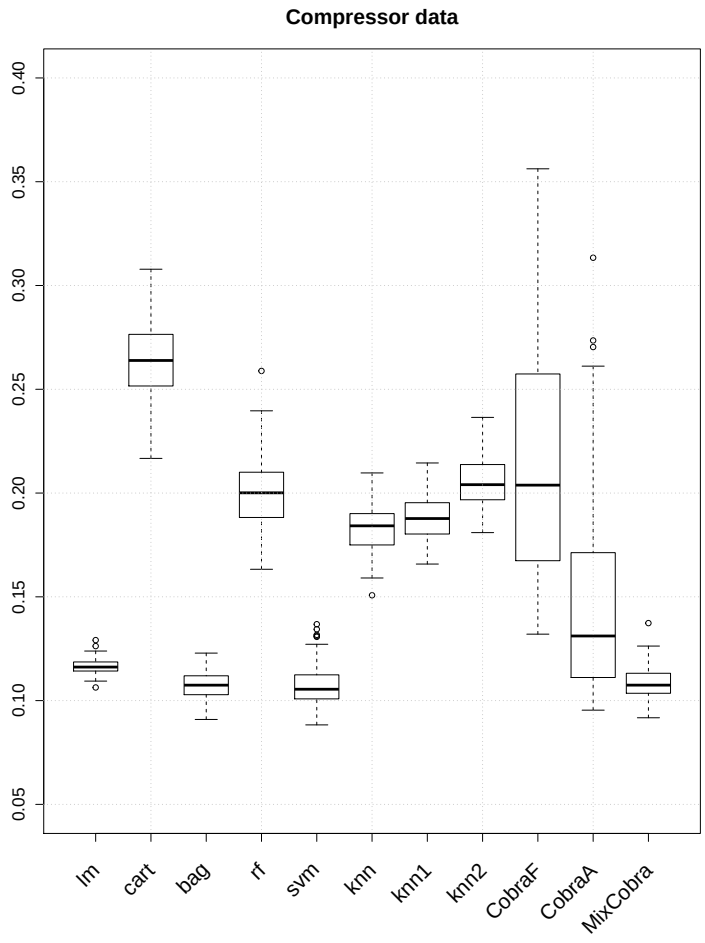


Figure IV.2.2 – Boxplots of the performances for the different estimators and the aggregation methods, for the air compressor equipment

Modeling Wind Turbines The second application aim to model 6 wind turbines on a wind plant in France. Each turbine is described by 5 variables, representing half-hourly information. The production of electricity is the target variable. The explanatory variables are the wind power, the wind direction (sine, cosine), and the temperature. For each wind turbine, each variable is first transformed to have zero mean and unit variance.

Table IV.2.4 presents the performances obtained in modeling the 6 wind turbines

(W_k, $1 \leq k \leq 6$), for $n = 1000$ operational data points, using the same set of $p = 8$ estimators as in the previous paragraph and the 3 aggregations methods. We observe that `MixCobra` shows very satisfactory results. Besides, for all wind turbines, the `bagging` estimator provides in average the best performances. However, Table IV.2.5 presents the number of runs over the $N = 100$ repetitions where every given estimator provides the smallest error. The model which provides the smallest error is never the same, and is alternatively either `bagging` or `svm`, with proportions differing for each wind turbine.

	W1	W2	W3	W4	W5	W6
`lm`	14.39	18.22	16.65	15.72	15.16	16.97
	(0.9)	(1.2)	(1.2)	(1.7)	(1.2)	(1.8)
`cart`	18.39	19.53	19.46	19.10	19.02	19.22
	(1.7)	(1.8)	(1.9)	(2.0)	(1.7)	(1.8)
`bag`	9.13	9.95	9.61	9.35	9.75	10.17
	(0.7)	(1.2)	(1.0)	(0.7)	(0.8)	(1.1)
`rf`	11.43	16.32	15.41	13.89	15.12	14.87
	(0.9)	(2.7)	(2.2)	(1.8)	(2.0)	(1.7)
`svm`	9.28	13.55	12.72	11.12	12.01	12.00
	(1.0)	(3.5)	(3.5)	(2.3)	(2.1)	(2.6)
`knn` 2	13.00	15.59	14.65	14.62	15.75	15.70
	(0.6)	(1.5)	(1.1)	(1.0)	(1.2)	(1.1)
`knn` 5	12.12	15.26	13.90	13.60	14.95	14.75
	(0.7)	(1.8)	(1.1)	(1.0)	(1.3)	(1.3)
`knn` 10	12.69	16.78	15.00	14.41	15.93	15.66
	(0.8)	(1.8)	(1.3)	(1.2)	(1.4)	(1.5)
`CobraF`	18.83	29.55	31.76	24.01	24.73	23.92
	(4.5)	(10.6)	(10.9)	(8.5)	(9.2)	(8.4)
`CobraA`	14.15	17.76	17.38	16.43	17.71	15.29
	(5.5)	(8.1)	(9.8)	(7.4)	(9.0)	(7.2)
`MixCobra`	9.02	10.47	10.04	9.45	10.37	10.57
	(0.6)	(1.4)	(1.0)	(0.8)	(0.9)	(1.1)

Table IV.2.4 – Average test performance (and standard deviation into brackets) for the different estimators and the aggregation methods, for the wind turbines (1 unit $= 10^{-2}$).

	W1	W2	W3	W4	W5	W6
`lm`	0	0	0	0	0	0
`cart`	0	0	0	0	0	0
`bag`	52	89	88	76	93	77
`rf`	0	0	0	0	0	0
`svm`	48	11	12	24	7	23
`knn` 2	0	0	0	0	0	0
`knn` 5	0	0	0	0	0	0
`knn` 10	0	0	0	0	0	0

Table IV.2.5 – Number of runs where each estimator provides the best model for the wind turbines (smallest test error).

Increasing the dimension or the number of estimators Since our combining method relies on a kernel rule, it is expected to be affected in some sense by the curse of dimensionality. However, an interesting feature of our approach is that the term based on the distance between the entries, $(X - x)/\alpha$, and the term involving the preliminary estimations, $(\mathbf{f}(X) - \mathbf{f}(x))/\beta$, are not affected the same way, since the "dimension" in not the same in both cases. Indeed, in the first case, the dimension is the actual dimension d of the space $\mathbb{R}^d$ containing the entries, whereas in the second case, the role of the dimension is played by the number p of individual estimators. In fact, the final combined estimator shows an interesting flexibility through the calibration of α and β. When the dimension d increases, the method will give more weight to the combining part, which is not affected in itself by increasing d and may only be affected through the particular behavior of the initial estimators considered. Conversely, for reasonable values of d, the effect of an increase of the number p of estimators may be balanced by the distance term between the entries.

To study the effect of dimensionality of the inputs, the original compressor observations are embedded in successive high-dimensional spaces of size $p = 11, 16, 21, 26$, by artificially adding to the 6 initial variables $5, 10, 15, 20$ independent random variables uniformly distributed.

Table IV.2.6 shows the performances computed using $N = 100$ replications as above, with again $2/3$ of the observations randomly chosen to calibrate the estimators and $1/3$ of the remaining observations used to estimate the performances of the different estimators. The impact of increasing the dimension differs depending on the estimator. We observe, as expected, that the performances of the `knn` estimators

strongly deteriorate as the dimension increases. On the contrary, the performances of the `lm` and `cart` estimators are stable. Table IV.2.6 shows that `MixCobra` keep performing quite well, even if the performances of some estimators decrease, and adapts to the increase of dimensionality.

	E1 $d = 6$	E2, $d = 11$	E3, $d = 16$	E4, $d = 21$	E5, $d = 26$
`lm`	11.6 (0.4)	11.7 (0.3)	11.8 (0.4)	11.9 (0.4)	11.8 (0.3)
`cart`	26.4 (1.9)	26.0 (2.0)	26.6 (1.9)	26.8 (2.2)	26.2 (1.8)
`bag`	10.7 (0.6)	11.4 (0.7)	12.2 (0.7)	12.3 (0.6)	12.6 (0.6)
`rf`	20.1 (1.6)	17.0 (1.3)	17.0 (1.5)	18.8 (1.4)	18.2 (1.7)
`svm`	10.8 (1.0)	14.8 (1.1)	16.4 (0.7)	16.6 (0.7)	16.2 (0.7)
`knn` 2	18.4 (1.2)	41.3 (2.1)	52.9 (2.3)	60.3 (2.8)	65.1 (3.4)
`knn` 5	18.8 (1.0)	36.5 (1.9)	45.5 (1.8)	51.9 (2.5)	56.6 (3.1)
`knn` 10	20.5 (1.1)	36.3 (2.1)	44.5 (1.9)	50.3 (2.1)	54.4 (3.6)
`MixCobra`	10.8 (0.7)	13.2 (0.9)	13.9 (0.8)	14.5 (0.9)	14.1 (0.4)

Table IV.2.6 – Average test performance (and standard deviation into brackets) for the different estimators and `Mixcobra`, for the initial air compressor data embedded into high-dimensional spaces, adding respectively 5, 10, 15, 20, 25 random variables.

Chapter IV.3

Clustering for predictive models

This chapter corresponds to a collaboration with Mathilde Mougeot and Sothea Has, for whom it is a part of his PhD thesis work, published in the Journal Statistical Computation and Simulation *(Fischer et al., 2021).*

IV.3.1 Introduction

IV.3.1.1 Context

The performance of a supervised learning model depends not only on the choice of the model itself but also on the quality of the data set used to learn a model. The frequent expression "garbage in, garbage out (GIGO)" highlights that nonsense or incomplete input data produces nonsense output as it is difficult to build an accurate model when some information is missing.

Yet, in practical problems, several fields useful for processing or understanding data may be missing for some reasons. For instance, in hiring processes, the use of information about individuals, such as gender, ethnicity, place of residence, is not allowed for ethic reasons, to avoid discrimination. Similarly, when high school students apply for further studies in higher education, not every information can be considered for selection. Besides, the General Data Protection Regulation (GDPR) text regulates data processing in the European Union since May 2018. It strengthens the French Data Protection Act, establishing rules on the collection and use of data on French territory (Tikkinen-Piri et al., 2018). As a result, contextual data that could

characterize individuals a little too precisely is often missing in available databases. Moreover, in an industrial context, not all recorded fields are made available for data processing for confidentiality reasons. For example, in the automotive industry, GPS data could be a valuable tool to provide services such as predictive vehicle maintenance. However, it is difficult to use such data as they are extremely sensitive. To sum up, in various areas, databases containing individual information have to respect anonymization rules before being analyzed.

Mining such databases can then be a particularly complex task as some critical fields are missing. In this context, the modalities of a missing qualitative variable correspond to several underlying groups of observations, which are a priori unknown but should be meaningful for designing a predictive model. In this case, an appropriate approach consists of using a two-step procedure: the clusters are computed in the first step and, in the second step, a predictive model is fit for each cluster. This two-step procedure has already been used, for instance, to approximate time evolution curves in the context of nuclear industry in Auder and Fischer (2012), to forecast electricity consumption using high-dimensional regression mixture models by Devijver et al. (2020), or to cluster multiblocks before PLS regression by Keita et al. (2015). The final performance of the model may strongly depend on the first step, since different configurations of clusters may lead to quite different global models. Finding an appropriate configuration of clusters is not an easy task which often requires a deep data investigation and/or human expertise.

IV.3.1.2 Presentation of the model

To build accurate predictive models in situations where the contextual data are missing, and to avoid an unfortunate choice of clusters, we propose to aggregate several instances of the two-step procedures where each instance corresponds to a particular clustering. Our strategy is characterized by three steps, each based on a fairly simple procedure. The first step aims at clustering the input data into several groups and is based on the well-known k-means algorithm. As the underlying group structures are unknown and may be complex, a given Bregman divergence is used as a distortion measure in the k-means algorithm. As already mentioned in Chapters III.1 and III.2, these proximity measures are an interesting clustering tool thanks to the correspondence with distributions of the exponential family. In the second step, for each divergence, a very simple predictive model is fit per cluster. The final step provides an adaptive global predictive model by aggregating the models corresponding to the different Bregman divergences, thanks to the consensus idea

presented in Chapters IV.1 and IV.2. We name this procedure the *KFC* procedure for k-means/Fit/Consensus.

In the sequel, we observe $(X_1, Y_1), \ldots, (X_n, Y_n)$, supposed to be realizations from a random pair (X, Y) with values in $\mathbb{R}^d \times \{0, 1\}$ (binary classification) or in $\mathbb{R}^d \times \mathbb{R}$ (regression). For the clustering step, we consider 4 Bregman divergences: Squared Euclidean distance (Euclid), I-divergence (I-div), Logistic (Logit) and Itakura-Saito (IS) divergences.

IV.3.2 Aggregation

IV.3.2.1 Description of the different aggregation methods

We enumerate in this section the different aggregation methods that will be used in our procedure, introduced by Mojirsheibani (1999, 2000) or discussed in Chapters IV.1 and IV.2. For classification and for regression, 3 procedures are considered each time : weights based on indicator functions, smoother kernel-based weights, weights involving inputs and outputs. Let $d_{\mathcal{H}}$ stand for the Hamming distance: $d_{\mathcal{H}}(u, v)$ measures the number of disagreements in the coordinates of the vectors u and v. For $x \in \mathbb{R}^d$, let $\mathbf{f}(x) = (f_1(x), \ldots, f_p(x))$ denote the vector of predictions for x given by p preliminary estimators. Let K denote a kernel, with $K_h(x) = K(x/h)$. Moreover, for $u \in \mathbb{R}^d$ and $v \in \mathbb{R}^p$, $K(u, v)$ stands for the kernel K applied to the concatenation of u and v.

For regression, we define, for $\ell \in \{1, 2, 3\}$:

$$Comb_\ell^R(x) = \frac{1}{n} \sum_{i=1}^{n} W_{n,i}^{(\ell)}(x) Y_i,$$

where

$$W_{n,i}^{(1)}(x) = \frac{\mathbf{1}_{\left\{\sum_{s=1}^{p} \mathbf{1}_{\{|f_s(X_i) - f_s(x)| < \varepsilon\}} \geq p\alpha\right\}}}{\sum_{j=1}^{n} \mathbf{1}_{\left\{\sum_{s=1}^{p} \mathbf{1}_{\{|f_s(X_j) - f_s(x)| < \varepsilon\}} \geq p\alpha\right\}}},$$

$$W_{n,i}^{(2)}(x) = \frac{K_h(\mathbf{f}(X_i) - \mathbf{f}(x))}{\sum_{j=1}^{n} K_h(\mathbf{f}(X_j) - \mathbf{f}(x))},$$

$$W_{n,i}^{(3)}(x) = \frac{K\left(\frac{X_i - x}{\alpha}, \frac{\mathbf{f}(X_i) - \mathbf{f}(x)}{\beta}\right)}{\sum_{j=1}^{n} K\left(\frac{X_j - x}{\alpha}, \frac{\mathbf{f}(X_j) - \mathbf{f}(x)}{\beta}\right)}.$$

The definition chosen for $W_{n,i}^{(1)}$ is the one for which unanimity of the initial estimators is not required.

For classification, we set:

$$Comb_1^C(x) = \begin{cases} 1 & \text{if } \sum_{i=1}^n \mathbf{1}_{\{\mathbf{f}(X_i)=\mathbf{f}(x)\}}\mathbf{1}_{\{Y_i=1\}} > \sum_{i=1}^n \mathbf{1}_{\{\mathbf{f}(X_i)=\mathbf{f}(x)\}}\mathbf{1}_{\{Y_i=0\}} \\ 0 & \text{otherwise,} \end{cases}$$

$$Comb_2^C(x) = \begin{cases} 1 & \text{if } \sum_{i=1}^n (2Y_i-1)K_h\left(d_{\mathcal{H}}(\mathbf{f}(X_i),\mathbf{f}(x))\right) > 0 \\ 0 & \text{otherwise,} \end{cases}$$

$$Comb_3^C(x) = \begin{cases} 1 & \text{if } \sum_{i=1}^n (2Y_i-1)K\left(\frac{X_i - X}{\alpha}, \frac{\mathbf{f}(X_i)-\mathbf{f}(x)}{\beta}\right) > 0 \\ 0 & \text{otherwise.} \end{cases}$$

IV.3.2.2 Considered kernels

Apart from the naive one, we will use the following kernels:

1. Gaussian kernel: for a given $\sigma > 0$ and for all $x \in \mathbb{R}^d$,

$$K(x) = e^{-\frac{|x|^2}{2\sigma^2}}.$$

2. Triangular kernel: for all $x \in \mathbb{R}^d$,

$$K(x) = (1 - |x|_1)\mathbf{1}_{\{|x|_1 \leq 1\}}.$$

   where $|.|_1$ is the ℓ_1-norm and is defined by: $|x|_1 = \sum_{i=1}^d |X_i|$

3. Epanechnikov kernel: for all $x \in \mathbb{R}^d$,

$$K(x) = (1 - |x|^2)\mathbf{1}_{\{|x| \leq 1\}}.$$

   where $|.|$ is the ℓ-norm and is defined by: $|x| = \left(\sum_{i=1}^d X_i^2\right)^{1/2}$

4. Bi-weight kernel: for all $x \in \mathbb{R}^d$,

$$K(x) = (1 - |x|^2)^2\mathbf{1}_{\{|x| \leq 1\}}.$$

5. Tri-weight kernel: for all $x \in \mathbb{R}^d$,

$$K(x) = (1 - |x|^2)^3 \mathbf{1}_{\{|x| \leq 1\}}.$$

The kernels are plotted in dimension 1 in Figure IV.3.1.

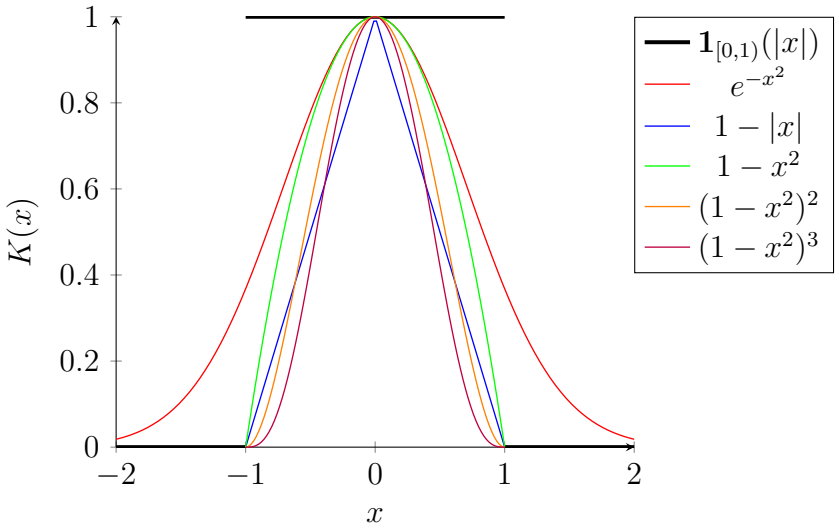


Figure IV.3.1 – Examples of kernels

IV.3.3 The KFC procedure

The three steps of the KFC strategy may be described as follows.

1. *k-means.* The input observations are first clustered using the k-means clustering algorithm with a Bregman divergence.
2. *Fit.* For every Bregman divergence, a dedicated predictive model is fit per cluster. Since the clustering step should help building an accurate model, we deliberately choose simple models to fit on a group, namely linear regression for regression models, logistic regression for classification.
3. *Consensus.* As neither the distribution nor the clustering structure of the input data is known, it is not clear in advance which divergence will be the most efficient. Thus, we propose to combine all the estimators corresponding to the different divergences, in order to take the best advantage of the clustering step.

For the combination task, we try the different procedures described above. Practically, the different parameters in the combining methods are optimized on a grid, using cross-validation.

Once the candidate model, which is the collection of all the local models constructed on the corresponding clusters, is fitted, in order to make a prediction for a new observation x, we first affect x to the closest cluster for each divergence, which yields one prediction per divergence, and then, perform aggregation.

IV.3.4 Simulated data

In this section, we analyze the behavior of the strategy on several simulated data sets in classification or regression problems.

IV.3.4.1 Description

In both classification and regression cases, we simulate 5 different kinds of input data sets. We consider 2-dimensional data sets where the two predictors (X_1, X_2) are simulated according to Exponential (Expo hereafter), Poisson (Poiss), Geometric (Geom) and Gaussian (Gauss) distribution respectively. The remaining data set is 3-dimensional, with predictors (X_1, X_2, X_3), distributed according to a Gaussian distribution. We use 1500 data points for training and 450 for testing. Each time, there are $k = 3$ balanced clusters of 500 observations for training and 150 for testing. The different distribution parameters used in the simulations are listed in Table IV.3.1.

Distribution	Parameter	Cluster 1	Cluster 2	Cluster 3
Expo	λ	$0.05; 0.5$	$0.5; 0.05$	$0.1; 0.1$
Poiss	λ	$3; 11$	$10; 2$	$13; 12$
Geom	p	$0.07; 0.35$	$0.55; 0.07$	$0.15; 0.15$
2D Gauss	μ	$4; 12$	$22; 9$	$10; 5$
	σ	$1; 1$	$2; 1$	$2; 2$
3D Gauss	μ	$6; 14; 6$	$5; 10; 15$	$8; 6; 14$
	σ	$1; 2; 1$	$2; 1; 2$	$1; 1; 2$

Table IV.3.1 – Parameters of the simulated data

The regression model in cluster k takes the form

$$Y^{(k)} = \beta_0^k + \sum_{j=1}^{d-1} \beta_j^k X_j^{(k)} + \varepsilon.$$

In classification, for cluster k, we set $Y^{(k)} = 0$ if

$$\frac{1 - e^{\beta_0^k + \sum_{j=1}^{d-1} \beta_j^k X_j^{(k)} + \varepsilon}}{1 + e^{\beta_0^k + \sum_{j=1}^{d-1} \beta_j^k X_j^{(k)} + \varepsilon}} \leq 0.$$

In both situations, $\varepsilon \sim \mathcal{N}(0, 10)$. If necessary, we discard negative values or apply a normalization. In the sequel, the performance of a model is assessed over 20 replications.

IV.3.4.2 Numerical results

This section analyzes the performance of the KFC procedure for classification or regression on the simulated examples described in Section IV.3.4. Each example is simulated 20 times. The measures of performance are the missclassification error for classification and the RMSE (Root Mean Square Error) for regression. We present the average error over the 20 runs, with the standard deviation into brackets. We compare the results (column block "kernels" in the tables) with those without aggregation, obtained without any preliminary clustering (column "Single"), or when clustering with a given Bregman divergence (column block "Bregman divergence"). Epan, Gauss, Triang, Bi-wgt and Tri-wgt stand for Epanechnikov, Gaussian, Triangular, Bi-weight and Tri-weight kernels respectively. For each example, the first row corresponds to $Comb_1, Comb_2$, and the second to $Comb_3$.

Table IV.3.2 below contains the results in the classification context. Of course, all models built after a clustering step outperform the model without clustering. The combined classification methods perform generally better than or similarly to the best individual estimator. The results of $Comb_3^C$, in the second row, seem to be better compared to the ones of $Comb_1^C$ and $Comb_2^C$, in the first row. We also note that the Gaussian kernel seems to be a good choice.

In the regression case, the RMSE results are given in Table IV.3.3. We observe that, for geometric distribution, the estimator based on Generalized Kullback-Leibler Divergence outperforms the estimator built after clustering with Logistic divergence. Again, the performance of the estimators is globally improved by combining. It is clear that Gaussian kernel does the best job.

Distr.	**Single**	**Bregman divergence**				**Kernel**					
		Euclid	GKL	Logit	IS	Naive	Epan	Gauss	Triang	Bi-wgt	Tri-wgt
Expo	18.86 (1.70)	8.58 (1.77)	7.42 (1.55)	4.09 (1.08)	**3.92** (1.15)	3.49 (0.89)	3.51 (0.94)	**3.46** (0.88)	3.51 (0.94)	3.56 (0.91)	3.56 (0.91)
						2.91 (0.81)	2.63 (0.70)	2.49 (0.74)	2.70 (0.75)	2.56 (0.63)	**2.46** (0.66)
Poiss	46.93 (3.35)	9.19 (1.27)	**8.45** (1.24)	13.33 (1.84)	10.15 (1.47)	8.59 (1.37)	**8.51** (1.46)	**8.51** (1.47)	**8.51** (1.46)	8.52 (1.47)	8.52 (1.49)
						8.51 (1.28)	8.46 (1.11)	8.44 (1.17)	**8.42** (1.15)	8.57 (1.28)	8.44 (1.13)
Geom	19.90 (2.07)	12.57 (2.39)	4.71 (2.37)	**3.94** (1.15)	8.12 (1.57)	3.61 (1.15)	**3.60** (1.16)	**3.60** (1.16)	3.61 (1.15)	**3.60** (1.16)	**3.60** (1.16)
						3.76 (0.92)	3.52 (1.11)	**2.94** (0.93)	3.48 (1.09)	3.47 (1.11)	3.40 (1.06)
2D Gauss	49.00 (2.52)	**12.37** (1.55)	12.40 (1.50)	14.14 (1.44)	13.05 (1.61)	12.87 (1.60)	12.82 (1.59)	**12.80** (1.56)	12.84 (1.57)	12.84 (1.57)	12.87 (1.60)
						12.02 (1.30)	12.11 (1.24)	12.06 (1.35)	12.11 (1.27)	12.09 (1.23)	12.10 (1.22)
3D Gauss	43.39 (2.52)	**10.77** (1.40)	10.99 (1.44)	11.74 (1.45)	11.56 (1.51)	11.08 (1.58)	11.01 (1.52)	**11.00** (1.50)	**11.00** (1.50)	11.04 (1.57)	11.03 (1.55)
						10.23 (1.40)	9.93 (1.47)	**9.76** (1.53)	10.04 (1.47)	9.83 (1.61)	9.84 (1.61)

Table IV.3.2 – Misclassification errors computed over 20 runs (1 unit = 10^{-2})

Distr.	**Single**	**Bregman divergence**				**Kernel**					
		Euclid	GKL	Logit	IS	Naive	Epan	Gauss	Triang	Bi-wgt	Tri-wgt
Expo	106.58 (7.13)	68.74 (6.84)	57.06 (7.37)	44.54 (7.37)	**44.46** (10.96)	55.11 (15.85)	51.14 (13.31)	**40.21** (14.40)	52.99 (13.12)	50.24 (13.74)	50.64 (14.41)
						56.34 (17.48)	52.62 (17.82)	**39.12** (14.98)	51.31 (19.55)	51.20 (19.69)	51.98 (20.12)
Poiss	26.76 (1.11)	10.16 (1.91)	**8.22** (2.25)	16.72 (1.61)	12.15 (1.86)	8.88 (1.65)	9.18 (1.98)	**8.43** (2.18)	8.85 (2.06)	8.84 (2.03)	8.76 (2.03)
						9.73 (2.25)	9.61 (1.86)	**9.13** (1.92)	9.64 (1.91)	9.40 (1.86)	9.43 (1.93)
Geom	70.45 (4.52)	29.99 (5.95)	**18.33** (7.34)	22.94 (6.21)	31.94 (5.19)	36.39 (13.81)	32.49 (13.49)	**21.51** (11.79)	31.48 (14.31)	31.44 (13.51)	30.89 (12.21)
						31.83 (12.88)	27.90 (14.20)	**17.82** (12.58)	26.82 (13.28)	28.45 (14.02)	24.58 (13.21)
2D Gauss	21.98 (1.20)	**5.63** (1.26)	6.46 (1.81)	19.36 (1.11)	9.38 (1.86)	7.09 (2.55)	6.57 (1.78)	**5.57** (0.49)	6.20 (1.72)	6.41 (1.76)	6.33 (1.75)
						9.75 (1.30)	7.70 (2.24)	**6.42** (1.49)	7.45 (2.42)	7.47 (2.28)	7.34 (2.31)
3D Gauss	53.55 (1.74)	**19.89** (3.49)	20.93 (2.97)	23.71 (2.70)	22.96 (2.74)	18.16 3.42)	18.20 (3.45)	**16.94** (4.06)	18.25 (3.41)	18.05 (3.50)	18.00 (3.49)
						19.24 (3.54)	18.52 (4.02)	**17.51** (3.64)	18.64 (4.37)	18.19 (3.91)	18.42 (3.68)

Table IV.3.3 – RMSE computed over 20 runs

IV.3.5 Real data

In this section, we study the performance of the KFC procedure on real data and study the robustness of the procedure with respect to the number k of clusters. The goal is to model the electrical power consumption of an air compressor equipment, using data from the same source as in Chapter IV.2 (Cadet et al., 2005). The target is the electrical power of the machine, and 6 explanatory variables are available: air temperature, input pressure, output pressure, flow, water temperature. We use $N = 2000$ hourly observations of a working air compressor, and consider 20 random partitions of the sample in training (80%) and testing sets. As the number of clusters is unknown, we perform the KFC algorithm with different values of the number of clusters $k \in \{1, 2, ..., 8\}$. For the aggregation step, we use a Gaussian kernel. The average RMSE and standard deviations are summarized in Table IV.3.4. Note that for a simple linear model ($k = 1$), the average RMSE is 178.67 with standard error deviation 5.47. We observe that the performance of the individual estimators improves as the number k of clusters increases. Note that $Comb_3^R$ outperforms $Comb_2^R$. Regardless of the number of clusters, the combination step allows to reduce the RMSE in each case to approximately the same level. Hence, our strategy may be interesting even without the knowledge of the number of clusters.

k	Euclid	GKL	Logistic	IS	$Comb_2^R$	$Comb_3^R$
2	158.85	158.90	159.35	158.96	153.34	**116.69**
	(6.42)	(6.48)	(6.71)	(6.41)	(6.72)	(5.86)
3	157.38	157.24	156.99	157.24	153.69	**117.45**
	(6.95)	(6.84)	(6.65)	(6.85)	(6.64)	(5.55)
4	154.33	153.96	153.99	154.07	152.09	**117.16**
	(6.69)	(6.74)	(6.45)	(7.01)	(6.58)	(5.99)
5	153.18	153.19	152.95	152.25	151.05	**117.55**
	(6.91)	(6.77)	(6.57)	(6.70)	(6.76)	(5.90)
6	151.16	151.67	151.89	151.75	150.27	**117.74**
	(6.91)	(6.96)	(6.62)	(6.57)	(6.82)	(5.86)
7	151.08	150.99	152.81	151.85	150.46	**117.58**
	(6.77)	(6.84)	(7.11)	(6.61)	(6.87)	(6.15)
8	151.27	151.09	152.07	150.90	150.21	**117.91**
	(7.17)	(7.01)	(6.65)	(6.96)	(7.03)	(5.83)

Table IV.3.4 – Average RMSE for the air compressor data

Part V

Interdisciplinary collaborations

Chapter V.1

Modeling wind energy production

This chapter corresponds to a numerical study Fischer et al. (2017) carried out in collaboration with Lucie Montuelle, Mathilde Mougeot and Dominique Picard, in the frame of the ANR project FOREWER (Modeling, forecasting and risk evaluation of wind energy production), as part of the postdoc subject of Lucie Montuelle (now Data Scientist at the French electricity transmission system operator RTE).

