THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par *l'Université Toulouse III - Paul Sabatier*
Discipline ou spécialité : *Mathématiques*

Présentée et soutenue par *Nabil RACHDI*
Le *5 Décembre 2011*

Titre : *Apprentissage Statistique et Computer Experiments*
- Approche quantitative du risque et des incertitudes en modélisation -

---

JURY
Jean-Marc AZAIS - Président
Jean-Claude FORT - Diecteur de thèse
Thierry KLEIN - Diecteur de thèse
Luc PRONZATO - Rapporteur
Fabien MANGEANT - Encadrant

Ecole doctorale : *EDMITT*
Unité de recherche : *IMT - Equipe de Statistiques et Probabilités*
Directeur(s) de Thèse : *Jean-Claude FORT (Paris V) et Thierry KLEIN (Toulouse III)*
Rapporteurs : *Luc PRONZATO (CNRS-I3S) et Stéphane BOUCHERON (Paris VII)*

Université de Toulouse

EADS

THÈSE

présentée pour obtenir le grade de

DOCTEUR EN SCIENCES DE L'UNIVERSITÉ TOULOUSE III

Spécialité : **Mathématiques Appliquées - Statistiques**

par

Nabil RACHDI

APPRENTISSAGE STATISTIQUE ET COMPUTER EXPERIMENTS

APPROCHE QUANTITATIVE DU RISQUE ET DES INCERTITUDES EN MODÉLISATION

Rapporteurs :	M. Luc	**PRONZATO**	I3S, CNRS Sophia-Antipolis
	M. Stéphane	**BOUCHERON**	Université Paris VII

Soutenue publiquement le **5 Décembre 2011** devant le jury composé de

M. Jean-Marc	**AZAIS**	Université Toulouse III	Examinateur
M. Jean-Claude	**FORT**	Université Paris V	Directeur de thèse
M. Thierry	**KLEIN**	Université Toulouse III	Directeur de thèse
M. Luc	**PRONZATO**	CNRS Sophia-Antipolis	Rapporteur
M. Fabien	**MANGEANT**	EADS Innovation Works	Encadrant

Institut de Mathématiques de Toulouse CNRS UMR 5219
Équipe de Statistique et Probabilités

INSTITUT DE MATHÉMATIQUES DE TOULOUSE

À ma tendre et chère famille.

Dans l'effort que nous faisons pour comprendre le monde, nous ressemblons quelque peu à l'homme qui essaie de comprendre le mécanisme d'une montre fermée. Il voit le cadran et les aiguilles en mouvement, il entend le tic-tac, mais il n'a aucun moyen d'ouvrir le boîtier. S'il est ingénieux il pourra se former quelque image du mécanisme, qu'il rendra responsable de tout ce qu'il observe, mais il ne sera jamais sûr que son image soit la seule capable d'expliquer ses observations. Il ne sera jamais en état de comparer son image avec le mécanisme réel, et il ne peut même pas se représenter la possibilité ou la signification d'une telle comparaison. Mais le chercheur croit certainement qu'à mesure que ses connaissances s'accroîtront, son image de la réalité deviendra de plus en plus simple et expliquera des domaines de plus en plus étendus de ses impressions sensibles.

A. Einstein & L. Infeld, *L'évolution des idées en physique*

Remerciements

Mes premiers remerciements vont à mes directeurs de thèse, Jean-Claude Fort et Thierry Klein, qui m'ont fait découvrir, durant mon Master 2, les *mathématiques théoriques appliquées* à travers de longues discussions sur le traitement amont de certaines problématiques industrielles. Jean-Claude, je te remercie pour ta généreuse disponibilité et tes conseils avisés. Tu répondais toujours présent quand j'en avais besoin et nos discussions ont toujours été riches et chaleureuses. Thierry, malgré la distance, tu as toujours été présent quand il le fallait, merci pour ton aide et pour les pistes porteuses que tu m'a suggérées durant ces années de recherche.

Un grand merci à Fabien Mangeant pour m'avoir suivi tout au long de ce travail. Tes remarques ont été très constructives et nos conversations m'ont permis de prendre du recul sur mon travail. Je souhaite également remercier Thierry Druot pour m'avoir accueilli durant un mois au sein de l'équipe Avant-Projet d'Airbus à Toulouse, ce fut pour moi une expérience très enrichissante.

Je tiens à remercier les membres du jury qui ont accepté de juger ce travail. Merci aux rapporteurs, à Luc Pronzato pour ses remarques très détaillées qui ont beaucoup contribué à améliorer la qualité du manuscript, et à Stéphane Boucheron pour ses précieux conseils. Merci également à Jean-Marc Azaïs d'avoir accepté de faire partie du jury.

Merci à tous mes collègues d'EADS : Vincent, Ariane, Benoit, Pierre, Michel, Anabelle, Vassili, Régis et Gilles. Sans oublier les toulousains : Jayant, Nolwenn, Guillaume, Jérome, Stéphane et Fabien. Toujours chez EADS, je ne pourrais omettre de remercier chaleureusement Eric Duceau pour son écoute, et Isabelle Terrasse pour sa gentillesse. J'adresse une pensée amicale à mes collègues d'IMACS : Jean-Loup, Sofiane, Fanny et Denis. Merci à vous tous pour les bons moments passés durant ces dernières années.

Toute ma gratitude va à ma famille et à ma belle-famille, leur soutien constant a grandement contribué à la réalisation de ce travail. En particulier, je remercie du fond du cœur mon père et ma mère qui m'ont soutenu tout au long de mes études. S'il n'y a ne serait-ce qu'une phrase dont je puisse être fier dans cette thèse, je la leur dédie. Merci à mon frère, à mes beaux-frères et à mes sœurs pour leurs encouragements et leur attention. Un clin d'œil à ma nièce et mes deux petits neveux, qui ont été une véritable source d'énergie. Je vous embrasse tous très fort.

Enfin, je remercie mon épouse de m'avoir supporté ces trois années. Merci pour tout ce que tu as fait pour moi durant cette aventure, je te dédie ce projet.

Table des matières

Chapitre 1

Introduction générale

Sommaire

Dans les affaires incertaines et douteuses il nous faut suspendre nos actions, en attendant que brille une plus grande lumière ; mais si l'occasion d'agir ne souffre aucun retard, entre deux solutions il nous faudra toujours choisir celle qui semble mieux adaptée, plus sûre, mieux réfléchie et plus probable, même si ni l'une ni l'autre ne mérite en fait ces adjectifs.

J. Bernoulli, *Ars Conjectandi*

1.1 Contexte général : processus de prise de décision

La prise en compte des risques et des incertitudes est un passage obligé dans un processus de prise de décision pour un grand nombre de domaines tels que la finance, le secteur médical, l'environnement ou encore l'industrie. Dans ce dernier secteur en particulier, il y a trois grandes phases d'un programme industriel avant d'aboutir à une prise de décision (voir Figure 1.1) :

1. *Analyse de la situation,*
2. *Recueil des éléments de décision,*
3. *Utilisation de la théorie de la décision .*

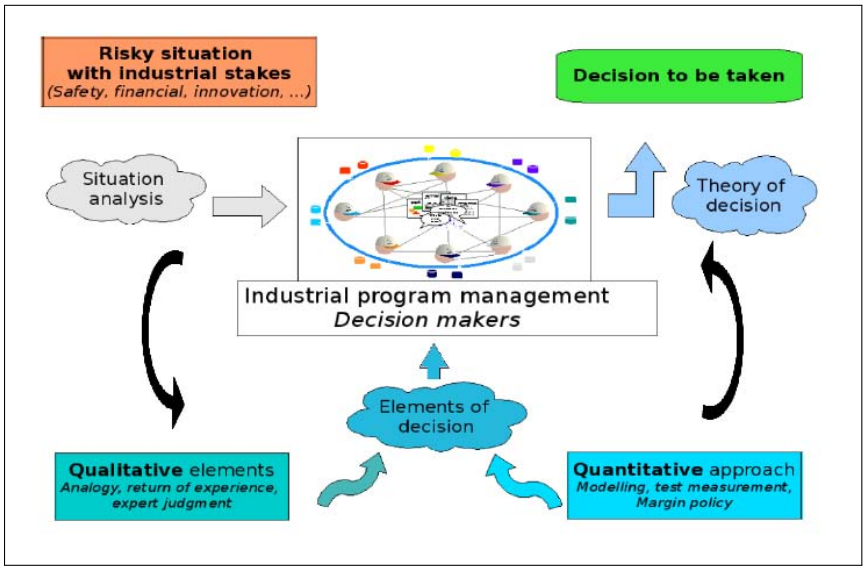


FIGURE 1.1 – Schéma d'un processus de prise de décision.

Bien que les applications et les phénomènes mis en jeu puissent être de nature très différente d'un secteur à un autre (par exemple le médical et l'aéronautique), le principe de modélisation et d'intégration des incertitudes reste très semblable.

Dans cette thèse, nous nous concentrerons principalement sur la phase "Éléments de décision" dont le but est de fournir un cadre permettant l'intégration d'outils aussi bien *qualitatifs* que *quantitatifs* tout en spécifiant les ressources nécessaires.
Par "qualitatifs", on entend ce qui se réfère aux retours d'experience (Rex), à l'intuition ou encore au jugement d'expert. Ces méthodes ne reposent pas sur des tests ou des modèles numériques.
Par "quantitatifs", on désigne les méthodes basées sur une représentation théorique et/ou simplifiée de la compréhension que l'on a de la situation. C'est plus particulièrement dans ce cadre là que s'inscrivent les présents travaux.

Un besoin industriel important dans la conception d'un produit, on parle de *cycle de vie* de la conception, est la garantie d'une certaine sécurité vis-à-vis des choix considérés tout au long de ce cycle. Ce besoin est d'autant plus important que le programme industriel définissant la conception peut être assez complexe, on parle de *système complexe*, où interagissent plusieurs

disciplines avec certaines règles propres au programme considéré. La philosophie du "principe de précaution" est prépondérante dans une telle situation et elle se traduit en particulier par deux faits :

- **La prise de marge :** adoption de coefficients de "sécurité" cooptés par des experts
- **L'analyse du pire cas :** évaluer la plage de fonctionnement du produit dans des conditions extrêmes.

La juxtaposition de la prise de marge et de l'analyse du pire cas a pour but d'assurer un maximum de sécurité et de prévoir une éventuelle défaillance au cours du cycle.
Une notion complémentaire à celle du principe de précaution est la notion de *conception robuste*, où l'enjeu repose sur la garantie de performance du produit tout en minimisant sa sensibilité aux perturbations (dispersions de fabrication, vieillissement, variations d'environnement etc...).

Beaucoup d'industriels ont pris conscience de la nécessité d'une démarche systématique et quantitative du principe de précaution et de la conception robuste sous incertitude avec l'objectif d'aller vers une logique de "démonstration", concernant par exemple le pire cas et la prise de marge.

Les modèles mathématiques sont couramment utilisés tant dans le monde académique que dans le monde industriel. Ils ont d'une part la vocation de modéliser des phénomènes complexes, et d'autre part ils contribuent à une meilleure compréhension de ces phénomènes. Un des enjeux majeurs de la modélisation est d'être le plus "réaliste" possible, que ce soit en termes de précision des modèles utilisés ou bien en termes de qualité (et quantité) de l'information disponible. L'intégration de la méconnaissance, et des *incertitudes* en général, dans la modélisation de phénomènes complexes s'est beaucoup développée cette dernière décennie en gardant à l'esprit trois points fondamentaux :

- **Réalisme** : la modélisation d'un phénomène variable par une représentation déterministe peu être assez réductrice,
- **Robustesse** : on va chercher un modèle qui "résiste" à des petites modifications de sa structure,
- **Rapidité d'execution** : le temps de calcul, comprenant éventuellement la préparation de données etc..., est un paramètre important d'un modèle numérique car un modèle coûteux en temps de calcul est difficilement exploitable pour une analyse d'incertitude.

La motivation principale de nos travaux provient de la *méthodologie incertitude* communément adoptée dans l'industrie, dont le livre [2] détaille les différentes composantes dites A,B,C et D (Figure 1.2). Dans cette thèse, nous proposons un cadre théorique assez général pour l'étude et l'analyse de modèles numériques en présence de données incertaines.

1.2 Les Computer Experiments

1.2.1 Histoire et définition

Les grandes avancées des moyens de calculs depuis les années 50 ont fait naître le domaine du *Scientific Computing* ou *Computational Science*, qui est une science à part entière, dont l'objet est en substance l'analyse des modèles mathématiques implémentés dans les ordinateurs. En particulier, les *Expériences Simulées* ou *Computer Experiments* ont connu un développement

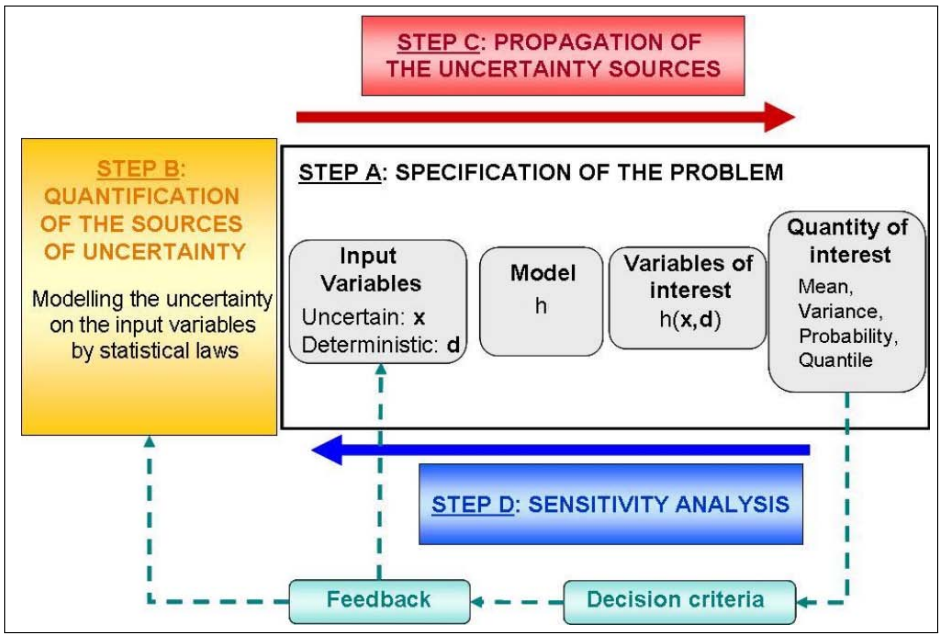


FIGURE 1.2 – Approche méthodologique pour le traitement des incertitudes dans les modèles numériques.

grandissant, jouant un rôle crucial pour la compréhension de phénomènes complexes.

Plusieurs définitions d'un *computer experiment*[1] peuvent être trouvées dans la littérature, considérons par exemple celle donnée dans [4] :

" *Suppose that a mathematical theory exists, e.g, a set of differential equations, that relates the output of a complex physical process to a set of input variables. Suppose also that a numerical method exists for accurately solving the mathematical system. The presence of these two elements with appropriate computer hardware and software to implement the numerical methods allows one to conduct a* **computer experiment** *[...]* "

1.2.2 La simulation : un pont entre l'expérience et la théorie

La simulation - au sens computer experiment - vient alors naturellement se glisser entre "expérience" et "théorie", ce qui est une position favorable à un compromis entre les "mesures expérimentales" d'une part, et les "modèles et connaissances a priori" d'autre part. On qualifie la simulation de pont entre l'expérience et la théorie alors qu'on aimerait un pont entre l'expérience et la réalité. Nous qualifierons le chaînon manquant entre théorie et réalité de **validation de modèle** qui peut être étudié soit grâce à une connaissance partielle de la réalité, soit grâce à un jugement d'expert. En ce qui nous concerne, la "réalité" est donnée par une variable Y, appelée *variable d'intérêt*, représentant un certain phénomène que l'on supposera réel et appartenant à un espace $\mathcal{Y} \subset \mathbb{R}$. Il y a ainsi 4 niveaux, de l'expérience à la réalité, que l'on illustre dans la Figure 1.3.
La validation de modèle précédemment évoquée tend à justifier le choix et l'emploi du modèle $\mathcal{M} = \{x \mapsto h(x,\theta), \theta \in \Theta\}$. L'erreur induite est communément appelée **erreur d'approximation** ou **erreur systématique**. L'usage d'un computer experiment vient rajouter une erreur supplémentaire appelée **erreur d'estimation**. L'expérience fournit des données

1. Dans toute la suite, on préfèrera l'appellation anglo-saxone "computer experiment" à "expérience simulée"

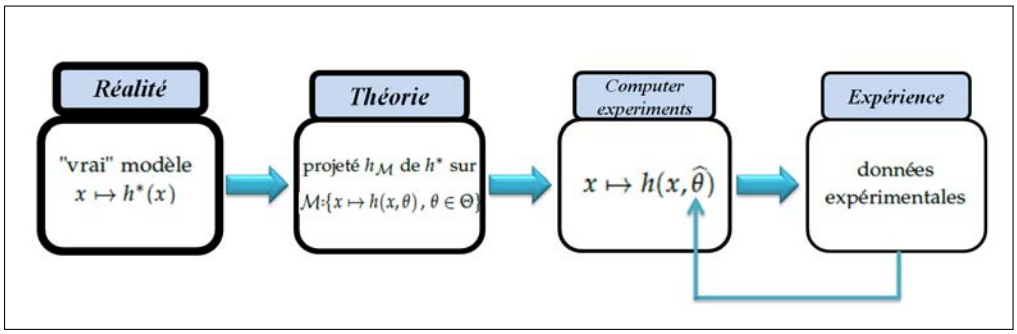


FIGURE 1.3 – De la réalité à l'expérience.

brutes, "qui parlent d'elles-mêmes" (Silverman [5]), et une analyse portant uniquement sur des données expérimentales peut conduire dans certains cas complexes à une erreur (de généralisation) non négligeable. D'autres types d'erreur peuvent encore apparaitrent selon la modélisation considérée, par exemple lorsque l'on intègre des *sources d'incertitudes* (cf. Section 1.3).
Nous pouvons ainsi constater le rôle central des computers experiments, d'autant plus qu'ils ne sont pas figés, c'est à dire qu'ils laissent place, par exemple, à l'intégration de nouveaux moyens de calcul, à un raffinement des modèles, à l'implémentation d'algorithmes plus performants etc...
En outre, la simulation nous permet aussi l'accès à des *régions critiques* qui ne sont pas ou très difficilement accessibles expérimentalement. Et au-delà de ces situations "critiques" en réalité, la simulation, si elle est bien menée, peut améliorer notre compréhension du phénomène considéré en se plaçant dans des configurations "impossibles" en réalité. Cette souplesse fait la force des computers experiments qui deviennent alors un outil puissant, propice à des progrès dans diverses disciplines.

1.2.3 Computer experiment et modèle boîte noire

Une des composantes d'un computer experiment est l'utilisation d'une *boîte noire*, terme introduit dans les années 30 par le mathématicien allemand W. Cauer [1] dans le domaine de l'*Electrical Engineering*. Peu à peu, ce concept s'est très vite étendu à d'autres domaines tels que la physique, la finance, la cryptographie ou encore la philosophie. Ce concept appliqué à un computer experiment découpe en trois parties la phase de calcul : les entrées, la fonction de transfert et les sorties, voir Figure 1.4.

FIGURE 1.4 – Illustration d'un boîte noire

Bien que simple, cette vision nécessite de définir en particulier la nature de chacune des entrées, ce qui en pratique n'est pas direct. Par exemple, selon [4] on peut avoir différents types de variables d'entrées : les variables de contrôle, variables d'environnement, variables de modèle etc... À ces variables peuvent s'ajouter des variables *incertaines*, *bruitées*. Souvent, on scinde les entrées en deux parties : les entrées déterministes et les entrées incertaines. Parmi les entrées déterministes, il peut y avoir des entrées "fixes" pour la configuration de calcul considérée : par exemple celles connues avec précision, et des entrées "variables" communé-

ment appelées *paramètres* ou *tuning parameters*. Sans perte de généralité, on pourra supposer que les entrées "fixes" font parties de la fonction de transfert et considérer comme entrées déterministes simplement les paramètres. Selon la configuration de l'étude, une variable peut passer de déterministe "fixe" à paramètre à déterminer, ou bien, une variable déterministe peut dans une autre configuration être incertaine.
En somme, pour notre propos nous noterons une fonction boîte noire (que l'on appellera aussi *modèle numérique*) comme suit :

$$h : \mathcal{X} \times \Theta \longrightarrow \mathcal{Y} \qquad (\mathbf{x}, \boldsymbol{\theta}) \longmapsto h(\mathbf{x}, \boldsymbol{\theta}) \tag{1.1}$$

où $\mathcal{X} \subset \mathbb{R}^d$ est l'espace des entrées dont on a une méconnaissance, et $\Theta \subset \mathbb{R}^k$ est l'espace des paramètres. On a supposé que la fonction h est à valeur dans l'espace $\mathcal{Y}$ du phénomène d'intérêt Y.
Dans ce travail, on considérera les situations où les $\mathbf{x}_i$ produisant les "vraies" donnée Y_i (cf. Section 1.2.4) ne sont pas observés. Ce qui correspond, par exemple, à la situation où les entrées du modèles h sont différentes des conditions expérimentales.
Par ailleurs, on pourra voir Y comme une certaine fonction de $\mathbf{x} \in \mathcal{X}$, par exemple $Y = h^*(\mathbf{x}) + \eta^*$, où la fonction h^* est inconnue et η^* est une variable aléatoire (bruit).
Dans toute la suite, la fonction h sera supposée "exploitable", c'est à dire demandant un temps de calcul raisonnable. En effet, il se peut qu'un computer experiment soit gourmand en temps de calcul et/ou que les moyens nécessaires aient un coût non négligeable, par exemple lorsque l'on modélise "finement" le phénomène Y qui nous intéresse. Dans ce cas, la fonction h sera par exemple une modélisation moins fine du phénomène ou bien un certain *métamodèle*[2]. Autrement dit, de manière générale la fonction h sera vue soit comme un programme informatique (simplification d'une fonction boîte noire coûteuse) dont les paramètres $\boldsymbol{\theta}$ peuvent avoir une signification physique, soit comme une fonction "purement" mathématique, par exemple

$$h(\mathbf{x}, \boldsymbol{\theta}) = \sum_{l=1}^{k} \phi_l(\mathbf{x})\, \theta_l\,, \tag{1.2}$$

où les ϕ_l sont des fonctions connues.

1.2.4 Natures des "vraies" données $Y_1, ..., Y_n \in \mathcal{Y}$

En pratique, les données de "références", $Y_1, ..., Y_n$, aussi appelées "réelles" ou encore "expérimentales" sont des données auxquelles nous attachons une grande importance car toute inférence sera basée sur ces dernières. Selon les problèmes posés, la "réalité" représentée par ces données peut avoir des natures différentes : par exemple, les mesures $Y_1, ..., Y_n$ peuvent être issues d'essais en vol, en soufflerie, sur maquettes etc... Ces données peuvent aussi provenir d'un code de calcul très coûteux (moyens, temps de calcul etc...) sensé modéliser le phénomène d'intérêt Y, code très peu exploitable. Dans ce dernier cas, la fonction h est bien souvent un substitut au code coûteux. Il y a ainsi deux niveaux de réalité : la "vraie vie" et "ce qui est coûteux", que l'on confondra dans nos travaux bien qu'il convienne de garder cette nuance à l'esprit.

Trois points fondamentaux restent à élucider, dans l'ordre :

2. terme désigné pour "modèle simplifié"

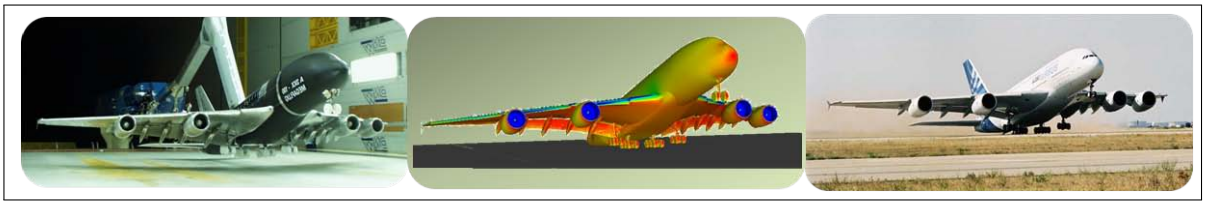


FIGURE 1.5 – Exemple de "vraies" données, de gauche à droite : essais en soufflerie, simulations numériques et essais réels.

1. la prise en compte des erreurs et des incertitudes (sur $\mathbf{x}$, sur le modèle h etc...),
2. le but recherché de l'analyse,
3. l'estimation des paramètres .

1.3 Prise en compte des incertitudes et de leurs erreurs

" Uncertainty is not an accident of the scientific method, but its substance."
Peter Hoeg (1995), romancier Danois.

1.3.1 Les incertitudes

Le *management* des incertitudes dans les modèles numériques s'est largement développé dans le monde industriel ces 10 dernières années. La formation de groupes de recherche tels que ESReDA [3], le GDR MASCOT NUM, le GT Incertitudes ou encore l'IMdR [4] ont contribué à une avancée significative dans ce domaine. En effet, la maîtrise de l'incertitude est une étape clé en recherche appliquée car les modèles numériques sont très souvent au centre de processus de décision, avec des enjeux industriels importants. Le terme "incertitude" concerne en réalité ce que l'on ne "maîtrise pas" dans les données d'entrées $\mathbf{x} \in \mathcal{X}$ mais concerne également l'incertitude sur le modèle numérique h lui-même (choix de la modélisation ...). Pour plus de précisions sur ces aspects de sources d'incertitudes, on renvoie au livre [2] traitant des incertitudes dans le contexte industriel.

On modélisera simplement la donnée incertaine $\mathbf{x}$ par une variable aléatoire $\mathbf{X}$ appartenant à un espace de probabilité $(\mathcal{X}, \mathcal{B}, P^{\mathbf{x}})$. Par suite, la fonction boîte noire (1.1) devient

$$h : (\mathcal{X}, \mathcal{B}, P^{\mathbf{x}}) \times \Theta \longrightarrow \mathcal{Y} \qquad (\mathbf{X}, \boldsymbol{\theta}) \longmapsto h(\mathbf{X}, \boldsymbol{\theta}) , \tag{1.3}$$

qu'on pourra qualifier de *modèle numérique stochastique* (ou simplement *modèle stochastique*).

Par ailleurs, on modélise également le phénomène Y (réalité physique) par une variable aléatoire de mesure Q inconnue de densité f (par rapport à la mesure de Lebesgue).

1.3.2 Différentes composantes des incertitudes

Incertitudes sur le modèle stochastique
La considération d'un modèle stochastique du type (1.3) pour représenter le phénomène

3. European Safety, Reliability & Data Association
4. Institut pour la Maîtrise des Risques

d'intérêt Y induit la considération de plusieurs niveaux de modélisation dont on donne une brève description :

- **"vrai" modèle** h^* **:** définit le phénomène Y (par exemple $Y = h^*(\mathbf{X}) + \eta^*$),
- **modèle théorique** h_{th} **:** correspond au niveau de compréhension et de simplification du problème,
- **modèle numérique** h_{num} **:** solution numérique du modèle théorique (choix d'un schéma numérique etc...),
- **modèle implémenté** h **:** implémentation du modèle numérique sur ordinateur.

Le passage d'un niveau à l'autre introduit une erreur de modélisation. Par rapport à la Figure 1.3, la nouveauté est l'introduction du modèle "numérique" entre Computer Experiments et Théorie. On se permettra l'appellation "numérique" même pour les modèles "implémentés".

Incertitudes sur les entrées.
En plus des erreurs précédentes peut s'introduire un autre type d'erreur : l'**erreur sur les entrées X**, autrement dit l'erreur sur la distribution $P^{\mathbf{x}}$. Cela représente un réel enjeu en pratique. Pour notre propos, cette considération est implicite dans l'écriture du modèle stochastique $h(\mathbf{X}, \boldsymbol{\theta})$ où $\mathbf{X} \sim P^{\mathbf{x}}$. En effet, soit $P^{\mathbf{x},*}$ la "vraie" distribution (inconnue) de la variable $\mathbf{X}$ et notons $F^*_{\mathbf{x}}$ et $F_{\mathbf{x}}$ les fonctions de distribution associées aux mesures $P^{\mathbf{x},*}$ et $P^{\mathbf{x}}$, respectivement. Supposons que $F^*_{\mathbf{x}}$ appartient à une famille $\{F_{\mathbf{x},\boldsymbol{\beta}}\,,\, \boldsymbol{\beta} \in B \subset \mathbb{R}^b\}$. Alors, sous certaines conditions de continuité on peut toujours écrire

$$\widetilde{h}(\mathbf{X}, \widetilde{\boldsymbol{\theta}}) = h(F^{-1}_{\mathbf{x},\boldsymbol{\beta}} \circ F_{\mathbf{x}}(\mathbf{X}), \boldsymbol{\theta})\,,$$

avec $\widetilde{\boldsymbol{\theta}} = (\boldsymbol{\theta}, \boldsymbol{\beta})^T$ et $\widetilde{h}$ donné par la relation précédente. Afin de garder une écriture homogène dans la suite de nos développements, on considérera toujours la distribution des entrées connue, quitte à effectuer le changement $h = \widetilde{h}$ et $\boldsymbol{\theta} = \widetilde{\boldsymbol{\theta}}$. Nous insistons sur le fait que cette simplification d'écriture ne néglige en aucun cas l'effet du choix de la distribution $P^{\mathbf{x}}$ en pratique, bien au contraire.

Après cela, il reste à définir le but de l'analyse (du calcul) que l'on qualifiera de *post-traitement*, et également, à déterminer une procédure d'estimation des paramètres. Ces deux derniers points, avec le fait que la fonction h peut n'être connue que sur des jeux entrées/sorties (boîte noire), constituent l'objet principal de cette thèse.

1.4 Post-traitements

En pratique, on rencontre essentiellement deux post-traitements : le **problème inverse** et la **prédiction**.

1.4.1 Problème inverse

Dans le cas d'un problème inverse, le but est d'identifier le paramètre $\boldsymbol{\theta}$ qui est bien souvent une grandeur physique pertinente, à l'aide du modèle stochastique h et de données

$Y_1, ..., Y_n$. Par exemple, notons $f_{\boldsymbol{\theta}}$ la densité de probabilité associée à la sortie du modèle stochastique $h(\mathbf{X}, \boldsymbol{\theta})$ (variable aléatoire) et rappelons que f est la densité associée au phénomène Y. Considérons la dissemblance de Kullback-Leibler sur les densités

$$KL(f_1, f_2) = \int_{\mathcal{Y}} \log\left(\frac{f_1}{f_2}\right)(y)\, f_1(y)\, dy\,, \quad f_1,\ f_2 \text{ deux densités sur} \mathcal{Y}\,,$$

avec les propriétés $KL(f_1, f_2) \geq 0$ et $KL(f_1, f_2) = 0 \Leftrightarrow f_1 = f_2$.
Une manière d'optimiser le paramètre $\boldsymbol{\theta}$ est de minimiser la quantité

$$KL(f, f_{\boldsymbol{\theta}})\,. \tag{1.4}$$

Deux difficultés apparaissent alors : d'une part f est inconnue et d'autre part $f_{\boldsymbol{\theta}}$ n'est pas supposée être connue analytiquement car h est une fonction boîte noire. L'idée est de remplacer f et $f_{\boldsymbol{\theta}}$ par des versions empiriques : i.e f sera remplacée par sa "*densité empirique* $f^n{}'' = \frac{1}{n}\sum_{i=1}^n \delta_{Y_i}$ et $f_{\boldsymbol{\theta}}$ sera remplacée par une *densité simulée* basée sur un échantillon $\left(h(\mathbf{X}'_j, \boldsymbol{\theta})\right)_{j=1,...,m}$ où $\mathbf{X}'_j \sim P^{\mathbf{x}}$, par exemple à l'aide d'une reconstitution par noyau que l'on note $f^m_{\boldsymbol{\theta}}$. Ainsi, on obtient une version empirique de la quantité (1.4) donnée par

$$KL(f^n, f^m_{\boldsymbol{\theta}}) = -\log(n) - \frac{1}{n}\sum_{i=1}^n \log(f^m_{\boldsymbol{\theta}})(Y_i)\,, \tag{1.5}$$

que l'on va utiliser pour l'estimation du paramètre. On obtient

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\Theta}{\text{Argmin}}\, KL(f^n, f^m_{\boldsymbol{\theta}}) = \underset{\boldsymbol{\theta}\in\Theta}{\text{Argmin}} -\frac{1}{n}\sum_{i=1}^n \log(f^m_{\boldsymbol{\theta}})(Y_i)\,. \tag{1.6}$$

Nous nous distinguons du cadre statistique classique où le statisticien "pose" un modèle sur les densités $\{f_{\boldsymbol{\theta}}\,,\ \boldsymbol{\theta} \in \Theta\}$. L'esprit adopté dans ces travaux de thèse est de "laisser s'exprimer" le modèle stochastique $\mathbf{X} \mapsto h(\mathbf{X}, \boldsymbol{\theta})$, ce qui se traduit en pratique par de la simulation. Le Chapitre 3 est consacré à l'étude générale de telles procédures en utilisant les processus empiriques dans le cadre de la M-estimation.

1.4.2 Prediction

Dans le cas d'un problème de prédiction, le regard est porté sur la variable d'intérêt Y et l'objectif est de déterminer, ou plutôt de prédire, un comportement particulier du phénomène Y par l'étude du modèle stochastique h . La prédiction passe inévitablement par une étape d'estimation des paramètres et on peut établir que la prédiction inclut un problème inverse, mais à une petite nuance près : en prédiction on ne cherche pas de réalité physique aux paramètres, contrairement à un "vrai" problème inverse, on s'assure plutôt que les paramètres satisfont à un certain critère de qualité de prédiction. Citons Kennedy et O'Hagan [3]

"*It may be that the physically true value of a calibration parameter gives a worse fit, and less accurate future prediction, than other value. It is dangerous to interpret the estimates of* $\boldsymbol{\theta}$ *obtained by calibration as estimates of the true physical values of those parameters.*"

Par conséquent, un problème de prédiction consiste en la donnée
– d'une quantité à prédire relative au phénomène Y : moyenne, quantile, probabilité de dépassement, densité, espérance conditionnelle etc... que l'on appellera *quantité d'intérêt* ou *caractéristique*,

– et d'une procédure d'estimation des paramètres.
En pratique, cela revient à calculer un paramètre $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$ puis à utiliser le modèle stochastique $\mathbf{X} \mapsto h(\mathbf{X}, \widehat{\boldsymbol{\theta}})$ afin de simuler la quantité d'intérêt recherchée.

Dans cette thèse, nous nous sommes aussi penchés sur l'influence de la méthode d'estimation des paramètres sur la performance (de prédiction) de la quantité d'intérêt recherchée. Nous parlerons de *dualité*. Nous tenons à souligner que la notion de dualité introduite dans ces travaux est différente de celle employée classiquement en mathématiques. Pour notre propos, la dualité estimation-prédiction est à entendre comme une relation de réciprocité entre méthodes d'estimation et prédiction.

1.5 Post-traitement et estimation des paramètres, une dualité ?

Il est fréquent en pratique d'utiliser un seul jeu de paramètres issu d'une certaine procédure d'estimation (souvent par moindres carrés) pour prédire certains comportements de la variable d'intérêt Y.
Une question se pose alors :
existe-t-il une procédure d'estimation "optimale" pour prédire une quantité d'intérêt donnée ?
Cette question est étudiée au Chapitre 4 où en particulier nous tenterons d'élucider l'exemple donné Figure 1.6 : supposons que l'on veuille prédire la densité de Y (en rouge) à partir des données $(\mathbf{x}_1, Y_1), ...(\mathbf{x}_n, Y_n)$ et d'un modèle $(\mathbf{X}, \boldsymbol{\theta}) \mapsto h(\mathbf{X}, \boldsymbol{\theta})$, où les $\mathbf{x}_i$ sont des observations i.i.d sous $P^{\mathbf{x}}$. On trace en Figure 1.6 la densité (par méthode à noyau) de la variable aléatoire $h(\mathbf{X}, \widehat{\boldsymbol{\theta}})$ pour 3 méthodes d'estimation de $\widehat{\boldsymbol{\theta}}$, en particulier la méthode de régression (en cyan) et la méthode minimisant la divergence de Kullback-Leibler donnée en (1.6) (en bleu). On note clairement dans cet exemple que la méthode par minimisation de la divergence de

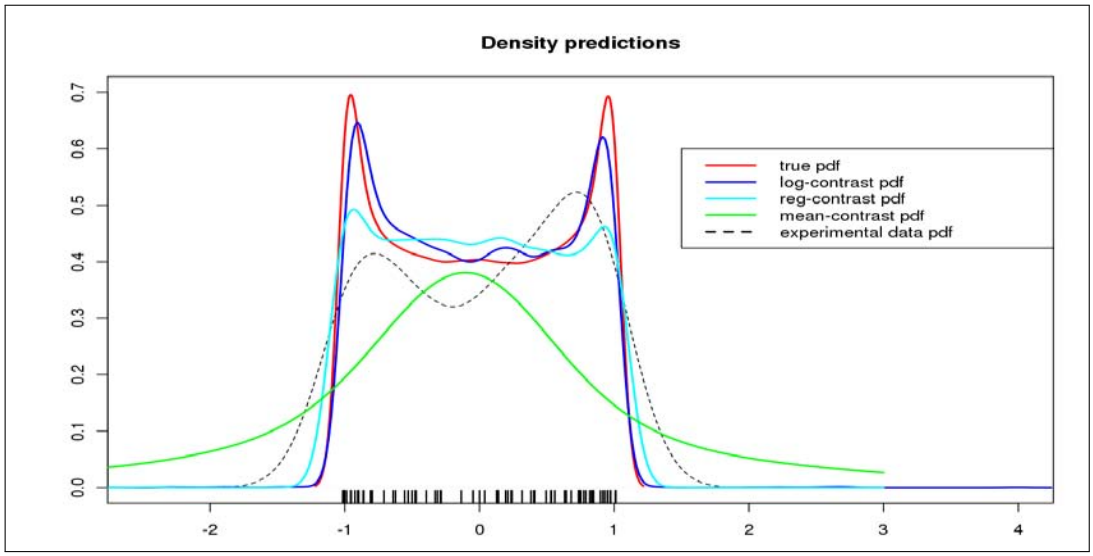


FIGURE 1.6 – Prédiction de la densité f avec estimation par minimisation de KL (en bleu), avec estimation par régression (en cyan) et avec estimation par moyenne (en vert)

Kullback-Leibler donne un meilleur résultat que celle issue de la régression. Cela s'explique en partie par le fait que la métrique de Kullback-Leibler agit sur les densités alors que la méthode de regression porte sur la norme $L_2(P^{\mathbf{x}})$ opérant sur les espérances conditionnelles. Comme le but de l'analyse est la prédiction d'une densité, il paraît alors assez naturel qu'une

méthode d'estimation "faite pour cela" soit plus performante. Nous donnerons plus de détails et d'exemples au Chapitre 4.

1.6 Approche apprentissage statistique

Nos travaux s'inscrivent dans le cadre de l'*Apprentissage Statistique*, propice à l'étude et à l'analyse de comportements stochastiques, aussi complexes soient-ils, à travers des outils tels que les inégalités de concentrations et les processus empiriques. Dans le Chapitre 2, nous donnons un formalisme pour l'étude des computer experiments à travers la théorie de l'apprentissage statistique. C'est dans ce cadre que nous étudierons une grande classe de procédures statistiques intégrant la simulation (par exemple (1.6)) via la notion de *fonction de contraste*. Puis, c'est encore dans ce cadre que nous formaliserons la *dualité* estimation-prédiction dans les computer experiments.

1.7 Objectifs et structure de la thèse

Ce travail de thèse a pour objectifs de :

- Formaliser la notion de quantité d'intérêt,
- Adapter le cadre de l'apprentissage statistique à l'analyse de modèles numériques en présence d'incertitude,
- Proposer et étudier théoriquement des méthodes d'estimation de paramètres et de prédiction de quantités d'intérêt.

La thèse comporte 8 chapitres organisés comme suit.

Dans le Chapitre 2, un bref rappel sur l'apprentissage statistique est présenté. Ensuite, on donne deux définitions d'une quantité d'intérêt : la première est une définition fonctionnelle et la seconde est la caractérisation comme argument minimum d'une fonction de contraste moyennée (que l'on appellera risque) que l'on aura définie au préalable. La notion de pénalisation de contraste est également évoquée, avec une illustration dans le cas de la *ridge regression*. Après cela, on donne les définitions propres à l'usage de computer experiments dans le cadre de l'apprentissage statistique, et on présente les problèmes liés à l'étude des quantités d'intérêt correspondantes.

Dans le Chapitre 3, on propose une méthode d'estimation des paramètres dans les modèles boîte noire stochastiques basée sur la M-estimation. L'originalité de cette méthode réside dans le fait que le M-estimateur dépend non seulement des n données observées mais également des m données de simulation. On montre le Théorème 3.4.1 donnant une borne du risque du M-estimateur qui n'est pas exactement de l'ordre de $1/\sqrt{n}$ dans les cas réguliers, mais comporte un terme supplémentaire due à la nécessité de la simulation.

Le Chapitre 4 est consacré à l'étude de prédictions de quantités d'intérêt définies dans le Chapitre 2. On se focalise plus particulièrement sur la *dualité* estimation-prédiction où l'estimation est à entendre au sens du Chapitre 3. Notre propos sera illustré d'exemples académiques (voir par exemple la Proposition 4.5.1) ainsi que d'applications numériques simples qui montrent en particulier que la procédure d'estimation par régression n'est pas "optimale" pour certaines quantités d'intérêt recherchées.

Dans le Chapitre 5, on propose un algorithme stochastique permettant le calcul pratique des estimateurs issus des procédures statistiques définies dans le Chapitre 3 conduisant à minimiser une fonction qui peut être fortement non convexe. L'idée clé de notre algorithme est de convoluer la fonction à minimiser avec une fonction "dynamique" (qui varie en fonction du temps) bien choisie.

Les Chapitres 6 et 7 proposent des applications dans un contexte industriel de la méthodologie développée dans les chapitres précédents.

Au Chapitre 6 on trouve une application des procédures d'estimation précédemment définies pour l'identification d'un paramètre de consommation moteur en ingénierie des turbines. À partir de masses de fuel consommées en croisière par un avion commercial, on propose d'inférer de manière robuste le paramètre d'intérêt à l'aide d'un modèle bruité prenant en entrée des variables aéronautiques (quelques-unes sont incertaines) dont le paramètre d'intérêt. On montre en particulier le Théorème 6.6.1 qui est une application du Théorème 3.4.1 du Chapitre 3.

Le contenu du Chapitre 7 est une première approche de formalisation pour traiter la combinaison de données expérimentales et simulées, pouvant être de nature différente, afin de prédire le comportement d'un phénomène non observable. Cette réflexion a été motivée par deux faits : le premier est la volonté des constructeurs aéronautiques de passer des constructions en alliages d'aluminium aux matériaux composites, le second porte sur l'évaluation de la *POD* (probability of detection) dans le domaine du contrôle non destructif où l'on dispose de données expérimentales et simulées pour un matériau de référence, et des données de simulations pour le matériau d'intérêt. Le but étant de "construire" des données proches des données expérimentales du matériau d'intérêt (inobservables en pratique).

Bibliographie

[1] W. Cauer. Die verwirklichung von wechselstromwiderständen vorgeschriebener frequenzabhängigkeit. *Electrical Engineering (Archiv fur Elektrotechnik)*, 17(4) :355–388, 1926.

[2] E. de Rocquigny, N. Devictor, and S. Tarantola, editors. *Uncertainty in industrial practice*. John Wiley.

[3] M.C. Kennedy and A. O'Hagan. Bayesian calibration of computer models. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 63(3) :425–464, 2001.

[4] T.J. Santner, B.J. Williams, and W. Notz. *The design and analysis of computer experiments*. Springer Verlag, 2003.

[5] B.W. Silverman. *Density estimation for statistics and data analysis*. Chapman & Hall/CRC, 1986.

Chapitre 2

Apprentissage Statistique et Computer Experiments

Sommaire

Résumé du Chapitre

Dans ce chapitre, nous proposons un formalisme provenant de l'apprentissage statistique pour l'étude et l'analyse des computers experiments sous incertitudes. Nous tenterons d'expliciter clairement les deux principales phases d'une telle étude : l'estimation (calibration) des paramètres et la prédiction d'une quantité d'intérêt liée au phénomène considéré, tout en évoquant une sorte de *dualité* entre ces deux phases. Nos propos seront illustrés par des exemples académiques et numériques.

[AVERTISSEMENT]

Les 4 premières sections de ce Chapitre traitent de généralités sur l'apprentissage statistique. Le lecteur familier avec cette discipline peut directement se rendre à la Section 2.5 p.31 où nous posons la problématique de nos développements.(Un détour sur la formalisation d'une quantité d'intérêt p.19 est toutefois conseillé).

2.1 Introduction

Les moyens de calcul et de modélisation étant de plus en plus performants, l'utilisation des *Computers Experiments* se révèle être un outil indispensable pour l'accompagnement de jugements et/ou de prises de décision relatifs à une caractéristique d'un phénomène donné. Toutefois, comme nous l'avons vu dans le chapitre précédent, cette modélisation comporte des *erreurs* qui peuvent soit lui être intrinsèque (erreur d'approximation, bruit numérique...) soit provenir d'une méconnaissance de certaines variables intervenant dans le modèle. Il est donc nécessaire de formaliser et de quantifier (dans la mesure du possible) ces erreurs afin de mettre en exergue leurs effets sur l'objectif voulu.

2.2 Brève introduction à l'apprentissage statistique

La théorie de l'apprentissage statistique (*Statistical Learning*) est née des travaux précurseurs de V. Vapnik et A. Chervonenkis [8],[9], entre les années 1960 et 1990. L'apprentissage statistique est une branche du *Machine Learning* qui lui même est une branche de l'Intelligence Artificielle, que l'on ne définira pas, mais dont on retiendra en substance qu'il s'agit de développer des algorithmes permettant à un système d'"apprendre" à partir de données empiriques. Le mot "apprendre" peut avoir plusieurs significations : par exemple, à partir d'une base de données entrées/sorties (*training set*) aussi appelée *base d'apprentissage*, le système prédit la sortie d'une nouvelle entrée. On peut aussi vouloir décrire le comportement (de la sortie) d'un système dans un contexte donné : comportement moyen, extrême, global etc... Dans ce cas, nous parlerons de prédiction de quantités d'intérêt que l'on étudiera au Chapitre 4.
L'apprentissage statistique est un cadre théorique assez souple qui permet une analyse rigoureuse et formelle d'un phénomène avec certaines garanties sur la performance des algorithmes utilisés. Ce cadre permet de prendre en compte toute amélioration éventuelle de l'information et des outils de modélisation.
Une introduction à la théorie de l'apprentissage statistique est donnée par O. Bousquet, S. Boucheron et G. Lugosi dans [3], ainsi que dans les références qui y figurent.

En résumé, le cadre général de l'apprentissage cimente les trois points suivants :

1. Observation du phénomène
2. Modélisation du phénomène observé
3. Prédiction du phénomène (ou d'une caractéristique) à l'aide de la modélisation.

Une méthodologie d'apprentissage statistique dépend fortement de la nature des données observées, des techniques de modélisation utilisées et de ce que l'on veut prédire. En particulier, une étape préliminaire à la prediction est l'estimation des paramètres (aussi appelée étape de calibration) intervenant dans la modélisation du phénomène. Ce sera l'objet du Chapitre 3.

Nous présentons maintenant un cadre classique dans lequel s'inscrivent beaucoup de problèmes d'apprentissage.

2.2.1 Cadre paramétrique classique

Le cadre de l'apprentissage statistique est généralement introduit dans un contexte non paramétrique. Nous choisirons d'introduire les quelques notions suivantes dans un cadre paramétrique afin de mieux cerner le but de notre propos. Toutefois, nous adopterons un cadre

plus formel à la section 2.3 pour définir quelques notions qui vont nous être utiles dans les chapitres qui suivront. Ensuite, nous repasserons à un cadre paramétrique à la section 2.5 où il s'agit de poser les problématiques sur lesquelles nous nous sommes penchés concernant l'étude des computer experiments dans le cadre de l'apprentissage statistique.

Considérons une base de données $Z_1 = (\mathbf{X}_1, Y_1), ..., Z_n = (\mathbf{X}_n, Y_n)$ que nous supposons i.i.d de loi inconnue $Q^{\mathbf{z}}$, avec $\mathbf{X}_i \in \mathcal{X} \subset \mathbb{R}^d$ et $Y_i \in \mathcal{Y} \subset \mathbb{R}$.
Le but est de construire un modèle (aussi appelée *fonction de prédiction*) qui prédit la sortie Y associée à toute nouvelle entrée $\mathbf{X}$, où la donnée $Z = (\mathbf{X}, Y)$ sera supposée distribuée selon $Q^{\mathbf{z}}$ et indépendante des observations.

On se donne une famille paramétrique de modèles

$$\{\mathbf{x} \in \mathcal{X} \mapsto h(\mathbf{x}, \boldsymbol{\theta}) \in \mathcal{Y}\,, \quad \boldsymbol{\theta} \in \Theta\}\,,$$

où le modèle $h(\mathbf{x}, \boldsymbol{\theta})$ peut être - soit de nature purement mathématique, i.e décomposition sur une base polynômiale etc... - soit issu d'une modélisation physique du phénomène, i.e Computer Experiments (cf Chapitre 1).
Avant de pouvoir prédire, il est nécessaire de passer par une étape d'estimation des paramètres.
On note $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ une *fonction de perte* qui mesure la perte (la distance) entre la sortie réelle y et la sortie prédite $y' = h(\mathbf{x}, \boldsymbol{\theta})$, par exemple :

– Classification :
$$l(y, y') = \mathbb{1}_{y \neq y'}$$
– Regression L_2 :
$$l(y, y') = (y - y')^2\,.$$

Le risque associé au choix d'un paramètre $\boldsymbol{\theta}$ dans Θ est donnée par

$$\mathcal{R}(\boldsymbol{\theta}) \quad = \quad \mathbb{E}_{Q^{\mathbf{z}}}\left(l(Y, h(\mathbf{X}, \boldsymbol{\theta}))\right)\,. \tag{2.1}$$

Le but de l'apprentissage est d'approcher au mieux le "meilleur" paramètre $\boldsymbol{\theta}$, i.e celui qui minimise le risque $\mathcal{R}(\boldsymbol{\theta})$, que l'on note

$$\boldsymbol{\theta}^\star = \underset{\boldsymbol{\theta} \in \Theta}{\text{Argmin}}\, \mathcal{R}(\boldsymbol{\theta})\,.$$

On note $\mathcal{R}(\boldsymbol{\theta}^\star)$ le risque *idéal* (i.e le plus petit risque atteignable). Cependant, un tel paramètre n'est pas accessible car il nécessite la minimisation du risque $\mathcal{R}(\cdot)$ qui dépend de la mesure $Q^{\mathbf{z}}$ inconnue.
Un *algorithme d'apprentissage* ou *procédure d'estimation* fournit un paramètre $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(Z_1, ..., Z_n)$ qui dépend des observations ($\widehat{\boldsymbol{\theta}}$ est donc aléatoire). La qualité d'une procédure d'estimation est donnée par l'étude du comportement de la variable aléatoire $\mathcal{R}(\widehat{\boldsymbol{\theta}})$, plus particulièrement par la quantité positive

$$\mathcal{R}(\widehat{\boldsymbol{\theta}}) - \mathcal{R}(\boldsymbol{\theta}^\star)$$

que l'on appelle *excès de risque* (sur le modèle $\{\mathbf{x} \in \mathcal{X} \mapsto h(\mathbf{x}, \boldsymbol{\theta}) \in \mathcal{Y}\,, \quad \boldsymbol{\theta} \in \Theta\}$). On donnera une définition plus générale de l'excès de risque à la section 2.3.6.

Minimisation du risque empirique

La procédure la plus classique d'estimation est la minimisation de la fonction suivante

$$\mathcal{R}_n(\boldsymbol{\theta}) \quad = \quad \mathbb{E}_{Q_n^{\mathbf{z}}}\left(l(Y, h(\mathbf{X}, \boldsymbol{\theta}))\right) \tag{2.2}$$

où $Q_n^{\mathbf{z}}$ (connue) est la mesure empirique de $Q^{\mathbf{z}}$ (inconnue), construite à partir de données $Z_1, ..., Z_n$, donnée par

$$Q_n^{\mathbf{z}} = \frac{1}{n}\sum_{i=1}^{n} \delta_{Z_i}\,.$$

On réécrit (2.2)

$$\mathcal{R}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} l(Y_i, h(\mathbf{X}_i, \boldsymbol{\theta}))\,,$$

et la procédure générale d'estimation s'écrit alors

$$\widehat{\boldsymbol{\theta}} \quad = \quad \underset{\boldsymbol{\theta}\in\Theta}{\text{Argmin}} \frac{1}{n}\sum_{i=1}^{n} l(Y_i, h(\mathbf{X}_i, \boldsymbol{\theta}))\,. \tag{2.3}$$

On retrouve bien la procédure des moindres carrés en considérant la perte $l(y, y') = (y - y')^2$

$$\widehat{\boldsymbol{\theta}}_{reg} = \underset{\boldsymbol{\theta}\in\Theta}{\text{Argmin}} \frac{1}{n}\sum_{i=1}^{n} (Y_i - h(\mathbf{X}_i, \boldsymbol{\theta}))^2\,. \tag{2.4}$$

Remarque 2.2.1. Souvent, on rajoute un terme de pénalisation (ou de régularisation) à la procédure (2.3). Par exemple, considérons le cas de la regression (2.4) et supposons que $h(\mathbf{x}, \boldsymbol{\theta})$ est linéaire en $\boldsymbol{\theta}$, on établit les procédures *pénalisées* ci-dessous

$$\widehat{\boldsymbol{\theta}}_{ridge} \quad = \quad \underset{\boldsymbol{\theta}\in\Theta}{\text{Argmin}} \frac{1}{n}\sum_{i=1}^{n} (Y_i - h(\mathbf{X}_i, \boldsymbol{\theta}))^2 + \lambda \left\|\boldsymbol{\theta}\right\|_2^2\,, \quad \left\|\boldsymbol{\theta}\right\|_2 = \left(\sum_{l=1}^{k} \theta_l^2\right)^{1/2} \tag{2.5}$$

$$\widehat{\boldsymbol{\theta}}_{lasso} \quad = \quad \underset{\boldsymbol{\theta}\in\Theta}{\text{Argmin}} \frac{1}{n}\sum_{i=1}^{n} (Y_i - h(\mathbf{X}_i, \boldsymbol{\theta}))^2 + \lambda \left\|\boldsymbol{\theta}\right\|_1\,, \quad \left\|\boldsymbol{\theta}\right\|_1 = \sum_{l=1}^{k} |\theta_l|\,. \tag{2.6}$$

La première procédure est appelée *Ridge Regression* et la seconde *Lasso Regression*. On trouvera une étude assez étendue sur les techniques d'optimisation sous-jacentes à ces procédures dans [4].
Nous verrons dans la suite que les procédures (2.5) et (2.6) sont des cas particuliers d'estimateurs construits à partir de *contrastes pénalisés*. Voir la Section 2.4 pour le cadre général, et voir la Section 2.6 pour le cas paramétrique.

Approche asymptotique et non asymptotique

Une importante littérature a été (et est toujours) consacrée à l'étude de la procédure (2.3) en prenant en compte la "régularité" de la fonction de perte $l(\cdot, \cdot)$. Citons les travaux de P. Massart [6] ainsi que les références qui s'y trouvent.
Considérons le cas particulier où l'on formule les hypothèses suivantes

- $Y_i = h(\mathbf{X}_i, \boldsymbol{\theta}^\star) + \varepsilon_i\,, \quad i = 1, ..., n$
- $\varepsilon_i \sim \mathcal{N}(0, 1)$ et indépendants des $\mathbf{X}_i$.

L'estimateur $\widehat{\boldsymbol{\theta}}_{reg}$ (2.4) est celui du maximum de vraisemblance, il hérite donc des propriétés qui lui sont bien connues.
Dans le cas général, l'étude de telles procédures est établie en utilisant la théorie des *processus empiriques* que nous présentons en Annexe. Deux types d'études sont envisageables :

- *Etude non asymptotique :*

  on rappelle que l'excès de risque d'une procédure d'estimation s'écrit $\mathcal{R}(\widehat{\boldsymbol{\theta}}) - \mathcal{R}(\boldsymbol{\theta}^\star)$ et est positif. L'objet d'une étude non asymptotique est de déterminer des *bornes supérieures* de l'excès de risque, par exemple en moyenne

  $$\mathbb{E}\mathcal{R}(\widehat{\boldsymbol{\theta}}) - \mathcal{R}(\boldsymbol{\theta}^\star) \leq B_n$$

  ou bien une borne valable avec probabilité au moins $1 - \epsilon$

  $$\mathcal{R}(\widehat{\boldsymbol{\theta}}) - \mathcal{R}(\boldsymbol{\theta}^\star) \leq B_n^\epsilon \,,$$

  etc...
  Les bornes B_n et B_n^ϵ sont des constantes qui peuvent dépendre de n et la remarque importante est que ces inégalités sont valables pour tout $n \geq 1$. Ce type d'inégalités quantifient l'erreur de prédiction relative à la procédure $\widehat{\boldsymbol{\theta}}$ par rapport à celle donnée par $\boldsymbol{\theta}^\star$.
  L'avantage, parmi d'autres, que peut présenter une telle approche en pratique est sa validité même si l'on dispose de peu de données observées. Toutefois, il convient de souligner que les bornes intervenant dans ces inégalités sont délicates à calculer et sont souvent "pessimistes" (i.e trop élevées). Nous verrons plus précisément comment obtenir de telles inégalités dans le Chapitre 3 .

- *Etude asymptotique :*
  (C'est un sujet que nous n'avons pas encore abordé dans ce travail).
  Ici, il s'agit d'étudier les propriétés de convergence de l'estimateur (de la procédure) $\widehat{\boldsymbol{\theta}}$. Les résultats espérés sont du type (*Théorème Central Limite*)

  $$r_n(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star) \rightsquigarrow \mathcal{N}(0, \Sigma) \,,$$

  où r_n est appelé *vitesse de convergence* (par exemple, $r_n = \sqrt{n}$) et Σ est une matrice de variance-covariance. Le symbole "$\rightsquigarrow$" désigne généralement la *convergence étroite*, qui correspond à la *convergence en loi* dans le cas de variables aléatoires.

Dans cette thèse, nous adoptons le cadre non asymptotique.

2.2.2 Prédiction de modèle

Une fois la procédure d'estimation établie (et étudiée), le paramètre $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(Z_1, ..., Z_n)$ (où on rappelle que $Z_i = (\mathbf{X}_i, Y_i)$) va désormais servir à la prédiction souhaitée. Pour l'exemple de la régression, nous avions

$$\widehat{\boldsymbol{\theta}}_{reg} = \underset{\boldsymbol{\theta} \in \Theta}{\text{Argmin}} \frac{1}{n} \sum_{i=1}^{n} (Y_i - h(\mathbf{X}_i, \boldsymbol{\theta}))^2 \,,$$

ainsi la prédiction de la sortie y^* associée à l'entrée $\mathbf{x}^*$ est

$$h(\mathbf{x}^*, \widehat{\boldsymbol{\theta}}_{reg}) \,.$$

Dans ce que nous venons de voir, le terme de "prédiction" signifie "à une entrée, je prédis la sortie". Nous pouvons qualifier ce type de prédiction de *prédiction de modèle* dans le sens où on cherche à reconstruire le lien entre l'*entrée* $\mathbf{X}$ et la *sortie* Y. C'est le cas en régression ou en classification par exemple. Autrement dit, on cherche à *apprendre* la **structure** qui lie $\mathbf{X}$ à Y. La théorie de l'apprentissage statistique permet de bien appréhender les problématiques de prédiction de ce genre.
En pratique nous sommes très souvent amenés à vouloir prédire non pas "une sortie" mais un "comportement de la sortie". De plus, il convient de prendre en compte que les données $\mathbf{X}_i$ qui produisent les Y_i ne sont pas nécessairement observées.

2.2.3 Prédiction de caractéristiques ou de quantités d'intérêt

Certains comportements d'un phénomène aléatoire peuvent être critiques au vu du contexte dans lequel ils interviennent. Il est donc indispensable de savoir prédire une caractéristique jouant un rôle crucial, que ce soit en termes de sécurité, financier ou autre.
En gardant les mêmes notations que précédemment, il s'agit cette fois de prédire une caractéristique de la variable aléatoire Y :
– moyenne
– quantile
– probabilité de dépassement
– distribution de densité
– queue de distribution
– etc...
Par exemple, intéressons-nous à la *quantité d'intérêt* [1]

$$\mathbb{P}(Y > s)\,. \tag{2.7}$$

Cette quantité intervient souvent en fiabilité où le comportement extrême d'un phénomène (mécanique, thermique etc...) peut avoir des conséquences néfastes sur un système.
Rappelons que l'on dispose des observations $Z_1 = (\mathbf{X}_1, Y_1), ..., Z_n = (\mathbf{X}_n, Y_n)$ et d'un modèle paramétrique

$$\{\mathbf{x} \in \mathcal{X} \mapsto h(\mathbf{x}, \boldsymbol{\theta}) \in \mathcal{Y}\,, \quad \boldsymbol{\theta} \in \Theta\}\,.$$

Afin de prédire la quantité (2.7), il paraîtrait assez naturel d'estimer le paramètre $\boldsymbol{\theta}$ par régression, $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}_{reg}$, donné par la procédure (2.4). Ensuite, une prédiction de (2.7) serait

$$\mathbb{P}(h(\mathbf{X}, \widehat{\boldsymbol{\theta}}_{reg}) > s)\,, \tag{2.8}$$

où la probabilité est prise sous la loi de $\mathbf{X}$. Il faut souligner qu'on a utilisé une procédure de prédiction de modèle donnant $\widehat{\boldsymbol{\theta}}_{reg}$ pour prédire une quantité portant sur la variable Y. (L'estimateur $\widehat{\boldsymbol{\theta}}_{reg}$ trouvait sa légitimité lorsque l'on voulait prédire la sortie Y associée à une entrée $\mathbf{X}$).

Alors, une question se pose :

Existe-t-il une procédure d'estimation "optimale" pour prédire une quantité d'intérêt ?

Ou bien, de manière générale,

Quelle est la contribution de l'erreur due à la procédure d'estimation sur l'erreur de prédiction ?

1. le mot *quantité d'intérêt* est très souvent employé en pratique pour désigner une caractéristique d'un aléa. Nous proposerons une formalisation à la Définition 2.3.1 .

Nous tenterons d'élucider cette question dans le Chapitre 4 où nous formaliserons ce que nous entendons par "erreur de procédure" et "erreur de prédiction". Dans le petit exemple que nous sommes en train de mener, une "erreur de prédiction" serait

$$Err_{\mathbb{P}(Y>s)}(\widehat{\boldsymbol{\theta}}_{reg}) = \Big(\mathbb{P}(h(\mathbf{X}, \widehat{\boldsymbol{\theta}}_{reg}) > s) - \mathbb{P}(Y > s)\Big)^2 .$$

La question de procédure "optimale" pose le problème de l'existence d'une procédure d'estimation $\widehat{\boldsymbol{\theta}}_{opt}$ telle que

$$Err_{\mathbb{P}(Y>s)}(\widehat{\boldsymbol{\theta}}_{opt}) \leq Err_{\mathbb{P}(Y>s)}(\widehat{\boldsymbol{\theta}}_{reg}), \tag{2.9}$$

avec la même base de données.

Nous reviendrons sur ce type d'interrogation après avoir défini quelques notions générales utiles à nos développements.

2.3 Formalisme général

Dans tout le manuscript, nous garderons les notations suivantes : $Q^{\mathbf{z}}$ est la distribution du couple $Z = (\mathbf{X}, Y)$, Q la distribution de Y et $P^{\mathbf{x}}$ celle de $\mathbf{X}$. On supposera $P^{\mathbf{x}}$ connue (ou bien un échantillon de "grande taille" est disponible).

2.3.1 Espace caractéristique $\mathcal{F}$ et quantité d'intérêt $\rho_{\mathcal{F}}$

On note $(\mathcal{F}, \|\cdot\|_{\mathcal{F}})$ un *espace caractéristique* (*feature space*) de $Q^{\mathbf{z}}$ c'est à dire l'espace correspondant à une caractéristique qu'on notera $\rho_{\mathcal{F}} \in \mathcal{F}$ de la mesure $Q^{\mathbf{z}}$. Par exemple, on peut être intéressé par la prédiction de l'espérance conditionnelle $\mathbb{E}(Y/\mathbf{X} = \mathbf{x})$ qui peut être qualifiée de caractéristique "structurelle" (on cherche la structure qui lie les entrées aux sorties). Dans ce cas, l'espace $\mathcal{F}$ sera un sous-espace des fonctions de $\mathcal{X}$ dans $\mathcal{Y}$ et $\rho_{\mathcal{F}}(\mathbf{x}) = \mathbb{E}(Y/\mathbf{X} = \mathbf{x})$. Ou bien, on peut être intéressé par des caractéristiques "marginales" : par exemple, la densité de la variable Y, sa moyenne, un quantile, une probabilité de dépassement etc... (voir Tableau 2.1 page 24). Ici, l'espace $\mathcal{F}$ sera soit un sous-espace des fonctions de $\mathcal{Y}$ dans $\mathbb{R}$ (dans le cas de la densité par exemple), soit une partie de $\mathbb{R}$ (pour des quantités scalaires telles que la moyenne, la variance etc...).
Nous donnons ci-après une définition d'une *caractéristique* ou *quantité d'intérêt* définie sur un espace de probabilité (ou un espace de fonctions de distribution).

Définition 2.3.1. Quantité d'intérêt.
Soit un ensemble Ξ quelconque.
On note Π_Ξ l'ensemble des mesures de probabilité (par rapport à la mesure de Lebesgue) sur Ξ et $\mathbb{F}_\Xi$ l'ensemble des fonctions de distribution sur Ξ. Pour tout espace $\mathcal{F}$, on définit la **caractéristique** ou **quantité d'intérêt** dans $\mathcal{F}$ comme une application (une fonctionnelle sur Π_Ξ) $\rho_{\mathcal{F}} : \Pi_\Xi \longrightarrow \mathcal{F}$ ou bien $\rho_{\mathcal{F}} : \mathbb{F}_\Xi \longrightarrow \mathcal{F}$ (une fonctionnelle sur $\mathbb{F}_\Xi$). En particulier, une quantité d'intérêt sera dite ***μ*-linéaire** si

$$\mu \in \Pi_\Xi , \quad \rho_{\mathcal{F}}(\mu) := \int_\Xi w_{\mathcal{F}}(\xi)\, \mu(d\xi) ,$$

où $w_{\mathcal{F}} : \Xi \to \mathcal{F}$ est une application qui dépend de la caractéristique $\rho_{\mathcal{F}}$ considérée. De manière équivalente, en écrivant $\mu(d\xi) = d\,F(\xi)$, on dira que $\rho_{\mathcal{F}}$ est F**-linéaire** si

$$F \in \mathbb{F}_\Xi , \quad \rho_{\mathcal{F}}(F) := \int_\Xi w_{\mathcal{F}}(\xi)\, d\,F(\xi) .$$

En fonction du contexte, on choisira l'écriture $\rho_{\mathcal{F}}(\mu)$ ou bien $\rho_{\mathcal{F}}(F)$.

Par exemple, la moyenne et la variance sont des quantités d'intérêt μ-linéaires alors qu'un quantile ne l'est pas. Donnons quelques exemples dans le cas où $\mu = Q^{\mathbf{z}}$ (avec $F_{\mathbf{z}}$ la fonction de distribution associée) et $\Xi = \mathcal{Z} = \mathcal{X} \times \mathcal{Y}$.

Moyenne de Y : $\mathbb{E}_Q(Y)$.
Dans ce cas $\mathcal{F} = \mathbb{R}$ et en posant $w_{\mathcal{F}} : \mathcal{Z} \to \mathbb{R}$ tel que

$$w_{\mathcal{F}}(\mathbf{x}, y) = y\,,$$

on a bien $\rho_{\mathcal{F}}(Q^{\mathbf{z}}) = \int_{\mathcal{X}\times\mathcal{Y}} y\, Q^{\mathbf{z}}(d\mathbf{x}, dy) = \mathbb{E}_Q(Y)$.

Probabilité de dépassement : $\mathbb{P}(Y > s)$.
Dans ce cas $\mathcal{F} = [0,1]$ et en posant $w_{\mathcal{F}} : \mathcal{Z} \to [0,1]$ tel que

$$w_{\mathcal{F}}(\mathbf{x}, y) = \mathbb{1}_{y>s}\,,$$

on a $\rho_{\mathcal{F}}(Q^{\mathbf{z}}) = \mathbb{P}(Y > s)$.

Espérance conditionnelle : $\mathbf{u} \mapsto \mathbb{E}(Y/\mathbf{X} = \mathbf{u})$.
Dans ce cas $\mathcal{F} = \{$ functions $\mathcal{X} \to \mathcal{Y}\}$ et en posant

$$w_{\mathcal{F}}(\mathbf{x}, y)(\mathbf{u}) = \frac{y}{p^{\mathbf{x}}(\mathbf{x})}\delta_{\mathbf{u}}(\mathbf{x})\,,$$

où $p^{\mathbf{x}}$ est la densité de $\mathbf{X}$, on retrouve bien $\rho_{\mathcal{F}}(Q^{\mathbf{z}})(\mathbf{u}) = \mathbb{E}(Y/\mathbf{X} = \mathbf{u})$.