The numerical results presented in this chapter are the ones published in the article. Nevertheless, further experiments conducted in another context tend to show that different results might be obtained by optimizing the methods a little differently, so that the order that appears here between the various tested procedures should not be considered as absolute.

V.1.1 Introduction

In the first quarter of 2020, wind energy represents 10,8% of the French electricity production. Since electricity can hardly be stored, forecasting tools are essential to appropriately balance the production of the different renewable energies. The purpose of this chapter is to quantify the modeling performances of wind production at a farm scale, using real operational data provided by the company Maïa Eolis (today Engie Green).

A possible approach relies on physical models of a wind turbine related to analytical equations. Here, our contribution pertains to the more flexible and robust to noise bunch of strategies base on statistical or data mining methods. The aim is

to model the power production by learning the phenomenon directly on the data. This point of view has been investigated in the literature for instance Kramer et al. (2013); Sideratos and Hatziargyriou (2012) and in Quan et al. (2014), with neural networks. A special network, called extreme learning machine, has been used in Wan et al. (2014) for probabilistic interval forecasting. The k-nearest neighbor method has been studied for wind power modeling by Kusiak et al. (2009). In Mangalova and Shesterneva (2016), the k-nearest neighbor algorithm is used for probabilistic forecasts in the frame of the Global Energy Forecasting Competition 2014.

Support vector machines for regression have been proposed in this context in Kramer and Gieseke (2011), whereas Kusiak et al. (2009) provides a comparison between several data-mining approaches. Besides, time series-based models have also contributed to the field of wind power forecast (see, e.g., Milligan et al. (2004); Wu et al. (2014)). For an overview of different modeling and forecasting methods for wind power, the reader may further refer to the surveys Costa et al. (2008); Foley et al. (2012); Giebel et al. (2011); Jung and Broadwater (2014).

We investigate and compare different techniques for modeling the electrical power for 3 wind farms in France. Each time, we first model the electrical power of each wind turbine of the farm using local inputs coming from sensors directly installed on each wind turbine. The predictive power of the farm is then given by the sum of the predictive powers computed for each wind turbine. In a second step, we quantify the modeling performances by using more global inputs as may be provided by a meteorologist forecaster as, for example, Météo France. The goal is to quantify the performance of the different models running in an operational environment, using only average input information at a farm scale.

The chapter is organized as follows. In Section V.1.2, we describe the data set. Section V.1.3 introduces the different methods investigated. Section V.1.4 presents and discusses the modeling performances obtained using the local information on each turbine. The results found when replacing this information by the more global one, relying on averages, are given in Section V.1.5.

V.1.2 Data set

As already mentioned above, the data set has been provided by Maïa Eolis (today Engie Green). It comes from 3 different farms with 4 to 6 turbines, in the North and East of France, from 2011 to 2014. In a farm, each wind turbine provides 10 minute measurements of electrical power, wind speed, wind direction, temperature, as well

as an indicator of the working state of the turbine. The electrical power output of the whole farm is also provided on a 10 minute basis.

To detect freeze, wind speed is measured on each turbine both by a classical anemometer and a heated one. Since more measures are available from the heated anemometer, the study has been conducted with this data. Wind direction is provided by a weather vane and has been recoded into two variables corresponding to the cosine and the sine of the angle. The state of the turbine may correspond to start, stop or full working of the turbine, depending on the wind speed and maintenance operations. For the sake of simplicity, this study focuses on fully operating times. Besides, the data has been averaged over 30 minutes in order to slightly smooth the signals. However, it should be stressed that most often the results obtained on a 10 minute basis are quite similar to those presented in the sequel.

Taking advantage of the 30 minutes averages, two additional variables have been introduced: the variance of the wind speed, and the variance of the wind direction over 30 minutes. The second variable (complex-valued), has been decomposed into its real part and its imaginary part, leading to a total of 7 explanatory variables.

V.1.3 Predictive methods

In this section, our aim is to model the farm power using the 7 explanatory variables. More precisely, the variables are observed at time t and the sum of the power of each turbine of the farm at time t is predicted. Then, the estimated farm power is computed by summing the estimated turbine powers. The error is calculated at the farm scale.

Our intention is to compare parametric statistical methods inspired from the related physical equation, to more agnostic nonparametric machine learning methods, which can easily accommodate non-linear modeling as well as dependence between variables, which is the case here.

Theoretical equation

According to theoretical studies on wind turbines (see, e.g., Lydia et al., 2014), the delivered power obeys the following equation :

$$P(W) = \frac{1}{2}\rho S c_p W^3, \tag{V.1.1}$$

where W is the wind speed, ρ the air density, S the rotor surface, which is the area swept by the blades, and c_p the power coefficient, corresponding to the fraction of wind energy that the wind turbine is able to extract.

Figure V.1.1 shows for a wind turbine the electrical power versus wind-speed as well as versus the cube of this quantity, with two theoretical curves corresponding to two different values of c_p: the maximal theoretical value (16/27, red curve), and a more realistic value given in Table 8 of Carrillo et al. (2013) (blue curve). The third curve (in green) is provided by the turbine builder, based on his experiments. Notice that the data cloud is quite dispersed. Although the theoretical curves correspond to some trend, there is room for improvement to provide a better prediction.

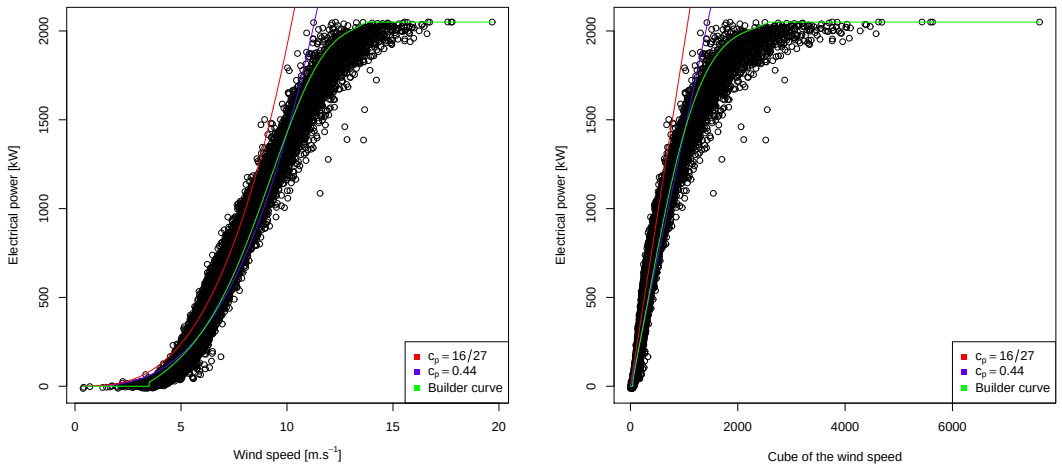


Figure V.1.1 – Empirical observations for a wind turbine and theoretical power curves for different power coefficient values, compared to the turbine builder curve

Prediction by "persistence"

The so-called "persistence" predictor uses the last observation as prediction: if Y_t denotes the electric production at time t, the predicted production at time t is given by $\hat{Y}_t = Y_{t-1}$. This very naive method will serve as a benchmark to asses the added value of statistical predictions.

Parametric methods

Several methods have been tested to approximate the power curve and model the production. In this section, we present the parametric statistical methods, directly inspired from the physical equation.

Parametric modeling according to the wind speed only We first investigated the simplest parametric models, namely linear regression and logistic regression, with the wind speed as unique explanatory variable. If the predicted power at time t is denoted by $\hat{Y}_t$, the linear model is given by

$$\hat{Y}_t = \hat{a}_0 + \hat{a}_1 W_t,$$

where W_t denotes the wind speed at time t, and $\hat{a}_0$ and $\hat{a}_1$ are computed using ordinary least squares (OLS). The logistic regression model may be written

$$\hat{Y}_t = \frac{\hat{C}}{1 + \exp(\hat{a}_0 + \hat{a}_1 W_t)},$$

where the parameters $\hat{a}_0$, $\hat{a}_1$, $\hat{C}$ are obtained using maximum likelihood.

Introducing a third degree polynomial of the wind speed in the logistic regression has also been considered to mimic more closely Equation (V.1.1). More precisely, the model is then defined by:

$$\hat{Y}_t = \frac{\hat{C}}{1 + \exp(\hat{a}_0 + \hat{a}_1 W_t + \hat{a}_2 W_t^2 + \hat{a}_3 W_t^3)},$$

where $\hat{a}_i$, $i = 0, \ldots, 3$ and $\hat{C}$ are estimated parameters. This model is called polynomial logistic regression in the sequel.

Parametric modeling using more variables Linear regression, logistic regression and polynomial logistic regression have also been studied with the full set of variables, using not only wind speed as a predictor, but also wind direction (coded by its cosine and sine : $D^{\cos}$ and $D^{\sin}$), temperature T and the variances of the wind speed W^S and direction, $D^{S,\mathfrak{Re}}$ and $D^{S,\mathfrak{Im}}$:

$$\hat{Y}_t = \hat{a}_0 + \hat{a}_1 W_t + \hat{a}_2 D_t^{\cos} + \hat{a}_3 D_t^{\sin} + \hat{a}_4 T_t + \hat{a}_5 W_t^S + \hat{a}_6 D_t^{S,\mathfrak{Re}} + \hat{a}_7 D_t^{S,\mathfrak{Im}} \quad \text{(V.1.2)}$$

$$\hat{Y}_t = \hat{C}\left[1 + \exp\left(\hat{a}_0 + \hat{a}_1 W_t + \hat{a}_2 D_t^{\cos} + \hat{a}_3 D_t^{\sin} + \hat{a}_4 T_t + \hat{a}_5 W_t^S + \hat{a}_6 D_t^{S,\mathfrak{Re}} + \hat{a}_7 D_t^{S,\mathfrak{Im}}\right)\right]^{-1}$$

$$\hat{Y}_t = \hat{C}\left[1 + \exp\left(\hat{a}_0 + \sum_{k=1}^{3} \hat{a}_{1,k} (W_t)^k + \hat{a}_2 D_t^{\cos} + \hat{a}_3 D_t^{\sin} + \hat{a}_4 T_t + \hat{a}_5 W_t^S + \hat{a}_6 D_t^{S,\mathfrak{Re}} + \hat{a}_7 D_t^{S,\mathfrak{Im}}\right)\right]^{-1} \quad \text{(V.1.3)}$$

In the last equation (V.1.3), corresponding to polynomial logistic regression, only the wind speed W_t occurs in the expression as a polynomial of order 3, to be consistent with the theoretical equation (V.1.1). A Lasso penalized version of equation (V.1.2) has been investigated as well (Tibshirani, 1994). More specifically, the coefficients $\hat{a}_1, \ldots, \hat{a}_7$ are estimated with OLS, under the ℓ^1 constraint $\sum_{j=1}^{7} |a_j| \leq \kappa$ for some constant $\kappa > 0$.

Non-parametric methods

Non-parametric methods may be very useful to model complex and non-linear phenomena. In general, they do not lead to closed-form expressions as in the previous section.

kNN The k-nearest-neighbor procedure consists in computing

$$\hat{Y}_t = \frac{1}{k} \sum_{j=1}^{k} Y_{(j)},$$

where $Y_{(j)}$ corresponds to the wind power of the j-th nearest neighbor of the observation at time t, according to the Euclidean distance of the variables at time t. The features are standardize to have mean 0 and variance 1 since they are measured in different units. The number k of neighbors is optimized on a grid.

CART, bagging and RF Tree-based methods are also applied. Growing a binary tree based on the CART algorithm (Breiman et al. (1984)) consists in defining rectangular regions by recursive splitting, the last set of domains corresponding to the

leafs of the tree. The prediction is provided by the average value in the leaf in which the observation falls. To avoid over-fitting, the tree is pruned by cross-validation. To reduce variance, we also use bagging (Breiman (1996)), which consists in averaging the predictions of several trees grown on bootstrap samples: for B bootstrap samples, the predicted value is given by

$$\hat{Y}_t = \sum_{b=1}^{B} \hat{Y}_t^b, \tag{V.1.4}$$

where $\hat{Y}_t^b$ denotes the prediction for the b-th bootstrap sample. To produce more diversity in the trees to be averaged, random forests only consider, for each split, a smaller number of randomly chosen variables. The trees are not pruned (Breiman, 2001).

SVM Through a nonlinear kernel function, the support vector machine method for regression maps the inputs into a higher dimensional feature space, where an optimal hyperplane is constructed (see for instance Drucker et al., 1997). The regularization parameter of the method is calibrated on a grid.

In the next section, all the experiments have been conducted using the R software. The previous procedures are implemented respectively in the packages `lars`, `kernlab`, `FNN`, `rpart` and `randomForest`. For random forests, the default parameters, advocated by Breiman, were used: 500 trees were grown in each forest and the size of the subset of randomly chosen variables, commonly denoted by *mtry*, is the floor of the third of the number of variables. Note that bagging is a particular case of random forests where *mtry* equals the total number of variables.

V.1.4 Modeling performance

As usual, the data set is split into a training and a test set. In order to quantify the variability of the predictive ability, several test sets are used: an average performance, as well as a standard deviation, are then computed. More precisely, the procedures are trained on around 8000 instant-points and 10 data sets of 724 points are used to evaluate the performances. The error criterion is the Root Mean Squared Error (RMSE), defined between a vector of predictions $\hat{Y}$ and a vector of observed wind

power productions Y by

$$RMSE(\hat{Y}) = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(\hat{Y}_t - Y_t)^2}.$$

Another quantity of interest for industries is the error as percentage of the installed power (denoted by % of IP in the tables below), also called "normalized RMSE" in this context: the RMSE is divided by the theoretical power of the farm.

As already mentioned, procedures have been investigated either using the wind speed variable only or using all available variables. Figure V.1.2 and Table V.1.1 present the results obtained for one wind farm located in the East of France. Observe that all the methods show a much better performance than the naive persistence method, substantially reducing the average error, with generally a lower standard deviation.

The methods using only the wind speed as predictor show pretty good performances. The error is reduced by more than a factor of 2 compared to persistence. The polynomial logistic regression shows a very good performance, in accordance with the physical equations, with, however, somewhat high variability. The SVM and bagging methods show the best results with also better stability. Some of the parametric methods behave better when adding more variables, namely the wind direction, the variances of the wind speed and direction, and the temperature, without, however, equaling the performance of polynomial logistic regression. The results with Lasso are the same as with classical linear regression, all variables being kept. This may be due to strong correlation between predictors. The CART algorithm does not take advantage of the additional variables and seems to choose its cuts only according to the wind speed, which may be explained by the prevailing importance of the wind speed over other measures. The best kNN method and the best SVM procedure, tested with several kernels, are not very competitive here. Bagging outperforms all the investigated statistical models, including random forests with default settings. This is due to the importance of the wind speed relative to the other variables. Comparing CART and bagging highlights the advantages of bootstrapping and averaging. This step allows to reduce the error by a third, when dealing with all the predictors. Note that, according to the Panorama de l'électricité renouvelable, French industries obtain in 2015 a root mean squared error of 2.4% of the installed power of farm productions, which illustrates the benefits of using bagging (1.65%).

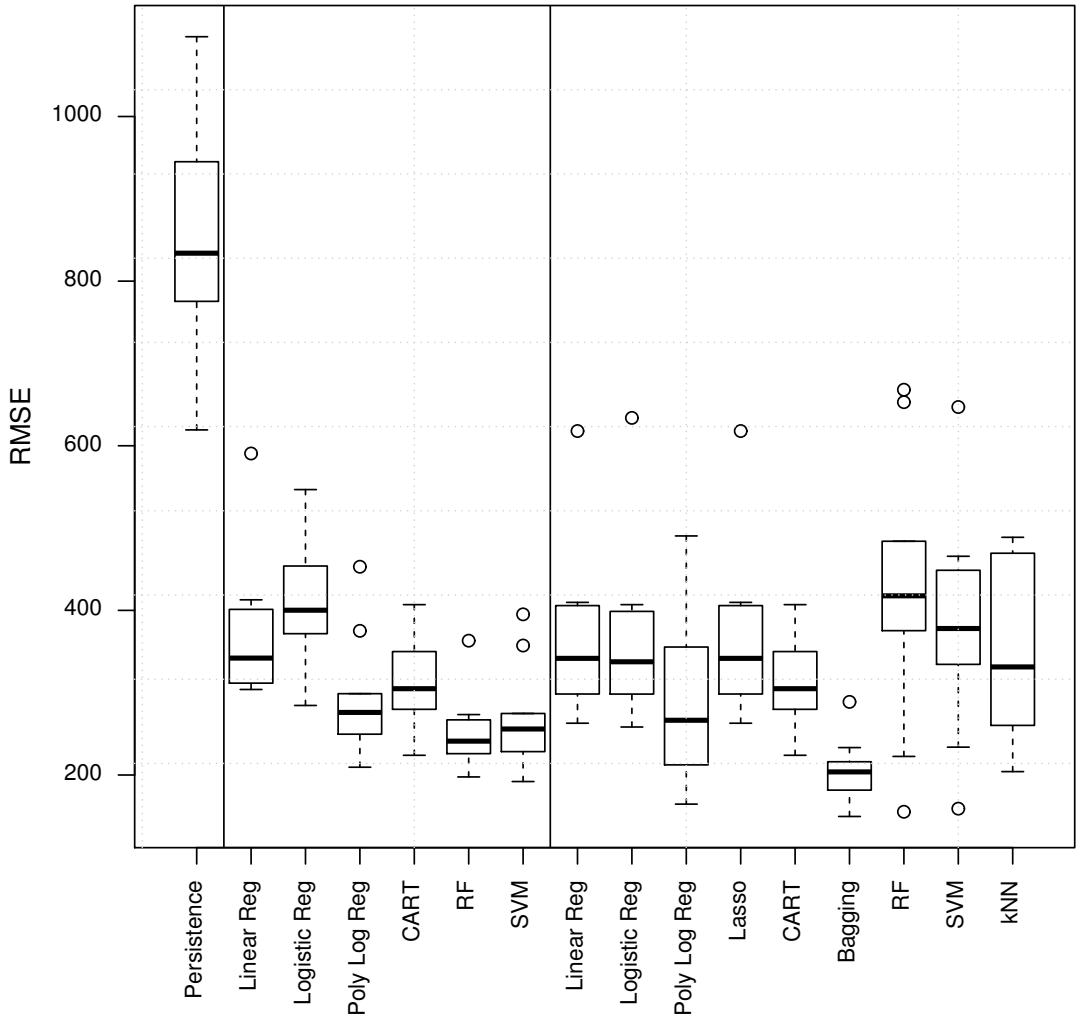


Figure V.1.2 – Boxplots of the RMSE

	Method	RMSE	% of IP
	Persistence	855.52 (141.14)	6.96
using wind speed only	Linear Regression	373.61 (86.91)	3.04
	Logistic Regression	404.86 (76.74)	3.29
	Polynomial Log. Reg.	290.36 (73.87)	2.36
	CART	314.46 (57.74)	2.56
	Bagging (=RF)	**250.52 (46.52)**	**2.04**
	SVM	269.94 (64.21)	2.19
using all variables	Linear Regression	364.21 (102.39)	2.96
	Logistic Regression	362.76 (107.58)	2.95
	Polynomial Log. Reg.	292.57 (100.53)	2.38
	Lasso	364.21 (102.39)	2.96
	CART	314.46 (57.74)	2.56
	Bagging	**203.50 (39.72)**	**1.65**
	RF	425.78 (161.53)	3.46
	SVM	382.16 (134.34)	3.11
	kNN (k=2)	355.29 (109.96)	2.89

Table V.1.1 – Average RMSE with standard deviation into brackets and % of installed power

Comparison of different farms For the two other farms, located in the North of France, the hierarchy between the procedures is quite similar. The procedure ranking first most often is bagging.

Next, we compared the farms using a common test set, with observed variables available at the same time for each farm, with at least one turbine fully operational. The test set has been divided into ten subsets of 1440 instant-points, each covering a period of around thirty days, to quantify the average performance and its variability. The training set consists in around 7200 instant-points, satisfying a ratio of 83% of the data dedicated to learning and 17% used for test.

Only the best procedure, bagging, has been applied. We also compare the results with those derived from the turbine builder's power curve, represented by the green curve in Figure V.1.1. Figure V.1.3 highlights the good results of bagging on the first and the third farms. It performs reasonably well on the second farm, but is not as good as the power curve's builder. It may be explained by the difference between the wind speed in the training sample and in the test set. Few high wind speed levels are observed in the training sample on the second farm compared to the test sample, so the bagging prediction may not be accurate.

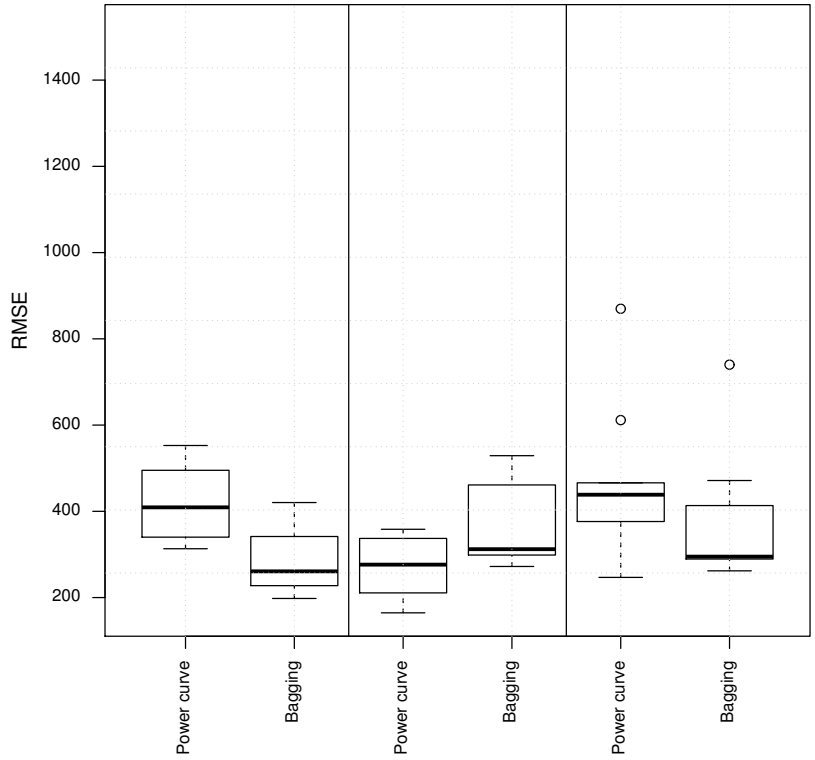


Figure V.1.3 – Comparison of the RMSE for the turbine builder's power curve and the bagging procedure on several farms using local measures

V.1.5 Towards forecast : a stability investigation

Forecasting electrical power requires two steps: providing forecasts of the explanatory variables, and constructing an accurate model to plug these forecasts, which is our aim here. If we have at hand an efficient model, then the performance of a forecasting procedure of electrical wind power will directly depend on wind forecasts. These meteorological forecasts may be obtained thanks to ensemble methods based on numerical computations, mostly based on the Navier-Stokes equations, like the climate reanalyzes performed by the European Centre for Medium-Range Weather Forecasts (ECMWF) or the French Weather Agency Météo France.

At a farm scale, as already mentioned, many observations are recorded in real time on each wind turbine. Thanks to the analysis conducted in the previous section, we are able to identify an accurate model, built with this kind of observations. In a wind farm, in general, two wind turbines are at a distance of about 300m from each other. However, concerning numerical models, the finest grid resolution for forecast of wind and temperature provided by the French Weather Agency Météo France is for instance brought by the AROME model, which proposes a resolution of about 1.5 km (5 times larger). Consequently, an interesting question is also to quantify the predictive power not using very local information, but information on a much broader scale. To mimic the scale of meteorological wind forecasts, we decided to introduce virtual sensors: for each variable, a global information is computed by averaging all the localized variables coming from the set of turbines installed on the wind farm. This makes it possible to quantify the loss of accuracy due to the replacement of localized data by global information and is a first step towards forecasting.

The results are available in Table V.1.2. The deterioration of the prediction can easily be seen in Figure V.1.4. We observe that polynomial logistic regression is remarkably robust, performing similarly to the context with local measures, contrary to SVM and kNN. When only wind speed is considered, polynomial logistic regression competes with bagging, whereas the latter outperforms all the considered procedures when dealing with all the variables.

	Method	Mean of RMSE	Sd of RMSE	% of IP
	Persistence	855.52	141.14	6.96
using wind speed only	Linear Regression	393.09	77.25	3.20
	Logistic Regression	541.37	103.15	4.40
	Polynomial Log. Reg.	**288.28**	**75.23**	**2.34**
	CART	349.17	53.20	2.84
	Bagging (=RF)	293.26	48.96	2.38
using all variables	Linear Regression	387.71	89.73	3.15
	Logistic Regression	524.30	92.58	4.26
	Polynomial Log. Reg.	297.16	92.79	2.42
	Lasso	387.44	89.86	3.15
	CART	349.17	53.20	2.84
	Bagging	**228.75**	**43.35**	**1.86**
	RF	447.77	161.84	3.64
	SVM	424.15	143.02	3.45
	kNN	428.05	125.84	3.48

Table V.1.2 – Modeling performances using deteriorated wind measures (average)

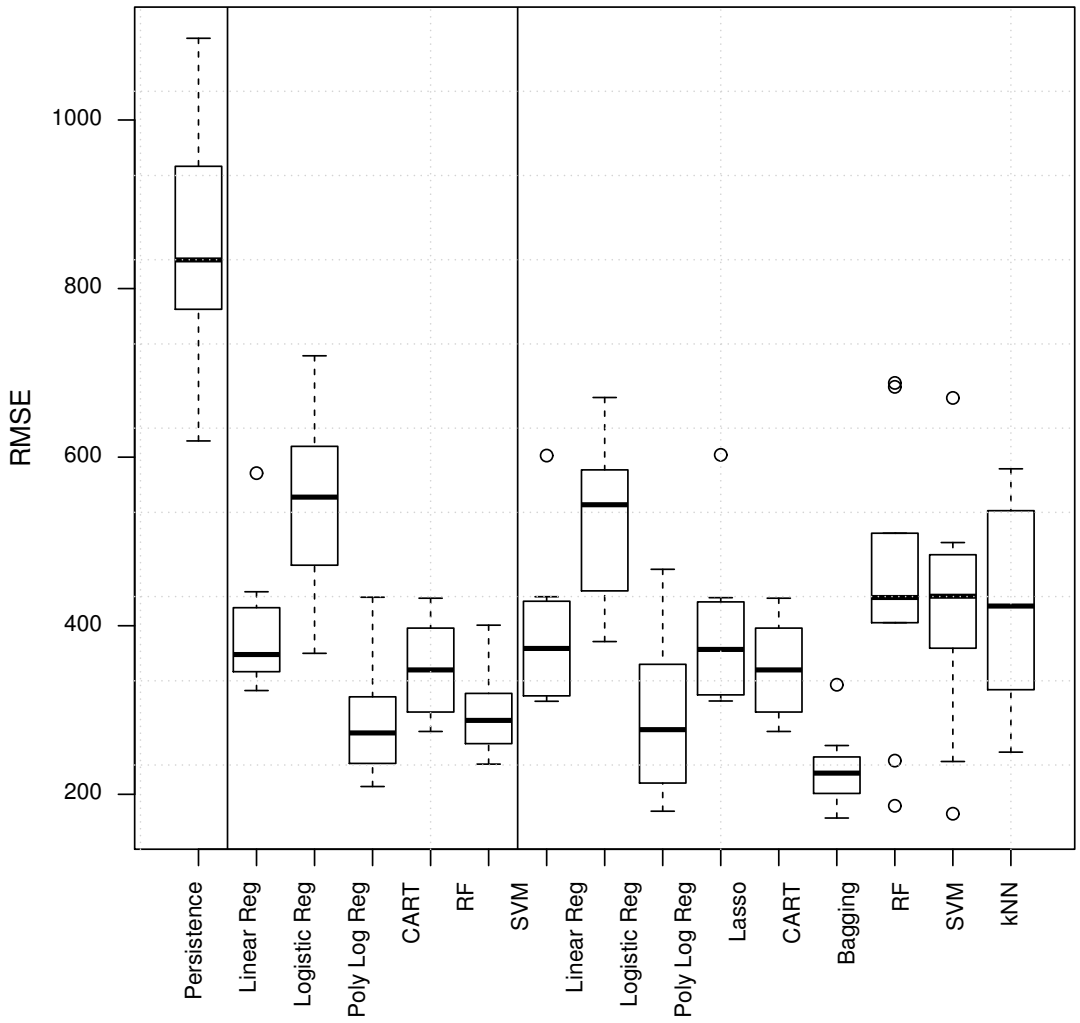


Figure V.1.4 – Boxplots of the RMSE for the different procedures using averaged measures

Comparison of different farms Just as in the previous framework, bagging and the turbine builder's power curve prediction have been tested on several farms. Figure V.1.5 stresses the good results obtained with bagging, which seems robust to data averaging.

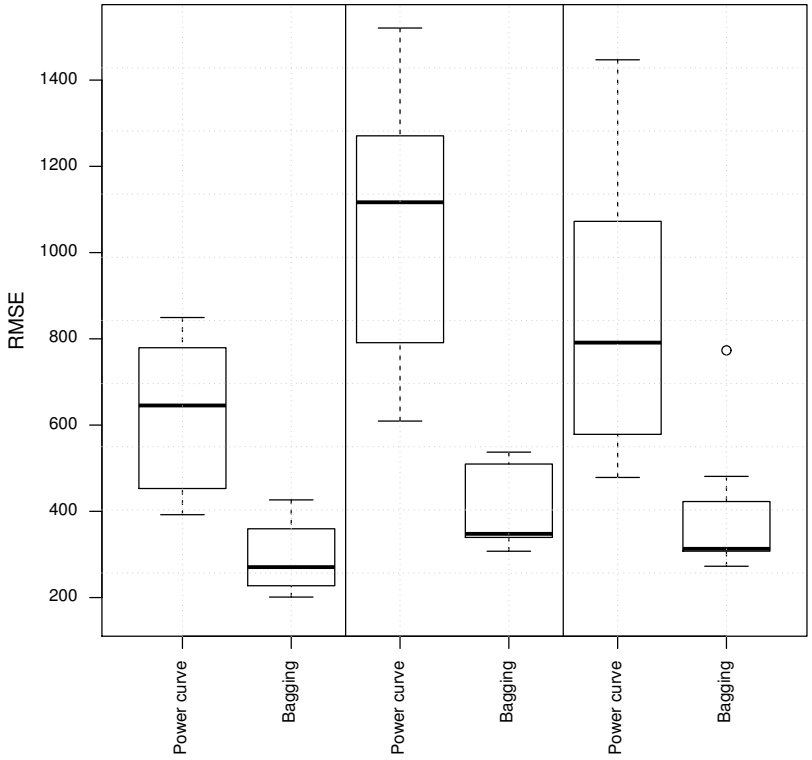


Figure V.1.5 – Comparison of the RMSE for the turbine builder's power curve and the bagging procedure on several farms using averaged measures

V.1.6 Conclusion and perspectives

Depending on the wind farm, a method which is the best with true local wind information inputs may not perform well any more when using averaged data designed to mimic meteorological wind forecasts. So, despite the good performance of the constructor power curve for one farm, bagging shows to be more robust when turning to aggregated data.

More generally, this observation raised the following question: in the frame of this work, the data, provided by the company Maïa Eolis, comes from 3 wind farms, all located in the North and East of France, the first turbines installed by the company

being essentially located in these regions, but it could be of prime interest to have access to wind data from farms in other regions of France. Note that, in the comparison of data-mining approaches conducted by Kusiak et al. (2009) for US wind farm data, the kNN method, which does not perform particularly well in our study, appears to outperform the other methods.

Here, we calibrated our models using stationary data, that is, data corresponding to full functioning of the wind turbines. A complementary work may be to enrich our models including the time slots where wind turbines are working in a non stationary regime (corresponding essentially to start-up regime). This would allow to compute predictions over a (complete) long time of use, taking into account transitory phenomena of a turbine.

Regarding effective forecasting, an intermediate step has to be accomplished before simply plugging in our models the information brought by Météo France or the ECMWF. Indeed, these previsions suffer inevitably from a bias due to several causes, which has to be corrected in order to build accurate forecasting at the end. For example, the wind speed prevision is provided by the meteorologists for a given height, but the wind speed at the height of the wind turbine may be different. This shows that it is necessary to elaborate a so-called downscaling method, in other words to find the best possible relationship between real wind at a wind turbine and the meteorological wind forecasts at hand.

Chapter V.2

Downscaling wind speed: SIRTA data

This chapter corresponds to a collaboration with Mathilde Mougeot and Riwal Plougonven (LMD, École Polytechnique), in the frame of the 3-year projects of Mehdi Kechiar and Côme de Lassus Saint-Geniès and the PhD work of Bastien Alonzo at École Polytechnique. The corresponding article Alonzo et al. (2018) has been published by Springer as a chapter of the monograph Renewable Energy: Forecasting and Risk Management.

V.2.1 Introduction

Surface wind speed is a meteorological variable of considerable importance because it impacts human activities in a number of ways, including damage to buildings, fallen tower cranes, and injuries due to objects carried by intense winds. Recently, the significant development of wind energy has created a new motivation and demand for estimations of the wind speed near the surface.

According to the GWEC (Global Wind Energy Council), 54GW wind capacity has been installed in 2016, corresponding to an increase of 12.6% of the total installed capacity. Worldwide, the number of wind farms increases each year, feeding the electrical network with a larger amount of energy. For instance, in 2016, France has seen its highest capacity growth rate ever recorded. This sharp increase of connected wind power has for example allowed the network to receive 8.6 GW from wind power plants, on November 20th, corresponding to 17.9% of the energy produced this day as reported by the French electricity transmission system operator RTE (Réseau

de Transport de l'Électricité). It is very important to be able to anticipate energy production, in order to plan maintenance operations and manage the balance between energy production and consumption. Moreover, the evolution of the energy market regulation, with the end of feeding-in tariffs, make this anticipation crucial for wind energy producers.

The question of precisely estimating wind speed at specific locations has received renewed interest from the wind-energy sector since appropriate forecasts at specific location of a wind farm are required. Very different approaches have been considered for forecasting the wind speed at locations of wind farms for different lead times: for short lead times of minutes to a few hours, statistical learning models trained with the locally observed wind speed have been developed using a variety of techniques (e.g. Chang, 2014; Foley et al., 2012; Tascikaraoglu and Uzunoglu, 2014; Wang et al., 2011). For longer lead times, from half a day to several days, outputs from numerical weather prediction models have been used (Lazic et al., 2014; Mejia et al., 2018; Ranaboldo et al., 2013). The increasing need for accurate forecasts of the surface wind speed fortunately comes with the improvement of the NWP models describing and forecasting atmospheric motions. The skill of NWP models is continuously increasing (Bauer et al., 2015), as well as their spatial resolutions. They constitute undoubtedly a key source of information for surface wind speed forecasts.

Nevertheless, these models are not necessarily performing uniformly well for all atmospheric variables. If we can be very confident in the ability of NWP models to represent several variables, some others may not be so reliable. More specifically, variables such as mid-tropospheric pressure which reflect the large-scale mass distribution, are well understood physically (Vallis, 2006) and efficiently modeled numerically. However, variables tied to phenomena occurring on smaller scales (such as cloud-cover or near-surface winds) depend much more directly on processes that are "parameterized", represented by a simplified process. In contrast to large-scale motions (governed by the Navier-Stokes equations), parameterizations are generally partly rooted in physical arguments, but also in large part empirical. According to Haiden et al. (2018), surface wind components from NWP models still include significant errors. When comparing output from a numerical model to a local measurement, there will always be several sources of error: model error, any modeling involving a simplification (several physical processes, especially parameterized ones, are not well modeled), representativity error, corresponding to the contrast between the values over a grid box and the value at a specific point, numerical error (even if we were describing only processes governed by well-established physical laws, discretization is unavoidable). It is worth pointing out that finer scale models inevitably

come with a higher numerical cost.

Strategies to estimate surface winds, or other meteorological variables, from the output of Numerical Weather Prediction models (NWP) or climate models have been developed in several contexts, with different motivations, and leading to different methodologies and applications.As NWP models perform now very well in predicting large-scale systems, there is a strong motivation for taking advantage of both NWP outputs and local variables at an observation site. Actually estimating a local quantity based on output of a numerical weather prediction model and past observations at a given location, mostly via linear models, has been an active field of research for half a century, called Model Output Statistics (Glahn and Lowry, 1972). Nowadays, it is common for operational centers to carry out MOS to provide forecasts of quantities where observations are available (Baars and Mass, 2005; Kang et al., 2011; Schmeits et al., 2005; Wilson and Vallée, 2002; Zamo et al., 2014).

The task of building a procedure to estimate a variable sensitive to small scales based on information on the large-scale flow is known as downscaling in meteorology and climatology. When used in the context of climate projections, the aim is to generate plausible time series of local variables in climate change scenarios Wilby and Dawson (2013). Downscaling has also been applied to surface winds. In this framework, different studies have shown the importance of a certain set of variables to predict surface wind speed. Among them, markers of large-scale systems (geopotential height, pressure fields) and boundary layer stability drivers (surface temperature, boundary layer height, wind and temperature gradient) can be cited (Davy et al., 2010; Devis et al., 2013; Salameh et al., 2009). In terms of methodology, several methods have been investigated, including linear regression, support vector regression or neural networks (Jung and Broadwater, 2014; Soman et al., 2010). The most common method used is linear regression, with a central issue being the choice of explanatory variables. For instance, Ranaboldo et al. (2013) present a stepwise procedure to identify the most relevant variables to forecast 10m wind speed at two locations, showing that variables describing the flow lead to the best performances.

The European Center for Medium-range Weather Forecasts (ECMWF) is an intergovernmental operational center that provides medium-range weather forecasts on a global scale, and has the largest repository of archived global weather data. The model has reached a spatial resolution of 0.125° in latitude and longitude, corresponding to a resolution of about 9km in the horizontal. While this is a fine resolution for a global numerical weather prediction model, this remains coarse-grained when comparing 10m wind speed to measurements at one specific location, given for instance the sensitivity to the local topography. Note that ECMWF now provides a

100m wind speed output variable, developed mainly for wind energy applications. Since surface variables such as 10m and 100m wind speed may not be sufficiently well represented in the ECMWF model, surface wind speed should be corrected by using the robust information given by some observed variables. Here, we use surface wind speed observed at 10m, 100m over a long period of 5 years at SIRTA observation platform (Haeffelin et al., 2005). Our aim is, in particular, to explore how different statistical models perform in modeling and forecasting the 10m and 100m wind speed using informations of ECMWF analyses and forecasts outputs at different horizons.

Below, Section V.2.3 focuses on modeling, and Section V.2.4 is dedicated to forecasts. Before that, Section V.2.2 presents the methodology and the data.

V.2.2 Methodology and data

This section provides an overview of the statistical methodology and describes the data in more detail.

V.2.2.1 Methodology

Our aim is to model and forecast the real observed wind speed from NWP model outputs. Our approach will rely on ECMWF analyses, the best estimate of the atmospheric state at a given time using a model and observations (Kalnay, 2003), and ECMWF forecasts. More specifically, analyses are obtained by assimilating observed data from within a time window around the corresponding time to previous forecasts made by the numerical weather prediction model. In what follows, the observed wind speed is the target and the variables retrieved from ECMWF are potential explanatory features. Because of the complexity of meteorological phenomena, statistical modeling provides an appropriate framework for corrections of representativity errors and modeling of site-dependent variability.

The methods chosen are linear regression, because it is a simple, widely used technique, and random forests, which, to our knowledge, were not deeply studied in the framework of downscaling surface wind speed. Among the meteorological variables, some of them provide more important information linked to the target than others, and some of them may be correlated. For linear regression, performing variable selection, either by applying a stepwise procedure or via the ℓ_1 Lasso regularization method (Tibshirani, 1994), may be useful to keep only the most important variables.

Regarding random forests, variables are somehow processed automatically, and, moreover, nonlinear relations are handled very well. Therefore, the comparison of the two different kinds of statistical models, parametric and nonparametric, should be very instructive.

V.2.2.2 Data

Observed wind speed In this chapter, we use observations of the wind speed at the SIRTA observation platform (Haeffelin et al., 2005). Surface wind speed at 10m height from anemometer recording is available at 5-minutes frequency. The wind speed at 100m height from Lidar recording is available at 10-minutes frequency. Both data span for 5 years from 2011 to 2015. We filter observations by a sinusoidal function over a 6-hour window centered at 00h, 06h, 12h and 18h to obtain a 6-hourly observed wind speed to be compared to the NWP model outputs available at this time frequency. We found that the resulting time series are not sensitive to the filter function. We also try different filtering windows, concluding that 6-hours is the best to compare to the NWP model outputs. Due to some missing data, two final time series of 5049 filtered observations are computed (over 7304 if all data were available). SIRTA observatory is based 20km in south of Paris on the Saclay plateau ($48.7°N$ and $2.2°E$). Figure V.2.1 shows the SIRTA observation platform location, marked by the red point on the map, and its close environment. Regarding the relief near SIRTA, observe that a forest is located at about 50m north to the measurement devices. To the south, buildings can be found at about 300m from the SIRTA observatory. In the east-west axis, no close obstacle are encountered. Further south, the edge of the Saclay plateau shows a vertical drop of about 70m, from 160m on top to 90m at the bottom.

ECMWF Analyses Variables are retrieved from ECMWF analyses at 4 points around the SIRTA platform. Given the spatial resolution of ECMWF analyses, topography is smoothed compared to the real one. As the surface wind speed is very influenced by the terrain, the modeled surface wind speed is not necessarily close to the observed wind speed. The data spans from 01/01/2011 to 31/12/2015 at the 6-hour frequency. It is sampled at each date where a filtered sampled observation is available.

The near surface wind speed at a given location can be linked to different phenomena. The large-scale circulation brings the flow to the given location explaining the slowly varying wind speed. The wind speed in altitude, the geopotential height,

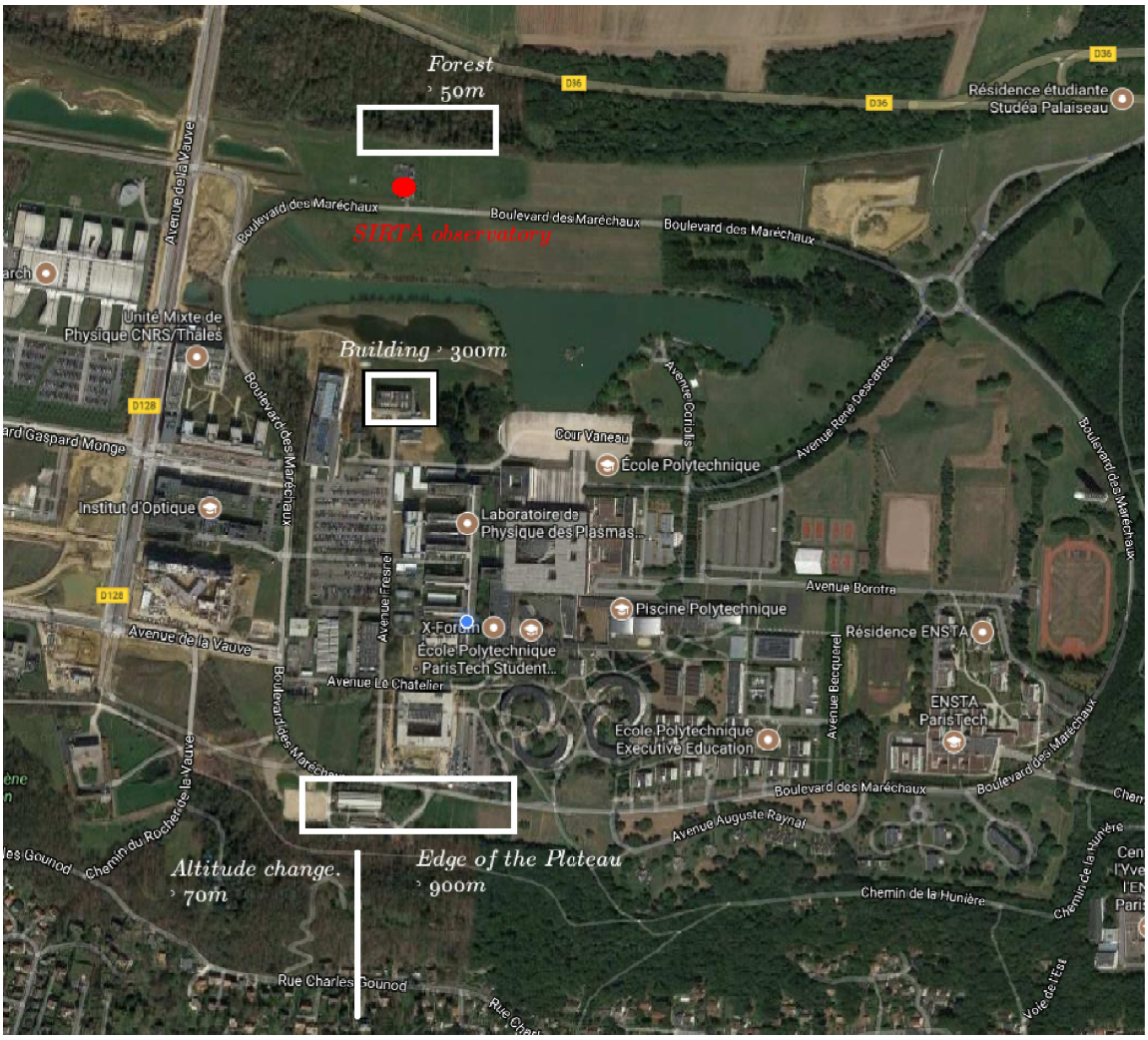


Figure V.2.1 – Map of the SIRTA observation platform and its surroundings

the vorticity, the flow divergence, sometimes the temperature, can be markers of large systems like depressions, fronts, storms, or high pressure systems explaining a large part of the low frequency variations of the surface wind speed (Table V.2.2). At a finer scale, what is happening in the boundary layer is very important to explain the intra-day variations of the wind speed. The state and stability of the boundary layer can be derived from surface variables describing the exchanges inside the layer. Exchanges are driven mostly by temperature gradient and wind shear that develop turbulent flow (Table V.2.3). These variables are computed from the raw data. Thermodynamic variables like surface, skin, and dew point temperatures and surface heat fluxes can also inform on the stability of the boundary layer, as well as its height and dissipation on its state (Table V.2.1). In the tables, the zonal wind speed is the component along the local parallel of latitude, whereas the meridional wind speed is the component along the local meridian.

In the end, 20 output variables are retrieved from ECMWF analyses at the 4 points around the SIRTA observatory and at different pressure levels. Note that we restrict the study to local variables (at the location of measurements or in the column above). It might also be possible to take advantage from larger scale information (Davy et al., 2010; Zamo et al., 2016). The choice of taking 4 points around the SIRTA platform has the advantage of being simple and straightforward. Providing instead the explanatory variables by their interpolated value at SIRTA and the two components of their gradient does not lead to significantly different results.

Altitude (m)	Variable	Unit
10m/100m	Norm of the wind speed	$m.s^{-1}$
10m/100m	Zonal wind speed	$m.s^{-1}$
10m/100m	Meridional wind speed	$m.s^{-1}$
2m	Temperature	K
2m	Dew point Temperature	K
Surface	Skin temperature	K
Surface	Mean sea level pressure	Pa
Surface	Surface pressure	Pa
–	Boundary layer height	m
–	Boundary layer dissipation	$J.m^{-2}$
Surface	Surface latent heat flux	$J.m^{-2}$
Surface	Surface sensible heat flux	$J.m^{-2}$

Table V.2.1 – Surface variables

Variable	Unit
Zonal wind speed	$m.s^{-1}$
Meridional wind speed	$m.s^{-1}$
Geopotential height	$m^2.s^{-2}$
Divergence	s^{-1}
Vorticity	s^{-1}
Temperature	K

Table V.2.2 – Altitude variables (pressure levels: 1000hPa, 925hPa, 850hPa, 700hPa, 500hPa)

Pressure level (hPa)	Variable	Unit
10m to 925hPa	Wind shear	$m.s^{-1}$
10m to 925hPa	Temperature gradient	K

Table V.2.3 – Computed variables

ECMWF forecasts For the wind speed forecast in Section V.2.4, we will apply our model to the year 2015 of forecasts from ECMWF model. A forecast is launched every day at 00:00 UTC. The time resolution retained is 3 hours and the maximum lead time is 5 days. The same variables as for the analyses are retrieved at the same points around the SIRTA platform.

V.2.3 The relationship between analyzed and observed winds

V.2.3.1 10m/100m wind speed variability comparison

We compare the observed wind speed at 10m and 100m with the 10m and 100m wind speed output of the ECMWF analyses at the closest grid point, respectively. No significant difference can be found when using other grid points, or the mean of the four surrounding locations.

Figure V.2.2 shows the probability density function (pdf) of the wind speed coming from ECMWF analyses and observations, as well as an example of a time series of corresponding wind speeds. It appears that the 10m wind speed from ECMWF

analyses displays a systematic bias by overestimating the 10m observed wind speed (Figure V.2.2, a and b). The wind at 100m is comparatively well modeled in terms of variations in the time series, but also in terms of distribution (Figure V.2.2, c and d). It seems that the errors mainly come from the overestimation of peaked wind speeds and the underestimation of low wind speeds (Figure V.2.2, c and d). As 10m wind speed is very influenced by even low topography and surrounding obstacles, which are smoothed or not represented in ECMWF analyses, some of its variations are not well described. The effect of the topography and terrain specificity have less impact on the 100m wind speed, so that it is much better represented in ECMWF analyses.

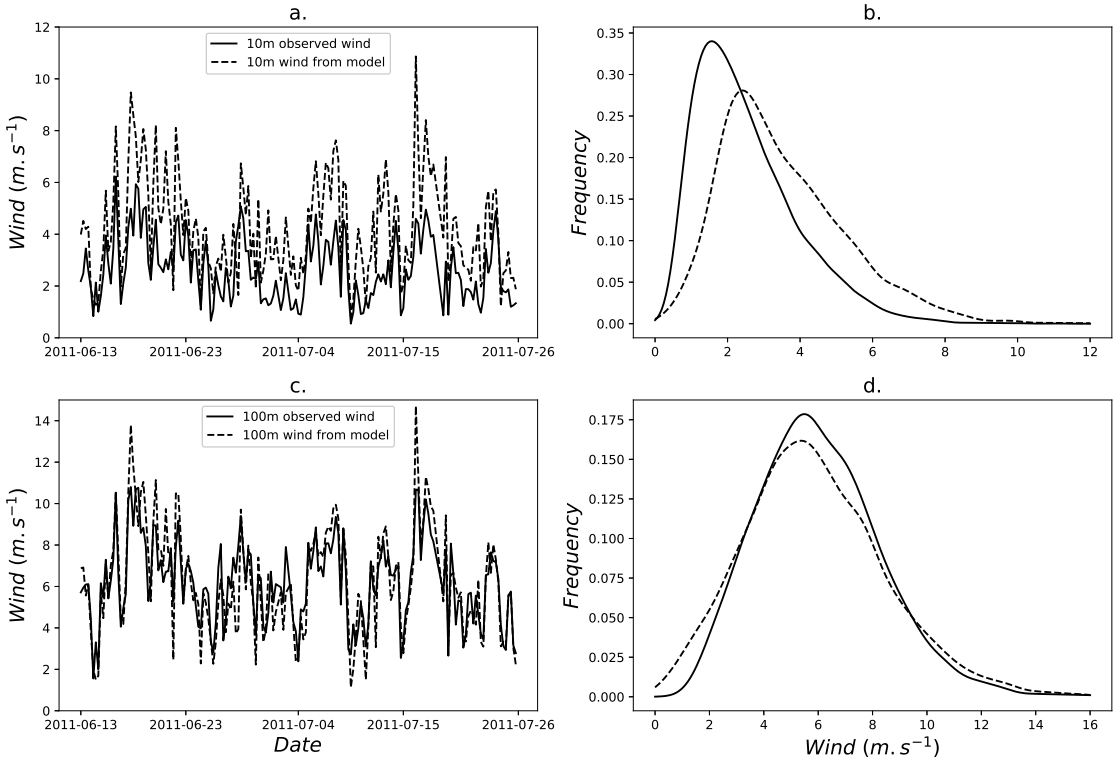


Figure V.2.2 – Comparison between observed wind data and ECMWF analyses: 10m (top) and 100m (bottom) wind speed time series in summer 2011 (panels a and c, respectively) and pdfs corresponding to the 5 years of data (panels b and d)

The ability of the ECMWF model to represent the observed wind speed is quantified in Table V.2.4 by the mean deviation (Dev), the Root Mean Square Error (RMSE), and the correlation coefficient (Corr). Here, the deviation for the i^{th} observation is $(y_i - x_i)$,

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - y_i)^2}{n}}, \qquad \text{Corr} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}},$$

where x_i is the wind speed from the NWP model and y_i the observed wind speed, n is the number of samples (x_i, y_i), $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

Periods	Dev		RMSE		Corr	
	F10	F100	F10	F100	F10	F100
2011-2015	-1.00	0.14	1.41	1.01	0.82	0.93
2011	-1.19	0.04	1.59	1.06	0.80	0.91
2012	-0.94	0.23	1.31	1.03	0.85	0.92
2013	-1.13	0.06	1.52	0.93	0.82	0.94
2014	-0.88	0.26	1.30	1.00	0.80	0.93
2015	-0.87	0.14	1.30	0.97	0.82	0.94
Winter	-0.97	0.04	1.41	0.97	0.83	0.94
Spring	-1.11	0.27	1.56	1.02	0.71	0.90
Summer	-0.92	0.33	1.31	1.04	0.80	0.91
Fall	-1.04	-0.10	1.36	1.00	0.87	0.93

Table V.2.4 – Mean deviation, RMSE, and correlation performed by ECMWF analyses for modeling the 10m and 100m wind speed

No clear improvement of the ECMWF analyses over the years from 2011 to 2015 can be detected in Table V.2.4. The correlation stays quite constant over the years for both 10m and 100m wind speeds. The mean deviation and RMSE seem to decrease for the 10m wind speed, with nevertheless a notable good score in 2012. The variations of performance may come only from changes in the predictability of the weather over Europe (Folland et al., 2012). Seasonal variations of the performance of ECMWF analyses can be seen, especially on the correlation between the observed and modeled wind speed. At both 10m and 100m, the analyzed wind speed is better correlated with the observations in winter and fall than in spring and summer. In all cases, the scores shown are better for the 100m wind speed than for the 10m wind speed.

Variations of the performance of the ECMWF analyses in representing the observed wind speed are evidenced by Figure V.2.3. The figure shows the 10m wind speed from ECMWF analyses as a function of the 10m observed wind speed for different directions of the analyzed wind. It is obvious that the errors made by the numerical model differ regarding the direction of the wind, which can be easily linked to the specificity of the terrain. Indeed, when a north wind is recorded, it has been blocked by the forest north of the anemometer. The same happens for south winds with the building situated further and which influence is thus not as substantial as the forest. Figure V.2.4 displays the same as Figure V.2.3 but for the 100m wind speed. It seems that there is no more dependence of the performance of the ECMWF analyses regarding the direction of the 100m wind speed ; it appears to be not significantly impacted by the surrounding forests and buildings.