Densité de Y : $v \mapsto f(v)$.
Dans ce cas $\mathcal{F} =\{$ densités sur $\mathcal{Y}\}$ et en posant

$$w_{\mathcal{F}}(\mathbf{x}, y)(v) = \delta_v(y)\,,$$

on a $\rho_{\mathcal{F}}(Q^{\mathbf{z}})(v) = f(v)$. Ou encore, $\rho_{\mathcal{F}}(F_{\mathbf{z}})(v) = \frac{d}{dy}F_{\mathbf{z}}(y)|_{y=v}$.

Quantile d'ordre α de Y : q_α .
Dans ce cas $\mathcal{F} = \mathbb{R}$ et en notant F_Y la fonction de distribution de Y, on a $\rho_{\mathcal{F}}(F_Y) = F_Y^{-1}(\alpha)$ (en fait $= \inf_{u\in\mathbb{R}}\{u\,,\, F_Y(\alpha) \geq u\}$)

Comme la mesure $Q^{\mathbf{z}}$ du phénomène d'intérêt est fixée, on notera pour toute la suite une quelconque quantité d'intérêt associée à $Q^{\mathbf{z}}$

$$\rho_{\mathcal{F}} := \rho_{\mathcal{F}}(Q^{\mathbf{z}})\,,$$

où $\mathcal{F}$ est l'espace où vit la caractéristique.

2.3.2 Un détour sur une notion de régularité de quantités d'intérêt

La Définition 2.3.1 suggère d'introduire une notion de *régularité de quantités d'intérêt* basée sur la "régularité" de l'application $\rho_{\mathcal{F}} : \mathbb{F}_{\Xi} \longrightarrow \mathcal{F}$. La difficulté est bien évidemment le fait que les espaces $\mathbb{F}_{\Xi}$ et $\mathcal{F}$ ne sont pas des espaces standards et qu'il faut définir une notion de dérivabilité entre ces espaces (qu'il faut au préalable normer...).
Supposons que Ξ est un intervalle de $\mathbb{R}$ et munissons l'espace $\mathcal{F}$ d'une norme $\|\cdot\|_{\mathcal{F}}$. Ensuite, notons $(\mathbb{D}, \|\cdot\|)$ l'espace de Skorohod (i.e des fonctions càdlàg) muni d'une métrique $\|\cdot\|$ pouvant être la distance de Levy, de Prohorov ou encore de Kolmogorov, parmi d'autres. Il y a alors un sens à vouloir définir une notion de dérivabilité de l'application $\rho_{\mathcal{F}} : \mathbb{F}_{\Xi} \subset \mathbb{D} \longrightarrow \mathcal{F}$. D'après la littérature, il y a trois principales notions de dérivation fonctionnelle, de la plus forte à la plus faible : différentiabilité au sens de Fréchet, de Hadamard ou encore de Gâteau. Considérons cette dernière.
Soit une distribution $F \in \mathbb{F}_{\Xi}$ et G un élément de $\mathbb{D}$, la dérivée au sens de Gâteau de $\rho_{\mathcal{F}}$ en F dans la direction G est donnée par

$$\lim_{t \to 0} \frac{\rho_{\mathcal{F}}(F + t\,G) - \rho_{\mathcal{F}}(F)}{t} := D_F(G) \in \mathcal{F}\,,$$

quand une telle limite existe. La "régularité" peut par exemple être interprétée au vu du comportement de $G \mapsto D_F(G)$. Une direction G (ou perturbation) bien connue en Statistique Robuste [5] est la direction

$$G_{\xi_0} = \mathbb{1}_{[\xi_0, +\infty[} - F\,,$$

donnant lieu à la définition de la *fonction d'influence* γ

$$\gamma : \xi \mapsto D_F(G_\xi) \tag{2.10}$$

qui mesure la sensibilité à l'introduction d'une masse de Dirac.

Remarque 2.3.1. La quantité $\rho_{\mathcal{F}}(F)$ sera dite *robuste* (pour F) si la fonction d'influence γ est **bornée**. En effet, pour une quantité robuste, on s'attend à ce qu'une variation infinitésimale des données ($F \leftrightarrow F + tG$) n'ait pas beaucoup d'impact.

On renvoie à [5] pour plus de précisions concernant le domaine de la robustesse statistique. Ici, nous tentons seulement de comprendre ce qui peut se cacher derrière la notion (qui n'est pas tout à fait claire) de "régularité de quantités d'intérêt". Il est bien connu que la moyenne n'est pas une quantité robuste alors qu'intuitivement on sent que c'est une quantité "simple" ... régulière ?
Dans le cas où $\rho_{\mathcal{F}}(F)$ est **F-linéaire**, on a

$$D_F(G) = \int_{\Xi} w_{\mathcal{F}}(\xi)\, dG(\xi)\,,$$

et en particulier

$$D_F(G_{\xi_0}) = w_{\mathcal{F}}(\xi_0) - \rho_{\mathcal{F}}(F)\,.$$

On en déduit que pour des quantités d'intérêt **F-linéaires** on a $\gamma : \xi \mapsto w_{\mathcal{F}}(\xi) - \rho_{\mathcal{F}}(F)$.
Donnons trois exemples de fonctions γ.
Soit X une variable aléatoire sur un ensemble Ξ de distribution F, de densité f.
Dans le cas de la moyenne, $w_{\mathcal{F}}(\xi) = \xi$ d'où

$$\gamma_{moy} : \xi \mapsto \xi - \mathbb{E}(X)$$

(γ_{moy} n'est pas bornée d'où la non robustesse de la moyenne).
Dans le cas de la variance, $w_{\mathcal{F}}(\xi) = (\xi - \mathbb{E}(X))^2$ d'où

$$\gamma_{var} : \xi \mapsto (\xi - \mathbb{E}(X))^2 - \mathrm{Var}(X)\,.$$

Pour la quantité d'intérêt - quantile d'ordre α (q_α) - $\rho_{\mathcal{F}}(F)$ **n'est pas F-linéaire** puisque $\rho_{\mathcal{F}}(F) = F^{-1}(\alpha)$. Le calcul de γ n'est pas aussi direct que précédemment, il faut revenir à la dérivation au sens de Gâteau. En notant $F_t = F + t\,G_\xi = (1-t)F + t\,\mathbb{1}_{[\xi,+\infty[}$, on a par définition

$$\gamma_{q_\alpha}(\xi) = \lim_{t\to 0} \frac{\rho_{\mathcal{F}}(F_t) - \rho_{\mathcal{F}}(F)}{t} = \frac{F_t^{-1}(\alpha) - F^{-1}(\alpha)}{t}\,,$$

ou encore

$$\gamma_{q_\alpha}(\xi) = \frac{d}{dt} F_t^{-1}(\alpha)|_{t=0}\,.$$

Pour obtenir cette dérivée, partons de l'identité

$$\alpha = F_t \circ F_t^{-1}(\alpha) = (1-t)\,F \circ F_t^{-1}(\alpha) + t\,\mathbb{1}_{[\xi,+\infty[}(F_t^{-1}(\alpha))\,,$$

que l'on dérive

$$0 = -F \circ F_t^{-1}(\alpha) + (1-t)\frac{d}{dt}F_t^{-1}(\alpha)\,f(F_t^{-1}(\alpha)) + \mathbb{1}_{[\xi,+\infty[}(F_t^{-1}(\alpha)) + t\frac{d}{dt}\mathbb{1}_{[\xi,+\infty[}(F_t^{-1}(\alpha))\,,$$

où $f = dF/d\xi$. En prenant $t = 0$, il vient

$$0 = -\alpha + \frac{d}{dt}F_t^{-1}(\alpha)|_{t=0}\,f(F^{-1}(\alpha)) + \mathbb{1}_{[\xi,+\infty[}(F^{-1}(\alpha))\,,$$

et on obtient finalement

$$\gamma_{q_\alpha} : \xi \mapsto \frac{\alpha - \mathbb{1}_{]-\infty,q_\alpha]}(\xi)}{f(q_\alpha)}\,.$$

On trace à la Figure 2.1 les différentes fonctions γ calculées dans le cas où X est une variable gaussienne standard en considérant $\alpha = 0.5$ (i.e q_α est la médiane).
On retrouve bien le fait que la moyenne et la variance ne sont pas des quantités *robustes* car γ_{moy} et γ_{var} ne sont pas bornées, alors que γ_{q_α} l'est. Autrement-dit la médiane est robuste. Cependant, la courbe γ_{q_α} présente une singularité en 0 (en fait en q_α) alors que γ_{moy} et γ_{var} sont tout à fait régulières.

Une étude plus rigoureuse et plus approfondie permettrait peut être de mieux comprendre l'intérêt du phénomène de *régularité* que nous avons tenté de mettre en évidence.

2.3.3 Fonction de contraste Ψ et de risque $\mathcal{R}_\Psi$

Après avoir défini ce qu'était une caractéristique (2.3.1), nous allons maintenant vouloir un moyen de la "qualifier" (de la contraster).

Definition 2.3.1. Contraste et Risque.
Un $\mathcal{F}$-contraste (ou contraste s'il n'y a pas d'ambiguïté) est une application du type

$$\begin{aligned} \Psi : \mathcal{F} &\longrightarrow L_1(Q^{\mathbf{z}}) \\ \rho &\longmapsto \Psi(\rho,\cdot) : (\mathbf{x},y) \in \mathcal{X}\times\mathcal{Y} \longmapsto \Psi(\rho,(\mathbf{x},y)) \end{aligned} \tag{2.11}$$

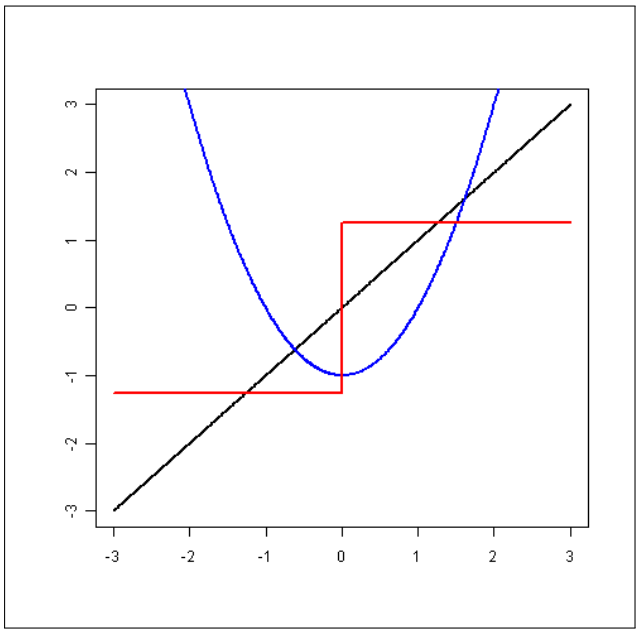


FIGURE 2.1 – Tracé des courbes γ_{moy} (noir), γ_{var} (bleue) et γ_{q_α} (rouge).

tel qu'il existe un **unique** élément $\rho_{\mathcal{F}} \in \mathcal{F}$ tel que

$$\rho_{\mathcal{F}} = \underset{\rho \in \mathcal{F}}{\text{Argmin}}\, \mathcal{R}_\Psi(\rho) \tag{2.12}$$

où

$$\mathcal{R}_\Psi(\rho) := \mathbb{E}_{Q^Z}\, \Psi\left(\rho\,,\, Z\right)$$

est appelé Ψ-risque (ou risque) de ρ.

Dans la littérature, l'élément $\rho_{\mathcal{F}}$ est souvent qualifié de *cible*.
On donne quelques exemples de contrastes.

Example 2.3.1. Exemples de contrastes.
- $\mathcal{F}$ = {fonctions de $\mathcal{X}$ dans $\mathcal{Y}$}
  regression-contrast

$$\Psi\left(\rho, (\mathbf{x}, y)\right) = (y - \rho(\mathbf{x}))^2$$

- $\mathcal{F} = \mathbb{R}$
  mean-contrast

$$\Psi\left(\rho, (\mathbf{x}, y)\right) = \Psi\left(\rho, y\right) = (y - \rho)^2$$

- $\mathcal{F}$={fonctions de densités de $\mathcal{Y}$}
  log**-contrast**

$$\Psi\left(\rho, (\mathbf{x}, y)\right) = \Psi\left(\rho, y\right) = -\log \rho(y)$$

  L_2-**contrast**

$$\Psi\left(\rho, (\mathbf{x}, y)\right) = \Psi\left(\rho, y\right) = ||\rho||_2^2 - 2\,\rho(y)$$

- etc...

Remarque 2.3.2. On notera que la définition du contraste est générale (dans notre contexte) dans le sens où elle inclut les contrastes à valeurs dans $L_1(Q)$ (on rappelle que Q est la mesure de probabilité associée à Y), c'est le cas du log-contrast, L_2-contrast et *mean*-contrast donnés en exemple. La justification triviale est bien sûr le fait que

$$L_1(Q) \subset L_1(Q^{\mathbf{z}}) \,,$$

puis qu'une fonction définie sur $\mathcal{Y}$ peut être vue comme une fonction définie sur $\mathcal{X} \times \mathcal{Y}$. Ce point de vue sera commode lorsque l'on analysera des procédures statistiques issues de divers contrastes (Chapitre 4). On pourra également se restreindre au contraste à valeurs dans $L_1(Q)$. Ce sera le cas dans le Chapitre 3.
Un contraste peut aussi être défini à valeur dans $L_1(\mu)$ pour une quelconque mesure μ sur un espace T.

D'après la remarque précédente, le risque associé à un contraste Ψ à valeurs dans $L_1(Q)$ s'écrit

$$\mathcal{R}_\Psi(\rho) = \mathbb{E}_{Q^{\mathbf{z}}} \Psi(\rho, Z) = \mathbb{E}_Q \Psi(\rho, Y) \,.$$

Dans le Tableau 2.1, on donne un récapitulatif de quelques espaces de caractéristiques avec les cibles et les contrastes associés.

Nature de la prédiction	Espace $\mathcal{F}$	Caractéristique $\rho_{\mathcal{F}} \in \mathcal{F}$	$\mathcal{F}$-contraste
Esp. conditionnelle	$\mathcal{F} = L_2(P^{\mathbf{x}})$	$\rho_{\mathcal{F}} : \mathbf{x} \mapsto \mathbb{E}(Y/\mathbf{X} = \mathbf{x})$	$\Psi(\rho,(\mathbf{x},y)) = (y - \rho(\mathbf{x}))^2$
Densité	$\mathcal{F} = \{$ densités sur $\mathcal{Y}\}$	$\rho_{\mathcal{F}} : y \mapsto f(y)$	$\Psi(\rho, y) = -\log(\rho)(y)$
Moyenne	$\mathcal{F} = \mathbb{R}$	$\rho_{\mathcal{F}} = \mathbb{E}(Y) \in \mathbb{R}$	$\Psi(\rho, y) = (y - \rho)^2$
Probabilité	$\mathcal{F} = [0,1]$	$\rho_{\mathcal{F}} = \mathbb{P}(Y > s) \in [0,1]$	$\Psi(\rho, y) = (1_{y>s} - \rho)^2$

TABLE 2.1 – Exemple d'espaces caractéristiques et de quantités d'intérêt associées à un contraste.

2.3.4 De la fonction de perte $l(\cdot,\cdot)$ à la notion de contraste

Revenons un instant au cas de la régression avec la fonction de perte donnée par

$$l(y, h(\mathbf{x}, \boldsymbol{\theta})) = (y - h(\mathbf{x}, \boldsymbol{\theta}))^2 \,.$$

On peut voir cette *perte* comme étant le contraste suivant
– $\mathcal{F} = \{\mathbf{x} \mapsto h(\mathbf{x}, \boldsymbol{\theta}),\ \boldsymbol{\theta} \in \Theta\}$: un élément de $\mathcal{F}$ est donc $\rho = h(\cdot, \boldsymbol{\theta})$
– on a

$$\Psi(h(\cdot, \boldsymbol{\theta}), (\mathbf{x}, y)) = (y - h(\mathbf{x}, \boldsymbol{\theta}))^2 \,.$$

Nous préférons l'écriture d'une "perte" sous le formalisme de contraste car cela met en évidence les caractéristiques des modèles numériques qui interviennent dans la procédure statistique utilisée. Cela donne également un cadre assez général permettant l'étude de procédures de prédiction que l'on verra au Chapitre 4.

Dans la section qui suit, nous allons voir comment une quantité d'intérêt (ou caractéristique) relative à la mesure $Q^{\mathbf{z}}$ (espérance conditionnelle $\mathbf{x} \mapsto \mathbb{E}(Y/\mathbf{X} = \mathbf{x})$, moyenne de Y, densité de Y, etc...) peut être caractérisée par un contraste.

2.3.5 Caractérisation de la quantité d'intérêt par contraste

Dans la précédente définition du contraste nous avons en particulier mentionné l'existence d'un unique élément $\rho_{\mathcal{F}} \in \mathcal{F}$, appelé cible, tel que

$$\rho_{\mathcal{F}} = \underset{\rho \in \mathcal{F}}{\text{Argmin}}\, \mathcal{R}_\Psi(\rho)\,.$$

En fait, cet élément dépend exclusivement du contraste Ψ (et de la mesure $Q^{\mathbf{z}}$, mais qui est fixée). De ce fait, une quantité d'intérêt (ou caractéristique) $\rho_{\mathcal{F}}$ appartenant à un espace $\mathcal{F}$ peut être *implicitement* définie par

$$\exists\, \Psi\, :\, \mathcal{F} \to L_1(Q^{\mathbf{z}})\,, \quad \rho_{\mathcal{F}} = \underset{\rho \in \mathcal{F}}{\text{Argmin}}\, \mathcal{R}_\Psi(\rho)\,, \tag{2.13}$$

sous condition d'existence d'un tel contraste. Ce sera le cas dans tout ce qui suivra.

Remarque 2.3.3. Il s'agit en quelque sorte d'une vision inverse de la définition d'un contraste, où il est demandé l'unique existence de $\rho_{\mathcal{F}}$, autrement dit, l'unique existence d'une "quantité d'intérêt".

Par conséquent, une quantité d'intérêt peut être - soit vue par sa définition donnée en (2.3.1), c'est à dire l'image d'une mesure de probabilité par une certaine application - ou bien, une quantité d'intérêt peut être vue comme un *argmin* d'un certain contraste moyenné (= risque). Nous verrons que ces deux visions sont complémentaires dans la problématique de prédiction de quantités d'intérêt (cf. Chapitre 4).

2.3.6 Modèle, projeté, risque absolu et idéal et excès de risque

Nous donnons quelques définitions générales.

Soit $\Psi\, :\, \mathcal{F} \longrightarrow L_1(Q^{\mathbf{z}})$ un $\mathcal{F}$-contraste et $\mathcal{R}_\Psi(\rho) = \mathbb{E}_{Q^{\mathbf{z}}}\, \Psi\,(\rho\,,\, Z)$ le risque associé. Rappelons également que

$$\rho_{\mathcal{F}} = \underset{\rho \in \mathcal{F}}{\text{Argmin}}\, \mathcal{R}_\Psi(\rho)\,.$$

On appelle **risque absolu** le risque donné par

$$\mathcal{R}^{\mathcal{F}} := \mathcal{R}_\Psi(\rho_{\mathcal{F}}) = \inf_{\rho \in \mathcal{F}} \mathcal{R}_\Psi(\rho)\,. \tag{2.14}$$

On appelle **modèle** un sous-ensemble $F \subset \mathcal{F}$ de $\mathcal{F}$.

De manière générale, nous considérerons un modèle F de la forme suivante.
Soit $\pi_{\mathcal{Z}} \subset \Pi_{\mathcal{Z}}$ un sous-ensemble de l'ensemble des mesures de probabilités sur $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, posons

$$F = \{\rho_{\mathcal{F}}(\mu)\,,\, \mu \in \pi_{\mathcal{Z}}\} \subset \mathcal{F}\,, \tag{2.15}$$

où $\rho_{\mathcal{F}}(\mu)$ est donné par (2.3.1). En pratique, l'ensemble $\pi_{\mathcal{Z}}$ nous sera donné par un computer experiment (cf Section 2.5).

Remarque 2.3.4. La définition (2.3.1) d'une quantité d'intérêt permet de "construire" un modèle F à partir d'un ensemble de mesures $\pi_{\mathcal{Z}}$ ou à partir d'un ensemble de fonctions de distribution $F_{\mathcal{Z}}$.

On appelle **projeté** de $\rho_{\mathcal{F}}$ sur F une caractéristique $\rho_F \in F$ qui satisfait

$$\rho_F := \underset{\rho \in F}{\text{Argmin}}\, \mathcal{R}_\Psi(\rho) \tag{2.16}$$

On appelle **risque idéal** le risque donné par

$$\mathcal{R}^F := \mathcal{R}_\Psi(\rho_F) = \inf_{\rho \in F} \mathcal{R}_\Psi(\rho)\,. \tag{2.17}$$

Notons que ρ_F n'est pas nécessairement unique. S'il est unique, ρ_F peut être vu comme la "meilleure" prédiction que l'on puisse avoir.
Le risque $\mathcal{R}_\Psi(\rho) = \mathbb{E}_{Q^{\mathbf{z}}} \Psi(\rho, Z)$ dépend de la mesure $Q^{\mathbf{z}}$ qui est inconnue. L'idée est de minimiser un autre risque.

On appelle **risque empirique** $\widehat{\mathcal{R}}_\Psi(\rho)$, qui dépend des observations, une approximation de $\mathcal{R}_\Psi(\rho)$.

On appelle **prédiction** l'estimateur

$$\widehat{\rho}_F = \underset{\rho \in F}{\text{Argmin}}\, \widehat{\mathcal{R}}_\Psi(\rho) \tag{2.18}$$

qui sera alors la prédiction effective de $\rho_{\mathcal{F}}$. La qualité de l'estimateur $\widehat{\rho}_F$ vis-à-vis de son objectif $\rho_{\mathcal{F}}$ est mesurée par son *excès de risque*, défini ci-après.

Définition 2.3.2. Excès de risque.
Soit Ψ un $\mathcal{F}$-contraste et $\mathcal{R}_\Psi$ son risque associé. L'*excès de risque* d'un élément $\rho \in \mathcal{F}$ est donné par

$$\mathcal{E}_\Psi(\rho) := \mathcal{R}_\Psi(\rho) - \mathcal{R}^{\mathcal{F}} \tag{2.19}$$

où $\mathcal{R}^{\mathcal{F}}$ est le *risque absolu* (2.14).

Nature de la prédiction	Caractéristique $\rho_{\mathcal{F}}$	Excès de risque $\mathcal{E}_\Psi$
Esp. conditionnelle	$\rho_{\mathcal{F}} : \mathbf{x} \mapsto \mathbb{E}(Y/\mathbf{X} = \mathbf{x})$	$\mathcal{E}_\Psi(\rho) = \lVert\rho_{\mathcal{F}} - \rho\rVert^2_{L_2(P^{\mathbf{x}})}$
Densité	$\rho_{\mathcal{F}} : y \mapsto f(y)$	$\mathcal{E}_\Psi(\rho) = KL(\rho_{\mathcal{F}}, \rho)$
Moyenne	$\rho_{\mathcal{F}} = \mathbb{E}(Y)$	$\mathcal{E}_\Psi(\rho) = (\rho_{\mathcal{F}} - \rho)^2$
Probabilité	$\rho_{\mathcal{F}} = \mathbb{P}(Y > s)$	$\mathcal{E}_\Psi(\rho) = (\rho_{\mathcal{F}} - \rho)^2$

TABLE 2.2 – Exemples d'excès de risque.

Le Tableau 2.2 donne quelques exemples d'excès de risque. On peut vérifier sur ces exemples que les contrastes donnés sont bien définis, en particulier que $\rho_{\mathcal{F}}$ est bien unique. En effet, remarquons tout d'abord que

$$\rho_{\mathcal{F}} = \underset{\rho \in \mathcal{F}}{\text{Argmin}}\, \mathcal{R}_\Psi(\rho) = \underset{\rho \in \mathcal{F}}{\text{Argmin}}\, \mathcal{E}_\Psi(\rho)$$

car le risque absolu $\mathcal{R}^{\mathcal{F}}$ est une constante. Ensuite, notons que les quantités $\|\rho_{\mathcal{F}} - \rho\|^2_{L_2(P^{\mathbf{x}})}$, $KL(\rho_{\mathcal{F}}, \rho)$ et $(\rho_{\mathcal{F}} - \rho)^2$ sont positives, et nulles si et seulement si $\rho = \rho_{\mathcal{F}}$. D'où l'unicité de $\rho_{\mathcal{F}}$.
La quantité $KL(\rho_1, \rho_2)$ est la dissemblance de Kullback-Leibler donnée par

$$KL(\rho_1, \rho_2) = \int_{\mathcal{Y}} \log\left(\frac{\rho_1}{\rho_2}\right)(y)\, \rho_1(y)\, dy\,.$$

L'excès de risque de l'estimateur $\widehat{\rho}_{\mathcal{F}}$ (2.18) est donnée par

$$\mathcal{E}_{\Psi}(\widehat{\rho}_F) = \mathcal{R}_{\Psi}(\widehat{\rho}_F) - \mathcal{R}^{\mathcal{F}}\,. \tag{2.20}$$

On remarque que cet excès de risque peut se décomposer en la somme de deux quantités positives

$$\mathcal{E}_{\Psi}(\widehat{\rho}_F) = \underbrace{\mathcal{R}_{\Psi}(\widehat{\rho}_F) - \mathcal{R}^F}_{\text{terme de variance}} + \underbrace{\mathcal{R}^F - \mathcal{R}^{\mathcal{F}}}_{\text{terme de biais}}\,. \tag{2.21}$$

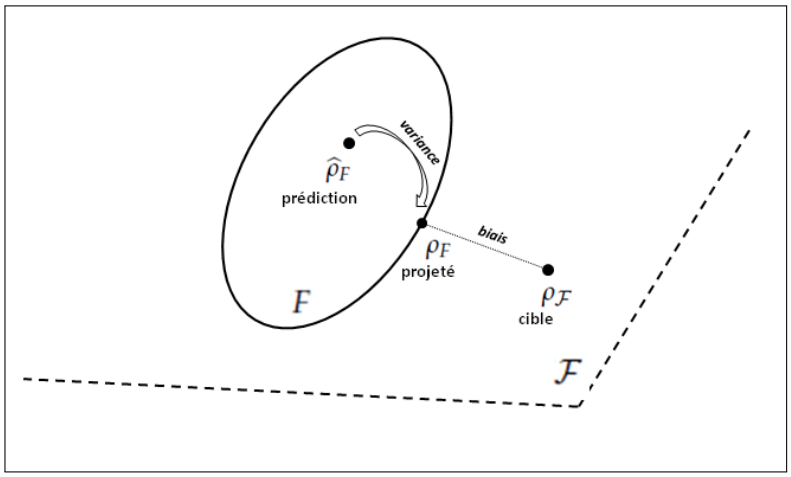


FIGURE 2.2 – Composantes générales de la prédiction

L'enjeu principal de l'apprentissage statistique est de fournir une procédure statistique (i.e $\widehat{\rho}_F$) avec le plus petit excès de risque possible. Par la décomposition (2.21), on pourrait vouloir minimiser deux termes, mais ceux-ci ont des comportements antagonistes.
En effet, il y a tout d'abord le *terme de biais* donné par $\mathcal{R}^F - \mathcal{R}^{\mathcal{F}} \geq 0$ qui mesure la qualité d'approximation du modèle F pour la quantité d'intérêt $\rho_{\mathcal{F}} \in \mathcal{F}$. Par définition de ρ_F (2.16), ce terme décroît avec F pour l'inclusion. Une appréciation du terme de biais peut être obtenue par un jugement d'expert.
Le *terme de variance*, donné par $\mathcal{R}_{\Psi}(\widehat{\rho}_F) - \mathcal{R}^F$, mesure la "complexité" du modèle F relativement au contraste Ψ et à la mesure $Q^{\mathbf{z}}$. Ainsi, un modèle "trop grand" mènera à une variance importante et à un petit biais. Un "petit" modèle aura une variance faible et un grand biais.
Il s'agit ici d'un *compromis biais-variance*. Cette problématique est au coeur des principes de *sélection de modèles* portant sur le choix de F , on pourra consulter [6] pour plus de précisions à ce sujet.

Pour notre propos, avant le "choix" même d'un modèle F, nous nous sommes intéressés à la *nature* d'un tel modèle fourni par un computer experiment (e.g boîte noire...), et aux procédures statistiques résultantes qui sont non standards.
Les notions générales précédemment données seront appliquées aux computer experiments à la Section 2.5. Avant, évoquons la notion de *pénalisation*.

2.4 Pénalisation de contraste

Le contenu de cette section est inspiré des travaux de P. Massart [6] sur la sélection de modèles. Nous présentons ici des notions qui vont nous être utiles pour le cadre paramétrique à la Section 2.5. Nous tenons également à souligner que notre but premier n'est pas la sélection de modèle mais plutôt l'estimation paramétrique, comme nous le verrons dans les sections suivantes. Ainsi, la notion de pénalisation que nous utiliserons par la suite portera sur l'espace des paramètres Θ donné par l'étude de computer experiments. Ce sera l'objet de la Section 2.6. Avant cela, nous donnons quelques définitions et notations dans un cadre plus général.

2.4.1 Motivations et définitions

Dans les sections qui précèdent, nous avons vu qu'une caractéristique $\rho_{\mathcal{F}} \in \mathcal{F}$ (inconnue) telle que

$$\rho_{\mathcal{F}} = \underset{\rho\in\mathcal{F}}{\text{Argmin}}\, \mathcal{R}_\Psi(\rho)\,, \quad \mathcal{R}_\Psi(\rho) = \mathbb{E}_{Q^{\mathbf{z}}}\Psi(\rho, Z)$$

est prédite par

$$\widehat{\rho}_F = \underset{\rho\in F}{\text{Argmin}}\, \widehat{\mathcal{R}}_\Psi(\rho)\,, \tag{2.22}$$

où $F \subset \mathcal{F}$ est un modèle et $\widehat{\mathcal{R}}_\Psi$ un risque empirique associé au risque $\mathcal{R}_\Psi$ (inconnu).
Rappelons que le projeté de $\rho_{\mathcal{F}} \in \mathcal{F}$ sur F est donné par

$$\rho_F = \underset{\rho\in F}{\text{Argmin}}\, \mathcal{R}_\Psi(\rho)\,.$$

Il est alors légitime de comprendre d'où provient l'erreur d'estimation entre la caractéristique que l'on calcule $\widehat{\rho}_{\mathcal{F}}$ et son objectif $\rho_{\mathcal{F}}$. Il est aisé de noter que la seule différence est d'avoir remplacé le risque inconnu $\mathcal{R}_\Psi$ par un risque empirique $\widehat{\mathcal{R}}_\Psi$ disponible.
Ecrivons

$$\mathcal{R}_\Psi(\rho) = \widehat{\mathcal{R}}_\Psi(\rho) + \Big(\mathcal{R}_\Psi(\rho) - \widehat{\mathcal{R}}_\Psi(\rho)\Big)\,,$$

et posons

$$\text{pen}^{id}_\Psi(\rho) := \mathcal{R}_\Psi(\rho) - \widehat{\mathcal{R}}_\Psi(\rho)\,. \tag{2.23}$$

Alors nous avons

$$\rho_F = \underset{\rho\in F}{\text{Argmin}} \Big(\widehat{\mathcal{R}}_\Psi(\rho) + \text{pen}^{id}_\Psi(\rho)\Big)\,.$$

Autrement dit, dans la procédure de minimisation donnant $\widehat{\rho}_F$ (2.22) il "manquerait" le terme $\text{pen}^{id}_\Psi(\rho)$ pour "tomber" sur la cible ρ_F (i.e $\widehat{\rho}_F = \rho_F$).
En fait, la quantité $\text{pen}^{id}_\Psi(\rho)$ ci-dessus est qualifiée de **pénalité idéale**, terme repris de [1], qui représente la quantité manquante à la procédure de minimisation (2.22) pour annihiler l'erreur de substitution due au fait de remplacer $\mathcal{R}_\Psi(\rho)$ par $\widehat{\mathcal{R}}_\Psi(\rho)$.
Malheureusement, la quantité $\text{pen}^{id}_\Psi(\rho)$ est inconnue de l'utilisateur car elle dépend notamment de la mesure $Q^{\mathbf{z}}$. Ainsi, deux stratégies se présentent : soit on ne pénalise pas, soit on cherche à établir une procédure permettant de prendre en compte le "terme manquant" dans la procédure de minimisation.
La première stratégie conduit souvent à un problème de *surapprentissage*. C'est à dire que l'on va *apprendre* une caractéristique $\widehat{\rho}_F$ qui va "coller" excessivement aux données. On donnera

une illustration dans la Section 4.15.
La deuxième stratégie semble plus sage car elle permet une généralisation de l'information apportée par les données, ce qui améliore les performances de prédiction.

Il est à noter que, dans nos propos, la pénalité idéale aura très souvent une espérance nulle

$$\forall\, \rho \in \mathcal{F}\,, \quad \mathbb{E}(\mathrm{pen}_\Psi^{id}(\rho)) = 0\,,$$

ce qui n'est pas le cas en sélection de modèle. Notre objectif étant l'estimation paramétrique où le "vrai" risque a été remplacé par un risque empirique. Nous verrons cela à la Section 2.6.

Définition 2.4.1. Fonction de pénalisation et contraste pénalisé.
Soit un $\mathcal{F}$-contraste Ψ.
On appelle *fonction de pénalisation* sur $\mathcal{F}$ (ou *pénalisation*) une application

$$\begin{array}{rccc} \mathrm{pen}_\Psi : & \mathcal{F} & \longrightarrow & \mathbb{R} \\ & \rho & \longmapsto & \mathrm{pen}_\Psi(\rho)\,. \end{array} \tag{2.24}$$

On appelle *contraste pénalisé* un contraste

$$\Psi_{\mathrm{pen}}(\rho, z) = \Psi(\rho, z) + \mathrm{pen}_\Psi(\rho)\,. \tag{2.25}$$

Bien que la pénalité idéale $\mathrm{pen}_\Psi^{id}(\rho) := \mathcal{R}_\Psi(\rho) - \widehat{\mathcal{R}}_\Psi(\rho)$ soit inconnue, on peut tenter de l'estimer. Les récents travaux de [6] et [1] donne une méthodologie permettant de construire une pénalité "proche" de la pénalité idéale dans le cadre de la sélection de modèle. Nous n'irons pas au niveau de détail présent dans ces travaux, nous nous contenterons d'en extraire les grandes lignes et de les inscrire dans nos développements.

2.4.2 Choix d'une pénalité

Le choix d'une pénalité "convenable" sera guidé par le lemme suivant.

Lemme 2.4.1. *Soit Ψ un $\mathcal{F}$-contrast. Considérons la pénalité idéale pen_Ψ^{id} (2.23) et soit une pénalité $\mathrm{pen}_\Psi : \mathcal{F} \to \mathbb{R}$ telle que*

$$\text{pour tout } \rho \in F \subset \mathcal{F}\,, \quad \mathrm{pen}_\Psi(\rho) \geq \mathrm{pen}_\Psi^{id}(\rho)\,. \tag{2.26}$$

En considérant le contraste pénalisé $\Psi_{\mathrm{pen}}(\rho, z) = \Psi(\rho, z) + \mathrm{pen}_\Psi(\rho)$ et $\widehat{\rho}_F = \mathrm{Argmin}_{\rho \in F}\, \widehat{\mathcal{R}}_{\Psi_{\mathrm{pen}}}(\rho)$, nous obtenons

$$\mathcal{R}_\Psi(\widehat{\rho}_F) \leq \inf_{\rho \in F} \left(\mathcal{R}_\Psi(\rho) + (\mathrm{pen}_\Psi - \mathrm{pen}_\Psi^{id})(\rho) \right)\,. \tag{2.27}$$

Démonstration. Soit $\rho \in F$, on a

$$\begin{array}{rcl} \mathcal{R}_\Psi(\widehat{\rho}_F) & = & \widehat{\mathcal{R}}_\Psi(\widehat{\rho}_F) + \mathrm{pen}_\Psi^{id}(\widehat{\rho}_F) \\ & = & \widehat{\mathcal{R}}_\Psi(\widehat{\rho}_F) + \mathrm{pen}_\Psi(\widehat{\rho}_F) + (\mathrm{pen}_\Psi^{id} - \mathrm{pen}_\Psi)(\widehat{\rho}_F)\,. \end{array}$$

Or, par définition de $\widehat{\rho}_F$, pour tout $\rho \in F$

$$\widehat{\mathcal{R}}_\Psi(\widehat{\rho}_F) + \mathrm{pen}_\Psi(\widehat{\rho}_F) = \widehat{\mathcal{R}}_{\Psi_{\mathrm{pen}}}(\widehat{\rho}_F) \leq \widehat{\mathcal{R}}_{\Psi_{\mathrm{pen}}}(\rho) = \widehat{\mathcal{R}}_\Psi(\rho) + \mathrm{pen}_\Psi(\rho)\,.$$

Puis, en écrivant $\widehat{\mathcal{R}}_\Psi(\rho) = \mathcal{R}_\Psi(\rho) - \mathrm{pen}_\Psi^{id}(\rho)$, il vient

$$\mathcal{R}_\Psi(\widehat{\rho}_F) \quad \leq \quad \mathcal{R}_\Psi(\rho) + (\mathrm{pen}_\Psi - \mathrm{pen}_\Psi^{id})(\rho) + (\mathrm{pen}_\Psi^{id} - \mathrm{pen}_\Psi)(\widehat{\rho}_F)\,.$$

Enfin, si pen_Ψ satisfait (2.26) alors le troisième terme du membre de droite de la dernière inégalité est négatif, on le majore donc par 0.
Ce qui donne, pour tout $\rho \in F$

$$\mathcal{R}_\Psi(\widehat{\rho}_F) \leq \mathcal{R}_\Psi(\rho) + (\mathrm{pen}_\Psi - \mathrm{pen}_\Psi^{id})(\rho)\,,$$

et en prenant l'infimum sur F, on a le résultat voulu. □

L'inégalité (2.27) du lemme précédent nous informe de combien on "s'écarte" du risque idéal $\mathcal{R}^F = \inf_{\rho\in F} \mathcal{R}_\Psi(\rho)$. En effet, on peut écrire (avec une borne supérieure plus pessimiste)

$$\mathcal{R}^F \leq \mathcal{R}_\Psi(\widehat{\rho}_F) \leq \mathcal{R}^F + \sup_{\rho\in F} \Big((\mathrm{pen}_\Psi - \mathrm{pen}_\Psi^{id})(\rho) \Big)\,.$$

Ainsi, tout l'enjeu du choix d'une pénalité est d'en choisir une dont on peut garantir qu'elle reste "uniformément proche" de la pénalité idéale, autrement dit, on voudrait que

$$\mathrm{pen}_\Psi^{id} \leq \mathrm{pen}_\Psi \leq (1+\delta)\,\mathrm{pen}_\Psi^{id} \quad \text{uniformément sur } F\,,$$

avec $\delta > 0$ le plus petit possible.
L'étude d'une telle pénalité se fait en regardant la pénalité idéale pen_Ψ^{id} comme un processus empirique, dont le contrôle peut se faire grâce aux *inégalité de concentrations*, on pourra consulter le chapitre des auteurs S. Boucheron, O. Bousquet et G. Lugosi [2] et les références qui y figurent.

Remarque 2.4.1. Si les fluctuations de la pénalité idéale sont très minimes, par example

$$\mathrm{pen}_\Psi^{id} \approx \text{constante}\,,$$

alors, ne pas pénaliser ne sera pas très préjudiciable quant à la performance de la prédiction.

En somme, nous venons de voir qu'une "bonne" pénalité devrait avoir un comportement proche de la pénalité idéale. Or, cette dernière pénalité dépend également des données, c'est pourquoi en pratique on considérera des pénalités de la forme

$$\mathrm{pen}_\Psi(\rho) = K\,\mathrm{pen}_{shape}(\rho) \tag{2.28}$$

où K est communément appelée **constante de calibration** (aux données) et pen_{shape} représente **la forme** de la pénalité.

2.4.3 Entre "pénalisation" et "régularisation"

La notion de pénalisation, comme nous l'avons définie, a donc pour but d'améliorer la qualité de la prédiction dont il est question, en cherchant à compenser au maximum l'erreur due à la substitution de quantités déterministes (inconnues) par des quantités empiriques basées sur des données (connues). Dans ce cas, les considérations sont purement statistiques dans le sens où on ne se pose pas le problème de résolution *algorithmique* du projeté $\widehat{\rho}_F$.
La notion de *régularisation*, quant à elle, intervient dans les problèmes mal-posés pour justement les rendre bien-posés. Il s'agit de régulariser le problème inverse dont il est question.

La méthode la plus utilisée est la méthode de *Tikhonov* [7].
Ces deux notions sont fondamentalement liées et un comportement naturel du praticien sera de penser à la régularisation lorsqu'il pénalise, ou au contraire, penser à la pénalisation (i.e pénalité idéale) lorsqu'il s'interesse au problème d'optimisation.

Nous utiliserons l'exemple de la *ridge regression* à la Section 2.7 où l'on verra également cette méthode sous l'angle de la pénalité idéale, autre manière de justifier l'usage d'une pénalité L_2 .

2.5 Apprentissage avec un Computer Experiment

Après les généralités précédemment établies, nous présentons maintenant le cadre particulier dans lequel nous nous plaçons.

2.5.1 Quelques rappels

Soit un échantillon $Z_1 = (\mathbf{X}_1, Y_1), ..., Z_n = (\mathbf{X}_n, Y_n)$ i.i.d de loi jointe $Q^{\mathbf{z}}$. On note $P^{\mathbf{x}}$ la loi de $\mathbf{X}$ et Q la loi de Y de densité f par rapport à la mesure de Lebesgue. (Les données $\mathbf{X}_1, ..., \mathbf{X}_n$ peuvent ne pas être observées).
On rappelle l'écriture d'une fonction boîte noire donnée en introduction :

$$\begin{array}{rcl} h \,:\, (\mathcal{X}, \mathcal{B}, P^{\mathbf{x}}) \times \Theta & \longmapsto & \mathcal{Y} \\ (\mathbf{X}, \boldsymbol{\theta}) & \longrightarrow & h(\mathbf{X}, \boldsymbol{\theta}) \end{array}$$

où $\mathcal{X} \subset \mathbb{R}^d$ est l'*espace des entrées* et $\Theta \subset \mathbb{R}^k$ l'*espace des paramètres*.
On notera $p^{\mathbf{x}}$ la densité de probabilité associée à $P^{\mathbf{x}}$.
Par ailleurs, rappelons que nous supposons l'écriture

$$Y = h^*(\mathbf{X}) + \eta^* \,,$$

pour une certaine fonction h^* et une certaine variable aléatoire η^* inconnues.
Considérons les variables aléatoires

$$Y_{\boldsymbol{\theta},\eta} = h(\mathbf{X}, \boldsymbol{\theta}) + \eta \,, \quad \boldsymbol{\theta} \in \Theta \,, \tag{2.29}$$

où η est une variable aléatoire de densité g_η qu'on supposera centrée. On retiendra que $Y_{\boldsymbol{\theta},\eta}$ dépend de h.

En toute rigueur, il conviendrait de définir un espace probabilisé (assez vaste) $(\Omega, \mathcal{A}, \mathbb{P})$ équipé des variables aléatoires $\mathbf{X}$, Y et η .

La variable aléatoire η représente l'erreur de modélisation et on pourra émettre certaines hypothèses (en plus de $\mathbb{E}(\eta) = 0$) pour mener une analyse statistique (par exemple, on pourrait en plus supposer que $\mathbb{E}(\eta/\mathbf{X}) = 0$ ...). Les variables $Y_{\boldsymbol{\theta},\eta}$, $\boldsymbol{\theta} \in \Theta$ constituent alors un *modèle stochastique* du phénomène qui nous intéresse, Y, et même du couple $(\mathbf{X}, Y)$.

Notons le couple

$$Z_{\boldsymbol{\theta},\eta} := (\mathbf{X}, Y_{\boldsymbol{\theta},\eta}) \,, \tag{2.30}$$

et $Q^{\mathbf{z}_{\boldsymbol{\theta},\eta}}$ sa mesure de probabilité associée.
On vérifie immédiatement que

$$Q^{\mathbf{z}_{\boldsymbol{\theta},\eta}}(d\mathbf{x}, dy) = p^{\mathbf{x}}(\mathbf{x})\, g_{\eta/\mathbf{x}=\mathbf{x}}(y - h(\mathbf{x}, \boldsymbol{\theta}))\, d\mathbf{x}\, dy\,, \tag{2.31}$$

où $g_{\eta/\mathbf{x}=\mathbf{x}}$ est la densité de η conditionnellement à $\mathbf{X} = \mathbf{x}$.

En reprenant les notations de la section précédente, le but est de prédire une caractéristique (ou quantité d'intérêt) $\rho_{\mathcal{F}} \in \mathcal{F}$ de la mesure $Q^{\mathbf{z}}$.

2.5.2 Modèle F donné par computer experiments

D'après la remarque (2.3.4), la caractérisation par contraste d'une quantité d'intérêt nous donne l'existence d'un contraste $\Psi : \mathcal{F} \to L_1(Q^{\mathbf{z}})$ tel que

$$\rho_{\mathcal{F}} = \underset{\rho \in \mathcal{F}}{\operatorname{Argmin}}\, \mathcal{R}_{\Psi}(\rho)\,,$$

où

$$\mathcal{R}_{\Psi}(\rho) = \mathbb{E}_{Q^{\mathbf{z}}} \Psi(\rho, Z)\,.$$

Ensuite, la définition "directe" d'une quantité d'intérêt comme étant une application de l'espace des mesures de probabilité sur $\mathcal{X} \times \mathcal{Y}$ dans $\mathcal{F}$ (2.3.1), donne un modèle $F \subset \mathcal{F}$ qui s'écrit

$$F = \{\rho_{\mathcal{F}}(\mu)\,,\, \mu \in \pi_{\mathcal{Z}}\} \subset \mathcal{F}\,,$$

pour un certain ensemble $\pi_{\mathcal{Z}}$ de mesures sur $\mathcal{Z}$, .
Il semble naturel de prendre comme ensemble de mesures

$$\pi_{\mathcal{Z}} = \{Q^{\mathbf{z}_{\boldsymbol{\theta},\eta}}\,,\, \boldsymbol{\theta} \in \Theta\}\,,$$

où $Q^{\mathbf{z}_{\boldsymbol{\theta},\eta}}$, donné en (2.31), est la mesure associée au modèle stochastique (2.29), $Y_{\boldsymbol{\theta},\eta} = h(\mathbf{X}, \boldsymbol{\theta}) + \eta$.
Ainsi, le modèle F s'écrit

$$F = \{\rho \in \rho_{\mathcal{F}}(Q^{\mathbf{z}_{\boldsymbol{\theta},\eta}})\,,\, \boldsymbol{\theta} \in \Theta\}\,.$$

Par souci de simplification, nous noterons désormais

$$\rho_{\mathcal{F}}(\boldsymbol{\theta}) := \rho_{\mathcal{F}}(Q^{\mathbf{z}_{\boldsymbol{\theta},\eta}})\,,$$

et nous garderons à l'esprit que $\rho_{\mathcal{F}}(\boldsymbol{\theta})$ dépend de la variable aléatoire η (et de h aussi). Nous le rappellerons lorsque ce sera nécessaire.

Enfin, pour tout ce qui va suivre, un modèle F s'écrira toujours de manière générale

$$F = \{\rho_{\mathcal{F}}(\boldsymbol{\theta})\,,\, \boldsymbol{\theta} \in \Theta\}\,. \tag{2.32}$$

Nous donnons quelques exemples détaillés à la Section 2.5.4.

En outre, nous avons vu que le projeté ρ_F de $\rho_{\mathcal{F}}$ sur F est donné par

$$\rho_F = \underset{\rho \in F}{\operatorname{Argmin}}\, \mathcal{R}_{\Psi}(\rho)\,. \tag{2.33}$$

Or, le modèle F étant donné par (2.32), le projeté s'écrit également

$$\rho_F = \rho_{\mathcal{F}} \left(\underset{\boldsymbol{\theta}\in\Theta}{\text{Argmin}}\, \mathcal{R}_\Psi(\rho_{\mathcal{F}}(\boldsymbol{\theta})) \right) .$$

Par abus de notation, nous noterons

$$\mathcal{R}_\Psi(\boldsymbol{\theta}) := \mathcal{R}_\Psi(\rho_{\mathcal{F}}(\boldsymbol{\theta})) .$$

Ainsi,

$$\rho_F = \rho_{\mathcal{F}} \left(\underset{\boldsymbol{\theta}\in\Theta}{\text{Argmin}}\, \mathcal{R}_\Psi(\boldsymbol{\theta}) \right) . \tag{2.34}$$

Afin de ne pas rajouter de complexité à notre étude, nous formulons l'hypothèse suivante.

Hypothèse 2.5.1. Pour tout contraste Ψ, on supposera que le projeté ρ_F est **unique**.

2.5.3 Ψ-minimiseur et Ψ-estimateur

On appelle **Ψ-minimiseur** un paramètre satisfaisant

$$\boldsymbol{\theta}_\Psi = \underset{\boldsymbol{\theta}\in\Theta}{\text{Argmin}}\, \mathcal{R}_\Psi(\boldsymbol{\theta}) . \tag{2.35}$$

Remarque 2.5.1. Le paramètre $\boldsymbol{\theta}_\Psi$ n'est pas nécessairement unique, i.e le modèle $F = \{\rho_{\mathcal{F}}(\boldsymbol{\theta}) , \boldsymbol{\theta} \in \Theta\}$ n'est pas nécessairement *identifiable*. On conservera tout de même la notation $\boldsymbol{\theta}_\Psi$ même s'il s'agit d'un ensemble. Le Ψ-minimiseur $\boldsymbol{\theta}_\Psi$ est en fait incalculable, on construit alors une procédure d'estimation induite.

Le projeté ρ_F de $\rho_{\mathcal{F}} \in \mathcal{F}$ sur $F \subset \mathcal{F}$ s'écrit

$$\rho_F = \rho_{\mathcal{F}} (\boldsymbol{\theta}_\Psi) .$$

Nous sommes ainsi amenés à une étude paramétrique où l'essentiel du travail réside dans l'estimation du paramètre $\boldsymbol{\theta}_\Psi$.

Soit $\widehat{\mathcal{R}}_\Psi(\boldsymbol{\theta})$ un risque empirique associé à $\mathcal{R}_\Psi(\boldsymbol{\theta})$ (qui dépend de $Q^{\mathbf{z}}$ donc inconnu).

On appelle **Ψ-estimateur** (d'un Ψ-minimiseur) un estimateur qui satisfait

$$\widehat{\boldsymbol{\theta}}_\Psi = \underset{\boldsymbol{\theta}\in\Theta}{\text{Argmin}}\, \widehat{\mathcal{R}}_\Psi(\boldsymbol{\theta}) . \tag{2.36}$$

Finalement, la prédiction de $\rho_{\mathcal{F}}$ est donnée par

$$\widehat{\rho_{\mathcal{F}}} = \rho_{\mathcal{F}}(\widehat{\boldsymbol{\theta}}_\Psi) \in F . \tag{2.37}$$

Le risque empirique $\widehat{\mathcal{R}}_\Psi(\boldsymbol{\theta})$ est en fait la version empirique (i.e connue, calculable) du risque $\mathcal{R}_\Psi(\boldsymbol{\theta}) = \mathbb{E}_{Q^{\mathbf{z}}} \Psi (\rho , Z)$.
En effet, tout d'abord la mesure $Q^{\mathbf{z}}$ est supposée inconnue, nous ne disposons que d'un échantillon $Z_1, ..., Z_n$. Le risque empirique prend alors la forme

$$\widehat{\mathcal{R}}_\Psi(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \Psi (\rho_{\mathcal{F}}(\boldsymbol{\theta}), Z_i) .$$

Ensuite, il se peut que la caractéristique $\rho_{\mathcal{F}}(\boldsymbol{\theta})$ ne soit pas analytiquement calculable, c'est le cas pour les computer experiments (cf. exemple de la densité ci-après). Le risque empirique s'écrirait alors

$$\widehat{\mathcal{R}}_{\Psi}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \Psi\left(\rho_{\mathcal{F}}^{m}(\boldsymbol{\theta}), Z_i\right) ,$$

où $\rho_{\mathcal{F}}^{m}(\boldsymbol{\theta})$ est une approximation de $\rho_{\mathcal{F}}(\boldsymbol{\theta})$ basée sur un m-échantillon

$$\mathcal{X}^m = (\mathbf{X}'_1, ..., \mathbf{X}'_m)$$

i.i.d de loi $P^{\mathbf{x}}$, et que l'on supposera dans toute la suite indépendant des $(Z_1 = (\mathbf{X}_1, Y_1), ..., Z_n = (\mathbf{X}_n, Y_n))$. En fait, l'approximation est plutôt basée sur un m-échantillon

$$((h(\mathbf{X}'_j, \boldsymbol{\theta}) + \eta'_j))_{1\leq j\leq m} .$$

Nous verrons cela en détail dans le Chapitre 3.
(Par souci de simplification, quand il n'y aura pas d'ambiguïté on pourra omettre les données η'_j , i.e on considérera simplement $((h(\mathbf{X}'_j, \boldsymbol{\theta})))_{1\leq j\leq m}$.)

Dans la suite de nos développements, un Ψ-estimateur sera soit de la forme

$$\widehat{\boldsymbol{\theta}}_{\Psi} = \underset{\boldsymbol{\theta}\in\Theta}{\text{Argmin}} \frac{1}{n}\sum_{i=1}^{n} \Psi\left(\rho_{\mathcal{F}}(\boldsymbol{\theta}), Z_i\right) \tag{2.38}$$

si $\rho_{\mathcal{F}}(\boldsymbol{\theta})$ est calculable (ex. espérance conditionnelle, fonction h triviale etc...), ou bien

$$\widehat{\boldsymbol{\theta}}_{\Psi} = \underset{\boldsymbol{\theta}\in\Theta}{\text{Argmin}} \frac{1}{n}\sum_{i=1}^{n} \Psi\left(\rho_{\mathcal{F}}^{m}(\boldsymbol{\theta}), Z_i\right) . \tag{2.39}$$

2.5.4 Exemples

Voici deux petits exemples qui illustrent nos propos. Le premier est classique et a pour but d'asseoir nos notations. Le second sera une des principales motivations de nos travaux.

Prédiction de l'espérance conditionnelle

Cet exemple classique consiste en la prédiction de

$$\rho_{\mathcal{F}} : \mathbf{x} \mapsto \mathbb{E}(Y/\mathbf{X} = \mathbf{x}) \in \mathcal{F} = L_2(P^{\mathbf{x}}) ,$$

avec le contraste

$$\Psi(\rho, (\mathbf{x}, y)) = (y - \rho(\mathbf{x}))^2 .$$

Considérons l'ensemble de mesures $\{Q^{\mathbf{z}_{\boldsymbol{\theta},\eta}} , \boldsymbol{\theta} \in \Theta\}$ (2.31) où

$$Q^{\mathbf{z}_{\boldsymbol{\theta},\eta}}(d\mathbf{x}, dy) = p^{\mathbf{x}}(\mathbf{x})\, g_{\eta/\mathbf{X}=\mathbf{x}}(y - h(\mathbf{x}, \boldsymbol{\theta}))\, d\mathbf{x}\, dy .$$

Dans ce cas, l'"application" quantité d'intérêt est donnée par

$$\mu \longmapsto \rho_{\mathcal{F}}(\mu) = \int_{\mathcal{Z}} w_{\mathcal{F}}(x, y)\, \mu(dx\, dy)$$

avec

$$w_{\mathcal{F}}(\mathbf{x}, y)(\mathbf{u}) = \frac{y}{p^{\mathbf{x}}(\mathbf{x})}\delta_{\mathbf{u}}(\mathbf{x}) .$$

Le modèle F correspondant est

$$F = \{\rho_{\mathcal{F}}(\boldsymbol{\theta})\,,\ \boldsymbol{\theta} \in \Theta\}$$

avec pour tout $\mathbf{u} \in \mathcal{X}$

$$\begin{aligned}
\rho_{\mathcal{F}}(\boldsymbol{\theta})(\mathbf{u}) &= \rho_{\mathcal{F}}(Q^{\mathbf{z}_{\boldsymbol{\theta},\eta}})(\mathbf{u}) \\
&= \int_{\mathcal{Z}} w_{\mathcal{F}}(\mathbf{x},y)(\mathbf{u})\, Q^{\mathbf{z}_{\boldsymbol{\theta},\eta}}(d\mathbf{x},dy) \\
&= \int_{\mathcal{Z}} \frac{y}{p^{\mathbf{x}}(\mathbf{x})}\delta_{\mathbf{u}}(\mathbf{x})\, p^{\mathbf{x}}(\mathbf{x})\, g_{\eta/\mathbf{X}=\mathbf{x}}(y - h(\mathbf{x},\boldsymbol{\theta}))\, d\mathbf{x}\, dy \\
&= \int_{\mathcal{Y}} \frac{y}{p^{\mathbf{x}}(\mathbf{u})}\, p^{\mathbf{x}}(\mathbf{u})\, g_{\eta/\mathbf{X}=\mathbf{u}}(y - h(\mathbf{u},\boldsymbol{\theta}))\, dy \\
&= \int_{\mathcal{Y}} y\, g_{\eta/\mathbf{X}=\mathbf{u}}(y - h(\mathbf{u},\boldsymbol{\theta}))\, dy\,.
\end{aligned}$$

En posant le changement de variable $y' = y - h(\mathbf{u},\boldsymbol{\theta})$, il vient

$$\begin{aligned}
\rho_{\mathcal{F}}(\boldsymbol{\theta})(\mathbf{u}) &= \int_{\mathcal{Y}} (y + h(\mathbf{u},\boldsymbol{\theta}))\, g_{\eta/\mathbf{X}=\mathbf{u}}(y)\, dy \\
&= h(\mathbf{u},\boldsymbol{\theta}) + \mathbb{E}(\eta/\mathbf{X} = \mathbf{u})\,.
\end{aligned}$$

Bien entendu, ceci se retrouve directement en remplaçant Y dans $\mathbf{x} \mapsto \mathbb{E}(Y/\mathbf{X} = \mathbf{x})$ par sa modélisation $Y_{\boldsymbol{\theta},\eta}$, ce qui donne immédiatement

$$\mathbb{E}(Y_{\boldsymbol{\theta},\eta}/\mathbf{X} = \mathbf{x}) = \mathbb{E}(h(\mathbf{X},\boldsymbol{\theta}) + \eta/\mathbf{X} = \mathbf{x}) = h(\mathbf{x},\boldsymbol{\theta}) + \mathbb{E}(\eta/\mathbf{X} = \mathbf{x})\,.$$

Maintenant, supposons que η est indépendant de $\mathbf{X}$, alors la quantité $\rho_{\mathcal{F}}(\boldsymbol{\theta})$ est donnée simplement par

$$\rho_{\mathcal{F}}(\boldsymbol{\theta})\ :\ \mathbf{u} \mapsto h(\mathbf{u},\boldsymbol{\theta}) \quad (\text{car}\,\mathbb{E}(\eta) = 0)\,.$$

Il en résulte la procédure d'estimation suivante

$$\begin{aligned}
\boldsymbol{\theta}_{\Psi} &= \underset{\boldsymbol{\theta}\in\Theta}{\text{Argmin}}\, \mathbb{E}_{Q^{\mathbf{z}}} \Psi(\rho_{\mathcal{F}}(\boldsymbol{\theta}), Z = (\mathbf{X}, Y)) \\
&= \underset{\boldsymbol{\theta}\in\Theta}{\text{Argmin}}\, \mathbb{E}_{Q^{\mathbf{z}}} (Y - \rho_{\mathcal{F}}(\boldsymbol{\theta})(\mathbf{X}))^2 \\
&= \underset{\boldsymbol{\theta}\in\Theta}{\text{Argmin}}\, \mathbb{E}_{Q^{\mathbf{z}}} (Y - h(\mathbf{X},\boldsymbol{\theta}))^2\,,
\end{aligned}$$

qui n'est autre qu'une procédure de regression L_2.
D'où l'estimateur des moindres carrés

$$\widehat{\boldsymbol{\theta}}_{\Psi} = \underset{\boldsymbol{\theta}\in\Theta}{\text{Argmin}}\, \frac{1}{n}\sum_{i=1}^{n} (Y_i - h(\mathbf{X}_i,\boldsymbol{\theta}))^2\,.$$

Il vient que $\widehat{\rho_{\mathcal{F}}}\ :\ \mathbf{x} \mapsto h(\mathbf{x},\widehat{\boldsymbol{\theta}}_{\Psi})$ est la prédiction de $\rho_{\mathcal{F}}\ :\mathbf{x} \mapsto \mathbb{E}(Y/\mathbf{X} = \mathbf{x})$.

Maintenant, supposons que nous voulons prédire une autre caractéristique que l'espérance conditionnelle.

Prédiction de la densité f de Y

Nous voulons prédire

$$\rho_{\mathcal{F}} = f \in \mathcal{F} = \{\text{densités sur } \mathcal{Y}\}$$

avec le contraste

$$\Psi(\rho, y) = -\log(\rho)(y)\,.$$

Dans ce cas, l'espace F est constitué des fonctions de $\mathcal{Y}$, $v \mapsto \rho_{\mathcal{F}}(\boldsymbol{\theta})(v)$, $\boldsymbol{\theta} \in \Theta$ données comme suit. Soit $v \in \mathcal{Y}$ et considérons

$$w_{\mathcal{F}}(\mathbf{x}, y)(v) = \delta_v(y)\,.$$

On a

$$\begin{aligned}
\rho_{\mathcal{F}}(\boldsymbol{\theta})(v) &= \rho_{\mathcal{F}}(Q^{\mathbf{z}_{\boldsymbol{\theta},\eta}})(v) \\
&= \int_{\mathcal{Z}} w_{\mathcal{F}}(\mathbf{x}, y)(v)\, Q^{\mathbf{z}_{\boldsymbol{\theta},\eta}}(d\mathbf{x}, dy) \\
&= \int_{\mathcal{Z}} \delta_v(y)\, p^{\mathbf{x}}(\mathbf{x})\, g_{\eta/\mathbf{X}=\mathbf{x}}(y - h(\mathbf{x}, \boldsymbol{\theta}))\, d\mathbf{x}\, dy \\
&= \int_{\mathcal{X}} p^{\mathbf{x}}(\mathbf{x})\, g_{\eta/\mathbf{X}=\mathbf{u}}(v - h(\mathbf{x}, \boldsymbol{\theta}))\, d\mathbf{x}\,.
\end{aligned}$$

Par exemple, supposons que la variable η est indépendante de $\mathbf{X}$ et que $g_{\eta/\mathbf{X}=\mathbf{u}} = g_\eta$ est la densité d'une gaussienne centrée de variance σ_η^2, alors

$$\rho_{\mathcal{F}}(\boldsymbol{\theta})(v) = \int_{\mathcal{X}} p^{\mathbf{x}}(\mathbf{x})\, K_{\sigma_\eta}(v - h(\mathbf{x}, \boldsymbol{\theta}))\, d\mathbf{x}\,,$$

avec K_{σ_η} le noyau gaussien

$$K_{\sigma_\eta}(y) = \frac{1}{\sqrt{2\,\pi}\,\sigma_\eta}\, e^{-y^2/2\,\sigma_\eta^2}\,.$$

En fait, $\rho_{\mathcal{F}}(\boldsymbol{\theta})$ est une approximation par convolution gaussienne de la *densité image* de la densité $p^{\mathbf{x}}$ par l'application $\mathbf{x} \mapsto h(\mathbf{x}, \boldsymbol{\theta})$, ce qui a bien un sens puisque le but est de prédire la densité $\rho_{\mathcal{F}} = f$ de la variable Y.

La procédure d'estimation est donnée par

$$\boldsymbol{\theta}_\Psi = \operatorname*{Argmin}_{\boldsymbol{\theta}\in\Theta} \mathbb{E}_{Q^{\mathbf{z}}} \Psi(\rho_{\mathcal{F}}(\boldsymbol{\theta}), Y) \tag{2.40}$$

$$= \operatorname*{Argmin}_{\boldsymbol{\theta}\in\Theta} -\mathbb{E}_Q \log(\rho_{\mathcal{F}}(\boldsymbol{\theta}))(Y)\,. \tag{2.41}$$

Contrairement au cas de l'espérance conditionnelle, la caractéristique $\rho_{\mathcal{F}}(\boldsymbol{\theta})$ n'est pas connue analytiquement, sauf pour des cas triviaux. En effet, le modèle numérique $h(\cdot, \cdot)$ peut être très compliqué et/ou être connu simplement sur des jeux d'entrées/sorties.
Il est alors naturel d'approcher

$$\rho_{\mathcal{F}}(\boldsymbol{\theta})(v) = \int_{\mathcal{X}} p^{\mathbf{x}}(\mathbf{x})\, K_{\sigma_\eta}(v - h(\mathbf{x}, \boldsymbol{\theta}))\, d\mathbf{x}$$

par l'estimateur à noyau suivant

$$\rho_{\mathcal{F}}^m(\boldsymbol{\theta})(v) = \frac{1}{m} \sum_{j=1}^{m} K_{\sigma_\eta}(v - h(\mathbf{X}'_j, \boldsymbol{\theta}))\,, \quad \mathbf{X}'_1, ..., \mathbf{X}'_m \quad i.i.d \sim p^{\mathbf{x}}\,.$$

On obtient alors un Ψ-estimateur de la forme donnée en (2.39)

$$\widehat{\boldsymbol{\theta}}_\Psi = \underset{\boldsymbol{\theta}\in\Theta}{\text{Argmin}} -\frac{1}{n}\sum_{i=1}^{n} \log(\rho_{\mathcal{F}}^m(\boldsymbol{\theta}))(Y_i)\,. \tag{2.42}$$

Il vient que $\widehat{\rho_{\mathcal{F}}} : y \mapsto \rho_{\mathcal{F}}^m(\widehat{\boldsymbol{\theta}}_\Psi)(y)$ est la prédiction de $\rho_{\mathcal{F}} = f$.
Ce type de procédure sera étudié de manière générale dans le Chapitre 3.

Désormais, nous ne détaillerons plus nécessairement la quantité $\rho_{\mathcal{F}}(\boldsymbol{\theta})$ comme nous venons de le faire. On donnera directement la quantité d'intérêt dont il s'agit comme dans les exemples suivants.

Example 2.5.1. Exemple de caractéristiques

- $\rho_{\mathcal{F}}(\boldsymbol{\theta})(\cdot) = h(\cdot, \boldsymbol{\theta})$ ("caractéristique = espérance conditionnelle $\mathbf{x} \mapsto h(\mathbf{x}, \boldsymbol{\theta})$")
- $\rho_{\mathcal{F}}(\boldsymbol{\theta}) = \mathbb{E}_{P^X}(h(\mathbf{X}, \boldsymbol{\theta}))$ ("caractéristique = moyenne")
- $\rho_{\mathcal{F}}(\boldsymbol{\theta}) = \mathbb{P}(h(\mathbf{X}, \boldsymbol{\theta}) > y_0)$ ("caractéristique = probabilité de dépassement")
- $\rho_{\mathcal{F}}(\boldsymbol{\theta}) = pdf\ of\ h(\mathbf{X}, \boldsymbol{\theta})$ ("caractéristique = densité")
- etc...

2.5.5 Un Ψ-estimateur dépend des données disponibles

Un Ψ-minimiseur $\boldsymbol{\theta}_\Psi$

$$\boldsymbol{\theta}_\Psi = \underset{\boldsymbol{\theta}\in\Theta}{\text{Argmin}}\, \mathcal{R}_\Psi(\boldsymbol{\theta}) = \underset{\boldsymbol{\theta}\in\Theta}{\text{Argmin}}\, \mathbb{E}_{Q^Z}\Psi(\rho_{\mathcal{F}}(\boldsymbol{\theta}), Z)$$

est une grandeur abstraite dans le sens où la procédure qui lui est sous-jacente va permettre de construire un Ψ-estimateur $\widehat{\boldsymbol{\theta}}_\Psi$. Or, le passage de l'un à l'autre nécessite des données **relatives au contraste** Ψ considéré.
En effet, si on dispose des données jointes

$$Z_1 = (\mathbf{X}_1, Y_1), ..., Z_n = (\mathbf{X}_n, Y_n)\,,$$

alors on pourra construire un quelconque estimateur $\widehat{\boldsymbol{\theta}}_\Psi$ à partir d'un contraste dépendant de la donnée $(\mathbf{x}, y)$: du type $\Psi(\rho, (\mathbf{x}, y))$, où bien, à partir d'un contraste dépendant uniquement de la donnée y : du type $\Psi(\rho, y)$.
Cependant, si on ne dispose que des données

$$Y_1, ..., Y_n\,,$$

par exemple, parce que les données $\mathbf{X}_1, ..., \mathbf{X}_n$ ne sont pas observées, alors on ne pourra pas calculer un estimateur $\widehat{\boldsymbol{\theta}}_\Psi$ dont le contraste Ψ dépend de la données $(\mathbf{x}, y)$. En d'autres termes, on ne pourra par exemple pas calculer l'estimateur des moindres carrés

$$\widehat{\boldsymbol{\theta}}_\Psi = \underset{\boldsymbol{\theta}\in\Theta}{\text{Argmin}}\, \frac{1}{n}\sum_{i=1}^{n}(Y_i - h(\mathbf{X}_i, \boldsymbol{\theta}))^2\,.$$

Seul les estimateurs basés sur un contraste dépendant de la donnée y pourront être calculés. Nous donnons au Chapitre 6 une application d'une telle procédure.