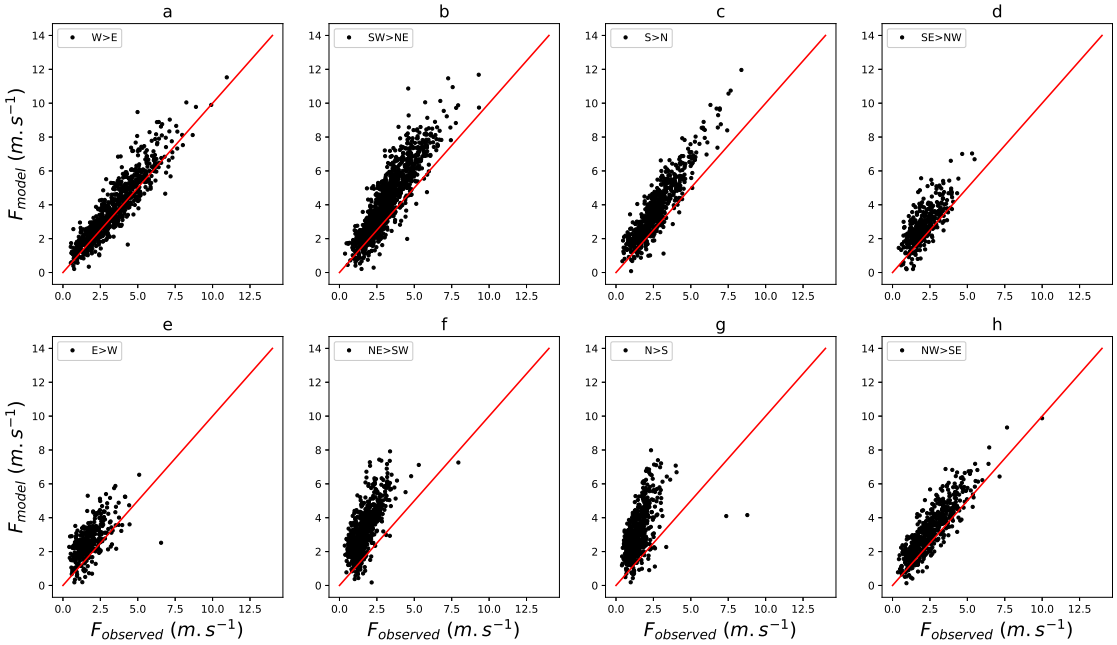


Figure V.2.3 – 10m wind speed from ECMWF analyses as a function of the 10m observed wind speed, given cardinal directions: a=west, b=southwest, c=south, d=southeast, e=east, f=northeast, g=north, h=northwest

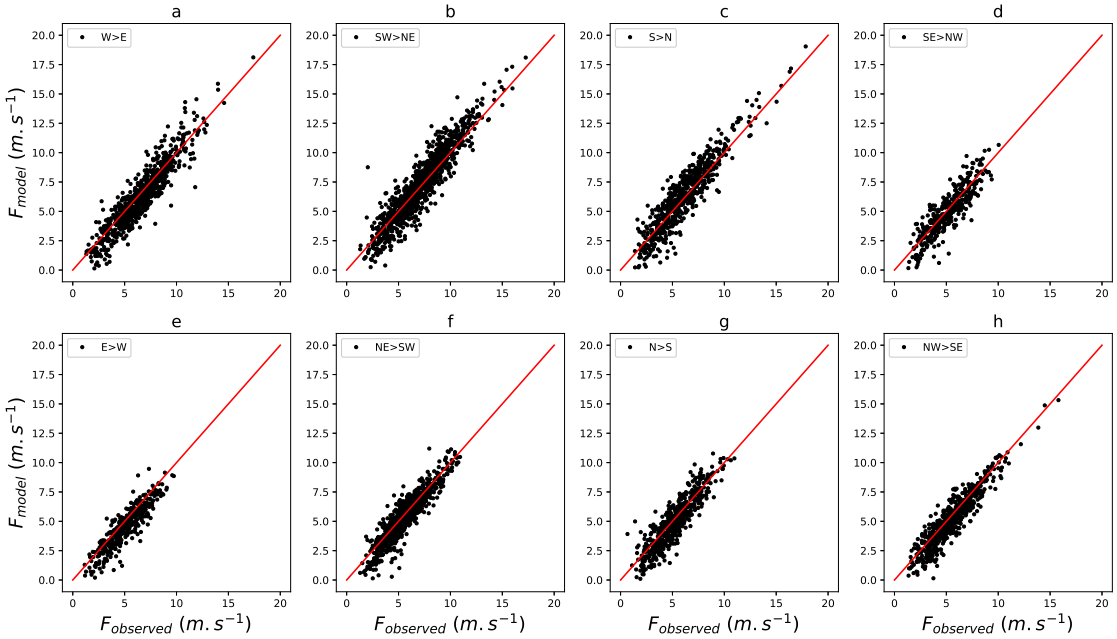


Figure V.2.4 – 100m wind speed from ECMWF analyses as a function of the 10m observed wind speed, given cardinal directions: a=west, b=southwest, c=south, d=southeast, e=east, f=northeast, g=north, h=northwest

V.2.3.2 Reconstruction of the observed wind speed based on NWP outputs

In the sequel, a k-fold cross-validation is performed over 10 different periods taken within the 5 years of analyses and observations. Each time, statistical downscaling models are trained on a given period of about 4500 data points and applied over the remaining period of about 500 data points to reconstruct the 10m and 100m wind speed. Table V.2.5 summarizes the different statistical models considered: linear model with only the 10m wind, with all variables, with stepwise or Lasso selection, and random forests. Moreover, the models may be fitted conditionally to the direction of the wind speed. All models are evaluated in terms of RMSE and correlation with the observed wind speed.

Model type	Variables	Direction	Name
Linear	F10	No	LR_F
Linear	All	No	LR_A
Linear	Stepwise	No	LR_{SW}
Linear	Lasso	No	LR_{La}
Linear	F10	Yes	LR_F^{dir}
Linear	All	Yes	LR_A^{dir}
Linear	Stepwise	Yes	LR_{SW}^{dir}
Random forest	All	No	RF_A
Random forest	All	Yes	RF_A^{dir}

Table V.2.5 – Statistical models used to downscale 10m and 100m wind speed

10m wind speed reconstruction Figure V.2.6 shows results for the reconstruction of the 10m wind speed. First, by using only wind speed with a linear model LR_F, RMSE is reduced by about 40%, but the correlation stays constant. Adding other variables to linear model (i.e. LR_A, LR_{SW} and LR_{La}) allows to reduce the RMSE by 60%, and to significantly improve correlation from 0.80 to 0.91 between reconstructed wind speed and observed one. Random forests RF_A perform slightly better than these linear models. Note that variables retained by stepwise selection are very diverse (wind speed, large scale variables, boundary layer state drivers), while Lasso mainly selects wind speed and wind component variables, thus using redundant information. Analyzing the main variables used by random forests shows that much weight is put on wind component, highlighting the dependence of the error on the 10m wind speed regarding its direction.

Fitting a linear model in each direction (model denoted by LR_{SW}^{dir}), with associated stepwise variable selection, performs better than any other model (Figure V.2.3). We obtain a significant improvement of the RMSE and correlation scores. As expected regarding Figure V.2.3 (g), the best improvement is retrieved for north wind speed and is of more than $0.1 m.s^{-1}$ compared to LR_{SW}. No improvement is found for east winds, surely because the number of data is too small. Fitting a random forest in each direction does not improve results, probably because the direction is already well handled by this model by using the zonal and meridional component of the wind. A big advantage of random forests is that it does not require to explore the data in depth beforehand for extracting appropriate and relevant features as inputs to the model.

Figure V.2.5 shows the time series of the 10m observed wind speed and the NWP model wind speed output for summer 2011 (panel a) and the probability density function corresponding to the entire period, 2011 to 2015 (panel b). Panels c and e show respectively time series of the reconstructed 10m wind speed by LR_{SW}^{dir} (red line) and LR_{SW} (blue line), and by RF_A^{dir} (magenta line) and RF_A (cyan line). Panels d and f show the corresponding pdfs. All statistical models allow for a good bias correction. All models underestimate the small quantiles of the distribution and give a distribution very peaked around the mode. High percentiles are however well reconstructed.

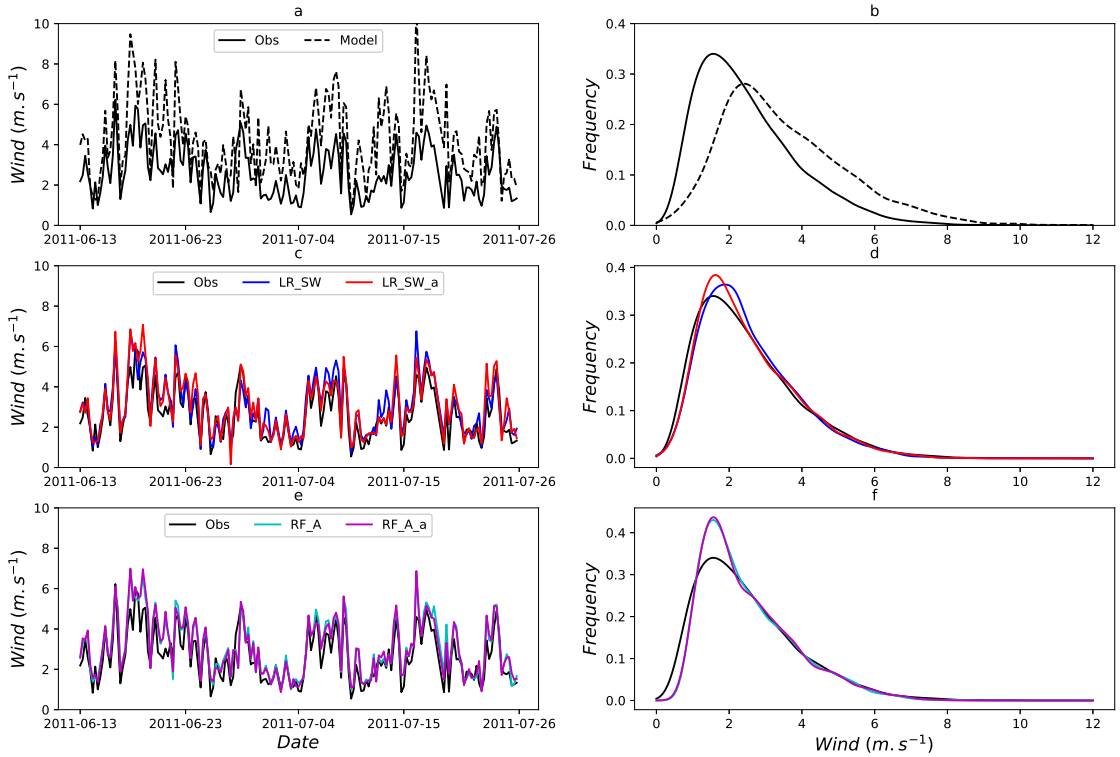


Figure V.2.5 – Timeseries and pdfs of the observed 10m wind speed (straight black line), ECMWF analyses (dotted black line) (a and b), linear models LR_{SW} (blue) and LR_{SW}^{dir} (red) (c and d), random forest models RF_A (cyan) and RF_A^{dir} (magenta) (e and f)

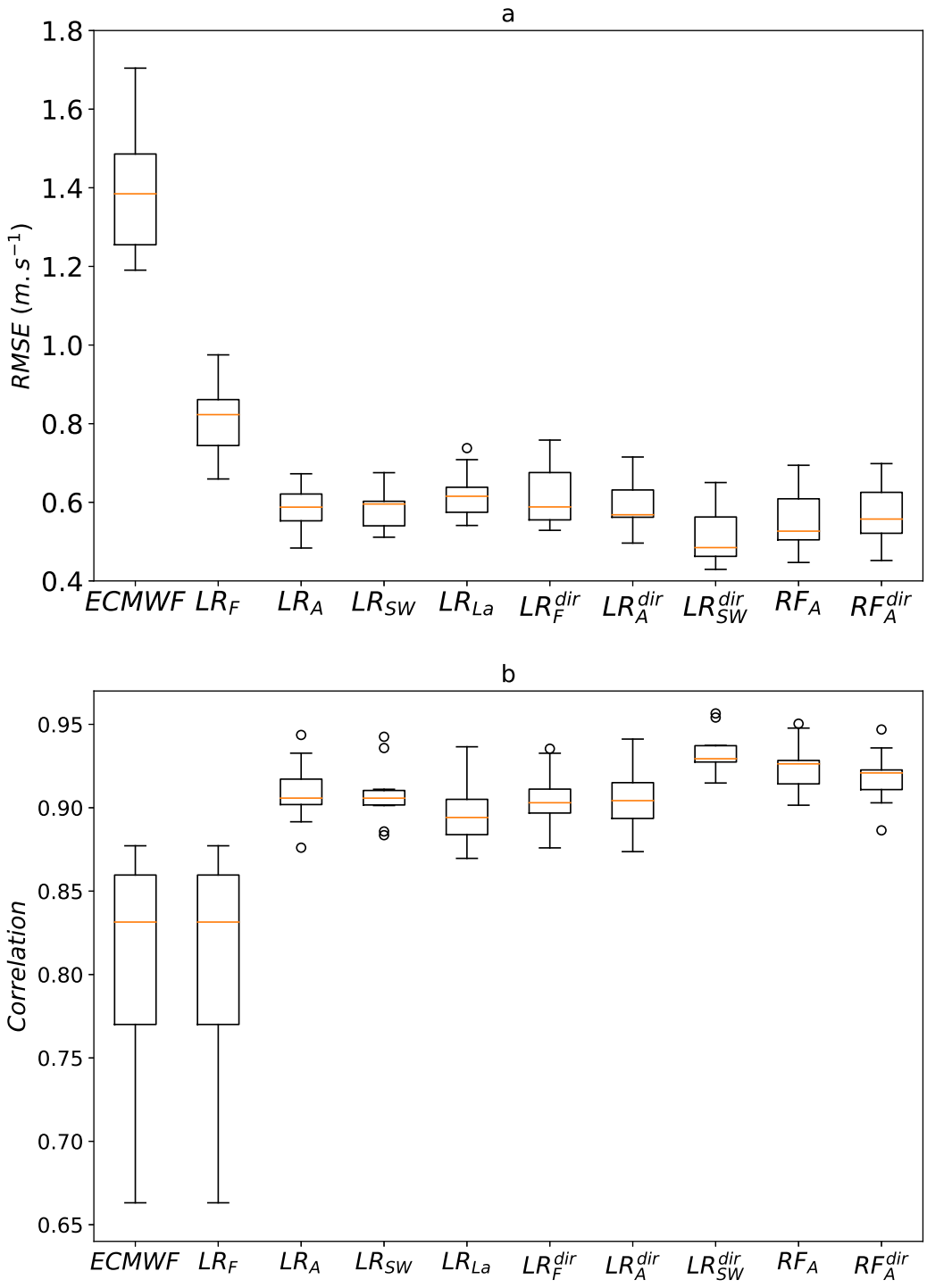


Figure V.2.6 – RMSE and correlation results when reconstructing 10m wind speed with the statistical models compared to the ECMWF analyses

100m wind speed reconstruction Figure V.2.7 shows the 100m wind speed reconstruction results. The model LR_F reduces the RMSE of about 15% corresponding to $0.14m.s^{-1}$ and the best model LR_{SW}^{dir} by 23% corresponding to $0.23m.s^{-1}$. The correlation is improved from 0.92 to 0.94. Adding the direction dependence to linear model with only 100m wind speed (LR_F^{dir}) does not improve the results, since the error on the 100m wind speed does not depend on the direction. Using all explanatory variables (LR_A^{dir}) leads to a strong over-fitting, but the model with stepwise variable selection (LR_{SW}^{dir}) performs significantly better. In the case of 100m wind speed, the best improvement is found for east wind speeds. For random forests, the information on the direction does not improve the results, as already noticed for 10m wind speed reconstruction. The most important variables for random forests and stepwise selection are the 100m wind speed and the wind shear in the boundary layer, whereas the Lasso technique mainly selects 100m wind speed.

Figure V.2.8 shows the time series of 100m observed wind speed and the NWP model wind speed output for summer 2011 (panel a) and the probability density function corresponding to the entire period from 2011 to 2015 (panel b). Panels c and e show respectively time series of the reconstructed 100m wind speed by LR_{SW}^{dir} (red line) and LR_{SW} (blue line), and by RF_A^{dir} (magenta line) and RF_A (cyan line). Panels d and f show the corresponding pdfs. Some peaked wind speeds are less overestimated after statistical downscaling. As for the 10m wind speed, statistical models underestimate the small quantiles of the distribution and give a distribution peaked around the mode.

As a conclusion, we observe that the 100m wind speed is already well represented in ECMWF analyses, with a good correlation and no systematic bias. Nevertheless, statistical models reduces the RMSE on the 10m wind speed between 40% and 65%, and between 15% and 23% for the 100m wind speed, improving at the same time the correlation between the observed wind speed and the reconstructed one. Note that random forests give, without specific calibration, results comparable to those of the best linear models.

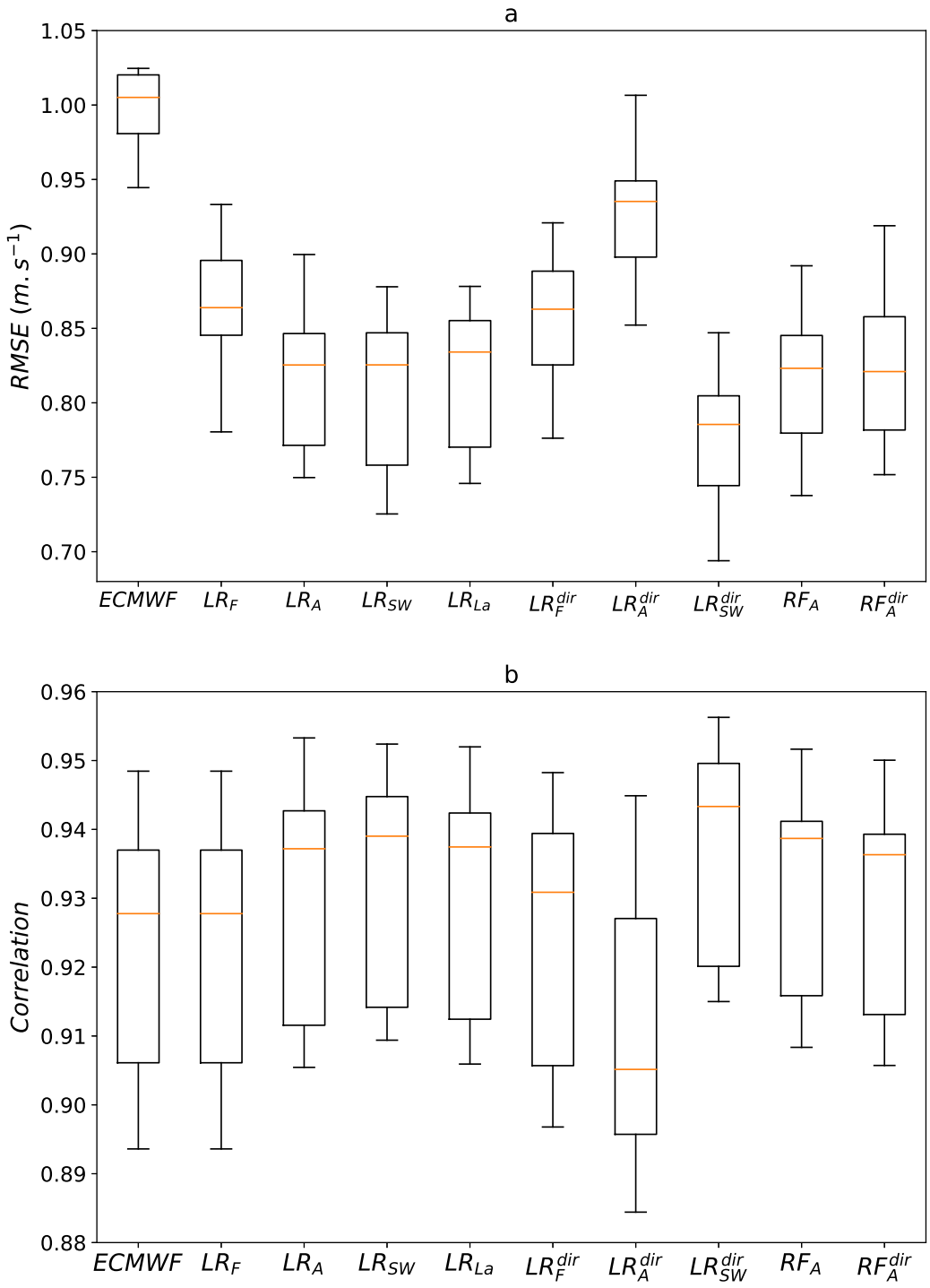


Figure V.2.7 – RMSE and correlation results when reconstructing 100m wind speed with the statistical models compared to the ECMWF analyses

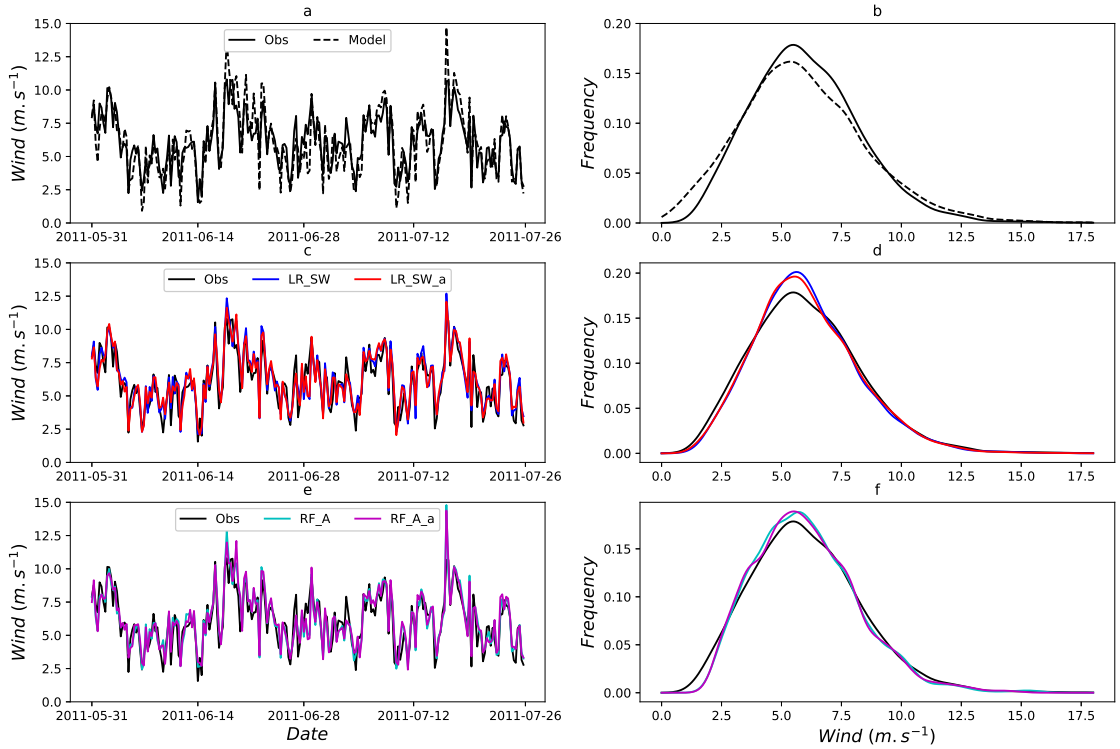


Figure V.2.8 – Timeseries and pdfs of the observed 100m wind speed (straight black line), ECMWF analyses (dotted black line) (a and b), linear models LR_{SW} (blue) and LR_{SW}^{dir} (red) (c and d), random forest models RF_A (cyan) and RF_A^{dir} (magenta) (e and f)

V.2.4 Forecasts of surface winds

In this section, we use the previous statistical models based on the knowledge of the observed wind speed and the outputs of ECMWF analyses to forecast wind speed at five days horizon with a frequency of 3 hours. We have access to 1 year of ECMWF forecasts in 2015. We train these statistical models on ECMWF analyses from 2011 to 2014, and apply the resulting model to the forecasts. Figures V.2.9 and V.2.10 show respectively the RMSE averaged over the 365 days for the 10m and 100m wind speed. A strong diurnal cycle of the performances of both ECMWF forecasts and downscaled statistical predictions of the 10m wind speed is evidenced. This diurnal cycle seems to be observed also for 100m wind speed forecasts, but with a less important amplitude. As the data set is trained on the ECMWF analyses, it

seems that diurnal cycle is better represented in ECMWF analyses than in ECMWF forecasts. This could be explained by the data assimilation system that may help to correct errors coming from NWP model parameterizations.

Adding the dependence with the direction (LR_{SW}^{dir}) allows for a significant reduction of the RMSE. Random forests RF_A show the best performance. Here, the robustness of the nonparametric method proves to be a very valuable feature. For 100m wind speed forecasts, all statistical models are quite similar.

For both 10m and 100m wind speed forecasts, statistical downscaling models allow for significant improvements, at any lead time from 3 hours to 5 days. Here, the models were trained on analyses, which may not be optimal. Training directly on ECMWF forecasts, for each lead time separately, should deeply improve results, by removing the displayed diurnal cycle, and also letting the increase in RMSE with the lead time be less sharp.

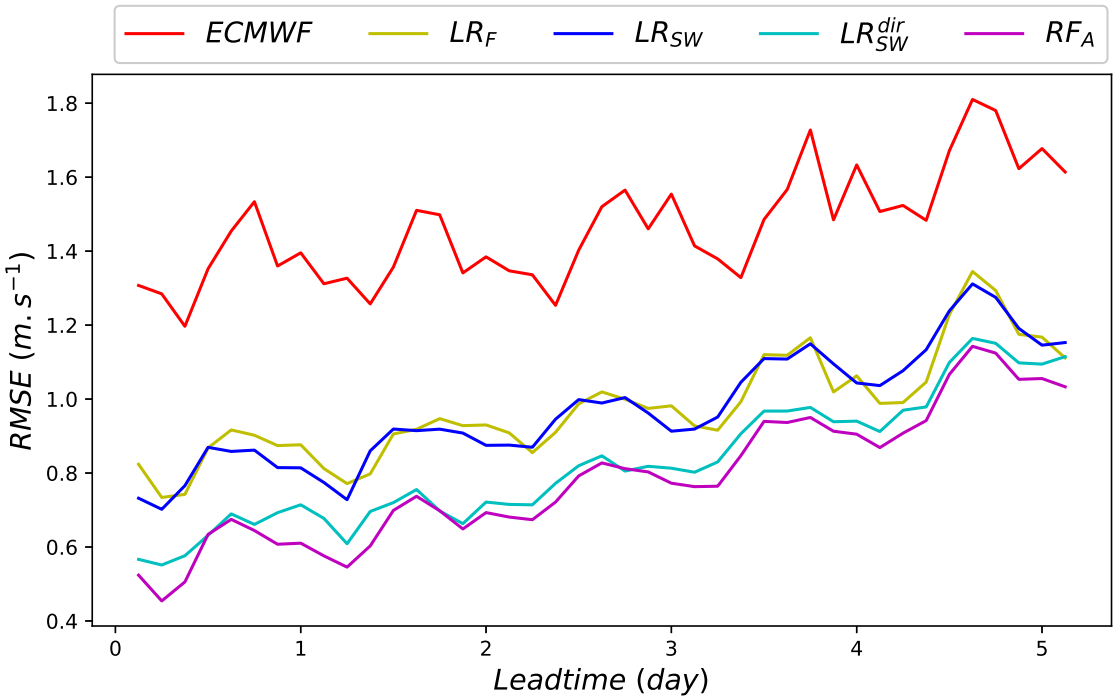


Figure V.2.9 – RMSE, computed between the 10m observed wind speed, and the 10m forecast wind speed, averaged over the entire set of forecasts

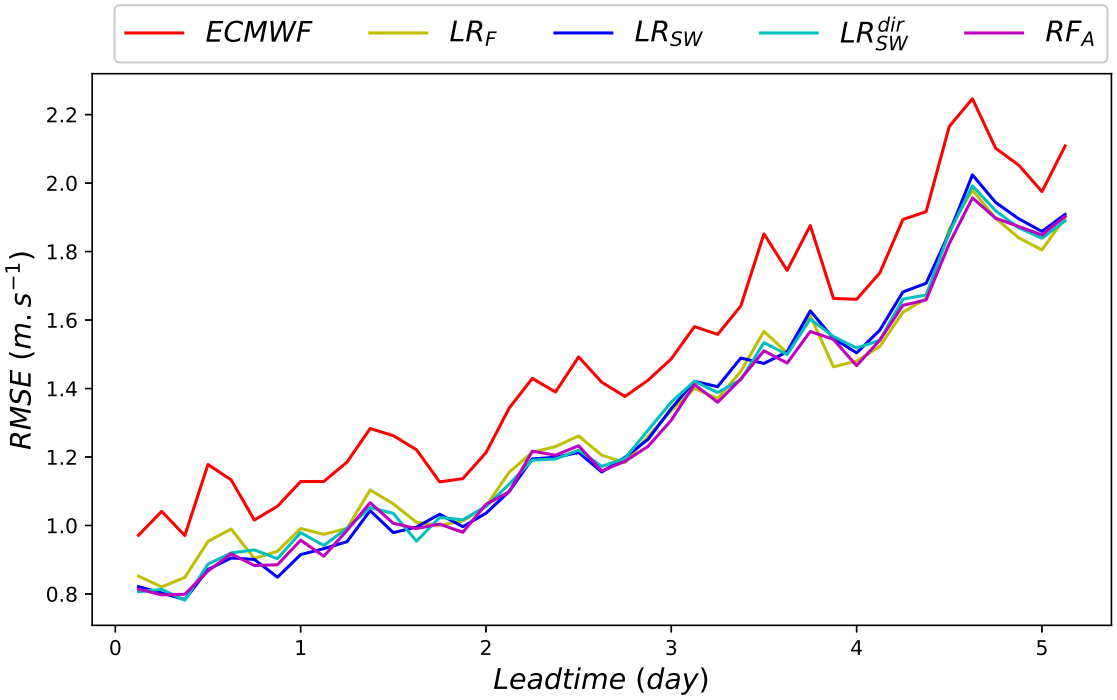


Figure V.2.10 – RMSE, computed between the 100m observed wind speed, and the 100m forecast wind speed, averaged over the entire set of forecasts

Chapter V.3

Downscaling wind speed: data over France

This chapter is the result of a collaboration with Mathilde Mougeot and Riwal Plougonven, in the frame of the 3-year projects of Rebeca Doctors and Lishan Liao and the Master 1 internship of Naveen Goutham at École Polytechnique. The corresponding article Goutham et al. (2021) has been published in the journal Boundary-layer meteorology.

V.3.1 Introduction

As mentioned in Chapter V.2, numerical weather prediction models constitute an important source of information on surface flows. However, as the latter are turbulent and strongly influenced by small-scale features absent in the limited representation of these models, the modeled surface wind components, when compared with local observations at a given site, generally contain large errors, including biases. As investigated in Chapter V.2, for a given site where observations are available for a long enough interval, these observations may be used to correct the biases and errors of the model for that location. The purpose is to refine the determination of low-level wind speed, at locations where observations are available, using statistical learning models to link information from a numerical weather prediction model and from past observations.

We aim to explore and improve the estimation for the local 10m wind speed from recent outputs of the ECMWF model over stations in France sampling different geographical settings. Specific issues considered are the performance of the numerical weather prediction model and the improvement gained by using parametric and nonparametric models.

The data and methods used are described in Section V.3.2. The performance of the numerical weather prediction model and of its combinations with different post-processing models are assessed and compared in Section V.3.3. In Section V.3.4, focusing on the best model, we reduce the number of explanatory variables and identify what seems, for all stations, to constitute the most informative list of variables. Finally, we discuss the issue of forecasting for one station in Section V.3.5.

V.3.2 Data and Methodology

The data includes wind speed observations and, as in Chapter V.2, numerical weather prediction model outputs from ECMWF.

The observations come from the Integrated Surface Database, which is a global database of observed meteorological data available at an hourly time resolution (Smith et al., 2011). About 400 weather stations in France provide their meteorological data in the Integrated Surface Database. In order to better train the models, we retained stations with over 90% of available data for a span of eight years from 2010 to 2017. The retrieved observed data comes from 171 stations well distributed across mainland France and Corsica.

The aim is to model the 10m wind speed at these meteorological stations in France from the outputs of the ECMWF model. The ECMWF explanatory variables are presented in Tables V.3.1 to V.3.3 below. The local surface wind speed is related to the synoptic-scale flow. The large-scale (synoptic) systems like depressions, fronts, and storms are described in terms of physical variables at different pressure levels in Table V.3.1. However, the intra-day wind speed variations that occur in the boundary layer may not be fully explained by the synoptic flows. Table V.3.2 collects variables that convey information about the stability of the boundary layer. Other important variables that convey information about the vertical-exchange processes in the boundary layer are the vertical wind shear and the temperature gradient (Table V.3.3).

Pressure level (hPa)	Variable	Unit
1000/925/850/500	Zonal wind component	$m.s^{-1}$
1000/925/850/500	Meridional wind component	$m.s^{-1}$
1000/925/850/500	Geopotential height	$m^2\,s^{-2}$
1000/925/850/500	Divergence	s^{-1}
1000/925/850/500	Vorticity	s^{-1}
1000/925/850/500	Temperature	K

Table V.3.1 – Explanatory variables, at different pressure levels

Altitude	Variable	Unit
10m/100m	Wind speed	$m.s^{-1}$
10m/100m	Zonal wind component	$m.s^{-1}$
10m/100m	Meridional wind component	$m.s^{-1}$
2m	Temperature	K
Surface	Skin temperature	K
–	Mean sea level pressure	Pa
Surface	Surface pressure	Pa
–	Boundary layer height	m
–	Boundary layer dissipation	$J.m^{-2}$
Surface	Surface latent heat flux	$J.m^{-2}$
Surface	Surface sensible heat flux	$J.m^{-2}$

Table V.3.2 – Explanatory variables retrieved from surface variables of the numerical weather prediction model. The last three variables are accumulated over the last six hours.

Vertical level	Variable	Unit
10m to 100m	Bulk wind shear	$m.s^{-1}$
1000hPa to 925hPa	Bulk wind shear	$m.s^{-1}$
1000hPa to 925hPa	Temperature difference	K

Table V.3.3 – Explanatory variables computed as differences in the vertical between two height or pressure levels

The data used in the statistical models for a specific station is obtained via linear interpolation from the closest grid points of the ECMWF model surrounding that station. We also computed north–south, east–west, and diagonal gradients around each station, estimated using finite differences. We found that the north–south and east–west gradients were more significant than the diagonal gradients. Hence, for each variable, we retained its value interpolated at the station location and the two gradients (north–south and east–west) to feed into the machine learning models. This leads to 117 explanatory variables for each station.

The time period covered by the data set is April 2010 – December 2017. In order for the observed data to match the six-hour frequency of the ECMWF model outputs, we only consider two-hour averaging windows centered at 00:00, 06:00, 12:00 and 18:00 UTC.

The ability of the ECMWF model to represent the observed wind speed is quantified by the root mean square error (RMSE) denoted by $E_{w,obs}$, and Pearson's correlation coefficient $\rho_{w,obs}$, defined as follows:

$$E_{w,obs} = \sqrt{\frac{\sum_{t\in\mathcal{S}} \left(y_t^w - y_t^{obs}\right)^2}{|\mathcal{S}|}}, \quad \rho_{w,obs} = \frac{\sum_{t\in\mathcal{S}} \left(y_t^w - \bar{y}^w\right)\left(y_t^{obs} - \bar{y}^{obs}\right)}{\sqrt{\sum_{t\in\mathcal{S}} \left(y_t^w - \bar{y}^w\right)^2}\sqrt{\sum_{t\in\mathcal{S}} \left(y_t^{obs} - \bar{y}^{obs}\right)^2}}.$$

Here, w stands for the time series from the ECMWF analyses and *obs* for the observed wind speed, $\mathcal{S}$ denotes the set of indices of the data, $|\mathcal{S}|$ the number of elements of $\mathcal{S}$, and $\bar{y} = \frac{1}{|\mathcal{S}|}\sum_{t\in\mathcal{S}} y_t$ is the mean of the time series y.

Figure V.3.1 shows the RMSE and correlation coefficient for the 10m wind speed between the observations and the ECMWF analyses, for all the meteorological stations under consideration in France. Figure V.3.1 (a) shows that the RMSE of the wind speed from the ECMWF analyses exceeds $1.0 m.s^{-1}$ for most of the inland stations: the minimum at an individual station is $0.95 m.s^{-1}$, the maximum is $4.58 m.s^{-1}$. The average over all stations is $1.74 m.s^{-1}$, with a standard deviation of $0.79 m.s^{-1}$. The coastal stations in the west, south and Corsica have a higher RMSE, greater than $2 m.s^{-1}$. In Figure V.3.1 (b), we see that the correlation coefficient for inland stations in the north is about 0.8, whereas for stations in the south and along the coasts it hardly reaches 0.7 and can be as low as 0.4. Note that, because of the higher RMSE and the lower correlation coefficient found along the coasts, special attention was paid to these stations during interpolation to check if the location of grid points from the model has an effect. Upon careful examination, it was noticed that the location of grid points has no significant influence. The poorer performance may be due to factors that contribute to the difficulty of modeling wind speed at the coast.

These include the discontinuity in surface conditions and the ensuing complexity of the boundary layer, and also possibly local phenomena such as sea breeze.

We computed the annual averages of the RMSE and correlation coefficient of the ECMWF analyses over all stations (see Figure V.3.2). An improvement of the performance of the model in the year 2014 is observed, resulting from changes in the ECMWF model, notably a modification of the parametrization of surface drag and the increase of the vertical resolution from 91 to 137 levels in June 2013 (Riddaway, 2013). Nevertheless, these changes did not have an impact on the correlation coefficient. The average RMSE and the correlation coefficient for the time period 2010–2017 are $1.74 m.s^{-1}$ and 0.68 respectively. Extreme values for the RMSE are $1.82 m.s^{-1}$ (in 2010) and $1.68 m.s^{-1}$ (in 2016), and the average is $1.74 m.s^{-1}$. The correlation coefficient is stable during this period. The median of the RMSE for all stations is $1.42 m.s^{-1}$, smaller than the average, as can be expected for a positive variable which can be very large in locations where the model performs very poorly. These errors are significant given that the time-averaged wind speed averaged over all stations is $3.4 m.s^{-1}$. More precisely, we calculated the ratio of the RMSE to the time-averaged wind speed for each station. The overall mean of these ratios is 0.52, implying that any significant decrease of the error is worthwhile.

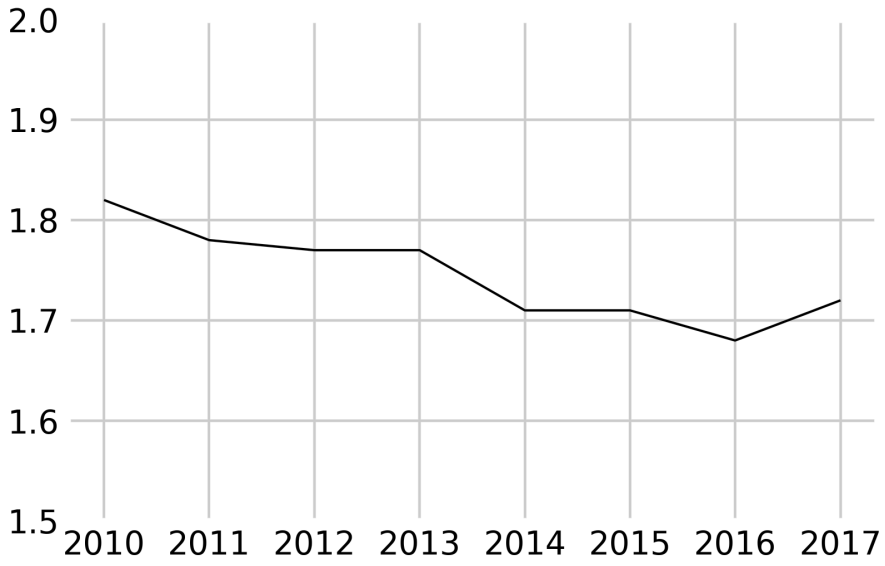
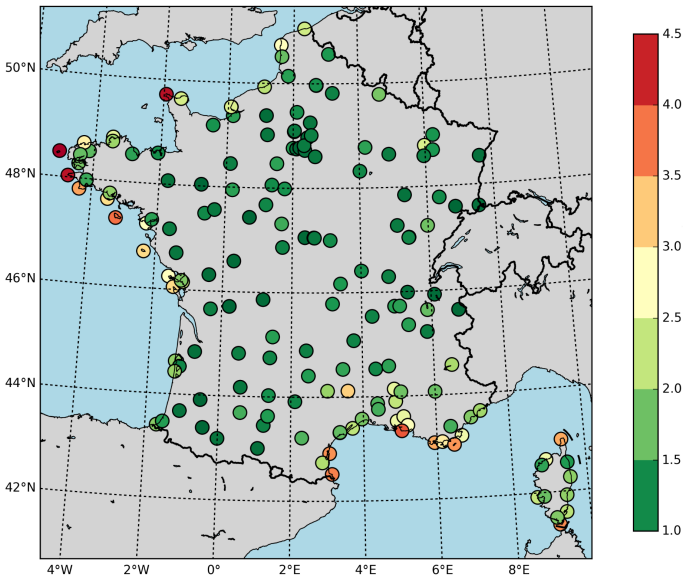
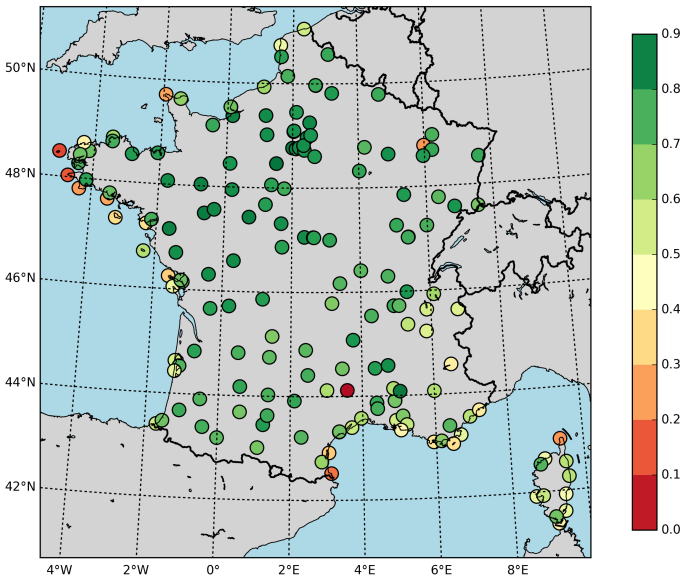


Figure V.3.2 – RMSE ($m.s^{-1}$) of the ECMWF analyses for the wind speed at 10m over all the stations in France for the years 2010 to 2017.



(a)



(b)

Figure V.3.1 – (a) RMSE and (b) correlation coefficient for the 10m wind speed between the observations and the ECMWF analyses

V.3.3 Comparison of different parametric and nonparametric models

The parametric models implemented are linear regression with all explanatory variables (hereafter LRall), linear regression with stepwise selection of variables (hereafter LRstep), and with ℓ^1-regularization (hereafter Lasso). The nonparametric models are random forest (hereafter RF) and gradient boosting (hereafter GB) with all variables, and k-nearest neighbors (hereafter k-NN) using the 10most important explanatory variables provided by the random forest model. Results from another k-NN model, with only five wind speed-related explanatory variables, were not satisfactory and are not reported here. The models are summarized in Table V.3.4. Random forest models with less variables are discussed in Section V.3.4. In order to train and test the different models, 10-fold cross-validation is used.

Machine learning model	Name
Linear regression with all variables	LRall
Linear regression with stepwise selection	LRstep
Linear regression with Lasso regularization	Lasso
Random forest with all variables	RF
Gradient boosting with all variables	GB
k-nearest neighbors, 10 variables	k-NN

Table V.3.4 – The different models

V.3.3.1 Modeling performances for one station

We first study one station, Le Havre-Octeville (49.53°N, 0.08°E). The station is located on the coast, in northern France, and is the northernmost station on the Greenwhich meridian. Figure V.3.4 displays the time series and scatter plot of 10m observed and modeled wind speed over a few weeks in 2010. This station was chosen as qualitatively representative of the overall results, with a rather pronounced, but not exceptional, improvement. Other individual stations typically display the same ordering of the performances of the different models, but with rather weaker contrasts for inland stations, and with comparable or greater improvements for many coastal stations. The 10m wind speed from the ECMWF analyses shows a significant error, as illustrated in the time series (purple line of Figure V.3.4). The machine learning models (green and yellow lines in the time series) closely follow the observed wind

speed (black line in the time series), suggesting improvements for the RMSE and the correlation coefficient over the ECMWF analyses. The scatter plot shows that the ECMWF model usually overestimates the wind speed over $4m.s^{-1}$ as can be seen from the scatter plot (represented by purple dots). The implemented models generally underestimate the wind speed over $5m.s^{-1}$ (illustrated by green and yellow dots). Figure V.3.3 shows the RMSE and correlation coefficient of all the models (over 8 years) for the reconstruction of the 10m wind speed at the considered station. The RMSE of the ECMWF analyses is high at $2.3m.s^{-1}$ whereas the correlation coefficient is low at about 0.7. All the implemented models imply improvement, resulting in RMSE reduced to values between $1.05m.s^{-1}$ and $1.35m.s^{-1}$, and the correlation coefficient increased to values between 0.73 and 0.86. Among the implemented models, three groups can be distinguished. The first group includes the linear regression models which reduce the RMSE by about 44% and increase the correlation coefficient by about 6%. The second noticeable group consists of the tree-based machine learning models, which give the best performance, reducing the RMSE by 55% and increasing the correlation coefficient by 22%. The performance of the k-NN model is intermediate between the first and the second groups with an improvement in the RMSE and the correlation coefficient by 50% and 15% over the ECMWF analyses.

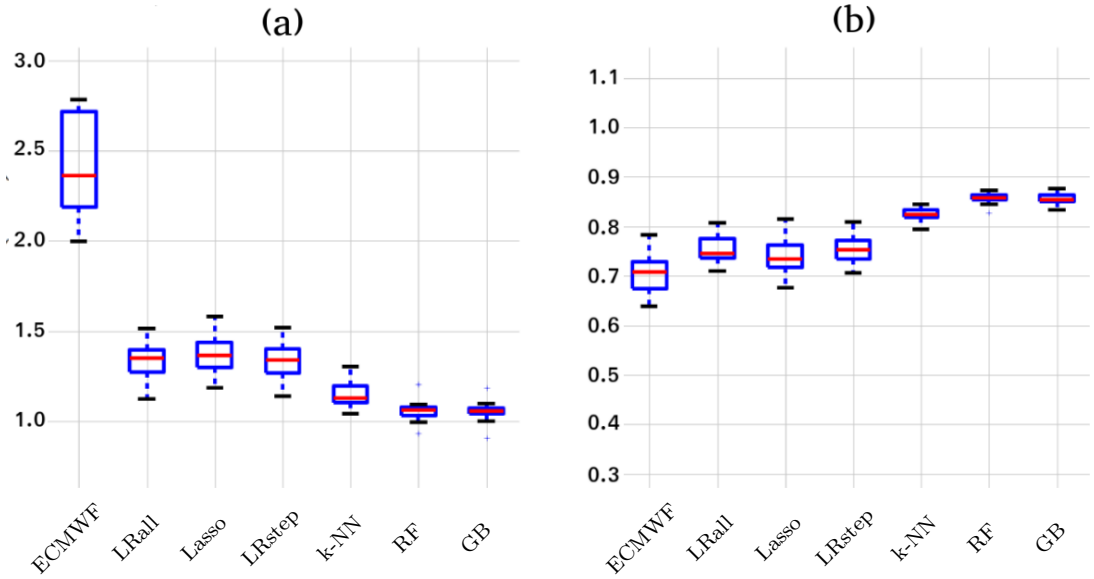


Figure V.3.3 – (a) RMSE and (b) correlation coefficient of all models for the station Le Havre-Octeville in France

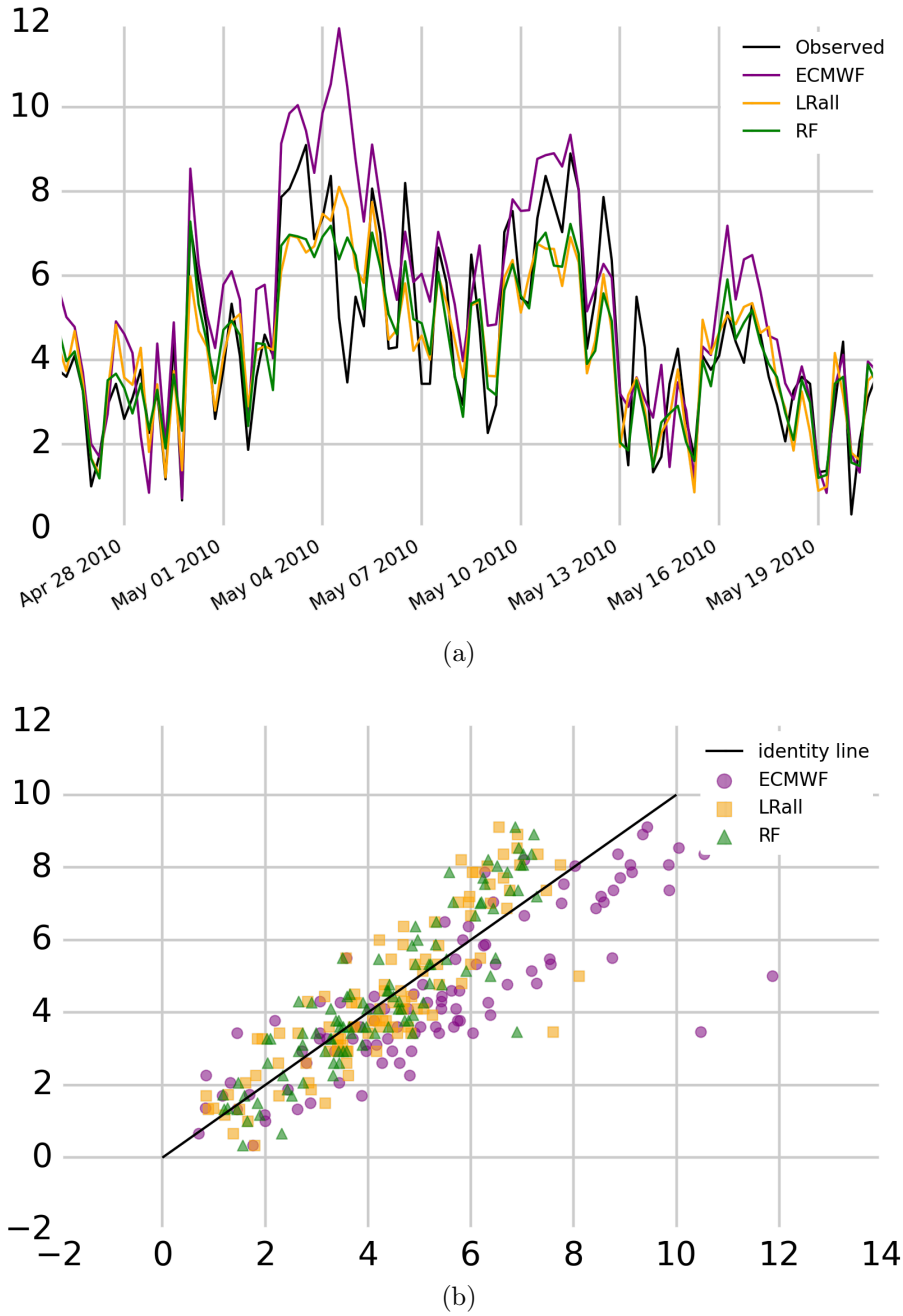


Figure V.3.4 – Wind speed at 10m at Le Havre-Octeville station: (a) time series of the observed and modeled wind speed and (b) observed vs modeled wind speed

V.3.3.2 Performance of the parametric and nonparametric models over France

Figure V.3.5 displays the RMSE and the correlation coefficient of all models for all the stations in France. All the models generally perform better than the ECMWF analyses in representing the 10m wind speed. Overall, the parametric models (LRall, LRstep, and Lasso) reduce the RMSE relative to the ECMWF analyses by 25% and increase the correlation coefficient by 8%; all the models reduce the inter-quartile range of the RMSE by approximately 50% and that of the correlation coefficient by 20%. For about 25% of the stations, the parametric models lead to an RMSE lower than the minimum RMSE represented by the ECMWF model. About 25% of the stations in the ECMWF model display a RMSE higher than the highest value represented by the parametric models. The correlation coefficient of about 50% of the stations in the parametric models are above the third quartile of the ECMWF model. On the whole, the tree-based nonparametric models, such as the random forests and gradient boosting models significantly reduce the RMSE relative to ECMWF analyses by 33% and increase the correlation coefficient by 15%. Both reduce the inter-quartile range of the RMSE by roughly 60% and the correlation coefficient by 50%. About 50% of the stations in the tree-based nonparametric models have RMSE lower than the lowest value and correlation coefficient higher than the highest value of the ECMWF model. The RMSE and the correlation coefficient of about 75% of the stations in the random forests and gradient boosting models are well within the first quartile and above the third quartile of the ECMWF model, respectively. To conclude, these models seem to provide positive results with minimal effort.

V.3.3.3 Geographical pattern

The improvements obtained by the machine learning models are not homogeneous geographically. To illustrate this, Figure V.3.6 shows the percentage change in the RMSE and the correlation coefficient for the LRall model with respect to the ECMWF analyses for stations in France. It is clear that the LRall model performs significantly better than the ECMWF model everywhere. The strongest reductions in the RMSE of at least 30% are present on the western coast, the southern coast, and Corsica where the ECMWF analyses perform poorly (see Figure V.3.1). In general, the RMSE of inland stations decreases by 15%, with a few local stations showing reductions of up to 60%. The correlation coefficients follow a similar pattern with largest increases seen on the coastal stations including Corsica. On average, in-

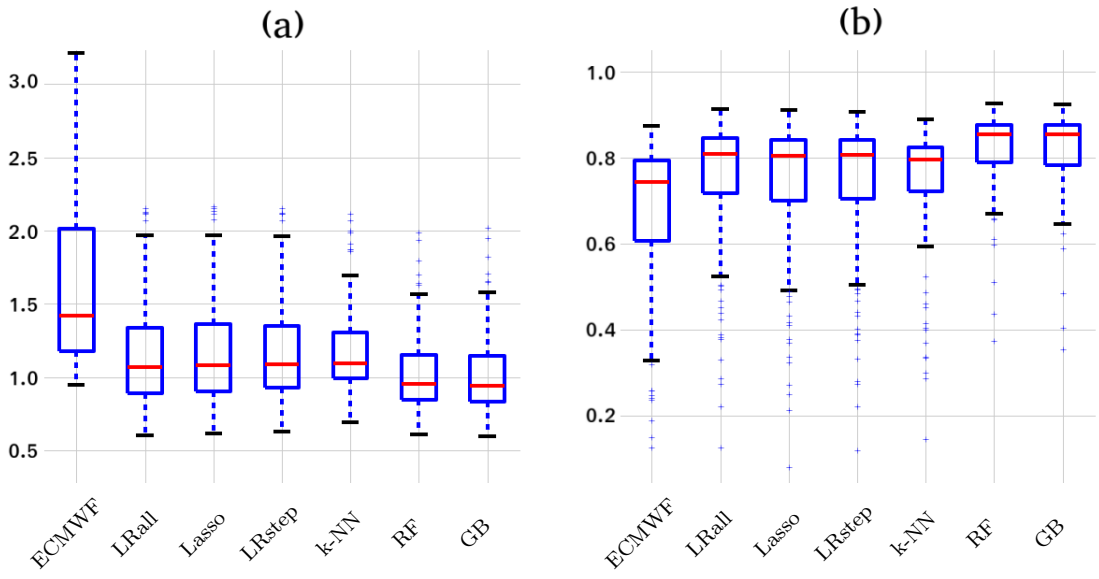
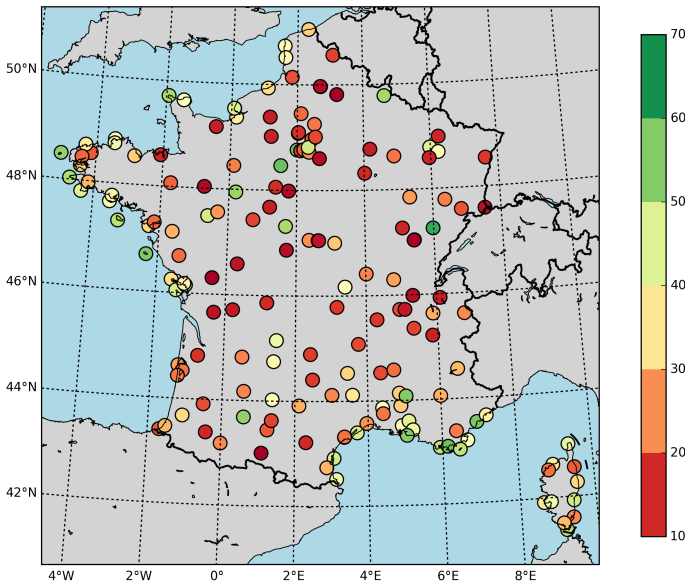


Figure V.3.5 – (a) RMSE and (b) correlation coefficient of all models for all the stations in France

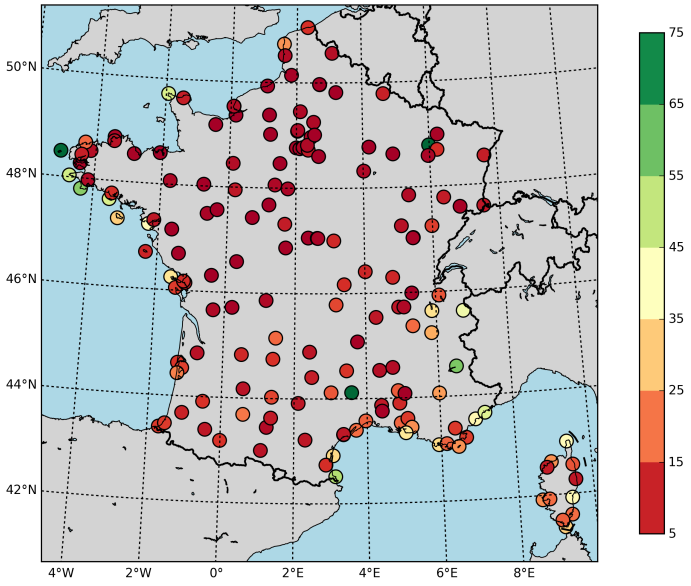
land stations show an increase of 6% for the correlation coefficient. The other two parametric models show a similar pattern.

The k-NN model performs heterogeneously. The largest reductions of the RMSE are found at the coastal stations including Corsica, whereas a few inland stations yield increases of the RMSE relative to the ECMWF analyses. The mean reduction of the RMSE for the k-NN model at the coastal stations is larger than that of the parametric models. Yet because of the poor performance of the k-NN model for the inland stations, parametric models outperform the k-NN model overall. The results for the correlation coefficient confirm these conclusions.

The performance of tree-based models shows a pattern similar to that of the parametric models but with even larger improvements. Changes in RMSE and correlation coefficients relative to the ECMWF model indicate improvement everywhere. The largest changes are found on the western coast, the southern coast, and Corsica with an average reduction of the RMSE of 50% and an increase of the correlation coefficient of 70%. In general, the RMSE of inland stations decreases by 25% with a few stations showing stronger reductions of up to 60%. The correlation coefficient increases by 12% on average for the inland stations.



(a)



(b)

Figure V.3.6 – Percentage change from ECMWF analyses to LRall model for (a) RMSE and (b) correlation coefficient

V.3.4 Relevance of the different explanatory variables

V.3.4.1 Reducing the list of explanatory variables

In order to develop a simplified and more explainable model, the relevance of explanatory variables in random forests for each station in France is analyzed. It is found that the variables relating to wind speed dominate the rank table for most of the stations. It is also noted that the ranking of explanatory variables is unique to each station with the importance value in every station dropping typically between the 40th and the 50th variables. This motivated an implementation of the random forest model with only 50 important explanatory variables specific to each station (compared with 117 explanatory variables). This model is named RF50. The performance of the model is not degraded, rather very slightly enhanced; more importantly, it is found that over 50% of the original explanatory variables do not provide useful information. Redundancy in the explanatory variables results from the very high correlation between many explanatory variables. The RF50 model reduces the list of explanatory variables for each station, but in a way specific to each station, and it therefore requires a station-specific analysis.

A more generic approach should use the same list of explanatory variables for all the stations. Figure V.3.7 shows the number of occurrences of the 50 most important explanatory variables for stations in France, obtained from the analysis of the lists of 50 most important variables for the 171 stations. For readability, we indicate with colors the categories of variables. Note that the explanatory variables based on pressure (to be more precise on the geopotential taken on isobaric surfaces) include the horizontal gradients. These gradients approximate the geostrophic wind components, where geostrophic wind denotes the theoretical wind that would result from an exact balance between the Coriolis force and the pressure gradient force. Hence, the gradients are very related to wind components.

Let us mention that 107 of the original 117 explanatory variables appear at least for one station's list of 50 most important variables.

A model based on a more generic approach named RF50C with 50 explanatory variables common to all stations was developed and it performs as well as RF50 (Figure V.3.8).