2.5.6 Ψ-excès de risque dans le cas paramétrique

Une définition générale de l'excès de risque a été donné à la Définition 2.3.2. Nous allons donner une définition plus spécifique à notre cadre d'étude.

Soit $\Psi : \mathcal{F} \to L_1(Q^{\mathbf{z}})$ un $\mathcal{F}$-contraste et

$$\mathcal{R}_\Psi(\boldsymbol{\theta}) = \mathbb{E}_{Q^{\mathbf{z}}}\left(\Psi(\rho_{\mathcal{F}}(\boldsymbol{\theta}), Z)\right)$$

son risque associé sur $F = \{\rho_{\mathcal{F}}(\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\} \subset \mathcal{F}$.
On rappelle que le risque absolu est donné par

$$\mathcal{R}^{\mathcal{F}} = \underset{\rho \in \mathcal{F}}{\operatorname{Argmin}}\, \mathcal{R}_\Psi(\rho)\,, \quad \mathcal{R}_\Psi(\rho) = \mathbb{E}_{Q^{\mathbf{z}}}\left(\Psi(\rho, Z)\right)\,. \tag{2.43}$$

Dans ce contexte paramétrique, le risque idéal s'écrit

$$\mathcal{R}^{F} = \inf_{\boldsymbol{\theta} \in \Theta} \mathcal{R}_\Psi(\boldsymbol{\theta})\,. \tag{2.44}$$

Comme le modèle F est décrit par l'ensemble Θ, nous préférerons l'écriture

$$\mathcal{R}^{F} = \mathcal{R}^{\Theta}_{\Psi}\,,$$

où Ψ est le $\mathcal{F}$-contrast considéré ($F \subset \mathcal{F}$).

Définition 2.5.1. Ψ-excès de risque paramétrique.
Soit $\Psi : \mathcal{F} \to L_1(Q^{\mathbf{z}})$ un $\mathcal{F}$-contraste. Le Ψ-excès de risque (ou excès de risque s'il n'y a pas d'ambiguïté) d'un élément $\boldsymbol{\theta} \in \Theta$ est donné par

$$\mathcal{E}_\Psi(\boldsymbol{\theta}) := \mathcal{R}_\Psi(\boldsymbol{\theta}) - \mathcal{R}^{\mathcal{F}}\,. \tag{2.45}$$

Nous retrouvons la décomposition en un terme de bias et un terme de variance en écrivant

$$\mathcal{E}_\Psi(\boldsymbol{\theta}) = \underbrace{\mathcal{R}_\Psi(\boldsymbol{\theta}) - \mathcal{R}^{\Theta}_{\Psi}}_{\text{terme de variance}} + \underbrace{\mathcal{R}^{\Theta}_{\Psi} - \mathcal{R}^{\mathcal{F}}}_{\text{terme de biais}}\,.$$

Dans la suite, nous nous focaliserons davantage sur le terme de variance que l'on pourra qualifier de Ψ-excès de risque sur Θ (ou sur F) et que l'on notera

$$\mathcal{E}^{\Theta}_{\Psi}(\boldsymbol{\theta}) := \mathcal{R}_\Psi(\boldsymbol{\theta}) - \mathcal{R}^{\Theta}_{\Psi}\,. \tag{2.46}$$

Remarque 2.5.2. Performance d'unΨ-estimateur.
La performance (statistique) d'un Ψ-estimateur $\widehat{\boldsymbol{\theta}}_\Psi$ est donnée par l'étude de son Ψ-excès de risque sur Θ

$$\mathcal{E}^{\Theta}_{\Psi}(\widehat{\boldsymbol{\theta}}_\Psi)\,,$$

qui est d'autant plus proche de zéro que $\widehat{\boldsymbol{\theta}}_\Psi$ est performant. Tout l'enjeu est de déterminer un "bon" majorant (cf. Chapitre 3).
Toutefois, il convient de ne pas négliger le terme de biais induit par les hypothèses de modélisation.

2.6 Quelques mots sur la pénalisation de contraste dans le cas paramétrique

Revenons maintenant sur la notion de pénalisation vue à la Section 2.4 et spécifions cette notion dans le cadre paramétrique que nous venons d'établir.

Le but est de calculer un Ψ-estimateur

$$\widehat{\boldsymbol{\theta}}_\Psi = \underset{\boldsymbol{\theta}\in\Theta}{\text{Argmin}}\, \widehat{\mathcal{R}}_\Psi(\boldsymbol{\theta})\,, \tag{2.47}$$

où le risque empirique $\widehat{\mathcal{R}}_\Psi$ peut prendre les deux formes suivantes

$$\widehat{\mathcal{R}}_\Psi(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \Psi\left(\rho_{\mathcal{F}}(\boldsymbol{\theta}), Z_i\right) \quad \text{où} \quad \widehat{\mathcal{R}}_\Psi(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \Psi\left(\rho^m_{\mathcal{F}}(\boldsymbol{\theta}), Z_i\right)\,. \tag{2.48}$$

Rappelons que $\rho^m_{\mathcal{F}}(\boldsymbol{\theta})$ est une approximation de $\rho_{\mathcal{F}}(\boldsymbol{\theta})$ basée sur un m-échantillon $\mathcal{X}^m = (\mathbf{X}'_1, ..., \mathbf{X}'_m)$ i.i.d de loi $P^{\mathbf{x}}$. Nous allons tenter de voir comment améliorer, dans la mesure du possible, la performance du Ψ-estimateur $\widehat{\boldsymbol{\theta}}_\Psi$ (2.47) avec la même base de données.

En suivant le même raisonnement que dans la Section 2.4, la **pénalité idéale** est donnée par

$$\text{pen}^{id}_\Psi(\boldsymbol{\theta}) = \left(\mathcal{R}_\Psi - \widehat{\mathcal{R}}_\Psi\right)(\boldsymbol{\theta})\,. \tag{2.49}$$

(Rappelons que la notion de "pénalité idéale" est différente de celle employée en sélection de modèle.)
Dans notre contexte, cette pénalité (inconnue) vaut, ou bien

$$\text{pen}^{id}_\Psi(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \left(\mathbb{E}_{Q^{\mathbf{z}}}\Psi\left(\rho_{\mathcal{F}}(\boldsymbol{\theta}), Z\right) - \Psi\left(\rho_{\mathcal{F}}(\boldsymbol{\theta}), Z_i\right)\right) \tag{2.50}$$

ou bien

$$\text{pen}^{id}_\Psi(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \left(\mathbb{E}_{Q^{\mathbf{z}}}\Psi\left(\rho_{\mathcal{F}}(\boldsymbol{\theta}), Z\right) - \Psi\left(\rho^m_{\mathcal{F}}(\boldsymbol{\theta}), Z_i\right)\right)\,. \tag{2.51}$$

Quitte à rajouter une constante positive indépendante de $\boldsymbol{\theta}$, on peut supposer que $\text{pen}^{id}_\Psi > 0$. On rappelle qu'une pénalité, définie en (2.4.1) à la Section 2.4, est la donnée d'une application

$$\begin{aligned} \text{pen}_\Psi\,:\, \mathcal{F} &\longrightarrow \mathbb{R} \\ \rho &\longmapsto \text{pen}_\Psi(\rho)\,. \end{aligned}$$

En pratique, la pénalisation est faite sur un modèle $F \subset \mathcal{F}$ de la forme $F = \{\rho_{\mathcal{F}}(\boldsymbol{\theta})\,,\, \boldsymbol{\theta} \in \Theta\} \subset \mathcal{F}$. Ainsi, une pénalité sur F est de la forme $\text{pen}_\Psi(\rho_{\mathcal{F}}(\boldsymbol{\theta}))$ et par abus de notation nous écrirons

$$\text{pen}_\Psi(\boldsymbol{\theta}) := \text{pen}_\Psi(\rho_{\mathcal{F}}(\boldsymbol{\theta}))\,.$$

D'après le Lemme 2.4.1, une pénalité pen_Ψ devrait satisfaire

$$\text{pen}^{id}_\Psi \leq \text{pen}_\Psi \leq (1+\delta)\,\text{pen}^{id}_\Psi \quad \text{uniformément sur } \Theta\,,$$

avec $\delta > 0$ le plus petit possible.

La notion de pénalité à laquelle nous nous intéressons doit être entendue comme une pénalité "naturelle", c'est à dire comme un terme sensé "réduire" l'erreur que l'on commet en remplaçant le risque $\mathcal{R}_\Psi(\boldsymbol{\theta})$ par sa (une) version empirique $\widehat{\mathcal{R}}_\Psi(\boldsymbol{\theta})$. Cette notion de pénalité naturelle peut être à l'origine de certaines pénalités *a priori* utilisées en pratique, cf. exemple de la ridge regression à la Section 2.7.

Un contraste pénalisé sera donc de la forme

$$\Psi_{\text{pen}}(\rho_{\mathcal{F}}(\boldsymbol{\theta}), z) = \Psi(\rho_{\mathcal{F}}(\boldsymbol{\theta}), z) + \text{pen}_\Psi(\boldsymbol{\theta})\,. \tag{2.52}$$

Les pénalités seront cherchées sous la forme

$$K\,\text{pen}_{shape}(\boldsymbol{\theta})\,,$$

où $K \in \mathbb{R}$ est une constante de calibration qui, en quelque sorte, "fidélise" la pénalisation aux données (la pénalité idéale dépend des données). Puis, $\text{pen}_{shape}(\boldsymbol{\theta})$ est une fonction de $\boldsymbol{\theta}$ seulement.

Remarque 2.6.1. La constante K à déterminer n'est pas nécessairement positive comme nous pourrions être tenté de le penser. En effet, une pénalité de la forme $\boldsymbol{\theta} \mapsto K\,\text{pen}_{shape}(\boldsymbol{\theta})$ est sensée se comporter comme la pénalité idéale, par exemple (2.50), qui est nulle en espérance.

La constante K peut être **positive** ou **négative**. La calibration de la constante K ne sera pas considérée, bien que ce soit une étape clé quant à l'efficacité d'une procédure d'estimation avec pénalisation. Toutefois, on pourra s'inspirer des travaux menés en sélection de modèle, par exemple [1].

Par conséquent, les estimateurs

$$\widehat{\boldsymbol{\theta}}_{\Psi_{\text{pen}}} = \underset{\boldsymbol{\theta}\in\Theta}{\text{Argmin}}\, \widehat{\mathcal{R}}_\Psi(\boldsymbol{\theta}) + K\,\text{pen}_{shape}(\boldsymbol{\theta})$$

sont fonctions de K, on notera $\widehat{\boldsymbol{\theta}}_{\Psi_{\text{pen}}} = \widehat{\boldsymbol{\theta}}_{\Psi_{\text{pen}}}(K)$.

De manière générale, on notera $\mathcal{K}$ un sous-ensemble (donné) de $\mathbb{R}$ tel que $K \in \mathcal{K}$.

2.7 Ridge regression et pénalité idéale

Dans cette section, nous tentons de donner une interprétation de la ridge regression, voir par exemple [4] Section 3.4.1, en termes de pénalité idéale. Autrement dit, nous essayons de comprendre l'utilisation de la pénalité $K\,\|\boldsymbol{\theta}\|_2^2$ avec $K > 0$ à travers la pénalité idéale donnée par le problème de régression.

2.7.1 Pénalisation $L_2(P^{\mathbf{x}})$

Soit Ψ_{reg} le $\mathcal{F}$-contraste ($\mathcal{F} = L_2(P^{\mathbf{x}})$)

$$\Psi_{reg}(\rho, (\mathbf{x}, y)) = (y - \rho(\mathbf{x}))^2\,.$$

Considérons le modèle $F = \{\rho_{\mathcal{F}}(\boldsymbol{\theta}) \,:\, \mathbf{x} \mapsto h(\mathbf{x}, \boldsymbol{\theta})\}$.
L'estimateur de regression est donné par

$$\widehat{\boldsymbol{\theta}}_{\Psi_{reg}} = \underset{\boldsymbol{\theta}\in\Theta}{\text{Argmin}}\, \frac{1}{n}\sum_{i=1}^{n} \Psi_{reg}\left(\rho_{\mathcal{F}}(\boldsymbol{\theta}), Z_i\right) = \underset{\boldsymbol{\theta}\in\Theta}{\text{Argmin}}\, \frac{1}{n}\sum_{i=1}^{n} \left(Y_i - h(\mathbf{X}_i, \boldsymbol{\theta})\right)^2\,.$$

Soit la pénalité $L_2(P^{\mathbf{x}})$ sur $\mathcal{F}$, donnée par

$$\text{pen}_{\Psi_{reg}} : \rho \in \mathcal{F} \longmapsto K \|\rho\|^2_{L_2(P^{\mathbf{x}})}, \quad K \in \mathbb{R}.$$

L'estimateur de regression sous le contraste pénalisé $\Psi_{\text{pen}}(\rho, z) = \Psi_{reg}(\rho, z) + \text{pen}_{\Psi_{reg}}(\rho)$ est donné par

$$\widehat{\boldsymbol{\theta}}_{\Psi_{pen}} = \underset{\boldsymbol{\theta}\in\Theta}{\text{Argmin}} \frac{1}{n}\sum_{i=1}^{n} (Y_i - h(\mathbf{X}_i, \boldsymbol{\theta}))^2 + K \|h(\cdot, \boldsymbol{\theta})\|^2_{L_2(P^{\mathbf{x}})}. \tag{2.53}$$

Supposons maintenant que h s'écrit

$$h(\mathbf{x}, \boldsymbol{\theta}) = \sum_{l=1}^{k} \phi_l(\mathbf{x})\, \boldsymbol{\theta}_l, \tag{2.54}$$

où les fonctions $\phi_l : \mathcal{X} \to \mathbb{R}$, $l = 1, ..., k$, forment une base orthonormale par rapport à la mesure $P^{\mathbf{x}}$. Alors, on vérifie aisément que pour tout $\boldsymbol{\theta} \in \Theta$

$$\|h(\cdot, \boldsymbol{\theta})\|^2_{L_2(P^{\mathbf{x}})} = \|\boldsymbol{\theta}\|^2_2,$$

et l'estimateur (2.53) s'écrit

$$\widehat{\boldsymbol{\theta}}_{\Psi_{pen}} = \widehat{\boldsymbol{\theta}}(K) = \underset{\boldsymbol{\theta}\in\Theta}{\text{Argmin}} \frac{1}{n}\sum_{i=1}^{n} (Y_i - h(\mathbf{X}_i, \boldsymbol{\theta}))^2 + K \|\boldsymbol{\theta}\|^2_2, \quad K \in \mathbb{R}. \tag{2.55}$$

On retrouve l'estimateur de *ridge regression* (2.5) en considérant une constante K positive.

Notons une entrée $\Phi(\mathbf{X}) = (\phi_1(\mathbf{X}), ..., \phi_k(\mathbf{X}))$ et $\Phi = \begin{pmatrix} \Phi(\mathbf{X}_1) \\ \cdot \\ \cdot \\ \cdot \\ \Phi(\mathbf{X}_n) \end{pmatrix}$.

L'estimateur (2.55) s'écrit alors

$$\widehat{\boldsymbol{\theta}}(K) = \underset{\boldsymbol{\theta}\in\Theta}{\text{Argmin}} \|\mathbf{Y} - \Phi\,\boldsymbol{\theta}\|^2_2 + K \|\boldsymbol{\theta}\|^2_2, \quad K \in \mathbb{R}. \tag{2.56}$$

2.7.2 Légitimité de la pénalisation $K\|\boldsymbol{\theta}\|^2_2$ avec $K > 0$

En reprenant ce qui a été dit à la Section (2.4.3), une première justification de la pénalité $K\|\boldsymbol{\theta}\|^2_2$, dans le cas où $K > 0$, réside dans le faite qu'elle améliore le conditionnement du problème tout en le gardant convexe (cf. méthode de Tikhonov).
Par ailleurs, tentons de voir s'il y a "un lien", une justification, entre la pénalité $\text{pen}_{\Psi_{reg}}(\boldsymbol{\theta}) = K\|\boldsymbol{\theta}\|^2_2$ et la pénalité idéale $\text{pen}^{id}_{\Psi_{reg}}$. Le fait de considérer $K > 0$ sera étudié dans un second temps.
Arrêtons-nous un instant sur $\text{pen}^{id}_{\Psi_{reg}}$.
Dans ce cas, la pénalité idéale est

$$\text{pen}^{id}_{\Psi_{reg}}(\boldsymbol{\theta}) \;=\; \frac{1}{n}\sum_{i=1}^{n} \left(\mathbb{E}_{Q^{\mathbf{z}}} \Psi_{reg}\left(\rho_{\mathcal{F}}(\boldsymbol{\theta}), Z\right) - \Psi_{reg}\left(\rho_{\mathcal{F}}(\boldsymbol{\theta}), Z_i\right)\right), \quad \rho_{\mathcal{F}}(\boldsymbol{\theta}) = h(\cdot, \boldsymbol{\theta}).$$

En considérant l'hypothèse sur h (2.54), on montre facilement que

$$\text{pen}^{id}_{\Psi_{reg}}(\boldsymbol{\theta}) \;=\; \boldsymbol{\theta}^T A_n \boldsymbol{\theta} + b_n \boldsymbol{\theta} + c_n, \tag{2.57}$$

où

$$A_n = I_k - \frac{1}{n}\sum_{i=1}^{n} \Phi^T(\mathbf{X}_i)\Phi(\mathbf{X}_i)\,, \quad \Phi(\mathbf{X}_i) = (\phi_1(\mathbf{X}_i), ..., \phi_k(\mathbf{X}_i))$$

puis $b_n = \frac{2}{n}\sum_{i=1}^{n}\left(Y_i\Phi(\mathbf{X}_i) - \mathbb{E}(Y\Phi(\mathbf{X}))\right)$ et $c_n = \frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{E}(Y^2) - Y_i^2\right)$.
I_k est la matrice identité de dimension k.
La constante c_n ne dépend pas de $\boldsymbol{\theta}$, sans perte de généralité on écrit

$$\text{pen}^{id}_{\Psi_{reg}}(\boldsymbol{\theta}) \quad = \quad \boldsymbol{\theta}^T A_n \boldsymbol{\theta} + b_n\boldsymbol{\theta}\,. \tag{2.58}$$

Cette pénalité demeure toujours inconnue car même si la matrice A_n est parfaitement connue, le vecteur b_n dépend de la loi du couple $(\mathbf{X}, Y)$ et est donc inconnu.
Toutefois, notons λ^n_{min} et λ^n_{max} la plus petite et la plus grande valeur propre de A_n, respectivement, qui sont réelles car A_n est symétrique. Il est clair que

$$\lambda^n_{min}\|\boldsymbol{\theta}\|_2^2 + b_n\boldsymbol{\theta} \leq \text{pen}^{id}_{\Psi_{reg}}(\boldsymbol{\theta}) \leq \lambda^n_{max}\|\boldsymbol{\theta}\|_2^2 + b_n\boldsymbol{\theta}\,.$$

Par cette dernière inégalité, il est naturel de considérer une pénalité de la forme

$$\text{pen}_{\Psi}(\boldsymbol{\theta}) = K\,\|\boldsymbol{\theta}\|_2^2\,,$$

où K appartient à un certain intervalle $\mathcal{K} \subset \mathbb{R}$. Notons que, pour l'instant, aucun élément déterminant assure ou justifie que K soit positif (même si cela peut être intuitif). En effet, nous n'avons pas nécessairement $\lambda^n_{min} \geq 0$. D'après le Lemme 2.4.1, une "bonne" pénalité $\text{pen}_{\Psi_{reg}}$ devrait satisfaire

$$\text{pen}^{id}_{\Psi_{reg}} \leq \text{pen}_{\Psi_{reg}} \leq (1+\delta)\,\text{pen}^{id}_{\Psi_{reg}} \quad \text{uniformément sur } \Theta\,,$$

avec $\delta > 0$ le plus petit possible. Autrement dit, la "forme" de la pénalité $\text{pen}_{\Psi_{reg}}$ devrait être proche de celle de $\text{pen}^{id}_{\Psi_{reg}}$. Justifier la pénalité $\text{pen}_{\Psi}(\boldsymbol{\theta}) = K\,\|\boldsymbol{\theta}\|_2^2$, $K \in \mathcal{K}$ par la considération précédente est un travail assez délicat. Nous proposons dans ce cas précis d'analyser le risque $\mathcal{R}_{\Psi_{reg}}(\widehat{\boldsymbol{\theta}}(K))$ en tant que fonction de K, et en particulier nous justifierons le fait de considérer $K > 0$.
On montre la proposition suivante :

Proposition 2.7.1. *Il existe $\delta > 0$ tel que pour tout $K \in [0, \delta]$,*

$$\mathbb{E}_{(\mathbf{X}_i, Y_i)_{i=1..n}}\left(\mathcal{R}_{\Psi_{reg}}(\widehat{\boldsymbol{\theta}}(K))\right) \leq \mathbb{E}_{(\mathbf{X}_i, Y_i)_{i=1..n}}\left(\mathcal{R}_{\Psi_{reg}}(\widehat{\boldsymbol{\theta}}_{\Psi_{reg}})\right)\,.$$

La preuve est donnée à la Section 2.9.

En regardant $\widehat{\boldsymbol{\theta}}(K)$ comme un Ψ_K-estimateur $\widehat{\boldsymbol{\theta}}_{\Psi_K} = \widehat{\boldsymbol{\theta}}(K)$, où $\Psi_K(\rho_{\mathcal{F}}(\boldsymbol{\theta}), \mathbf{z}) = \Psi_{reg}(\rho_{\mathcal{F}}(\boldsymbol{\theta}), \mathbf{z}) + K\,\|\boldsymbol{\theta}\|_2^2$, on a montré qu'il existe un (des) K tel que

$$\mathbb{E}_{(Y_i, \mathbf{X}_i)_{i=1..n}}\left(\mathcal{R}_{\Psi_{reg}}(\widehat{\boldsymbol{\theta}}_{\Psi_K})\right) \leq \mathbb{E}_{(Y_i, \mathbf{X}_i)_{i=1..n}}\left(\mathcal{R}_{\Psi_{reg}}(\widehat{\boldsymbol{\theta}}_{\Psi_{reg}})\right)\,,$$

autrement dit, il existe une procédure qui est meilleure (en moyenne) que celle de la régression classique dont le contraste associé Ψ_{reg} est celui qui caractérise la grandeur d'intérêt visée (i.e l'espérance conditionnelle). Dans le Chapitre 4, la procédure relative au contraste Ψ_K sera qualifiée de *procédure croisée* (car on veut prédire une grandeur caractérisée par Ψ_{reg}) et celle relative à Ψ_{reg} de *procédure classique*. Intuitivement, on pourrait s'attendre à ce que les

procédures classiques donnent de meilleurs résultats, en termes de risque de prédiction, que les autres procédures dont l'estimation des paramètres peut être assez "étrangère" vis-à-vis de la prédiction cherchée. Ici, le contraste Ψ_K n'est pas si "étrange" que cela pour prédire l'espérance conditionnelle puisqu'il n'est qu'une version pénalisée du contraste Ψ_{reg}.
Nous verrons ceci un peu plus en détail dans le Chapitre 4 où un résultat similaire à la Proposition 2.7.1, la Proposition 4.5.1, sera établi. Nous donnerons également plusieurs exemples numériques.

2.8 Enjeux de l'apprentissage d'un computer experiment

2.8.1 Enjeux pour les problématiques de prédiction

Les deux exemples de la Section 2.5.4 tendent à souligner le fait que si l'on veut prédire, par le biais d'un computer experiment, une caractéristique liée au phénomène Y autre que l'espérance conditionnelle, des procédures statistiques non standards sont à mettre en oeuvre.
Par ailleurs, on fait remarquer qu'une réponse partielle peut être apportée à une question posée à la Section 2.2.3, où il s'agissait de savoir s'il existait des procédures "optimales" pour la prédiction de quantités d'intérêt autre que l'espérance conditionnelle (e.g moyenne, quantile, dépassement de seuil etc...). En effet, d'après (2.13), toute quantité d'intérêt $\rho_{\mathcal{F}}$ peut être vue comme l'*argmin* de l'espérance d'un contraste Ψ à déterminer. Ensuite, le schéma de prédiction est similaire à celui donné en exemple plus haut, notamment celui de la prédiction de la densité f.

► Un premier travail consiste à étudier les estimateurs du type (2.42) dans un cadre général. Ce sera l'objet du Chapitre 3.

Ensuite, rappelons l'exemple donné à la section (2.2.3) où l'on s'interrogeait sur le sens de la quantité

$$\mathbb{P}(h(\mathbf{X}, \widehat{\boldsymbol{\theta}}_{reg}) > s) \quad (\widehat{\boldsymbol{\theta}}_{reg} = \text{moindres carrés}),$$

afin de prédire

$$\rho_{\mathcal{F}} = \mathbb{P}(Y > s).$$

Il est vrai qu'une procédure de régression paramétrique peut sembler assez "générale", voire robuste, vis-à-vis d'autres prédictions que nous pouvons calculer avec le modèle stochastique

$$\mathbf{X} \longmapsto h(\mathbf{X}, \widehat{\boldsymbol{\theta}}_{reg}).$$

Mais qu'en est-il quantitativement ?

► Avec le formalisme de la Section 2.5, le problème peut se poser comme suit. Supposons que l'on veuille prédire une caractéristique $\rho_{\mathcal{F}^p} \in \mathcal{F}^p$ et soit Ψ^p un $\mathcal{F}^p$-contraste qui caractérise $\rho_{\mathcal{F}^p}$. Soit Ψ un quelconque $\mathcal{F}$-contraste et notons $\widehat{\boldsymbol{\theta}}_{\Psi^p}$ et $\widehat{\boldsymbol{\theta}}_{\Psi}$ les estimateurs associés au contrastes Ψ^p et Ψ, respectivement.
Il semble alors naturel de s'interroger sur la performance de $\widehat{\boldsymbol{\theta}}_{\Psi^p}$ et $\widehat{\boldsymbol{\theta}}_{\Psi}$ vis-à-vis de la prédiction voulue $\rho_{\mathcal{F}^p}$. Autrement dit, on se pose la question du signe de la difference suivante

$$\mathcal{R}_{\Psi^p}(\widehat{\boldsymbol{\theta}}_{\Psi}) - \mathcal{R}_{\Psi^p}(\widehat{\boldsymbol{\theta}}_{\Psi^p}).$$

Ce sera l'objet du Chapitre 4.

2.8.2 Enjeux pour les problèmes inverses stochastiques

Les Ψ-estimateurs du type

$$\widehat{\boldsymbol{\theta}}_{\Psi} = \underset{\boldsymbol{\theta}\in\Theta}{\text{Argmin}} \frac{1}{n}\sum_{i=1}^{n} \Psi\left(\rho_{\mathcal{F}}^{m}(\boldsymbol{\theta}), Y_i\right)$$

peuvent intervenir dans le cas où on dispose d'observations $Y_1, ..., Y_n$ et d'un modèle numérique *bruité* $h(\mathbf{X}, \boldsymbol{\theta})$, où $\mathbf{X}$ est le bruit et $\boldsymbol{\theta}$ les paramètres à estimer. L'étude de ce type d'estimateur est donnée dans le Chapitre 3. Une application sera également donnée dans le Chapitre 6 où il s'agira de caractériser une performance aéronautique par un problème inverse à partir de mesures et d'un modèle comportant des variables incertaines.

2.9 Preuve de la Proposition 2.7.1

Calculons l'espérance du risque $\mathcal{R}_{\Psi_{reg}}(\widehat{\boldsymbol{\theta}}(K))$ sous les données $(\mathbf{X}_i, Y_i)_{i=1..n}$. Considérons une re-paramétrisation du problème. La matrice Φ admet une décomposition en valeurs singulières

$$\Phi = U\, D\, V^T,$$

où U et V sont des matrices orthogonales de taille $n \times k$ et $k \times k$, respectivement. D est une matrice diagonale à élément positifs ou nuls $d_1, ..., d_k$ appelés *valeurs singulières*. Ces trois matrices dépendent des données aléatoires $\mathbf{X}_{1..n} = \mathbf{X}_1, ..., \mathbf{X}_n$ qu'on pourra geler. Maintenant, remplaçons les entrées $\Phi(\mathbf{x})$ par les entrées $\Phi(\mathbf{x})\, V$ (composantes principales), ou encore, considérons la matrice des entrées $\widetilde{\Phi} = \Phi\, V$ à la place de Φ. Ainsi, l'estimateur (2.56) s'écrit

$$\begin{aligned} \widehat{\boldsymbol{\theta}}(K) &= \left(\widetilde{\Phi}^T\widetilde{\Phi} + K\, I_k\right)^{-1} \widetilde{\Phi}^T \mathbf{Y} \\ &= \left(D^2 + K\, I_k\right)^{-1} D\, U^T\, \mathbf{Y}, \end{aligned} \tag{2.59}$$

où I_k est la matrice identité de dimension k et $\mathbf{Y} = (Y_1, ..., Y_n)^T$. Supposons de plus qu'il existe un $\boldsymbol{\theta}^*$ tel que

$$Y_i = \widetilde{\Phi}(\mathbf{X}_i)\, \boldsymbol{\theta}^* + \epsilon_i, \quad i = 1, ..., n$$

où les ϵ_i sont i.i.d centrés et indépendants de $\mathbf{X}$, de variance σ^2. Alors pour tout $\theta \in \Theta$,

$$\mathcal{R}_{\Psi_{reg}}(\boldsymbol{\theta}) = \sigma^2 + \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^2.$$

En particulier, après calcul nous obtenons sans difficulté que

$$\mathcal{R}_{\Psi_{reg}}(\widehat{\boldsymbol{\theta}}(K)) = \sigma^2 + \sum_{l=1}^{k} \frac{(\theta_l^*)^2 K^2 + \sigma^2\, d_l^2}{(K + d_l^2)^2} + 2\, K\, (D^2 + K\, I_k)^{-2} D\, \boldsymbol{\epsilon} \tag{2.60}$$

où $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_n)^T$. Par conséquent, en prenant l'espérance conditionnellement à $\mathbf{X}_{1..n} = \mathbf{x}_{1..n}$ (où $\mathbf{x}_{1..n} = \mathbf{x}_1, ...\mathbf{x}_n$) le troisième terme est nul et on a simplement

$$\mathbb{E}_{Y_{1..n}/\mathbf{X}_{1..n}}\left(\mathcal{R}_{\Psi_{reg}}(\widehat{\boldsymbol{\theta}}(K))\right) = \sigma^2 + \sum_{l=1}^{k} \frac{(\theta_l^*)^2 K^2 + \sigma^2\, d_l^2}{(K + d_l^2)^2}.$$

La dépendance en les données $\mathbf{X}_{1..n}$ de $\mathbb{E}_{Y_{1..n}/\mathbf{X}_{1..n}}\left(\mathcal{R}_{\Psi_{reg}}(\widehat{\boldsymbol{\theta}}(K))\right)$ ne se fait qu'au travers des d_l, $l = 1, ..., k$ qui, on le rappelle, sont les valeurs singulières de la matrice Φ.
Notons que lorsque $K = 0$, i.e pas de pénalisation, la quantité précédente vaut

$$\mathbb{E}_{Y_{1..n}/\mathbf{X}_{1..n}}\left(\mathcal{R}_{\Psi_{reg}}(\widehat{\boldsymbol{\theta}}(0))\right) = \sigma^2(1 + \text{trace}(D^{-2})),$$

qui est le risque de l'estimateur classique des moindres carrés. Désormais, l'enjeu est de comparer les valeurs que prend $\mathbb{E}_{Y_{1..n}/\mathbf{x}_{1..n}}\left(\mathcal{R}_{\Psi_{reg}}(\widehat{\boldsymbol{\theta}}(K))\right)$ lorsque K varie, avec le risque $\sigma^2(1+\text{trace}(D^{-2}))$ donné lorsqu'on ne pénalise pas. Pour simplifier les calculs, nous présenterons les variations à l fixé en considérant l'application

$$K \mapsto \frac{(\theta_l^*)^2K^2+\sigma^2 d_l^2}{(K+d_l^2)^2}\,, \quad K \in \mathbb{R}/\{-d_l^2\}\,,$$

dont les valeurs seront comparées à $R_0 = \frac{\sigma^2}{d_l^2}$. Notons que R_0 est la variance de la l^{ieme} composante de l'estimateur classique des moindres carrés.
Etudions donc les variations de cette fonction en posant $s := \sigma^2$, $t := (\theta_l^*)^2$, $d := d_l^2$ et

$$g(K) = \frac{t\,K^2+s\,d}{(K+d)^2}\,, \quad K \in \mathbb{R}/\{-d\}\,,$$

avec $R_0 = \frac{s}{d}$. Distinguons deux cas.

Si $\mathbf{t=0}$, on a $g(K) = \frac{s\,d}{(K+d)^2}$ et on obtient le tableau de variations suivant

K	$-\infty$		$-2d$		$-d$		0		$+\infty$
$g'(K)$		$+$				$-$			
g	0	↗	R_0	↗	$+\infty$	$+\infty$	↘	R_0	↘ 0

Ainsi, on a $g(K) \leq R_0$ si et seulement si $K \in \mathcal{K}_1 :=]-\infty, -2\,d] \cup [0, +\infty[$.

Si $\mathbf{t>0}$, le numérateur de la dérivée de $K \mapsto g(K)$ vaut

$$2\,d\,(t\,K^2+(t\,d-s)K-s\,d)\,,$$

et le Δ du second facteur est donné par

$$\Delta = (t\,d+s)^2\,.$$

La dérivée s'annule donc deux fois : en $K_1 = -d$ (valeur interdite) et en $K_2 = s/t$.
Pour savoir quand g coupe la droite $y = R_0$ on doit considérer trois sous-cas :

- $\mathbf{0 < t < R_0}$:

K	$-\infty$		$\frac{2s}{t-R_0}$		$-d$		0		$\frac{s}{t}$		$+\infty$
$g'(K)$		$+$				$-$			0	$+$	
g	t	↗	R_0	↗	$+\infty$	$+\infty$	↘	R_0	↘ $\frac{R_0}{1+R_0/t}$	↗	t

Dans ce cas, $g(K) \leq R_0$ si et seulement si $K \in \mathcal{K}_2 :=]-\infty, \frac{2s}{t-R_0}] \cup [0, +\infty[$.

- $\mathbf{t = R_0}$:

K	$-\infty$		$-d$	0	$\frac{s}{t}$	$+\infty$
$g'(K)$	$+$		$\|$	$-$	0	$+$
g	$t = R_0$	$+\infty$	$+\infty$	R_0	$\frac{R_0}{1+R_0/t}$	$t = R_0$

Dans ce cas, $g(K) \leq R_0$ si et seulement si $K \in \mathcal{K}_3 := [0, +\infty[$.

- $\mathbf{t} > \mathbf{R_0}$:

K	$-\infty$		$-d$	0	$\frac{s}{t}$	$\frac{2s}{t-R_0}$	$+\infty$
$g'(K)$	$+$		$\|$	$-$	0	$+$	
g	t	$+\infty$	$+\infty$	R_0	$\frac{R_0}{1+R_0/t}$	R_0	t

Dans ce cas, $g(K) \leq R_0$ si et seulement si $K \in \mathcal{K}_4 := [0, \frac{2s}{t-R_0}]$.

On illustre ces variations dans le Tableau 2.3 et à la Figure 2.3.
En pratique, on ne connaît pas a priori de conditions sur t et par souci de robustesse on ne peut se placer dans l'un des cas précédents. Si on note $\mathcal{K}(t) = \cap_{i=1}^{4} \mathcal{K}_i$ l'intersection des ensembles $\mathcal{K}_i$ (qui dépend de t), on a que $\mathcal{K}(t) \neq \varnothing$ et $\mathcal{K}(t) \subset [0, +\infty[$, cela pour tout s, d et t. Autrement dit, quelque soit les hypothèses sur les données, il existe un $\delta > 0$ tel que pour tout $K \in [0, \delta]$ on a que $g(K) \leq R_0$. Ou encore, en sommant sur les $l = 1, ..., k$, il vient que pour $K \in [0, \delta]$

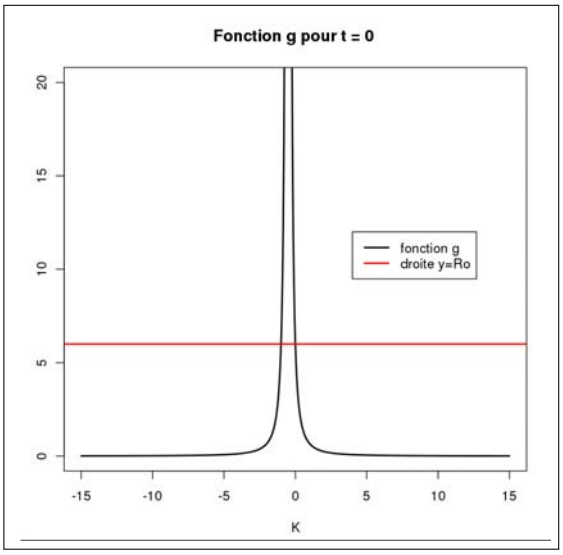
$$\mathbb{E}_{Y_{1..n}/\mathbf{X}_{1..n}} \left(\mathcal{R}_{\Psi_{reg}}(\widehat{\boldsymbol{\theta}}(K)) \right) \leq \mathbb{E}_{Y_{1..n}/\mathbf{X}_{1..n}} \left(\mathcal{R}_{\Psi_{reg}}(\widehat{\boldsymbol{\theta}}_{\Psi_{reg}}) \right) ,$$

et même

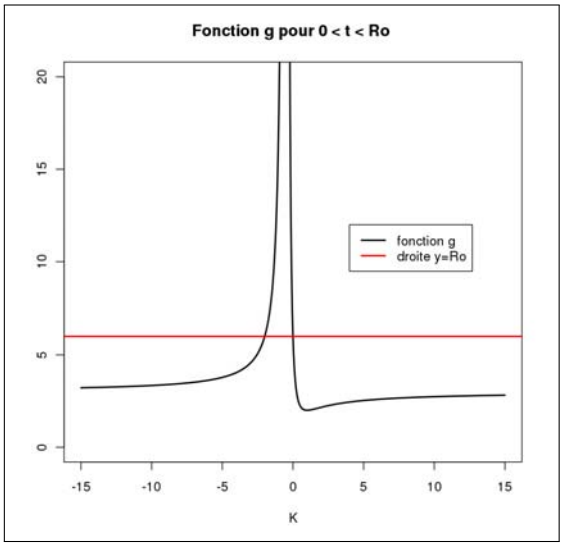
$$\mathbb{E}_{(Y_i,\mathbf{X}_i)_{i=1..n}} \left(\mathcal{R}_{\Psi_{reg}}(\widehat{\boldsymbol{\theta}}(K)) \right) \leq \mathbb{E}_{(Y_i,\mathbf{X}_i)_{i=1..n}} \left(\mathcal{R}_{\Psi_{reg}}(\widehat{\boldsymbol{\theta}}_{\Psi_{reg}}) \right) ,$$

avec $\widehat{\boldsymbol{\theta}}_{\Psi_{reg}} = \widehat{\boldsymbol{\theta}}(0)$ l'estimateur classique des moindres carrés (sans pénalisation).

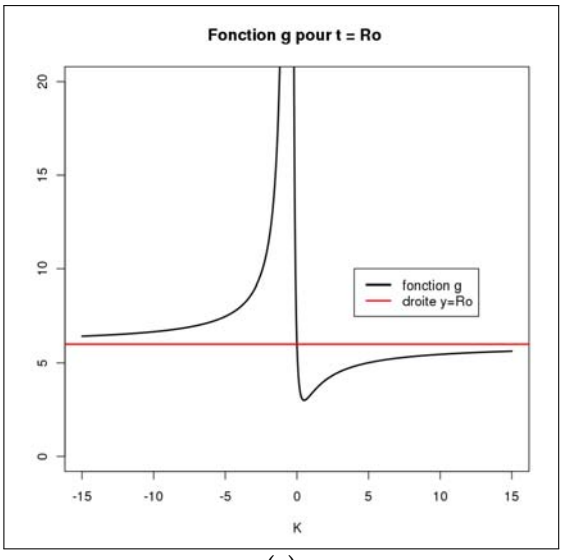
Ce qui conclut la preuve de la Proposition 2.7.1.



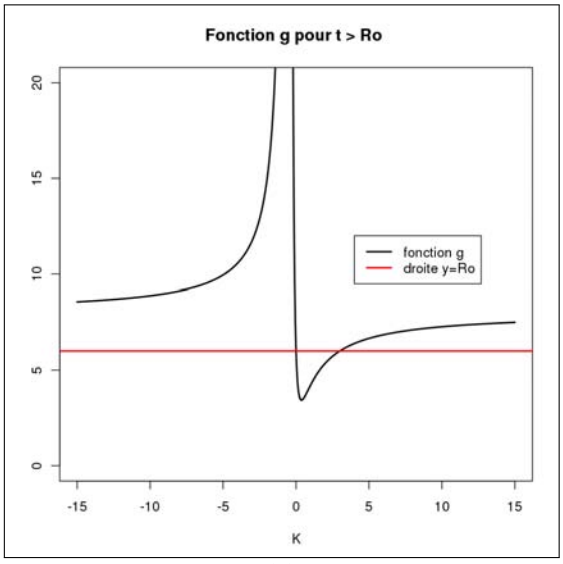
(a)



(b)



(c)



(c)

FIGURE 2.3 – Fonction g pour différentes valeurs de t

Bibliographie

[1] S. Arlot and P. Massart. Data-driven calibration of penalties for least-squares regression. *The Journal of Machine Learning Research*, 10 :245–279, 2009.

[2] S. Boucheron, O. Bousquet, and G. Lugosi. (chapter) concentration inequalities. *Machine Learning Summer School 2003*, 3176 :169–207, 2004.

[3] O. Bousquet, S. Boucheron, and G. Lugosi. (chapter) introduction to statistical learning theory. *Machine Learning Summer School 2003*, 3176 :208–240, 2004.

[4] T. Hastie, R. Tibshirani, and J.H. Friedman. *The elements of statistical learning : data mining, inference, and prediction*. Springer Verlag, 2009.

[5] P.J. Huber. *Robust statistics*. Wiley-Interscience, 1981.

[6] P. Massart. *Concentration inequalities and model selection : Ecole d'Eté de Probabilités de Saint-Flour XXXIII-2003*. Springer Verlag, 2007.

[7] A.N. Tikhonov. *On the stability of inverse problems*, volume 39. 1943.

[8] V. Vapnik. *Statistical learning theory*. Wiley New York, 1998.

[9] V. Vapnik. *The nature of statistical learning theory*. Springer Verlag, 2000.

Chapitre 3

Bornes de risque de M-estimateurs construits à partir de modèles boîte noire

Sommaire

Résumé du Chapitre

L'objet de ce chapitre est l'estimation de paramètres dans les modèles boîte noire $(\mathbf{X}, \boldsymbol{\theta}) \mapsto h(\mathbf{X}, \boldsymbol{\theta})$ à partir des observations $Y_1, ..., Y_n$ et $\mathbf{X}'_1, ..., \mathbf{X}'_m$, avec $\mathbf{X}'_j \sim \mathbf{X}$. Nous présentons une procédure générale d'estimation basée sur la *M-estimation*, tenant compte de l'aspect simulation numérique.

Risk bound for new M-estimation problems

Nabil Rachdi [1], Jean-Claude Fort [2], Thierry Klein [3]

Abstract
In this paper, we develop new algorithms for parameter estimation in the case of models type Input/Output in order to represent and to characterize a phenomenon Y. From experimental data $Y_1, ..., Y_n$ supposed to be i.i.d from Y, we prove a risk bound qualifying the proposed procedures in terms of the number of experimental data n, computing budget m and model complexity. The methods we present are general enough which should cover a wide range of applications.

3.1 Introduction

As in many statistical problems, we are interested in investigating the stochastic behavior of a random variable Y. We have at disposal an i.i.d sample $Y_1, ..., Y_n$. These data come from experiments that could be real or the result of a computer code. In an industrial context, it is not rare that the size of the available set of data is small. This is due either to the cost of each real experiment or to the very long time needed for each run of a simulation code. It is encountered in various field of industry : meteorology, oil extraction, nuclear security, aeronautic, mechanical engineering etc...
Besides these costly experiments or codes, various reduced models are available. Even if they still are complicated, one can use them to simulate in a reasonable computing time and obtain large samples from simulations.
These reduced models depend on unknown parameters that need to be estimated. So that the reduced models take the following form : $(\mathbf{x}, \boldsymbol{\theta}) \in \mathcal{X} \times \Theta \mapsto h(\mathbf{x}, \boldsymbol{\theta})$. Generally the variables $\mathbf{x}$ used to built the reduced models are not the same that have been measured (if they have) as experimental conditions during the experiments leading to the data $Y_1, ..., Y_n$. That is why in this work we do not suppose that the available data are couples of input/ouput variables $(\mathbf{x}_i, Y_i)$: the variables $\mathbf{x}_i$ are not available or are not the same used in the recuded models. Thus our only experimental data are the Y_i's.
Let us take an example of particular interest coming from EADS [4] Research department : the effect of an electromagnetic field on the behavior of an aircraft. When lightning or an electromagnetic field strikes an aircraft, sensors measure data corresponding to the intensity of such field in various part of the aircraft. The data recorded are dispersed due to the intrinsic variability of the phenomenon. In our framework, information of one sensor is represented by the sample $Y_1, ..., Y_n$. On another side, we have at disposal several computer codes h modeling the electromagnetic field in function of input variables $\mathbf{x}$ and parameters $\boldsymbol{\theta}$ that can be tuned. The result of such a computer code is a function $h(\mathbf{x}, \boldsymbol{\theta})$. The variables $\mathbf{x}$ will be modeled by random variables, for instance it could be variables describing the atmospheric conditions, the angles of the lightning $w.r.t.$ the aircraft,....The vector parameter $\boldsymbol{\theta}$ is part of

1. Institut de Mathématiques de Toulouse - EADS Innovation Works, 92152 Suresnes
2. Université Paris Descartes, 45 rue des saints pères, 75006 Paris
3. Institut de Mathématiques de Toulouse, 118 route de Narbonne F-31062 Toulouse
4. EADS : European Aeronautic Defense and Space Company

the model and will be estimated. In this case the computer codes have various degrees of complexity. Actually, one has at disposal a set of models $\mathcal{H}$ covering all available models : from the simplest to the most complicated. Hence, another important issue would be to "select" a model among the set $\mathcal{H}$ for a specific use. We don't treat this aspect in this paper, we work with only one model h.

In general these reduced models remains complex in the following meaning. For $\boldsymbol{\theta} \in \Theta$, let us consider a feature of the random model output $h(\mathbf{X}, \boldsymbol{\theta})$, for instance its mean, its variance, a quantile or its probability density function. We will say that h is a *complex model* if the feature we are interested in is analytically *unreachable* in $\boldsymbol{\theta}$. *Complex models* can arise from several ways. For example, the function $h(\cdot, \boldsymbol{\theta})$ can have a complicated form due to the high complexity of the modeling, or the function can be a *black box* function input/output and so, not with an analytical form. This situation is very common in engineering, where complex models exist and are only known through simulations. This aspect is the principal motivation of our work.
In this context, shortly speaking, our goal is to construct a *Random Simulator*, $\mathbf{X} \mapsto h(\mathbf{X}, \widehat{\boldsymbol{\theta}})$ with $\mathbf{X}$ some random variable, predicting as well as possible some feature of the distribution of the observed data $Y_1, ..., Y_n$.
This is not very far from the framework of (Auffray, Barbillon, Marin et Pierre Barbillon) where they look for good metamodels of a time consuming black-box in order to evaluate probability of rare events by simulation. Yet, in this paper we are not interested in building or analyzing metamodels, but we try to optimally use experimental data and simulated data of a given metamodel, which are not directly coupled. See also (Pierre Barbillon, Gilles Celeux, Agnès Grimaud, Yannick Lefebvre, Étienne De Rocquigny) [1].

This paper is the theoretical part of a work on industrial applications in the field of "Uncertainty Management" [3]. We aim at studying a data-dependent model which outputs are "close to" some observed data (*experimental data*). The results we present are theoretical in that the estimation procedures we propose don't include practical implementations. The same is true for the modeling aspect : we deal with (input/output) models without specifying what can be done in practice. We do not deal with the pertinence of the possible reduced models (*metamodels*) (see [11, 25, 18, 20]). The impact of modeling technics will be treated in a forthcoming paper where we will apply some results obtained in this study in an industrial context.

The main tool of our development is the empirical processes theory. This theory constitutes a mathematical toolbox of asymptotics statistics and more recently non-asymptotic statistics. It was first explored in the 1950's by the work on Functional Central Limit Theorem [5]. Along the years, the development of empirical processes theory increased successfully thanks to work of many contributors, R.M. Dudley [6], D. Pollard [17], P. Gaenssler [7], Galen R. Shorack and Jon A. Wellner [19] and others. More recently, many references give a general overview of this theory with its applications to statistics, for example [24, 22, 13]. Empirical processes give power tools for evaluating statistical estimation and inference problems. Essential developments of non-asymptotic theory have been done in the last decade by the use of concentration inequalities to derive risk bounds [21], [16] and [12] among others. Our work directly derives from these advances.

Estimation based on minimizing a function was introduced by Huber in 1964 [9] where he proposed generalizing maximum likelihood estimation. The estimators resulting are called

M-estimator ("M" for minimizing or maximizing) [10]. The class of M-estimators is a broad class because many estimation procedure can be viewed as M-estimation, maximum likelihood and least-squares estimators are some of the most important examples. Asymptotic properties of these estimators were widely studied in a general context, and many authors like [22] or [23] used empirical processes theory which turn out to be a very valuable tool.

We present a general method where the criterion to minimize depends on both experimental and simulated data. This paper is organized as follows. In Section 3.2 we describe our general framework. In Section 3.3 we present the inverse problem. In Section 3.4 we establish Theorem 3.4.1 providing a risk bound for inverse problems based on both experimental and simulated data. In Section 3.5 we give some comments. In Section 3.6 and 3.7 we discuss about constants in Theorem 3.4.1. Section 3.8 is devoted to the proofs.

3.2 General setting

3.2.1 The model

- Probabilistic modeling.
Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space. We assume that all random variables are defined on this probability space.
Let a complex phenomenon modeled by a random real valued variable $Y \in \mathcal{Y}$, with distribution unknown Q and f the associated (Lebesgue) density function. Let us assume that $\mathcal{Y} \subset [-M, M]$, $M > 0$.
Let us suppose that a n-sample $Y_1, ..., Y_n$ is available : we call it *experimental data*.
Next, we suppose that this complex phenomenon can be approximately represented by the outputs $h(\mathbf{x}, \boldsymbol{\theta})$ of a *reduced model* h.

$$\begin{array}{rccl} h : & \mathcal{X} \times \Theta & \longrightarrow & \mathcal{Y} \\ & (\mathbf{x}, \boldsymbol{\theta}) & \longmapsto & h(\mathbf{x}, \boldsymbol{\theta}) \end{array}$$

where $\mathcal{X} \subset \mathbb{R}^d$ (*input space*), $\Theta \subset \mathbb{R}^k$ compact (*parameter space*).
We equip the input space $\mathcal{X}$ with a probability measure $P^{\mathbf{x}}$ which forms a probability space $(\mathcal{X}, \mathcal{B}, P^{\mathbf{x}})$. The probability measure $P^{\mathbf{x}}$ is not supposed to be known, we will only assume having at disposal a sample drawn from this distribution. In the case where $P^{\mathbf{x}}$ is known, without loss of generality, one can simply consider the uniform distribution on $[0, 1]$ provided to apply a well known probabilistic transformation.
The input vector is a random vector $\mathbf{X}$ defined on this space, and so, the output vector $h(\mathbf{X}, \boldsymbol{\theta})$ is a random real valued variable, for each $\boldsymbol{\theta} \in \Theta$. We emphasize that the $\mathbf{X}_j'$'s are variables used to produce model outputs but are not the inputs that gave the Y_i's. In practice, the data $\mathbf{X}_1', ..., \mathbf{X}_m'$ either arise from simulations of the random variable $\mathbf{X}$ with known distribution $P^{\mathbf{x}}$ or in some cases from a large data base (from experiments etc...).
The space $\mathcal{Y}$ is equipped with a σ-algebra $\mathcal{E}$ so as to ensure the measurability of the functions

$$\begin{array}{rccl} h(\cdot, \boldsymbol{\theta}) : & (\mathcal{X}, \mathcal{B}, P^{\mathbf{x}}) & \longrightarrow & (\mathcal{Y}, \mathcal{E}) \\ & \mathbf{X} & \longmapsto & h(\mathbf{X}, \boldsymbol{\theta}) \,. \end{array}$$

Moreover, we suppose given m realizations of the input random vector $\mathbf{X}$,

$$\mathbf{X}_1', ..., \mathbf{X}_m'$$

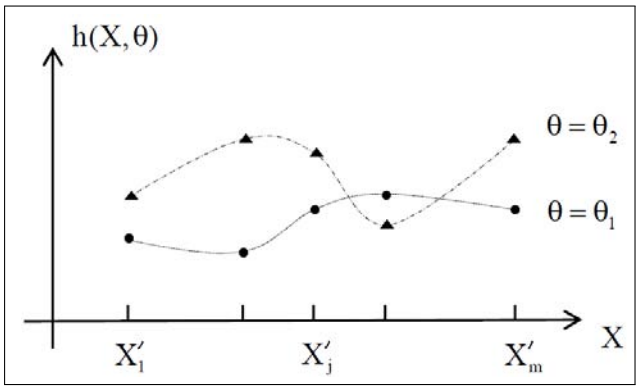


FIGURE 3.1 – Example of model outputs with 2 different parameters.

which provides m outputs called *simulated data*

$$h(\mathbf{X}'_1, \boldsymbol{\theta}), ..., h(\mathbf{X}'_m, \boldsymbol{\theta}) \quad \textit{for all} \quad \boldsymbol{\theta} \in \Theta.$$

In this paper, we develop a general method for estimating the parameter $\boldsymbol{\theta}$ based on the *training data* made of the experiments results $Y_1, ..., Y_n$ and the simulated inputs of the reduced model h, $\mathbf{X}'_1, ..., \mathbf{X}'_m$. The outputs of the model will depend on the parameter $\boldsymbol{\theta}$ to be estimated. The method we propose is general enough to include some specific problems met in practice. Indeed, two kinds of statistical analysis involving inverse problems can be considered : *Identification* and *Prediction*.

- Identification.
This analysis consists in estimating the "true" parameter $\boldsymbol{\theta}^\star$. It aims at estimating "physical" parameters having a real signification like dimensions or material properties for instance.

- Prediction.
In prediction, one wants to estimate a parameter $\boldsymbol{\theta}^\star$ (not necessarily unique) in order to predict the random phenomenon Y. One hopes that $h(\mathbf{X}, \boldsymbol{\theta}^\star) \approx Y$, in the sense that its distribution shares some feature with those of the distribution of Y : the same mean, variance, probabilty tail or the same probability density function.
Here, the parameter $\boldsymbol{\theta}^\star$ may have no real (physical) signification. For instance it is the case when using reduced models given by a Multi-Layer Perceptron, where the parameter is the values of the connections.

3.2.2 Model performance

Tools for evaluating the model performance

Let us introduce some tools to evaluate the quality of a model h parameterized by $\boldsymbol{\theta} \in \Theta$.

- Feature of probability measure, model, contrast and Risk function.

A *feature* of the distribution μ is a quantity of the form $\rho_{\mathcal{F}}(\mu) \in \mathcal{F}$ where $\mathcal{F}$ is called the *feature space* [5].

5. Voir la Définition (2.3.1) du Chapitre 2.

Notice that the feature space $\mathcal{F}$ can be either a scalar space (mean, threshold probability, etc...) or a functional space (density distribution, cumulative distribution function). The former case is part of the later one, identifying scalar to constant fuctions.
We equip the feature space $\mathcal{F}$ with the norm $||\cdot||_{\mathcal{F}}$ which can be a L_r-norm ($r \geq 1$) when $\mathcal{F}$ is a space of functions defined on $\mathcal{Y}$.
In all what follows, we denote by $\rho_h(\boldsymbol{\theta})$ a feature of the distribution of the random model output $h(\mathbf{X}, \boldsymbol{\theta})$.
We call **model** (feature space) a subset $F \subset \mathcal{F}$. In particular, we will deal with a model induced by h given by

$$F_{h,\boldsymbol{\theta}} = \{\rho_h(\boldsymbol{\theta}),\, \boldsymbol{\theta} \in \Theta\} \subset \mathcal{F}. \tag{3.1}$$

Definition 3.2.1. Contrast and risk function. [6]
A **contrast function** (with value in $L_1(Q)$) is any function

$$\begin{aligned} \Psi : \mathcal{F} &\longrightarrow L_1(Q) \\ \rho &\longmapsto \Psi(\rho, \cdot) : y \in \mathcal{Y} \longmapsto \Psi(\rho, y), \end{aligned} \tag{3.2}$$

such that

$$\rho^* = \underset{\rho \in \mathcal{F}}{\operatorname{Argmin}}\, \mathbb{E}_Y\, \Psi(\rho, Y)$$

is *unique*.
We call **risk function** the application

$$\forall\, \rho \in \mathcal{F}, \quad \mathcal{R}_\Psi(\rho) := \mathbb{E}_Y\, \Psi(\rho, Y)\,.$$

On the model $F_{h,\boldsymbol{\theta}} \subset \mathcal{F}$, we denote the risk by

$$\mathcal{R}_\Psi(h, \boldsymbol{\theta}) := \mathbb{E}_Y\, \Psi(\rho_h(\boldsymbol{\theta}),\, Y)\,. \tag{3.3}$$

Next, for a random variable ξ, we use the notation $\mathbb{E}_\xi$ for the expectation *w.r.t.* the variable ξ.

Example 3.2.1. Some classical features, associated contrasts and classical risk functions
- $\mathcal{F} = \mathbb{R}$ (constant functions) : we may consider $\rho(\mu) = \int u\, \mu(du) = \mathbb{E}_\mu(\xi)$ (mean), $\rho(\mu) = \int \mathbb{1}_{[s,+\infty[}(u)\mu(du) = \mu(\xi > s)$ (exceeding probability), etc...

  Mean-contrast

$$\Psi(\rho, y) = (y - \rho)^2, \quad \mathcal{R}_\Psi(h, \boldsymbol{\theta}) = (\mathbb{E}(Y) - \rho_h(\boldsymbol{\theta}))^2 + Var(Y)$$

- $\mathcal{F} = \{\textit{set of density functions}\}$

  log**-contrast**

$$\Psi(\rho, y) = -\log \rho(y), \quad \mathcal{R}_\Psi(h, \boldsymbol{\theta}) = KL(f, \rho_h(\boldsymbol{\theta})) - \mathbb{E}(\log(Y))\,,$$

  where $KL(g_1, g_2) = \int \log(\frac{g_1}{g_2})(y)\, g_1(y)\, dy$

  L_2-**contrast**

$$\Psi(\rho, y) = ||\rho||_2^2 - 2\,\rho(y), \quad \mathcal{R}_\Psi(h, \boldsymbol{\theta}) = \|\rho_h(\boldsymbol{\theta}) - f\|_2^2 - \|f\|_2^2\,.$$

6. Dans cet article, nous considérons uniquement les contrastes à valeurs dans $L_1(Q)$, autrement dit qui dépendent uniquement de la donnée y.

- etc...
In view of that examples, it make sense to investigate models h or/and parameters $\boldsymbol{\theta}$ providing small risk values. Here we restrict to parameters.

3.3 Inverse Problem.

Our goal is to compute a parameter $\boldsymbol{\theta} \in \Theta$ making the risk function $\mathcal{R}_\Psi(h, \boldsymbol{\theta})$ as small as possible.

- *Oracle.*

We want to estimate a parameter $\boldsymbol{\theta}^\star$ minimizing the risk (3.3), i.e

$$\boldsymbol{\theta}^\star \in \operatorname*{Argmin}_{\boldsymbol{\theta}\in\Theta} \mathcal{R}_\Psi(h, \boldsymbol{\theta})\,. \tag{3.4}$$

In the literature, the parameter $\boldsymbol{\theta}^\star$ is also called the *oracle*. This term was introduced by Donoho and Johnstone [4].
Notice that it may exist more than one parameter minimizing the risk $\mathcal{R}_\Psi(h, \boldsymbol{\theta})$. The minimal risk we can reach is $\mathcal{R}_\Psi(h, \boldsymbol{\theta}^\star)$, also called *ideal risk*.

However, the risk function $\mathcal{R}_\Psi(h, \boldsymbol{\theta})$ is uncomputable (hence $\boldsymbol{\theta}^\star$) for two reasons. First, the measure Q is unknown, and second, because we are dealing with complex models.

We aim at computing a parameter $\widehat{\boldsymbol{\theta}}$ that performs as well as the oracle $\boldsymbol{\theta}^\star$, that is

$$\mathcal{R}_\Psi(h, \widehat{\boldsymbol{\theta}}) \approx \mathcal{R}_\Psi(h, \boldsymbol{\theta}^\star)\,.$$

In what follows, we establish a risk bound of the form

$$\mathcal{R}_\Psi(h, \widehat{\boldsymbol{\theta}}) \leq C\, \mathcal{R}_\Psi(h, \boldsymbol{\theta}^\star) + \Delta\,.$$

We propose the following estimation procedure to built $\widehat{\boldsymbol{\theta}}$.

As Q is unknown, we replace it by its empirical version

$$Q_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{Y_i}$$

based on $Y_1, ..., Y_n$. The approximation of the risk becomes

$$\frac{1}{n}\sum_{i=1}^{n} \Psi\left(\rho_h(\boldsymbol{\theta})\,,\, Y_i\right)\,.$$

Then, it remains the feature $\rho_h(\boldsymbol{\theta})$ which is supposed analytically intractable (for each $\boldsymbol{\theta}$). We propose to estimate the feature as follows.

- *Plug-in estimator.*

We denote by $\rho_h^m(\boldsymbol{\theta})$ a *plug-in* estimator of $\rho_h(\boldsymbol{\theta})$ based on $h(\mathbf{X}'_1, \boldsymbol{\theta}), ..., h(\mathbf{X}'_m, \boldsymbol{\theta})$. We suppose that $\rho_h^m(\boldsymbol{\theta})$ takes the following form

$$\rho_h^m(\boldsymbol{\theta}) := \frac{1}{m} \sum_{j=1}^{m} \widetilde{\rho}(h(\mathbf{X}'_j, \boldsymbol{\theta})) \tag{3.5}$$

where $\frac{1}{m}\widetilde{\rho} : \mathcal{Y} \to \mathcal{F}$ is a *weight function* depending on the contrast Ψ considered.
For simplicity, we may also call $\widetilde{\rho}$ weight function.

Example 3.3.1. Examples of weight functions.
- *mean-contrast*

$$\frac{1}{m}\widetilde{\rho}(y) = \frac{y}{m}$$

- log-*contrast or L_2-contrast density estimation*

$$\frac{1}{m}\widetilde{\rho}(y)(\cdot) = \frac{1}{m} K_b(\cdot - y)$$

where $K_b(\cdot - y) = \frac{1}{b}K(\frac{\cdot - y}{b})$ for a kernel $K()$ and a bandwidth b (See Figure (3.2) for an illustration). It means that the method of estimation relies on kernel density estimation.

Another choice is to use an expansion on a given (truncated) L_2-basis, $(\varphi_j, 0 \leq j \leq L)$, which leads to the weight function

$$\frac{1}{m}\widetilde{\rho}(y)(\cdot) = \frac{1}{m} \sum_{j=1}^{L} \varphi_j(\cdot)\varphi_j(y)$$

Remark 3.3.1. The weight function $\frac{1}{m}\widetilde{\rho}(y)$ evaluated at $y \in \mathcal{Y}$ can be either a scalar value ($\frac{1}{m}$ for the mean) or a function (for the density).
So that without loss of generality, one can see the weight function $\frac{1}{m}\widetilde{\rho}(y)$ at a point $y \in \mathcal{Y}$ as a function,

$$\widetilde{\rho}(y) \, : \, \lambda \in \mathcal{Y} \longmapsto \widetilde{\rho}(y)(\lambda) \, .$$

For instance, in the case where $\frac{1}{m}\widetilde{\rho}(y) = \frac{y}{m}$, the function $\widetilde{\rho}(y)(\lambda)$ is constant in λ.

In the sequel the examples of density estimation will be carried out from the kernel method. We chose this method because it is simple to write and so very popular in the uncertainty management in industrial context. Notice that we will assume some adaptivity of our kernel estimator by choosing a bandwidth $b = b_m$ that will depend on m.

Definition 3.3.1. We denote by $\sigma_h^m(\boldsymbol{\theta})$, called *simulation error*, the error committed estimating the feature $\rho_h(\boldsymbol{\theta})$ by the estimator $\rho_h^m(\boldsymbol{\theta})$,

$$\sigma_h^m(\boldsymbol{\theta}) := ||\rho_h^m(\boldsymbol{\theta}) - \rho_h(\boldsymbol{\theta})||_{\mathcal{F}} \, .$$

By triangular inequality and the fact that $\mathbb{E}_{\mathbf{X}}\widetilde{\rho}(h(\mathbf{X}, \boldsymbol{\theta})) = \mathbb{E}_{\mathbf{X}_{1..m}}\rho_h^m(\boldsymbol{\theta})$, it holds

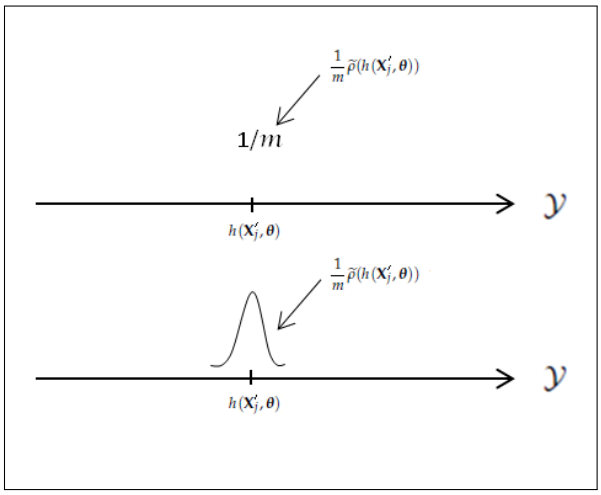


FIGURE 3.2 – Example of weight function in the case of the mean (top) and the case of the density (bottom).

$$
\begin{aligned}
\sigma_h^m(\boldsymbol{\theta}) &= ||\rho_h^m(\boldsymbol{\theta}) - \rho_h(\boldsymbol{\theta})||_{\mathcal{F}} \\
&= ||\rho_h^m(\boldsymbol{\theta}) - \mathbb{E}_{\mathbf{X}_{1..m}}\rho_h^m(\boldsymbol{\theta}) + \mathbb{E}_{\mathbf{X}_{1..m}}\rho_h^m(\boldsymbol{\theta}) - \rho_h(\boldsymbol{\theta})||_{\mathcal{F}} \\
&= ||\rho_h^m(\boldsymbol{\theta}) - \mathbb{E}_{\mathbf{X}}\widetilde{\rho}(h(\mathbf{X},\boldsymbol{\theta})) + \mathbb{E}_{\mathbf{X}}\widetilde{\rho}(h(\mathbf{X},\boldsymbol{\theta})) - \rho_h(\boldsymbol{\theta})||_{\mathcal{F}} \\
&\leq ||\rho_h^m(\boldsymbol{\theta}) - \mathbb{E}_{\mathbf{X}}\widetilde{\rho}(h(\mathbf{X},\boldsymbol{\theta}))||_{\mathcal{F}} + ||\mathbb{E}_{\mathbf{X}}\widetilde{\rho}(h(\mathbf{X},\boldsymbol{\theta})) - \rho_h(\boldsymbol{\theta})||_{\mathcal{F}} \\
&= \left|\left|\frac{1}{m}\sum_{j=1}^{m}\left[\widetilde{\rho}(h(\mathbf{X}'_j,\boldsymbol{\theta})) - \mathbb{E}_{\mathbf{X}}\widetilde{\rho}(h(\mathbf{X},\boldsymbol{\theta}))\right]\right|\right|_{\mathcal{F}} + B_h^m(\boldsymbol{\theta}) \qquad (3.6)
\end{aligned}
$$

with

$$
B_h^m(\boldsymbol{\theta}) := ||\mathbb{E}_{\mathbf{X}}\widetilde{\rho}(h(\mathbf{X},\boldsymbol{\theta})) - \rho_h(\boldsymbol{\theta})||_{\mathcal{F}} \qquad (3.7)
$$

the *bias error*.

The first term in the right hand side of inequality (3.6) is a *variance* (random) term, and the second is a *bias* (deterministic) term.