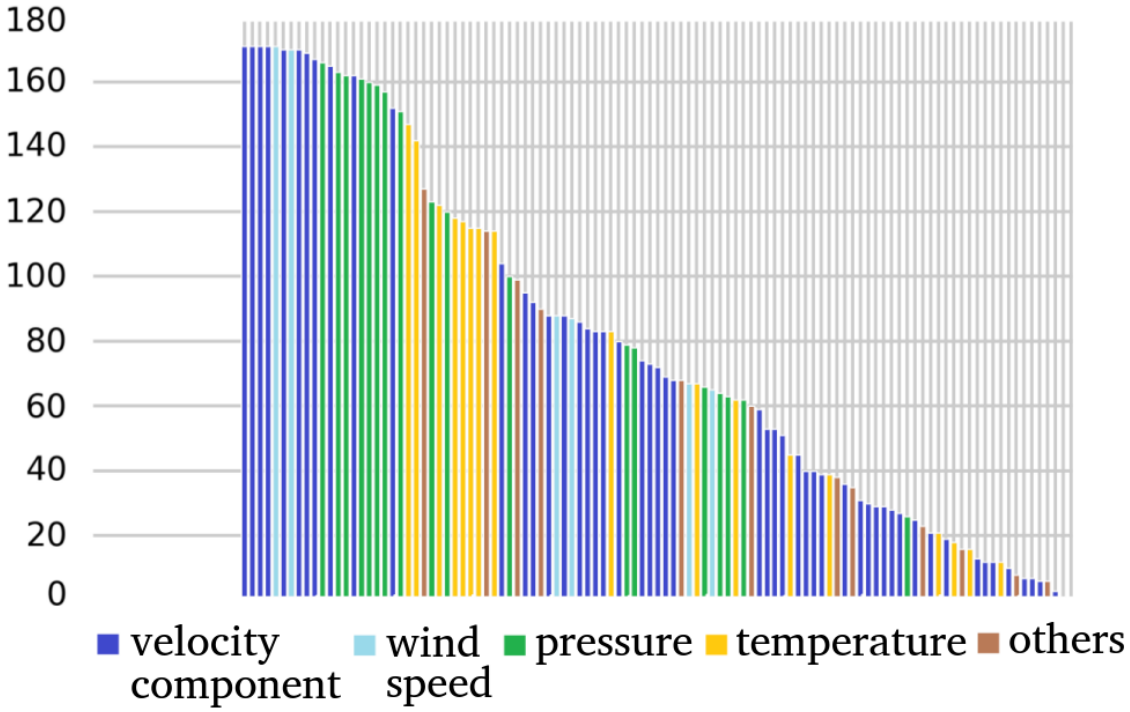


Figure V.3.7 – Number of occurrences of explanatory variables for the RF50 model for all the stations in France

To investigate how much the list of variables can be shortened, another model RF25C, with the 25 most important explanatory variables was developed. At this point, the performance begins to degrade marginally: the RF25C model performs as well as the full model, with only a 1% increase of the RMSE overall. However, the RF10C model with the 10 most important variables not only produces an increase of the RMSE by 8% and a decrease of the correlation coefficient by 2%, but it also yields an increase of the inter-quartile range for the RMSE and the correlation coefficient by 13% and 11%, respectively. Nonetheless, the RF10C model performs significantly better than all the parametric models described in Section V.3.3.2.

Further analyzing the list of explanatory variables, we find that the wind speed at 100m ($F100$), wind speed at 10m ($F10$) and bulk wind shear between 10m and 100m are the three most significant variables that provide crucial information from the synoptic flow at any given location. However, using only these three variables (RF3C) significantly degrades the performance of the model.

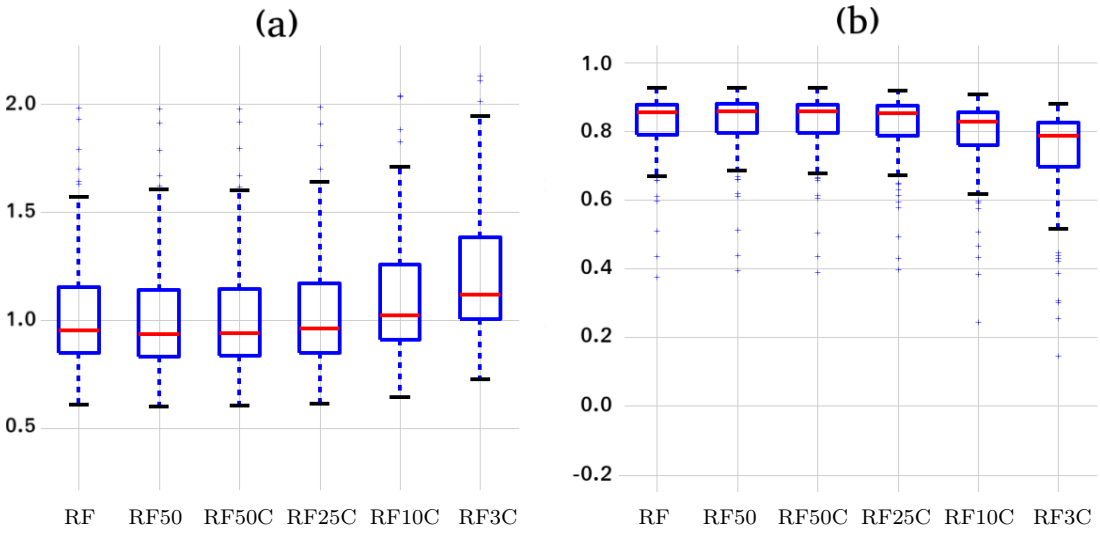
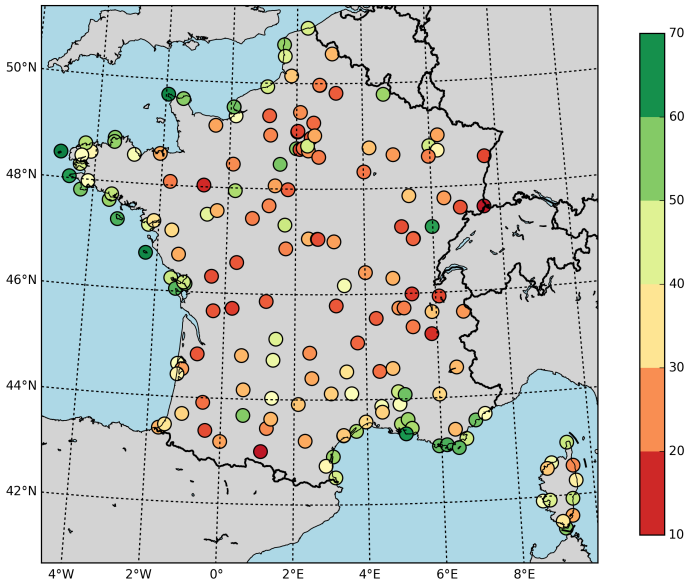


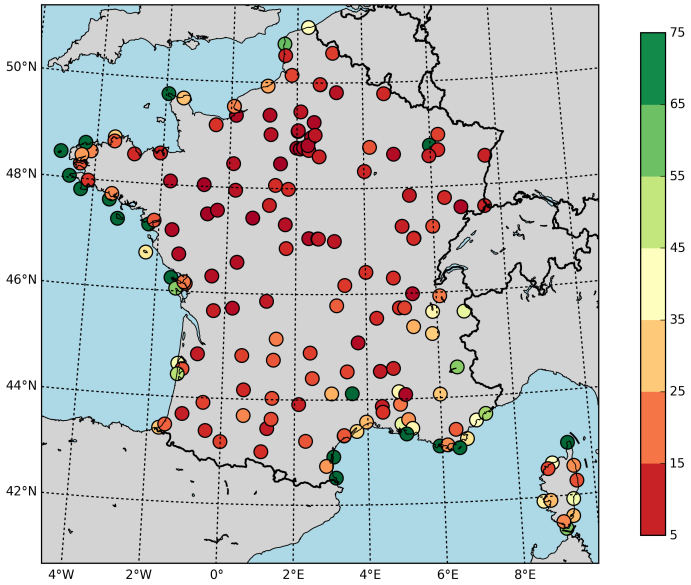
Figure V.3.8 – (a) RMSE and (b) correlation coefficient of the model using random forests with different lists of explanatory variables for all the stations in France

Regarding the spatial distribution, the percentage change of the RMSE and the correlation coefficient of the RF25C model with respect to the ECMWF analyses is shown in Figure V.3.9. From Figure V.3.9 (a) it can be noticed that the RMSE of inland stations in the north of France is reduced by 30% on average. The RMSE for stations in the inland south decreases on average by 40%. The largest decreases of up to 80% are found for coastal stations in the west, the south and Corsica. Results for the correlation coefficient in Figure V.3.9 (b) confirm these conclusions. The correlation coefficient for stations in the inland north and inland south show an increase of 15% and 22%, respectively, and for coastal stations the correlation coefficient shows an increase of 60%.

In conclusion, the full random forest model used an unnecessarily long list of explanatory variables, which is not detrimental to its performance. The RF50 model with 50 explanatory variables specific to each station performs slightly better. The RF50C model with 50 common explanatory variables performs as well as the RF50 model but is generic in nature. The RF25C model is simple and robust with just 25 important explanatory variables and is comparable to the full model in performance. Hence, the RF25C model is a good compromise between performance and simplicity. It is instructive to analyze the list of 25 explanatory variables retained.



(a)



(b)

Figure V.3.9 – Percentage change from ECMWF analyses to RF25C model for (a) RMSE and (b) correlation coefficient

V.3.4.2 List of significant variables

The following are the most significant explanatory variables.

Top 9 list:

- The two horizontal wind components and the wind speed at 10m and 100m above ground (6 variables),
- The wind shear between 10m and 100m (1 variable),
- The two horizontal wind components at 500hPa (2 variables).

Top 25 list (in addition to the previous 9 variables):

- The horizontal wind components at 850 hPa and 925 hPa (4 variables),
- The horizontal components of the gradients of geopotential at 925hPa, 850hPa and 500hPa (6 variables),
- The horizontal components of the gradients of mean sea level pressure (2 variables),
- Skin temperature,
- Temperature at 2m,
- The boundary-layer height,
- One of the components of the gradient of surface pressure.

The top 25 list includes 9 variables corresponding to components of the gradient of geopotential or pressure. These variables are strongly correlated to the wind components because geostrophic balance is an excellent approximation at these latitudes. Yet, it is noteworthy that the RF10C model, which does not include these explanatory variables, did not perform as well as the RF25C model. Finally, it is interesting that few variables describing temperature and boundary-layer parameters appear in the top 25 list. As seen from Figure V.3.7, variables ranked between 25 and 35 mainly describe the temperature and the boundary layer, and the comparable performances of the RF25C and the RF50C models suggest that their contribution is minor. To conclude, it is striking that the most relevant variables almost all describe the flow (wind components, wind speed, geopotential gradient). It was expected that, given the importance of thermal and convective processes in the boundary layer, the inclusion of information on the temperature and stratification would be helpful. It turns out that this information does not significantly modify the performance of the models. A possible explanation is that the numerical weather prediction model of the ECMWF already describes rather well the surface flow, and the vertical shear in the first 100m already encompasses the relevant information on the stratification and mixing in the boundary layer.

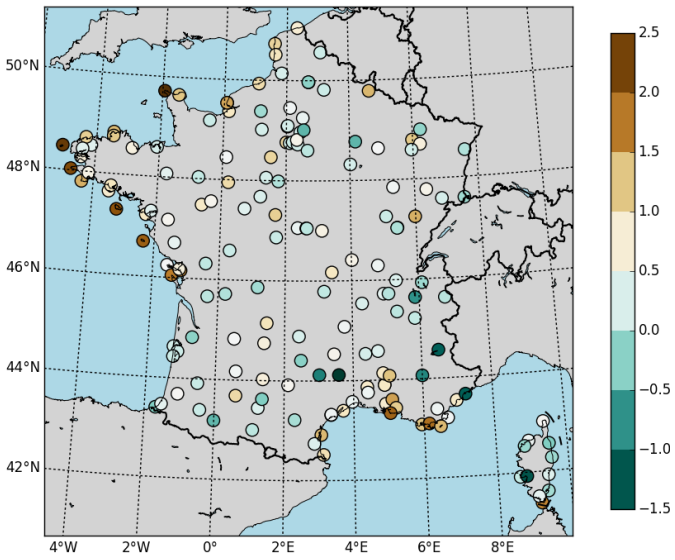
V.3.4.3 Bias

Let us assess the performance of the machine learning models in terms of reduction of the bias present in the ECMWF analyses. The bias of the surface wind speed from the ECMWF analyses is shown in Figure V.3.10 (a). The locations of the largest RMSE (see Figure V.3.1) coincide with those of the largest bias. There is mostly a positive bias over coastal stations. There are also a few inland stations displaying a significant negative bias, associated with unusually large RMSE. Over the whole set of stations, the bias is on average $0.47 m.s^{-1}$. For individual stations, the biases range from -1.61 to $2.50 m.s^{-1}$. As expected, the machine learning models prove very efficient at removing the bias. As an illustration, Figure V.3.10 (b) displays the bias for the RF25C model, which is uniformly negligible, the average bias being $0.004 m.s^{-1}$. The bias for individual stations is very weak, ranging for all but two stations from -0.01 to $0.02 m.s^{-1}$. The two outliers are stations in the Alps, with biases between $0.02 m.s^{-1}$ and $0.04 m.s^{-1}$.

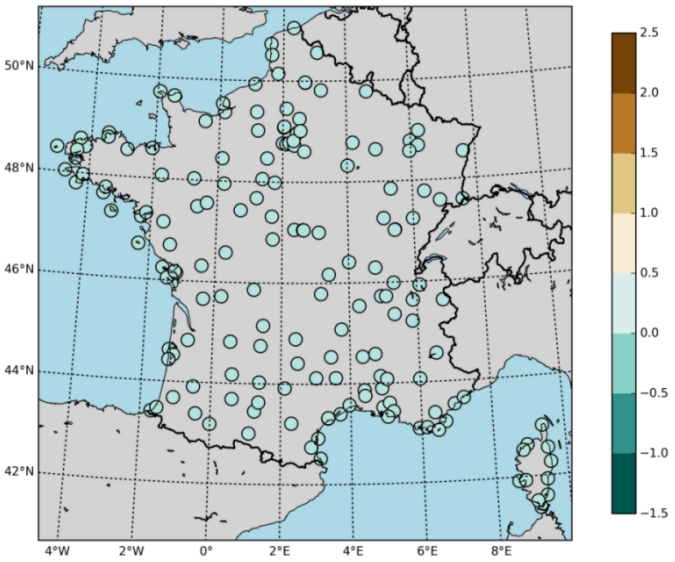
V.3.5 Exploratory forecast test

This section intends to improve the forecasts of the surface wind speed from the outputs of the ECMWF model, as investigated at the end of Chapter V.2. Note that this provides only a lower bound on the potential accuracy of forecasts, because the machine learning models are not trained on the forecasts.

The ECMWF high-resolution global-forecast model is run twice a day at a base time of 00:00 and 12:00 UTC and each run forecasts the weather up to 10 days. We limit this study to the station Le Havre-Octeville, already used in Section V.3.3.1. Appropriately, the ECMWF forecast data were retrieved at lead times of 0h, 3h, 6h, 12h, and 24h, where 0h corresponds to that of the analyses. The machine learning models used to reconstruct the wind speed from these forecasts are the same as described and used previously, they have been trained using model outputs from the analyses.Figure V.3.11 shows the RMSE and the correlation coefficient of the wind speed at 10m from ECMWF forecasts at various lead times for the station Le Havre-Octeville. As seen previously, the RMSE is rather large (nearly $2.5 m.s^{-1}$), and it remains fairly constant over the first 24h of the forecast. The RF25C model trained on the analyses is applied to the outputs of the ECMWF forecasts at lead times from 3 to 24h. The RMSE and correlation coefficient of the reconstructed wind speed are shown in Figure V.3.12. The RMSE is dramatically reduced (down to $1.2 m.s^{-1}$ or less): the average reduction in RMSE over all the lead times is about 55%, and the



(a)



(b)

Figure V.3.10 – Bias ($m.s^{-1}$) in the surface wind speed for (a) the ECMWF analyses and (b) the estimated surface wind speed using the RF25C model

average increase in correlation coefficient is about 21%. Hence, the improvements thanks to machine learning carry over to forecasts. As already mentioned in the previous chapter, The results could be further improved by applying a model that is trained separately for each lead time directly on the forecasts.

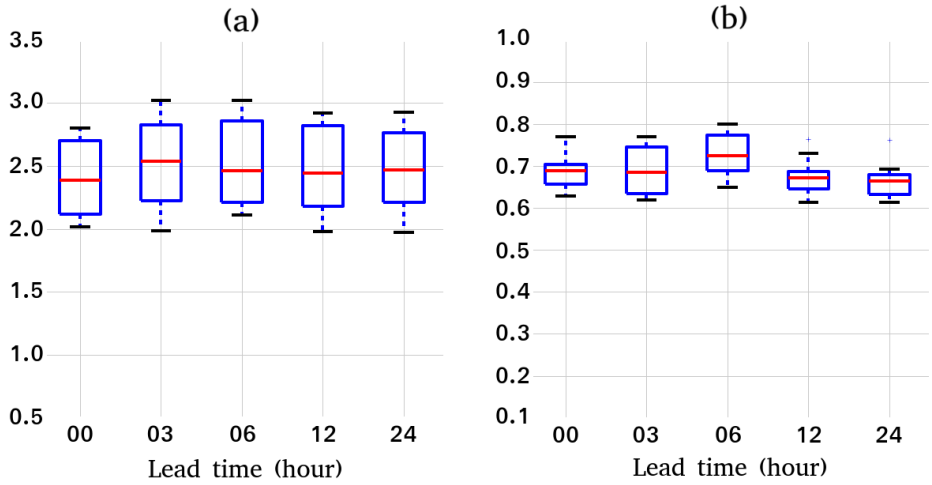


Figure V.3.11 – (a) RMSE ($m.s^{-1}$) and (b) correlation coefficient of the ECMWF forecast wind speed at 10m at various time horizons at Le Havre-Octeville.

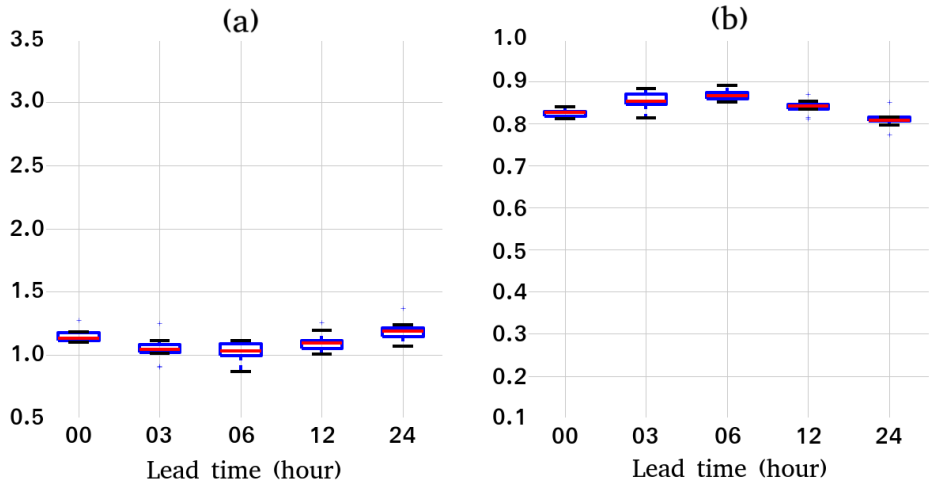


Figure V.3.12 – (a) RMSE ($m.s^{-1}$) and (b) correlation coefficient of wind speed at 10m from the RF25C model at various time horizons at Le Havre-Octeville

Chapter V.4

Some other applied collaborations

This chapter briefly presents three other interdisciplinary collaborations, mainly related to supervision activities.

V.4.1 Computer science for cancer pharmacology

The first year after having completed my PhD, I had the opportunity to participate in the supervision of the statistical part of the Computer Science PhD thesis of Maya Alsheh-Ali, on topics applied to biology, at Université Paris Descartes. More specifically, Maya was working with pharmacologists and she studied the structure of tumoral images. After her PhD defense, Maya joint her family in Sweden, to hold a position at Karolinska Institutet.

In Alsheh-Ali et al. (2013), we propose an automatic method to quantitatively describe the spatial organization governing populations of biological objects, such as cells, which exist in stationary histology images. The goal is to be able to compare different tumoral models in order to evaluate potential therapies. We compare two animal models of colorectal cancer thanks to spatial statistics tools and a functional analysis of variance method. We obtain that there are significant differences in the considered statistics depending on cancer model, and on the day after tumor implant.

V.4.2 Statistical tools for phonetics

Nicolas Ballier, professor at the English Department at Université de Paris, contacted me for a collaboration "statistics and linguistics", and in particular, proposed me to co-supervise the PhD thesis of Maëlle Amand in the frame of a call for projects "dual-culture doctoral contracts" (contrats doctoraux double culture de l'Université Sorbonne Paris Cité). The goal of the thesis was to study audio data, the so-called "NECTE corpus", from a sociological and quantitative point of view, which required a statistical learning analysis. Now, Maëlle is assistant professor in "Phonetics, phonology and English linguistics" at Limoges.

More precisely, the goal was to solve a vowel classification issue from recordings from 1960s. A part of the recordings has indeed been annotated by an "expert", meaning that each vowel in this part has a label, assigned by ear. The question was whether we are able to label the rest of the data using modern machine learning techniques.

During the PhD, the statistical aspects finally deviated from the initial classification question, and focused more on data analysis applied to sociological studies, and associated visualization methods. Moreover, beside her main PhD subject Maëlle was also interested, through collaborations with other students, in statistical testing frameworks applied to studies regarding English-Language Learners.

The original statistical learning question was explored further recently, during the end-of-study internship of Clarisse Thiard, third-year student at École Nationale Supérieure Agronomique de Rennes (Thiard, 2021).

V.4.3 Modeling for astrophysics

Recently, I had the opportunity to collaborate with a team from the French Alternative Energies and Atomic Energy Commission (Commissariat à l'énergie atomique et aux énergies alternatives, CEA) on a modeling problem in the field of radiation belt physics.

The general purpose is the study of the Van Allen radiation belts, two zones of energetic charged particles surrounding Earth, coming from solar wind or cosmic rays. The interaction in the inner magnetosphere between whistler-mode waves and energetic radiation belt electrons results in the so-called pitch-angle diffusion process causing electron precipitation onto the atmosphere. Here, the pitch angle

denotes the angle between the particle's velocity vector and the local magnetic field vector. The process can be described by a pitch-angle diffusion or Fokker–Planck coefficient. The computation of this parameter, relying on quasi-linear theory, is extremely time-consuming as soon as it is performed at high temporal resolution from satellite measurements of ambient wave and plasma properties. However, the knowledge of pitch-angle diffusion coefficients is important for understanding the behavior of the Van Allen radiation belts. Using various interpolation and machine learning strategies, our goal in Kluth et al. (2021) consists in building a global model of pitch-angle diffusion coefficients for storm conditions based on data corresponding to storms observed by the NASA Van Allen Probes, two robotic spacecraft dedicated to the study of the Van Allen radiation belts.

Part VI

Perspectives

Chapitre VI.1

Quelques prolongements

Pour conclure, nous présentons quelques directions de recherche, sur lesquelles nous prévoyons de travailler, à plus ou moins long terme.

VI.1.1 Courbes principales

VI.1.1.1 Propriétés de régularité supplémentaires

Une première question qui se pose naturellement est celle de la relation précise entre la version contrainte de la définition de courbe principale, étudiée ici, et sa version variationnelle, c'est-à-dire le problème dans lequel on considère un critère pénalisé au lieu d'utiliser une contrainte directe de longueur. C'est alors la valeur de la constante multiplicative devant la pénalité qui induit la longueur. L'équivalence des deux problèmes serait par exemple appréciable d'un point de vue algorithmique. Sinon, s'ils sont distincts, cela signifie que chaque question sur les courbes principales pourrait en fait être traitée dans le cadre de chacune des deux formulations du problème.

Dans l'article Delattre et Fischer (2020) (Chapitre II.3), nous exploitons la condition du premier ordre, condition de point critique. Il serait intéressant de pouvoir également tirer profit de l'information contenue dans la condition du second ordre. De plus, nous avons montré qu'une courbe principale est toujours injective en dimension $d = 2$: une question ouverte concerne ce qui se passe en dimension $d > 2$.

Un autre point consiste à se demander si l'on peut obtenir certaines informa-

tions sur la distribution de $\hat{t}$. En outre, les courbes principales pour les distributions classiques, comme la distribution uniforme par exemple, ont une forme particulière, rappelant un labyrinthe, avec des couloirs de largeur apparemment égale, comme illustré par la Figure VI.1.1. Là encore, tout résultat théorique serait très instructif.

L'analogie avec la quantification vectorielle, basée sur un critère assez similaire, pose également d'autres questions. Par exemple, en considérant la situation où la longueur tend vers l'infini, il serait intéressant, d'un point de vue théorique, de comprendre s'il est possible de démontrer un résultat similaire au théorème de Zador.

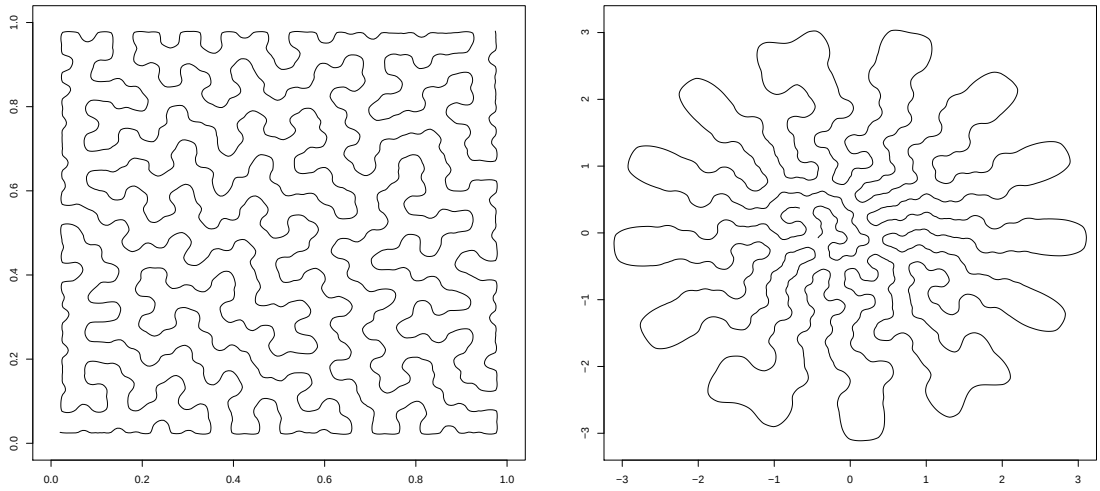


FIGURE VI.1.1 – Deux exemples de courbe principale avec contrainte de longueur : (a) Loi uniforme sur le carré $[0,1]^2$ (b) Loi normale standard

VI.1.1.2 Vitesse de convergence en estimation avec petit bruit

À la suite de l'article Delattre et Fischer (2021) (Chapitre II.4), où nous établissons, dans un modèle additif, la convergence d'une suite de courbes empiriques optimales vers la courbe générative sous des hypothèses faibles sur le bruit et la courbe, nous nous intéressons à un cadre où une certaine régularité sur la courbe inconnue et des conditions appropriées sur le bruit permettent d'obtenir une vitesse de convergence. On observe des données distribuées selon le modèle suivant :

$$X_i = g(U_i) + \varepsilon_i, \quad i = 1, \dots, n,$$

où la courbe $g : [0,1] \to \mathbb{R}^d$ est inconnue, les ε_i sont des variables aléatoires indépendantes, telles que $E[|\varepsilon_i|] \leq m$ et $Var(|\varepsilon_i|) \leq \sigma^2$ pour tout $i = 1, \ldots, n$, et les U_i, $i = 1, \ldots, n$, sont des variables aléatoires indépendantes de loi $\mu_i \geq c\lambda$ sur $[0,1]$. On suppose que g est rectifiable et $|g(t)| = \mathscr{L}(g)$ dt-a.e. De plus, on fait l'hypothèse que $\text{reach}(\text{Im}g) \geq r > 0$. Pour rappel, le *reach* de g est le rayon maximal d'une boule que l'on peut faire rouler le long de la courbe (Federer, 1959). Le but est de construire dans ce contexte une suite de courbes $\hat{f}_n$, telle que $\text{Im}\hat{f}_n$ converge vers $\text{Im}g$ en probabilité, en distance de Hausdorff, avec la longueur $\mathscr{L}(\hat{f}_n)$ asymptotiquement égale à $\mathscr{L}(g)$, et de calculer la vitesse de convergence associée. Il s'agit d'une convergence au sens de la double asymptotique où la taille de l'échantillon n est grande et l'erreur moyenne m est petite.

VI.1.1.3 Vitesse de convergence en apprentissage statistique

Nous avons obtenu une vitesse de convergence améliorée pour les courbes principales en apprentissage statistique. Néanmoins, nous ne savons pas si cette vitesse peut encore être améliorée. Une suite naturelle serait la recherche d'une borne inférieure minimax.

VI.1.1.4 Point de vue algorithmique

Une autre direction de recherche relative à notre contexte de courbes principales consiste à étudier les performances d'un algorithme permettant de calculer des approximations des estimateurs que nous avons définis. Pour l'instant, les résultats que nous avons obtenus sur des exemples, en utilisant un algorithme de descente de gradient, semblent pertinents et prometteurs.

Notons que l'implémentation de stratégies d'initialisation astucieuses représente une question importante en soi. En particulier, il faut éviter la formation de brins inutiles, sur lesquels aucune observation ne se projette.

À partir du cadre considéré dans Delattre et Fischer (2021) (Chapitre II.4), on fixe $V(x) = x^2$, et on considère la version variationnelle du critère Δ_n : le critère pénalisé est donné par

$$\Delta_n(f) + \lambda \int_0^1 |f'(t)|^2 dt.$$

Une courbe optimale basée sur cette approche variationnelle peut être approximée en

pratique par une ligne polygonale avec $k \geq 1$ sommets $v_1, \ldots, v_k \in \mathbb{R}^d$, en minimisant

$$\frac{1}{n}\sum_{i=1}^{n}\left[\min_{1\leq j\leq k}|X_i - v_j|^2\right] + \lambda\,(k-1)\sum_{j=1}^{k-1}|v_{j+1} - v_j|^2.$$

Observons que les seules projections sur la ligne polygonale intervenant dans le critère sont les projections sur les sommets $v_1, \ldots, v_k$. Le calcul de $v_1, \ldots, v_k$ peut alors être effectué par un algorithme de descente de gradient stochastique. Pour la sélection de la longueur, nous faisons croître le coefficient λ jusqu'à ce qu'un minimum soit atteint pour le critère global utilisé pour définir la longueur $\hat{L}_n$ (voir Theorem II.4.3.1). On peut noter que la définition de $\hat{L}_n$ est beaucoup plus simple, par exemple, lorsque $c = 1$. Dans ce cas particulier, $\mathcal{M}_c$ contient une seule loi de probabilité, la loi uniforme, de sorte que la distance à évaluer est en fait la distance à la loi uniforme. Dans ce cas, on peut choisir pour $\mathcal{D}$ la distance de Wasserstein L^1, ce qui correspond simplement à la distance L^1 entre fonctions de répartition.

VI.1.1.5 Autres objets géométriques en statistique

Nous souhaitons également étudier d'autres « objets principaux », tels que les graphes ou les surfaces, mais aussi aborder diverses questions reliant la statistique et la géométrie, un sujet qui sera bientôt particulièrement mis en lumière, lors du Trimestre Thématique« Geometry and Statistics in Data Sciences », qui se tiendra à l'Institut Henri Poincaré à l'automne 2022.

Par exemple, nous chercherons à définir des contraintes appropriées permettant d'étendre les résultats de Delattre et Fischer (2020) aux surfaces, et nous considérerons l'estimation de surfaces. D'une manière générale, la perspective de nouvelles collaborations avec différents collègues intéressés par les domaines de la statistique et de la géométrie est extrêmement motivante.

VI.1.2 Classification non supervisée, déconvolution

VI.1.2.1 Clustering spectral

Une collaboration sur le *clustering* spectral a débuté avec Ilaria Giulini (LPSM, Université de Paris) et Mathilde Mougeot. Notre objectif est double. La première préoccupation consiste à valoriser en pratique tout le potentiel de l'approche introduite

dans Giulini (2016), avec choix adaptatif du nombre final de classes, cette procédure ayant été principalement étudiée d'un point de vue théorique jusqu'à présent. Un deuxième projet consiste à employer l'algorithme de *clustering* spectral obtenu pour effectuer une agrégation de partitions dans plusieurs contextes. En effet, le *clustering* spectral, qui exploite les éléments propres d'une matrice laplacienne construite à partir d'une matrice de distances, se prête très bien à l'agrégation : il s'agit de construire une matrice de similarité basée sur le nombre de fois où des couples d'observations ont été classés dans un même groupe par les méthodes de partitionnement préliminaires.

L'objectif peut être d'obtenir une moyenne, un résumé de l'information, ce qui est pertinent dans des situations pratiques où plusieurs partitions différentes apparaissent naturellement, ou de conférer une certaine robustesse au partitionnement, puisqu'une partition construite en combinant différentes distances aura la propriété d'être relativement robuste par rapport à la mesure de proximité.

De plus, dans un contexte de grande dimension, appliquer une certaine méthode de partitionnement à des projections des données est une idée naturelle. Il est alors intéressant d'étudier, d'un point de vue théorique et pratique, l'apport d'une agrégation de partitions obtenues pour plusieurs dimensions de projection relativement petites.

VI.1.2.2 Déconvolution Wasserstein

Pour compléter les résultats de déconvolution de l'article Dedecker et al. (2015) (Chapitre III.4), pour des lois de probabilité ordinairement régulières, nous nous intéressons au cas de la dimension $d > 1$. Une extension de nos résultats pourrait s'appuyer sur une inégalité obtenue par Fournier et Guillin (2015), qui semble appropriée pour obtenir des vitesses de convergence minimax en toute dimension.

VI.1.3 Agrégation

VI.1.3.1 Données fonctionnelles et de grande dimension

Dans de nombreuses applications, les observations à traiter sont de grande dimension ou fonctionnelles. Nous prévoyons d'étudier l'adaptation de nos stratégies de combinaison d'estimateurs à ces cadres. Une première tentative pour aborder le

cas de la grande dimension, utilisant des projections aléatoires, est proposée dans la thèse de doctorat de Sothea Has, co-encadré avec Mathilde Mougeot (Has, 2021). Plusieurs idées, reposant sur l'introduction de distances plus adaptées dans les poids, ou sur une sélection de variables préliminaire, basée par exemple sur une analyse des corrélations entre la variable cible et les différentes variables explicatives, méritent d'être explorées. Par ailleurs, une collaboration a débuté avec Pamela Llop (Universidad Nacional del Litoral, Argentine) pour étendre notre procédure d'agrégation aux données fonctionnelles.

VI.1.3.2 Contexte de données massives

Nous réfléchissons également à l'adaptation de notre méthode dans une situation de données massives. En effet, être obligé, pour construire les poids, de vérifier des conditions impliquant toutes les observations va poser problème lorsque la taille de l'échantillon devient très grande. À cet égard, des stratégies de sous-échantillonnage peuvent être envisagées, comme initié lors du stage de Master 1 d'Adam Mourjane, élève de deuxième année à l'École Normale Supérieure (Mourjane, 2021).

VI.1.4 Collaboration en physique pour le climat

La collaboration avec Riwal Plougonven va se poursuivre, notamment à travers notre projet "Guider les paramétrisations des modèles de climat par le triptyque observations - IA - simulations", qui vient d'être accepté dans le cadre d'un appel à projets de l'Institut des Mathématiques pour la Planète Terre (IMPT). En bref, notre objectif est d'améliorer la description des ondes de gravité internes dans les modèles de climat en introduisant une méthodologie d'apprentissage statistique comme dans Alonzo et al. (2018); Goutham et al. (2021) (Chapitres V.2 et V.3). En effet, dans l'atmosphère, les ondes de gravité internes sont des mouvements à très petite échelle qui sont difficiles à intégrer dans les modèles climatiques. Cependant, elles jouent un rôle crucial dans la circulation atmosphérique au-dessus de 15-20 km. Dans le cadre de notre projet, nous disposerons d'un jeu de données précieux dans ce contexte, provenant de mesures par ballons pressurisés stratosphériques, et nous utiliserons des méthodes statistiques pour estimer les ondes de gravité grâce aux observations des ballons, à partir de la connaissance de la circulation à grande échelle, telle qu'elle serait décrite par un modèle climatique.

Chapter VI.2

Some extensions

As a conclusion to this manuscript, we present some directions for future research, on which we plan to work, in the long or short term.

VI.2.1 Principal curves

VI.2.1.1 Additional regularity properties

A issue of interest is to learn about the relationship between the constrained version of the principal curve definition, studied in this document, and its variational counterpart, that is the problem in which, instead of using a direct length-constraint, a penalized criterion is considered, with the multiplicative weight on the length penalty as smoothness parameter. For instance, if the two problems turn out to be equivalent, it may be nice from an algorithmic point of view. Otherwise, if they are distinct, it means that each question about principal curves could actually be addressed under each of the two problem formulations.

In Delattre and Fischer (2020) (Chapter II.3), we exploit the first order condition of being a critical point. It would be interesting to be able also to take advantage of the information contained in the second order condition. Moreover, we have shown that a two-dimensional principal curve is always injective. An open question concerns what happens in dimension $d > 2$.

Another problem is whether some information can be obtained about the distribution of $\hat{t}$. Furthermore, principal curves for classical distributions, such as the

uniform one for example, have a particular shape, reminding a labyrinth, with corridors of seemingly equal width, as illustrated in Figure VI.2.1. Again, any theoretical result would be very informative.

The analogy with vector quantization, based on a quite similar criterion, also brings further questions. For instance, considering the situation where the length tends to infinity, it would be interesting, from a theoretical point of view, to understand if a result of the kind of Zador's Theorem can be shown.

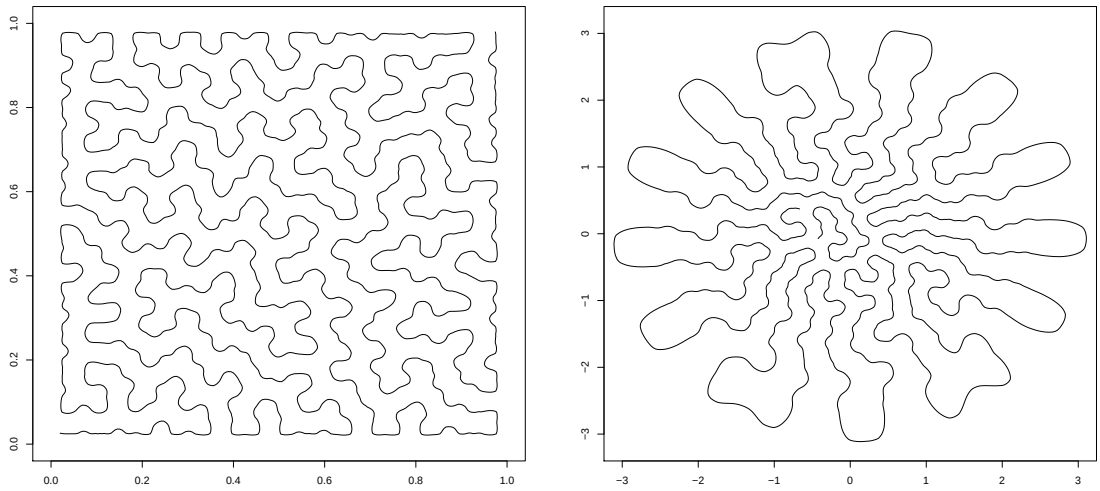


Figure VI.2.1 – Two examples of principal curves with length constraint: (a) Uniform distribution over the square $[0,1]^2$ (b) Standard Gaussian distribution

VI.2.1.2 Rates of convergence in estimation under small noise

As a continuation of the article Delattre and Fischer (2021) (Chapter II.4), where we establish, in an additive model, the convergence of a sequence of empirically optimal curves to the generative curve under weak assumptions on the noise and on the curve, we are interested in a framework where some regularity on the unknown curve and appropriate conditions on the noise allow to derive rates of convergence . We observe data distributed according to the following model:

$$X_i = g(U_i) + \varepsilon_i, \quad i = 1, \dots, n,$$

where the curve $g : [0,1] \to \mathbb{R}^d$ is unknown, the ε_i are independent random variables such that $E[|\varepsilon_i|] \leq m$ and $Var(|\varepsilon_i|) \leq \sigma^2$ for every $i = 1, \ldots, n$, and the U_i, $i = 1, \ldots, n$, are independent random variables with distribution $\mu_i \geq c\lambda$ on $[0,1]$. We assume that g is rectifiable and $|g(t)| = \mathscr{L}(g)$ dt-a.e. Moreover, we suppose that $\text{reach}(\text{Im}g) \geq r > 0$. Recall that the reach of g is the maximal radius of a ball rolling along the curve (Federer, 1959). The aim is to build in this context a sequence of curves $\hat{f}_n$ such that $\text{Im}\hat{f}_n$ converges in probability to $\text{Im}g$, in Hausdorff distance, with the length $\mathscr{L}(\hat{f}_n)$ asymptotically equal to $\mathscr{L}(g)$, and to compute the rate of convergence. Here, convergence is to be understood in the sense of the double asymptotic where the sample size n is large and the expected error m is small.

VI.2.1.3 Rate of convergence in statistical learning

We have obtained an improved rate of convergence for principal curves in statistical learning. Nevertheless, we do not know whether this rate can be further improved. A natural continuation is to search for a minimax lower bound.

VI.2.1.4 Computational point of view

Another research direction about our principal curve context is to study the performance of an algorithm allowing to compute estimators. For now, the results obtained on examples using a gradient descent algorithm seem relevant and promising.

Note that designing smart initialization strategies is a important question in itself. In particular, the formation of unnecessary strands, without observations projecting onto it, should be avoided.

Starting from the framework in Delattre and Fischer (2021) (Chapter II.4), set $V(x) = x^2$, and consider the variational version of the criterion Δ_n: the penalized criterion is given by

$$\Delta_n(f) + \lambda \int_0^1 |f'(t)|^2 dt.$$

Based on this variational approach, an optimal curve may be approximated in practice by a polygonal line with $k \geq 1$ vertices $v_1, \ldots, v_k \in \mathbb{R}^d$, by minimizing

$$\frac{1}{n} \sum_{i=1}^{n} \left[\min_{1 \leq j \leq k} |X_i - v_j|^2 \right] + \lambda (k-1) \sum_{j=1}^{k-1} |v_{j+1} - v_j|^2.$$

Observe that the only projections onto the polygonal line involved in the criterion are projections onto the vertices $v_1, \ldots, v_k$. The computation of $v_1, \ldots, v_k$ may then be performed by a stochastic gradient descent algorithm. For the length selection, we let the weight λ grow until a minimum is reached for the global criterion used for defining the length $\hat{L}_n$ (see Theorem II.4.3.1). Note that the definition of $\hat{L}_n$ is much simpler, for instance, if $c = 1$. In this particular case, $\mathcal{M}_c$ contains a single distribution, the uniform one, so that the distance to be assessed is the distance to the uniform distribution. In this case, the L^1 Wasserstein metric may be chosen for the distance $\mathcal{D}$, which simply corresponds to the L^1 distance between cumulative distribution functions.

VI.2.1.5 Other geometric objects in statistics

We will also focus on further principal objects, such as graphs or surface, as well as on various questions linking statistics and geometry, a topic which will be in the spotlight soon, at the Institut Henri Poincaré Thematic Quarter "Geometry and Statistics in Data Sciences" in fall 2022.

For instance, we will try to define appropriate constraints allowing to extend the results in Delattre and Fischer (2020) to surfaces, and we will investigate estimation of surfaces. More generally, the possibility of starting new collaborations with different colleagues interested in the fields of statistics and geometry is very motivating.

VI.2.2 Clustering, deconvolution

VI.2.2.1 Spectral clustering

A collaboration about spectral clustering has started with Ilaria Giulini (LPSM, Université de Paris) and Mathilde Mougeot. Our goal is twofold. The first concern consists in bringing out the full practical potential of the clustering approach introduced in Giulini (2016), with adaptive choice of the final number of clusters. Indeed, this procedure has mainly been studied from a theoretical point of view up to now. A second project is to employ the obtained spectral clustering algorithm to perform clustering aggregation in several contexts. Indeed, spectral clustering, which exploits the eigenstructure of a Laplacian matrix built from a distance matrix, lends itself very well to aggregation via the construction of a similarity matrix encoding how often pairs of points have been clustered together by preliminary methods.

The aim may be to obtain an average, a summary of the information, which is relevant in practical situations in which several instances of clustering naturally emerge, or to enhance a certain robustness of the partitioning, since a partition built by combining different distances will have the property of being relatively robust with respect to the proximity measure.

Furthermore, in a high-dimensional context, applying a certain clustering method to projections of the data is a natural idea. It is then worth studying, from a theoretical and from a computational point of view, the benefit of aggregating clustering results obtained for several small dimension projections.

VI.2.2.2 Wasserstein deconvolution

To complete the deconvolution results in Dedecker et al. (2015) (Chapter III.4), for ordinary smooth distributions, we are interested in the dimension $d > 1$ situation. An extension of our results could be based on an inequality obtained by Fournier and Guillin (2015), which seems appropriate for deriving minimax convergence rates in any dimension.

VI.2.3 Aggregation

VI.2.3.1 High-dimensional and functional data

In many applications the observations to be processed are high-dimensional or functional data. We plan to study adaptation of our combining strategies to these frameworks. A first attempt to address the high-dimensional case, using random projections, is proposed in the PhD of Sothea Has, co-supervised with Mathilde Mougeot (Has, 2021). Several ideas, relying on the introduction of well-suited distances in the weights, or on a preliminary feature selection, based for example on an analysis of the correlations between the output and the different explanatory variables, seem worth exploring. Besides, a collaboration has started with Pamela Llop (Universidad Nacional del Litoral, Argentina) to extend our aggregation procedure to functional data.

VI.2.3.2 Big data framework

We are also thinking about adapting our method to the big data situation. Indeed, checking conditions involving all data points to construct our weights could be intractable as soon as the sample size becomes very large. In this regard, sub-sampling strategies may be considered, as was initiated during the Master 1 internship of Adam Mourjane, second-year student at École Normale Supérieure (Mourjane, 2021).

VI.2.4 Collaboration in physics for climate

The collaboration with Riwal Plougonven will go on, in particular through our recent accepted project "Triptych observations - AI - simulations to guide the parameterizations of climate models" in the frame of the Institute of Mathematics for Planet Earth (*Guider les paramétrisations des modèles de climat par le triptyque observations - IA - simulations*, Institut des Mathématiques pour la Planète Terre, IMPT). In short, our purpose is to improve the description of internal gravity waves in climate models by introducing a statistical learning methodology as in Alonzo et al. (2018); Goutham et al. (2021) (Chapters V.2 and V.3). Indeed, in the atmosphere, internal gravity waves are very small-scale movements that are difficult to integrate into climate models. However, they play a crucial role in the atmospheric circulation above 15-20 km. In the frame of our project, we will have at hand a valuable data set in this context, coming from balloon measurements, and we will use statistical methods to estimate gravity waves thanks to the balloon observations, from the knowledge of the large-scale flow, as it would be described by a climate model.

Publications

B. Alonzo, R. Plougonven, M. Mougeot, A. Fischer, A. Dupre, and P. Drobinski. From numerical weather prediction outputs to accurate local surface wind speed: statistical modelling and forecasts. In *Renewable Energy: Forecasting and Risk Management*, Springer Proceedings in Mathematics & Statistics, 2018.

M. Alsheh-Ali, J. Seguin, A. Fischer, N. Mignet, L. Wendling, and T. Hurtut. Comparison of the spatial organization in colorectal tumors using second-order statistics and functional ANOVA. In *8th Int. Symp. Image Signal Process. Anal. (ISPA)*, pages 257–261, 2013.

B. Auder and A. Fischer. Projection-based curve clustering. *J. Stat. Comput. Simul.*, 82:1145–1168, 2012.

G. Biau and A. Fischer. Parameter selection for principal curves. *IEEE Trans. Inform. Theory*, 1924-1939:58, 2012.

G. Biau, A. Fischer, B. Guedj, and J. Malley. Cobra: A combined regression strategy. *J. Multivar. Anal.*, 146:18–28, 2016.

C. Brécheteau, A. Fischer, and C. Levrard. Robust Bregman clustering. *Ann. Statist.*, 49:1679–1701, 2021.

J. Dedecker, A. Fischer, and B. Michel. Improved rates for Wasserstein deconvolution with ordinary smooth error in dimension one. *Electron. J. Stat.*, 9:234–265, 2015.

S. Delattre and A. Fischer. On principal curves with a length constraint. *Ann. Inst. Henri Poincaré (B) Probab. Stat.*, 56:2108–2140, 2020.

S. Delattre and A. Fischer. Estimation via length-constrained generalized empirical principal curves under small noise, 2021.

A. Fischer. Quantization and clustering with Bregman divergences. *J. Multivar. Anal.*, 101:2207–2221, 2010.

A. Fischer. On the number of groups in clustering. *Stat. Probab. Lett.*, 81:1771–1781, 2011.

A. Fischer. Selecting the length of a principal curve within a Gaussian Model. *Electron. J. Stat.*, 7:342–363, 2013.

A. Fischer. Deux méthodes d'apprentissage non supervisé : synthèse sur la méthode des centres mobiles et présentation des courbes principales. *J. Soc. Fr. Stat.*, 155: 2–35, 2014.

A. Fischer. On two extensions of the vector quantization scheme. *J. Soc. Fr. Stat.*, 156:51–75, 2015.

A. Fischer and M. Mougeot. Aggregation using input-output trade-off. *J. Statist. Plann. Inference*, 200:1–19, 2019.

A. Fischer and D. Picard. On change-point estimation under Sobolev sparsity. *Electron. J. Stat.*, 14:1648–1689, 2020.

A. Fischer, L. Montuelle, M. Mougeot, and D. Picard. Statistical learning for wind power: a modeling and stability study towards forecasting. *Wind Energy*, 20: 2037–2047, 2017.

A. Fischer, S. Has, and M. Mougeot. A clusterwise supervised learning procedure based on the aggregation of distances. *J. Stat. Comput. Simul.*, 91:2307–2327, 2021.

N. Goutham, B. Alonzo, A. Dupré, R. Plougonven, R. Doctors, L. Liao, M. Mougeot, A. Fischer, and P. Drobinski. Using machine learning methods to improve surface wind from the outputs of a numerical weather prediction model. *Bound.-Layer Meteorol.*, 179:133–161, 2021.

G. Kluth, J. F. Ripoll, S. Has, A. Fischer, M. Mougeot, and E. Camporeale. Machine learning methods applied to the global modeling of event-driven pitch angle diffusion coefficients during high-speed streams, 2021.

Bibliography

E. Aamari and C. Levrard. Stability and minimax optimality of tangential Delaunay complexes for manifold reconstruction. *Discrete Comput. Geom.*, 59:923–971, 2018.

E. Aamari and C. Levrard. Nonasymptotic rates for manifold, tangent space and curvature estimation. *Ann. Statist.*, 47:177–204, 2019.

H. Akaike. Information theory and an extension of the maximum likelihood principle. In *Proceedings of the 2nd IEEE Int. Symp. Inf. Theory*, pages 267–281, 1973.

Y. Alber and D. Butnariu. Convergence of Bregman projection methods for solving consistent convex feasibility problems in reflexive Banach spaces. *J. Optim. Theory Appl.*, 92:33–61, 1997.

A. D. Alexandrov and Y. G. Reshetnyak. *General Theory of Irregular Curves*. Mathematics and its Applications. Kluwer Academic Publishers, Dordrecht, 1989.

S. Arlot and P. Massart. Data-driven calibration of penalties for least-squares regression. *J. Mach. Learn. Res.*, 10:245–279, 2009.

T. Arnold and L. Tilton. *Humanities Data in R: Exploring Networks, Geospatial Data, Images, and Text*. Quantitative Methods in the Humanities and Social Sciences. Springer International Publishing, 2015.

J.-Y. Audibert. Aggregated estimators and empirical complexity for least square regression. *Ann. Inst. Henri Poincaré (B) Probab. Stat.*, 40:685–736, 2004.

A. Aue, S. Hörmann, L. Horváth, M. Reimherr, et al. Break detection in the covariance structure of multivariate time series models. *Ann. Statist.*, 37:4046–4087, 2009.

J. A. Baars and C. F. Mass. Performance of National Weather Service forecasts compared to operational, consensus and weighted model output statistics. *Wea. Forecast.*, 20:1034–1047, 2005.

N. Balakrishnan and M. Mojirsheibani. A simple method for combining estimates to improve the overall error rates in classification. *Comput. Stat.*, pages 1–17, 2015.

A. Banerjee, X. Guo, and H. Wang. On the optimality of conditional expectation as a Bregman predictor. *IEEE Trans. Inform. Theory*, 51, 2005a.

A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh. Clustering with Bregman divergences. *J. Mach. Learn. Res.*, 6:1705–1749, 2005b.

J. D. Banfield and A. E. Raftery. Ice floe identification in satellite images using mathematical morphology and clustering about principal curves. *J. Am. Stat. Assoc*, 87:7–16, 1992.

A. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probab. Theory Relat. Fields*, 113:301–413, 1999.

P. L. Bartlett, S. Boucheron, and G. Lugosi. Model selection and error estimation. *Mach. Learn.*, 48:85–113, 2001.

M. Basseville and I. V. Nikiforov. *Detection of abrupt changes: Theory and applications*. Lecture Notes in Mathematics. Prentice Hall, 1993.

P. Bauer, A. Thorpe, and G. Brunet. The quiet revolution of numerical weather prediction. *Nature*, 525:47–55, 2015.

L. Birgé. Model selection via testing: An alternative to (penalized) maximum likelihood estimators. *Ann. Inst. Henri Poincaré (B) Probab. Stat.*, 42:273–325, 2006.

L. Birgé and P. Massart. Gaussian model selection. *J. Eur. Math.*, 3:203–268, 2001.

L. Birgé and P. Massart. Minimal penalties for Gaussian model selection. *Probab. Theory Relat. Fields*, 138:33–73, 2007.

L. Birgé and P. Massart. From model selection to adaptive estimation. In D. Pollard, E. Torgersen, and G. Yang, editors, *Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics*, pages 55–87. Springer, New York, 1997.

S. Bobkov and M. Ledoux. *One-dimensional empirical measures, order statistics and Kantorovich transport distances*. Memoirs of the American Mathematical Society. AMS, 2019.

L. M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Comput. Math. & Math. Phys.*, 7:200–217, 1967.

L. Breiman. Bagging predictors. *Mach. Learn.*, 24:123–140, 1996.

L. Breiman. Random forests. *Mach. Learn.*, 45:5–32, 2001.

L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and regression trees*. Wadsworth Statistics/Probability Series. Wadsworth Advanced Books and Software, Belmont, CA, 1984.

B. E. Brodsky and B. S. Darkhovsky. *Nonparametric methods in change-point problems*. Kluwer Academic Publisher, The Netherlands, 1993.

C. Brunsdon. Path estimation from GPS tracks. In *Proceedings of the 9th Int. Conf. GeoComput., National Centre for Geocomputation, National University of Ireland, Maynooth, Eire*, 2007.

A. Bücher, I. Kojadinovic, T. Rohmer, and J. Segers. Detecting changes in cross-sectional dependence in multivariate time series. *J. Multivar. Anal.*, 132:111–128, 2014.

F. Bunea, A. B. Tsybakov, and M. H. Wegkamp. Aggregation for Gaussian regression. *Ann. Statist.*, 35:1674–1697, 2007.

G. Buttazzo and E. Stepanov. Optimal transportation networks as free Dirichlet regions for the Monge-Kantorovich problem. *Ann. Sc. Norm. Sup. Pisa Cl. Sci.*, II(4):631–678, 2003.

G. Buttazzo, E. Oudet, and E. Stepanov. Optimal transportation problems with free Dirichlet regions. *Prog. Nonlinear Differ. Equ. Their Appl.*, 51:41–65, 2002.

G. Buttazzo, E. Mainini, and E. Stepanov. Stationary configurations for the average distance functional and related problems. *Control Cybern.*, 38:1107–1130, 2009.

C. Butucea and B. Tsybakov. Sharp optimality in density deconvolution with dominating bias. I. *Theory Probab. Appl.*, 52:24–39, 2008a.

C. Butucea and B. Tsybakov. Sharp optimality in density deconvolution with dominating bias. II. *Theory Probab. Appl.*, 52:237–249, 2008b.

O. Cadet, C. Harper, and M. Mougeot. Monitoring energy performance of compressors with an innovative auto-adaptive approach. In *Proceedings of the ISA Expo, Chicago*, 2005.

B. S. Caffo, C. M. Crainiceanu, L. Deng, and C. W. Hendrix. A case study in pharmacologic colon imaging using principal curves in single photon emission computed tomography. *J. Am. Stat. Assoc*, 103:1470–1480, 2008.

C. Caillerie and B. Michel. Model selection for simplicial approximation. *Found. Comput. Math.*, 11:707–731, 2011.

C. Caillerie, F. Chazal, J. Dedecker, and B. Michel. Deconvolution for the Wasserstein metric and geometric inference. *Electron. J. Stat.*, 5:1394–1423, 2011.

H. Cardot, P. Cenac, and P.-A. Zitt. Efficient and fast estimation of the geometric median in Hilbert spaces with an averaged stochastic gradient algorithm. *Bernoulli*, 19:18–43, 2013.

G. Carlsson. Topology and data. *Bull. Amer. Math. Soc.*, 46:255–308, 2009.

E. Carlstein, H. Müller, and D. Siegmund, editors. *Change-point problems*, volume 23. Institute of Mathematical Statistics, 1994.

C. Carrillo, A. F. O. Montaño, J. Cidrás, and E. Díaz-Dorado. Review of power curve modelling for wind turbines. *Renew. Sustain. Energy Rev.*, 21:572–581, 2013.

R. J. Carroll and P. Hall. Optimal rates of convergence for deconvolving a density. *J. Amer. Statist. Assoc.*, 83:1184–1186, 1988.

O. Catoni. *Statistical Learning Theory and Stochastic Optimization*. Lectures on Probability Theory and Statistics, École d'Été de Probabilités de Saint-Flour XXXI - 2001, Lecture Notes in Mathematics. Springer, 2004.

A. Cauchy. Mémoire sur la rectification des courbes et la quadrature des surfaces courbes, 1850.

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, New York, 2006.

K. Chang and J. Ghosh. Principal curves for nonlinear feature extraction and classification. *SPIE Applications of Artificial Neural Networks in Image Processing III*, 3307:120–129, 1998.

W. Y. Chang. A literature review of wind forecasting methods. *J. Power Energy Eng.*, 2:161–168, 2014.

F. Chazal, D. Cohen-Steiner, and Q. Mérigot. Geometric inference for probability measures. *Found. Comput. Math.*, 11:733–751, 2011.

F. Chazal, L. J. Guibas, S. Y. Oudot, and P. Skraba. Persistence-based clustering in Riemannian manifolds. *Journal of the ACM*, 60:1–38, 2013.