Assumption 3.3.1. We assume that the plug-in estimator $\rho_h^m(\boldsymbol{\theta})$ (4.2) is uniformly asymptotically unbiased, i.e it exists some constant $B_h(m)$ depending on h and m such that the bias error (3.7) satisfies

$$
\sup_{\boldsymbol{\theta}\in\Theta} B_h^m(\boldsymbol{\theta}) < B_h(m) < \infty\,,
$$

and $B_h(m) \to 0$ with m.

Finally, the criterion we propose to minimize has the form

$$
\frac{1}{n}\sum_{i=1}^{n}\Psi\left(\frac{1}{m}\sum_{j=1}^{m}\widetilde{\rho}(h(\mathbf{X}'_j,\boldsymbol{\theta}))\,,\,Y_i\right)\,,
$$

which provides the estimator

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\Theta}{\text{Argmin}} \frac{1}{n}\sum_{i=1}^{n} \Psi\left(\frac{1}{m}\sum_{j=1}^{m}\widetilde{\rho}(h(\mathbf{X}'_j,\boldsymbol{\theta}))\,,\, Y_i\right)\,, \tag{3.8}$$

or

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\Theta}{\text{Argmin}} \sum_{i=1}^{n} \Psi\left(\frac{1}{m}\sum_{j=1}^{m}\widetilde{\rho}(h(\mathbf{X}'_j,\boldsymbol{\theta}))\,,\, Y_i\right)\,.$$

In the various cases we mentioned it gives $\widehat{\boldsymbol{\theta}}_M = \text{Argmin}_{\boldsymbol{\theta}\in\Theta} \sum_{i=1}^{n}\left(\sum_{j=1}^{m}\left(Y_i - h(\mathbf{X}'_j,\boldsymbol{\theta})\right)\right)^2$ for the *mean-contrast*, $\widehat{\boldsymbol{\theta}}_{\log} = \text{Argmin}_{\boldsymbol{\theta}\in\Theta} -\sum_{i=1}^{n}\log\left(\sum_{j=1}^{m} K_b(Y_i - h(\mathbf{X}'_j,\boldsymbol{\theta}))\right)$ for the log-*contrast*, $\widehat{\boldsymbol{\theta}}_{L_2} = \text{Argmin}_{\boldsymbol{\theta}\in\Theta}\left\{\left\|\sum_{j=1}^{m} K_b(\,\cdot - h(\mathbf{X}'_j,\boldsymbol{\theta}))\right\|_2^2 - \frac{2}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} K_b(\,Y_i - h(\mathbf{X}'_j,\boldsymbol{\theta}))\right\}$ for the L_2-*contrast*.

Remark 3.3.2. 1. The estimator $\widehat{\boldsymbol{\theta}}$ depends on the model h, the number of experimental data n and the number of simulation data m .
2. The number of simulations m have to be thought greater than n (number of experimental data). It appears natural to think that experimental data are difficult to obtain whereas simulated data are more reachable.

We recall that the issue is the statistical properties of this procedure taking into account the two kinds of data : experimental and simulated data, which is non classical in statistics.

Once we define the procedure for computing $\widehat{\boldsymbol{\theta}}$, we have to qualify the *quality* of this procedure.
It's the topic of the following section.

3.4 Main Result

In this section, we aim at establishing a risk bound which provides a qualification of the estimation procedure previously defined.
We recall that

$$\mathcal{R}_\Psi(h,\boldsymbol{\theta}) = \mathbb{E}_Y\,\Psi\left(\rho_h(\boldsymbol{\theta}), Y\right)\,,$$

$$\boldsymbol{\theta}^\star \in \underset{\boldsymbol{\theta}\in\Theta}{\text{Argmin}}\, \mathcal{R}_\Psi(h,\boldsymbol{\theta})\,,$$

and

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\Theta}{\text{Argmin}} \frac{1}{n}\sum_{i=1}^{n} \Psi\left(\frac{1}{m}\sum_{j=1}^{m}\widetilde{\rho}(h(\mathbf{X}'_j,\boldsymbol{\theta}))\,,\, Y_i\right)\,.$$

Now, we give some definitions and notations useful for setting the Theorem 3.4.1.

Denote by

$$\mathbb{G}_n = \sqrt{n}(Q_n - Q)$$

and

$$\mathbb{K}_m^{\mathbf{x}} = \sqrt{m}(P_m^{\mathbf{x}} - P^{\mathbf{x}})\,,$$

the Q-empirical process (based on $Y_1, ..., Y_n$) and the $P^{\mathbf{x}}$-empirical process (based on $\mathbf{X}'_1, ..., \mathbf{X}'_m$), respectively.
Let the classes of functions

$$\mathcal{W}_{(\widetilde{\rho},\Psi)} = \{y \in \mathcal{Y} \mapsto \Psi(\widetilde{\rho}(\lambda)\,,\, y)\,,\, \lambda \in \mathcal{Y}\}\,, \tag{3.9}$$

$$\mathcal{P}_{(\widetilde{\rho},h)} = \{\mathbf{x} \in \mathcal{X} \mapsto \widetilde{\rho}(h(\mathbf{x},\boldsymbol{\theta}))(\lambda)\,,\, (\boldsymbol{\theta},\lambda) \in \Theta \times \mathcal{Y}\}\,. \tag{3.10}$$

	$\mathcal{W}_{(\widetilde{\rho},\Psi)}$	$\mathcal{P}_{(\widetilde{\rho},h)}$	A_Ψ
mean-contrast	$y \mapsto (y-\lambda)^2$, $\lambda \in \mathcal{Y}$	$\mathbf{x} \mapsto h(\mathbf{x},\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$	$4\,M$
log-contrast	$y \mapsto -\log\left(K_b(y-\lambda)\right)$, $\lambda \in \mathcal{Y}$	$\mathbf{x} \mapsto K_b(\lambda - h(\mathbf{x},\boldsymbol{\theta}))$, $(\lambda,\boldsymbol{\theta}) \in \Theta \times \mathcal{Y}$	$\Vert f\Vert_2/\eta$
L_2-contrast	$y \mapsto \Vert K_b(\,\cdot - \lambda)\Vert_2 - 2\,K_b(y-\lambda)$, $\lambda \in \mathcal{Y}$	*idem*	$2\,(\Vert f\Vert_2 + B)$

TABLE 3.1 – Example of classes of functions and constant A_Ψ (see Section 3.6.1).

Next, we use the following notation : let P be some measure and $\mathcal{G}$ a class of real valued functions. We denote by

$$P\,g := \int g(u)P(du) \qquad g \in \mathcal{G}$$

and

$$\Vert P\Vert_{\mathcal{G}} := \sup_{g\in\mathcal{G}} |P\,g|\,.$$

With this notation, for a class of functions $\mathcal{G}_{\mathcal{Y}} : \mathcal{Y} \to \mathbb{R}$ we have

$$\begin{aligned}
\mathbb{G}_n\,g &= \int_{\mathcal{Y}} g(u)\mathbb{G}_n(du) \\
&= \sqrt{n}\int_{\mathcal{Y}} g(u)(Q_n - Q)(du) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(g(Y_i) - \mathbb{E}(g(Y))\right)\,.
\end{aligned}$$

Also, for a class of functions $\mathcal{G}_{\mathcal{X}} : \mathcal{X} \to \mathbb{R}$

$$\mathbb{K}^{\mathbf{x}}_m\,g = \frac{1}{\sqrt{m}}\sum_{j=1}^{m}\Big(g(\mathbf{X}'_j) - \mathbb{E}(g(\mathbf{X}))\Big)\,.$$

Remark 3.4.1. The quantities $\Vert\mathbb{G}_n\Vert_{\mathcal{G}_{\mathcal{Y}}}$ and $\Vert\mathbb{K}^{\mathbf{x}}_m\Vert_{\mathcal{G}_{\mathcal{X}}}$ are nonnegative real valued random variables.

In our applications, the class of functions $\mathcal{G}_{\mathcal{Y}}$ is $\mathcal{W}_{(\widetilde{\rho},\Psi)}$ and $\mathcal{G}_{\mathcal{X}}$ is $\mathcal{P}_{(\widetilde{\rho},h)}$, respectively defined in (3.9) and (3.10).

Definition 3.4.1. Tightness.
Let $(\xi_l)_{l\geq 1}$ be a sequence of real valued random variables defined on the probability space

$(\Omega, \mathcal{A}, \mathbb{P})$.
This sequence is tight if for all $\varepsilon > 0$, it exists some compact $\mathcal{K}^\varepsilon \subset \mathbb{R}$ such that

$$\forall\, l \geq 1\,, \quad \mathbb{P}\left(\xi_l \in \mathcal{K}^\varepsilon\right) \geq 1 - \varepsilon\,.$$

In particular, if the ξ_l are nonnegative, the sequence is tight if for all $\varepsilon > 0$ it exists some constant $\bar{K}^\varepsilon \geq 0$ such that

$$\forall\, l \geq 1\,, \quad \mathbb{P}\left(\xi_l \leq \bar{K}^\varepsilon\right) \geq 1 - \varepsilon\,.$$

We make the following assumptions.

Assumption 3.4.1. Let $\mathcal{W}^m_{(\widetilde{\rho},\Psi)}$ be the class of functions

$$\mathcal{W}^m_{(\widetilde{\rho},\Psi)} = \{y \in \mathcal{Y} \mapsto \Psi(\frac{1}{m}\sum_{j=1}^{m}\widetilde{\rho}(\lambda_j)\,,\, y)\,,\, (\lambda_j)_{1\leq j\leq m} \in \mathcal{Y}^m\}\,.$$

We assume that it exists some universal constant $\gamma > 0$ such that

$$\|\mathbb{G}_n\|_{\mathcal{W}^m_{(\widetilde{\rho},\Psi)}} \leq \gamma\, \|\mathbb{G}_n\|_{\mathcal{W}_{(\widetilde{\rho},\Psi)}}\,,$$

where $\mathcal{W}_{(\widetilde{\rho},\Psi)}$ is given in (3.9).

This assumption may be explained by the fact that the "complexity" of the class of functions $\mathcal{W}^m_{(\widetilde{\rho},\Psi)}$ is "close" to the complexity of $\mathcal{W}_{(\widetilde{\rho},\Psi)}$ which can be viewed as $\mathcal{W}_{(\widetilde{\rho},\Psi)} = \mathcal{W}^{m=1}_{(\widetilde{\rho},\Psi)}$. In other words, we assume that the summation of functions $\widetilde{\rho}(\lambda_j)$ hasn't a significant impact on the behavior of $\|\mathbb{G}_n\|_{\mathcal{W}^m_{(\widetilde{\rho},\Psi)}}$.

Assumption 3.4.2. Let us assume that the contrast Ψ satisfies

- for all $\rho_1, \rho_2 \in \mathcal{F}$
$$\mathbb{E}_Y\left|\Psi(\rho_1, Y) - \Psi(\rho_2, Y)\right| \leq A_\Psi\, ||\rho_1 - \rho_2||_{\mathcal{F}}$$
with a constant $A_\Psi < \infty$ independent of ρ_1, ρ_2.

The contrasts given in Example 3.2.1 satisfy Assumption 3.4.2 under right conditions on the distribution of Y.

Theorem 3.4.1. ***Risk bound for Parameter Estimation.***
Under the Assumptions (3.4.1), (3.4.2) and (3.3.1), suppose that the sequences of random variables $||\mathbb{G}_n||_{\mathcal{W}_{(\widetilde{\rho},\Psi)}}$ *and* $||\mathbb{K}^{\mathbf{x}}_m||_{\mathcal{P}_{(\widetilde{\rho},h)}}$ *are tight. Denote by* $\bar{K}^\varepsilon_{(\widetilde{\rho},\Psi)}$ *and* $\bar{K}^\varepsilon_{(\widetilde{\rho},h)}$ *the associated constants, uniform (or decreasing) in n and m, respectively.*
Let the feature space $\mathcal{F}$ *be equipped with either the absolute value norm, or some* L_r *norm.*
Then, for all $\varepsilon > 0$*, with probability at least* $1 - 2\varepsilon$ *it holds*

$$\mathcal{R}_\Psi(h, \widehat{\boldsymbol{\theta}}) \leq \inf_{\boldsymbol{\theta}\in\Theta}\left(\mathcal{R}_\Psi(h, \boldsymbol{\theta})\right) + \frac{K^\varepsilon_{(\widetilde{\rho},\Psi)}}{\sqrt{n}}\left(1 + \sqrt{\frac{n}{m}}(K^\varepsilon_{(\widetilde{\rho},h)} + B_m)\right)$$

where the constants $K^\varepsilon_{(\widetilde{\rho},\Psi)}$, $K^\varepsilon_{(\widetilde{\rho},h)}$ *depend on* $\bar{K}^\varepsilon_{(\widetilde{\rho},\Psi)}$, $\bar{K}^\varepsilon_{(\widetilde{\rho},h)}$, A_Ψ, M *and* r. B_m *is a bias factor depending on* $B_h(m)$.

3.5 Some comments

It is of interest to compare the methodology we develop with the classical framework where the feature $\rho_h(\boldsymbol{\theta})$ of the random model output $h(\mathbf{X}, \boldsymbol{\theta})$ is analytically tractable. In this case, the estimation procedure (3.8) is classically

$$\widehat{\boldsymbol{\theta}}_n = \underset{\boldsymbol{\theta}\in\Theta}{\text{Argmin}} \, \frac{1}{n}\sum_{i=1}^{n} \Psi\left(\rho_h(\boldsymbol{\theta}), Y_i\right),$$

and we can derive immediately a risk bound.

Proposition 3.5.1. ***Basic risk bound.***
It holds that

$$\mathcal{R}_\Psi(h, \widehat{\boldsymbol{\theta}}_n) \leq \inf_{\boldsymbol{\theta}\in\Theta} \left(\mathcal{R}_\Psi(h, \boldsymbol{\theta})\right) + \frac{2}{\sqrt{n}} \|\mathbb{G}_n\|_{\widetilde{\mathcal{W}}_\Psi}, \tag{3.11}$$

where

$$\widetilde{\mathcal{W}}_\Psi = \left\{ y \in \mathcal{Y} \mapsto \Psi(\rho_h(\boldsymbol{\theta}), y), \, \boldsymbol{\theta} \in \Theta \right\}.$$

Proof. The proof comes from a classical calculus in M-estimation, see for example [23] (p. 46) . □

Most of statistical procedures, as likelihood, regression, classification etc... can be written like (3.11). Such procedures have been widely studied with a large literature available. Recently, authors use the Empirical Processes theory (see [22, 23, 24, 13] among others) to derive limit theorems. Indeed, the asymptotic (and non-asymptotic) properties of the estimator $\widehat{\boldsymbol{\theta}}_n$ can be given from the behavior of the residual term $\frac{2}{\sqrt{n}}\|\mathbb{G}_n\|_{\widetilde{\mathcal{W}}_\Psi}$. In particular, for *identification* problem (i.e $\boldsymbol{\theta}^\star$ is unique), consistency and rate of convergence are derived from the fluctuations of the random variable $\|\mathbb{G}_n\|_{\widetilde{\mathcal{W}}_\Psi}$, see for example [22].
Suppose for a moment that it exists some constant (uniform in n) such that with high probability

$$\|\mathbb{G}_n\|_{\widetilde{\mathcal{W}}_\Psi} \leq \frac{K}{2},$$

then by inequality (3.11), with high probability

$$\mathcal{R}_\Psi(h, \widehat{\boldsymbol{\theta}}_n) \leq \inf_{\boldsymbol{\theta}\in\Theta} \left(\mathcal{R}_\Psi(h, \boldsymbol{\theta})\right) + \frac{K}{\sqrt{n}}. \tag{3.12}$$

Thus, depending on whether the constant K is sharp or not, one can bound properly the estimation error. To compute such (sharp) constant K is difficult in general, we can refer to [14, 21, 24, 15].

Inequality (3.11) can not be applied to our framework because the induced procedure $\widehat{\boldsymbol{\theta}}_n$ involves the quantity $\rho_h(\boldsymbol{\theta})$ which is untractable for *complex models*.
The result of Theorem 3.4.1 is non-asymptotic, i.e valid for all $n \geq 1$ and $m \geq 1$ under mentioned assumptions. The fundamental point of this theorem is the "*concentration of the measure phenomenon*" (Ledoux [14], Billingsley [2]). It derives from our assumptions, more precisely, when we supposed the tightness of the sequences of the random variables $\|\mathbb{G}_n\|_{\mathcal{W}_{(\tilde{\rho},\Psi)}}$ ($Y_{1..n}$-dependent) and $\|\mathbb{K}_m^{\mathbf{x}}\|_{\mathcal{P}_{(\tilde{\rho},h)}}$ ($\mathbf{X}_{1..m}$-dependent). Moreover, we insist on the fact that the

constants $\bar{K}^{\varepsilon}_{(\widetilde{\rho},\Psi)}$ (that bounds $\|\mathbb{G}_n\|_{\mathcal{W}_{(\widetilde{\rho},\Psi)}}$) and $\bar{K}^{\varepsilon}_{(\widetilde{\rho},h)}$ (that bounds $\|\mathbb{K}^{\mathbf{x}}_m\|_{\mathcal{P}_{(\widetilde{\rho},h)}}$) are uniform (or decreasing) in n and m, respectively. The advantage of this uniformity is the explicit expression of the *residual* term

$$\frac{K^{\varepsilon}_{(\widetilde{\rho},\Psi)}}{\sqrt{n}}\left(1+\sqrt{\frac{n}{m}}(K^{\varepsilon}_{(\widetilde{\rho},h)}+B_m)\right) \tag{3.13}$$

depending on the data (n and m) on one hand, and on the constants $\bar{K}^{\varepsilon}_{(\widetilde{\rho},\Psi)}$, $\bar{K}^{\varepsilon}_{(\widetilde{\rho},h)}$ and B_m on the other hand. However, although the existence of such constants are proved or supposed, their computation is more tedious. Indeed, we need results about tail bounds for Gaussian and Empirical Processes. We will discuss in Section 3.6.3 how to compute properly such constants using concentration inequalities. Let us assume for a moment the existence of these constants.
We showed that the estimation procedure $\widehat{\boldsymbol{\theta}}$ defined in (3.8) "mimic" the ideal risk $\mathcal{R}_\Psi(h,\boldsymbol{\theta}^\star) = \inf_{\boldsymbol{\theta}\in\Theta}(\mathcal{R}_\Psi(h,\boldsymbol{\theta}))$ up to the residual term (3.13). Making $m \longrightarrow +\infty$, this residual becomes simply $\frac{K^{\varepsilon}_{(\widetilde{\rho},\Psi)}}{\sqrt{n}}$ which has the same form as those found in classical cases (3.12). We find the usual rate of convergence $\sqrt{n}$.
In our purpose, the factor

$$\left(1+\sqrt{\frac{n}{m}}(K^{\varepsilon}_{(\widetilde{\rho},h)}+B_m)\right) > 1$$

we call *simulation factor*, is due to the simulations used to estimate the feature $\rho_h(\boldsymbol{\theta})$ of the random output $h(\mathbf{X},\boldsymbol{\theta})$ by a plug-in estimator $\rho^m_h(\boldsymbol{\theta})$ we defined in (4.2).
It appears that for fixed n, one should have a number of simulation data m greater than n.

Remark 3.5.1. The term $\inf_{\boldsymbol{\theta}\in\Theta}(\mathcal{R}_\Psi(h,\boldsymbol{\theta}))$ in Theorem 3.4.1 appears as the best (smallest) error one can make. This kind of error is commonly called *approximation error* or *systematic error*. It can be understood as the "distance" between the *a priori* knowledge one has with the observed phenomenon.

The *ideal risk* is $\inf_{\boldsymbol{\theta}\in\Theta}(\mathcal{R}_\Psi(h,\boldsymbol{\theta}))$. It represents a "distance" between the "target" and the "best" information available, see Remark 3.5.1. Let's consider the case of density estimation. If this term is supposed equal to zero, it means that we believe that the density f belongs to the family of densities $\{\rho_h(\boldsymbol{\theta}),\ \boldsymbol{\theta}\in\Theta\}$. In this case we obtain for example (L_2-contrast)

$$\|\rho_h(\widehat{\boldsymbol{\theta}}_{L_2}) - f\|_2^2 \leq \frac{K^{\varepsilon}_{(\widetilde{\rho},L_2)}}{\sqrt{n}}\left(1+\sqrt{\frac{n}{m}}(K^{\varepsilon}_{(\widetilde{\rho},h)}+B_m)\right).$$

However, such *a priori* has to be made with precautions. It is necessary to verify that the model h is able to reach a sufficiently large range of distributions, that one we believe f belongs to.

3.6 About the constants in Theorem 3.4.1

3.6.1 Constant A_Ψ

We will show how we obtain the constants A_Ψ in Table (3.1). Let us recall that $\mathcal{Y}\in[-M,M]$.

- *mean-contrast.*
Let $y \in \mathcal{Y}$, $\rho_1, \rho_2 \in \mathcal{F} \subset \mathcal{Y}$. We have

$$\begin{aligned} \left|(y-\rho_1)^2 - (y-\rho_2)^2\right| &= |\rho_1 - \rho_2|\,|2y - (\rho_1+\rho_2)| \\ &\leq |\rho_1 - \rho_2|\,4M\,. \end{aligned}$$

- log-*contrast.*
Let $y \in \mathcal{Y}$, $\rho_1, \rho_2 \in \mathcal{F}$, with $\mathcal{F}$ some set of density functions.
Moreover, suppose that it exists some $\eta > 0$ such that

$$\forall\, \rho \in \mathcal{F} \quad \rho > \eta\,.$$

By Taylor Lagrange formula, it exists some $\tau \in (\rho_1(y), \rho_2(y))$ such that

$$\begin{aligned} |\log(\rho_1(y)) - \log(\rho_2(y))| &= \frac{1}{\tau}\,|\rho_1(y) - \rho_2(y)| \\ &\leq \frac{1}{\eta}\,|\rho_1(y) - \rho_2(y)| \end{aligned}$$

since $\rho > \eta$ for all $\rho \in \mathcal{F}$ and $\tau > \eta$.
Taking the expectation under the measure Q (with Lebesgue density f) involves the quantity $\mathbb{E}_Y(|\rho_1(Y) - \rho_2(Y)|)$ in the right member. By Cauchy-Schwarz inequality

$$\mathbb{E}_Y(|\rho_1(Y) - \rho_2(Y)|) \leq \|\rho_1 - \rho_2\|_2\,\|f\|_2\,,$$

so

$$\mathbb{E}_Y\,|\log(\rho_1(Y)) - \log(\rho_2(Y))| \leq \frac{\|f\|_2}{\eta}\,\|\rho_1 - \rho_2\|_2\,.$$

- L_2-*contrast.*
Let $y \in \mathcal{Y}$, $\rho_1, \rho_2 \in \mathcal{F}$, with $\mathcal{F}$ be some set of density functions.
Suppose that it exists some $B > 0$ such that

$$\sup_{\rho \in \mathcal{F}} \|\rho\|_2 < B\,.$$

By triangular inequality

$$\begin{aligned} \left|(\|\rho_1\|_2^2 - 2\rho_1(y)) - (\|\rho_2\|_2^2 - 2\rho_2(y))\right| &\leq \left|\|\rho_1\|_2^2 - \|\rho_2\|_2^2\right| + 2\,|\rho_2(y) - \rho_2(y)| \\ &\leq \|\rho_1 - \rho_2\|_2^2 + 2\,|\rho_2(y) - \rho_2(y)|\,. \end{aligned}$$

Taking the expectation under Q and by Cauchy-Schwarz inequality (as before) yields

$$\begin{aligned} \mathbb{E}_Y\left|(\|\rho_1\|_2^2 - 2\rho_1(Y)) - (\|\rho_2\|_2^2 - 2\rho_2(Y))\right| &\leq \|\rho_1 - \rho_2\|_2^2 + 2\,\|\rho_1 - \rho_2\|_2\,\|f\|_2 \\ &\leq \|\rho_1 - \rho_2\|_2\,(\|\rho_1 - \rho_2\|_2 + 2\,\|f\|_2) \\ &\leq 2\,(B + \|f\|_2)\,\|\rho_1 - \rho_2\|_2\,. \end{aligned}$$

The two previous conditions on the densities (uniformly lower bounded or upper bounded in L_2) are restrictive. Yet many densities with a fixed compact support belong to one or the other set, which allows to chose between the two contrasts. Of course it needs an a priori information, which is not always available or true.

3.6.2 Constant $B_h(m)$

When the *plug-in* estimator $\rho_h^m(\boldsymbol{\theta})$ is unbiased, the bias term $B_h^m(\boldsymbol{\theta})$ defined in (3.7) is zero for all $\boldsymbol{\theta} \in \Theta$ and $m > 0$, hence $B_h(m) = 0$ too. This is not the case for the kernel density estimation.

We study the example of the kernel estimator, *i.e.* when the weight function $\widetilde{\rho}$ is a function of the form

$$\widetilde{\rho}(y)(\cdot) = K_b(\cdot - y)$$

where $K_b(\cdot - y) = \frac{1}{b}K(\frac{\cdot - y}{b})$ for some kernel $K()$ and some bandwidth b.
Consider that $\|\cdot\|_{\mathcal{F}} = \|\cdot\|_2$, then for all $\boldsymbol{\theta} \in \Theta$ we have

$$\begin{aligned} B_h^m(\boldsymbol{\theta}) &= \|\mathbb{E}_{\mathbf{X}}(K_b(\cdot - h(\mathbf{X}, \boldsymbol{\theta}))) - \rho_h(\boldsymbol{\theta})\|_2 \\ &= \left(\int_{\mathcal{Y}} \left(\int_{\mathcal{X}} \left(K_b(y - h(x, \boldsymbol{\theta})) - \rho_h(\boldsymbol{\theta}) \right) P^{\mathbf{x}}(dx) \right)^2 dy \right)^{1/2} . \end{aligned}$$

Theorem 24.1 in [23] (p. 345) states that : Let $\xi_1, ..., \xi_m \in \mathcal{Y}$ be an i.i.d sample drawn from a probability density function g and $K : \mathcal{Y} \to \mathbb{R}^+$ some function (kernel). Denote by

$$\widehat{g}(y) = \frac{1}{m} \sum_{j=1}^{m} \frac{1}{b} K\left(\frac{y - \xi_m}{b} \right) .$$

Assume that : $\|g''\|_2 < +\infty$, $\int y\, K(y)\, dy = 0$, $I = \int y^2\, K(y)\, dy < +\infty$, then it exists a constant C_g such that for all $b > 0$

$$\mathbb{E}_{\xi_{1..m}} \|\widehat{g} - g\|_2^2 \leq C_g \left(\frac{1}{m\, b} + b^4 \right) .$$

In particular, the bias term $\|\mathbb{E}_{\xi_{1..m}} \widehat{g} - g\|_2$ is bounded above by

$$\frac{I\, \|g''\|_2}{\sqrt{3}}\, b^2 .$$

Applying with $g = \rho_h(\boldsymbol{\theta})$ satisfying the assumptions of this Theorem, we obtain

$$B_h^m(\boldsymbol{\theta}) \leq \frac{I\, \|\rho_h''(\boldsymbol{\theta})\|_2}{\sqrt{3}}\, b^2 .$$

If $\sup_{\boldsymbol{\theta} \in \Theta} \|\rho_h''(\boldsymbol{\theta})\|_2$ is finite, it justifies the existence of $B_h(m) = \sup_{\boldsymbol{\theta} \in \Theta} B_h^m(\boldsymbol{\theta})$.

3.6.3 Constants $\bar{K}^{\varepsilon}_{(\widetilde{\rho},\Psi)}$ and $\bar{K}^{\varepsilon}_{(\widetilde{\rho},h)}$

We detail the arguments for computing the constants $\bar{K}^{\varepsilon}_{(\widetilde{\rho},\Psi)}$ and $\bar{K}^{\varepsilon}_{(\widetilde{\rho},h)}$. Since these constants are tightness constants relative to some empirical processes (see the assumptions of Theorem 3.4.1), we will give arguments with a generic empirical process $\mathbb{W}_p = \sqrt{p}(W_p - W)$ indexed by a generic class of functions $\mathcal{G}$.
Now, the goal is to compute some constant $K(\varepsilon)$ such that

$$\mathbb{P}(\|\mathbb{W}_p\|_{\mathcal{G}} \leq K(\varepsilon)) \geq 1 - \varepsilon \quad \text{for small } \varepsilon > 0 . \tag{3.14}$$

For this, we propose to use the work of T. Klein and E. Rio [12], in particular Theorem 1.1, which deals with right hand side deviations of the empirical process. They show that for an empirical process $\mathbb{W}_p$ indexed by a **countable** class of functions $\mathcal{G}$ with values in $[-1,1]$

$$\mathbb{P}\left(\sup_{g\in\mathcal{G}}\mathbb{W}_p(g)\geq\mathbb{E}(\sup_{g\in\mathcal{G}}\mathbb{W}_p(g))+t\right)\leq exp\left(-\frac{t^2}{2v+3\,t/\sqrt{p}}\right), \tag{3.15}$$

for all positive t and some constant v. They also give left hand side deviations.
In our purpose, we don't really work with $\sup_{g\in\mathcal{G}}\mathbb{W}_p(g)$ but rather with $\sup_{g\in\mathcal{G}}|\mathbb{W}_p(g)| = \|\mathbb{W}_p\|_{\mathcal{G}}$ corresponding to a two-side control. Hence, according to the work of T. Klein and E. Rio [12], it exists some function $\varphi_{\mathcal{G}}:\mathbb{R}_+\to[0,1]$ decreasing to zero such that for all positive t

$$\mathbb{P}\left(\|\mathbb{W}_p\|_{\mathcal{G}}\geq\mathbb{E}(\|\mathbb{W}_p\|_{\mathcal{G}})+t\right)\leq\varphi_{\mathcal{G}}(t)\,. \tag{3.16}$$

Another point is missing before we apply this result in our context, it is the fact that the result is valid for countable classes of functions, and so, we need to extend the Theorem 1.1 in [12]. We prove the following proposition.

Proposition 3.6.1. *Let $\mathbb{W}_p$ be an empirical process indexed by a class of functions $\mathcal{G}$ taking values in $[-1,1]$ and parameterized by a* **compact** *set C of $\mathbb{R}^l$, $l\geq 1$. Suppose that the application*

$$\lambda\in\mathcal{C}\longmapsto g_\lambda\in\mathcal{G}\subset L_2 \tag{3.17}$$

is continuous.
Then, it exists a function $\varphi_{\mathcal{G}}$ decreasing to zero (given by [12]) such that for all $t\geq 0$

$$\mathbb{P}\left(\|\mathbb{W}_p\|_{\mathcal{G}}\geq\mathbb{E}(\|\mathbb{W}_p\|_{\mathcal{G}})+t\right)\leq\varphi_{\mathcal{G}}(t)\,. \tag{3.18}$$

Proof. For simplicity, we prove the proposition with $\mathcal{G}=\mathcal{W}_{(\widetilde{\rho},\Psi)}$ where

$$\mathcal{W}_{(\widetilde{\rho},\Psi)}=\{y\in\mathcal{Y}\mapsto\Psi(\widetilde{\rho}(\lambda)\,,\,y)\,,\,\lambda\in\mathcal{Y}\}$$

(in fact we consider $\mathbb{W}_p=\mathbb{G}_p$) and take $\mathcal{Y}=[-M,M]$. Moreover, without loss of generality, suppose that the functions in $\mathcal{W}_{(\widetilde{\rho},\Psi)}$ take values in $[-1,1]$.
We define the sets $\mathcal{Y}^s=\{y_1^s,...,y_{i_s}^s\}$ for $s\geq 1$ recursively. $\mathcal{Y}^1=\{-M,0,M\}$, assuming that the set $\mathcal{Y}^s=\{y_1^s,...,y_{i_s}^s\}$ is built and reordering the elements in increasing order, we take the middle points $\tilde{y}_j^s=\frac{y_j^s+y_{j+1}^s}{2}$ and obtain $\widetilde{\mathcal{Y}}^s=\{\tilde{y}_j^s,\,i=1,...,i_{s-1}-1\}$. Then we define

$$\mathcal{Y}^{s+1}=\mathcal{Y}^s\cup\widetilde{\mathcal{Y}}^s$$

reordered to have increasing elements, and it holds $\mathrm{Card}(\mathcal{Y}^s)=2^s+1$.
Now, we define the classes of functions

$$\mathcal{W}^s_{(\widetilde{\rho},\Psi)}=\{y\in\mathcal{Y}\mapsto\Psi(\widetilde{\rho}(\lambda)\,,\,y)\,,\,\lambda\in\mathcal{Y}_s\}$$

noticing that for all $s\geq 1$,

$$\mathcal{W}^{s-1}_{(\widetilde{\rho},\Psi)}\subsetneq\mathcal{W}^s_{(\widetilde{\rho},\Psi)}\subsetneq\mathcal{W}_{(\widetilde{\rho},\Psi)}\,. \tag{3.19}$$

By this previous display and the fact that $\bigcup_{s\geq 1} \mathcal{Y}^s$ is dense in $[-M, M]$ and by the continuous assumption (3.17) , we have

$$\overline{\lim_{s\to\infty} \mathcal{W}^s_{(\widetilde{\rho},\Psi)}} = \overline{\bigcup_{s\geq 1} \mathcal{W}^s_{(\widetilde{\rho},\Psi)}} = \mathcal{W}_{(\widetilde{\rho},\Psi)}. \tag{3.20}$$

The classes of functions $\mathcal{W}^s_{(\widetilde{\rho},\Psi)}$, $s \geq 1$ are countable with values in $[-1, 1]$ and we may apply the inequality (3.16) to the classes $\mathcal{W}^s_{(\widetilde{\rho},\Psi)}$. We get for all $t \geq 0$ and $s \geq 1$

$$\mathbb{P}\left(\|\mathbb{W}_p\|_{\mathcal{W}^s_{(\widetilde{\rho},\Psi)}} \geq \mathbb{E}(\|\mathbb{W}_p\|_{\mathcal{W}^s_{(\widetilde{\rho},\Psi)}}) + t \right) \quad \leq \quad \varphi_s(t)\,. \tag{3.21}$$

We then prove that the two members of this inequality converge when $s \to \infty$.
Write the left member as follows

$$\begin{aligned} & \mathbb{P}\left(\|\mathbb{W}_p\|_{\mathcal{W}^s_{(\widetilde{\rho},\Psi)}} \geq \mathbb{E}(\|\mathbb{W}_p\|_{\mathcal{W}^s_{(\widetilde{\rho},\Psi)}}) + t \right) \\ = \quad & \mathbb{E}\left(\mathbb{1}_{\|\mathbb{W}_p\|_{\mathcal{W}^s_{(\widetilde{\rho},\Psi)}} \geq \mathbb{E}(\|\mathbb{W}_p\|_{\mathcal{W}^s_{(\widetilde{\rho},\Psi)}}) + t} \right) \\ = \quad & \mathbb{E}\left(\mathbb{1}_{\|\mathbb{W}_p\|_{\mathcal{W}^s_{(\widetilde{\rho},\Psi)}} - \mathbb{E}(\|\mathbb{W}_p\|_{\mathcal{W}^s_{(\widetilde{\rho},\Psi)}}) \geq t} \right). \end{aligned} \tag{3.22}$$

The inclusions (3.19) yields

$$\|\mathbb{W}_p\|_{\mathcal{W}^{s-1}_{(\widetilde{\rho},\Psi)}} \leq \|\mathbb{W}_p\|_{\mathcal{W}^s_{(\widetilde{\rho},\Psi)}} \leq \|\mathbb{W}_p\|_{\mathcal{W}_{(\widetilde{\rho},\Psi)}} \quad \forall s \geq 1\,,$$

so the sequence $\left(\|\mathbb{W}_p\|_{\mathcal{W}^s_{(\widetilde{\rho},\Psi)}} \right)_{s\geq 1}$ is increasing and bounded, thus it converges. By monotone convergence, we obtain that the sequence $\left(\mathbb{E}\left(\|\mathbb{W}_p\|_{\mathcal{W}^s_{(\widetilde{\rho},\Psi)}} \right) \right)_{s\geq 1}$ converges too provided that $\mathbb{E}(\|\mathbb{W}_p\|_{\mathcal{W}_{(\widetilde{\rho},\Psi)}}) < \infty$. Thus, the sequence $\left(\|\mathbb{W}_p\|_{\mathcal{W}^s_{(\widetilde{\rho},\Psi)}} - \mathbb{E}\left(\|\mathbb{W}_p\|_{\mathcal{W}^s_{(\widetilde{\rho},\Psi)}} \right) \right)_{s\geq 1}$ converges, and by dominated convergence the quantity (3.22) converges to the wanted limit

$$\mathbb{E}\left(\mathbb{1}_{\|\mathbb{W}_p\|_{\mathcal{W}_{(\widetilde{\rho},\Psi)}} - \mathbb{E}(\|\mathbb{W}_p\|_{\mathcal{W}_{(\widetilde{\rho},\Psi)}}) \geq t} \right) = \mathbb{P}\left(\|\mathbb{W}_p\|_{\mathcal{W}_{(\widetilde{\rho},\Psi)}} \geq \mathbb{E}(\|\mathbb{W}_p\|_{\mathcal{W}_{(\widetilde{\rho},\Psi)}}) + t \right).$$

For the right member of (3.21), by similar arguments, it can be shown that $\varphi_s(t) \to \varphi(t) = \varphi_{\mathcal{G}}(t)$.
That concludes the proof. □

Next, since the function $t \mapsto \varphi_{\mathcal{G}}(t)$ is decreasing from $\mathbb{R}_+$ into $[0, 1]$, then it exists a unique function $\kappa_{\mathcal{G}} : [0, 1] \to \mathbb{R}_+$ such that

$$\forall\, t \geq 0 \quad \kappa^{-1}_{\mathcal{G}}(t) = \varphi_{\mathcal{G}}(t)\,. \tag{3.23}$$

Then, we can write (3.18) as follows, for all $\varepsilon \in]0, 1[$

$$\mathbb{P}\left(\|\mathbb{W}_p\|_{\mathcal{G}} \geq \mathbb{E}(\|\mathbb{W}_p\|_{\mathcal{G}}) + \kappa_{\mathcal{G}}(\varepsilon) \right) \leq \varepsilon$$

or equivalently

$$\mathbb{P}\left(\|\mathbb{W}_p\|_{\mathcal{G}} \leq \mathbb{E}(\|\mathbb{W}_p\|_{\mathcal{G}}) + \kappa_{\mathcal{G}}(\varepsilon) \right) \geq 1 - \varepsilon\,.$$

Thus, for a constant $K(\varepsilon)$ that should satisfy (3.14), i.e

$$\mathbb{P}(\|\mathbb{W}_p\|_{\mathcal{G}} \leq K(\varepsilon)) \geq 1 - \varepsilon\,,$$

one can take $K(\varepsilon)$ equal to

$$\mathbb{E}(\|\mathbb{W}_p\|_{\mathcal{G}}) + \kappa_{\mathcal{G}}(\varepsilon)\,.$$

But, the quantity $\mathbb{E}(\|\mathbb{W}_p\|_{\mathcal{G}})$ remains not tractable. We propose to bound it.
Indeed, *maximal inequalities* allow to bound such quantities in terms of *entropy integrals*. Although these methods are known to be not sharp, the bounds we will obtain are of interest for our purpose.

Before, let recall some useful notations from [24] (p. 83-85).[7]
Let $\mathcal{G}$ be a class of functions and W some probability measure. We denote $G : y \mapsto G(y)$ an *envelope function* of the class $\mathcal{G}$. The bracketing number is $N_{[]}(\epsilon, \mathcal{G}, L_2(W))$ and the entropy with bracketing is the logarithm of the bracketing number. Last, the bracketing integral is defined as

$$J_{[]}(\delta, \mathcal{G}, L_2(W)) := \int_0^{\delta} \sqrt{\log\, N_{[]}(\epsilon, \mathcal{G}, L_2(W))}d\epsilon\,.$$

Now we apply Corollary 19.35 of [23] (p. 288), it holds that

$$\mathbb{E}(\|\mathbb{W}_p\|_{\mathcal{G}}) \quad \leq \quad a_{\mathcal{G}}\, J_{[]}(\|G\|_{2,W}, \mathcal{G}, L_2(W))\,, \tag{3.24}$$

where
– $a_{\mathcal{G}}$ is some universal constant
– G is an envelop function of $\mathcal{G}$ and

$$\|G\|_{2,W} = \left(\int G^2\, W(dy)\right)^{1/2}.$$

Remark 3.6.1. The quantity $J_{[]}(\|G\|_{2,W}, \mathcal{G}, L_2(W))$ is computable if one has the bracketing numbers $N_{[]}(\epsilon, \mathcal{G}, L_2(W))$ $(\forall\, \epsilon > 0)$, see examples in Section 3.7 below.

Finally, setting

$$K(\varepsilon) = a_{\mathcal{G}}\, J_{[]}\left(\|G\|_{2,Q}, \mathcal{G}_{(\widetilde{\rho},\Psi)}, L_2(W)\right) + \kappa_{\mathcal{G}}(\varepsilon) \tag{3.25}$$

provides the claimed constant. In particular, we should take $\mathcal{G} = \mathcal{W}_{(\widetilde{\rho},\Psi)}$ $(W = Q)$ and $\mathcal{G} = \mathcal{P}_{(\widetilde{\rho},h)}$ $(W = P^{\mathbf{x}})$ in order to compute $\bar{K}^{\varepsilon}_{(\widetilde{\rho},\Psi)}$ and $\bar{K}^{\varepsilon}_{(\widetilde{\rho},h)}$, respectively.

3.7 Constants $\bar{K}^{\varepsilon}_{(\widetilde{\rho},\Psi)}$ and $\bar{K}^{\varepsilon}_{(\widetilde{\rho},h)}$ in particular cases

3.7.1 $\bar{K}^{\varepsilon}_{(\widetilde{\rho},\Psi)}$ for the Mean-contrast

Here we use the computations that can be found in [24].
Recall that in this case

$$\mathcal{W}_{(\widetilde{\rho},\Psi)} = \{y \mapsto (y-\lambda)^2\,,\, \lambda \in \mathcal{Y}\}\,.$$

This class is uniformly bounded by $4\,M^2$, we take the envelop function $G = 4\,M^2$. Then, we have

$$|(y-\lambda_1)^2 - (y-\lambda_2)^2| \leq |\lambda_1 - \lambda_2|\, F(y)\,,$$

with $F(y) = |2y + 2\,M|$.
Following the lines of [24] we obtain the following constant :

$$\bar{K}^{\varepsilon}_{(\widetilde{\rho},\Psi)} = 8\, a_1\, \sqrt{\pi}\, M^2 + \kappa_1(\varepsilon)\,. \tag{3.26}$$

7. Les notions qui suivent sont décrites dans l'Annexe.

3.7.2 $\bar{K}^{\varepsilon}_{(\widetilde{\rho},h)}$ with the weight function $\widetilde{\rho}(y) = y$

In this case, the class of functions $\mathcal{P}_{(\widetilde{\rho},h)}$ is

$$\mathcal{P}_{(\widetilde{\rho},h)} = \{\mathbf{x} \in \mathcal{X} \mapsto h(\mathbf{x},\boldsymbol{\theta})\,,\ \boldsymbol{\theta} \in \Theta\} \quad (\mathcal{X} \subset \mathbb{R}^d)\,.$$

We assumed in the introduction that the models $\mathbf{x} \mapsto h(\mathbf{x},\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$ are uniformly bounded by M, thus denote by P an envelop of $\mathcal{P}_{(\widetilde{\rho},h)}$, take $P = M$.
Moreover, let us suppose that the models $\mathbf{x} \mapsto h(\mathbf{x},\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$ belong to the Hölder space $\mathbb{H}(\mathcal{X},\alpha,L)$ $(\alpha, L > 0)$ defined as

$$\mathbb{H}(\mathcal{X},\alpha,L) = \{g : \mathcal{X} \to \mathbb{R}\, \textit{continuous},\ \|g\|_\alpha \leq L\}$$

where

$$\|g\|_\alpha = \max_{|\nu|\leq\lfloor\alpha\rfloor} \sup_{x\in\mathcal{X}} |D^\nu g(x)| + \max_{\nu:|\nu|=\lfloor\alpha\rfloor} \sup_{x,x'\in\mathcal{X}} \frac{|D^\nu g(x) - D^\nu g(x')|}{\|x-x'\|^{\alpha-\lfloor\alpha\rfloor}}$$

with $\lfloor\alpha\rfloor$ the largest integer smaller than α, and the differential operator D^ν is defined as, for $\nu = (\nu_1,...,\nu_d) \in \mathbb{N}^d$

$$D^\nu = \frac{\partial^{|\nu|}}{\partial\nu_1^{\nu_1}...\partial\nu_d^{\nu_d}}, \quad \textit{and} \quad |\nu| = \sum_{i=1}^d \nu_i\,.$$

We aim at computing the entropy integral $J_{[]}\left(\|P\|_{2,Q},\, \mathcal{P}_{(\widetilde{\rho},h)},\, L_2(Q)\right)$ by integrating the entropy $\log N_{[]}(\epsilon,\, \mathcal{P}_{(\widetilde{\rho},h)},\, L_2(Q))$.
Corollary 2.7.2 in [24] (p. 157) gives an entropy bound for the Hölder space $\mathbb{H}(\mathcal{X},\alpha,1)$:

$$\log N_{[]}(\epsilon,\, \mathbb{H}(\mathcal{X},\alpha,1),\, L_2(Q)) \leq K\left(\frac{1}{\epsilon}\right)^{d/\alpha} \quad \forall\,\epsilon > 0\,, \tag{3.27}$$

where K depends on α, $\mathrm{diam}(\mathcal{X})$ and d.
Using (3.27) and the inequality

$$J_{[]}\left(\|P\|_{2,Q},\, \mathcal{P}_{(\widetilde{\rho},h)},\, L_2(Q)\right) \leq J_{[]}\left(\|P\|_{2,Q},\, \mathbb{H}(\mathcal{X},\alpha,L),\, L_2(Q)\right)\,,$$

it holds for $d < 2\alpha$

$$J_{[]}\left(\|P\|_{2,Q},\, \mathcal{P}_{(\widetilde{\rho},h)},\, L_2(Q)\right) \leq \sqrt{K}\int_0^M \left(\frac{L}{\epsilon}\right)^{d/2\alpha} d\epsilon\,,$$

hence

$$J_{[]}\left(\|P\|_{2,Q},\, \mathcal{P}_{(\widetilde{\rho},h)},\, L_2(Q)\right) \leq M\sqrt{K}\left(\frac{L}{M}\right)^{d/2\alpha} \frac{1}{1-d/2\alpha}\,.$$

Finally, under the condition $d < 2\,\alpha$, we get the constant

$$\bar{K}^{\varepsilon}_{(\widetilde{\rho},h)} = a_2\, M\sqrt{K}\left(\frac{L}{M}\right)^{d/2\alpha} \frac{1}{1-d/2\alpha} + \kappa_2(\epsilon)\,.$$

The condition $d < 2\alpha$ above, means that the dimension of the random input $\mathbf{X}$ (equal to d) is limited by the "smoothness" of the models $\mathbf{x} \mapsto h(\mathbf{x},\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$. The smoother the models are (i.e α large), the larger the dimension d can be.

Remark 3.7.1. To compute the constants $\bar{K}^{\varepsilon}_{(\widetilde{\rho},\Psi)}$ and $\bar{K}^{\varepsilon}_{(\widetilde{\rho},h)}$ is difficult but we have adopted a nonasymptotic point of view, so that these computations are necessary in order to give numerical values to the risk bounds.

3.8 Proofs

To prove the risk bound of Theorem (3.4.1), we need the following lemmas.

3.8.1 Preliminary lemmas

Lemma 3.8.1. *Consider the random functions*

$$y \mapsto \Psi(\rho_h^m(\boldsymbol{\theta})\,,\, y)\,, \quad \boldsymbol{\theta} \in \Theta \quad \textit{with} \quad \rho_h^m(\boldsymbol{\theta}) = \frac{1}{m}\sum_{j=1}^{m} \widetilde{\rho}(h(\mathbf{X}_j', \boldsymbol{\theta}))$$

We have (a.s.)

$$\sup_{\boldsymbol{\theta}\in\Theta} |\mathbb{G}_n\left(\Psi\left(\rho_h^m(\boldsymbol{\theta})\right)\right)| \leq \gamma\, ||\mathbb{G}_n||_{\mathcal{W}_{(\widetilde{\rho},\Psi)}}\,,$$

where $\mathcal{W}_{(\widetilde{\rho},\Psi)}$ is defined in (3.9).

Proof. Conditionally to $\mathbf{X}_1' = \mathbf{x}_1', ..., \mathbf{X}_m' = \mathbf{x}_m'$, we have trivially

$$\left\{ \rho_h^m(\boldsymbol{\theta}) = \frac{1}{m}\sum_{j=1}^{m} \widetilde{\rho}(h(\mathbf{x}_j', \boldsymbol{\theta}))\,,\, \boldsymbol{\theta} \in \Theta \right\} \subset \left\{ \frac{1}{m}\sum_{j=1}^{m} \widetilde{\rho}(\lambda_j)\,,\, (\lambda_j)_{1\leq j\leq m} \in \mathcal{Y}^m \right\}.$$

Hence it yields that

$$\sup_{\boldsymbol{\theta}\in\Theta} |\mathbb{G}_n\left(\Psi\left(\rho_h^m(\boldsymbol{\theta})\right)\right)| \quad \leq \quad ||\mathbb{G}_n||_{\mathcal{W}^m_{(\widetilde{\rho},\Psi)}}\,,$$

with $\mathcal{W}^m_{(\widetilde{\rho},\Psi)} = \{y \in \mathcal{Y} \mapsto \Psi(\frac{1}{m}\sum_{j=1}^{m} \widetilde{\rho}(\lambda_j)\,,\, y)\,,\, (\lambda_j)_{1\leq j\leq m} \in \mathcal{Y}^m\}$. By Assumption 3.4.1 we obtain that

$$\sup_{\boldsymbol{\theta}\in\Theta} |\mathbb{G}_n\left(\Psi\left(\rho_h^m(\boldsymbol{\theta})\right)\right)| \quad \leq \quad \gamma\, ||\mathbb{G}_n||_{\mathcal{W}_{(\widetilde{\rho},\Psi)}}\,,$$

for some universal constant $\gamma > 0$.
Finally, the right member does not depend on $\mathbf{x}_1', ..., \mathbf{x}_m'$, and the result follows. □

Remark 3.8.1. The left member of the inequality in the lemma (3.8.1) depends on the model h, contrary to the right member. Indeed, this last term depends only on the weight function with the associated contrast, and on n.

Lemma 3.8.2. *Consider the $P^{\mathbf{x}}$-empirical process $\mathbb{K}_m^{\mathbf{x}}$ and let $\|\cdot\|_{\mathcal{F}} = |\cdot|$ or $\|\cdot\|_r$ and define*

$$c = \begin{cases} 1 & \textit{if } \widetilde{\rho}(y) \textit{ is constant}, \quad \forall y \in \mathcal{Y}\,, \\ (2\,M)^{1/r} & \textit{else} \end{cases}.$$

We have

$$\sup_{\boldsymbol{\theta}\in\Theta} ||\mathbb{K}_m^{\mathbf{x}}\, \widetilde{\rho}(h(\cdot, \boldsymbol{\theta}))||_{\mathcal{F}} \leq c\, ||\mathbb{K}_m^{\mathbf{x}}||_{\mathcal{P}_{(\widetilde{\rho},h)}}\,,$$

where $\mathcal{P}_{(\widetilde{\rho},h)}$ is defined in (3.10).

Proof. Let us notice that the quantity

$$\mathbb{K}_m^{\mathbf{x}}\,\widetilde{\rho}(h(\cdot,\boldsymbol{\theta})) = \frac{1}{\sqrt{m}}\sum_{j=1}^{m}\left[\widetilde{\rho}(h(\mathbf{X}_j',\boldsymbol{\theta})) - \mathbb{E}_{\mathbf{X}}\widetilde{\rho}(h(\mathbf{X},\boldsymbol{\theta}))\right]$$

can be (up to a factor) either a sum of independent random real variables or a sum of independent random functions.

- If $\widetilde{\rho}(y) \in \mathbb{R}$ for all $y \in \mathcal{Y}$ (we have a sum of random variables).

Taking $\|\cdot\|_{\mathcal{F}} = |\cdot|$ the absolute value norm, it comes directly that

$$\begin{aligned}\sup_{\boldsymbol{\theta}\in\Theta}||\mathbb{K}_m^{\mathbf{x}}\,\widetilde{\rho}(h(\cdot,\boldsymbol{\theta}))||_{\mathcal{F}} &= \sup_{\boldsymbol{\theta}\in\Theta}\left|\frac{1}{\sqrt{m}}\sum_{j=1}^{m}\left[\widetilde{\rho}(h(\mathbf{X}_j',\boldsymbol{\theta})) - \mathbb{E}_{\mathbf{X}}\widetilde{\rho}(h(\mathbf{X},\boldsymbol{\theta}))\right]\right| \\ &= \|\mathbb{K}_m^{\mathbf{x}}\|_{\mathcal{P}_{(\widetilde{\rho},h)}}\end{aligned}$$

Remark 3.8.2. In this case, $\widetilde{\rho}(y)(\lambda) = \widetilde{\rho}(y)$ for all y and λ in $\mathcal{Y}$.
- If, for all $y \in \mathcal{Y}$, $\widetilde{\rho}(y)$ is a real valued function defined on $\mathcal{Y}$.

Take $\|\cdot\|_{\mathcal{F}} = \|\cdot\|_r$, $r \geq 1$, the L_r norm. By integration properties and the fact that

$$\sup_{z\geq 0} z^r = \left(\sup_{z\geq 0} z\right)^r,$$

we have

$$\begin{aligned}\sup_{\boldsymbol{\theta}\in\Theta}||\mathbb{K}_m^{\mathbf{x}}\,\widetilde{\rho}(h(\cdot,\boldsymbol{\theta}))||_r &= \sup_{\boldsymbol{\theta}\in\Theta}\left|\left|\frac{1}{\sqrt{m}}\sum_{j=1}^{m}\left[\widetilde{\rho}(h(\mathbf{X}_j',\boldsymbol{\theta})) - \mathbb{E}_{\mathbf{X}}\widetilde{\rho}(h(\mathbf{X},\boldsymbol{\theta}))\right]\right|\right|_r \\ &= \sup_{\boldsymbol{\theta}\in\Theta}\left(\int_{\mathcal{Y}}\left|\frac{1}{\sqrt{m}}\sum_{j=1}^{m}\left[\widetilde{\rho}(h(\mathbf{X}_j',\boldsymbol{\theta}))(\lambda) - \mathbb{E}_{\mathbf{X}}\widetilde{\rho}(h(\mathbf{X},\boldsymbol{\theta}))(\lambda)\right]\right|^r d\lambda\right)^{1/r} \\ &\leq \sup_{\boldsymbol{\theta}\in\Theta}\left(\int_{\mathcal{Y}}\left(\sup_{\lambda\in\mathcal{Y}}\left|\frac{1}{\sqrt{m}}\sum_{j=1}^{m}\left[\widetilde{\rho}(h(\mathbf{X}_j',\boldsymbol{\theta}))(\lambda) - \mathbb{E}_{\mathbf{X}}\widetilde{\rho}(h(\mathbf{X},\boldsymbol{\theta}))(\lambda)\right]\right|\right)^r d\lambda\right)^{1/r} \\ &= \sup_{\boldsymbol{\theta}\in\Theta}\sup_{\lambda\in\mathcal{Y}}\left|\frac{1}{\sqrt{m}}\sum_{j=1}^{m}\left[\widetilde{\rho}(h(\mathbf{X}_j',\boldsymbol{\theta}))(\lambda) - \mathbb{E}_{\mathbf{X}}\widetilde{\rho}(h(\mathbf{X},\boldsymbol{\theta}))(\lambda)\right]\right|\left(\int_{\mathcal{Y}} d\lambda\right)^{1/r} \\ &= (2\,M)^{1/r}\sup_{(\boldsymbol{\theta},\lambda)\in\Theta\times\mathcal{Y}}\left|\frac{1}{\sqrt{m}}\sum_{j=1}^{m}\left[\widetilde{\rho}(h(\mathbf{X}_j',\boldsymbol{\theta}))(\lambda) - \mathbb{E}_{\mathbf{X}}\widetilde{\rho}(h(\mathbf{X},\boldsymbol{\theta}))(\lambda)\right]\right|.\end{aligned}$$

Finally, notice that

$$\sup_{(\boldsymbol{\theta},y)\in\Theta\times\mathcal{Y}}\left|\frac{1}{\sqrt{m}}\sum_{j=1}^{m}\left[\widetilde{\rho}(h(\mathbf{X}_j',\boldsymbol{\theta}))(y) - \mathbb{E}_{\mathbf{X}}\widetilde{\rho}(h(\mathbf{X},\boldsymbol{\theta}))(y)\right]\right| = ||\mathbb{K}_m^{\mathbf{x}}||_{\mathcal{P}_{(\widetilde{\rho},h)}}$$

and the result follows. □

Remark 3.8.3. In the case where the weight function is a kernel $K_b(\cdot - \cdot)$, the quantity

$$\mathbb{K}_m^{\mathbf{x}}\widetilde{\rho}(h(\cdot,\boldsymbol{\theta})) = \frac{1}{\sqrt{m}}\sum_{j=1}^{m}\Big[K_b(\,\cdot - h(\mathbf{X}_j',\boldsymbol{\theta})) - \mathbb{E}_{\mathbf{X}}K_b(\,\cdot - h(\mathbf{X},\boldsymbol{\theta}))\Big]$$

is treated as a sum of independent random functions in the recent work of A. Goldenshluger and O. Lepski [8]. Here we have made the restrictive assumption that $\mathcal{Y} \subset [-M, M]$. A valuable challenge would be to extend our results to the unbounded case using [8].

3.8.2 Proof of Theorem (3.4.1)

Proof. We denote by

- $M(h,\boldsymbol{\theta}) = \mathcal{R}_\Psi(h,\boldsymbol{\theta}) = \mathbb{E}_Y\,\Psi\left(\rho_h(\boldsymbol{\theta}), Y\right)$
- $M_{n,}(h,\boldsymbol{\theta}) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\Psi\left(\rho_h(\boldsymbol{\theta}), Y_i\right)$
- $M_{,m}(h,\boldsymbol{\theta}) = \mathbb{E}_Y\,\Psi\left(\rho_h^m(\boldsymbol{\theta}), Y\right)$
- $M_{n,m}(h,\boldsymbol{\theta}) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\Psi\left(\rho_h^m(\boldsymbol{\theta}), Y_i\right)$
- $\mathbb{G}_n\Psi\left(\rho_h^m(\boldsymbol{\theta})\right) = \sqrt{n}\,\left(M_{n,m}(h,\boldsymbol{\theta}) - M_m(h,\boldsymbol{\theta})\right)$

where $\rho_h^m(\boldsymbol{\theta}) = \dfrac{1}{m}\displaystyle\sum_{j=1}^{m}\widetilde{\rho}(h(\mathbf{X}_j',\boldsymbol{\theta}))$ and recall that

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\Theta}{\text{Argmin}}\; M_{n,m}(h,\boldsymbol{\theta}) \quad \text{and} \quad \boldsymbol{\theta}^\star = \underset{\boldsymbol{\theta}\in\Theta}{\text{Argmin}}\; M(h,\boldsymbol{\theta})\,. \tag{3.28}$$

We have,

$$\begin{aligned}
& \mathcal{R}_\Psi(h,\widehat{\boldsymbol{\theta}}) \\
= \;& M(h,\widehat{\boldsymbol{\theta}}) - M_m(h,\widehat{\boldsymbol{\theta}}) + M_m(h,\widehat{\boldsymbol{\theta}}) - M_{n,m}(h,\widehat{\boldsymbol{\theta}}) + M_{n,m}(h,\widehat{\boldsymbol{\theta}}) \\
= \;& -\Big(M_m(h,\widehat{\boldsymbol{\theta}}) - M(h,\widehat{\boldsymbol{\theta}})\Big) - \frac{1}{\sqrt{n}}\mathbb{G}_n\Psi\left(\rho_h^m(\widehat{\boldsymbol{\theta}}_{n,m})\right) + \underbrace{M_{n,m}(h,\widehat{\boldsymbol{\theta}}) - M_{n,m}(h,\boldsymbol{\theta}^\star)}_{\leq 0\ (3.28)} + M_{n,m}(h,\boldsymbol{\theta}^\star) \\
\leq \;& -\Big(M_m(h,\widehat{\boldsymbol{\theta}}) - M(h,\widehat{\boldsymbol{\theta}})\Big) - \frac{1}{\sqrt{n}}\mathbb{G}_n\Psi\left(\rho_h^m(\widehat{\boldsymbol{\theta}})\right) + M_{n,m}(h,\boldsymbol{\theta}^\star) - M_m(h,\boldsymbol{\theta}^\star) + M_m(h,\boldsymbol{\theta}^\star) \\
\leq \;& -\Big(M_m(h,\widehat{\boldsymbol{\theta}}) - M(h,\widehat{\boldsymbol{\theta}})\Big) - \frac{1}{\sqrt{n}}\mathbb{G}_n\Psi\left(\rho_h^m(\widehat{\boldsymbol{\theta}})\right) + \frac{1}{\sqrt{n}}\mathbb{G}_n\Psi\left(\rho_h^m(\boldsymbol{\theta}^\star)\right) + M_m(h,\boldsymbol{\theta}^\star) \\
\leq \;& -\Big(M_m(h,\widehat{\boldsymbol{\theta}}) - M(h,\widehat{\boldsymbol{\theta}})\Big) + \frac{1}{\sqrt{n}}\mathbb{G}_n\left(\Psi\left(\rho_h^m(\boldsymbol{\theta}^\star)\right) - \Psi\left(\rho_h^m(\widehat{\boldsymbol{\theta}})\right)\right) \\
& + M_m(h,\boldsymbol{\theta}^\star) - M(h,\boldsymbol{\theta}^\star) + M(h,\boldsymbol{\theta}^\star) \\
\leq \;& \frac{1}{\sqrt{n}}\mathbb{G}_n\left(\Psi\left(\rho_h^m(\boldsymbol{\theta}^\star)\right) - \Psi\left(\rho_h^m(\widehat{\boldsymbol{\theta}})\right)\right) + \left(M_m(h,\boldsymbol{\theta}^\star) - M(h,\boldsymbol{\theta}^\star)\right) - \Big(M_m(h,\widehat{\boldsymbol{\theta}}) - M(h,\widehat{\boldsymbol{\theta}})\Big) \\
& + M(h,\boldsymbol{\theta}^\star) \\
\leq \;& \inf_{\boldsymbol{\theta}\in\Theta}\left(\mathcal{R}_\Psi(h,\boldsymbol{\theta})\right) + \frac{2}{\sqrt{n}}\sup_{\boldsymbol{\theta}\in\Theta}\left|\mathbb{G}_n\left(\Psi\left(\rho_h^m(\boldsymbol{\theta})\right)\right)\right| + 2\sup_{\boldsymbol{\theta}\in\Theta}\left|M_m(h,\boldsymbol{\theta}) - M(h,\boldsymbol{\theta})\right|
\end{aligned}$$

since $M(h,\boldsymbol{\theta}^\star) = \mathcal{R}_\Psi(h,\boldsymbol{\theta}^\star) = \inf_{\boldsymbol{\theta}\in\Theta}\left(\mathcal{R}_\Psi(h,\boldsymbol{\theta})\right)$.

Now, we want to bound the second and third terms in the right member of the last inequality.

Second term. The Lemma 3.8.1 provides that (a.s)

$$\sup_{\boldsymbol{\theta}\in\Theta} |\mathbb{G}_n\left(\Psi\left(\rho_h^m(\boldsymbol{\theta})\right)\right)| \leq \gamma ||\mathbb{G}_n||_{\mathcal{W}_{(\widetilde{\rho},\Psi)}}$$

where $\mathcal{W}_{(\widetilde{\rho},\Psi)} = \{\Psi(\widetilde{\rho}(\lambda)\,,\,\cdot)\,,\,\lambda\in\mathcal{Y}\}$. Thus the second term is bounded by $\frac{2\gamma}{\sqrt{n}}||\mathbb{G}_n||_{\mathcal{W}_{(\widetilde{\rho},\Psi)}}$.