F. Chazal, B. T. Fasy, F. Lecci, B. Michel, A. Rinaldo, and L. Wasserman. Subsampling methods for persistent homology. In *Proceedings of the 32nd Int. Conf. Machine Learning (ICML)*, volume 37, pages 2143–2151, 2014.

H. Chen and N. Zhang. Graph-based change-point detection. *Ann. Statist.*, 43: 139–176, 2015.

H. Cho and P. Fryzlewicz. Multiple-change-point detection for high dimensional time series via sparsified binary segmentation. *J. R. Stat. Soc., B: Stat. Methodol.*, 77: 475–507, 2015.

A. Cholaquidis, R. Fraiman, J. Kalemkerian, and P. Llop. A nonlinear aggregation type classifier. *J. Multivar. Anal.*, 146:269–281, 2016.

P. J. Corkeron, P. Anthony, and R. Martin. Ranging and diving behaviour of two 'offshore' bottlenose dolphins, Tursiops sp., off eastern Australia. *J. Mar. Biolog. Assoc. U.K.*, 84:465–468, 2004.

A. Costa, A. Crespo, J. Navarro, G. Lizcano, H. Madsen, and E. Feitosa. A review on the young history of the wind power short-term prediction. *Ren. Sust. Energy Rev.*, 12:1725–1744, 2008.

I. Cribben and Y. Yu. Estimating whole-brain dynamics by using spectral clustering. *J. R. Stat. Soc., C: Appl. Stat*, 66:607–627, 2017.

M. W. Crofton. On the theory of local probability, applied to straight lines at random in a plane. *Philos. Trans. R. Soc.*, 158:181–199, 1868.

I. Csiszár. Generalized projections for non-negative functions. *Acta Math. Hungarica*, 68:161–185, 1995.

M. Csörgő and L. Horváth. *Limit theorems in change-point analysis*. Wiley, 1997.

J. A. Cuesta-Albertos, A. Gordaliza, and C. Matrán. Trimmed k-means: an attempt to robustify quantizers. *Ann. Statist.*, 25:553–576, 1997.

E. Cule. ridge: Ridge regression with automatic selection of the penalty parameter, 2012. URL `http://cran.r-project.org/package=ridge`.

A. Dalalyan and A. B. Tsybakov. Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity. *Mach. Learn.*, 72:39–61, 2008.

I. Dattner, A. Goldenshluger, and A. Juditsky. On deconvolution of distribution functions. *Ann. Statist.*, 39:2477–2501, 2011.

R. J. Davy, M. J. Woods, C. J. Russell, and P. A. Coppin. Statistical downscaling of wind variability from meteorological fields. *Bound.-Layer Meteorol.*, 175:161–175, 2010.

G. De'ath. Principal curves: a new technique for indirect and direct gradient analysis. *Ecology*, 80:2237–2253, 1999.

J. Dedecker and B. Michel. Minimax rates of convergence for Wasserstein deconvolution with supersmooth errors in any dimension. *J. Multivar. Anal.*, 122:278–291, 2013.

E. del Barrio, E. Giné, and C. Matrán. The central limit theorem for the Wasserstein distance between the empirical and the true distributions. *Ann. Probab.*, 27:1009–1971, 1999.

E. del Barrio, E. Giné, and F. Utzet. Asymptotics for L^2 functionals of the empirical quantile process, with applications to tests of fit based on weighted Wasserstein distances. *Bernoulli*, 11:131–189, 2005.

P. Delicado. Another look at principal curves and surfaces. *J. Multivar. Anal.*, 77: 84–116, 2001.

S. Dereich, M. Scheutzow, and R. Schottstedt. Constructive quantization: Approximation by empirical measures. *Ann. Inst. Henri Poincaré (B) Probab. Stat.*, 49: 1183–1203, 2013.

E. Devijver, Y. Goude, and J. M. Poggi. Clustering electricity consumers using high-dimensional regression mixture models. *Appl. Stoch. Models Bus. Ind.*, 36: 159–177, 2020.

A. Devis, N. P. van Lipzig, and M. Demuzere. A new statistical approach to downscale wind speed distribution at a site in northern Europe. *J. Geophys. Res. Atmos.*, 118:2272–2283, 2013.

L. Devroye and A. Krzyżak. An equivalence theorem for l^1 convergence of the kernel regression estimate. *J. Statist. Plann. Inference*, 23:71–82, 1989.

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition.* Appl. Math. Springer, New York, 1996.

Y. Drier, M. Sheffer, and E. Domany. Pathway-based personalized analysis of cancer. *PNAS*, 110:6388–6393, 2013.

H. Drucker, C. J. C. Burges, L. Kaufman, A. J. Smola, and V. Vapnik. Support vector regression machines. In M. C. Mozer, M. I. Jordan, and T. Petsche, editors, *Adv. Neural Inf. Process. Syst. 9*, pages 155–161. MIT Press, 1997.

T. Duchamp and W. Stuetzle. Extremal properties of principal curves in the plane. *Ann. Statist.*, 24:1511–1520, 1996a.

T. Duchamp and W. Stuetzle. Geometric properties of principal curves in the plane. In H. Rieder, editor, *Robust Statistics, Data Analysis, and Computer Intensive Methods: in Honor of Peter Huber's 60th Birthday*, volume 109 of *Lecture Notes in Statistics*, pages 135–152. Springer-Verlag, New York, 1996b.

R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification.* Wiley-Interscience, New York, 2000.

R. M. Dudley. The sizes of compact subsets of Hilbert space and continuity of Gaussian processes. *J. Funct. Anal.*, 1:290–330, 1967.

R. P. W. Duin and E. Pekalska. *The Dissimilarity Representation For Pattern Recognition: Foundations And Applications*, volume 64. World scientific, 2005.

Š. S. Èbralidze. Inequalities for the probabilities of large deviations in terms of pseudomoments. *Teor. Verojatnost. i Primenen.*, 16:760–765, 1971.

J. Einbeck, G. Tutz, and L. Evers. Local principal curves. *Statist. Comput.*, 15: 301–313, 2005a.

J. Einbeck, G. Tutz, and L. Evers. Exploring multivariate data structures with local principal curves. In C. Weihs and W. Gaul, editors, *Classification - The Ubiquitous Challenge, Proceedings of the 28th Ann. Conf. of the Gesellschaft für Klassifikation, University of Dortmund*, Studies in Classification, Data Analysis, and Knowledge Organization, pages 256–263. Springer, Berlin, Heidelberg, 2005b.

F. Enikeeva and Z. Harchaoui. High-dimensional change-point detection with sparse alternatives. *Ann. Statist.*, 47:2051–2079, 2019.

S. Evert. A simple LNRE model for random character sequences. In *7èmes Journées Internationales d'Analyse Statistique des Données Textuelles (Louvain-la-Neuve*, pages 411–422, 2004.

J. Fan. On the optimal rates of convergence for nonparametric deconvolution problems. *Ann. Statist.*, 19:1257–1272, 1991a.

J. Fan. Global behavior of deconvolution kernel estimates. *Statist. Sinica*, 2:541–551, 1991b.

J. Fan. Adaptively local one-dimensional subproblems with application to a deconvolution problem. *Ann. Statist.*, 21:600–610, 1993.

H. Federer. Curvature measures. *Trans. Am. Math. Soc.*, 93:418–491, 1959.

A. M. Foley, P. G. Leahy, A. Marvuglia, and E. J. McKeogh. Current methods and advances in forecasting of wind power generation. *Renew. Energy*, 37:1–8, 2012.

C. K. Folland, A. A. Scaife, J. Lindesay, and D. B. Stephenson. How potentially predictable is northern European winter climate a season ahead? *Int. J. Climatol.*, 32:801–818, 2012.

N. Fournier and A. Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probab. Theory Relat. Fields*, 162:707–738, 2015.

H. Friedsam and W. A. Oren. The application of the principal curve analysis technique to smooth beamlines. In *Proceedings of the 1st Int. Workshop on Acc. Alignment*, 1989.

B. A. Frigyik, S. Srivastava, and M. R. Gupta. An introduction to functional derivatives. Technical report, Department of Electrical Engineering, University of Washington, Seattle, 2008a.

B. A. Frigyik, S. Srivastava, and M. R. Gupta. Functional Bregman divergence and Bayesian estimation of distributions. *IEEE Trans. Inform. Theory*, 54:5130–5139, 2008b.

H. Fritz, L. A. Garcia-Escudero, and A. Mayo-Iscar. tclust: An R package for a trimming approach to cluster analysis. *J. Stat. Softw.*, 47(12):1–26, 2012.

L. A. Garcí a Escudero, A. Gordaliza, C. Matrán, and A. Mayo-Iscar. A general trimming approach to robust cluster analysis. *Ann. Statist.*, 36:1324–1345, 2008.

C. R. Genovese, M. Perone-Pacifico, I. Verdinelli, and L. Wasserman. The geometry of nonparametric filament estimation. *J. Am. Stat. Assoc.*, 107:788–799, 2012a.

C. R. Genovese, M. Perone-Pacifico, I. Verdinelli, and L. Wasserman. Minimax manifold estimation. *J. Mach. Learn. Res.*, 13:1263–1291, 2012b.

C. R. Genovese, M. Perone-Pacifico, I. Verdinelli, and L. Wasserman. Manifold estimation and singular deconvolution under Hausdorff loss. *Ann. Statist.*, 40: 941–963, 2012c.

S. Gerber and R. Whitaker. Regularization-free principal curve estimation. *J. Mach. Learn. Res.*, 14:1285–1302, 2013.

A. Gersho and R. M. Gray. *Vector Quantization and Signal Compression*. Kluwer Academic Publishers, Norwell, 1992.

G. Giebel, R. Brownsword, G. Kariniotakis, M. Denhard, and C. Draxl. The state-of-the-art in short-term prediction of wind power: A literature overview. Technical report, ANEMOS.plus, 2011.

I. Giulini. Kernel spectral clustering, 2016. arXiv:1606.06519.

H. R. Glahn and D. A. Lowry. The use of model output statistics (mos) in objective weather forecasting. *J. Appl. Meteorol.*, 11:1203–1211, 1972.

A. Gordaliza. Best approximations to random variables based on trimming procedures. *J. Approx. Theory*, 64:162–180, 1991.

S. Graf and H. Luschgy. *Foundations of Quantization for Probability Distributions*. Lecture Notes in Mathematics. Springer-Verlag, Berlin, Heidelberg, 2000.

R. M. Gray, A. Buzo, A. H. Gray, and Y. Matsuyama. Distortion measures for speech processing. *IEEE Trans. Acoust.*, 28:367–376, 1980.

B. Guedj. *COBRA: Nonlinear Aggregation of Predictors*, 2013. URL http://cran.r-project.org/web/packages/COBRA.

L. Guibas, D. Morozov, and Q. Mérigot. Witnessed k-distance. *Discrete Comput. Geom.*, 49:22–45, 2013.

GWEC (Global Wind Energy Council). Global Wind Statistics, 2016.

L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A distribution-free theory of nonparametric regression*. Springer Series in Statistics. Springer, New York, 2002.

M. Haeffelin, L. Barthes, O. Bock, C. Boitel, S. Bony, D. Bouniol, H. Chepfer, M. Chiriaco, J. Cuesta, J. Delanoe, P. Drobinski, J. L. Dufresne, C. Flamant, M. Grall, A. Hodzic, F. Hourdin, F. Lapouge, Y. Lemaitre, A. Mathieu, Y. Morille, C. Naud, V. Noel, W. O'Hirok, J. Pelon, C. Pietras, A. Protat, B. Romand, G. Scialom, and R. Vautard. Sirta, a ground-based atmospheric observatory for cloud and aerosol research. *Ann. Geophys.*, 23:1–23, 2005.

M. Hahsler, M. Piekenbrock, and D. Doran. dbscan: Fast density-based clustering with R. *J. Stat. Softw.*, 91(1):1–30, 2019.

T. Haiden, M. Janousek, J.-R. Bidlot, R. Buizza, L. Ferranti, F. Prates, and F. Vitart. Evaluation of ECMWF forecasts, including the 2018 upgrade. *ECMWF Technical Memo.*, 831, 2018.

P. Hall and S. N. Lahiri. Estimation of distributions, moments and quantiles in deconvolution problems. *Ann. Statist*, 36:2110–2134, 2008.

S. Has. Agrégation consensuelle et mesures de distances pour l'apprentissage statistique. Thèse de doctorat, Sorbonne Université, 2021.

T. Hastie and B. Efron. lars: Least Angle Regression, Lasso and Forward Stagewise, 2012. URL http://cran.r-project.org/package=lars.

T. Hastie and W. Stuetzle. Principal curves. *J. Am. Stat. Assoc.*, 84:502–516, 1989.

L. Horváth and G. Rice. Extensions of some classical methods in change point analysis. *Test*, 23:219–255, 2014.

H. Hotelling. Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.*, 24:417–441, 498–520, 1933.

J. Huang, T. Zhang, and D. Metaxas. Learning with structured sparsity, 2011.

B. Jin, G. Pan, Q. Yang, and W. Zhou. On high-dimensional change point problem. *Sci. China Math.*, 59:2355–2378, 2016.

M. Jirak. Uniform change point tests in high dimension. *Ann. Statist.*, 43:2451–2483, 2015.

I. M. Johnstone and B. W. Silverman. Needles and straw in haystacks: empirical Bayes estimates of possibly sparse sequences. *Ann. Statist.*, 32:1594–1649, 2004.

L. Jones and C. Byrne. General entropy criteria for inverse problems, with applications to data compression, pattern classification, and cluster analysis. *IEEE Trans. Inform. Theory*, 36, 1990.

A. Juditsky and A. Nemirovski. Functional aggregation for nonparametric estimation. *Ann. Statist.*, 28:681–712, 2000.

J. Jung and R. P. Broadwater. Current status and future advances for wind speed and power forecasting. *Renew. Sustain. Energy Rev.*, 31:762–777, 2014.

E. Kalnay. *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge University Press, 2003.

J.-H. Kang, M.-S. Suh, K.-O. Hong, and C. Kim. Development of updateable model output statistics (UMOS) system for air temperature over South Korea. *Asia-Pac. J. Atmos. Sci.*, 47:199–211, 2011.

B. Kégl and A. Krzyżak. Piecewise linear skeletonization using principal curves. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24:59–74, 2002.

B. Kégl, A. Krzyżak, T. Linder, and K. Zeger. Learning and design of principal curves. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22:281–297, 2000.

N. Keita, S. Bougeard, and G. Saporta. Clusterwise multiblock PLS regression. In *CFE-CMStatistics 2015*, page 195, December 2015.

B. Kégl. *Principal Curves: Learning, Design, and Applications*. PhD thesis, Concordia University, Montréal, Québec, Canada, 1999.

A. Korostelev and O. Lepski. On a multi-channel change-point problem. *Math. Methods Statist.*, 17:187–197, 2008.

A. P. Korostelev. On minimax estimation of a discontinuous signal. *Theory Probab. Appl.*, 32:727– 730, 1987.

O. Kramer and F. Gieseke. Short-term wind energy forecasting using support vector regression. In *Proceedings of the 6th Int. Conf. Soft Computing Models in Industrial and Environmental Applications (SOCO)*, pages 271–280. Springer, 2011.

O. Kramer, F. Gieseke, and B. Satzger. Wind energy prediction and monitoring with neural computation. *Neurocomputing*, 109:84–93, 2013.

A. Kusiak, H. Zheng, and Z. Song. Wind farm power prediction: a data-mining approach. *Wind Energy*, 12:275–293, 2009.

M. Lavielle and G. Teyssière. Detection of multiple change-points in multivariate time series. *Lith. Math. J.*, 46:287–306, 2006.

L. Lazic, G. Pejanovic, M. Zivkovic, and L. Ilic. Improved wind forecasts for wind power generation using the Eta model and MOS (Model Output Statistics). *Energy*, 73:567–574, 2014.

A. Lemenant. About the regularity of average distance minimizers in $\mathbb{R}^2$. *J. Convex Anal.*, 18:949–981, 2011.

A. Lemenant. A presentation of the average distance minimizing problem. *J. Math. Sci.*, 181:820–836, 2012.

O. Lepski. On a problem of adaptive estimation in Gaussian white noise. *Theory Probab. Appl.*, 35:454–466, 1991.

O. Lepski. Asymptotically minimax adaptive estimation I. Upper bounds. Optimally adaptive estimates. *Theory Probab. Appl.*, 36:682–697, 1992.

O. Lepski. Asymptotically minimax adaptive estimation II. Schemes without optimal adaptation: Adaptive estimators. *Theory Probab. Appl.*, 37:433–448, 1993.

M. Lerasle. Optimal model selection in density estimation. *Ann. Inst. Henri Poincaré (B) Probab. Stat.*, 48:884–908, 2012.

S. Li. FNN: Fast Nearest Neighbor search algorithms and applications, 2013. URL http://cran.r-project.org/package=FNN.

A. Liaw and M. Wiener. randomForest: Breiman and Cutler's Random Forests for classification and regression, 2002. URL http://cran.r-project.org/package=randomForest.

T. Linder. Learning-theoretic methods in vector quantization. In L. Györfi, editor, *Principles of Nonparametric Learning*. Springer-Verlag, Wien, 2002.

S. P. Lloyd. Least squares quantization in PCM. *IEEE Trans. Inform. Theory*, 28: 129–137, 1982.

X. Y. Lu and D. Slepcev. Properties of minimizers of average-distance problem via discrete approximation of measures. *SIAM J. Math. Anal.*, 45:3114–3131, 2013.

X. Y. Lu and D. Slepcev. Average-distance problem for parameterized curves. *ESAIM: Control Optim. Calc. Var.*, 22:404–416, 2016.

M. Lydia, S. S. Kumar, A. I. Selvakumar, and G. E. Prem Kumar. A comprehensive review on wind turbine power curve modeling techniques. *Ren. Sust. Energy Rev.*, 30:452 – 460, 2014.

P. Mair, K. Hornik, and J. de Leeuw. Isotone optimization in R: pool-adjacent-violators algorithm (PAVA) and active set methods. *J. Stat. Softw.*, 32:1–24, 2009.

C. L. Mallows. Some comments on C_p. *Technometrics*, 15:661–675, 1973.

E. Mangalova and O. Shesterneva. K-nearest neighbors for GEFCom2014 probabilistic wind power forecasting. *Int. J. Forecast.*, 32, 2016.

C. Mantegazza and A. Mennucci. Hamilton-Jacobi equations and distance functions in Riemannian manifolds. *Appl. Math. Optim.*, 47:1–25, 2003.

K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate Analysis*. Academic Press, London, 1979.

P. Massart. *Concentration Inequalities and Model Selection*. École d'Été de Probabilités de Saint-Flour XXXIII - 2003, Lecture Notes in Mathematics. Springer, Berlin, Heidelberg, 2007.

L. Meier, S. van de Geer, and P. Bühlmann. High-dimensional additive modeling. *Ann. Statist.*, 37:3779 – 3821, 2009.

A. Meister. *Deconvolution Problems in Nonparametric Statistics*. Lecture Notes in Statistics. Springer, 2009.

J. Mejia, M. Giordano, and E. Wilcox. Conditional summertime day-ahead solar irradiance forecast. *Sol. Energy*, 163:610–622, 2018.

M. Milligan, M. Schwartz, and Y.-H. Wan. Statistical wind power forecasting for u.s. wind farms. In *17th Conf. Probab. Stat. Atmos. Sci., American Meteorological Society Annual Meeting*, 2004.

M. Mojirsheibani. Combining classifiers via discretization. *J. Am. Stat. Assoc*, 94: 600–609, 1999.

M. Mojirsheibani. A kernel-based combined classification rule. *Stat. Probab. Lett.*, 48:411–419, 2000.

M. Mojirsheibani. An almost surely optimal combined classification rule. *J. Multivar. Anal.*, 81:28–46, 2002a.

M. Mojirsheibani. A comparison study of some combined classifiers. *Comm. Statist. Simulation Comput.*, 31:245–260, 2002b.

S. J. N. Mosconi and P. Tilli. Γ-convergence for the irrigation problem. *J. Convex Anal.*, 12:145–158, 2005.

A. Mourjane. Agrégation consensuelle et applications aux contextes de grande dimension et big data. Mémoire de Master 1, École Normale Supérieure, 2021.

H. G. Müller. Change-points in nonparametric regression. *Ann. Statist.*, 20:737–671, 1992.

A. Nemirovski. *Topics in Non-Parametric Statistics*. École d'Été de Probabilités de Saint-Flour XXVIII - 1998. Springer, 2000.

F. Nielsen, J. D. Boissonnat, and R. Nock. Bregman Voronoi diagrams: properties, algorithms and applications. Technical report, INRIA, 2007.

U. Ozertem and D. Erdogmus. Locally defined principal curves and surfaces. *J. Mach. Learn. Res.*, 12:1249–1286, 2011.

E. S. Page. A test for a change in a parameter occurring at an unknown point. *Biometrika*, 42:523–527, 1955.

Panorama de l'électricité renouvelable. Réseau de Transport de l'Électricité (RTE), Syndicat des Énergies Renouvelables (SER), Électricité Réseau Distribution France (ERDF), Association des Distributeurs d'Électricité en France (ADEeF), 2015.

E. Paolini and E. Stepanov. Qualitative properties of maximum and average distance minimizers in $\mathbb{R}^n$. *J. of Math. Sci.*, 122:3290–3309, 2004.

K. Pearson. On lines and planes of closest fit to systems of points in space. *Philos. Mag.*, 2:559–572, 1901.

P. Polak and G. Wolansky. The lazy travelling salesman problem in $\mathbb{R}^2$. *ESAIM: Control Optim. Calc. Var.*, 13:538–552, 2007.

E. C. Polley and M. J. van der Laan. Super Learner in Prediction. Technical report, Tech. Rep. UC Berkeley, 2010.

E. C. Polley and M. J. van der Laan. SuperLearner: Super Learner Prediction, 2012. URL http://cran.r-project.org/package=SuperLearner.

P. Preuss, R. Puchstein, and H. Dette. Detection of multiple structural breaks in multivariate time series. *J. Am. Stat. Assoc.*, 110:654–668, 2015.

H. Quan, D. Srinivasan, and A. Khosravi. Short-term load and wind power forecasting using neural network-based prediction intervals. *IEEE Trans. Neural Netw. Learn. Syst.*, 25:303–315, 2014.

S. T. Rachev and L. Rüschendorf. *Mass transportation problems*, volume II of *Probability and its Applications*. Springer-Verlag, 1998.

M. Ranaboldo, G. Giebel, and B. Codina. Implementation of a model output statistics based on a meteorological variable screening for short-term wind 806 power forecasts. *Wind Energy*, 16:811–826, 2013.

K. Reinhard and M. Niranjan. Parametric subspace modeling of speech transitions. *Speech Commun.*, 27:19–42, 1999.

B. Riddaway. Newsletter No. 136 - Summer 2013. ECMWF, 2013.

B. Ripley. tree: Classification and regression trees, 2012. URL http://cran.r-project.org/package=tree.

Y. Ritov. Asymptotic efficient estimation of the change point with unknown distributions. *Ann. Statist.*, 18:1829–1839, 1990.

RTE (Réseau de Transport de l'Électricité). Annual Electricity Report, 2016.

M. J. Sabin and R. M. Gray. Global convergence and empirical consistency of the Generalized Lloyd Algorithm. *IEEE Trans. Inform. Theory*, 32:148–155, 1986.

T. Salameh, P. Drobinski, M. Vrac, and P. Naveau. Statistical downscaling of near-surface wind over complex terrain in southern france. *Meteorol. Atmos. Phys.*, 103: 253–265, 2009.

S. Sandilya and S. R. Kulkarni. Principal curves with bounded turn. *IEEE Trans. Inform. Theory*, 48:2789–2793, 2002.

A. Saumard. Optimal model selection in heteroscedastic regression using piecewise polynomial functions. *Electron. J. Stat.*, 7:1184–1223, 2013.

M. J. Schmeits, K. J. Kok, and D. H. Vodelezang. Probabilistic forecasting of (severe) thunderstorms in the Netherlands using model output statistics. *Wea. Forecast.*, 20:134–148, 2005.

G. Schwarz. Estimating the dimension of a model. *Ann. Statist.*, 6:33–73, 1978.

X. Shi, Y. Wu, and C. R. Rao. Consistent and powerful graph-based change-point test for high-dimensional data. *Proceedings of the Natl. Acad. Sci. USA (PNAS)*, 114:3873–3878, 2017.

A. N. Shiryaev. *Optimal Stopping Rules*. Springer-Verlag, New York, 1978.

G. Sideratos and N. D. Hatziargyriou. Probabilistic wind power forecasting using radial basis function neural networks. *IEEE Trans. Power Syst.*, 27:1788–1796, 2012.

A. Smith, N. Lott, and R. Vose. The integrated surface database: Recent developments and partnerships. *Bull. Am. Meteor. Soc.*, 92:704–708, 2011.

Y. S. Soh and V. Chandrasekaran. High-dimensional change-point estimation: Combining filtering with convex optimization. *Appl. Comput. Harmon. Anal.*, 43:122 – 147, 2017.

S. S. Soman, H. Zareipour, O. Malik, and P. Mandal. A review of wind power and wind speed forecasting methods with different time horizons. *IEEE North American Power Symposium (NAPS)*, pages 1–8, 2010.

C. Spearman. General intelligence, objectively determined and measured. *Am. J. Psychol.*, 15:201–293, 1904.

D. C. Stanford and A. E. Raftery. Finding curvilinear features in spatial point patterns: principal curve clustering with noise. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22:2237–2253, 2000.

A. Strehl and J. Ghosh. Cluster ensembles - A knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.*, 3:583–617, 2002.

T. Tarpey and B. Flury. Self-consistency: a fundamental concept in statistics. *Stat. Sci.*, 11:229–243, 1996.

A. Tascikaraoglu and M. Uzunoglu. A review of combined approaches for prediction of short-term wind speed and power. *Ren. Sust. Energy Rev.*, 34:243–254, 2014.

C. Thiard. Classification informatique de la prononciation des voyelles en anglais à partir d'enregistrements acoustiques issus du Tyneside Linguistic Survey. Mémoire de Master 2, École Nationale Supérieure Agronomique de Rennes, 2021.

R. Tibshirani. Principal curves revisited. *Statist. Comput.*, 2:183–190, 1992.

R. Tibshirani. Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc., B: Stat. Methodol.*, 58:267–288, 1994.

C. Tikkinen-Piri, A. Rohunen, and J. Markkula. EU General Data Protection Regulation: Changes and implications for personal data collecting companies. *Comput. Law Secur. Rev.*, 34:134–153, 2018.

A. B. Tsybakov. Optimal rates of aggregation. In B. Schölkopf and M. Warmuth, editors, *Computational Learning Theory and Kernel Machines*, Lecture Notes in Computer Science, pages 303–313. Springer, Heidelberg, 2003.

G. K. Vallis. *Atmospheric and Oceanic Fluid Dynamics*. Cambridge University Press, Cambridge, U.K., 2006.

M. J. van der Laan, E. C. Polley, and A. E. Hubbard. Super learner. *Stat. Appl. Genet. Mol. Biol.*, 6, 2007.

A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer series in Statistics. Springer, 1996.

J. J. Verbeek, N. Vlassis, and B. Kröse. A soft k-segments algorithm for principal curves. In *Proc. Int. Jt Conf. Neural Netw. 2001*, pages 450–456, 2001.

C. Villani. *Optimal Transport: Old and New*. Grundlehren Der Mathematischen Wissenschaften. Springer-Verlag, 2008.

C. Wan, Z. Xu, P. Pinson, Z. Y. Dong, and K. P. Wong. Probabilistic forecasting of wind power generation using extreme learning machine. *IEEE Trans. Power Syst.*, 29:1033–1044, 2014.

D. Wang, Y. Yu, and A. Rinaldo. Optimal change point detection and localization in sparse dynamic networks. *Ann. Statist.*, 49:203–232, 2021.

T. Wang and R. J. Samworth. High dimensional change point estimation via sparse projection. *J. R. Stat. Soc., B: Stat. Methodol.*, 80:57–83, 2018.

X. Wang, P. Guo, and X. Huang. A review of wind power forecasting models. *Energy Procedia*, 12:770–778, 2011.

M. H. Wegkamp. Model selection in nonparametric regression. *Ann. Statist.*, 31: 252–273, 2003.

R. L. Wilby and C. W. Dawson. The statistical downscaling model: insights from one decade of application. *Int. J. Climatol.*, 33:1707–1719, 2013.

L. J. Wilson and M. Vallée. The Canadian Updateable Model Output Statistics (UMOS) system: design and development tests. *Wea. Forecast.*, 17:206–222, 2002.

W. C. K. Wong and A. C. S. Chung. Principal curves to extract vessels in 3D angiograms. In *Proceedings of the 2008 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW'08)*, pages 1–8, 2008.

B. Wu, M. Song, K. Chen, Z. He, and X. Zhang. Wind power prediction system for wind farm based on auto regressive statistical model and physical model. *J. Renew. Sustain. Energy*, 6:013101, 2014.

Y. Yang. Combining different procedures for adaptive regression. *J. Multivar. Anal.*, 74:135–161, 2000.

Y. Yang. Adaptive regression by mixing. *J. Am. Stat. Assoc*, 96:574–588, 2001.

Y. Yang. Aggregating regression procedures to improve performance. *Bernoulli*, 10: 25–47, 2004.

M. Zamo, O. Mestre, P. Arbogast, and O. Pannecouke. A benchmark of statistical regression methods for short-term forecasting of photovoltaic electricity production, part I: Deterministic forecast of hourly production. *Sol. Energy*, 105:792–803, 2014.

M. Zamo, L. Bel, O. Mestre, and J. Stein. Improved gridded wind speed forecasts by statistical postprocessing of numerical models with block regression. *Weather and Forecasting*, 31:1929–1945, 2016.