Third term. We have

$$\begin{aligned} |M_m(h,\boldsymbol{\theta}) - M(h,\boldsymbol{\theta})| &= |\mathbb{E}_Y\left(\Psi\left(\rho_h^m(\boldsymbol{\theta}),Y\right) - \Psi\left(\rho_h(\boldsymbol{\theta}),Y\right)\right)| \\ &\leq \mathbb{E}_Y\left|\Psi\left(\rho_h^m(\boldsymbol{\theta}),Y\right) - \Psi\left(\rho_h(\boldsymbol{\theta}),Y\right)\right|. \end{aligned}$$

By Assumption 3.4.2, we obtain

$$|M_m(h,\boldsymbol{\theta}) - M(h,\boldsymbol{\theta})| \leq A_\Psi\,||\rho_h^m(\boldsymbol{\theta}) - \rho_h(\boldsymbol{\theta})||_{\mathcal{F}}\,, \tag{3.29}$$

for some positive constant A_Ψ.
Moreover, the inequality (3.6) yields

$$||\rho_h^m(\boldsymbol{\theta}) - \rho_h(\boldsymbol{\theta})||_{\mathcal{F}} \leq \left|\left|\frac{1}{m}\sum_{j=1}^{m}\left[\widetilde{\rho}(h(\mathbf{X}'_j,\boldsymbol{\theta})) - \mathbb{E}_{\mathbf{X}}\widetilde{\rho}(h(\mathbf{X},\boldsymbol{\theta}))\right]\right|\right|_{\mathcal{F}} + B_h^m(\boldsymbol{\theta})\,. \tag{3.30}$$

Equivalently, by considering the empirical process $\mathbb{K}_m^{\mathbf{x}} = \sqrt{m}(\mathbb{P}_m^{\mathbf{x}} - P^{\mathbf{x}})$, we obtain

$$\begin{aligned} ||\rho_h^m(\boldsymbol{\theta}) - \rho_h(\boldsymbol{\theta})||_{\mathcal{F}} &\leq \frac{1}{\sqrt{m}}\,||\mathbb{K}_m^{\mathbf{x}}\,\widetilde{\rho}(h(\cdot,\boldsymbol{\theta}))||_{\mathcal{F}} + B_h^m(\boldsymbol{\theta}) &(3.31)\\ &\leq \frac{1}{\sqrt{m}}\left(||\mathbb{K}_m^{\mathbf{x}}\,\widetilde{\rho}(h(\cdot,\boldsymbol{\theta}))||_{\mathcal{F}} + \sqrt{m}\,B_h^m(\boldsymbol{\theta})\right). &(3.32)\end{aligned}$$

Taking the *supremum* over Θ and combining the Lemma (3.8.2) and the Assumption 3.3.1 gives

$$\sup_{\boldsymbol{\theta}\in\Theta} ||\rho_h^m(\boldsymbol{\theta}) - \rho_h(\boldsymbol{\theta})||_{\mathcal{F}} \leq \frac{1}{\sqrt{m}}\left(c\,\|\mathbb{K}_m^{\mathbf{x}}\|_{\mathcal{P}_{(\widetilde{\rho},h)}} + \sqrt{m}\,B_h(m)\right).$$

Hence, in (3.29) we obtain

$$\sup_{\boldsymbol{\theta}\in\Theta} |M_m(h,\boldsymbol{\theta}) - M(h,\boldsymbol{\theta})| \leq \frac{A_\Psi}{\sqrt{m}}\left(c\,\|\mathbb{K}_m^{\mathbf{x}}\|_{\mathcal{P}_{(\widetilde{\rho},h)}} + \sqrt{m}\,B_h(m)\right).$$

Finally, the following bound holds for the procedure risk

$$\mathcal{R}_\Psi(h,\widehat{\boldsymbol{\theta}}) \leq \inf_{\boldsymbol{\theta}\in\Theta}\left(\mathcal{R}_\Psi(h,\boldsymbol{\theta})\right) + \frac{2\,\gamma}{\sqrt{n}}\,||\mathbb{G}_n||_{\mathcal{W}_{(\widetilde{\rho},\Psi)}} + 2\,\frac{A_\Psi}{\sqrt{m}}\left(c\,\|\mathbb{K}_m^{\mathbf{x}}\|_{\mathcal{P}_{(\widetilde{\rho},h)}} + \sqrt{m}\,B_h(m)\right).$$

Now, let us notice that for any 3 events E_1, E_2, E_3 we have by elementary probability calculus

$$\mathbb{P}(E_1) \leq \mathbb{P}(E_1\cap E_2\cap E_3) + \mathbb{P}(E_2^c) + \mathbb{P}(E_3^c)\,. \tag{3.33}$$

Take the following events

$$E_1 = \left\{\mathcal{R}_\Psi(h,\widehat{\boldsymbol{\theta}}) \leq \inf_{\boldsymbol{\theta}\in\Theta}\left(\mathcal{R}_\Psi(h,\boldsymbol{\theta})\right) + \frac{2\,\gamma}{\sqrt{n}}\,\|\mathbb{G}_n\|_{\mathcal{W}_{(\widetilde{\rho},\Psi)}} + 2\,\frac{A_\Psi}{\sqrt{m}}\left(c\,\|\mathbb{K}_m^{\mathbf{x}}\|_{\mathcal{P}_{(\widetilde{\rho},h)}} + \sqrt{m}\,B_h(m)\right)\right\}$$

$$E_2 = \left\{ \inf_{\boldsymbol{\theta}\in\Theta} \left(\mathcal{R}_\Psi(h,\boldsymbol{\theta})\right) + \frac{2\,\gamma}{\sqrt{n}} \left\|\mathbb{G}_n\right\|_{\mathcal{W}_{(\widetilde{\rho},\Psi)}} \leq \inf_{\boldsymbol{\theta}\in\Theta} \left(\mathcal{R}_\Psi(h,\boldsymbol{\theta})\right) + \frac{2\,\gamma}{\sqrt{n}}\, \bar{K}^{\varepsilon}_{(\widetilde{\rho},\Psi)} \right\}$$

and

$$E_3 = \left\{ 2\,\frac{A_\Psi}{\sqrt{m}} \left(c\,\|\mathbb{K}^{\mathbf{x}}_m\|_{\mathcal{P}_{(\widetilde{\rho},h)}} + \sqrt{m}\,B_h(m) \right) \leq 2\,\frac{A_\Psi}{\sqrt{m}} \left(c\,\bar{K}^{\varepsilon}_{(\widetilde{\rho},h)} + \sqrt{m}\,B_h(m) \right) \right\},$$

where $\bar{K}^{\varepsilon}_{(\widetilde{\rho},\Psi)}$ and $\bar{K}^{\varepsilon}_{(\widetilde{\rho},h)}$ are such that

$$\mathbb{P}_{Y_{1\ldots n}}(\|\mathbb{G}_n\|_{\mathcal{W}_{(\widetilde{\rho},\Psi)}} \leq \bar{K}^{\varepsilon}_{(\widetilde{\rho},\Psi)}) \geq 1-\varepsilon$$

and

$$\mathbb{P}_{\mathbf{X}_{1\ldots m}}(\|\mathbb{K}^{\mathbf{x}}_m\|_{\mathcal{P}_{(\widetilde{\rho},h)}} \leq \bar{K}^{\varepsilon}_{(\widetilde{\rho},h)}) \geq 1-\varepsilon$$

respectively (for all $\varepsilon > 0$).
Using the inequality (3.33) with the fact that $\mathbb{P}(E_2) = \mathbb{P}_{Y_{1\ldots n}}(\|\mathbb{G}_n\|_{\mathcal{W}_{(\widetilde{\rho},\Psi)}} \leq \bar{K}^{\varepsilon}_{(\widetilde{\rho},\Psi)})$ and $\mathbb{P}(E_3) = \mathbb{P}_{\mathbf{X}_{1\ldots m}}(\|\mathbb{K}^{\mathbf{x}}_m\|_{\mathcal{P}_{(\widetilde{\rho},h)}} \leq \bar{K}^{\varepsilon}_{(\widetilde{\rho},h)})$, we obtain

$$\mathbb{P}(E_1) \leq \mathbb{P}_{Y_{1\ldots n},\mathbf{X}_{1,\ldots,m}} \left(\mathcal{R}_\Psi(h,\widehat{\boldsymbol{\theta}}) \leq \inf_{\boldsymbol{\theta}\in\Theta} \left(\mathcal{R}_\Psi(h,\boldsymbol{\theta})\right) + \frac{2\,\gamma}{\sqrt{n}}\,\bar{K}^{\varepsilon}_{(\widetilde{\rho},\Psi)} + 2\,\frac{A_\Psi}{\sqrt{m}} \left(c\,\bar{K}^{\varepsilon}_{(\widetilde{\rho},h)} + \sqrt{m}\,B_h(m) \right) \right) + 2\,\varepsilon\,.$$

But note that $\mathbb{P}(E_1) = 1$, so

$$\mathbb{P}_{Y_{1\ldots n},\mathbf{X}_{1,\ldots,m}} \left(\mathcal{R}_\Psi(h,\widehat{\boldsymbol{\theta}}) \leq \inf_{\boldsymbol{\theta}\in\Theta} \left(\mathcal{R}_\Psi(h,\boldsymbol{\theta})\right) + \frac{2\,\gamma}{\sqrt{n}}\,\bar{K}^{\varepsilon}_{(\widetilde{\rho},\Psi)} + 2\,\frac{A_\Psi}{\sqrt{m}} \left(c\,\bar{K}^{\varepsilon}_{(\widetilde{\rho},h)} + \sqrt{m}\,B_h(m) \right) \right) \geq 1-2\,\varepsilon\,.$$

Equivalently, we have with probability at least $1-2\,\varepsilon$

$$\mathcal{R}_\Psi(h,\widehat{\boldsymbol{\theta}}) \leq \inf_{\boldsymbol{\theta}\in\Theta} \left(\mathcal{R}_\Psi(h,\boldsymbol{\theta})\right) + \frac{K^{\varepsilon}_{(\widetilde{\rho},\Psi)}}{\sqrt{n}} \left(1 + \sqrt{\frac{n}{m}} (K^{\varepsilon}_{(\widetilde{\rho},h)} + B_m) \right)$$

where

$$K^{\varepsilon}_{(\widetilde{\rho},\Psi)} = 2\,\gamma\,\bar{K}^{\varepsilon}_{(\widetilde{\rho},\Psi)}\,,$$

$$K^{\varepsilon}_{(\widetilde{\rho},h)} = A_\Psi\,c\,\frac{\bar{K}^{\varepsilon}_{(\widetilde{\rho},h)}}{\gamma\,\bar{K}^{\varepsilon}_{(\widetilde{\rho},\Psi)}}$$

and

$$B_m = \sqrt{m}\frac{A_\Psi}{\gamma\,\bar{K}^{\varepsilon}_{(\widetilde{\rho},\Psi)}} B_h(m)\,.$$

That concludes the proof. □

Bibliographie

[1] P. Barbillon, G. Celeux, A. Grimaud, Y. Lefebvre, and E. De Rocquigny. Nonlinear methods for inverse statistical problems. *Computational Statistics & Data Analysis*, 55(1) :132–142, 2011.

[2] P. Billingsley. *Convergence of probability measures*. Wiley New York, 1968.

[3] E. de Rocquigny, N. Devictor, and S. Tarantola, editors. *Uncertainty in industrial practice*. John Wiley.

[4] D.L. Donoho and J.M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3) :425–455, 1994.

[5] M.D. Donsker. Justification and extension of Doob's heuristic approach to the Kolmogorov-Smirnov theorems. *The Annals of mathematical statistics*, pages 277–281, 1952.

[6] R.M. Dudley. Weak convergence of measures on nonseparable metric spaces and empirical measures on euclidian spaces. *Illinois Journal of Mathematics*, 11 :109–126, 1966.

[7] P. Gaenssler. *Empirical Processes*. Institute of Mathematical Statistics, Hayward, CA, 1983.

[8] A. Goldenshluger and O. Lepski. Uniform bounds for norms of sums of independent random functions. *Arxiv preprint arXiv :0904.1950*, 2009.

[9] P.J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, pages 73–101, 1964.

[10] P.J. Huber. *Robust statistics*. Wiley-Interscience, 1981.

[11] J.P.C. Kleijnen. *Design and analysis of simulation experiments*. Springer Verlag, 2007.

[12] T. Klein and E. Rio. Concentration around the mean for maxima of empirical processes. *Annals of probability*, 33(3) :1060–1077, 2005.

[13] M.R. Kosorok. *Introduction to empirical processes and semiparametric inference*. Springer series in statistics, 2008.

[14] M. Ledoux. *The concentration of measure phenomenon*. AMS, 2001.

[15] P. Massart. *Concentration inequalities and model selection : Ecole d'Eté de Probabilités de Saint-Flour XXXIII-2003*. Springer Verlag, 2007.

[16] P. Massart and É. Nédélec. Risk bounds for statistical learning. *The Annals of Statistics*, 34(5) :2326–2366, 2006.

[17] D. Pollard. Empirical processes : theory and applications. *Regional Conference Series in Probability and Statistics Hayward*, 1990.

[18] T.J. Santner, B.J. Williams, and W. Notz. *The design and analysis of computer experiments*. Springer Verlag, 2003.

[19] G.R Shorack and J.A Wellner. *Empirical processes with applications to statistics*. Wiley Series in Probability and Statistics, 1986.

[20] C. Soize and R. Ghanem. Physical systems with random uncertainties : chaos representations with arbitrary probability measure. *SIAM Journal on Scientific Computing*, 26 :395–410, 2004.

[21] M. Talagrand. Sharper bounds for Gaussian and empirical processes. *The Annals of Probability*, 22(1) :28–76, 1994.

[22] S. van de Geer. *Empirical processes in M-estimation*. Cambridge University Press, 2000.

[23] A.W. van der Vaart. *Asymptotic statistics*. Cambridge University Press, 2000.

[24] A.W. van der Vaart and J.A. Wellner. *Weak Convergence and Empirical Processes*. Springer Series in Statistics, 1996.

[25] E. Vazquez. (PhD thesis) Modélisation comportementale de systèmes non-linéaires multivariables par méthodes à noyaux et applications. 2005.

Chapitre 4

Prediction de quantités d'intérêt par simulations numériques

Sommaire

Résumé du Chapitre

Dans le chapitre qui précède, nous avons présenté une procédure d'estimation basée sur des fonctions de *contraste*, prenant en compte la simulation de modèles numériques. Or, dans le cadre de la calibration par exemple, l'estimation des paramètres est suivie de simulations du modèle numérique sous ces paramètres, dans le but de calculer une certaine quantité d'intérêt relative à la variable Y. On parlera de *Prédiction*. Dans ce chapitre, nous nous intéressons à l'influence de la procédure d'estimation des paramètres sachant que l'on veut prédire une certaine quantité d'intérêt fixée. On parlera de *dualité* estimation-prédiction.

About Prediction with Computer Experiments

Nabil Rachdi [1], Jean-Claude Fort [2]

Abstract
In this study, we try to highlight a kind of *duality* between (the nature of) the estimation procedures and (the nature of) the forecasting of a quantity of interest. Indeed, contrast minimisation intrinsically characterizes a "unique" quantity of interest : for example, the least squares (or *reg*) contrast characterizes the conditional expectation or the log-contrast characterize the density function etc... Hence, a natural way for predicting is to use the contrast that characterizes the wanted prediction for parameters estimation. But in practice, one may make predictions with a model where the parameters have already been estimated by some contrast, not necessarily the "good" one. Then it appears challenging to investigate such procedures.
Our study is illustrated with some academic and numerical examples.

4.1 Introduction

In this paper, we are interested in the study of a quantity of interest of some "complex" random phenomenon Y. For instance, one may want to know some features like mean, quantile, probability density function etc... Often, we have at disposal numerical models representing - modeling - the phenomenon Y, with more or less accuracy. These quantities can play a crucial role in a decision making process, then it is important to control the way of computing such quantities : statistical error, model error, hypothesis judgment etc...
In this work, we try to formalize what can be done in practice for computing some quantity of interest thanks to stochastic numerical model simulations. The work in [2] points out this problem in industrial practice and a mathematical formalization have been proposed in [3].
Two kinds of studies may be interesting to investigate

- the stochastic behavior of Y which can model, in practice, a natural or physical phenomenon subjected to variability or uncertainty
- the characterization of "causes" producing the phenomenon Y, i.e the identification of experimental conditions.

The first problem will be called *Prediction problem* and the second *Inverse problem*.
In most of applications, the modeling of the *variable of interest* Y is made through models $\mathcal{H}$ viewed as black-boxes (e.g only known through input/output values)

$$\begin{array}{rcl} h : \mathcal{X} \times \Theta & \longmapsto & \mathcal{Y} \\ (\mathbf{x}, \boldsymbol{\theta}) & \longrightarrow & h(\mathbf{x}, \boldsymbol{\theta}) \,. \end{array}$$

Let us suppose for a moment that some knowledge gives us an input $\mathbf{x}_0$ corresponding to the experimental conditions considered. For each parameter $\boldsymbol{\theta} \in \Theta$, the model output $h(\mathbf{x}_0, \boldsymbol{\theta})$ is deterministic, whereas at this same configuration the variable of interest Y presents some

1. Institut de Mathématiques de Toulouse - EADS Innovation Works, 92152 Suresnes
2. Université Paris Descartes, 45 rue des saints pères, 75006 Paris

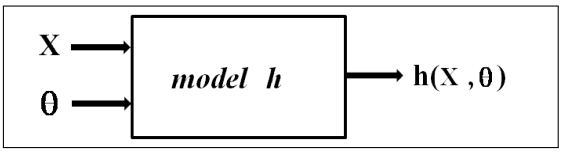


FIGURE 4.1 – black-box model

variability. This can be explained by the fact that we don't take into account the variability of experimental conditions in the modeling. Thus, in order to be more realistic compared with the observed phenomenon, we equip the input space $\mathcal{X}$ with a probability measure $P^{\mathbf{x}}$ which forms a probability space $(\mathcal{X}, \mathcal{B}, P^{\mathbf{x}})$, called *Input conditions*. Now, for each $\boldsymbol{\theta} \in \Theta$, the model output $h(\mathbf{X}, \boldsymbol{\theta})$ is random and with the same "nature" than the variable of interest Y (see figure (4.1)). At first glance, experimental conditions (producing Y) and input conditions may have no link in that the random variable Y may be independent of $\mathbf{X}$, we say that experimental and input conditions are *independent*. This is the case for very complex experimental conditions where only experimental data are available, for example in meteorology, oceanology etc ... However, in practice it happens that experimental and input conditions are of the same nature, for instance, the phenomenon Y is the result (output) of an expensive computer code or experiments, depending (among other factors) on $\mathbf{X}$. We will say that experimental and input conditions are *dependent*. This case provides an additional knowledge on experimental conditions, they are partially known.
In our study, both situations will be considered, we will see that the difference between experimental and input conditions can be reduced to a probabilistic dependence phenomenon.

The purpose of this work is to study prediction procedures that need estimation of parameters. Indeed, one may first compute some parameter $\widehat{\boldsymbol{\theta}}$ and then simulate the (stochastic) model $\mathbf{X} \mapsto h(\mathbf{X}, \widehat{\boldsymbol{\theta}})$ in order to approximate (to predict) some quantity of interest related to the random phenomenon Y. We will see that for a fixed quantity of interest, several estimators $\widehat{\boldsymbol{\theta}}$ can be computed and we aim at studying the effect of the estimation procedure on the prediction performance. Indeed, we will see that it exists a lot of manners for computing a $\widehat{\boldsymbol{\theta}}$. In particular, it is the result of the minimization of "cost functions" and an interesting question would be the choice of an "optimal" one. Roughly speaking, we try to answer to the questions

"Should the modeling, e.g the choice of a model h and/or parameter $\boldsymbol{\theta}$, depends on the quantity of interest ?" (we will talk about *Goal-Oriented modeling*)

"What can happen if the modeling doesn't take into account its final use ?"
(we will talk about *Cross modeling*)

4.2 Definitions and notations

4.2.1 General settings

Suppose that Y is a real-valued random variable, taking values in $\mathcal{Y}$, with distribution Q (unknown) and Lebesgue density f. Consider a class of models $\mathcal{H}$ defined as

$$\begin{array}{rcl} h \in \mathcal{H} \,:\, (\mathcal{X}, \mathcal{B}, P^{\mathbf{x}}) \times \Theta & \longmapsto & \mathcal{Y} \\ (\mathbf{X}, \boldsymbol{\theta}) & \longrightarrow & h(\mathbf{X}, \boldsymbol{\theta}) \end{array}$$

where $\mathcal{X} \subset \mathbb{R}^d$ is called *input space*, $(\mathcal{X}, \mathcal{B}, P^{\mathbf{x}})$ *input conditions* and $\Theta \subset \mathbb{R}^k$ (compact) is called *parameter space*.
Let $Q^{\mathbf{z}}$ be the joint distribution of the variable $Z = (\mathbf{X}, Y)$. In our developments, we will work at fixed model $h \in \mathcal{H}$. The probability measure $P^{\mathbf{x}}$ is not supposed to be known, we only need a (large) sample drawn from this distribution.

4.2.2 Goal

In this chapter, we will try to understand the *duality* between estimation procedures and prediction ones. More precisely, one question may be : given the feature one wants to predict, what estimation procedure to use ? Conversely, given parameters provided from some estimation procedure, what are the features one can hope to predict (reasonably) ?
In our framework, we formalize this as follows.

Let us denote by $\mathcal{F}^p$ the *prediction* feature space containing the feature $\rho^p = \rho_{\mathcal{F}^p}$ we want to predict. Then, let us consider the *prediction* model feature space relative to the measures induced by the random variables $h(\mathbf{X}, \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$

$$F = \{\rho_{\mathcal{F}^p}(\boldsymbol{\theta}),\ \boldsymbol{\theta} \in \Theta\} \subset \mathcal{F}^p. \tag{4.1}$$

The feature $\rho_{\mathcal{F}^p}(\boldsymbol{\theta})$ depends on the numerical model h and should be denoted by $\rho_{\mathcal{F}^p}(h, \boldsymbol{\theta})$. However, since h is fixed, for notational convenience we keep the notation $\rho_{\mathcal{F}^p}(\boldsymbol{\theta})$.

Example 4.2.1. Examples of features

- $\rho_{\mathcal{F}^p}(\boldsymbol{\theta})(\cdot) = h(\cdot, \boldsymbol{\theta})$ ("feature = model $\mathbf{x} \mapsto h(\mathbf{x}, \boldsymbol{\theta})$")
- $\rho_{\mathcal{F}^p}(\boldsymbol{\theta}) = \mathbb{E}_{P^{\mathbf{x}}}(h(\mathbf{X}, \boldsymbol{\theta}))$ ("feature = mean")
- $\rho_{\mathcal{F}^p}(\boldsymbol{\theta}) = \mathbb{P}(h(\mathbf{X}, \boldsymbol{\theta}) > y_0)$ ("feature = exceedance probability")
- $\rho_{\mathcal{F}^p}(\boldsymbol{\theta}) = pdf\ of\ h(\mathbf{X}, \boldsymbol{\theta})$ ("feature = density function")
- etc...

We recall that the model $h : \mathcal{X} \times \Theta \longrightarrow \mathcal{Y}$ can have a complicated form (solution of differential equations, finite elements code etc...) and thus is only known through input/output values. In this context, a feature $\rho_{\mathcal{F}^p}(\boldsymbol{\theta})$ is intractable. That's why we need to use a sample set

$$\mathcal{X}^m = \{\mathbf{X}'_1, ..., \mathbf{X}'_m\}, \quad \mathbf{X}'_j \text{ i.i.d from } P^{\mathbf{x}}$$

and then consider the simulated version of the model prediction (4.1)

$$\rho^m_{\mathcal{F}^p}(\boldsymbol{\theta}) := \frac{1}{m} \sum_{j=1}^{m} \widetilde{\rho}_{\mathcal{F}^p}(h(\mathbf{X}'_j, \boldsymbol{\theta})) \tag{4.2}$$

where $\widetilde{\rho}_{\mathcal{F}^p} : \mathcal{Y} \to \mathcal{F}^p$ is a *weight function* depending on the considered feature space $\mathcal{F}^p$.

Example 4.2.2. Examples of weight functions.

- for mean prediction

$$\widetilde{\rho}_{\mathcal{F}^p}(y) = y\,,$$

- for density prediction
$$\widetilde{\rho}_{\mathcal{F}^p}(y)(\cdot) = K_b(\cdot - y)\,,$$
where $K_b(\cdot - y) = \frac{1}{b}K(\frac{\cdot - y}{b})$ for a kernel $K()$ and a bandwidth b (other methods are available) ,
- for cumulative distribution function prediction
$$\widetilde{\rho}_{\mathcal{F}^p}(y)(\cdot) = \mathbb{1}_{y \leq \cdot}\,,$$
- etc ...

Then, we form the following set of predictors

$$F = \{\rho^m_{\mathcal{F}^p}(\boldsymbol{\theta}),\, \boldsymbol{\theta} \in \Theta\} \subset \mathcal{F}^p\,. \tag{4.3}$$

But it remains the estimation of the parameter $\boldsymbol{\theta}$, say $\widehat{\boldsymbol{\theta}}$, which will provide the prediction

$$\widehat{\rho^p} = \rho^m_{\mathcal{F}^p}(\widehat{\boldsymbol{\theta}})\,. \tag{4.4}$$

Hence, the questions we asked at the beginning of the section are about the estimation procedure giving $\widehat{\boldsymbol{\theta}}$ related to the considered prediction space $\mathcal{F}^p$.

In our developments, we will only focus on estimation procedures based on contrasts minimization.

4.3 Parameter estimation

In this subsection, we recall briefly how are built Ψ-estimators $\widehat{\boldsymbol{\theta}}_\Psi$ for some contrast Ψ.
The definitions and results are taken from Rachdi et al. [3]. [3]

Definition 4.3.1. Contrast and risk function
A contrast is an application

$$\begin{array}{rcl} \Psi : \mathcal{F} & \longrightarrow & L_1(Q^{\mathbf{z}}) \\ \rho & \longmapsto & \Psi(\rho, \cdot) : (\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y} \longmapsto \Psi(\rho, (\mathbf{x}, y)) \end{array}$$

such that $\rho_{\mathcal{F}} = \text{Argmin}_{\rho \in \mathcal{F}}\, \mathbb{E}_{Q^{\mathbf{z}}} \Psi(\rho, Z)$ is unique.
We denote the $\Psi-$risk of some model feature $\rho = \rho_{\mathcal{F}}(\boldsymbol{\theta})$ in $F \subset \mathcal{F}$ by

$$\mathcal{R}_\Psi(\boldsymbol{\theta}) := \mathbb{E}_{Q^{\mathbf{z}}} \Psi(\rho_{\mathcal{F}}(\boldsymbol{\theta}), Z)\,.$$

Example 4.3.1. Example of contrasts
- **mean-contrast**
$$\Psi\,(\rho, y) = (y - \rho)^2$$
- log**-contrast**
$$\Psi\,(\rho, y) = -\log \rho(y)$$

3. Voir aussi le Chapitre 3

- L_2-**contrast**

$$\Psi(\rho, y) = ||\rho||_2^2 - 2\,\rho(y)\,.$$

- **class-contrast** (Classification)

$$\Psi(\rho, z = (\mathbf{x}, y)) = \mathbb{1}_{\rho(\mathbf{x}) \neq y}\,,$$

- **reg-contrast** (Regression)

$$\Psi(\rho, z = (\mathbf{x}, y)) = (y - \rho(\mathbf{x}))^2\,,$$

- etc...

Remark 4.3.1. Notice that the Ψ-risk only depends on the $\mathcal{F}-$contrast Ψ.

The feature $\rho_{\mathcal{F}}(\boldsymbol{\theta})$ is uncomputable, so we use the approximation

$$\rho_{\mathcal{F}}^m(\boldsymbol{\theta}) := \frac{1}{m}\sum_{j=1}^{m} \widetilde{\rho}(h(\mathbf{X}'_j, \boldsymbol{\theta}))$$

where the sample $\mathbf{X}'_1, ..., \mathbf{X}'_m$ is i.i.d from $P^{\mathbf{x}}$ and $\frac{1}{m}\widetilde{\rho} : \mathcal{Y} \to \mathcal{F}$ is a weight function depending on the considered $\mathcal{F}$-contrast Ψ (see (4.2.2)).

Let us consider the $\Psi-$estimator

$$\widehat{\boldsymbol{\theta}}_{\Psi} = \underset{\boldsymbol{\theta}\in\Theta}{\text{Argmin}} \sum_{i=1}^{n} \Psi\left(\sum_{j=1}^{m} \widetilde{\rho}_{\mathcal{F}}(h(\mathbf{X}'_j, \boldsymbol{\theta}))\,,\, Y_i\right)\,. \tag{4.5}$$

Remark 4.3.2. We also consider the reg-contrast $\Psi = \Psi_{reg}$ providing the $\Psi_{reg}-$estimator given by

$$\begin{aligned}\widehat{\boldsymbol{\theta}}_{\Psi_{reg}} &= \underset{\boldsymbol{\theta}\in\Theta}{\text{Argmin}} \sum_{i=1}^{n} \Psi_{reg}\left(h(\cdot, \boldsymbol{\theta})\,,\, Z_i\right) \quad Z_i = (\mathbf{X}_i, Y_i) \\ &= \underset{\boldsymbol{\theta}\in\Theta}{\text{Argmin}} \sum_{i=1}^{n} \left(h(\mathbf{X}_i, \boldsymbol{\theta}) - Y_i\right)^2\,.\end{aligned}$$

In this case, we don't need simulation sample.

Finally, if one considers different contrasts, several parameters can be computed. The "quality" of some Ψ-estimator $\widehat{\boldsymbol{\theta}}_{\Psi}$ can be measured by its excess risk [4] (2.46) given by

$$\mathcal{E}_{\Psi}^{\Theta}(\widehat{\boldsymbol{\theta}}_{\Psi}) = \mathcal{R}_{\Psi}(\widehat{\boldsymbol{\theta}}_{\Psi}) - \mathcal{R}_{\Psi}^{\Theta}\,, \quad \mathcal{R}_{\Psi}^{\Theta} = \inf_{\boldsymbol{\theta}\in\Theta} \mathcal{R}_{\Psi}(\boldsymbol{\theta})\,.$$

For Ψ-estimators (4.5), we prove the following inequality in Rachdi et al. [3].

Theorem 4.3.1. ***Risk bound***
Let fix a model h. Under some assumptions on the contrast Ψ *and tightness conditions, we have for all* $\varepsilon > 0$*, with probability at least* $1 - 2\varepsilon$

$$0 \leq \mathcal{E}_{\Psi}^{\Theta}(\widehat{\boldsymbol{\theta}}_{\Psi}) \leq \frac{K^{\varepsilon}_{(\widetilde{\rho},\Psi)}}{\sqrt{n}}\left(1 + \sqrt{\frac{n}{m}}(K^{\varepsilon}_{(\widetilde{\rho},h)} + B_m)\right)$$

for some constants $K^{\varepsilon}_{(\widetilde{\rho},\Psi)}$, $K^{\varepsilon}_{(\widetilde{\rho},h)}$ *and a bias factor* B_m.

4. Voir Chapitre 2

4.4 Cross Prediction

4.4.1 Definitions

Now, considering (4.4), the purpose is to investigate the "quality" of predictions

$$\widehat{\rho^p} = \rho^m_{\mathcal{F}^p}(\widehat{\boldsymbol{\theta}}_\Psi)$$

by varying the Ψ-estimator $\widehat{\boldsymbol{\theta}}_\Psi$ (4.5), where Ψ is some $\mathcal{F}$-contrast.

In order to highlight important issues in what follows, we don't take into account the simulation aspect of the feature $\rho_{\mathcal{F}^p}(\boldsymbol{\theta})$. It means that we will deal with predictions of the form

$$\widehat{\rho^p} = \rho_{\mathcal{F}^p}(\widehat{\boldsymbol{\theta}}_\Psi) \in F \subset \mathcal{F}^p\,. \tag{4.6}$$

Remark 4.4.1. This last suggestion doesn't impact the following reasoning. It just simplifies the notations.

Remark 4.4.2. In some sense, the parameter $\widehat{\boldsymbol{\theta}}_\Psi$ represents the "way of modeling" or the "validity domain" of the numerical model h. Indeed, we have in mind that one modeling is always done for one purpose (which can be unknown or uncontrolled). Thus, it appears important to study the effect of some modeling on quantities of interest.

Let us suppose that

$$\exists\, \Psi^p \,:\, \mathcal{F}^p \to L_1(Q^{\mathbf{z}})\,, \quad \rho_{\mathcal{F}^p} = \underset{\rho\in\mathcal{F}^p}{\operatorname{Argmin}}\, \mathcal{R}_{\Psi^p}(\rho)\,,$$

which provides the prediction risk

$$\mathcal{R}_{\Psi^p}(\boldsymbol{\theta})\,(= \mathcal{R}_{\Psi^p}(\rho_{\mathcal{F}^p}(\boldsymbol{\theta}))) = \mathbb{E}_{Q^{\mathbf{z}}}\Psi^p(\rho_{\mathcal{F}^p}(\boldsymbol{\theta}), Z)\,.$$

The display (4.6) provides two cases :

– *first case* : the contrast Ψ is equal to Ψ^p

– *second case* : the contrast Ψ is any other contrast.

We formalize this by the following definition.

Definition 4.4.1. Cross prediction.
A **cross prediction** consists in predicting ρ^p by

$$\widehat{\rho^p} = \rho_{\mathcal{F}^p}(\widehat{\boldsymbol{\theta}}_\Psi)$$

where $\widehat{\boldsymbol{\theta}}_\Psi$ is a Ψ-estimator built from a Ψ-procedure with

$$\Psi \neq \Psi^p\,.$$

Else, we talk about **classical** prediction.

Remark 4.4.3. Cross prediction and convexification.
This notion of "cross prediction" is in the same spirit as the work of P. Barlett et al. [1] in classification. Indeed, instead of minimizing a classification contrast Ψ^p, the authors propose to use a "convex surrogate" contrast Ψ and they show theoretical results about rates of convergence of such procedure.
This technic may be viewed as a first kind of cross prediction we will present in the following.

We may consider **two kinds** of cross prediction :

▶ First, the contrast Ψ may be different from Ψ^p but defined on the same feature space $\mathcal{F} = \mathcal{F}^p$. For example, in the case of density prediction, one may take

$$\Psi^p(\rho, y) = -\log(\rho)(y)\,, \quad \text{(log-contrast)}\,,$$

and

$$\Psi_{L_2}(\rho, y) = \|\rho\|_2^2 - 2\,\rho(y)\,, \quad (L_2\text{-contrast})\,.$$

Both are different, and provide two predictions $\rho_{\mathcal{F}^p}(\widehat{\boldsymbol{\theta}}_{\Psi^p})$ and $\rho_{\mathcal{F}^p}(\widehat{\boldsymbol{\theta}}_{\Psi_{L_2}})$.
This case includes also the situation where we use penalized contrasts Ψ_{pen} given by

$$\Psi_{pen}(\rho, \mathbf{z}) = \Psi^p(\rho, \mathbf{z}) + K\,\mathrm{pen}_{shape}(\rho)\,.$$

Then, to compare Ψ_{pen} to Ψ^p amounts to compare penalized and unpenalized procedures.

▶ Second, the contrast Ψ may be different from Ψ^p **and not** defined on the (prediction) feature space $\mathcal{F}^p$. For example, let us consider again the density prediction with

$$\Psi^p(\rho, y) = -\log(\rho)(y) \quad \text{(log-contrast)}\,.$$

Now, suppose that the contrast Ψ is "totally different" from Ψ^p (in the spirit of predicting a density), for example

$$\Psi_{mean}(\rho, y) = (y - \rho)^2\,, \quad \rho = \mathbb{E}_{P^{\mathbf{X}}}(h(\mathbf{X}, \boldsymbol{\theta})) \quad \text{(mean-contrast)}\,.$$

Or, we can also consider

$$\Psi_{reg}(\rho, (\mathbf{x}, y)) = (y - \rho(\mathbf{x}))^2\,, \quad \rho = h(\cdot, \boldsymbol{\theta}) \quad \text{(reg-contrast)}\,.$$

We obtain the predictions $\rho_{\mathcal{F}^p}(\widehat{\boldsymbol{\theta}}_{\Psi^p})$, $\rho_{\mathcal{F}^p}(\widehat{\boldsymbol{\theta}}_{\Psi_{mean}})$ and $\rho_{\mathcal{F}^p}(\widehat{\boldsymbol{\theta}}_{\Psi_{reg}})$.

In our developments, we focus on the second case. The first one is also of interest. Intuitively, it seems that in the second case the cross prediction $\rho_{\mathcal{F}^p}(\widehat{\boldsymbol{\theta}}_{\Psi_{reg}})$ with regression parameter estimation is "better" than the cross prediction $\rho_{\mathcal{F}^p}(\widehat{\boldsymbol{\theta}}_{\Psi_{mean}})$ using mean-contrast estimation. This intuition can be explained by the fact that a "regression" procedure may bring more information than a "mean" one in order to predict a density function.
This will be illustrated in Section 4.6 by a toy numerical example, where we obtain the Figure 4.2.

Now, the issue is to study the performance of the predictions $\widehat{\rho^p} = \rho_{\mathcal{F}^p}(\widehat{\boldsymbol{\theta}}_{\Psi^p})$ and $\widehat{\rho^p} = \rho_{\mathcal{F}^p}(\widehat{\boldsymbol{\theta}}_{\Psi})$ (cross prediction). Since the performance is measured by the risk

$$\mathcal{R}_{\Psi^p}(\boldsymbol{\theta})\ (= \mathcal{R}_{\Psi^p}(\rho_{\mathcal{F}^p}(\boldsymbol{\theta}))) = \mathbb{E}_{Q^{\mathbf{Z}}}\Psi^p(\rho_{\mathcal{F}^p}(\boldsymbol{\theta}), Z)\,,$$

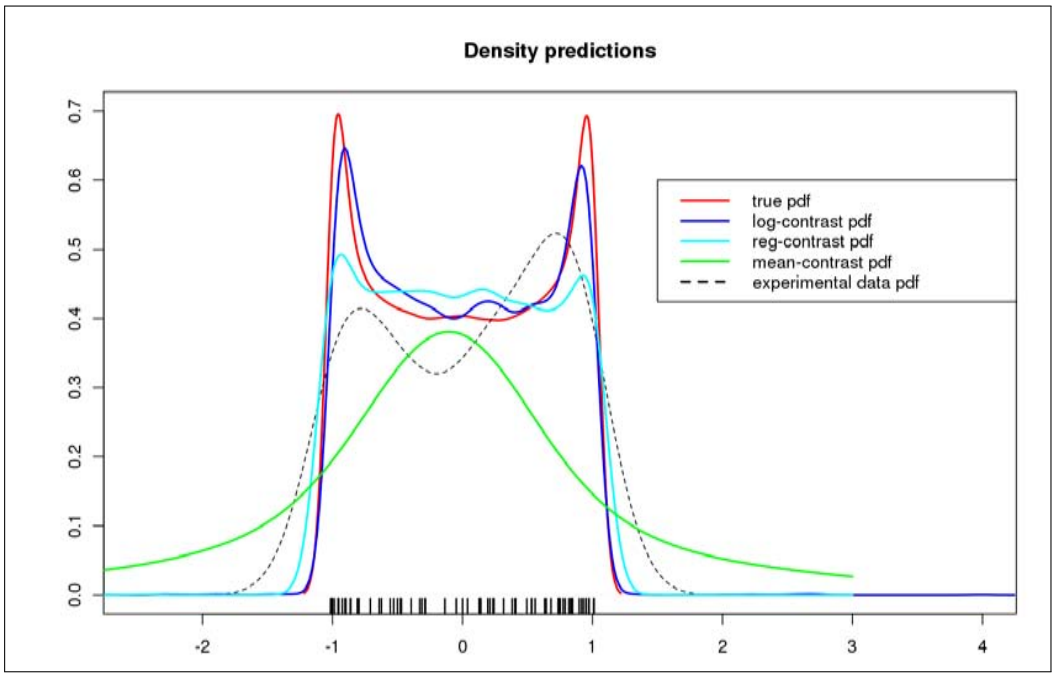


FIGURE 4.2 – Predictions of f from contrasts $\Psi^p = \Psi_{\log}$ (in blue), Ψ_{reg} (in cyan) and Ψ_{mean} (in green)

the final purpose is to "compare" the risks

$$\mathcal{R}_{\Psi^p}(\widehat{\boldsymbol{\theta}}_{\Psi^p}) \quad \text{and} \quad \mathcal{R}_{\Psi^p}(\widehat{\boldsymbol{\theta}}_{\Psi}) \,. \tag{4.7}$$

As a matter of fact, the challenge lies mainly in the *non-asymptotic* study of the latter quantities.

4.4.2 Non-asymptotic context

Let us recall that for any $\mathcal{F}$-contrast Ψ

$$\boldsymbol{\theta}_{\Psi} \in \operatorname*{Argmin}_{\boldsymbol{\theta}\in\Theta} \mathcal{R}_{\Psi}(\boldsymbol{\theta}) = \operatorname*{Argmin}_{\boldsymbol{\theta}\in\Theta} \mathbb{E}_{Q^z} \Psi(\rho_{\mathcal{F}}(\boldsymbol{\theta}), Z) \,, \tag{4.8}$$

and denote by $\widehat{\boldsymbol{\theta}}_{\Psi}$ a Ψ-estimator of $\boldsymbol{\theta}_{\Psi}$ (taking either the form (2.38) or (2.39)). We say that $\widehat{\boldsymbol{\theta}}_{\Psi}(n)$ is convergent if in some sense

$$\widehat{\boldsymbol{\theta}}_{\Psi}(n) \underset{n\to+\infty}{\longrightarrow} \boldsymbol{\theta}_{\Psi} \,.$$

Let us give the following obvious lemma.

Lemma 4.4.1. ***Sub-optimality of cross procedure***
Let ρ^p be some feature of interest of Y that belongs to a feature space $\mathcal{F}^p$, and denote by Ψ^p a $\mathcal{F}^p$-contrast. It holds that for all $\mathcal{F}$-contrast Ψ

$$\mathcal{R}_{\Psi^p}(\boldsymbol{\theta}_{\Psi^p}) \leq \mathcal{R}_{\Psi^p}(\boldsymbol{\theta}_{\Psi}) \,. \tag{4.9}$$

Proof. The proof is clear by definition of

$$\boldsymbol{\theta}_{\Psi^p} \in \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{Argmin}}\, \mathcal{R}_{\Psi^p}(\boldsymbol{\theta})\,.$$

Then for any parameters $\boldsymbol{\theta} \in \Theta$

$$\mathcal{R}_{\Psi^p}(\boldsymbol{\theta}_{\Psi^p}) \leq \mathcal{R}_{\Psi^p}(\boldsymbol{\theta})\,,$$

and in particular for some parameter $\boldsymbol{\theta}_\Psi \in \Theta$

$$\mathcal{R}_{\Psi^p}(\boldsymbol{\theta}_{\Psi^p}) \leq \mathcal{R}_{\Psi^p}(\boldsymbol{\theta}_\Psi)\,.$$

□

Hence, if the estimators $\widehat{\boldsymbol{\theta}}_{\Psi^p} = \widehat{\boldsymbol{\theta}}_{\Psi^p}(n)$ and $\widehat{\boldsymbol{\theta}}_\Psi = \widehat{\boldsymbol{\theta}}_\Psi(n)$ are both "convergent", one can hope that for sufficiently large n

$$\mathcal{R}_{\Psi^p}(\widehat{\boldsymbol{\theta}}_{\Psi^p}) \leq \mathcal{R}_{\Psi^p}(\widehat{\boldsymbol{\theta}}_\Psi)\,.$$

The asymptotic case would minimize the interest of our approach.
In practice, it is very difficult to have a *large* amount of data, due principally to costs and computing reasons. That's why we focus first on a non asymptotic approach, making sense to study the risks

$$\mathcal{R}_{\Psi^p}(\widehat{\boldsymbol{\theta}}_{\Psi^p}) \quad \text{and} \quad \mathcal{R}_{\Psi^p}(\widehat{\boldsymbol{\theta}}_\Psi)\,.$$

In some words, we attempt to formalize what can be done in practice. For example, one may be interested by some particular feature of Y belonging to $\mathcal{F}^p$ (for instance a set of density functions), and the estimation procedure can be "foreign" to the feature of interest (for instance, we compute a Ψ-estimator from a reg-contrast $\Psi = \Psi_{reg}$). See the numerical examples in Section 4.6. It would be natural to consider a contrast defined on the same feature space as the one of the wanted feature of interest. However, for computation limitations or other practical reasons, one may choose a cross procedure. Moreover, we will see that a classical procedure (i.e non cross) can provide overfitting problems (Section 4.15). That's why we propose to work on the quantification of the effect of such practice.

An interesting question would be what can happen when the (unknown) parameters $\boldsymbol{\theta}_\Psi$ and $\boldsymbol{\theta}_{\Psi^p}$ are replaced by their estimators $\widehat{\boldsymbol{\theta}}_\Psi$ and $\widehat{\boldsymbol{\theta}}_{\Psi^p}$, respectively, as mentioned in (4.4.1). In other words, do we have

$$\mathcal{R}_{\Psi^p}(\widehat{\boldsymbol{\theta}}_{\Psi^p}) \leq \mathcal{R}_{\Psi^p}(\widehat{\boldsymbol{\theta}}_\Psi) \quad \textit{(a.s or with high probability)}$$

or

$$\mathbb{E}\,\mathcal{R}_{\Psi^p}(\widehat{\boldsymbol{\theta}}_{\Psi^p}) \leq \mathbb{E}\,\mathcal{R}_{\Psi^p}(\widehat{\boldsymbol{\theta}}_\Psi)\,, \tag{4.10}$$

etc ... ?

This kind of result is very difficult to obtain in general. However, in some cases we can say something about.
In Section 4.10 we give some theoretical elements in order to investigate the behavior of the difference $\mathcal{R}_{\Psi^p}(\widehat{\boldsymbol{\theta}}_\Psi) - \mathcal{R}_{\Psi^p}(\widehat{\boldsymbol{\theta}}_{\Psi^p})$. Before, we present an academic example for which we prove an inequality of the form (4.10). Then, we give numerical examples through an academic setting. Finally, we tackle the overfitting problem.

Remark 4.4.4. In the case of a penalized procedure, one can hope to obtain

$$\mathcal{R}_{\Psi^p}(\widehat{\boldsymbol{\theta}}_{\Psi^p_{\text{pen}}}) - \mathcal{R}_{\Psi^p}(\widehat{\boldsymbol{\theta}}_{\Psi^p}) \leq 0\,,$$

which means that the cross procedure is better than the classical one. [5]

4.5 Academic example : mean prediction

Suppose that one may want to predict the mean (expectation) of the random phenomenon Y, i.e

$$\rho^p = \mathbb{E}_Q(Y) \in \mathcal{F}^p\,.$$

Suppose that we have at disposal a sample $(\mathbf{X}_1, Y_1), ..., (\mathbf{X}_n, Y_n)$ i.i.d from $Q^{\mathbf{z}}$ ($Z = (\mathbf{X}, Y) \sim Q^{\mathbf{z}}$). Next, let us consider a model

$$h\,:\,(\mathbf{X}, \boldsymbol{\theta}) \mapsto h(\mathbf{X}, \boldsymbol{\theta})$$

where $\mathbf{X} \sim P^{\mathbf{x}}$ and $\boldsymbol{\theta} \in \Theta$.
We aim at predicting the (unknown) quantity of interest $\rho^p = \mathbb{E}_Q(Y)$ by a quantity

$$\widehat{\rho^p} = \mathbb{E}_{P^{\mathbf{x}}}(h(\mathbf{X}, \widehat{\boldsymbol{\theta}}))$$

for some parameter $\widehat{\boldsymbol{\theta}}$.

Let us suppose that h is linear in $\boldsymbol{\theta} \in \Theta$

$$h(\mathbf{x}, \boldsymbol{\theta}) = \Phi(\mathbf{x}) \cdot \boldsymbol{\theta}\,,$$

where

$$\Phi(\mathbf{x}) := \big(1, \phi_1(\mathbf{x}), ..., \phi_p(\mathbf{x})\big)$$

is an orthonormal basis related to the probability measure $P^{\mathbf{x}}$, and suppose that it exists $\boldsymbol{\theta}^* \in \Theta$ such that

$$Y_i = \Phi(\mathbf{X}_i) \cdot \boldsymbol{\theta}^* + \varepsilon_i\,, \tag{4.11}$$

where the $\varepsilon_i{}'s$ are i.i.d random variables, independent of the $\mathbf{X}_i{}'s$, with zero mean and variance σ^2.

Let $\Psi^p = \Psi_{mean}$ be a $\mathcal{F}^p$-contrast and consider the following prediction risk

$$\forall \boldsymbol{\theta} \in \Theta\,, \quad \mathcal{R}_{\Psi^p}(\boldsymbol{\theta}) = (\rho^p - \mathbf{B} \cdot \boldsymbol{\theta})^2\,,$$

where $\mathbf{B} = \mathbb{E}_{P^{\mathbf{x}}}(\Phi(\mathbf{X}))$ (recall that $\rho^p = \mathbb{E}_Q(Y)$). Notice that by (4.11), we have $\rho^p = \mathbb{E}_Q(Y) = \mathbf{B} \cdot \boldsymbol{\theta}^*$.

Now, for the parameter estimation, let us consider two contrasts : a mean-contrast $\Psi_{mean} = \Psi^p$ which leads to a classical prediction (because we want to predict a mean), and a reg-contrast Ψ_{reg} (cross prediction). Notice that the feature

$$\rho_{\mathcal{F}^p}(\boldsymbol{\theta}) = \mathbb{E}_{P^{\mathbf{x}}}(h(\mathbf{X}, \boldsymbol{\theta})) = \mathbf{B} \cdot \boldsymbol{\theta}$$

5. Voir l'exemple de la ridge regression au Chapitre 2, Proposition 2.7.1.

is known in this case, thus we don't need simulation.
We have the two estimators

$$\widehat{\boldsymbol{\theta}}_{\Psi^p} = \underset{\boldsymbol{\theta}\in\Theta}{\text{Argmin}} \sum_{i=1}^{n} \left(Y_i - \mathbf{B}\,\boldsymbol{\theta}\right)^2 ,$$

and

$$\widehat{\boldsymbol{\theta}}_{\Psi_{reg}} = \underset{\boldsymbol{\theta}\in\Theta}{\text{Argmin}} \sum_{i=1}^{n} \left(Y_i - \Phi(\mathbf{X}_i)\cdot\boldsymbol{\theta}\right)^2 .$$

In this specific setting, we propose to "compare" the two random quantities $\mathcal{R}_{\Psi^p}(\widehat{\boldsymbol{\theta}}_{\Psi^p})$ and $\mathcal{R}_{\Psi^p}(\widehat{\boldsymbol{\theta}}_{\Psi_{reg}})$.

We prove the following proposition.

Proposition 4.5.1. *Let us consider the previous setting. It exists some constant $L > 0$ such that if $\dfrac{\|\boldsymbol{\theta}^*\|}{\sigma} \leq L$, then*

$$\mathbb{E}_{Y_{1..n}}\left(\mathcal{R}_{\Psi^p}(\widehat{\boldsymbol{\theta}}_{\Psi^p})\right) \leq \mathbb{E}_{(\mathbf{X}_i,Y_i)_{1..n}}\left(\mathcal{R}_{\Psi^p}(\widehat{\boldsymbol{\theta}}_{\Psi_{reg}})\right) .$$

The proof is given in Section 4.8.

The behavior of the difference

$$\mathcal{R}_{\Psi^p}(\widehat{\boldsymbol{\theta}}_{\Psi_{reg}}) - \mathcal{R}_{\Psi^p}(\widehat{\boldsymbol{\theta}}_{\Psi^p})$$

or more generally for any contrast Ψ

$$\mathcal{R}_{\Psi^p}(\widehat{\boldsymbol{\theta}}_{\Psi}) - \mathcal{R}_{\Psi^p}(\widehat{\boldsymbol{\theta}}_{\Psi^p})$$

is of interest in that it gives some robustness indications of the modeling (through parameter estimation) in relation to the wanted prediction. Intuitively, it appears that the difference above depends on some "distance between contrasts" Ψ^p and Ψ, and depends also on the performance of the estimators $\widehat{\boldsymbol{\theta}}_{\Psi^p}$ and $\widehat{\boldsymbol{\theta}}_{\Psi}$.
Theoretical aspects will be addressed in Section 4.10.

Now, let us give some numerical examples.

4.6 Numerical examples

In this section, we consider the following phenomenon

$$Y = \sin(X) + 0.01\,\varepsilon\,, \quad X,\,\varepsilon \sim \mathcal{N}(0,1) \quad \text{independents}\,. \tag{4.12}$$

We suppose for a moment that the writing above is unknown in order to apply our methodology. We will use it to compare the results.
The goal is to predict some features $\rho^p \in \mathcal{F}^p$ (unknown) of the random phenomenon Y (with probability measure Q). For this, let us assume that we have at disposal a sample $(X_1, Y_1), ..., (X_n, Y_n)$ and a numerical model

$$\begin{array}{rccl} h\,:\,\mathcal{X}\times\Theta & \longrightarrow & \mathcal{Y} \\ (x,\boldsymbol{\theta}) & \longmapsto & h(x,\boldsymbol{\theta}) = \theta_1 + x\,\theta_2 + x^3\,\theta_3\,. \end{array}$$

The uncertainty is modeled by equipping the space $\mathcal{X} \subset \mathbb{R}$ with a standard gaussian distribution, noted $P^{\mathbf{x}}$. That yields the stochastic model

$$\begin{aligned} h : (\mathcal{X}, \mathcal{B}, P^{\mathbf{x}}) \times \Theta &\longrightarrow \mathcal{Y} \\ (X, \boldsymbol{\theta}) &\longmapsto h(X, \boldsymbol{\theta}) = \theta_1 + X\,\theta_2 + X^3\,\theta_3\,. \end{aligned} \tag{4.13}$$

The model h is seen as a black-box function (see figure (4.1)).
We generate a simulation set $\mathcal{X}^m = \{X'_1, ..., X'_m\}$ ($X'_j \sim P^{\mathbf{x}}$) and form the model feature

$$\rho^m_{\mathcal{F}^p}(\boldsymbol{\theta}) := \frac{1}{m} \sum_{j=1}^{m} \widetilde{\rho}_{\mathcal{F}^p}(h(X'_j, \boldsymbol{\theta})) \in \mathcal{F}^p\,, \tag{4.14}$$

where the weight function $\widetilde{\rho}_{\mathcal{F}^p}$ depends on the considered feature ρ^p.
Then, any prediction of ρ^p will be of the form

$$\widehat{\rho^p} = \rho^m_{\mathcal{F}^p}(\widehat{\boldsymbol{\theta}}_\Psi)$$

where $\widehat{\boldsymbol{\theta}}_\Psi = (\widehat{\theta}_1, \widehat{\theta}_2, \widehat{\theta}_3)^T$ is some parameter resulting of some estimation procedure based on some contrast Ψ.
Let us consider the three procedures

$$\begin{aligned} \widehat{\boldsymbol{\theta}}_{\Psi_{\log}} &= \underset{\boldsymbol{\theta}\in\Theta}{\text{Argmin}} - \sum_{i=1}^{n} \log \left(\sum_{j=1}^{m} K_{b_\theta}(\, h(X'_j, \boldsymbol{\theta}) - Y_i) \right)\,, \\ \widehat{\boldsymbol{\theta}}_{\Psi_{reg}} &= \underset{\boldsymbol{\theta}\in\Theta}{\text{Argmin}} \sum_{i=1}^{n} \left(Y_i - h(X_i, \boldsymbol{\theta})\right)^2 \\ \widehat{\boldsymbol{\theta}}_{\Psi_{mean}} &= \underset{\boldsymbol{\theta}\in\Theta}{\text{Argmin}} \sum_{i=1}^{n} \left(\sum_{j=1}^{m} (Y_i - h(X'_j, \boldsymbol{\theta})) \right)^2 \end{aligned}$$

obtained from log-contrast, reg-contrast and mean-contrast, respectively.
For the numerical applications, we take $n = 70$ and $m = 10000$. We present in Table 4.1 the values of $\widehat{\boldsymbol{\theta}}_{\Psi_{\log}}$, $\widehat{\boldsymbol{\theta}}_{\Psi_{reg}}$ and $\widehat{\boldsymbol{\theta}}_{\Psi_{mean}}$.

Contrast	$\widehat{\boldsymbol{\theta}}_\Psi$
$\Psi = \Psi_{\log}$	$(\mathbf{0.0057}, \mathbf{1.025}, \mathbf{-0.163})$
$\Psi = \Psi_{reg}$	$(\mathbf{-0.0049}, \mathbf{0.9259}, \mathbf{-0.1048})$
$\Psi = \Psi_{mean}$	$(\mathbf{-0.0924}, \mathbf{0.6607}, \mathbf{0.5965})$

TABLE 4.1 – Values of Ψ-estimators for different contrasts Ψ.

Now, let us see the performances of some predictions using these different estimators.

4.6.1 Density probability prediction

Here, we aim at predicting the density of the variable Y.

We choose the weight function

$$\widetilde{\rho}_{\mathcal{F}^p}(y)(\cdot) = K_{b_\theta}(\,\cdot - y)\,,$$

where $K_{b_\theta}(\cdot - y) = \frac{1}{b_\theta} K(\frac{\cdot - y}{b_\theta})$ with a gaussian kernel $K()$. The bandwidth b_θ is computed from the sample $h(\mathbf{X}_1, \boldsymbol{\theta}), ..., h(\mathbf{X}_m, \boldsymbol{\theta})$ by the Silverman rule of thumb.
Then, any prediction of the density $f = \rho^p$ of Y will be of the form

$$\widehat{f} = \rho^m_{\mathcal{F}^p}(\widehat{\boldsymbol{\theta}}_\Psi)$$

for either $\widehat{\boldsymbol{\theta}}_\Psi = \widehat{\boldsymbol{\theta}}_{\Psi_{\log}}, \widehat{\boldsymbol{\theta}}_{\Psi_{reg}}$ or $\widehat{\boldsymbol{\theta}}_{\Psi_{\log}}$.

Let us consider the log-contrast $\Psi_{\log} = \Psi^p$ as a $\mathcal{F}^p$-contrast. Define the risk

$$\begin{aligned} \mathcal{R}_{\Psi^p}(\boldsymbol{\theta}) &= \mathbb{E}_Q \Psi^p(\rho_{\mathcal{F}^p}(\boldsymbol{\theta}), Y) \\ &= -\int_{\mathcal{Y}} \log(\rho_{\mathcal{F}^p}(\boldsymbol{\theta})(y))\, f(y)\, dy \end{aligned}$$

where $\rho_{\mathcal{F}^p}(\boldsymbol{\theta})$ is the density distribution of the random variable $h(X, \boldsymbol{\theta})$ and f the density of Y.

Notice that the first parameter $\widehat{\boldsymbol{\theta}}_{\Psi_{\log}} = \widehat{\boldsymbol{\theta}}_{\Psi^p}$ leads to a classical prediction whereas the two others lead to cross ones. It appears natural to compare the values of the prediction risk $\mathcal{R}_{\Psi^p}$ for each parameter estimation. We compute in Table 4.2 the different values of the risk for the three considered contrast procedures : $\Psi_{\log}$, Ψ_{reg} and Ψ_{mean} . We see that the classical procedure provides the smallest risk prediction value.

Remark 4.6.1. The risk values $\mathcal{R}_{\Psi^p}(\widehat{\boldsymbol{\theta}}_\Psi)$ may be seen up to a constant term. In fact, we should subtract the minimal risk $\mathcal{R}_{\Psi^p}(\boldsymbol{\theta}_{\Psi^p})$ in order to have a quantity close to a "distance".

	$\widehat{\boldsymbol{\theta}}_\Psi$	$\mathcal{R}_{\Psi^p}(\widehat{\boldsymbol{\theta}}_\Psi)$
$\Psi = \Psi^p$	$(\mathbf{0.0057}, \mathbf{1.025}, \mathbf{-0.163})$	**0.77**
$\Psi = \Psi_{reg}$	$(\mathbf{-0.0049}, \mathbf{0.9259}, \mathbf{-0.1048})$	**0.81**
$\Psi = \Psi_{mean}$	$(\mathbf{-0.0924}, \mathbf{0.6607}, \mathbf{0.5965})$	**1.37**

TABLE 4.2 – Risk values for density prediction at different parameters, with $\Psi^p = \Psi_{log}$.

To illustrate the results in the Table 4.2, we present in the figure (4.3) the different predictions $\widehat{f} = \rho^m_{\mathcal{F}^p}(\widehat{\boldsymbol{\theta}}_\Psi)$ of the probability density function f, using classical and cross procedures.

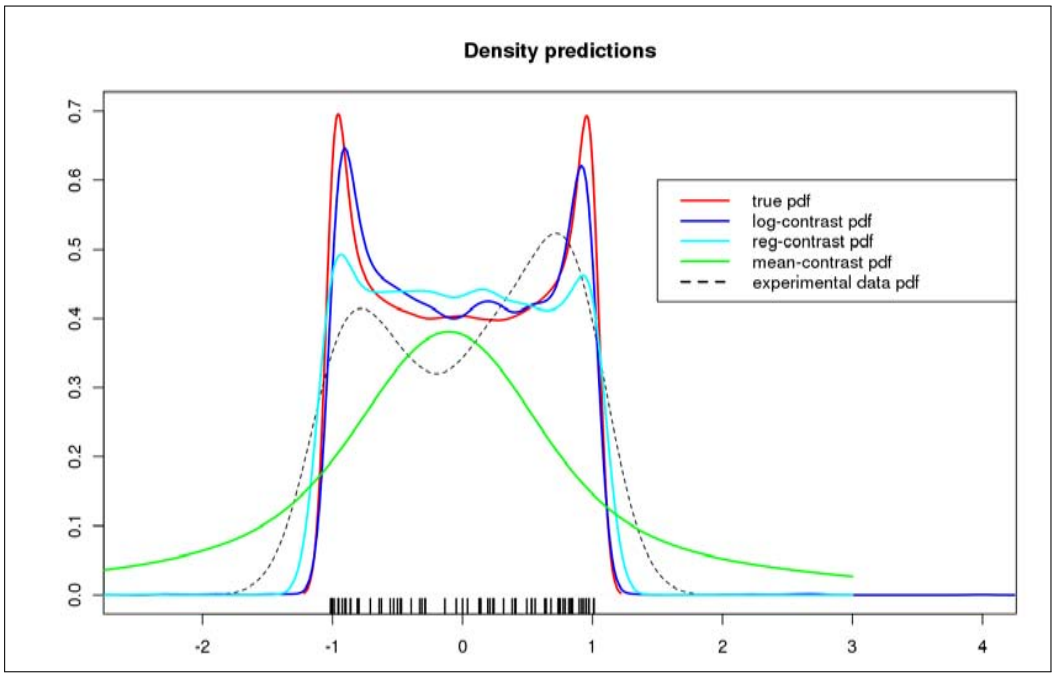


FIGURE 4.3 – Predictions of f from different contrasts

It seems that there is some kind of "ranking" of Ψ-estimators with respect to the nature of the prediction. As it seems natural, in this case the classical procedure provides a good prediction contrary to the procedure using the Ψ_{mean}-estimator. It can be understood as the fact that the Ψ_{mean}-estimator provides "local" information (on the mean), which may not be enough in order to describe the whole behavior of Y.

4.6.2 Conditional expectation prediction

Here, the goal is to predict

$$\rho^p \,:\, x \longmapsto \mathbb{E}(Y/X = x)\,.$$

The prediction is of the form

$$\widehat{\rho^p} \,:\, x \longmapsto h(x, \widehat{\boldsymbol{\theta}}_\Psi)\,,$$

for some Ψ-estimator $\widehat{\boldsymbol{\theta}}_\Psi$.
Let us consider the reg-contrast $\Psi_{reg} = \Psi^p$ as a $\mathcal{F}^p$-contrast. Define the risk

$$\begin{aligned}\mathcal{R}_{\Psi^p}(\boldsymbol{\theta}) &= \mathbb{E}_{Q^{\mathbf{z}}}\Psi^p(\rho_{\mathcal{F}^p}(\boldsymbol{\theta}), Z)\\ &= \int_{\mathcal{X}\times\mathcal{Y}} (y - \rho_{\mathcal{F}^p}(\boldsymbol{\theta})(x))\; Q^{\mathbf{z}}(dx, dy)\, dy\end{aligned}$$

where $\rho_{\mathcal{F}^p}(\boldsymbol{\theta})(x) = h(x, \boldsymbol{\theta})$.

	$\widehat{\boldsymbol{\theta}}_\Psi$	$\mathcal{R}_{\Psi^p}(\widehat{\boldsymbol{\theta}}_\Psi)$
$\Psi = \Psi^p$	$(\mathbf{-0.0049, 0.9259, -0.1048})$	**0.064**
$\Psi = \Psi_{\log}$	$(\mathbf{0.0057, 1.025, -0.163})$	**0.36**
$\Psi = \Psi_{mean}$	$(\mathbf{-0.0924, 0.6607, 0.5965})$	**6.18**

TABLE 4.3 – Risk values for the conditional expectation prediction from different Ψ-estimators, with $\Psi^p = \Psi_{reg}$.

The classical procedure using regression parameter gives the best result. The cross procedure with the Ψ_{log}-estimator follows with a poor performance on the "extreme" values, see Figure 4.4. As the case of density prediction, the model prediction using the Ψ_{mean}-estimator fails to predict the wanted quantity $x \mapsto \sin(x)$.

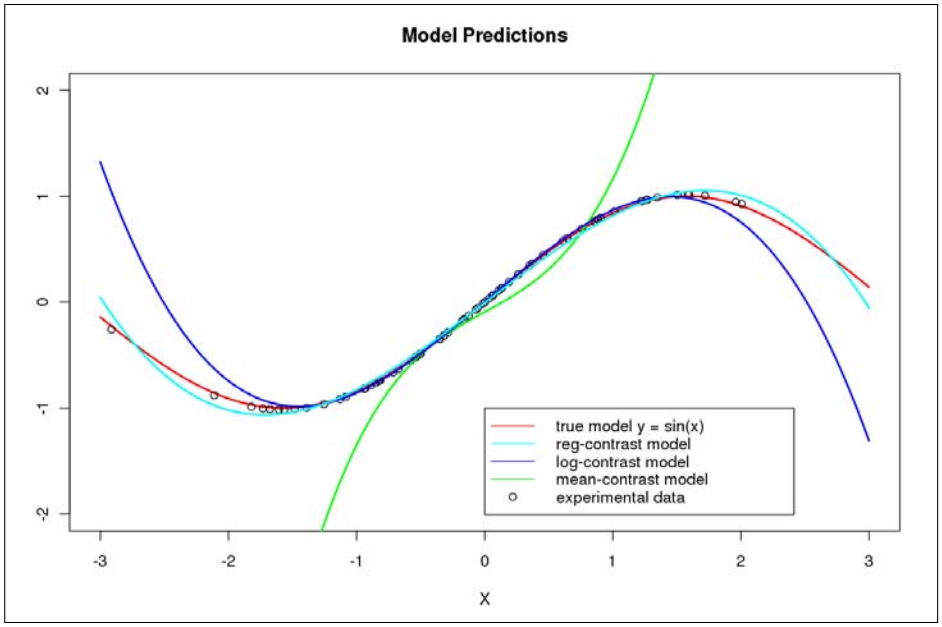


FIGURE 4.4 – Predictions of $x \mapsto \sin(x)$ from different contrasts

In the two previous examples, classical prediction procedures, i.e using the Ψ^p-estimator $\widehat{\boldsymbol{\theta}}_{\Psi^p}$, provide the best predictions.

However, it may happen that such procedures are not so performant because of *overfitting phenomenon*.

4.7 Overfitting phenomenon with exceeding probability prediction

The overfitting phenomenon is well known in statistical learning where one may not choose a large model class in order to avoid fitting data and have poor generalizing performance. In our context, we are interested in predicting a feature $\rho^p \in \mathcal{F}^p$ of the random variable Y. We showed the existence of two ways of prediction :

– Classical procedure : the estimation procedure is based on a $\mathcal{F}^p$-contrast
– Cross procedure : the estimation procedure is based on any $\mathcal{F}$-contrast other than a $\mathcal{F}^p$ one.

We also showed that classical procedures are better than cross ones in some cases (see preceding numerical examples).
However, classical procedures may lead to *overfitting* as we will see through the exceeding probability prediction example.

Let us recall that

$$\widehat{\boldsymbol{\theta}}_{\Psi^p} = \underset{\boldsymbol{\theta}\in\Theta}{\text{Argmin}} \sum_{i=1}^{n} \Psi\left(\sum_{j=1}^{m} \widetilde{\rho}_{\mathcal{F}^p}(h(\mathbf{X}'_j, \boldsymbol{\theta}))\,,\, Y_i\right)$$

which we roughly rewrite

$$\widehat{\boldsymbol{\theta}}_{\Psi^p} = \underset{\boldsymbol{\theta}\in\Theta}{\text{Argmin}}\ \mathcal{D}(\rho^n, \rho^m_{\mathcal{F}^p}(\boldsymbol{\theta}))\,,$$

where $\mathcal{D}(\cdot,\cdot)$ can be seen as a "distance" on $\mathcal{F}^p \times \mathcal{F}^p$, and ρ^n is an empirical prediction of ρ^p based on $Y_1, ..., Y_n$. The (classical) prediction is given by (4.4)

$$\widehat{\rho^p} = \rho^m_{\mathcal{F}^p}(\widehat{\boldsymbol{\theta}}_{\Psi^p})\,.$$

Next, denote by $F = \{\rho^m_{\mathcal{F}^p}(\boldsymbol{\theta})\,,\ \boldsymbol{\theta}\in\Theta\}$ the (approximated) model feature space where

$$\widehat{\rho^p} \in F \subset \mathcal{F}^p\,.$$

The overfitting occurs when for instance we have

$$\rho^n \in F\,, \tag{4.15}$$

this may be the consequence of considering a "large" model F. In this case, by definition of $\widehat{\boldsymbol{\theta}}_{\Psi^p}$, the prediction $\widehat{\rho^p}$ satisfies $\widehat{\rho^p} = \rho^n$.
In general, it's the reason why it is not recommended to choose a modeling providing a "large" model F, principally in order to avoid that $\rho^n \in F$. In some sense, the penalization procedures allow in the same time to "reduce" the model F in which we seek the prediction. An illustration of the penalization aspect will be addressed in Section 4.9.
Now, we aim at predicting $\rho^p = \mathbb{P}(Y > 0.5)$ in the same setting as in Subsection 4.6.

In this case, the feature space $\mathcal{F}^p$ is $[0,1]$. Let us consider the $\mathcal{F}^p$-contrast $\Psi_{prob} = \Psi^p$ defined as follows

$$\Psi_{prob}(y,\rho) = (\mathbb{1}_{y>0.5} - \rho)^2\,.$$

The associated risk is

$$\mathcal{R}_{\Psi_{prob}}(\rho) = \mathbb{E}_Q\left(\mathbb{1}_{Y>0.5} - \rho\right)^2\,. \tag{4.16}$$

One can easily check that $\rho^p = \text{Argmin}_{\rho\in\mathcal{F}^p}\,\mathcal{R}_{\Psi_{prob}}(\rho)$ is unique.
Considering that $\rho = \rho_{\mathcal{F}^p}(\boldsymbol{\theta}) = \mathbb{P}(h(X,\boldsymbol{\theta}) > 0.5)$, provides the (parametric) prediction risk

$$\begin{aligned}\mathcal{R}_{\Psi^p}(\boldsymbol{\theta}) &= \mathbb{E}_Q \Psi^p(\rho_{\mathcal{F}^p}(\boldsymbol{\theta}), Y)\\ &= \int_{\mathcal{Y}} \left(\mathbb{1}_{y>0.5} - \rho_{\mathcal{F}^p}(\boldsymbol{\theta})\right)^2 f(y)\,dy\,.\end{aligned}$$

The model feature is given by

$$\forall\,\boldsymbol{\theta}\in\Theta\,,\quad \rho^m_{\mathcal{F}^p}(\boldsymbol{\theta}) = \frac{1}{m}\sum_{j=1}^m \mathbb{1}_{h(X'_j,\boldsymbol{\theta})>s}\,.$$

Then, a classical procedure provides the prediction

$$\widehat{\rho^p} = \rho^m_{\mathcal{F}^p}(\widehat{\boldsymbol{\theta}}_{\Psi_{prob}}) = \frac{1}{m}\sum_{j=1}^m \mathbb{1}_{h(X'_j,\widehat{\boldsymbol{\theta}}_{\Psi_{prob}})>s}$$

where $\widehat{\boldsymbol{\theta}}_{\Psi_{prob}}$ is the Ψ_{prob}-estimator

$$\widehat{\boldsymbol{\theta}}_{\Psi_{prob}} = \underset{\boldsymbol{\theta}\in\Theta}{\text{Argmin}}\ \frac{1}{n}\sum_{i=1}^n \left(\mathbb{1}_{Y_i>0.5} - \frac{1}{m}\sum_{j=1}^m \mathbb{1}_{h(X'_j,\boldsymbol{\theta})>0.5}\right)^2\,. \tag{4.17}$$

We compute this estimator, and we find

$$\widehat{\boldsymbol{\theta}}_{\Psi_{prob}} = (\mathbf{0.238}, \mathbf{0.383}, \mathbf{0.876}) .$$

In Table 4.4, we compute the prediction risks $\mathcal{R}_{\Psi^p}(\widehat{\boldsymbol{\theta}}_\Psi)$ for different Ψ-estimators $\widehat{\boldsymbol{\theta}}_\Psi$. In fact, as the risk function $\boldsymbol{\theta} \mapsto \mathcal{R}_{\Psi^p}(\boldsymbol{\theta})$ takes small values, we compute relative Ψ_p-excess risks for each $\widehat{\boldsymbol{\theta}}_\Psi$, defined as follows

$$r(\widehat{\boldsymbol{\theta}}_\Psi) = \frac{\mathcal{R}_{\Psi^p}(\widehat{\boldsymbol{\theta}}_\Psi) - \mathcal{R}^\Theta_{\Psi^p}}{\mathcal{R}^\Theta_{\Psi^p}} , \tag{4.18}$$

where $\mathcal{R}^\Theta_{\Psi^p} = \inf_{\boldsymbol{\theta}\in\Theta} \mathcal{R}_{\Psi^p}(\boldsymbol{\theta})$.

	$\widehat{\boldsymbol{\theta}}_\Psi$	$r(\widehat{\boldsymbol{\theta}}_\Psi)$
$\Psi = \Psi_{\log}$	$(\mathbf{0.0057}, \mathbf{1.025}, -\mathbf{0.163})$	**1.3**%
$\Psi = \Psi_{reg}$	$(-\mathbf{0.0049}, \mathbf{0.9259}, -\mathbf{0.1048})$	**1.5**%
$\Psi = \Psi_{prob}$	$(\mathbf{0.238}, \mathbf{0.383}, \mathbf{0.876})$	**1.7**%
$\Psi = \Psi_{mean}$	$(-\mathbf{0.0924}, \mathbf{0.6607}, \mathbf{0.5965})$	**2.1**%

TABLE 4.4 – Relative Ψ_p-excess risk of $\widehat{\boldsymbol{\theta}}_\Psi$ with $\Psi^p = \Psi_{prob}$.

The log-contrast gives a good result although it is a cross procedure, see Figure 4.5 (a). The classical procedure is in the "third" position, after the cross procedure given by the reg-contrast. Thus, a cross procedure may not have always a negative effect on a prediction, provided that the learning procedure (which computes $\widehat{\boldsymbol{\theta}}_\Psi$) and the prediction are not too foreign (the mean-contrast doesn't provide a good prediction).
As in the previous examples, there is some "ranking" of contrasts related to the quantity of interest one wants to predict.

In particular, the probability estimation by Ψ_{prob} contrast, see Figure 4.5 (c), gives the same result as the empirical probability estimation

$$\rho^n = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{Y_i>0.5} .$$

Indeed (4.17) can also be written

$$\widehat{\boldsymbol{\theta}}_{\Psi_{prob}} = \underset{\boldsymbol{\theta}\in\Theta}{\text{Argmin}} \left(\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{Y_i>0.5} - \frac{1}{m}\sum_{j=1}^m \mathbb{1}_{h(X'_j,\boldsymbol{\theta})>0.5} \right)^2 ,$$

and with such criterion function, it's not surprising that $\frac{1}{m}\sum_{j=1}^m \mathbb{1}_{h(X'_j,\widehat{\boldsymbol{\theta}}_{\Psi_{prob}})>0.5} = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{Y_i>0.5}$.

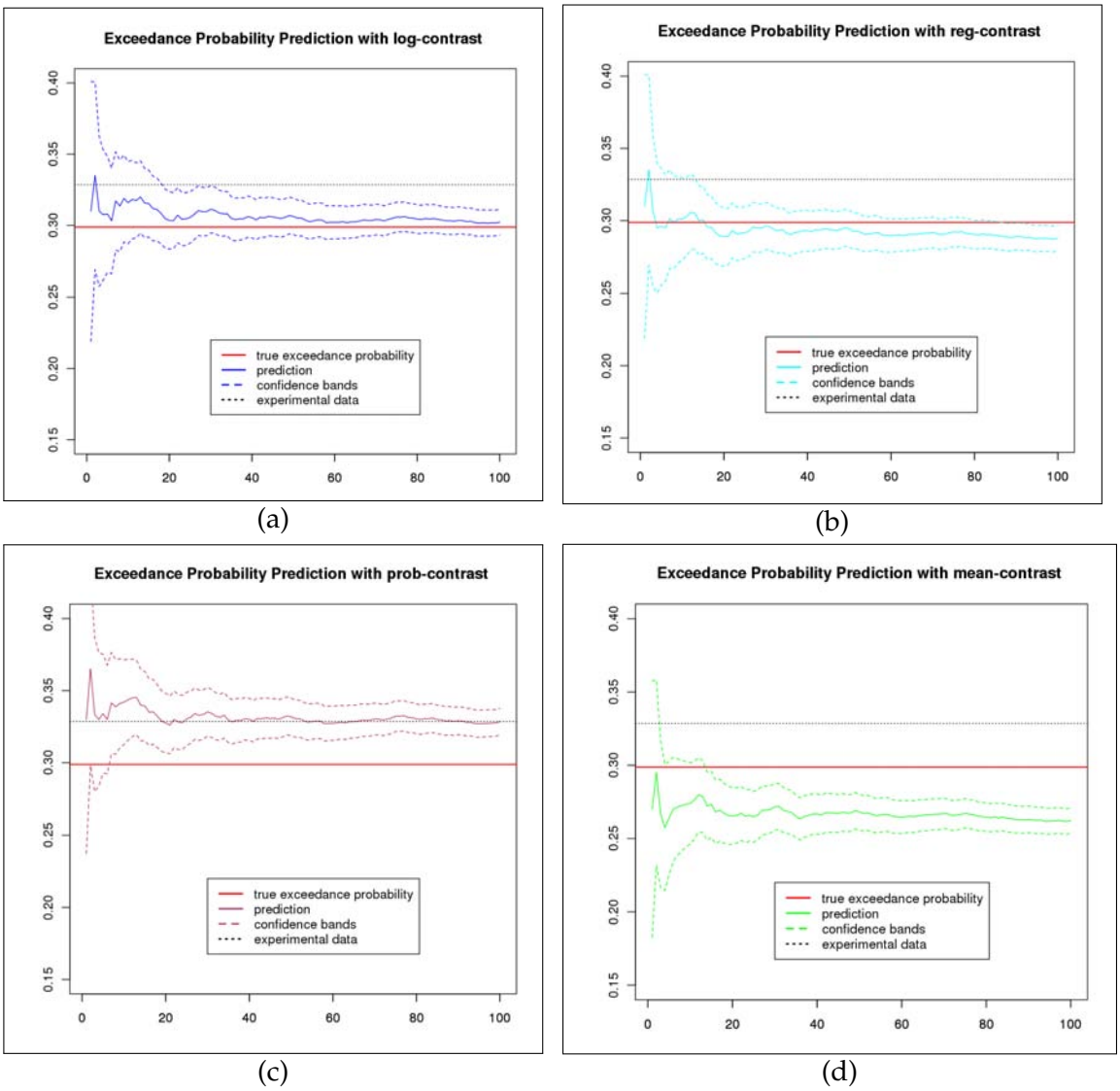


FIGURE 4.5 – Predictions of $\rho^p = \mathbb{P}(Y > 0.5)$ from four different contrasts. The graphs represent the convergence to the prediction $\frac{1}{m}\sum_{j=1}^{m} 1_{h(X'_j,\widehat{\boldsymbol{\theta}}_\Psi)>0.5}$ by plotting $\frac{1}{100\,m'}\sum_{j=1}^{100\,m'} 1_{h(X'_j,\widehat{\boldsymbol{\theta}}_\Psi)>0.5}$ for $m' \in \{1,...,100\}$ (increasing blocks).

4.8 Proof of Proposition 4.5.1

Classical prediction.

It is easy to see that

$$\widehat{\boldsymbol{\theta}}_{\Psi^p} = \{\boldsymbol{\theta} \in \Theta\,, \mathbf{B}\,\boldsymbol{\theta} = \frac{1}{n}\sum_i^n Y_i\}\,,$$

so

$$\mathbf{B}\,\widehat{\boldsymbol{\theta}}_{\Psi^p} = \frac{1}{n}\sum_i^n Y_i\,.$$

Then

$$\begin{aligned} \mathcal{R}_{\Psi^p}(\widehat{\boldsymbol{\theta}}_{\Psi^p}) &= \left(\rho^p - \mathbf{B}\cdot\widehat{\boldsymbol{\theta}}_{\Psi^p}\right)^2 \\ &= \left(\frac{1}{n}\sum_{i=1}^{n}(Y_i - \mathbf{B}\,\boldsymbol{\theta}^*)\right)^2 . \end{aligned} \tag{4.19}$$

By using the relation (4.11) and taking the expectation under $Y_1, ..., Y_n$ yields

$$\mathbb{E}_{Y_{1..n}}\left(\mathcal{R}_{\Psi^p}(\widehat{\boldsymbol{\theta}}_{\Psi^p})\right) = \frac{\sigma^2}{n} + \frac{1}{n}(\boldsymbol{\theta}^{*T}\,\mathbb{E}_{\mathbf{X}}(\Phi(X)^T\Phi(X))\,\boldsymbol{\theta}^*)\,. \tag{4.20}$$

Since $\Phi(\mathbf{X}) = (\phi_0(\mathbf{X}), ..., \phi_p(\mathbf{X}))$ ($\phi_0 = 1$) and the ϕ_l's are orthonormal functions w.r.t the distribution of $\mathbf{X}$, then $\mathbb{E}_{\mathbf{X}}\Phi(X)^T\Phi(X) = I_{p+1}$ where I_{p+1} is the identity matrix of size $p+1$. Finally, we obtain

$$\mathbb{E}_{Y_{1..n}}\left(\mathcal{R}_{\Psi^p}(\widehat{\boldsymbol{\theta}}_{\Psi^p})\right) = \frac{\sigma^2}{n} + \frac{\|\boldsymbol{\theta}^*\|^2}{n} = \frac{\sigma^2}{n}\left(1 + \frac{\|\boldsymbol{\theta}^*\|^2}{\sigma^2}\right) . \tag{4.21}$$

Cross procedure.

Let us work conditionally to $\mathbf{X}_1 = \mathbf{x}_1, ..., \mathbf{X}_n = \mathbf{x}_n$. For simplicity, we may write

$$\mathbb{E}_{Y_{1..n}}\left(\mathcal{R}_{\Psi^p}(\widehat{\boldsymbol{\theta}}_{\Psi_{reg}})\right) = \mathbb{E}_{(\mathbf{X}_i,Y_i)_{1..n}}\left(\mathcal{R}_{\Psi^p}(\widehat{\boldsymbol{\theta}}_{\Psi_{reg}})/\mathbf{X}_1 = \mathbf{x}_1, ..., \mathbf{X}_n = \mathbf{x}_n\right) .$$

It is well known that the regression estimator is

$$\widehat{\boldsymbol{\theta}}_{\Psi_{reg}} = \left(\Phi^T\,\Phi\right)^{-1}\Phi^T\,\mathbf{Y}\,,$$

where $\Phi = \begin{pmatrix} \phi_0(\mathbf{x}_1) & . & . & . & \phi_p(\mathbf{x}_1) \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ \phi_0(\mathbf{x}_n) & . & . & . & \phi_p(\mathbf{x}_n) \end{pmatrix}$ and $\mathbf{Y} = (Y_1, ..., Y_n)^T$.

Then, let us remark that

$$\begin{aligned} \mathcal{R}_{\Psi^p}(\widehat{\boldsymbol{\theta}}_{\Psi_{reg}}) &= \left(\rho^p - \mathbf{B}\cdot\widehat{\boldsymbol{\theta}}_{\Psi_{reg}}\right)^2 \\ &= \left(\mathbf{B}\cdot\widehat{\boldsymbol{\theta}}_{\Psi_{reg}} - \mathbb{E}_{Y_{1..n}}\left(\mathbf{B}\cdot\widehat{\boldsymbol{\theta}}_{\Psi_{reg}}\right)\right)^2 + \left(\mathbb{E}_{Y_{1..n}}\left(\mathbf{B}\cdot\widehat{\boldsymbol{\theta}}_{\Psi_{reg}}\right) - \rho^p\right) C \end{aligned}$$

for some random variable C. Since $\mathbb{E}_{Y_{1..n}}\left(\widehat{\boldsymbol{\theta}}_{\Psi_{reg}}\right) = \theta^*$ we have $\mathbb{E}_{Y_{1..n}}\left(\mathbf{B}\cdot\widehat{\boldsymbol{\theta}}_{\Psi_{reg}}\right) = \mathbb{E}_Q(Y) = \rho^p$. It holds that

$$\mathcal{R}_{\Psi^p}(\widehat{\boldsymbol{\theta}}_{\Psi_{reg}}) = \left(\mathbf{B}\cdot\widehat{\boldsymbol{\theta}}_{\Psi_{reg}} - \mathbb{E}_{Y_{1..n}}\left(\mathbf{B}\cdot\widehat{\boldsymbol{\theta}}_{\Psi_{reg}}\right)\right)^2 .$$

Now, taking the expectation under $Y_1, ..., Y_n$ yields

$$\mathbb{E}_{Y_{1..n}}\left(\mathcal{R}_{\Psi^p}(\widehat{\boldsymbol{\theta}}_{\Psi_{reg}})\right) = \mathrm{Var}_{Y_{1..n}}\left(\mathbf{B}\cdot\widehat{\boldsymbol{\theta}}_{\Psi_{reg}}\right)$$

that we develop

$$\begin{aligned}
\mathbb{E}_{Y_{1..n}}\left(\mathcal{R}_{\Psi^p}(\widehat{\boldsymbol{\theta}}_{\Psi_{reg}})\right) &= \mathrm{Var}_{Y_{1..n}}\left(\mathbf{B}\cdot\widehat{\boldsymbol{\theta}}_{\Psi_{reg}}\right)\\
&= \mathrm{Var}_{Y_{1..n}}\left(\mathbf{B}\cdot\left(\Phi^T\Phi\right)^{-1}\Phi^T\mathbf{Y})\right)\\
&= \mathrm{Var}_{\varepsilon_{1..n}}\left(\mathbf{B}\cdot\left(\Phi^T\Phi\right)^{-1}\Phi^T(\Phi\,\boldsymbol{\theta}^*+\mathbf{e})\right)\quad \mathbf{e}=(\varepsilon_1,...,\varepsilon_n)^T\quad (by\ (4.11))\\
&= \underbrace{\mathrm{Var}_{\varepsilon_{1..n}}\left(\mathbf{B}\cdot\left(\Phi^T\Phi\right)^{-1}\Phi^T\Phi\,\boldsymbol{\theta}^*\right)}_{=0}+\mathrm{Var}_{\varepsilon_{1..n}}\left(\mathbf{B}\cdot\left(\Phi^T\Phi\right)^{-1}\Phi^T\mathbf{e}\right)\\
&= \mathrm{Var}_{\varepsilon_{1..n}}\left(\mathbf{B}\cdot\left(\Phi^T\Phi\right)^{-1}\Phi^T\mathbf{e}\right).
\end{aligned}$$

Since the (real valued) random variable $\mathbf{B}\left(\Phi^T\Phi\right)^{-1}\Phi^T\mathbf{e}$ is centered, we have

$$\mathbb{E}_{Y_{1..n}}\left(\mathcal{R}_{\Psi^p}(\widehat{\boldsymbol{\theta}}_{\Psi_{reg}})\right)=\mathbb{E}_{\varepsilon_{1..n}}\left(\left(\mathbf{B}\left(\Phi^T\Phi\right)^{-1}\Phi^T\mathbf{e}\right)^2\right).$$

Let us notice that

$$\left(\mathbf{B}\left(\Phi^T\Phi\right)^{-1}\Phi^T\mathbf{e}\right)^2 = \mathbf{B}\left(\Phi^T\Phi\right)^{-1}\Phi^T\mathbf{e}\,\mathbf{e}^T\Phi\left(\Phi^T\Phi\right)^{-1}\mathbf{B}^T$$

and

$$\mathbb{E}(\mathbf{e}\,\mathbf{e}^T)=\sigma^2\,I_n\,.$$

Then, using the fact that $\phi_0=1$ and $\mathbb{E}_{P^{\mathbf{x}}}(\phi_j(\mathbf{X}))=0$, $j=1,...,p$ leads to

$$\mathbb{E}_{Y_{1..n}}\left(\mathcal{R}_{\Psi^p}(\widehat{\boldsymbol{\theta}}_{\Psi_{reg}})\right)=\sigma^2\,\mathbf{B}\,(\Phi^T\Phi)^{-1}\,\mathbf{B}^T$$

where $\mathbf{B}=(1,0,...,0)$ and

$$\Phi=\begin{pmatrix}1 & . & . & . & \phi_p(\mathbf{x}_1)\\ . & . & . & . & .\\ . & . & . & . & .\\ . & . & . & . & .\\ 1 & . & . & . & \phi_p(\mathbf{x}_n)\end{pmatrix}.$$

Hence, this previous expression can be simplified to

$$\mathbb{E}_{Y_{1..n}}\left(\mathcal{R}_{\Psi^p}(\widehat{\boldsymbol{\theta}}_{\Psi_{reg}})\right)=\sigma^2\,\left((\Phi^T\Phi)^{-1}\right)_{11},$$

where the notation $\left((\Phi^T\Phi)^{-1}\right)_{11}$ corresponds to the element of the matrix $(\Phi^T\Phi)^{-1}$ at the first line and the first column. Since for all nonsingular matrix A

$$A^{-1}=\frac{1}{det(A)}\,\mathrm{Com}(A)^T,$$

and taking $A=(\Phi^T\Phi)^{-1}$, we obtain

$$\begin{aligned}
\mathbb{E}_{Y_{1..n}}\left(\mathcal{R}_{\Psi^p}(\widehat{\boldsymbol{\theta}}_{\Psi_{reg}})\right) &= \sigma^2\,\frac{\left(\mathrm{Com}((\Phi^T\Phi))^T\right)_{11}}{det((\Phi^T\Phi))}\\
&= \sigma^2\,\frac{\mathrm{Cof}((\Phi^T\Phi))_{11}}{det((\Phi^T\Phi))}.
\end{aligned}$$

It could be written

$$\begin{aligned}\mathbb{E}_{Y_{1..n}}\left(\mathcal{R}_{\Psi^p}(\widehat{\boldsymbol{\theta}}_{\Psi_{reg}})\right) &= \frac{\sigma^2}{n}\left(1+\frac{n\,\mathrm{Cof}((\Phi^T\Phi))_{11}-det((\Phi^T\Phi))}{det((\Phi^T\Phi))}\right)\\ &= \frac{\sigma^2}{n}\left(1+\frac{Q}{det((\Phi^T\Phi))}\right),\end{aligned}$$

where

$$Q := n\,\mathrm{Cof}((\Phi^T\Phi))_{11}-det((\Phi^T\Phi))\,.$$

The quantity (4.22) has the same form as the quantity (4.21). Let us precise the value of $\frac{Q}{det((\Phi^T\Phi))}$, more precisely we will show that this ratio is positive (i.e Q is positive).
We have

$$\begin{aligned}(\Phi^T\Phi) &= \begin{pmatrix} n & \sum_{i=1}^n\phi_1(\mathbf{x}_i) & . & . & . & \sum_{i=1}^n\phi_p(\mathbf{x}_i)\\ \sum_{i=1}^n\phi_1(\mathbf{x}_i) & . & & & & \\ . & & . & & & \\ . & & & (\sum_{i=1}^n\phi_k(\mathbf{x}_i)\phi_l(\mathbf{x}_i))_{1\le k,l\le p} & & \\ . & & & & . & \\ \sum_{i=1}^n\phi_p(\mathbf{x}_i) & & & & & . \end{pmatrix}\\ &= n\times\begin{pmatrix} 1 & \overline{\phi_1} & . & . & . & \overline{\phi_p}\\ \overline{\phi_1} & . & & & & \\ . & & . & & & \\ . & & & (\overline{\phi_k\,\phi_l})_{1\le k,l\le p} & & \\ . & & & & . & \\ \overline{\phi_p} & & & & & . \end{pmatrix}\end{aligned}$$

where for a function g,

$$\overline{g}=\frac{1}{n}\sum_{i=1}^n g(\mathbf{x}_i)\,.$$

We denote by $M := (\overline{\phi_k\,\phi_l})_{1\le k,l\le p}$ and it is clear that

$$\mathrm{Cof}((\Phi^T\Phi))_{11}=n^p\,det(M)\,.$$

Moreover, using columns permutations and determinant properties, one can check that

$$det((\Phi^T\Phi))=n^{p+1}\left(det(M)-\sum_{\nu=1}^p det(M_\nu)\right),$$

where the matrix M_ν has the columns of the matrix M except the ν-th which is replaced by columns with elements $\left((\overline{\phi_k}\,\overline{\phi_\nu})_{1\le k\le p}\right)$.
Hence, the quantity Q is

$$\begin{aligned}Q &= n\,(n^p\,det(M))-n^{p+1}\left(det(M)-\sum_{\nu=1}^p det(M_\nu)\right)\\ &= n^{p+1}\sum_{\nu=1}^p det(M_\nu)\,,\end{aligned}$$

with

$$M_\nu = \begin{pmatrix} \overline{\phi_1^2} & \overline{\phi_1\,\phi_2} & \overline{\phi_1}\,\overline{\phi_\nu} & \overline{\phi_1\,\phi_p} \\ \overline{\phi_2\,\phi_1} & & \cdot & \\ \cdot & & \cdot & \\ \cdot & & (\overline{\phi_\nu})^2 & \\ \cdot & & \cdot & \\ \overline{\phi_p\,\phi_1} & & \overline{\phi_p}\,\overline{\phi_\nu} & \overline{\phi_p^2} \end{pmatrix}.$$

In order to prove that $Q \geq 0$, it suffices to prove that for all $\nu = 1, ..., p$

$$det(M_\nu) \geq 0\,. \tag{4.22}$$

Let us remark that if we prove that M_ν is a positive semi-definite matrix, i.e for all $\mathbf{z} \in \mathbb{R}^p$

$$\mathbf{z}^T\, M_\nu\, \mathbf{z} \geq 0\,,$$

then (4.22) is satisfied.
Let us introduce

$$\overline{\phi} := (\overline{\phi_1}, ..., \overline{\phi_p})\,,$$
$$N_\nu := \left(M_\nu - \overline{\phi}^T\,\overline{\phi}\right)^T,$$

the sets of indices

$$I := \{1, ..., p\} \quad I_{-\nu} = \{1, .., \nu - 1, \nu + 1, .., p\} \quad \nu \in \{1, ..., p-1\}\,,$$

and

$$C_{kl} := \overline{\phi_k\,\phi_l} - \overline{\phi_k}\,\overline{\phi_l} \quad k, l \in I\,.$$

By elementary calculus, we obtain

$$N_\nu = P_{\sigma_\nu} \begin{pmatrix} 0 & 0 & . & . & . & 0 \\ C_{2\nu} & & & & & \\ \cdot & & & & & \\ \cdot & & & C_{-\nu} & & \\ \cdot & & & & & \\ C_{p\nu} & & & & & \end{pmatrix} P_{\sigma_\nu}^T,$$

where P_{σ_ν} is the permutation matrix associated to the cycle

$$\sigma_\nu := \{\nu - 1, \nu - 2, ..., 1, \nu\}\,,$$

and

$$C_{-\nu} := (C_{kl})_{k,l \in I_{-\nu}}\,.$$

Denote by

$$\widetilde{N}_\nu := \begin{pmatrix} 0 & 0 & . & . & . & 0 \\ C_{2\nu} & & & & & \\ \cdot & & & & & \\ \cdot & & & C_{-\nu} & & \\ \cdot & & & & & \\ C_{p\nu} & & & & & \end{pmatrix}.$$

Let $\mathbf{v}_2, ..., \mathbf{v}_p$ be the $p-1$ eigen (orthonormalized) vectors of the matrix $C_{-\nu}$, it yields that the set of vectors $\mathbf{w}_i := \begin{pmatrix} 0 \\ \mathbf{v}_i \end{pmatrix}$ $i = 2, ..., p$ are $p-1$ eigen vectors of the matrix $\widetilde{N}_\nu$, and denote

by $\lambda_i, \quad i = 2, ..., p$ the corresponding eigen values. We complete this set of vectors with the vector $\mathbf{w}_1 = (1, 0, ..., 0)^T$ which forms an orthonormal basis of $\mathbb{R}^p$.
Let us prove that N_ν is positive semi-definite. Since permutation matrix is nonsingular, it is equivalent to prove that $\widetilde{N}_\nu$ is positive semi-definite.
Let $\mathbf{z} \in \mathbb{R}^p$ such that

$$\mathbf{z} = \sum_{i=1}^{p} \alpha_i \mathbf{w}_i \,,$$

we compute

$$\begin{aligned}
\mathbf{z}^T \widetilde{N}_\nu \mathbf{z} &= \left(\sum_{i=1}^{p} \alpha_i \mathbf{w}_i\right)^T \widetilde{N}_\nu \left(\sum_{i=1}^{p} \alpha_i \mathbf{w}_i\right) \\
&= \left(\alpha_1 \mathbf{w}_1 + \sum_{i=2}^{p} \alpha_i \mathbf{w}_i\right)^T \widetilde{N}_\nu \left(\alpha_1 \mathbf{w}_1 + \sum_{i=2}^{p} \alpha_i \mathbf{w}_i\right) \\
&= \left(\alpha_1 \mathbf{w}_1 + \sum_{i=2}^{p} \alpha_i \mathbf{w}_i\right)^T \alpha_1 \widetilde{N}_\nu \mathbf{w}_1 + \alpha_1 \mathbf{w}_1^T \widetilde{N}_\nu \left(\sum_{i=2}^{p} \alpha_i \mathbf{w}_i\right) + \left(\sum_{i=2}^{p} \alpha_i \mathbf{w}_i\right)^T \widetilde{N}_\nu \left(\sum_{i=2}^{p} \alpha_i \mathbf{w}_i\right) ,
\end{aligned}$$

but $\widetilde{N}_\nu \mathbf{w}_1 = 0$ and $\mathbf{w}_1$ is orthogonal to $\mathrm{Vect}(\mathbf{w}_2, ..., \mathbf{w}_p)$, so we obtain

$$\mathbf{z}^T \widetilde{N}_\nu \mathbf{z} = \sum_{i=2}^{p} \lambda_i \alpha_i^2 \geq 0 \,.$$

Hence N_ν is positive semi-definite.
Now, in order to prove that M_ν is also positive semi-definite, remark that

$$M_\nu = N_\nu^T + \overline{\phi}^T \overline{\phi} \,,$$

then M_ν is a sum of matrix positive semi-definite, thus M_ν too. In particular, each eigenvalue of M_ν is nonnegative, that leads to

$$det(M_\nu) \geq 0 \,,$$

for all $\nu = 1, ..., p$.

Finally, we showed that $Q = n^{p+1} \sum_{\nu=1}^{p} det(M_\nu)$ is nonnegative, and even positive, for all $\mathbf{x}_1, ..., \mathbf{x}_n$. Hence, denoting by

$$\beta_n = \mathbb{E}_{\mathbf{X}_{1..n}} \frac{Q}{det((\Phi^T \Phi))} > 0 \,,$$

and by the displays (4.21) and (4.22), it yields

$$\mathbb{E}_{(\mathbf{X}_i, Y_i)_{1..n}} \left(\mathcal{R}_{\Psi^p}(\widehat{\boldsymbol{\theta}}_{\Psi_{reg}})\right) - \mathbb{E}_{Y_{1..n}} \left(\mathcal{R}_{\Psi^p}(\widehat{\boldsymbol{\theta}}_{\Psi^p})\right) = \frac{\sigma^2}{n} \left(\beta_n - \frac{\|\boldsymbol{\theta}^*\|^2}{\sigma^2}\right) .$$

That concludes the proof.

4.9 Discussion sur le surapprentissage et la pénalisation de contraste

Dans les Sections 2.4 et 2.6, nous avons présenté la notion de pénalisation et sa motivation. Une grandeur importante étant la pénalité idéale relative à la procédure d'estimation.
Dans le cas de la section précédente, il était question d'une procédure de minimisation du type

$$\widehat{\boldsymbol{\theta}}_{\Psi} = \underset{\boldsymbol{\theta}\in\Theta}{\text{Argmin}}\ \widehat{\mathcal{R}}_{\Psi}(\boldsymbol{\theta})\,,$$

$$\widehat{\mathcal{R}}_{\Psi}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{1}_{Y_i>s} - \rho_{\mathcal{F}}(\boldsymbol{\theta})\right)^2\,,$$

avec $\rho_{\mathcal{F}}(\boldsymbol{\theta}) = \mathbb{P}(h(\mathbf{X},\boldsymbol{\theta}))$. Dans ce qui suit, nous ne tiendrons pas compte de la simulation de $\rho_{\mathcal{F}}(\boldsymbol{\theta})$.
Le risque théorique est $\mathcal{R}_{\Psi}(\boldsymbol{\theta}) = \mathbb{E}_Q\left(\mathbb{1}_{Y>s} - \rho_{\mathcal{F}}(\boldsymbol{\theta})\right)^2$.

Le but est de proposer une procédure pénalisée

$$\widehat{\boldsymbol{\theta}}_{\Psi_{\text{pen}}} = \underset{\boldsymbol{\theta}\in\Theta}{\text{Argmin}}\ \widehat{\mathcal{R}}_{\Psi}(\boldsymbol{\theta}) + \text{pen}_{\Psi}(\boldsymbol{\theta})\,, \tag{4.23}$$

afin d'améliorer la performance de l'estimateur $\widehat{\boldsymbol{\theta}}_{\Psi} = \widehat{\boldsymbol{\theta}}_{\Psi_{prob}}$ de la section précédente.

La pénalité idéale donnée par

$$\text{pen}^{id}_{\Psi}(\boldsymbol{\theta}) = (\mathcal{R}_{\Psi} - \widehat{\mathcal{R}}_{\Psi})(\boldsymbol{\theta})\,,$$

vaut dans ce cas, après calculs

$$\text{pen}^{id}_{\Psi}(\boldsymbol{\theta}) = 2\,a_n\,\rho_{\mathcal{F}}(\boldsymbol{\theta}) - a_n\,, \tag{4.24}$$

où $a_n = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{Y_i>s} - \mathbb{P}(Y>s)$. La constante (variable aléatoire) a_n ne dépendant pas de $\boldsymbol{\theta}$, la pénalité s'écrit

$$\text{pen}^{id}_{\Psi}(\boldsymbol{\theta}) = 2\,a_n\,\rho_{\mathcal{F}}(\boldsymbol{\theta})\,. \tag{4.25}$$

La seule inconnue de la pénalité idéale est donc la constante a_n, qui est en fait une variable aléatoire (dépend des Y_i) centrée et de variance v^2/n avec $v^2 = \mathbb{P}(y>s)(1-\mathbb{P}(y>s))$.
En approchant a_n par $G\,\frac{v}{\sqrt{n}}$, où G est une gaussienne standard, nous proposons la pénalité

$$\text{pen}_{\Psi}(\boldsymbol{\theta}) = K\,\frac{2\,\widehat{v}_n}{\sqrt{n}}\,\rho_{\mathcal{F}}(\boldsymbol{\theta})\,,$$

où K est une constante de calibration à déterminer et $\widehat{v}_n$ un estimateur de v.
On pourra supposer que la constante K prend une valeur se situant entre -3 et 3. La procédure d'estimation (4.23) s'écrit alors

$$\widehat{\boldsymbol{\theta}}_{\Psi_{\text{pen}}}(K) = \underset{\boldsymbol{\theta}\in\Theta}{\text{Argmin}}\ \frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{1}_{Y_i>s} - \rho_{\mathcal{F}}(\boldsymbol{\theta})\right)^2 + K\,\frac{2\,\widehat{v}_n}{\sqrt{n}}\,\rho_{\mathcal{F}}(\boldsymbol{\theta})\,. \tag{4.26}$$

En fait, avec les données utilisées dans la Section 4.15, la constante prend la valeur

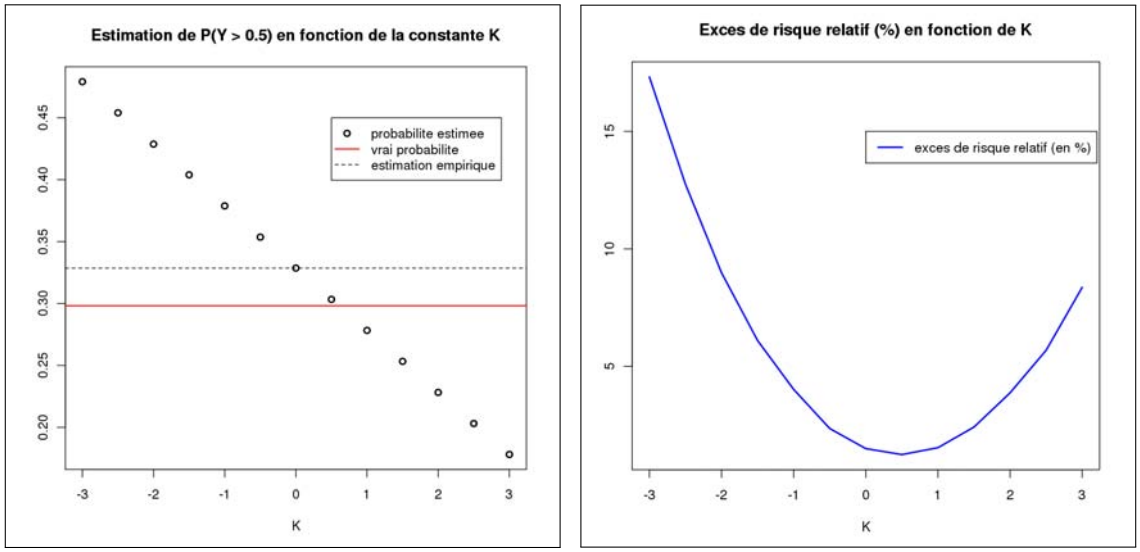
$$K^* = 0.541\,.$$

Autrement dit, l'estimateur $\widehat{\boldsymbol{\theta}}_{\Psi_{\text{pen}}}(K^*)$ est tel que

$$\mathbb{P}(h(\mathbf{X}, \widehat{\boldsymbol{\theta}}_{\Psi_{\text{pen}}}(K^*)) > 0.5) = \mathbb{P}(Y > 0.5)\,.$$

(Il n'y avait pas de biais de modèle.)

Bien que la calibration de la constante K ne fasse pas partie de notre propos, nous tenons néanmoins à donner quelques illustrations.



(a) (b)

FIGURE 4.6 – (a) Estimation de la probabilité $\mathbb{P}(Y > 0.5)$ par contraste pénalisé avec différentes constantes K. (b) Valeurs de l'excès de risque relatif r (4.18) aux différents paramètres $\widehat{\boldsymbol{\theta}}_{\Psi_{\text{pen}}}(K)$, $K \in \mathcal{K}$.

Dans la Figure 4.6 (a), on voit bien que lorsque $K = 0$ (pas de pénalisation), on retrouve l'estimation empirique basée sur les données Y_i, $i = 1, ..., n$. On retrouve également la constante $K^* = 0.541$, calculée à partir des données, qui minimise l'excès de risque relatif (donc l'excès de risque) à la Figure 4.6 (b).
La Figure 4.7 représente le tracé des paramètres $\widehat{\boldsymbol{\theta}}_{\Psi_{\text{pen}}}(K) = (\widehat{\theta}^1_{\Psi_{\text{pen}}}(K), \widehat{\theta}^2_{\Psi_{\text{pen}}}(K), \widehat{\theta}^2_{\Psi_{\text{pen}}}(K))$ en fonction de K (sur 31 valeurs).

4.10 Quelques éléments théoriques

Dans cette section, nous allons tenter d'élucider les questions soulevées dans les sections qui précèdent, notamment sur la dualité entre la procedure d'estimation des paramètres et la prédiction recherchée. Nous proposions l'étude de la quantité suivante

$$\mathcal{R}_{\Psi^p}(\widehat{\boldsymbol{\theta}}_{\Psi}) - \mathcal{R}_{\Psi^p}(\widehat{\boldsymbol{\theta}}_{\Psi^p})\,, \tag{4.27}$$

où Ψ^p est le contraste caractérisant la quantité d'intérêt considérée, et Ψ un contraste quelconque. Par exemple, dans un cas particulier nous avons montré que (Proposition 4.5.1)

$$\mathbb{E}_{(\mathbf{X}_i, Y_i)_{1..n}}\left(\mathcal{R}_{\Psi^p}(\widehat{\boldsymbol{\theta}}_{\Psi_{reg}}) - \mathcal{R}_{\Psi^p}(\widehat{\boldsymbol{\theta}}_{\Psi^p})\right) \geq 0\,.$$

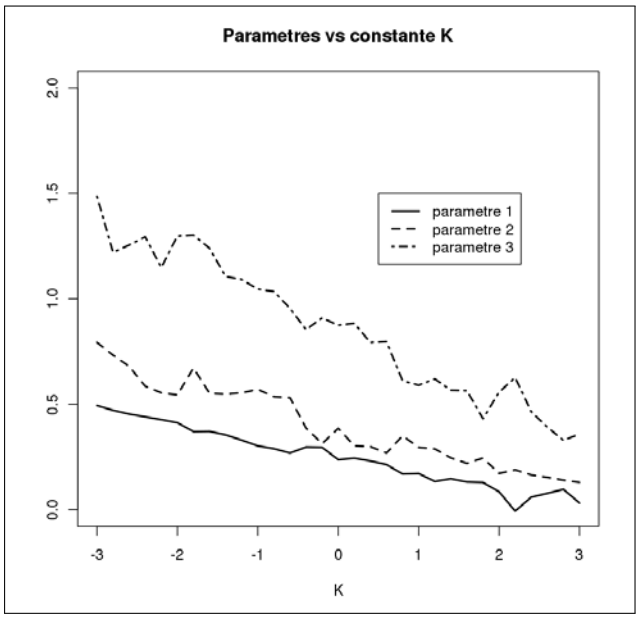


FIGURE 4.7 – Tracé de $K \mapsto \widehat{\theta}^i_{\Psi_{\text{pen}}}(K)$ pour $i = 1, 2$ et 3.

Cette inégalité signifie qu'en espérance, la procedure d'estimation $\widehat{\boldsymbol{\theta}}_{\Psi^p}$ ("classique") est "meilleure" que celle de la regression $\widehat{\boldsymbol{\theta}}_{\Psi_{reg}}$ pour la prédiction d'une moyenne. Toutefois, nous avons également vu qu'une procédure classique pouvait conduire à un surapprentissage défini dans la Section 4.15.

Intuitivement, deux ingrédients clé semblent expliquer les fluctuations de la différence (4.27) :

1. la performance des procédures d'estimation $\widehat{\boldsymbol{\theta}}_{\Psi^p}$ et $\widehat{\boldsymbol{\theta}}_{\Psi}$
2. la "distance" entre les contrastes Ψ^p et Ψ.

4.10.1 Performance d'un Ψ-estimateur

Nous avons vu, à la Remarque 2.5.2 du Chapitre 2, que la performance d'un Ψ-estimateur $\widehat{\boldsymbol{\theta}}_{\Psi}$ est donnée par son Ψ-excès de risque

$$\mathcal{E}^{\Theta}_{\Psi}(\widehat{\boldsymbol{\theta}}_{\Psi}) = \mathcal{R}_{\Psi}(\widehat{\boldsymbol{\theta}}_{\Psi}) - \mathcal{R}^{\Theta}_{\Psi}, \quad \left(\mathcal{R}^{\Theta}_{\Psi} = \inf_{\boldsymbol{\theta}\in\Theta} \mathcal{R}_{\Psi}(\boldsymbol{\theta})\right).$$

En effet, plus cette quantité est petite, plus l'estimateur est performant.

4.10.2 Dissemblance de contrastes

La notion de "distance" ou de "comparaison" de contrastes nous est apparue plusieurs fois jusqu'à présent. Notamment à la Section 4.6 d'exemples numériques où se dégageait une notion de proximité de contraste Ψ vis-à-vis de celui caractérisant la quantité d'intérêt que l'on veut prédire Ψ^p.

Dans ce qui suit, nous proposons une métrique indiquant la "proximité" entre deux contrastes. Rappelons les définitions suivantes.

– Risque : $\mathcal{R}_\Psi(\boldsymbol{\theta}) = \mathbb{E}_{Q^{\mathbf{z}}}\left(\Psi(\rho_{\mathcal{F}}(\boldsymbol{\theta}), Z)\right)$

– Modèle : $F = \{\rho_{\mathcal{F}}(\boldsymbol{\theta})\,,\ \boldsymbol{\theta} \in \Theta\} \subset \mathcal{F}$
On rappelle que
$$\rho_{\mathcal{F}}(\boldsymbol{\theta}) = \rho_{\mathcal{F}}(Q^{\mathbf{z}_{\boldsymbol{\theta},\eta}})\,, \quad Z_{\boldsymbol{\theta},\eta} = (\mathbf{X}, h(\mathbf{X},\boldsymbol{\theta}) + \eta)\,,$$
avec
$$Q^{\mathbf{z}_{\boldsymbol{\theta},\eta}}(d\mathbf{x}, dy) = p^{\mathbf{x}}(\mathbf{x})\, g_{\eta/\mathbf{X}=\mathbf{x}}(y - h(\mathbf{x},\boldsymbol{\theta}))\, d\mathbf{x}\, dy$$
et $\rho_{\mathcal{F}}(\cdot)$ est défini dans (2.3.1) et η une variable aléatoire centrée.

– Ψ-minimiseur : $\boldsymbol{\theta}_\Psi = \mathrm{Argmin}_{\boldsymbol{\theta}\in\Theta}\, \mathcal{R}_\Psi(\boldsymbol{\theta})$

– Projeté : $\rho_F = \rho_{\mathcal{F}}(\boldsymbol{\theta}_\Psi)$.

Définition 4.10.1. Dissemblance de contrastes.
Soit un modèle stochastique h (i.e $(\mathbf{X},\boldsymbol{\theta}) \mapsto h(\mathbf{X},\boldsymbol{\theta})\,,\ \mathbf{X} \sim P^{\mathbf{x}}\,,\ \boldsymbol{\theta} \in \Theta$) et soit la mesure $Q^{\mathbf{z}}$ sur $\mathcal{X} \times \mathcal{Y}$. Soit Ψ_1 et Ψ_2 un $\mathcal{F}_1$-contraste et un $\mathcal{F}_2$-contraste, respectivement. Notons $\mathcal{R}_{\Psi_1}$, $\mathcal{R}_{\Psi_2}$ leur risque respectif, puis $\boldsymbol{\theta}_{\Psi_2} = \mathrm{Argmin}_{\boldsymbol{\theta}\in\Theta}\, \mathcal{R}_{\Psi_2}(\boldsymbol{\theta})$.
On définit la dissemblance suivante

$$\Lambda^{[h,Q^{\mathbf{z}},\eta]}(\Psi_1,\Psi_2) := \max_{\boldsymbol{\theta}\in\boldsymbol{\theta}_{\Psi_2}} \mathcal{R}_{\Psi_1}(\boldsymbol{\theta}) - \mathcal{R}^{\Theta}_{\Psi_1}\,, \tag{4.28}$$

où $\mathcal{R}^{\Theta}_{\Psi_1} = \inf_{\boldsymbol{\theta}\in\Theta} \mathcal{R}_{\Psi_1}(\boldsymbol{\theta})$ et η est définie dans (2.29).

Remarque 4.10.1. 1 - La notion de proximité que nous avons défini est relative au modèle h, à la mesure $Q^{\mathbf{z}}$ et à la variable aléatoire η. Autrement dit, la modélisation et les hypothèses de l'étude ont une influence sur la "qualité" d'un contraste Ψ vis-à-vis de Ψ^p (contraste de "référence"). Nous verrons cela en exemple à la Section 4.10.3.

2 - La quantité $\Lambda^{[h,Q^{\mathbf{z}},\eta]}(\Psi_1,\Psi_2)$ n'est pas symétrique, c'est à dire que
$$\Lambda^{[h,Q^{\mathbf{z}},\eta]}(\Psi_1,\Psi_2) \neq \Lambda^{[h,Q^{\mathbf{z}},\eta]}(\Psi_2,\Psi_1)\,.$$

3 - Faire la comparaison, pour tout $\mathbf{z} \in \mathcal{Z}$
$$\Psi_1(\rho_1, z) - \Psi_2(\rho_2, z)\,, \quad \rho_1 \in \mathcal{F}_1,\, \rho_2 \in \mathcal{F}_2\,,$$
n'a pas de sens dans le cas où les espaces $\mathcal{F}_1$ et $\mathcal{F}_2$ sont différents.

Pour simplifier les notations, nous noterons s'il n'y a pas d'ambiguïté
$$\Lambda^{[h,Q^{\mathbf{z}},\eta]}(\Psi_1,\Psi_2) = \Lambda^{\eta}(\Psi_1,\Psi_2)\,.$$

Nous choisissons de mettre au premier plan la dépendance de cette dissemblance par rapport à la variable de modélisation η. Mais gardons à l'esprit que cette quantité est relative à h et à la mesure $Q^{\mathbf{z}}$ considérée.

Proposition 4.10.1. Propriétés de $\Lambda^{\eta}(\cdot,\cdot)$.
La dissemblance Λ^{η} (4.28) satisfait

1. *pour tout* Ψ_1, Ψ_2 , $\Lambda^{\eta}(\Psi_1,\Psi_2) \geq 0$

2. *pour tout* Ψ_1, Ψ_2 , $\Lambda^\eta(\Psi_1, \Psi_2) = 0 \Longleftrightarrow \boldsymbol{\theta}_{\Psi_2} \subset \boldsymbol{\theta}_{\Psi_1}$.

En particulier, si les minimiseurs $\boldsymbol{\theta}_{\Psi_1}$ *et* $\boldsymbol{\theta}_{\Psi_2}$ *sont uniques, alors*

$$\Lambda^\eta(\Psi_1, \Psi_2) = 0 \Longleftrightarrow \boldsymbol{\theta}_{\Psi_1} = \boldsymbol{\theta}_{\Psi_2} .$$

Remarquons que

$$\Lambda^\eta(\Psi_1, \Psi_2) = 0 \not\Rightarrow \Psi_1 = \Psi_2 .$$

En effet, il se peut que dans certains cas (par exemple pour certaines hypothèses sur η), la dissemblance Λ^η entre deux contrastes différents soit nulle.
Nous donnons ci-après un exemple où l'on retrouve un résultat bien connu.

4.10.3 Exemple

Dans cet exemple, nous allons supposer que la variable aléatoire η est indépendante de $\mathbf{X}$ et suit une loi normale centrée de variance σ^2 . Autrement dit, la densité g_η s'écrit

$$g_\eta(y) = \frac{1}{\sqrt{2\,\pi}\sigma}\, e^{-y^2/2\,\sigma^2} . \tag{4.29}$$

Premièrement, considérons le contraste $\Psi_{reg} : \mathcal{F}_{reg} \to L_1(Q^{\mathbf{z}})$ ($\mathcal{F}_{reg} = L_2(P^{\mathbf{x}})$) donné par

$$\forall\, \rho \in \mathcal{F}_{reg} , \quad \Psi_{reg}(\rho, (\mathbf{x}, y)) = (y - \rho(\mathbf{x}))^2 .$$

Sous les hypothèses précédentes et à partir de l'exemple déjà traité à la Section 2.5.4, on en déduit le modèle $F_{reg} \subset \mathcal{F}_{reg}$

$$F_{reg} = \{\mathbf{x} \mapsto \rho_{\mathcal{F}_{reg}}(\boldsymbol{\theta})(\mathbf{x}) = h(\mathbf{x}, \boldsymbol{\theta}) , \boldsymbol{\theta} \in \Theta\} .$$

Ensuite, considérons le contraste $\Psi_{\log} : \mathcal{F}_{\mathcal{Z}} \to L_1(Q^{\mathbf{z}})$ donné par

$$\forall\, \rho \in \mathcal{F}_{\mathcal{Z}} , \quad \Psi_{\log}(\rho, (\mathbf{x}, y)) = -\log(\rho)(\mathbf{x}, y) ,$$

où $\mathcal{F}_{\mathcal{Z}}$ est l'ensemble des densités sur $\mathcal{Z}$.

Explicitons le modèle $F_{\mathcal{Z}} \subset \mathcal{F}_{\mathcal{Z}}$ donné par l'application quantité d'intérêt.
Rappelons qu'une quantité d'intérêt sur $\mathcal{F}$ relative à une mesure μ sur $\mathcal{Z}$ est définie en (2.3.1) par

$$\rho_{\mathcal{F}}(\mu) = \int_{\mathcal{Z}} w_{\mathcal{F}}(x, y)\, \mu(dx, dy) .$$

Dans notre cas présent, $\mathcal{F} = \mathcal{F}_{\mathcal{Z}}$ et la fonction $w_{\mathcal{F}_{\mathcal{Z}}}$ est donnée par

$$w_{\mathcal{F}_{\mathcal{Z}}}(\mathbf{x}, y)(\mathbf{u}, v) = \delta_{\mathbf{u}}(\mathbf{x})\, \delta_v(y) .$$

Ensuite, on a $\mu = Q^{\mathbf{z}_{\boldsymbol{\theta},\eta}}$ où, d'après (2.31),

$$Q^{\mathbf{z}_{\boldsymbol{\theta},\eta}}(d\mathbf{x}, dy) = p^{\mathbf{x}}(\mathbf{x})\, g_{\eta/\mathbf{X}=\mathbf{x}}(y - h(\mathbf{x}, \boldsymbol{\theta}))\, d\mathbf{x}\, dy .$$

D'après l'hypothèse de début sur η (4.29), on a $g_{\eta/\mathbf{X}=\mathbf{x}} = g_\eta$.

Ainsi, pour tout $\boldsymbol{\theta} \in \Theta$ et $(\mathbf{u}, v) \in \mathcal{Z} = \mathcal{X} \times \mathcal{Y}$

$$\begin{aligned}
\rho_{\mathcal{F}_{\mathcal{Z}}}(\boldsymbol{\theta})(\mathbf{u}, v) &= \rho_{\mathcal{F}_{\mathcal{Z}}}(Q^{\mathbf{z}_{\boldsymbol{\theta},\eta}})(\mathbf{u}, v) \\
&= \int_{\mathcal{Z}} \delta_{\mathbf{u}}(\mathbf{x})\, \delta_v(y)\, p^{\mathbf{x}}(\mathbf{x})\, g_\eta(y - h(\mathbf{x}, \boldsymbol{\theta}))\, d\mathbf{x}\, dy \\
&= p^{\mathbf{x}}(\mathbf{u})\, g_\eta(v - h(\mathbf{u}, \boldsymbol{\theta})) \\
&= \frac{p^{\mathbf{x}}(\mathbf{u})}{\sqrt{2\,\pi}\sigma}\, e^{-(v-h(\mathbf{u},\boldsymbol{\theta}))^2/2\,\sigma^2}\,.
\end{aligned}$$

Enfin, il vient que

$$F_{\mathcal{Z}} = \{(\mathbf{x}, y) \mapsto \rho_{\mathcal{F}_{\mathcal{Z}}}(\boldsymbol{\theta})(\mathbf{x}, y) = \frac{p^{\mathbf{x}}(\mathbf{u})}{\sqrt{2\,\pi}\sigma}\, e^{-(v-h(\mathbf{u},\boldsymbol{\theta}))^2/2\,\sigma^2}\,,\ \boldsymbol{\theta} \in \Theta\}\,.$$

Calculons les risques $\mathcal{R}_{\Psi_{reg}}(\boldsymbol{\theta})$ et $\mathcal{R}_{\Psi_{\log}}(\boldsymbol{\theta})$.
On a

$$\begin{aligned}
\mathcal{R}_{\Psi_{reg}}(\boldsymbol{\theta}) &= \mathbb{E}_{Q^{\mathbf{z}}}\left(\Psi_{reg}(\rho_{\mathcal{F}_{reg}}(\boldsymbol{\theta}), Z)\right) \\
&= \mathbb{E}_{Q^{\mathbf{z}}}\left(Y - h(\mathbf{X}, \boldsymbol{\theta})\right)^2\,.
\end{aligned}$$

Ensuite,

$$\begin{aligned}
\mathcal{R}_{\Psi_{\log}}(\boldsymbol{\theta}) &= \mathbb{E}_{Q^{\mathbf{z}}}\left(\Psi_{\log}(\rho_{\mathcal{F}_{\mathcal{Z}}}(\boldsymbol{\theta}), Z)\right) \\
&= \mathbb{E}_{Q^{\mathbf{z}}}\left(-\log\left(\frac{p^{\mathbf{x}}(\mathbf{X})}{\sqrt{2\,\pi}\sigma}\, e^{-(Y-h(\mathbf{X},\boldsymbol{\theta}))^2/2\,\sigma^2}\right)\right) \\
&= -\mathbb{E}_{Q^{\mathbf{z}}} \log\left(\frac{p^{\mathbf{x}}(\mathbf{X})}{\sqrt{2\,\pi}\sigma}\right) + \frac{1}{2\,\sigma^2}\, \mathbb{E}_{Q^{\mathbf{z}}}\left(Y - h(\mathbf{X}, \boldsymbol{\theta})\right)^2 \\
&= a + b\, \mathcal{R}_{\Psi_{reg}}(\boldsymbol{\theta})\,,
\end{aligned}$$

où a et $b > 0$ sont des constantes indépendantes de $\boldsymbol{\theta}$.

Cette dernière égalité nous donne donc

$$\begin{aligned}
\boldsymbol{\theta}_{\Psi_{\log}} &= \operatorname*{Argmin}_{\boldsymbol{\theta}\in\Theta} \mathcal{R}_{\Psi_{\log}}(\boldsymbol{\theta}) \\
&= \operatorname*{Argmin}_{\boldsymbol{\theta}\in\Theta} \left(a + b\, \mathcal{R}_{\Psi_{reg}}(\boldsymbol{\theta})\right) \\
&= \operatorname*{Argmin}_{\boldsymbol{\theta}\in\Theta} \mathcal{R}_{\Psi_{reg}}(\boldsymbol{\theta})\,,
\end{aligned}$$

d'où

$$\boldsymbol{\theta}_{\Psi_{\log}} = \boldsymbol{\theta}_{\Psi_{reg}}\,.$$

Par conséquent, on vérifie immédiatement que

$$\Lambda^{\eta}(\Psi_{reg}, \Psi_{\log}) = \Lambda^{\eta}(\Psi_{\log}, \Psi_{reg}) = 0\,, \tag{4.30}$$

alors que $\Psi_{reg} \neq \Psi_{\log}$.

Ce résultat est généralement connu sous la forme :

sous l'hypothèse de bruit gaussien centré homoscédastique, l'estimateur des moindres carrés est l'estimateur du maximum de vraisemblance.

4.10.4 Remarque finale

Dans cette section, nous proposons une écriture de la différence

$$\mathcal{R}_{\Psi^p}(\widehat{\boldsymbol{\theta}}_{\Psi}) - \mathcal{R}_{\Psi^p}(\widehat{\boldsymbol{\theta}}_{\Psi^p})$$

afin de mieux cerner les éléments importants qui influent sur celle-ci.

Remarque 4.10.2. Si Ψ^p est un $\mathcal{F}^p$-contraste (de prédiction) et Ψ un $\mathcal{F}$-contraste (d'estimation). Notons $\widehat{\boldsymbol{\theta}}_{\Psi^p}$ et $\widehat{\boldsymbol{\theta}}_{\Psi}$ des estimateurs associés respectivement aux minimiseurs $\boldsymbol{\theta}_{\Psi^p}$ et $\boldsymbol{\theta}_{\Psi}$. Alors,

$$\mathcal{R}_{\Psi^p}(\widehat{\boldsymbol{\theta}}_{\Psi}) - \mathcal{R}_{\Psi^p}(\widehat{\boldsymbol{\theta}}_{\Psi^p}) \quad = \quad \mathcal{E}^{\Theta}_{\Psi^p}(\widehat{\boldsymbol{\theta}}_{\Psi}) - \mathcal{E}^{\Theta}_{\Psi^p}(\widehat{\boldsymbol{\theta}}_{\Psi^p})\,. \tag{4.31}$$

Et si $\boldsymbol{\theta}_{\Psi}$ est unique, alors

$$\mathcal{R}_{\Psi^p}(\widehat{\boldsymbol{\theta}}_{\Psi}) - \mathcal{R}_{\Psi^p}(\widehat{\boldsymbol{\theta}}_{\Psi^p}) = \Lambda^{\eta}(\Psi^p, \Psi) + \Delta_{\Psi^p}(\widehat{\boldsymbol{\theta}}_{\Psi}, \boldsymbol{\theta}_{\Psi}) - \mathcal{E}^{\Theta}_{\Psi^p}(\widehat{\boldsymbol{\theta}}_{\Psi^p})\,, \tag{4.32}$$

où $\Delta_{\Psi}(\widehat{\boldsymbol{\theta}}_{\Psi^p}, \boldsymbol{\theta}_{\Psi}) = \mathcal{R}_{\Psi^p}(\widehat{\boldsymbol{\theta}}_{\Psi}) - \mathcal{R}_{\Psi^p}(\boldsymbol{\theta}_{\Psi})\,.$

Cette remarque nous indique tout d'abord que la différence $\mathcal{R}_{\Psi^p}(\widehat{\boldsymbol{\theta}}_{\Psi}) - \mathcal{R}_{\Psi^p}(\widehat{\boldsymbol{\theta}}_{\Psi^p})$ est en fait la différence de deux excès de risques (positifs), correspondant respectivement à $\widehat{\boldsymbol{\theta}}_{\Psi}$ et $\widehat{\boldsymbol{\theta}}_{\Psi^p}$. Autrement dit, une *cross prediction* est performante si le Ψ-estimateur $\widehat{\boldsymbol{\theta}}_{\Psi}$ qui en résulte est plus performant (i.e a un plus petit excès de risque) que la procédure classique (adaptée) $\widehat{\boldsymbol{\theta}}_{\Psi^p}$.
L'excès de risque $\mathcal{E}^{\Theta}_{\Psi^p}(\widehat{\boldsymbol{\theta}}_{\Psi})$ comporte un **biais** intrinsèque donné par la quantité

$$\Lambda^{\eta}(\Psi^p, \Psi) \geq 0\,.$$

La part de variance de cet excès de risque est donnée par

$$\Delta_{\Psi}(\widehat{\boldsymbol{\theta}}_{\Psi^p}, \boldsymbol{\theta}_{\Psi}) = \mathcal{R}_{\Psi^p}(\widehat{\boldsymbol{\theta}}_{\Psi}) - \mathcal{R}_{\Psi^p}(\boldsymbol{\theta}_{\Psi})\,.$$

Grossièrement, un Ψ-estimateur $\widehat{\boldsymbol{\theta}}_{\Psi}$ "rapide" (par exemple $\Delta_{\Psi^p}(\widehat{\boldsymbol{\theta}}_{\Psi}, \boldsymbol{\theta}_{\Psi}) < \mathcal{E}^{\Theta}_{\Psi^p}(\widehat{\boldsymbol{\theta}}_{\Psi^p})$) mais tel que $\Lambda^{\eta}(\Psi^p, \Psi) > 0$ ne soit pas négligeable, n'améliore pas nécessairement la prédiction donnée par $\widehat{\boldsymbol{\theta}}_{\Psi^p}$.

4.11 Exemple d'application industrielle : Électromagnétisme

Dans cette section, nous proposons un schéma d'étude d'un problème industriel en électromagnétisme. Il s'agit d'analyser la réponse de certains composants d'un avion commercial lorsque ce dernier est exposé à l'emission d'une onde plane, qui représente en réalité une onde émise par une antenne.

4.11.1 Problématique

Nous disposons d'un code de calcul numérique que nous noterons

$$\bar{h} : \mathcal{X} \to \mathcal{Y}\,,$$

qui permet de calculer l'intensité d'un courant sur une structure à partir des caractéristiques (dans $\mathcal{X}$) d'une onde plane. Dans notre propos, l'onde plane sera prise sous la forme d'une

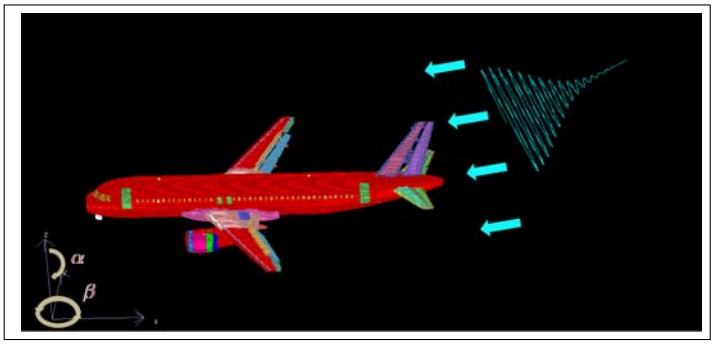


FIGURE 4.8 – Onde plane sur avion

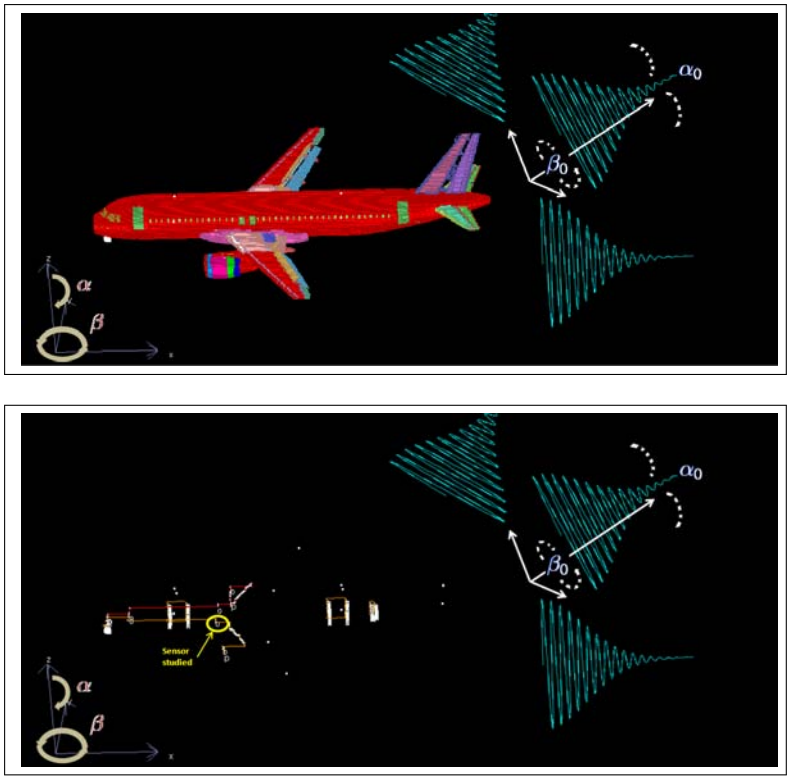


FIGURE 4.9 – Incertitude sur l'onde plane (haut) et capteur de courant (bas)

gaussienne modulée et les caractéristiques considérées sont les angles d'émission (α, β), c'est à dire que $\mathcal{X} = [0, 180] \times [0, 360]$. La structure que l'on étudie est un fil à l'intérieur de l'avion. (Représentation simplifiée d'une installation électrique).
Nous nous placerons dans le domaine temporel, puis nous nous focaliserons sur l'étude du **maximum en temps** du courant dans le fil (par la suite, on omettra le terme "maximum").

Nous disposons de $n = 16$ mesures expérimentales $I_{max}^{*,1}, ..., I_{max}^{*,n}$.

Soit $(\alpha, \beta) \in \mathcal{X}$, l'intensité donnée par le code numérique $\bar{h}$ est notée

$$I_{max} = \bar{h}(\alpha, \beta)\,.$$

Le but de l'étude est de calculer la probabilité que l'intensité I_{max}^* dépasse un certain seuil s sous une incertitude des angles (α, β),

$$\text{Quantité d'intérêt} = \mathbb{P}(I_{max}^* > s)\,.$$

Pour cela, nous allons utiliser le code numérique $\bar{h}$. La principale difficulté est que l'utilisation de $\bar{h}$ est extrêmement coûteuse en temps de calcul (de l'ordre de l'heure pour 1 run). On propose ci-après une approche possible pour l'étude du problème posé.

4.11.2 Approche possible

Afin d'illustrer nos propos, fixons nous un budget de $N = 20$ runs du modèle $\bar{h}$ et supposons qu'un jugement d'expert nous donne $\mathcal{X} = [28, 100] \times [100, 176]$, que l'on munira de la distribution uniforme $P^{\mathbf{x}}$ (on supposera l'indépendance des angles). Notons $(\mathbf{X}_1, I_{max}^1), ..., (\mathbf{X}_N, I_{max}^{20})$ les N données entrées/sorties utilisées pour les calculs, avec $\mathbf{X}_i = (\alpha_i, \beta_i) \sim P^{\mathbf{x}}$.

Construction d'un meta-modèle

Il est assez naturel d'envisager la construction d'un *meta-modèle* de $\bar{h}$ que l'on pourra exploiter de manière intensive. On construit alors une surface de réponse, notée $h : \mathcal{X} \to \mathcal{Y}$, obtenue par Krigeage à partir des données $(\mathbf{X}_1, I_{max}^1), ..., (\mathbf{X}_{20}, I_{max}^{20})$, voir Figures 4.10 et 4.12.
Ainsi, la quantité d'intérêt $\mathbb{P}(I_{max}^* > s)$ sera alors estimée (prédite) par une quantité du type

$$\mathbb{P}_{\widetilde{\mathbf{X}}}(h(\widetilde{\mathbf{X}}) > s)\,,$$

ou encore par Monter-Carlo intensif

$$\frac{1}{M}\sum_{j=1}^{M} \mathbb{1}_{h(\widetilde{\mathbf{x}}_j)>s}\,, \quad \widetilde{\mathbf{X}}_j \text{ i.i.d } \sim \widetilde{\mathbf{X}}\,,$$

où la variable aléatoire $\widetilde{\mathbf{X}}$ est à **déterminer**. En effet, il n'est pas dit que le choix de prendre $\widetilde{\mathbf{X}}$ distribué selon $P^{\mathbf{x}}$ soit le bon.

Calibration de l'incertitude

Modélisons l'incertitude $\widetilde{\mathbf{X}}$ par

$$\widetilde{\mathbf{X}} = \mathbf{a} + \Sigma\,\boldsymbol{\xi}\,,$$

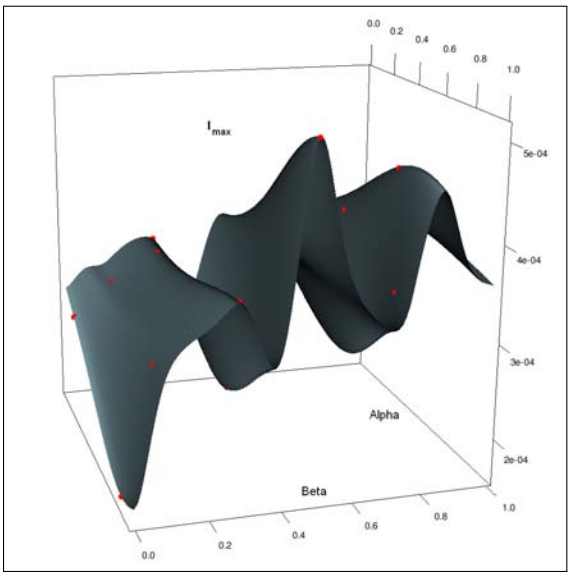


FIGURE 4.10 – Krigeage du code de calcul $\bar{h}$ à partir de 20 données

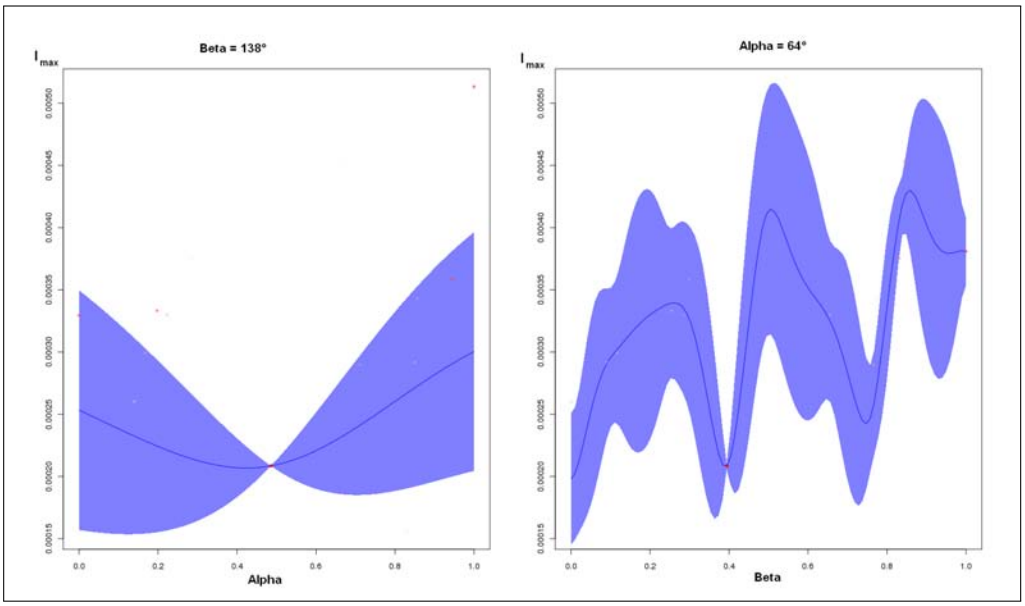


FIGURE 4.11 – Coupes du Krigeage et bandes de confiance à 95 %

où $\mathbf{a} = (\alpha, \beta)^T \in \mathbb{R}^2$, $\Sigma = \text{diag}(\sigma_1^2, \sigma_2^2)$ et $\boldsymbol{\xi}$ est une gaussienne standard dans $\mathbb{R}^2$. Par abus de notation, on écrira

$$h(\widetilde{\mathbf{X}}) = h(\boldsymbol{\xi}, \boldsymbol{\theta}), \quad \boldsymbol{\theta} = (\alpha, \beta, \sigma_1, \sigma_2).$$

La loi de $\boldsymbol{\xi}$ étant connue, la calibration consiste alors à estimer le paramètre $\boldsymbol{\theta}$. Pour cela, on rappelle que l'on dispose d'un échantillon $I_{max}^{*,1}, ..., I_{max}^{*,n}$. On se ramène alors à un problème inverse stochastique (cf. Chapitre 3). Par exemple, fixons $\alpha = 64^\circ$, $\sigma_1 = 0$ (i.e la composante en α est déterministe) et $\sigma_2 = 10$. En considérant le log-contraste, le paramètre β peut être estimé par

$$\widehat{\beta} = \underset{\beta}{\text{Argmin}} \sum_{i=1}^{n} \log \left(\sum_{j=1}^{m} K_{b_\beta}(I_{max}^{*,i} - h(\boldsymbol{\xi}_j, \beta)) \right), \quad \boldsymbol{\xi}_j \text{ i.i.d selon } \boldsymbol{\xi},$$

où K_{b_β} est le noyau gaussien de taille de fenêtre b_β égale à la règle de Silverman sur l'échantillon $h(\boldsymbol{\xi}_j, \beta)$, $j = 1, ..., m$.
Dans cet exemple, on obtient une valeur de $\widehat{\beta} = 135.87^\circ$.

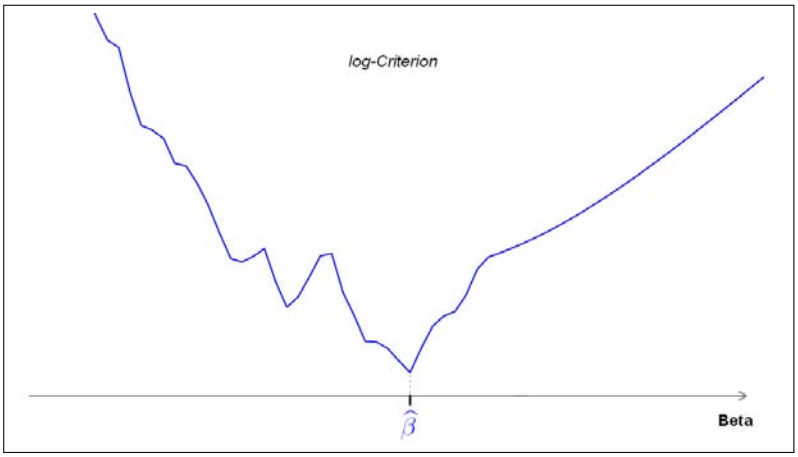


FIGURE 4.12 – Critère issu du log-contraste

Prédiction

Une fois le paramètre $\widehat{\boldsymbol{\theta}}$ calculé, par un tirage de Monte-Carlo intensif on calcule

$$\mathbb{P}_{\boldsymbol{\xi}}(h(\boldsymbol{\xi}, \widehat{\boldsymbol{\theta}}) > s)$$

qui sera alors candidat à la prédiction de $\mathbb{P}(I_{max}^* > s)$.
D'après ce qui a été établi dans ce chapitre, une question intéressante serait l'étude de l'influence de la procedure d'estimation des paramètres sur la quantité d'intérêt recherchée, ici $\mathbb{P}(I_{max}^* > s)$. De plus, il conviendrait également de discuter du choix du métamodèle ainsi que du choix des points $(\mathbf{X}_1, I_{max}^1), ..., (\mathbf{X}_N, I_{max}^N)$ servant à la construction du métamodèle (techniques de plans d'expériences adaptés à l'objectif recherché).

Les considérations de cette section peuvent constituer un début pour l'analyse du problème électromagnétique en question.

Bibliographie

[1] P.L. Bartlett, M.I. Jordan, and J.D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473) :138–156, 2006.

[2] E. de Rocquigny, N. Devictor, and S. Tarantola, editors. *Uncertainty in industrial practice*. John Wiley.

[3] N. Rachdi, J.C. Fort, and T. Klein. Risk bounds for new M-estimation problems . *hal 00537236*, 2010.

Chapitre 5

Algorithmes stochastiques pour modèles statistiques complexes

Sommaire

Résumé du Chapitre

Dans le domaine de la Statistique, beaucoup de problèmes reviennent à minimiser une fonction dépendant du problème statistique lui-même ainsi que de certaines hypothèses émises par le praticien. Parfois, la fonction à minimiser peut être très compliquée et comporter ainsi plusieurs difficultés pour la minimisation (minima locaux, bassins étroits etc...). Dans ce chapitre, nous proposons une méthode d'optimisation basée sur les algorithmes stochastiques. Cette méthode présente la particularité d'être *dynamique*. Plus précisément, nous allons permettre à la fonction à minimiser de "varier" en fonction du temps, afin d'éviter certaines irrégularités parasites.

5.1 Éléments sur les algorithmes stochastiques

5.1.1 Introduction

Les algorithmes stochastiques ont fait leur apparition dans les travaux précurseurs de H. Robbins et S. Monro dans les années 50 [9], sur l'utilisation de méthodes stochastiques pour l'optimisation.
Ces algorithmes ont souvent été pensés pour résoudre des problèmes d'optimisation difficiles, dont la particularité est la considération d'un aléa ou d'un bruit dans le problème en question.
Pour notre propos, nous nous baserons sur les méthodes données dans les ouvrages de M. Duflo [2] et [3], dont on donne les grandes lignes dans la section suivante.

5.1.2 L'algorithme stochastique

Définition 5.1.1. Algorithme stochastique.
Un algorithme stochastique est la donnée d'une suite $(\boldsymbol{\theta}_t)_{t\geq 0}$ à valeurs dans $\mathbb{R}^k$, adaptée à une filtration $\mathcal{F}_t$ et définie récursivement à partir d'un $\boldsymbol{\theta}_0 \in \mathbb{R}^k$ par

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \gamma_{t+1}\left(M(\boldsymbol{\theta}_t) + \pi_{t+1}\right), \quad t \geq 0\,, \tag{5.1}$$

pour une fonction $M : \mathbb{R}^k \to \mathbb{R}^k$, une suite $(\gamma_t)_{t\geq 1}$ déterministe positive et décroissante vers 0 vérifiant

$$\sum_{t\geq 1} \gamma_t = +\infty$$

et une suite $(\pi_t)_{t\geq 1}$ de variables aléatoires aussi appelées *perturbations*, qui satisfait pour tout $t \geq 0$
– π_t est $\mathcal{F}_t$-mesurable
– $\mathbb{E}(\pi_{t+1}\,/\mathcal{F}_t) = 0\,.$

Un résultat de convergence bien connu et crucial pour les méthodes récursives stochastiques est le théorème de Robbins-Monro, que l'on trouve par exemple dans [3] (Theorem 1.4.26 p. 29).

Théorème 5.1.1. Robbins-Monro.
Soit l'algorithme stochastique (5.1) avec la filtration $\mathcal{F}_t = \sigma(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_t)$, et notons $\|\cdot\|$ la norme euclidienne sur $\mathbb{R}^k$. Supposons que
(i) M est continue et il existe un $\boldsymbol{\theta}^* \in \mathbb{R}^k$ tel que

$$M(\boldsymbol{\theta}^*) = 0 \quad et \quad \forall\, \boldsymbol{\theta} \neq \boldsymbol{\theta}^*\,,\ M(\boldsymbol{\theta})\,(\boldsymbol{\theta} - \boldsymbol{\theta}^*) > 0$$

(ii) Il existe une constante $K > 0$ telle que pour tout $t \geq 0$

$$\mathbb{E}\left(\|M(\boldsymbol{\theta}_t) + \pi_{t+1}\|^2 / \mathcal{F}_t\right) \leq K(1 + \|\boldsymbol{\theta}_t\|^2)\,.$$

Alors si $\sum_{t\geq 1} \gamma_t^2 < +\infty$,

$$\boldsymbol{\theta}_t \underset{t\to+\infty}{\longrightarrow} \boldsymbol{\theta}^* \quad \textit{presque sûrement}\,.$$

Ce type d'algorithme a beaucoup été étudié, citons par exemple les travaux de J-C Fort et G. Pagès [4], [5] et les travaux de M. Benaïm [1], parmi d'autres. Ils ont pour but d'approcher une racine de l'équation

$$M(\boldsymbol{\theta}) = 0\,.$$

Par exemple, si on veut atteindre un niveau α d'une fonction croissante g, on prend $M(\boldsymbol{\theta}) = \alpha - g(\boldsymbol{\theta})$ (cf. Dosage par méthode de Robbins-Monro). Ou bien, si un problème consiste à minimiser une fonction H, alors on prend $M(\boldsymbol{\theta}) = \nabla H(\boldsymbol{\theta})$.
Notons que si on ne considère pas de perturbation $(\pi_t)_{t\geq 1}$, la méthode itérative (5.1) n'est autre qu'une méthode du gradient classique

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \gamma_{t+1}\, M(\boldsymbol{\theta}_t)\,, \quad t \geq 0\,.$$

En quelque sorte, la perturbation permettrait d'aller là où un algorithme du gradient ne peut pas. Par exemple, lorsqu'on est dans un puits, se permettre d'en visiter un autre, etc... Toutefois, la perturbation ne doit pas être trop "grande" afin d'assurer des propriétés de convergence. La nature de la perturbation varie en fonction des problèmes : par exemple, $M(\boldsymbol{\theta})$ est une quantité obtenue par expérience comportant un bruit, ou alors $M(\boldsymbol{\theta})$ est définie par une espérance que l'on ne sait pas (ou difficilement) calculer. C'est en particulier ce dernier aspect qui motive notre usage des algorithmes stochastiques à la Section 5.3.
En effet, supposons que nous voulons approcher le minimum d'une fonction $\boldsymbol{\theta} \mapsto H(\boldsymbol{\theta})$ telle que $H(\boldsymbol{\theta}) = \mathbb{E}_W(H(\boldsymbol{\theta}, W))$, où W est une variable aléatoire dans $\mathbb{R}^{k'}$, $k' \leq k$. Une méthode déterministe classique de descente de gradient va utiliser

$$M(\boldsymbol{\theta}) = \nabla\, H(\boldsymbol{\theta}) = \nabla\, \mathbb{E}_W(H(\boldsymbol{\theta}, W))\,,$$

et sous des conditions raisonnables on a

$$M(\boldsymbol{\theta}) = \mathbb{E}_W\left(\nabla_{\boldsymbol{\theta}}\, H(\boldsymbol{\theta}, W)\right)\,.$$

Si k' est grand, le calcul de $M(\boldsymbol{\theta})$ peut être difficile et les méthodes déterministes sont ainsi non praticables. En revanche, les méthodes stochastiques ne nécessitent que le calcul des $\nabla H(\boldsymbol{\theta}, w)$ (si le gradient est connu) ou des $H(\boldsymbol{\theta}, w)$ (si le gradient est inconnu), pour des $\boldsymbol{\theta}$ et w donnés par l'algorithme. Le fait de ne pas calculer l'espérance se paye par la présence d'un "bruit". On donne ci-après deux algorithmes stochastiques : le premier dans le cas où le gradient est connu et le second lorsqu'il ne l'est pas.

Gradient connu : algorithme de Robbins-Monro classique .

Notons le gradient $G(\boldsymbol{\theta}, w) = \nabla_{\boldsymbol{\theta}}\, H(\boldsymbol{\theta}, w)$, le problème consiste alors à chercher le (ou les) zéro de $M(\boldsymbol{\theta}) = \mathbb{E}(G(\boldsymbol{\theta}, W))$. Soit $(W_t)_{t\geq 1}$ une suite i.i.d de même loi que W et soit une suite $(\gamma_t)_{t\geq 1}$ satisfaisant les hypothèses de la Définition 5.1.1. Soit $\boldsymbol{\theta}_0 \in \mathbb{R}^k$ et $\mathcal{F}_t = \sigma(\boldsymbol{\theta}_0, W_1, ..., W_t)$, on vérifie facilement que la suite $(\boldsymbol{\theta}_t)_{t\geq 0}$ donnée par l'algorithme suivant

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \gamma_{t+1}\, G(\boldsymbol{\theta}_t, W_{t+1})\,, \quad t \geq 0$$

est bien un algorithme stochastique au sens de la Définition 5.1.1. En effet, il suffit d'écrire

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \gamma_{t+1}\left(\mathbb{E}(G(\boldsymbol{\theta}_t, W_{t+1})/\mathcal{F}_t) + \pi_{t+1}\right), \quad t \geq 0$$

avec $\pi_{t+1} = G(\boldsymbol{\theta}_t, W_{t+1}) - \mathbb{E}(G(\boldsymbol{\theta}_t, W_{t+1})/\mathcal{F}_t)$, et il est clair que $\mathbb{E}(G(\boldsymbol{\theta}_t, W_{t+1})/\mathcal{F}_t) = \mathbb{E}(G(\boldsymbol{\theta}_t, W)) = M(\boldsymbol{\theta}_t)$ et $\mathbb{E}(\pi_{t+1}/\mathcal{F}_t) = 0$ puisque $\boldsymbol{\theta}_t$ est $\mathcal{F}_t$-mesurable et W_{t+1} est indépendant de $\mathcal{F}_t$.

Gradient inconnu : algorithme de Kiefer-Wolfowitz .

Dans le cas où le gradient $G(\boldsymbol{\theta}, w)$ n'est pas connu analytiquement, l'algorithme de Kiefer-Wolfowitz, introduit en 1952 et motivé par celui de Robbins-Monro, propose d'approcher le gradient par différences finies. Soit $(W_t)_{t\geq 1}$ et $(\widetilde{W}_t)_{t\geq 1}$ deux suites mutuellement indépendantes, soit $(\delta_t)_{t\geq 0}$ une suite qui décroît vers 0 et notons $(e_l)_{1\leq l\leq k}$ la base canonique de $\mathbb{R}^k$. Un algorithme de Kiefer-Wolfowitz est donné par

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \gamma_{t+1}\, \widehat{G}(\boldsymbol{\theta}_t)\,,$$

avec $\widehat{G}(\boldsymbol{\theta}_t) = (\widehat{G}(\boldsymbol{\theta}_t)_1, ..., \widehat{G}(\boldsymbol{\theta}_t)_k)^T$ tel que

$$\widehat{G}(\boldsymbol{\theta}_t)_l = \frac{H(\boldsymbol{\theta}_t + \delta_t\, e_l, W_{t+1}) - H(\boldsymbol{\theta}_t - \delta_t\, e_l, \widetilde{W}_{t+1})}{2\,\delta_t}\,.$$

On a le résultat de convergence suivant.
Supposons que H est de classe $\mathcal{C}^2$ avec un minimum atteint en $\boldsymbol{\theta}^*$, puis que la hessienne $\nabla^2 H$ est Lispchitzienne et définie positive. Enfin, si les séquences $(\gamma_t)_{t\geq 1}$ et $(\delta_t)_{t\geq 0}$ satisfont

$$\sum_{t\geq 0}\gamma_t = +\infty \quad \text{et} \quad \sum_{t\geq 0}\left(\frac{\gamma_t}{\delta_t}\right)^2 < +\infty\,,$$

alors $\boldsymbol{\theta}_t \underset{t\to+\infty}{\longrightarrow} \boldsymbol{\theta}^*$ presque sûrement.

5.1.3 Vers un algorithme *dynamique*

La fonction H à minimiser présente en général de nombreux minima locaux. Notre idée est de la régulariser ("convexifier") en la convoluant par exemple avec un noyau gaussien. Cependant, il faut immédiatement noter que, dans le cas des computer experiments, H n'a pas d'expression analytique mais est seulement connue par le calcul en chaque $\boldsymbol{\theta}$. A fortiori, il en est de même pour toute régularisée. Pour contourner ce problème, nous observons que convoluer par une densité gaussienne revient à ajouter un bruit gaussien.
En effet, soit g_{σ^2} une densité de probabilité par rapport à la mesure de Lebesgue sur $\mathbb{R}^k$ et soit W^σ une variable aléatoire de densité g_{σ^2}, alors le produit de convolution entre H et g_{σ^2} en $\boldsymbol{\theta}$, noté $H \star g_{\sigma^2}(\boldsymbol{\theta})$, n'est autre que l'espérance $\mathbb{E}_{W^\sigma}(H(\boldsymbol{\theta} - W^\sigma))$ que l'on notera

$$H_\sigma(\boldsymbol{\theta}) = \mathbb{E}_{W^\sigma}(H(\boldsymbol{\theta} - W^\sigma))\,.$$

Exemple 5.1.1. Par exemple, on peut considérer un noyau gaussien multidimensionnel où g_{σ^2} est la densité gaussienne multivariée centrée, de matrice de covariance $\Sigma = \mathrm{diag}(\sigma_1^2, ..., \sigma_k^2)$ avec $\sigma^2 = (\sigma_1^2, ..., \sigma_k^2)$.

Par abus de notation on écrit $H(\boldsymbol{\theta} - W^\sigma) = H(\boldsymbol{\theta}, W^\sigma)$ et on retrouve l'écriture $H_\sigma(\boldsymbol{\theta}) = \mathbb{E}_{W^\sigma}(H(\boldsymbol{\theta}, W^\sigma))$ utilisée à la section précédente avec $k^{'} = k$. C'est ce qui motive l'utilisation d'algorithmes stochastiques comme ceux donnés précédemment (à σ fixé).
Bien entendu, notre objectif est de minimiser H et non pas H_σ. C'est ce qui explique l'adjectif *dynamique* : le caractère dynamique consiste à faire varier σ "au cours du temps" (de l'algorithme), et le faire tendre vers 0. Ceci est naturel puisque "$\lim_{\sigma\to 0} H_\sigma = H$", fonction dont nous cherchons le minimum.

5.2 Algorithmes stochastiques dynamiques pour le calcul de M-estimateurs de modèles statistiques complexes

A Dynamic Stochastic Algorithm for M-estimators computation

Nabil Rachdi [1], Jean-Claude Fort [2]

Abstract
Most of statistical procedures consist in estimating parameters by minimizing (or maximizing) some criterion. A minimizing parameter is also called in the literature *M-estimator*, [6]. Depending on the statistical problem and the available information, the criterion may be more or less complicated : non convex, no gradient, non smooth etc... Thus, it can be difficult in practice to compute an *M-estimator*. We propose a new algorithm to compute the parameters, mixing stochastic algorithms and smoothness technics. We will call it S^2Dyn for *Stochastic & Smooth Dynamic* algorithm.

5.3 Introduction

Let $H : \Theta \longrightarrow \mathbb{R}$ be some mapping, or criterion function, where $\Theta = \mathbb{R}^k$. Moreover, we assume that H has a unique global minimum $\boldsymbol{\theta}$ over Θ noted $\widehat{\boldsymbol{\theta}}$. Hence the problem is to compute

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\Theta}{\text{Argmin}}\, H(\boldsymbol{\theta})\,. \tag{5.2}$$

A classical example is the *maximum likelihood estimator* where H takes the form

$$H(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log(p_{\boldsymbol{\theta}}(Y_i))\,,$$

where $Y_1, ..., Y_n$ are i.i.d random variables drawn from a distribution Q and $\{p_{\boldsymbol{\theta}},\, \boldsymbol{\theta} \in \Theta\}$ is some family of density functions. In the case of a gaussian family of mean $\mu = \boldsymbol{\theta}$ and variance $\sigma^2 = 1$, the function H becomes (see Figure 5.1(a))

$$H(\boldsymbol{\theta}) = \sum_{i=1}^{n} (Y_i - \boldsymbol{\theta})^2 + C\,,$$

where C is a constant independent of $\boldsymbol{\theta}$. Here, one computes

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\Theta}{\text{Argmin}} \sum_{i=1}^{n} (Y_i - \boldsymbol{\theta})^2 = \frac{1}{n}\sum_{i=1}^{n} Y_i\,.$$

There are also cases where a simple Newton algorithm is enough to compute $\widehat{\boldsymbol{\theta}}$. However, the function H can be complicated, for instance if it is the result of a "complex" statistical modeling. Indeed, let us consider the estimators resulting in the statistical procedures introduced in the work of N. Rachdi et al. [8] :

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\Theta}{\text{Argmin}} \sum_{i=1}^{n} \Psi\left(\sum_{j=1}^{m} \widetilde{\rho}_{\mathcal{F}}(h(\mathbf{X}'_j, \boldsymbol{\theta}))\,,\, Y_i \right) \tag{5.3}$$

1. Institut de Mathématiques de Toulouse - EADS Innovation Works, 92152 Suresnes
2. Université Paris Descartes, 45 rue des saints pères, 75006 Paris

where $Y_1, ..., Y_n$ are i.i.d random variables drawn from a distribution Q, $\mathbf{X}'_1, ..., \mathbf{X}'_m$ are i.i.d random variables drawn from a distribution $P^{\mathbf{x}}$, $\mathcal{F}$ is some feature space, $\Psi : \mathcal{F} \to L_1(Q)$ is a $\mathcal{F}$-contrast, $\widetilde{\rho}_{\mathcal{F}} : \mathcal{Y} \to \mathcal{F}$ is some weight function and h is a computer code. Here, the function H is

$$H(\boldsymbol{\theta}) = \sum_{i=1}^{n} \Psi \left(\sum_{j=1}^{m} \widetilde{\rho}_{\mathcal{F}}(h(\mathbf{X}'_j, \boldsymbol{\theta})) \, , \, Y_i \right) .$$

For instance, let us consider the case of the log-contrast $\Psi(\rho, y) := -\log(\rho(y))$ and the weight function $\widetilde{\rho}_{\mathcal{F}}(y)(\cdot) := \frac{1}{\sqrt{2\pi}\, b} \, e^{\left(\frac{\cdot - y}{b}\right)^2}$ with a bandwidth b. The bandwidth b, in fact $b_{\boldsymbol{\theta}}$, is computed from the sample $h(\mathbf{X}'_j, \boldsymbol{\theta})$, $j = 1, ..., m$ for $\boldsymbol{\theta} \in \Theta$, by Silverman's rule-of-thumb :

$$b_{\boldsymbol{\theta}} = 1.06 \, m^{-1/5} \, \hat{\sigma}_{\boldsymbol{\theta}} \, ,$$

where $\hat{\sigma}_{\boldsymbol{\theta}}$ is the empirical standard deviation of the sample $h(\mathbf{X}'_j, \boldsymbol{\theta})$, $j = 1, ..., m$.
Finally H becomes

$$H(\boldsymbol{\theta}) = -\sum_{i=1}^{n} \log \left(\sum_{j=1}^{m} e^{\left(h(\mathbf{X}'_j, \boldsymbol{\theta}) - Y_i\right)^2 / b_{\boldsymbol{\theta}}^2} \right) + C, \tag{5.4}$$

where C is a constant independent of $\boldsymbol{\theta}$.
Let us recall that h is a computer code, viewed as a *black-box*[3], which represents a physical phenomenon. Typically, h gives solutions of differential equations etc... The computation of $\widehat{\boldsymbol{\theta}}$ is given by

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \Theta}{\text{Argmin}} \; -\sum_{i=1}^{n} \log \left(\sum_{j=1}^{m} e^{\left(h(\mathbf{X}'_j, \boldsymbol{\theta}) - Y_i\right)^2 / b_{\boldsymbol{\theta}}^2} \right) . \tag{5.5}$$

We should take into account two important issues. First, the function can be highly non-convex (with many local minima), second we don't have the analytical expression of the gradient. Figure 5.1(b) shows an example of function H resulting of this modeling.

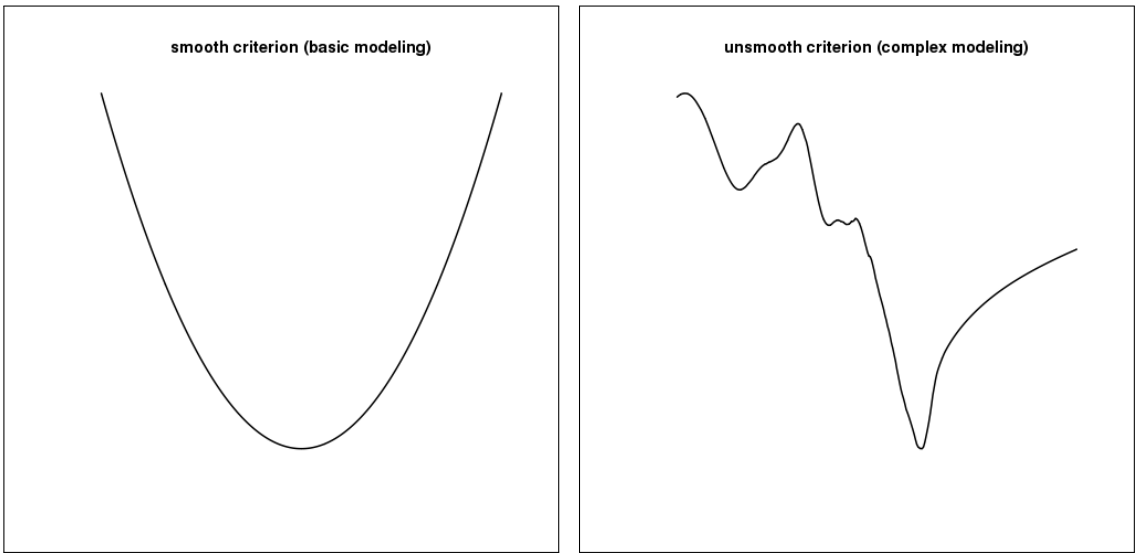
5.4 Smoothness and stochastic algorithm

In this section, we attempt to overcome irregularities or unsmoothness of the function H by making a convolution by some appropriate function, and we will see how naturally appears a stochastic algorithm.
The smoothness method is taken from [7] where the main idea is to minimize a modified function smoother than H while controlling the degree of smoothness, instead of minimizing directly H. However, when the modified function is a convolution of the criterion H with some function, one has to compute multi-dimensional integrals. This limits, in general, the use of such method in high dimensions. Moreover, in many applications, H is not analytically known, but only computable. In this paper we propose optimization procedures based on stochastic algorithms to compute the minimizer, which in any case overcomes the computation of the multi-dimensional integrals.

Let g_Σ be the gaussian probability density function w.r.t the Lebesgue measure on $\Theta = \mathbb{R}^k$, where Σ is the diagonal matrix $(k \times k)$ $\texttt{diag}(\sigma_1^2, ..., \sigma_k^2)$ and

$$\int_\Theta w \, g_\Sigma(w) \, dw = 0 \, .$$

3. black-box : function known only through its input and output values



(a) (b)

FIGURE 5.1 – (a) Function H for the mean of a gaussian with known variance. (b) Function H for a one dimensional complex statistical modeling.

For simplicity, we suppose that $\sigma^2 = \sigma_l^2$, $l = 1, ..., k$, where $\sigma > 0$, and we write $g_\Sigma = g_{\sigma^2}$. Let us denote by H_σ the convolution of H and g_{σ^2}

$$H_\sigma(\boldsymbol{\theta}) := \int_\Theta H(\boldsymbol{\theta} - w)\, g_{\sigma^2}(w) dw, \quad \sigma > 0\,. \tag{5.6}$$

By noting that $g_\sigma(\cdot) = \frac{1}{\sigma^k} g(\frac{\cdot}{\sigma})$, we show the following basic lemma.

Lemma 5.4.1. *Let $\mathcal{C}(\Theta)$ be the space of all continuous functions on Θ.*
If $H \in \mathcal{C}(\Theta)$, then

$$\forall\, \boldsymbol{\theta} \in \Theta\,, \quad H_\sigma(\boldsymbol{\theta}) \underset{\sigma \to 0}{\longrightarrow} H(\boldsymbol{\theta})\,.$$

If H is integrable, then

$$\forall\, \boldsymbol{\theta} \in \Theta\,, \quad H_\sigma(\boldsymbol{\theta}) \underset{\sigma \to +\infty}{\longrightarrow} 0\,.$$

The previous lemma shows the smoothing control of the function H by the transformation (5.6).

Remark 5.4.1. The probability density g_σ can be generalized to any other one defined as

$$g_\sigma(\cdot) = \frac{1}{\sigma^k} g\left(\frac{\cdot}{\sigma}\right)\,,$$

for any centered probability density g on $\mathbb{R}^k$.

Let us consider the following function as an academic example

$$H(\boldsymbol{\theta}) = \boldsymbol{\theta}^2 + a \sin(b\, \boldsymbol{\theta})\,, \tag{5.7}$$

for some constants $a > 0$ and $b > 0$. It is easy to show that

$$H_\sigma(\boldsymbol{\theta}) = \boldsymbol{\theta}^2 + a \sin(b\, \boldsymbol{\theta})\, e^{-(b\,\sigma)^2/2} + \sigma^2\,. \tag{5.8}$$

Remark 5.4.2. Notice that we have $H_\sigma(\boldsymbol{\theta}) \underset{\sigma \to 0}{\longrightarrow} H(\boldsymbol{\theta})$ but not $H_\sigma(\boldsymbol{\theta}) \underset{\sigma \to +\infty}{\longrightarrow} 0$. However, for large $\sigma > 0$, $H(\boldsymbol{\theta}) \approx \boldsymbol{\theta}^2 + \sigma^2$ which is very smooth.

Figure 5.2 shows the behavior of the function H_σ (with $a = 1$ and $b = 6$) with the parameter σ.

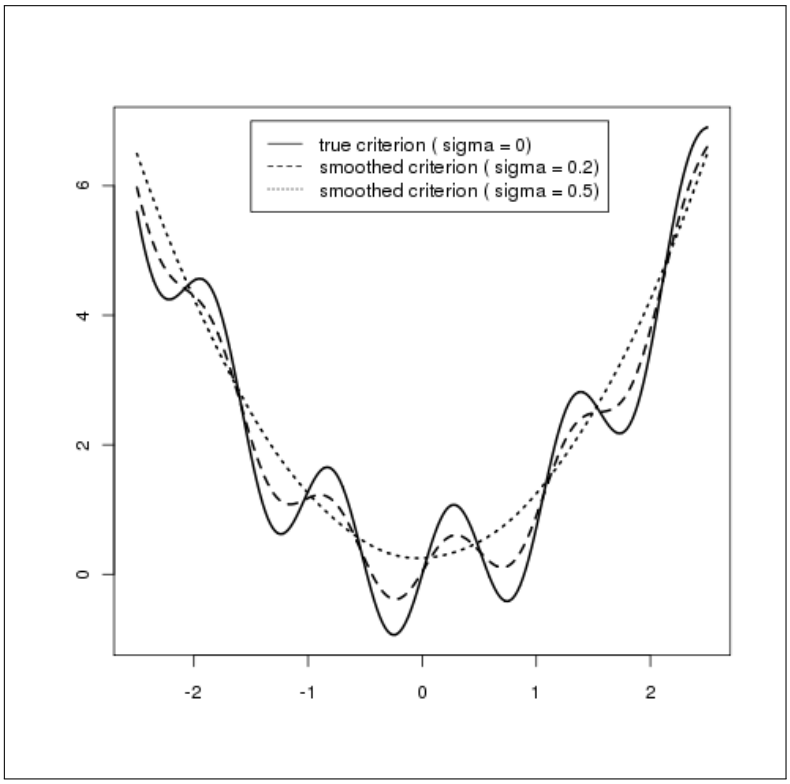


FIGURE 5.2 – Illustration of the transformation (5.8) for different values of σ, with $a = 1$ and $b = 6$.

Now the challenge is to compute the minimizer of H_σ which would be, a priori, more tractable than the minimizer of H. However, $H_\sigma(\boldsymbol{\theta})$ requires the knowledge of H which is supposed to be unknown analytically. Also, the computation of $H_\sigma(\boldsymbol{\theta})$ needs to integrate on $\Theta \in \mathbb{R}^k$ which can be difficult in general, especially if k is large. The following remark is the key of this work.
Notice that

$$H_\sigma(\boldsymbol{\theta}) = \int_\Theta H(\boldsymbol{\theta} - w)\, g_{\sigma^2}(w) dw = \mathbb{E}_{W^\sigma \sim g_{\sigma^2}} \left(H(\boldsymbol{\theta} - W^\sigma) \right) , \tag{5.9}$$

where $W^\sigma \sim g_{\sigma^2}$ means that W^σ is a random variable distributed from the density g_{σ^2}. For notational simplicity, we will denote abusively

$$H(\boldsymbol{\theta} - w) = H(\boldsymbol{\theta}, w) .$$

By the display (5.9), i.e $H_\sigma(\boldsymbol{\theta}) = \mathbb{E}_{W^\sigma \sim g_{\sigma^2}} \left(H(\boldsymbol{\theta}, W^\sigma) \right)$, we propose to use stochastic algorithms to compute the minimizer of H_σ. Given a sequence of random variables $(W_t^\sigma)_{t \geq 1}$ i.i.d from the distribution g_{σ^2}, and sequences of real numbers $(\gamma_t)_{t \geq 0}$, $(\delta_t)_{t \geq 0}$ decreasing to zero (both may depend on σ), we form the following Kiefer-Wolfowitz algorithms ([2] p. 53) for each

$\sigma > 0$:

$$(KW)\quad \begin{cases} \boldsymbol{\theta}_0^\sigma \in \Theta \\ \left(\widehat{\nabla}_t H(\boldsymbol{\theta}_t^\sigma)\right)_l = \dfrac{H(\boldsymbol{\theta}_t^\sigma + \delta_t e^l, W_t^\sigma) - H(\boldsymbol{\theta}_t^\sigma - \delta_t e^l, W_t^\sigma)}{2\,\delta_t} \\ \boldsymbol{\theta}_{t+1}^\sigma = \boldsymbol{\theta}_t^\sigma - \gamma_{t+1}\,\widehat{\nabla}_t H(\boldsymbol{\theta}_t^\sigma) \end{cases} \tag{5.10}$$

where $(e^l)_{l=1,\dots,k}$ is the canonical basis of $\mathbb{R}^k$. Let us notice that we use a single sequence $(W_t^\sigma)_{t\geq 1}$ and not a two independent sequences $(W_t^\sigma)_{t\geq 1}$ and $(\widetilde{W}_t^\sigma)_{t\geq 1}$ as the Kiefer-Wolfowitz algorithms are classically introduced.

Theorem 5.4.1. ***Classical Kiefer-Wolfowitz theorem (see Proposition 1.4.28 in [3])***
Let $f : \mathbb{R} \to \mathbb{R}$ be a function defined as $f(x) = \mathbb{E}(U(x))$ where $U(\cdot)$ is some random function. Let us define the iterative Kiefer-Wolfowitz procedure as follows

$$x_{t+1} = x_t - \gamma_{t+1}\frac{U(x_t + \delta_t) - U(x_t - \delta_t)}{\delta_t},$$

for some sequences $(\gamma_t)_t$ and $(\delta_t)_t$ decreasing to zero as $t \to +\infty$. If the three assumptions are satisfied

- $\sum_{t\geq 0}\gamma_t = +\infty$, $\sum_{t\geq 0}\left(\frac{\gamma_t}{\delta_t}\right)^2 < +\infty$.
- $\mathbb{E}(U^2(x)) \leq K(1 + x^2)$ *for some constant* K
- f *has a unique global minimum noted* x^**, is twice differentiable and strictly convex such that*

$$|f''(x)| \leq K(1 + |x|),$$

then $x_t \underset{t\to+\infty}{\longrightarrow} x^*$ (a.s).

Remark 5.4.3. Let us consider the (KW) algorithm in (5.10). Suppose that $H_\sigma \in \mathcal{C}^2(\Theta)$ and $\nabla^2 H_\sigma$ is Lipschitz and positive definite, and that the sequences $\gamma = (\gamma_t)_{t\geq 0}$, $\delta = (\delta_t)_{t\geq 0}$ decreasing to zero satisfy

- $\sum_{t\geq 0}\gamma_t = +\infty$
- $\sum_{t\geq 0}\left(\frac{\gamma_t}{\delta_t}\right)^2 < +\infty$.

Then, for all $\sigma > 0$

$$\boldsymbol{\theta}_t^\sigma \underset{t\to+\infty}{\longrightarrow} \widehat{\boldsymbol{\theta}^\sigma} := \underset{\boldsymbol{\theta}\in\Theta}{\text{Argmin}}\ H_\sigma(\boldsymbol{\theta}) \quad \textit{almost surely} \quad .$$

5.5 Stochastic & Smooth Dynamic algorithm

Let us consider the (KW) algorithm (5.10) with parameter σ depending on t. Let us denote by $\sigma := (\sigma_t)_{t\geq 0}$, $\gamma := (\gamma_t)_{t\geq 0}$, and $\delta := (\delta_t)_{t\geq 0}$ sequences of real numbers decreasing to zero. We propose the following S^2Dyn algorithm.

Algorithm 1 S^2Dyn algorithm

Require: $\sigma : t \mapsto \sigma_t$, γ, δ, $\boldsymbol{\theta}_0$, T_{dyn}

generate independent $W_1, ..., W_{T_{dyn}}$ with $W_t \sim g_{\sigma_t}$

for t from 0 to $T_{dyn} - 1$ **do**

$$\left(\widehat{\nabla}_t H(\boldsymbol{\theta}_t)\right)_l = \frac{H(\boldsymbol{\theta}_t + \delta_t e^l, W_{t+1}) - H(\boldsymbol{\theta}_t - \delta_t e^l, W_{t+1})}{2\,\delta_t}$$

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \gamma_{t+1}\,\widehat{\nabla}_t H(\boldsymbol{\theta}_t)$$

end for

return $\boldsymbol{\theta}_{T_{dyn}}$

Remark 5.5.1. The term of *Dynamic* means that the function H_{σ_t} changes in time, converging toward the "true" function H.

The function $\sigma : t \mapsto \sigma_t$ will be called *smoothing function* and we will see in the next section that its behavior is crucial for the convergence of our algorithm.

Remark 5.5.2. The stochastic process $\{\boldsymbol{\theta}_t, t \geq 0\}$ provided by the Algorithm 1 is a Markov Chain.

5.6 Simulation study

In this section, we test our algorithm on the 1D function (5.7) (with $a = 1$ and $b = 6$), and on the 2D Rosenbrock function.

5.6.1 1D example

$$H(\boldsymbol{\theta}) = \theta^2 + \sin(6\,\boldsymbol{\theta})\,.$$

H has a unique global minimum at $\boldsymbol{\theta} = -0.2424938$.
In order to be in the practical conditions mentioned in the introduction, see (5.3) and (5.5), we suppose that we don't have at disposal the gradient of H, because H is a black box, and that we cannot compute H_σ (which is in fact given by (5.8)).
Now let us consider the S^2Dyn algorithm with the following configurations.
Let the "time" be $T_{dyn} = 1500$ with time step $\Delta t = 5.10^{-2}$. Then, consider the following sequences : $\gamma_t = \dfrac{10^{-1}}{t}$ and $\delta_t = \dfrac{10^{-1}}{t^{0.4}}$ (notice that these sequences satisfy conditions of Theorem 5.4.1).

We present the evolution of $\boldsymbol{\theta}_t$ in t at 10 different starting points $\boldsymbol{\theta}_0$, for three smoothing functions.

In Figure 5.4, we consider the trivial smoothing function $\sigma_t = 0$ for all $t \in [0, T_{dyn}]$, i.e there is no *dynamic* during the time, $H_{\sigma=0} = H$. It amounts to local methods (we see that $\boldsymbol{\theta}_t$ converges to the nearest minimum).
In Figure 5.5 and Figure 5.6 we consider two others smoothing functions, where the first

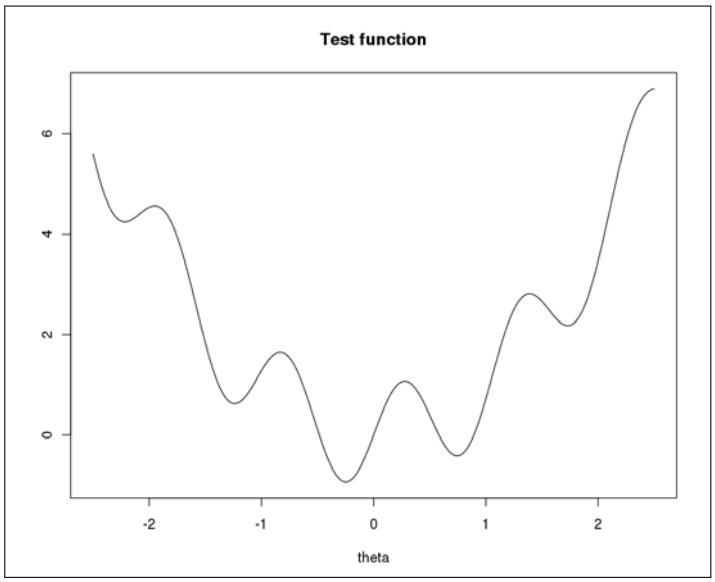


FIGURE 5.3 – test function $H(\boldsymbol{\theta}) = \theta^2 + \sin(6\,\boldsymbol{\theta})$.

function decreases rapidly and the other one decreases slowly. It appears that for a suitable function σ (which does not decrease too fast), the process $\{\boldsymbol{\theta}_t,\, t \in [0, T_{dyn}]\}$ converges (in some sense) to the minimum for all starting points (Figure 5.6 right).

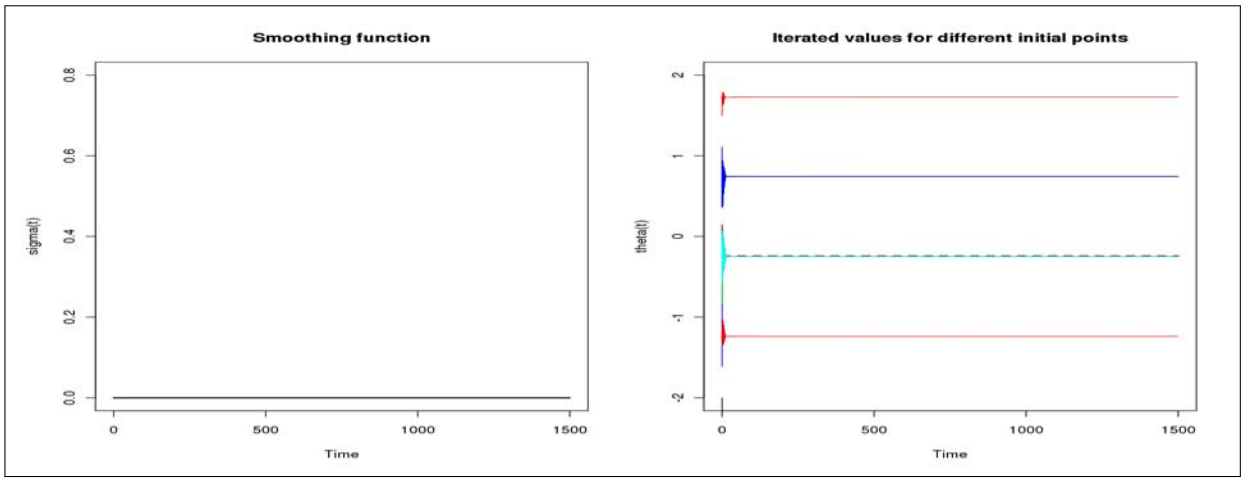


FIGURE 5.4 – Convergence of the S^2Dyn algorithm vs. behaviour of (decreasing rate of) the smoothness function σ, taken equal to 0. The graph on the right represents the convergence of the S^2Dyn for 10 initial points.

5.6.2 2D example

Now, let us consider the Rosenbrock function

$$H(\theta_1, \theta_2) = (\theta_1 - 1)^2 + 100\,(\theta_2 - \theta_1^2)^2\,.$$

H has a unique global minimum at $(\theta_1, \theta_2) = (1, 1)$.

We use the S^2Dyn algorithm with the following sequences : $\gamma_t = \dfrac{1}{10^3 + t^{0.6}}$ and $\delta_t = \dfrac{10^{-2}}{t^{0.4}}$.

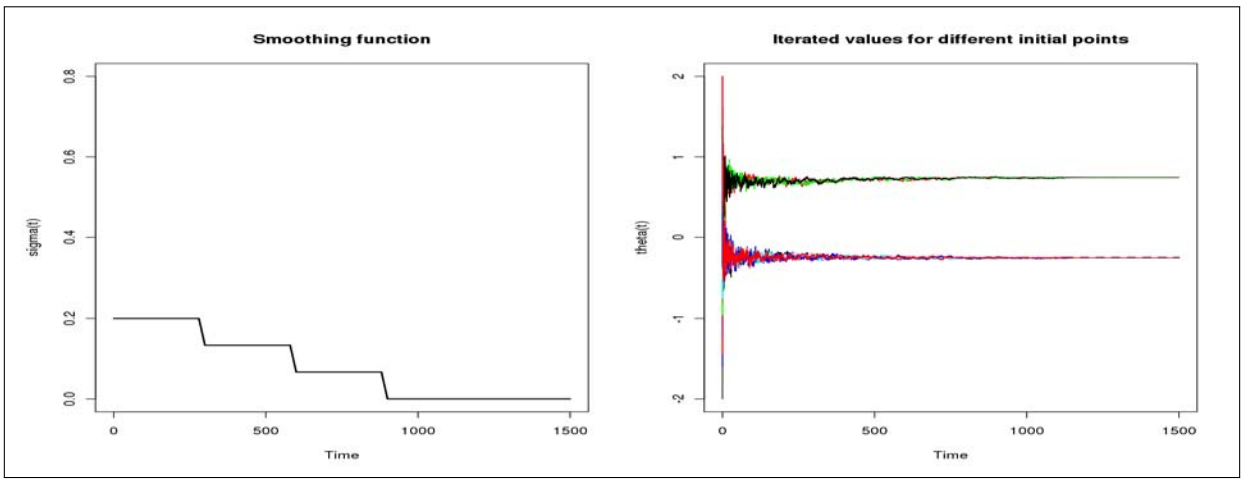


FIGURE 5.5 – Convergence of the S^2Dyn algorithm vs. behaviour of (decreasing rate of) the smoothness function σ, which decreases "rapidly". The graph on the right represents the convergence of the S^2Dyn for 10 initial points.

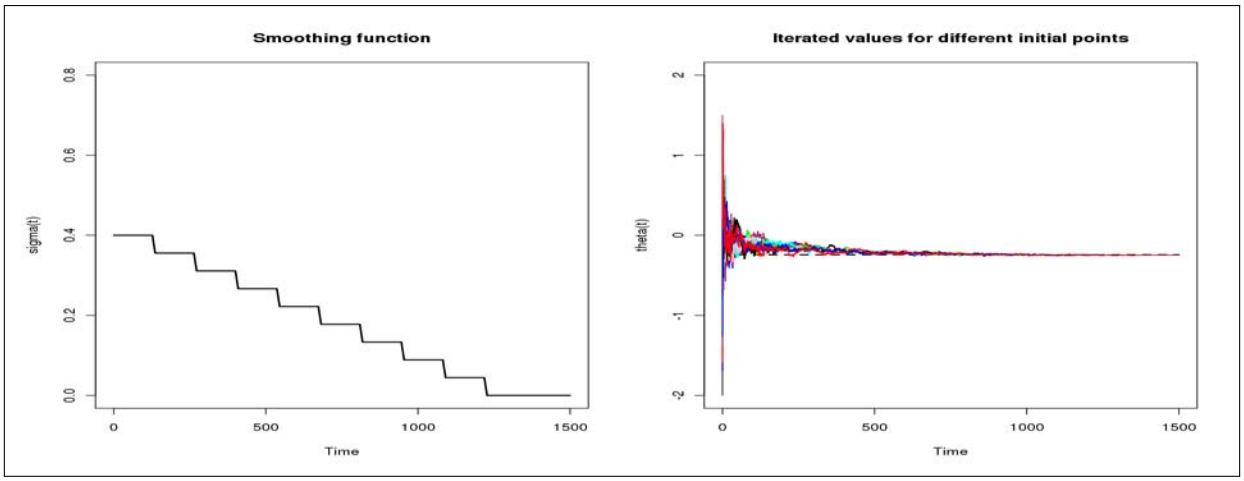


FIGURE 5.6 – Convergence of the S^2Dyn algorithm vs. behaviour of (decreasing rate of) the smoothness function σ, which decreases "slowly". The graph on the right represents the convergence of the S^2Dyn for 10 initial points.

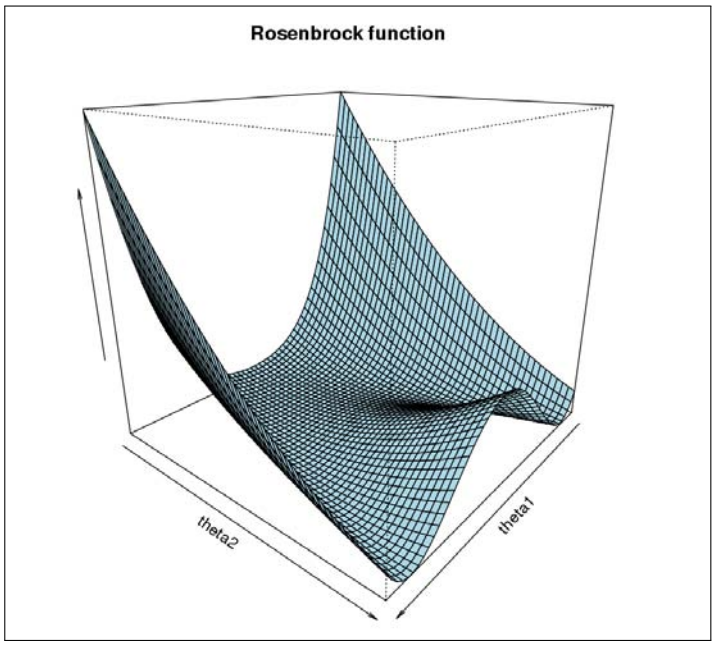
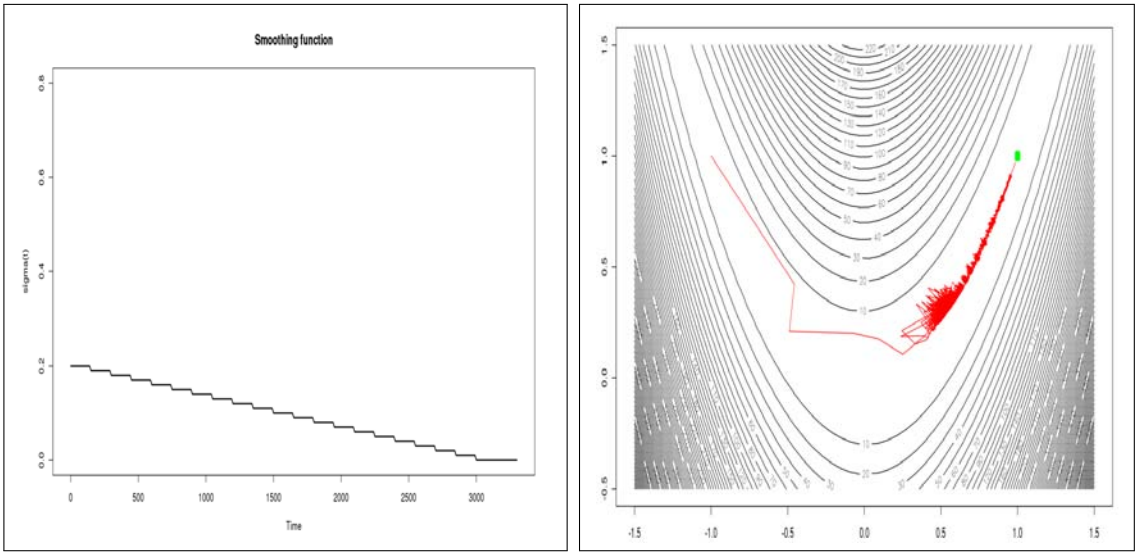


FIGURE 5.7 – Rosenbrock function

For $T_{dyn} = 3300$ and a time step $\Delta t = 5.10^{-2}$, the obtained minimum value is $(\theta_1, \theta_2)_{min} = (0.9856077, 0.9713646)$ and the Rosenbrock function evaluated at this point is $H_{min} = 2.07 \times 10^{-4}$. We check that if T_{dyn} is large enough (e.g $T_{dyn} \geq 5000$) we find $(\theta_1, \theta_2)_{min} = (1, 1)$.
Figure 5.8 shows the smoothing function used in the algorithm (5.8a) and the graph of convergence (5.8b).



(a) (b)

FIGURE 5.8 – (a) Smoothing function. (b) S^2Dyn algorithm applied to the Rosenbrock function.

5.7 Discussion

The algorithm we proposed seems to provide satisfying results on the toy examples studied. However, in order to complete our study, one has to investigate the limitations of such procedure : we have in mind for instance the case where the minimum is attained on a narrow valley located on the border of the domain. Then, an interesting issue would be to determine theoretical properties of the S^2dyn algorithm in the same spirit as Kiefer-Wolfowitz Theorem 5.4.1, where the specificity is the fact that σ depends on the time t, $\sigma = \sigma_t$. In particular, the rate of decreasing of σ_t may play a crucial rule.

Bibliographie

[1] M. Benaïm. Dynamics of stochastic approximation algorithms. *Seminaire de probabilites XXXIII*, pages 1–68, 1999.

[2] M. Duflo. *Algorithmes stochastiques*. Springer, 1996.

[3] M. Duflo. *Random iterative models*, volume 34. Springer Verlag, 1997.

[4] J.C. Fort and G. Pages. Asymptotic behavior of a markovian stochastic algorithm with constant step. *SIAM journal on control and optimization*, 37 :1456, 1999.

[5] J.C. Fort and G. Pagès. Decreasing step stochastic algorithms : As behaviour of weighted empirical measures. *Monte Carlo Methods and Applications*, 8(3) :237–270, 2002.

[6] P.J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, pages 73–101, 1964.

[7] J.J. Moré and Z. Wu. Global continuation for distance geometry problems. *Preprint Mcs-p, SIAM J. Optimization*, 7(7) :814–836, 1995.

[8] N. Rachdi, J.C. Fort, and T. Klein. Risk bounds for new M-estimation problems . *hal 00537236*, 2010.

[9] H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, pages 400–407, 1951.

Chapitre 6

Problème inverse stochastique : application à un modèle aéronautique

Sommaire

Résumé du Chapitre

Ce chapitre a pour but d'appliquer certaines notions rencontrées dans les chapitres précédents sur une problématique industrielle. On présente une application en ingénierie des turboréacteurs où le problème est la caractérisation robuste sous incertitude d'un paramètre d'intérêt du réacteur.

Stochastic Inverse Problem with Noisy Simulator - Application to aeronautic model -

Nabil Rachdi [1], Jean-Claude Fort [2], Thierry Klein [3]

Abstract
Inverse problem is a current practice in engineering where the goal is to identify parameters from observed data through numerical models. These numerical models, also called *Simulators*, are built to represent the phenomenon making possible the inference. However, such representation can include some part of variability or commonly called *uncertainty* (see [2]), due to some variables of the model. The phenomenon we study is the fuel mass needed to link two fixed countries with a commercial aircraft, where we only consider the *Cruise* phase [4].
From a data base of fuel mass consumptions during the cruise phase, we aim at identifying the *specific fuel consumption SFC* in a robust way, giving the uncertainty of the *cruise speed V* and the *lift-to-drag F*.
In this paper, we propose an estimation procedure based on *M*-estimation, taking into account this uncertainty.

Résumé
Le problème inverse est une pratique assez courante en ingénierie, où le but est de déterminer les causes d'un certain phénomène à partir d'observations de ce dernier. Le phénomène mis en jeu est représenté par un modèle numérique, dont certaines composantes peuvent comporter une part de variabilité (voir [2]). Le phénomène étudié est la masse de fuel nécessaire pour effectuer une liaison fixée avec un avion commercial, en ne considérant que la phase de *Croisière*. Le but étant, à partir de données de masses de fuel consommées en croisière, d'identifier de manière robuste la consommation spécifique *SFC* de la motorisation en tenant compte de l'incertitude sur la vitesse de croisière *V* et sur la *finesse F* de l'avion.
Dans cet article, nous proposons une procédure d'estimation basée sur la *M*-estimation, prenant en compte cette incertitude.

6.1 Introduction

One of engineering activities is to model real phenomena. Once a model is built (physical principles, state equations etc...), there are some parameters that have to be identified and some variables of the model may present some intrinsic variability. Hence, the identification of parameters should implicitly take into account the uncertainty of variables of the model. In this paper, we present a likelihood-based method to estimate aeronautic parameters in a *Fuel mass* model. We use an analytical model that can be viewed as a *black-box simulator*. From a data base $M_{fuel}^{*,1}, ..., M_{fuel}^{*,n}$ giving the mass of fuel consumed for n simulated lines

1. Institut de Mathématiques de Toulouse - EADS Innovation Works, 92152 Suresnes
2. Université Paris Descartes, 45 rue des saints pères, 75006 Paris
3. Institut de Mathématiques de Toulouse, 118 route de Narbonne F-31062 Toulouse
4. we don't consider the take-off and landing phases

between two fixed cities with a specific commercial aircraft, we aim at identifying the *specific fuel consumption* (SFC) which corresponds to a characteristic value of engines. The model we use depends in particular on the *cruise speed* (V) and on the *lift-to-drag ratio* (F). These variables present intrinsic variability in that the cruise speed may depend on atmospheric conditions, and the lift-to-drag is also subjected to variability potentially caused by turbulent phenomena. As a matter of fact, the identification of the parameter SFC should take into account the variability of the cruise speed and the lift-to-drag ratio.
In this paper, we propose an algorithm taken from the work of N. Rachdi *et al.* [4] allowing a characterization of SFC from the observed data $M_{fuel}^{*,1}, ..., M_{fuel}^{*,n}$ and model simulations.
This article is organized as follows. In Section 6.2 we describe the setting of the problem. In Section 6.3 we build the algorithm for the inverse problem with a Maximum-Likelihood based method. In Section 6.4 we apply the algorithm given in Section 6.3. In Section 6.5 we try to illustrate the effect of modeling conditions, particularly the random modeling on SFC and the number of observed data. In Section 6.6 we establish the Theorem 6.6.1 providing an upper bound of the estimation error of the proposed algorithm. Section 6.7 is devoted to the proof of the Theorem 6.6.1.

6.2 General setting

6.2.1 Observations

In our study, the data $M_{fuel}^{*,1}, ..., M_{fuel}^{*,n}$ were generated from an aeronautic software which simulates gas turbine configurations used for power generation. In particular, it can simulate the consumed mass of fuel at some configuration of engines, altitude, speed, atmospheric conditions etc... See Figure 6.1.

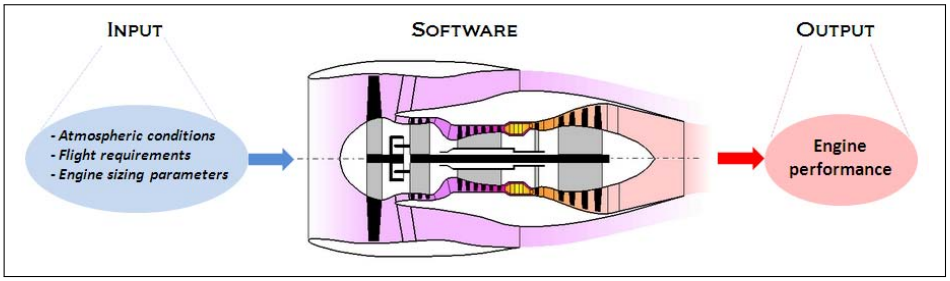


FIGURE 6.1 – Aeronautic software

For our purpose, we have generated $n = 32$ data by varying atmospheric conditions. The data are given in the Table 6.1.

Reference Fuel Masses [kg]							
7918	7671	7719	7839	7912	7963	7693	7815
7872	7679	8013	7935	7794	8045	7671	7985
7755	7658	7684	7658	7690	7700	7876	7769
8058	7710	7746	7698	7666	7749	7764	7667

TABLE 6.1 – Simulated mass of fuel consumptions from aeronautic software

Next, we will suppose that the observations $M_{fuel}^{*,1}, ..., M_{fuel}^{*,n}$ are drawn from an unknown

probability distribution Q with associated Lebesgue density f with support

$$\mathcal{I} := [M_{inf} = 7600, M_{sup} = 8100] .$$

The difference $M_{sup} - M_{inf} = 500\,kg$ have to be thought as an overconsumption about approximatively 7% .

6.2.2 The aeronautic model

We recall that we are interested in identifying the specific fuel consumption SFC that is a significant factor determining the fuel efficiency of a particular engine. The *Fuel mass* model we consider is given by the Bréguet formula :

$$M_{fuel} = \left(M_{empty} + M_{pload}\right) \left(e^{\frac{SFC \cdot g \cdot Ra}{V \cdot F}\, 10^{-3}} - 1\right) . \tag{6.1}$$

The fixed variables are

- M_{empty} : *Empty weight* = basic weight of the aircraft (excluding fuel and passengers)
- M_{pload} : *Payload* = maximal carrying capacity of the aircraft
- g : Gravitational constant
- Ra : *Range* = distance traveled by the aircraft

The uncertain variables already mentioned in the introduction are

- V : *Cruise speed* = aircraft speed between ascent and descent phase
- F : *Lift-to-drag ratio* = aerodynamic coefficient

The Table 6.2 gives the values of fixed variables and nominal values considered for uncertain variables.

input	value or nominal value	unit
M_{empty}	42600	kg
M_{pload}	19900	kg
g	9.8	m/s^2
Ra	3000	km
V_{nom}	231	m/s
F_{nom}	19	–

TABLE 6.2 – Values of Fuel mass model inputs

6.2.3 Noise modeling

As said in the introduction, we have to take into account the uncertainty of the cruise speed V and the lift-to-drag F. Giving the nominal value of each variable in the Table 6.2, an expert judgment provides the uncertainty bounds in Table 6.3.

variable	nominal value	min	max
V	**231**	226	234
F	**19**	18.7	19.05

TABLE 6.3 – Minimal and maximal values of uncertain variables

The uncertainty on the cruise speed V represents a relative difference of arrival time of 8 minutes.

We propose to model the uncertainties as presented in Table 6.4.

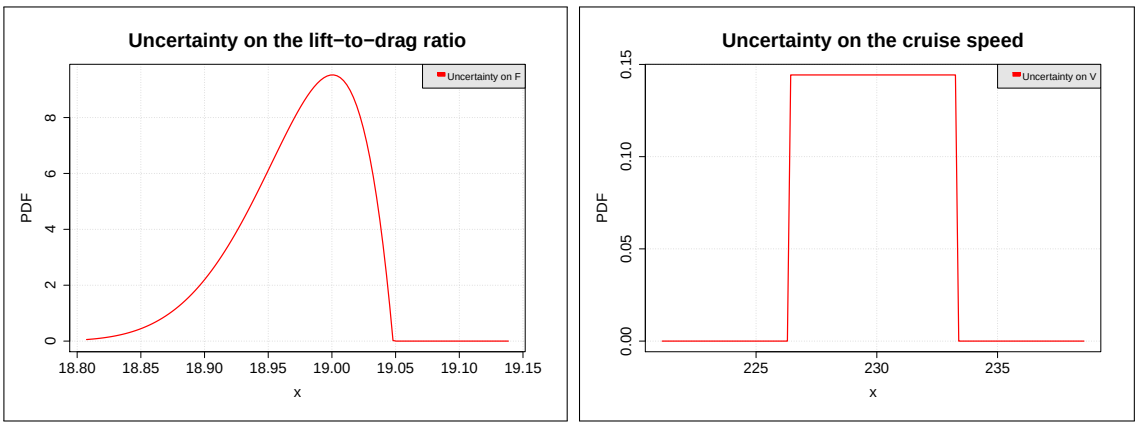
variable	distribution	parameter
V	*Uniform*	(V_{min}, V_{max})
F	*Beta*	$(7, 2, F_{min}, F_{max})$

TABLE 6.4 – Uncertainty modeling

The probability density function of a beta distribution on $[a, b]$ with shape parameters (α, β) is

$$g_{(\alpha,\beta,a,b)}(x) = \frac{(x-a)^{(\alpha-1)}(b-x)^{\beta-1}}{(b-a)^{\beta-1}B(\alpha,\beta)} \mathbb{1}_{[a,b](x)} ,$$

where $B(\cdot,\cdot)$ is the beta function.
The Figure 6.2 shows the probability density functions of V and F.



(a) (b)

FIGURE 6.2 – (a) Uncertainty on F. (b) Uncertainty on V.

In order to emphasize the "noisy" feature of the variables V and F, we will use the writing
– $V = V_{nom} + \epsilon_V$
– $F = F_{nom} + \epsilon_F$
where ϵ_V is a centered uniform random variable on the interval $[\epsilon_V^{min}, \epsilon_V^{max}]$ with

$$\epsilon_V^{min} = V_{min} - V_{nom} \quad and \quad \epsilon_V^{max} = V_{max} - V_{nom} .$$

The variable ϵ_F, supposed to be independent of ϵ_V, is a beta random variable on the interval $[\epsilon_F^{min}, \epsilon_F^{max}]$ with shape parameters $(7, 2)$ where

$$\epsilon_F^{min} = F_{min} - F_{nom} \quad and \quad \epsilon_F^{max} = F_{max} - F_{nom} .$$

6.2.4 Robust identification of *SFC*

In our developments, we won't consider that the parameter *SFC* is deterministic but it will be supposed random. Indeed, beyond the fact of obtaining a value of the specific fuel consumption, we also want to take into account its own variability in order to have a robust characterization of this parameter.

Let us suppose that SFC is a gaussian random variable

$$SFC \sim \mathcal{N}(\mu_{SFC}, \sigma^2_{SFC}) \tag{6.2}$$

with unknown parameters μ_{SFC} and σ_{SFC}.
We give the following ranges of variation

$$\mu_{SFC} \in [15, 20] \quad \text{and} \quad \sigma_{SFC} \in]0, 1] \,.$$

For what follows, let us consider the writing

$$SFC = \mu_{SFC} + \sigma_{SFC}\, \epsilon_{SFC} \,, \quad \epsilon_{SFC} \sim \mathcal{N}(0, 1) \,.$$

Now, our problem amounts to estimate the location parameter μ_{SFC} and the standard deviation σ_{SFC} .

6.2.5 Statistical modeling

Let us denote by $(\mathcal{E}, \mathbb{P}^{\boldsymbol{\epsilon}})$ the probability space associated to the *noise vector*

$$\boldsymbol{\epsilon} = (\epsilon_{SFC}, \epsilon_V, \epsilon_F)^T \,,$$

and denote the vector of parameters by

$$\boldsymbol{\theta} = (\mu_{SFC}, \sigma_{SFC})^T \,.$$

Then, let us consider the mass of fuel M_{fuel} as the function

$$M_{fuel} = h(\boldsymbol{\epsilon}, \boldsymbol{\theta}) \,,$$

where $h : (\mathcal{E}, \mathbb{P}^{\boldsymbol{\epsilon}}) \times \Theta \to \mathcal{I}_h$ is given by

$$h(\boldsymbol{\epsilon}, \boldsymbol{\theta}) = \left(M_{empty} + M_{pload}\right) \left(exp\left(\frac{(\mu_{SFC} + \sigma_{SFC}\, \epsilon_{SFC}) \cdot g \cdot Ra}{(V_{nom} + \epsilon_V) \cdot (F_{nom} + \epsilon_F)} \cdot 10^{-3} \right) - 1 \right) \tag{6.3}$$

with

$$\Theta = [15, 20] \times]0, 1]$$

(for compactness reason, the interval $]0, 1]$ can be replaced by $[s, 1]$ with a small $s > 0$) and $\mathcal{I}_h$ is the interval

$$\mathcal{I}_h = h(\mathcal{E}, \Theta) = [M^h_{inf}, M^h_{sup}] \,. \tag{6.4}$$

We denote by $|\mathcal{I}_h|$ its length.

Remark 6.2.1. Simulations give us that

$$\mathcal{I} \subset \mathcal{I}_h \,,$$

where $\mathcal{I} = [M_{inf}, M_{sup}]$ is the observation interval given above.

Now, the purpose is to estimate the parameter $\boldsymbol{\theta} \in \Theta$ from the set of data $M^{*,1}_{fuel}, ..., M^{*,n}_{fuel}$. In the next section, we propose an estimation procedure taken from [4].

6.3 Parameter estimation

The framework previously set allows us to apply the procedures developed in [4]. In particular, we choose to work with the log $-$contrast which can be understood as a Maximum Likelihood based estimation.

Let us suppose that the sample $M_{fuel}^{*,1}, ..., M_{fuel}^{*,n}$ is drawn from a distribution Q such that

$$Q \in \{Q_{\boldsymbol{\theta}}\,,\, \boldsymbol{\theta} \in \Theta\}$$

where $Q_{\boldsymbol{\theta}}$ is the *pushforward measure* of $\mathbb{P}^{\boldsymbol{\epsilon}}$ by the (measurable) application $\mathbf{u} \mapsto h(\mathbf{u}, \boldsymbol{\theta})$. Then, denoting by $\rho_{\boldsymbol{\theta}}$ the Lebesgue density associated to the measure $Q_{\boldsymbol{\theta}}$, the maximum likelihood procedure is given by

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\Theta}{\text{Argmin}} -\frac{1}{n}\sum_{i=1}^{n} \log(\rho_{\boldsymbol{\theta}}(M_{fuel}^{*,i}))\,. \tag{6.5}$$

However, the above procedure is unfeasible because the density $\rho_{\boldsymbol{\theta}}$ is not analytically tractable. As suggested in [4], we replace $\rho_{\boldsymbol{\theta}}$ by an estimator noted $\rho_{\boldsymbol{\theta}}^m$ and built as follows.

Let $\boldsymbol{\epsilon}_1, ..., \boldsymbol{\epsilon}_m$ be m random variables i.i.d from $\mathbb{P}^{\boldsymbol{\epsilon}}$ and consider the kernel smoothing

$$\rho_{\boldsymbol{\theta}}^m(\cdot) \;=\; \frac{1}{m}\sum_{j=1}^{m} K_{b_{\boldsymbol{\theta}}^m}\left(\cdot - h(\boldsymbol{\epsilon}_j, \boldsymbol{\theta})\right)\,, \tag{6.6}$$

where $K_{b_{\boldsymbol{\theta}}^m}$ is the gaussian kernel

$$K_{b_{\boldsymbol{\theta}}^m}(x) = \frac{1}{\sqrt{2\,\pi}\, b_{\boldsymbol{\theta}}^m} e^{-\frac{x^2}{2\,(b_{\boldsymbol{\theta}}^m)^2}}\,,$$

and $b_{\boldsymbol{\theta}}^m$ is computed from the sample $h(\boldsymbol{\epsilon}_j, \boldsymbol{\theta})$, $j = 1, ..., m$ for $\boldsymbol{\theta} \in \Theta$, by Silverman's rule-of-thumb :

$$b_{\boldsymbol{\theta}}^m = 1.06\, m^{-1/5}\, \hat{\sigma}_{\boldsymbol{\theta}}\,. \tag{6.7}$$

The quantity $\hat{\sigma}_{\boldsymbol{\theta}}$ is the empirical standard deviation of the sample $h(\boldsymbol{\epsilon}_j, \boldsymbol{\theta})$, $j = 1, ..., m$

$$\hat{\sigma}_{\boldsymbol{\theta}} = \frac{1}{m}\sum_{j=1}^{m}\left(h(\boldsymbol{\epsilon}_j, \boldsymbol{\theta}) - \frac{1}{m}\sum_{j=1}^{m} h(\boldsymbol{\epsilon}_j, \boldsymbol{\theta})\right)^2\,.$$

By replacing $\rho_{\boldsymbol{\theta}}$ by $\rho_{\boldsymbol{\theta}}^m$ in (6.5) and simplifying by the multiplying constant $1/n$ yields the estimation procedure

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\Theta}{\text{Argmin}} -\sum_{i=1}^{n} \log\left(\frac{1}{m}\sum_{j=1}^{m} K_{b_{\boldsymbol{\theta}}^m}\left(h(\boldsymbol{\epsilon}_j, \boldsymbol{\theta}) - M_{fuel}^{*,i}\right)\right)\,. \tag{6.8}$$

In the following section, we provide a numerical analysis using the algorithm given by (6.8). Theoretical aspect will be addressed in Section 6.6.

6.4 Numerical study : first approach

6.4.1 Estimation

Setting

$$J(\boldsymbol{\theta}) = -\sum_{i=1}^{n} \log\left(\frac{1}{m}\sum_{j=1}^{m} K_{b_{\boldsymbol{\theta}}^m}\left(h(\boldsymbol{\epsilon}_j,\boldsymbol{\theta}) - M_{fuel}^{*,i}\right)\right) \quad \text{with} \quad \boldsymbol{\theta} = (\mu_{SFC},\sigma_{SFC})^T,$$

our problem is a minimization problem where we want to compute

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\Theta}{\text{Argmin}}\, J(\boldsymbol{\theta}).$$

We recall that $n = 32$, the data $(M_{fuel}^{*,i})_{i=1,...,n}$ are given in Table 6.1, we choose $m = 10000$ and for $j = 1,...,m$, $\boldsymbol{\epsilon}_j \sim \mathbb{P}^{\boldsymbol{\epsilon}}$ where

$$\mathbb{P}^{\boldsymbol{\epsilon}}(du,dv,dw) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}\, g_{(7,2,\epsilon_F^{min},\epsilon_F^{max})}(v)\, \frac{1}{\epsilon_V^{max}-\epsilon_V^{min}} \mathbb{1}_{[\epsilon_V^{min},\epsilon_V^{max}]}(w)\, du\, dv\, dw.$$

This optimization procedure can be solved by Quasi-Newton methods. We present in Table 6.5 the obtained results. In Figure 6.3 we show the resulting probability density function of

estimator	value of $J(\widehat{\boldsymbol{\theta}})$	estimated SFC location	estimated SFC dispersion
$\widehat{\boldsymbol{\theta}}$	$J(\widehat{\boldsymbol{\theta}}) = 199.465$	$\widehat{\mu}_{\mathbf{SFC}} = \mathbf{17.397}$	$\widehat{\sigma}_{\mathbf{SFC}} = \mathbf{0.201}$

TABLE 6.5 – SFC characterization parameters

SFC given by $\mathcal{N}(\widehat{\mu}_{SFC},\widehat{\sigma}_{SFC})$.

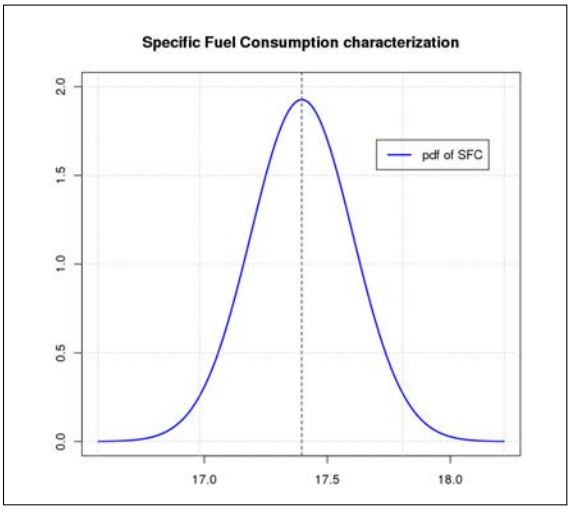
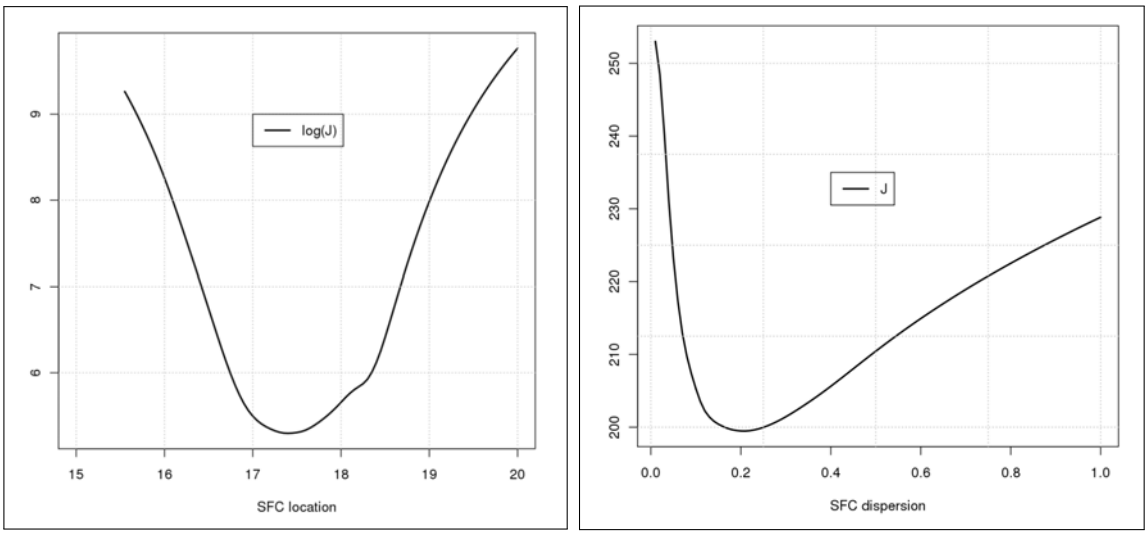


FIGURE 6.3 – Estimated Specific Fuel Consumption distribution.

We present in Figure 6.4 profile views of the criterion function $\boldsymbol{\theta} = (\mu_{SFC},\sigma_{SFC}) \mapsto J(\boldsymbol{\theta})$, first at $\sigma_{SFC} = \widehat{\sigma}_{SFC}$ (Figure 6.4(a), we show $\log(J(\boldsymbol{\theta}))$) and then at $\mu_{SFC} = \widehat{\mu}_{SFC}$ (Figure 6.4(b)). We notice that the minimum $\widehat{\boldsymbol{\theta}} = (\widehat{\mu}_{SFC},\widehat{\sigma}_{SFC})$ is well located.



(a) (b)

FIGURE 6.4 – (a) Profile view of $\log(J)$ at $\sigma_{SFC} = \widehat{\sigma}_{SFC}$. (b) Profile view of J at $\mu_{SFC} = \widehat{\mu}_{SFC}$.

6.4.2 Comparison with reference sample

In order to compare the results obtained in the previous subsection, we need some reference sample of Specific Fuel Consumption values at the same simulation conditions. The aeronautic software, described in the introduction, provides a sample of SFC values of size about 200 with the characteristics given in Table 6.6. The data in Table 6.6 have to be compared with those in Table 6.5 where the mean and standard deviation are $\widehat{\mu}_{SFC} = 17.397$ and $\widehat{\sigma}_{SFC} = 0.201$, respectively. The Table 6.7 provides the associated relative errors and the Figure 6.5 shows the histogram of the reference sample and the estimated distribution of SFC obtained in Figure 6.3.

	Mean	Stand. dev.
Reference sample	**17.49**	**0.57**

TABLE 6.6 – Reference sample characteristics

	Reference sample	Estimated SFC (6.3)	Relative error
Mean	**17.49**	**17.397**	**0.5 %**
Stand. dev.	**0.57**	**0.201**	**60.6 %**

TABLE 6.7 – Relative errors of the mean and deviation between reference SFC sample and the estimated model (6.3)

It seems that the location of the variable of interest SFC is well reached whereas the standard deviation estimation provides an error of 60%. The "error" has roughly two origins :

Statistical error : this error is due to the limited number of data : $n = 32$ for observed masses of fuel $M^{*,i}_{fuel}$ and $m = 10000$ for the ϵ_j's.

Model error : this error is relative to the use of Fuel mass model (6.1) with uncertain variables V and F (Figure 6.2), and includes the gaussian hypothesis for SFC (6.2). Thus, the model error can be separated into 2 parts : *physical* model error and *uncertainty modeling* error.

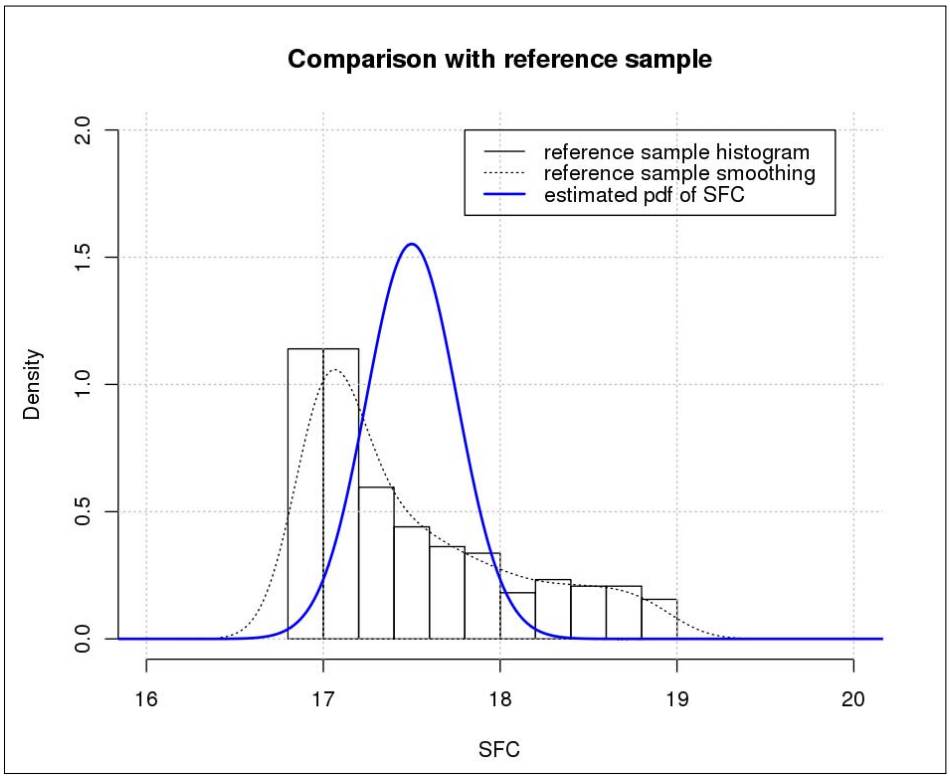


FIGURE 6.5 – Reference and estimated SFC distributions.

In Figure 6.5 we see that the SFC doesn't behave like a gaussian variable. This can be qualified as *model error*.
However, if one just wants to estimate the mean value of SFC, the gaussian hypothesis doesn't have a significative impact (0.5% of error). Nevertheless, if one wants more information about SFC, other modeling tools are needed to allow a robust characterization approach.

In the next subsection, we will discuss the *uncertainty modeling*, more precisely, the gaussian hypothesis for SFC given by (6.2).

6.5 On the probabilistic modeling of SFC

6.5.1 Considering Wiener-Hermite representation in the previous analysis

The characterization of a random variable by the mean and the standard deviation only, could be too approximative in order to study the whole behavior of the variable. In this study, we have made an *a priori* (a model) on the variable of interest SFC. In (6.2) we supposed that

$$SFC \sim \mathcal{N}(\mu_{SFC}, \sigma^2_{SFC}) \,,$$

which we rewrite

$$SFC = \mu_{SFC} + \sigma_{SFC}\, \xi \,, \quad \xi \sim \mathcal{N}(0,1) \,. \tag{6.9}$$

We will see that this gaussian hypothesis on SFC is a particular case of a more general representation.

The so called *Wiener Chaos Expansion*, developed in the 30's by Wiener [8], gives a representation of any second-order random variable Z :

$$Z = \sum_{l=0}^{\infty} z_l \mathrm{Y}_l((\xi_k)_{k\geq 1}) \,, \quad \text{(with convergence in } L_2(\mathbb{P}) \text{)} \tag{6.10}$$

where $(\xi_k)_{k\geq 1}$ is a (infinite) sequence of independent standard normal random variables and the Y_l's are the multivariate Hermite polynomials. This expansion is also called *Wiener-Hermite expansion*.
In practice, we have to consider a finite sequence $(\xi_1, ..., \xi_M)$ where M is called the *order* of the expansion, and the sum in (6.10) is truncated at p which is the *degree* of the expansion.
Hence, considering all M-dimensional Hermite polynomials of degree lower than p, the representation (6.10) is approximated by

$$Z \simeq Z^{p,M} = \sum_{l=0}^{P-1} z_l \mathrm{Y}_l(\boldsymbol{\xi}) \,, \quad \boldsymbol{\xi} = (\xi_1, ..., \xi_M) \,, \tag{6.11}$$

where

$$P = \frac{(M+p)!}{M!\, p!} \,.$$

The integer P corresponds to the number of coefficients to be estimated.
Moreover, one can notice that by orthogonality arguments in (6.11), we have

$$\mathbb{E}(Z^{p,M}) = z_0 = \mathbb{E}(Z) \tag{6.12}$$

and

$$\mathrm{Var}(Z^{p,M}) = \sum_{l=1}^{P-1} z_l^2 \,. \tag{6.13}$$

Let us notice that by the trivial decomposition

$$Z^{p,M} = Z + \left(Z^{p,M} - Z \right) \,,$$

each choice of p and M will induce a *model error*

$$mod_{err} := Z^{p,M} - Z \,.$$

We illustrate this aspect concerning SFC in the next subsection.

6.5.2 Application to the Specific Fuel Consumption

In our purpose, if we suppose that $\mathbb{E}(SFC^2) < \infty$ (it is implicitly supposed in the gaussian hypothesis), we can set the following modeling

$$SFC^{p,M} = \sum_{l=0}^{P-1} z_l \mathrm{Y}_l(\boldsymbol{\xi}) \,, \quad \boldsymbol{\xi} = (\xi_1, ..., \xi_M) \,, \quad M,\, p \geq 1$$

which we rewrite by (6.12)

$$SFC^{p,M} = \mu_{SFC} + \sum_{l=1}^{P-1} z_l \mathrm{Y}_l(\boldsymbol{\xi}) \,, \quad \boldsymbol{\xi} = (\xi_1, ..., \xi_M) \,, \quad M,\, p \geq 1 \,. \tag{6.14}$$

It appears now that the gaussian representation (6.9) is the particular case of (6.14) with $p = 1$ and $M = 1$. Moreover, in view of the Wiener representation (6.10), the gaussian one (6.9) may lead to a large approximation (if SFC is not gaussian) and thus contribute to a non negligible model error. We clearly see this in the Figure 6.5 where the reference data don't seem to be drawn from a gaussian distribution.
As a matter of fact, one can hope to reduce the model error (described in the previous subsection), at least the error corresponding to SFC modeling, by considering a less restrictive representation (6.14) with some appropriate order $M \geq 1$ and degree $p \geq 1$.
Let us consider the Wiener-Hermite expansion of order $M = 2$ and degree $p = 2$

$$SFC^{2,2} = \mu_{SFC} + \sum_{l=1}^{5} \theta_l \Upsilon_l(\boldsymbol{\xi}), \quad \boldsymbol{\xi} = (\xi_1, \xi_2)$$

or

$$SFC^{2,2} = \mu_{SFC} + \theta_1\xi_1 + \theta_2\xi_2 + \theta_3\xi_1\xi_2 + \theta_4(\xi_1^2 - 1) + \theta_5(\xi_2^2 - 1), \tag{6.15}$$

that leads to estimate $P = \frac{(2+2)!}{2!2!} = 6$ coefficients. The Table 6.8 shows the result obtained by the algorithm developed in the previous section where we change the function $h(\boldsymbol{\epsilon}, \boldsymbol{\theta})$ in (6.3) replacing $\sigma_{SFC}\epsilon_{SFC}$ by $\theta_1\xi_1 + \theta_2\xi_2 + \theta_3\xi_1\xi_2 + \theta_4(\xi_1^2 - 1) + \theta_5(\xi_2^2 - 1)$, with $\boldsymbol{\epsilon} = (\xi_1, \xi_2, \epsilon_V, \epsilon_F)$ and $\boldsymbol{\theta} = (\mu_{SFC}, \theta_1, ..., \theta_5)$.

	θ_0	θ_1	θ_2	θ_3	θ_4	θ_5
$SFC^{2,2}$	**17.470**	**0.047**	**0.054**	**0.182**	**0.103**	**0.063**

TABLE 6.8 – SFC characterization parameters

	Reference sample	from $SFC^{2,2}$	Relative error
Mean	**17.49**	**17.470**	**0.11 %**
Stand. dev.	**0.57**	**0.230**	**59.65 %**

TABLE 6.9 – Relative errors of the mean and standard deviation with $SFC^{2,2}$

Let us compare the relative errors of the first two statistical moments by considering $SFC^{1,1}$ (i,e the gaussian hypothesis Table 6.7) and $SFC^{2,2}$ (Table 6.9).
The Wiener-Hermite modeling seems to improve the mean estimation of SFC whereas the standard deviation is worst estimated in the two cases with an error of about 60%. There is no significative difference of the two methods regarding the first two moments. However, the behavior of density functions corresponding to $SFC^{1,1}$ (see Figure 6.5) and $SFC^{2,2}$ is clearly not the same. We present in the Figure 6.6 the result obtained when SFC is modeled by a Wiener expansion of order $M = 2$ and degree $p = 2$.
The distribution of SFC given by the Wiener expansion in Figure 6.6 seems to have a behavior close to the reference sample one, despite the fact that there is a non negligible bias. As mentioned in the previous subsection, this is due to the statistical and model errors. Indeed, let us recall that we have at disposal $n = 32$ reference fuel masses from which we characterize the Specific Fuel Consumption. It would be interesting to see what happens by adding reference fuel masses, i.e by reducing the statistical error.

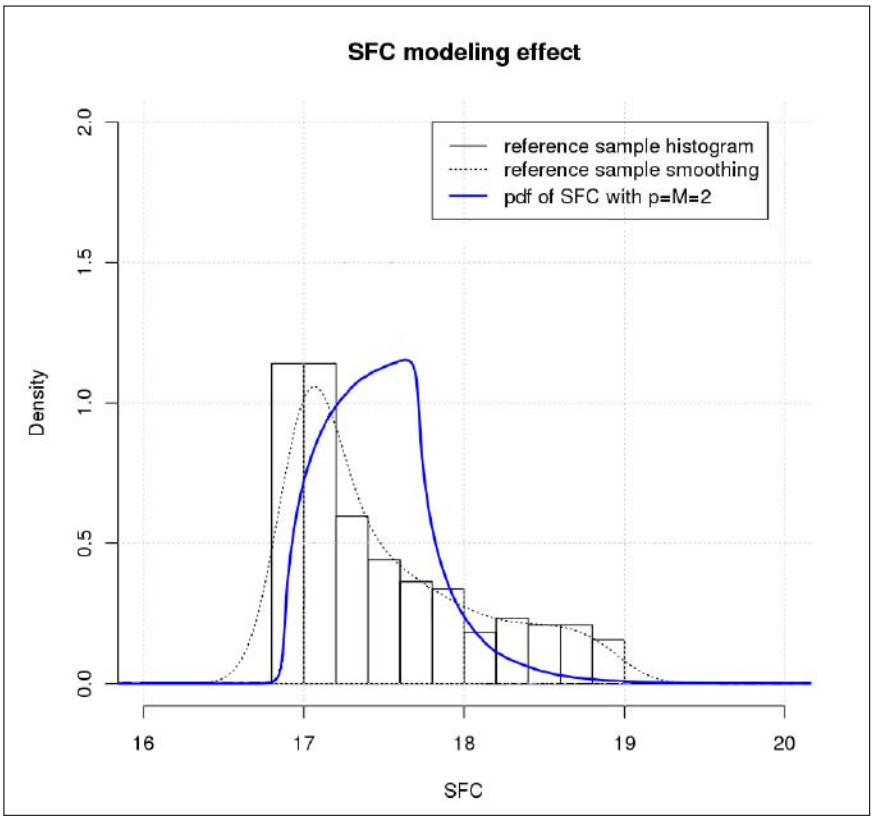


FIGURE 6.6 – Estimations of SFC probability density with a Wiener Expansion $p = M = 2$.

6.5.3 Wiener-Hermite analysis with augmented reference fuel mass sample

The Figure 6.7 shows the characterization of SFC obtained by a Wiener expansion of order $M = 2$ and degree $p = 2$ from an augmented reference fuel mass sample ($n = 82$, i.e we added 50 data obtained from the aeronautic software in the same initial conditions).
The Table 6.10 gives the coefficients corresponding to this simulation.

	θ_0	θ_1	θ_2	θ_3	θ_4	θ_5
$SFC^{2,2}$	**17.50**	**0.281**	**0.008**	**0.012**	**0.191**	**0.219**

TABLE 6.10 – SFC characterization parameters from augmented fuel mass sample

	Reference sample	from $SFC^{2,2}$	Relative error
Mean	**17.49**	**17.50**	**0.06 %**
Stand. dev.	**0.57**	**0.404**	**29.12 %**

TABLE 6.11 – Relative errors of the mean and standard deviation with $SFC^{2,2}$ from augmented fuel mass sample

Hence, by adding reference data we improve significatively the characterization of SFC on the first two statistical moments as well as on the whole probability density function of SFC.

Now, let us consider some knowledge on the SFC modeling through some expert judgment inducing a statistical modeling.

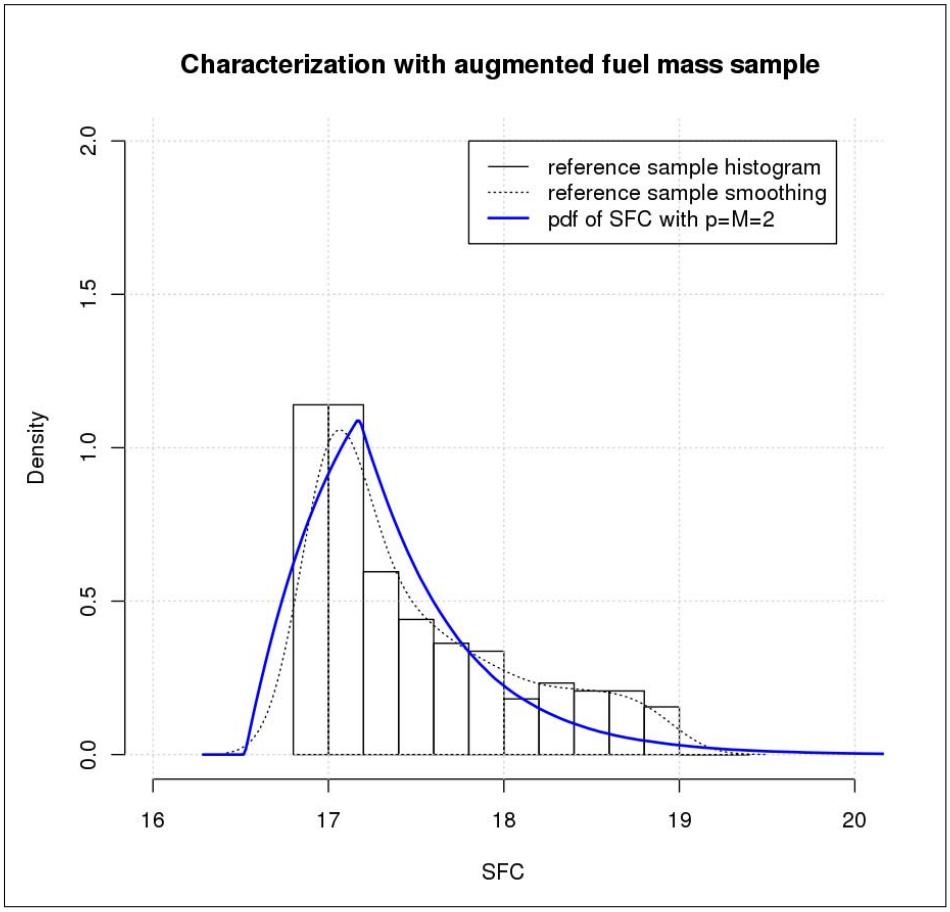


FIGURE 6.7 – Characterization of SFC with an augmented sample of fuel mass

6.5.4 Analysis with a "good" a priori knowledge

In the previous analyses, we only consider truncated Wiener-Hermite expansions which is more a mathematical hypothesis than a knowledge brought to the modeling. Suppose now that an expert judgment says that the distribution of the Specific Fuel Consumption is of exponential form. Mathematically, it turns out to suppose that the probability density of SFC belongs to the family

$$\left\{ p(u;\boldsymbol{\theta}) = \theta_2\, e^{-\theta_2(u-\theta_1)}\, \mathbb{1}_{[\theta_1,+\infty[}\,, \quad \boldsymbol{\theta} = (\theta_1,\theta_2) \in \mathbb{R}_+ \times \mathbb{R}_+^* \right\}.$$

One can check that this suggestion induces the modeling

$$SFC^{exp} = \theta_1 - \frac{1}{\theta_2}\log(\xi)\,, \quad \xi \sim \mathcal{U}([0,1])\,, \tag{6.16}$$

where $\mathcal{U}([0,1])$ is the uniform distribution on the interval $[0,1]$. The representation (6.16) is quite different from the one provided by the Wiener-Hermite expansions (see (6.9) and (6.15)).

Remark 6.5.1. Since the random variable SFC^{exp} is of finite variance, the modeling (6.16) could be seen as a particular case of Wiener expansion (6.10). An approximation would be given by choosing some order M^{exp} and degree p^{exp} in (6.11).

In what follows, we present the results of the numerical analysis corresponding to the case treated here, considering $n = 32$ and $n = 82$.

For $n = 32$.

	θ_1	θ_2
SFC^{exp}	**17.23**	**3.45**

TABLE 6.12 – Estimation of $\boldsymbol{\theta} = (\theta_1, \theta_2)$ when $n = 32$

	Reference sample	from SFC^{exp} ($n = 32$)	Relative error
Mean	**17.49**	**17.52**	**0.17 %**
Stand. dev.	**0.57**	**0.29**	**49.12 %**

TABLE 6.13 – Relative errors of the mean and standard deviation between reference SFC sample and SFC^{exp} when $n = 32$.

For $n = 82$.

	θ_1	θ_2
SFC^{exp}	**16.95**	**2**

TABLE 6.14 – Estimation of $\boldsymbol{\theta} = (\theta_1, \theta_2)$ when $n = 82$

We see clearly that the informative knowledge contribute to improve significatively the characterization of the Specific Fuel Consumption. With $n = 82$ fuel mass data, the results are rather satisfying regarding Figure 6.8 and Table 6.15.

	Reference sample	from SFC^{exp} ($n = 82$)	Relative error
Mean	**17.49**	**17.45**	**0.23 %**
Stand. dev.	**0.57**	**0.501**	**12.1 %**

TABLE 6.15 – Relative errors of the mean and standard deviation between reference SFC sample and SFC^{exp} when $n = 82$.

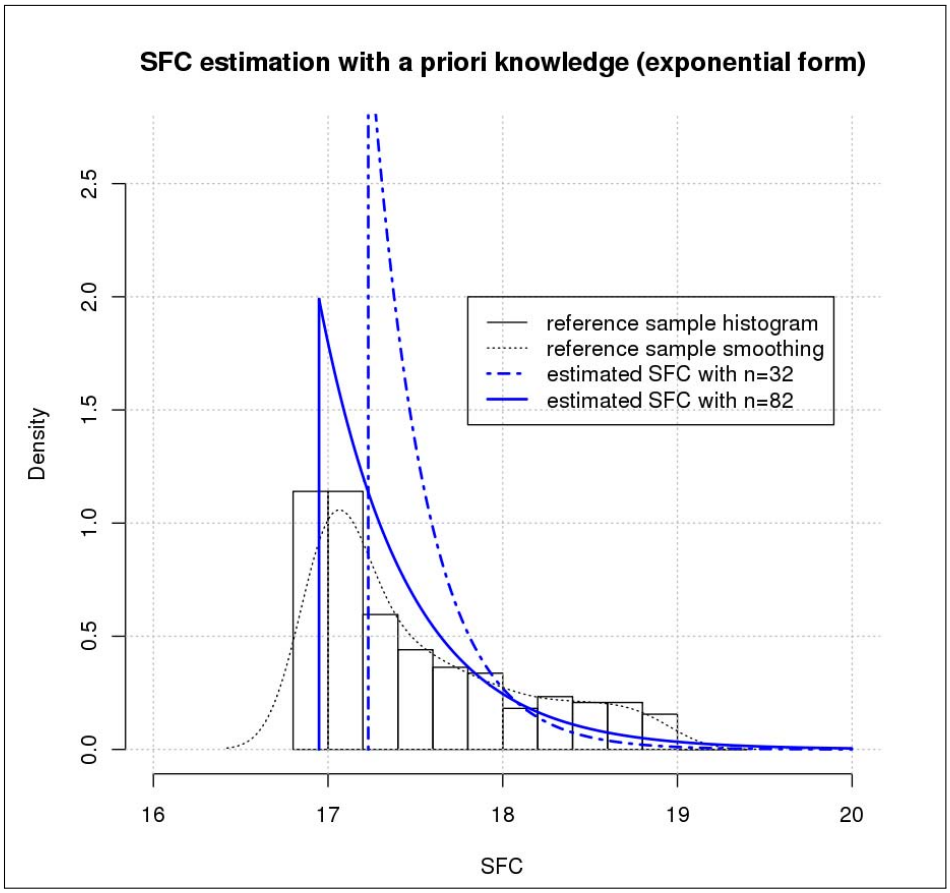


FIGURE 6.8 – Characterization of SFC with an exponential hypothesis

6.5.5 Conclusion

In this section, we tried to illustrate the effect of the modeling conditions for SFC characterization. In particular, we have shown the impact of a "model error" through the modeling of the random variable SFC and we also illustrated how the statistical error, through the number of fuel mass data, appears in the performance of the estimation.
In all cases, we computed some parameter $\widehat{\boldsymbol{\theta}}$ which we compared to the one given by a reference sample, which we can note $\boldsymbol{\theta}^*$. If we supposed that there is no model error, i.e the only error is due to limited number of data, it makes sense to investigate the difference $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|$. It is the topic of the following section.

6.6 Theoretical result

In this paper, the study of the procedure performance (6.8) will be non-asymptotic, i.e for a fixed number of observations $M_{fuel}^{*,i}$ (n) and a fixed number of variables $\boldsymbol{\epsilon}_j$ (m). The asymptotic study is let to a forthcoming work.

The quality of such estimation procedure can be investigated by giving an upper bound of the distance between the *reachable parameter* $\widehat{\boldsymbol{\theta}}$ and the *best parameter* $\boldsymbol{\theta}^*$ (unknown) which can be seen as the parameter obtained if one has infinitely many observations $M_{fuel}^{*,i}$ and variables $\boldsymbol{\epsilon}_j$. More precisely,

$$\boldsymbol{\theta}^* \quad = \quad \underset{\boldsymbol{\theta}\in\Theta}{\text{Argmin}}\, \mathbb{E}_Q \log\left(\rho_{\boldsymbol{\theta}}(M_{fuel}^*)\right)\,, \tag{6.17}$$

where roughly speaking, $\mathbb{E}_Q \log\left(\rho_{\boldsymbol{\theta}}(M_{fuel}^*)\right)$ is the "limit" of the quantity

$$\frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{1}{m}\sum_{j=1}^{m} K_{b_{\boldsymbol{\theta}}^m}\left(h(\boldsymbol{\epsilon}_j,\boldsymbol{\theta}) - M_{fuel}^{*,i}\right)\right)$$

in (6.8) when n and m go to infinity.
We consider the model $h(\boldsymbol{\epsilon},\boldsymbol{\theta})$ given in (6.3), but what follows can be generalized to any other one.
Then, denoting by $\|\cdot\|$ the Euclidian norm in $\mathbb{R}^2$, it makes sense to bound the quantity

$$\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|^2\,.$$

Let us denote by

$$\mathcal{R}(\boldsymbol{\theta}) := \mathbb{E}_Q \log\left(\rho_{\boldsymbol{\theta}}(M_{fuel}^*)\right)\,,$$

and by f the Lebesgue density associated to the measure Q.

Assumption 6.6.1. Let us consider the following assumptions.
- **A1** The map $\boldsymbol{\theta} \mapsto \mathcal{R}(\boldsymbol{\theta})$ is twice differentiable with

  $$\nabla\mathcal{R}(\boldsymbol{\theta}^*) = 0$$

  and has a symmetric positive definite Hessian matrix $\nabla^2\mathcal{R}$. Let us denote by $\lambda_{min} > 0$ the smallest eigenvalue of the set of matrices $\{\nabla^2\mathcal{R}(\boldsymbol{\theta}),\, \boldsymbol{\theta}\in\Theta\}$.
- **A2** It exists $\eta > 0$ such that for all $\boldsymbol{\theta}\in\Theta$, the density probability of $h(\boldsymbol{\epsilon},\boldsymbol{\theta})$ we noted $\rho_{\boldsymbol{\theta}}$, satisfies

  $$\rho_{\boldsymbol{\theta}} > \eta\,.$$

- **A3** For all $\boldsymbol{\theta} \in \Theta$, the second derivative of $\rho_{\boldsymbol{\theta}}$, we note $\rho''_{\boldsymbol{\theta}}$, exists and

$$C := \sup_{\boldsymbol{\theta}\in\Theta} \|\rho''_{\boldsymbol{\theta}}\|_2 < +\infty\,.$$

- **A4** We suppose that

$$0 < \delta < \inf_{\boldsymbol{\theta}\in\Theta} \hat{\sigma}_{\boldsymbol{\theta}} \quad \text{and} \quad \sup_{\boldsymbol{\theta}\in\Theta} \hat{\sigma}_{\boldsymbol{\theta}} < \sigma < +\infty\,,$$

  where $\hat{\sigma}_{\boldsymbol{\theta}}$ is defined in (6.7).

Theorem 6.6.1. *Let us consider the estimator $\widehat{\boldsymbol{\theta}}$ in (6.8) and the Assumptions (6.6.1). Then, for all $0 < \tau < 1/2$, with probability at least $1 - 2\,\tau$*

$$\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|^2 \leq c_1 \sqrt{\frac{\log(a_1\tau^{-1})}{n}} + \frac{c_2\sqrt{\log(a_2\tau^{-1})} + c_3\, m^{1/10}}{\sqrt{m}}\,,$$

for some constants c_1, c_2, c_3, a_1 and a_2.

The risk bound of this theorem seems surprising because we obtain a rate of $n^{1/4}$, whereas one expects a rate close to $\sqrt{n}$ for the treated parametric problem. This can be explained by the fact that, by the Assumption A1, we have

$$\mathcal{R}(\widehat{\boldsymbol{\theta}}) - \mathcal{R}(\boldsymbol{\theta}^*) \approx \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|^2\,.$$

Indeed, if $\mathcal{R}(\widehat{\boldsymbol{\theta}}) - \mathcal{R}(\boldsymbol{\theta}^*) \approx 1/\sqrt{n}$ [5] then obviously $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\| \approx n^{-1/4}$.
However, the $\sqrt{n}$-rate can be reached by considering the approach of Corollary 5.53 (pp. 77) in [6] where an additional assumption is made on the risk function $\boldsymbol{\theta} \mapsto \mathcal{R}(\boldsymbol{\theta})$. More precisely, this assumption relies on the function $\boldsymbol{\theta} \mapsto \Psi(\rho(\boldsymbol{\theta}))$ which is supposed to satisfy a Lipschitz condition.

6.7 Proof of Theorem 6.6.1

By Assumption A1, we have the Taylor-Lagrange formula

$$\mathcal{R}(\widehat{\boldsymbol{\theta}}) = \mathcal{R}(\boldsymbol{\theta}^*) + \frac{1}{2}\,(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^T\,\nabla^2\mathcal{R}(\xi)\,(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\,, \tag{6.18}$$

for some $\xi \in \Theta$.
Then, we will use the following lemma

Lemma 6.7.1. Rayleigh's quotient. *Let H be a real symmetric matrix $p \times p$ and denote by $\lambda_1 < \ldots < \lambda_p$ the ordered eigenvalues of H.*
It holds that for all $\mathbf{x} \in \mathbb{R}^p - \{0\}$

$$\lambda_1 \leq \frac{\mathbf{x}^T H\,\mathbf{x}}{\mathbf{x}^T\,\mathbf{x}} \leq \lambda_p\,.$$

Now, applying this lemma with $H = \nabla^2\mathcal{R}(\xi)$ and $\mathbf{x} = (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$ yields

$$\lambda_{min}\,\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|^2 \leq (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^T\,\nabla^2\mathcal{R}(\xi)\,(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\,,$$

5. Voir Théorème 3.4.1 du Chapitre 3

where $\lambda_{min} > 0$ is the smallest eigenvalue of the set of matrices $\{\nabla^2\mathcal{R}(\boldsymbol{\theta}),\, \boldsymbol{\theta} \in \Theta\}$.
Then, using this last inequality with the equality (6.18) gives the following inequality

$$\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|^2 \leq \frac{2}{\lambda_{min}} \left(\mathcal{R}(\widehat{\boldsymbol{\theta}}) - \mathcal{R}(\boldsymbol{\theta}^*)\right). \tag{6.19}$$

The problem turns to bound the positive quantity $\mathcal{R}(\widehat{\boldsymbol{\theta}}) - \mathcal{R}(\boldsymbol{\theta}^*)$, where $\widehat{\boldsymbol{\theta}}$ is given by (6.8). Such bound can be investigated by applying Theorem 4.1 in [4] [6], which is a general result. We will search to compute constants K_1^τ and K_2^τ such that, with high probability

$$\mathcal{R}(\widehat{\boldsymbol{\theta}}) - \mathcal{R}(\boldsymbol{\theta}^*) \leq \frac{2\,\|f\|_2}{\eta} \left(\frac{1}{\sqrt{n}} \frac{\eta}{2\,\delta^2\,\|f\|_2}\,\gamma\,K_1^\tau + \frac{1}{\sqrt{m}}\,\frac{1}{\sqrt{2\,\pi}\,\delta}\,K_2^\tau + \frac{1}{m^{2/5}}\,\frac{C\,(1.06\,\sigma)^2}{\sqrt{3}}\right). \tag{6.20}$$

For our purpose, the main work is to compute the concentration constants K_1^τ and K_2^τ derived from [4] in the following particular case.

6.7.1 On concentration constants K_1^τ and K_2^τ

First of all, let us recall some definitions and notations relative to empirical processes.

Definition 6.7.1. Empirical process. Let W be some probability measure on some space T and let us suppose given a k i.i.d sample $\xi_1, ..., \xi_k$ drawn from W. Let us denote by W_k the empirical measure

$$W_k := \frac{1}{k}\sum_{i=1}^{k} \delta_{\xi_i}$$

and $\mathcal{G}$ some class of real valued functions $g : T \to \mathbb{R}$.
We call W-empirical process indexed by $\mathcal{G}$ the following application

$$\begin{aligned} \mathbb{G}_k : \mathcal{G} &\longrightarrow \mathbb{R} \\ g &\longmapsto \mathbb{G}_k := \sqrt{k}\int_T g(t)\ (W_k - W)\,(dt), \end{aligned}$$

also written

$$\mathbb{G}_k\, g := \frac{1}{\sqrt{k}}\sum_{i=1}^{k} \left(g(\xi_k) - \mathbb{E}_W(g(\xi))\right).$$

We denote the supremum of an empirical process by

$$\|\mathbb{G}_k\|_{\mathcal{G}} := \sup_{g\in\mathcal{G}} |\mathbb{G}_k\, g|.$$

Following the proof lines of Theorem 2.1, the Table 2 p.11 in [4] (giving classes of functions) and considering the inequality (6.20), one can check that K_1^τ is defined as

$$\text{for all } n \geq 1, \quad \mathbb{P}(\|\mathbb{U}_n\|_{\mathcal{A}} \leq K_1^\tau) \geq 1 - \tau \tag{6.21}$$

where $\mathbb{U}_n$ is the Q-empirical process ($Q_n = \frac{1}{n}\sum_{i=1}^n \delta_{M_{fuel}^{*,i}}$) indexed by the class of functions

$$\mathcal{A} = \{y \in \mathcal{I} \longmapsto (y - \lambda)^2,\, \lambda \in \mathcal{I}_h\} \tag{6.22}$$

6. Voir aussi le Théorème 3.4.1 dans le Chapitre 3

where we recall

$$\mathcal{I} = [M_{inf}, M_{sup}] \quad \text{and} \quad \mathcal{I}_h = h(\mathcal{E}, \Theta)\,.$$

Similarly, the constant K_2^τ is defined as follows

$$\text{for all } m \geq 1\,, \quad \mathbb{P}(\|\mathbb{V}_m\|_{\mathcal{B}} \leq K_2^\tau) \geq 1 - \tau \tag{6.23}$$

where $\mathbb{V}_m$ is the P^ϵ-empirical process ($P_m^\epsilon = \frac{1}{m}\sum_{j=1}^m \delta_{\epsilon_j}$) indexed by the class of functions

$$\mathcal{B} = \{\mathbf{x} \in \mathcal{E} \longmapsto e^{-(h(\mathbf{x},\boldsymbol{\theta})-\lambda)^2/2b^2}, \quad (\boldsymbol{\theta}, \lambda, b) \in \Theta \times \mathcal{I}_h \times [\delta, \sigma]\}\,. \tag{6.24}$$

By the writings (6.21) and (6.23), the constants K_1^τ and K_2^τ arise from the "*concentration of the measure phenomenon*" (see [3], [1]). More precisely, these constants characterize the *tightness* of the sequences of random variables $\|\mathbb{U}_n\|_{\mathcal{A}}$ (which is $(M_{fuel}^{*,i})_{i=1,\dots,n}$ dependent) and $\|\mathbb{V}_m\|_{\mathcal{B}}$ (which is $(\boldsymbol{\epsilon}_j)_{j=1,\dots,m}$ dependent).

Now, we aim at computing (upper bound) these constants using concentration inequalities where the classes of functions $\mathcal{A}$ and $\mathcal{B}$ will play a crucial role. In particular, we will apply the following theorem which is Theorem 2.14.9 in [7].
Before, let us recall the definition of the *bracketing numbers* (taken from [7] p. 83-85).

Definition 6.7.2. Bracketing numbers. Let $\mathcal{G}$ be some class of functions on T and denote by W a probability measure on T.
Given two functions l, u, the bracket $[l, u]$ is the set of all functions g with $l \leq g \leq u$. An ε-bracket is a bracket $[l, u]$ with $||u - l||_{2,W} < \varepsilon$. The *bracketing number* $N_{[]}(\varepsilon, \mathcal{G}, L_2(W))$ is the minimum number of ε-brackets needed to cover the class of functions $\mathcal{G}$.
The *entropy with bracketing* is the logarithm of the bracketing number.

Remark 6.7.1. The bracketing numbers measure the "size", the complexity of a class of functions.

Theorem 6.7.1. *Let $\mathcal{G}$ be a uniformly bounded class of (measurable) functions $g : T \to [0, 1]$ and denote by W a probability measure on T. If the class $\mathcal{G}$ satisfies, for some constants K and L*

$$N_{[]}(\varepsilon, \mathcal{G}, L_2(W)) \leq \left(\frac{K}{\varepsilon}\right)^L \quad \textit{for every } 0 < \varepsilon < K\,. \tag{6.25}$$

Then, for every $t > 0$,

$$\mathbb{P}(\|\mathbb{G}_k\|_{\mathcal{G}} > t) \leq \left(\frac{D\,t}{\sqrt{L}}\right)^L e^{-2t^2}\,,$$

for a constant D that only depends on K.

The proof of this theorem can be found in [5].

Now, let K^τ be a constant (to determine) which satisfies

$$\mathbb{P}(\|\mathbb{G}_k\|_{\mathcal{G}} \leq K^\tau) \geq 1 - \tau\,.$$

This is equivalent to

$$\mathbb{P}(\|\mathbb{G}_k\|_{\mathcal{G}} > K^\tau) \leq \tau\,. \tag{6.26}$$

By Theorem 6.7.1, applied with $t = K^\tau$, we have

$$\mathbb{P}(\|\mathbb{G}_k\|_{\mathcal{G}} > K^\tau) \leq \left(\frac{D\,K^\tau}{\sqrt{L}}\right)^L e^{-2(K^\tau)^2}\,. \tag{6.27}$$

Hence, the constant K^τ can be taken such that

$$\left(\frac{D\,K^\tau}{\sqrt{L}}\right)^L e^{-2(K^\tau)^2} \leq \tau\,,$$

which is similar to

$$(K^\tau)^2 - \frac{L}{2}\log(K^\tau) \geq \frac{\log(a_{L,D}\,\tau^{-1})}{2}\,, \quad \text{with } a_{L,D} = \left(\frac{D}{\sqrt{L}}\right)^L. \tag{6.28}$$

Then, for small enough $\tau > 0$, let us consider the constant

$$K^\tau = \sqrt{\frac{\log(a_{L,D}\,\tau^{-1})}{2}} \tag{6.29}$$

which satisfies (6.28).

Finally, we see that the constant K^τ can be characterized (only) by the class of functions $\mathcal{G}$ through the constants D and L provided by Theorem 6.7.1.

In our purpose, the classes of interest are $\mathcal{A}$ and $\mathcal{B}$ defined in (6.22) and (6.24), respectively. Next, one can easily check that these classes are uniformly bounded and it is suitable to work with normalized classes

$$\bar{\mathcal{A}} = \alpha_{\mathcal{A}} + \frac{1}{\beta_{\mathcal{A}}}\mathcal{A}\,, \tag{6.30}$$

$$\bar{\mathcal{B}} = \alpha_{\mathcal{B}} + \frac{1}{\beta_{\mathcal{B}}}\mathcal{B}\,, \tag{6.31}$$

such that all the functions take values in $[0,1]$.

Now, we have to prove that the classes $\bar{\mathcal{A}}$ and $\bar{\mathcal{B}}$ have polynomial bracketing numbers following (6.25). This will give the constants $L_{\bar{\mathcal{A}}}$, $D_{\bar{\mathcal{A}}}$ and $L_{\bar{\mathcal{B}}}$, $D_{\bar{\mathcal{B}}}$ needed to identify the key constants K_1^τ and K_2^τ defined in (6.21) and (6.23), respectively.

6.7.2 Characterization of $L_{\bar{\mathcal{A}}}$, $D_{\bar{\mathcal{A}}}$, $L_{\bar{\mathcal{B}}}$, $D_{\bar{\mathcal{B}}}$

We consider the Theorem 2.7.11 in [7] (p. 164) which deals with classes that are Lipschitz in a parameter. It reads :

Theorem 6.7.2. *Let $\mathcal{G} = \{t \in T \mapsto g_s(t)\,,\, s \in S\}$ be a class of functions satisfying*

$$\textit{for all } t \in T\,, s, s' \in S\,, \quad |g_s(t) - g_{s'}(t)| \leq d(s,s')\,G(t)\,,$$

for some metric d on S and some function $G : t \mapsto G(t)$.
Then, for any norm

$$N_{[]}(2\varepsilon\,\|G\|, \mathcal{G}, \|\cdot\|) \leq N(\varepsilon, S, d)\,,$$

where $N(\varepsilon, S, d)$ is the minimal number of balls $\{r\,,\, d(r,s) < \varepsilon\}$ of radius ε needed to cover the set S.

In what follows, we detail the case of the class $\bar{\mathcal{A}}$. The case of the class $\bar{\mathcal{B}}$ is exactly in the same spirit.

Let us recall that Q is the probability measure considered on $\mathcal{I}$ (observation space) and that we have

$$\bar{\mathcal{A}} = \{f_\lambda \,:\, y \in \mathcal{I} \longmapsto \alpha_{\mathcal{B}} + \frac{1}{\beta_{\mathcal{A}}}(y-\lambda)^2,\, \lambda \in \mathcal{I}_h\}\,,$$

where $\mathcal{I} = [M_{inf}, M_{sup}]$ and $\mathcal{I}_h = [M^h_{inf}, M^h_{sup}]$ (with $\mathcal{I} \subset \mathcal{I}_h$).
So

$$|f_\lambda(y) - f_\lambda(y)| = \frac{1}{\beta_{\mathcal{A}}}|(y-\lambda_1)^2 - (y-\lambda_2)^2| \leq |\lambda_1 - \lambda_2|\, F(y)\,,$$

with $F(y) = \frac{2}{\beta_{\mathcal{A}}}(y + M^h_{sup})$, and by Theorem 6.7.2 applied with $\|\cdot\| = \|\cdot\|_{2,Q}$, it holds that

$$N_{[]}(\epsilon,\, \bar{\mathcal{A}},\, L_2(Q)) \leq N\left(\frac{\epsilon}{2\,\|F\|_{2,Q}}, \mathcal{I}_h, |\cdot|\right)\,.$$

Moreover, since

$$\|F\|_{2,Q} \leq \sup_{y\in\mathcal{I}} F(y)\, \|f\|_2$$

where f is the density associated to the measure Q, and using the fact that $\mathcal{I} \subset \mathcal{I}_h$, we obtain that

$$\|F\|_{2,Q} \leq \frac{4}{\beta_{\mathcal{A}}} M^h_{sup}\, \|f\|_2\,.$$

This last inequality yields

$$N\left(\frac{\varepsilon}{2\,\|F\|_{2,Q}}, \mathcal{I}_h, |\cdot|\right) \leq N\left(\frac{\beta_{\mathcal{A}}\,\varepsilon}{8\, M^h_{sup}\, \|f\|_2}, \mathcal{I}_h, |\cdot|\right)\,.$$

Since $\mathcal{I}_h = [M^h_{inf}, M^h_{sup}]$, the quantity (covering number) in the right member is bounded by

$$\frac{8\,|\mathcal{I}_h|\, M^h_{sup}\, \|f\|_2}{\beta_{\mathcal{A}}\,\varepsilon}\,, \quad |\mathcal{I}_h| = M^h_{sup} - M^h_{inf}.$$

We finally get

$$N_{[]}(\varepsilon,\, \bar{\mathcal{A}},\, L_2(Q)) \leq \frac{8\,|\mathcal{I}_h|\, M^h_{sup}\, \|f\|_2}{\beta_{\mathcal{A}}\,\varepsilon}\,,$$

that is

$$N_{[]}(\varepsilon,\, \bar{\mathcal{A}},\, L_2(Q)) \leq \left(\frac{K_{\bar{\mathcal{A}}}}{\varepsilon}\right)^{L_{\bar{\mathcal{A}}}}\,,$$

with

$$L_{\bar{\mathcal{A}}} = 1$$

and

$$K_{\bar{\mathcal{A}}} = \frac{8\,|\mathcal{I}_h|\, M^h_{sup}\, \|f\|_2}{\beta_{\mathcal{A}}} \quad \text{that determines } D_{\bar{\mathcal{A}}} \text{ by [5]}\,.$$

A similar work gives the constant $L_{\bar{\mathcal{B}}} = 1$ and a constant $D_{\bar{\mathcal{B}}}$.

6.7.3 End of the proof

By the previous subsection, we get the constants $K^\tau_{\bar{\mathcal{A}}}$ and $K^\tau_{\bar{\mathcal{B}}}$ given by (6.29) with associated constants L and D :

$$K^\tau_{\bar{\mathcal{A}}} = \sqrt{\frac{\log(a_1\,\tau^{-1})}{2}}, \quad a_1 = a_{L_{\bar{\mathcal{A}}},D_{\bar{\mathcal{A}}}} = D_{\bar{\mathcal{A}}} \tag{6.32}$$

$$K^\tau_{\bar{\mathcal{B}}} = \sqrt{\frac{\log(a_2\,\tau^{-1})}{2}}, \quad a_2 = a_{L_{\bar{\mathcal{B}}},D_{\bar{\mathcal{B}}}} = D_{\bar{\mathcal{B}}} \tag{6.33}$$

where initially $a_{L,D} = \left(\frac{D}{\sqrt{L}}\right)^L$ (by (6.28)).
But, the constants of interest K^τ_1 and K^τ_2 defined in (6.21) and (6.23) are relative to non normalized classes $\mathcal{A}$ and $\mathcal{B}$. Let us remark that if $\bar{\mathcal{G}} = \alpha + \frac{1}{\beta}\mathcal{G}$, then

$$\|\mathbb{G}_k\|_{\bar{\mathcal{G}}} = \frac{1}{\beta}\,\|\mathbb{G}_k\|_{\mathcal{G}}\,. \tag{6.34}$$

Now, let us denote by $K^\tau_{\bar{\mathcal{G}}}$ the constant that satisfies

$$\mathbb{P}(\|\mathbb{G}_k\|_{\bar{\mathcal{G}}} \le K^\tau_{\bar{\mathcal{G}}}) \ge 1-\tau\,,$$

and denote by $K^\tau_{\mathcal{G}}$ the constant that satisfies

$$\mathbb{P}(\|\mathbb{G}_k\|_{\mathcal{G}} \le K^\tau_{\mathcal{G}}) \ge 1-\tau\,.$$

By (6.34), it is easy to check that we can take

$$K^\tau_{\mathcal{G}} = \beta\,K^\tau_{\bar{\mathcal{G}}}\,.$$

We deduce that

$$K^\tau_1 = \beta_{\mathcal{A}}\,K^\tau_{\bar{\mathcal{A}}}$$

and

$$K^\tau_2 = \beta_{\mathcal{B}}\,K^\tau_{\bar{\mathcal{B}}}\,.$$

Finally, by (6.19) and (6.20) we have with probability $1-2\tau$

$$\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|^2 \le \frac{4\,\|f\|_2}{\lambda_{min}\,\eta}\left(\frac{1}{\sqrt{n}}\frac{\eta}{2\,\delta^2\,\|f\|_2}\,\gamma\,K^\tau_1 + \frac{1}{\sqrt{m}}\,\frac{1}{\sqrt{2\,\pi}\,\delta}\,K^\tau_2 + \frac{1}{m^{2/5}}\,\frac{C\,(1.06\,\sigma)^2}{\sqrt{3}}\right)$$

which we rewrite

$$\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|^2 \le \frac{\sqrt{2}\,c_1}{\sqrt{n}}K^\tau_{\bar{\mathcal{A}}} + \frac{\sqrt{2}\,c_2}{\sqrt{m}}K^\tau_{\bar{\mathcal{B}}} + \frac{c_3}{m^{1/5}}$$

with corresponding constants c_1, c_2 and c_3 and $K^\tau_{\bar{\mathcal{A}}}, K^\tau_{\bar{\mathcal{B}}}$ are given by (6.32) and (6.33).

That concludes the proof of Theorem 6.6.1.

Bibliographie

[1] P. Billingsley. *Convergence of probability measures*. Wiley New York, 1968.

[2] E. de Rocquigny, N. Devictor, and S. Tarantola, editors. *Uncertainty in industrial practice*. John Wiley.

[3] M. Ledoux. *The concentration of measure phenomenon*. AMS, 2001.

[4] N. Rachdi, J.C. Fort, and T. Klein. Risk bounds for new M-estimation problems . *hal 00537236*, 2010.

[5] M. Talagrand. Sharper bounds for Gaussian and empirical processes. *The Annals of Probability*, 22(1) :28–76, 1994.

[6] A.W. van der Vaart. *Asymptotic statistics*. Cambridge University Press, 2000.

[7] A.W. van der Vaart and J.A. Wellner. *Weak Convergence and Empirical Processes*. Springer Series in Statistics, 1996.

[8] N. Wiener. The homogeneous chaos. *American Journal of Mathematics*, 60(4) :897–936, 1938.

Chapitre 7

Combinaison de données expérimentales et données simulées par une approche apprentissage statistique

Sommaire

Résumé du Chapitre

Dans ce chapitre, on propose quelques méthodes concernant l'exploitation de données expérimentales et simulées, pouvant être de nature différente, afin d'"aider" à la simulation de données dans une certaine configuration d'intérêt.

Combining Experimental and Simulated data with a statistical learning approach

Nabil Rachdi [1]

Abstract
The study of complex phenomena where both experimental and simulated data are available is very common in engineering, where complex models exist and where experimental data of interest are difficult or even impossible to obtain. However, beyond these costly models, one or more modeling may be workable in that we may have at disposal both experimental and simulated data. Then, the challenge is to "combine" these data in order to analyse the phenomenon of interest. This paper aims at giving first motivations of the statistical learning approach to tackle that problem.

Keywords : statistical learning, experiments, simulations, calibration, uncertainty quantification

(*We owe thanks to Vincent Feuillard for fruitful discussions, Engineer R&D at EADS Innovation Works)

7.1 Introduction

This work has two main motivations, especially in aeronautics.
First, the majority of the aircrafts are manufactured with *aluminium alloys* for certain advantages like the specific rigidity of the material and its behavior with variations of temperature. However, an important challenge in aeronautic is the mass of the product and so, since the last decade, engineers focus on other technologies providing a satisfying rigidity ratio. The *composite* materials seem to have promising properties and are actually in the interest of industrial and academic researchers. In this context, one may has at disposal experimental and simulated data for aluminium alloys materials, and also simulated data for composite ones. But, in order to study composite materials behaviors, one hopes to "construct" data close to experimental data corresponding to composite ones which are not available in practice (since non-manufactured or for costs reasons).
Another motivation of our work is the so called *Model-Assisted approach to Probability Of Detection*, with abbreviation *MAPOD*, in the field of Non Destructive Technics. MAPOD is in fact a working group established in 2004 by the U.S Air Force whose goal is to provide new methodologies in order to analyze the probability of detection (POD), see the reference paper [1], integrating simulated data generated from some (assistant) model. In few words, the computation of a POD of some defect in a material at some configuration $\mathcal{C}$ can be quite expensive, because it requires the fabrication of costly experiments which replicates the material at the configuration of interest. Hence, there is no available experimental data at the configuration $\mathcal{C}$. However it is possible to have simulated data (by computer simulations for instance) at that configuration $\mathcal{C}$. Moreover, it happens that one can also have at disposal

1. Institut de Mathématiques de Toulouse - EADS Innovation Works, 92152 Suresnes

experimental and simulated data at some configuration $\mathcal{C}_0$ (called reference). For instance, in [3] the configuration $\mathcal{C}$ would correspond to the data on Aluminium where only simulated data are available, and the reference configuration $\mathcal{C}_0$ would correspond to the data on Titanium, where both experimental and simulated data are computable. (Let us remark that in the Non Destructive Control context, the Aluminium data are difficult to obtain, contrary to the context which compares Aluminium constructions and Composite ones). So, one of the goals of non destructive technics is to predict Aluminium experimental data in order to compute the POD.
Therefore, we need a method taking into account these three sets of data in order to infer the data set of interest. The approach adopted in [2] consists in using a *Transfer Function* to provide the sought experimental data. The goal of this paper is to give first elements to generalize this approach which is a particular case of what we call *global transformations*. We propose an alternative approach based on statistical learning leading to promising results.

7.2 General settings

Denote by

$$\mathbf{Y}^{exp(0)} := (Y_1^{exp(0)}, ..., Y_{n_{exp}}^{exp(0)})^T$$

and by

$$\mathbf{Y}^{sim(0)} := (Y_i^{sim((0)}, ..., Y_{n_{sim}}^{sim((0)})^T$$

the sets of experimental and simulated data at some *reference* configuration $\mathcal{C}_0$. Then, let us suppose that we want to generate "experimental" data $\mathbf{Y}^{exp(c)} := \{Y_i^{exp(c)}, i = 1, ..., n_{exp}\}$ provided from some configuration $\mathcal{C}$, for given simulated data at this configuration

$$\mathbf{Y}^{sim(c)} := (Y_i^{sim(c)}, ..., Y_{n_{sim}}^{sim(c)})^T .$$

For simplicity, we deal with real valued data. Suppose that for $i = 1, ..., n_{exp} = n_{sim} = n$,
– $Y_i^{exp(0)} = h^{exp(0)}(x_i) + \varepsilon_i^{exp(0)}$,
– $Y_i^{sim(0)} = h^{sim(0)}(x_i) + \varepsilon_i^{sim(0)}$,
– $Y_i^{sim(c)} = h^{sim(c)}(x_i) + \varepsilon_i^{sim(c)}$
where $h^{exp(0)}$, $h^{sim(0)}$ and $h^{sim(c)}$ are three workable computer experiments depending on some covariate x, and $\varepsilon_i^{exp(0)}$, $\varepsilon_i^{sim(0)}$, $\varepsilon_i^{sim(c)}$ are some noise variables. In this study, the noise won't be considered. The goal is to predict the "experimental" data $Y_i^{exp(c)}$ at the configuration $\mathcal{C}$, which can be viewed as $Y_i^{exp(c)} = h^{exp(c)}(x_i) + \varepsilon_i^{exp(c)}$ but with completely unknown function $h^{exp(c)}$. This is principally the reason why classical regression method based on the sample $(x_1, Y_1^{exp(c)}), ..., (x_n, Y_n^{exp(c)})$ is unfeasible (since $Y_i^{exp(c)}$ is unknown). In a first approach, we present *global* methods which consists in seeking a *transformation* $\widehat{\mathcal{T}} : \mathbb{R}^n \to \mathbb{R}^n$ satisfying

$$\widehat{\mathcal{T}} = \underset{\mathcal{T} \in \mathbb{T}_{\mathbb{R}^n}}{\text{Argmin}} \left\| \mathbf{Y}^{exp(0)} - \mathcal{T}(\mathbf{Y}^{sim(0)}) \right\|^2 . \tag{7.1}$$

Then, the idea is to "export" this transformation (induced by the configuration $\mathcal{C}_0$) to the configuration $\mathcal{C}$ in order to compute $\widehat{\mathcal{T}}(\mathbf{Y}^{sim(c)})$ supposed to estimate the wanted experimental data $\mathbf{Y}^{exp(c)}$. There may be several ways to investigate a transformation. We present in Section 7.4 an extension of global transformations.

Secondly, we present a statistical approach in Section 7.6 which differs from global transformations in that the statistical methods we present have to determine the general structure that links experimental and simulated data (at configuration $\mathcal{C}_0$). It amounts to seek a *model* $\mathcal{T} : \mathbb{R} \to \mathbb{R}$ satisfying

$$\widehat{\mathcal{T}} = \underset{\mathcal{T}\in\mathbb{T}_{\mathbb{R}}}{\text{Argmin}} \left\| \mathbf{Y}^{exp(0)} - \begin{pmatrix} \mathcal{T}(Y_1^{sim(0)}) \\ \cdot \\ \cdot \\ \mathcal{T}(Y_n^{sim(0)}) \end{pmatrix} \right\|^2 . \tag{7.2}$$

The displays (7.1) and (7.2) allow well to understand the main difference between the two approaches. In the first, the transformation acts "point by point" whereas in the second transformation, called model, we attempt to *learn* the general mechanism that links simulated and experimental data. We expect that the statistical approach is more robust than the global one. In this paper, we don't investigate theoretical considerations. Both approaches will be illustrated by numerical examples in Section 7.5 and 7.7, respectively, with comparisons in the latter section.

7.3 Approach by global transformations

Here, the purpose is to seek some application $\widehat{\mathcal{T}}$ satisfying

$$\widehat{\mathcal{T}} = \underset{\mathcal{T}\in\mathbb{T}}{\text{Argmin}} \left\| \mathbf{Y}^{exp(0)} - \mathcal{T}(\mathbf{Y}^{sim(0)}) \right\|^2 , \tag{7.3}$$

where $\mathbb{T}$ is some set of applications $\mathcal{T} : \mathbb{R}^n \to \mathbb{R}^n$, we call *(tranformation) model*.
Notice that here it makes sense to compare $Y_i^{exp(0)}$ with its simulated version $Y_i^{sim(0)}$, for all $i = 1, ..., n$, because we supposed the presence of some covariate x such that $Y_i^{exp(0)} = h^{exp(0)}(x_i)$ and $Y_i^{sim(0)} = h^{sim(0)}(x_i)$.
For example, consider the following model

$$\mathbb{T}_1 = \{\mathbf{y} = (y_1, ..., y_n)^T \mapsto \mathcal{T}_{\boldsymbol{\theta}}(\mathbf{y}) = \mathbf{y} + \boldsymbol{\theta}\,,\ \boldsymbol{\theta} \in \mathbb{R}^n\} . \tag{7.4}$$

By the display (7.3), it remains to seek the parameter

$$\begin{aligned} \widehat{\boldsymbol{\theta}} &= \underset{\boldsymbol{\theta}\in\mathbb{R}^n}{\text{Argmin}} \left\| \mathbf{Y}^{exp(0)} - \mathcal{T}_{\boldsymbol{\theta}}(\mathbf{Y}^{sim(0)}) \right\|^2 \\ &= \underset{\boldsymbol{\theta}\in\mathbb{R}^n}{\text{Argmin}} \sum_{i=1}^{n} (Y_i^{exp(0)} - \theta_i - Y_i^{sim(0)})^2 \\ &= \left(Y_1^{exp(0)} - Y_1^{sim(0)}, ..., Y_n^{exp(0)} - Y_n^{sim(0)} \right)^T . \end{aligned}$$

Hence, we get the transformation

$$\widehat{\mathcal{T}}_1 = \mathcal{T}_{\widehat{\boldsymbol{\theta}}} : \mathbf{y} \longmapsto \mathbf{y} + \widehat{\boldsymbol{\theta}} . \tag{7.5}$$

Let us give another example of global transformation. Take the following model

$$\mathbb{T}_2 = \{\mathbf{y} = (y_1, ..., y_n)^T \mapsto \mathcal{T}_{\boldsymbol{\theta}}(\mathbf{y}) = D(\boldsymbol{\theta})\,\mathbf{y}\,,\ \boldsymbol{\theta} \in \mathbb{R}^n\} , \tag{7.6}$$

where $D(\boldsymbol{\theta})$ is a diagonal matrix with diagonal elements $(\theta_1, ..., \theta_n)$. By the display (7.3), it remains to seek the parameter

$$\begin{aligned} \widehat{\boldsymbol{\theta}} &= \underset{\boldsymbol{\theta}\in\mathbb{R}^n}{\text{Argmin}} \left\| \mathbf{Y}^{exp(0)} - \mathcal{T}_{\boldsymbol{\theta}}(\mathbf{Y}^{sim(0)}) \right\|^2 \\ &= \underset{\boldsymbol{\theta}\in\mathbb{R}^n}{\text{Argmin}} \sum_{i=1}^{n} (Y_i^{exp(0)} - \theta_i Y_i^{sim(0)})^2 \\ &= \left(\frac{Y_1^{exp(0)}}{Y_1^{sim(0)}}, ..., \frac{Y_n^{exp(0)}}{Y_n^{sim(0)}} \right)^T . \end{aligned}$$

Then, we get the following transformation

$$\widehat{\mathcal{T}}_2 \,:\, \mathbf{y} \longmapsto D(\widehat{\boldsymbol{\theta}})\,\mathbf{y}\,. \tag{7.7}$$

Let us remark that for both transformations (7.5) and (7.7), $\widehat{\mathcal{T}}_1(\mathbf{Y}^{sim(0)}) = \widehat{\mathcal{T}}_2(\mathbf{Y}^{sim(0)}) = \mathbf{Y}^{exp(0)}$, whereas the transformations $\widehat{\mathcal{T}}_1$ and $\widehat{\mathcal{T}}_2$ are different. Moreover, in these specific cases $\widehat{\mathcal{T}}_1$ and $\widehat{\mathcal{T}}_2$ are **pointwise linear transformations**. For example, consider $n = 1$ and the data $(y^{sim(0)}, y^{exp(0)})$. We have for all $y \in \mathbb{R}$

$$\widehat{\mathcal{T}}_1(y) = y + y^{exp(0)} - y^{sim(0)} \quad \text{and} \quad \widehat{\mathcal{T}}_2(y) = \frac{y^{exp(0)}}{y^{sim(0)}}\, y\,.$$

We give an illustration in Figure 7.1.

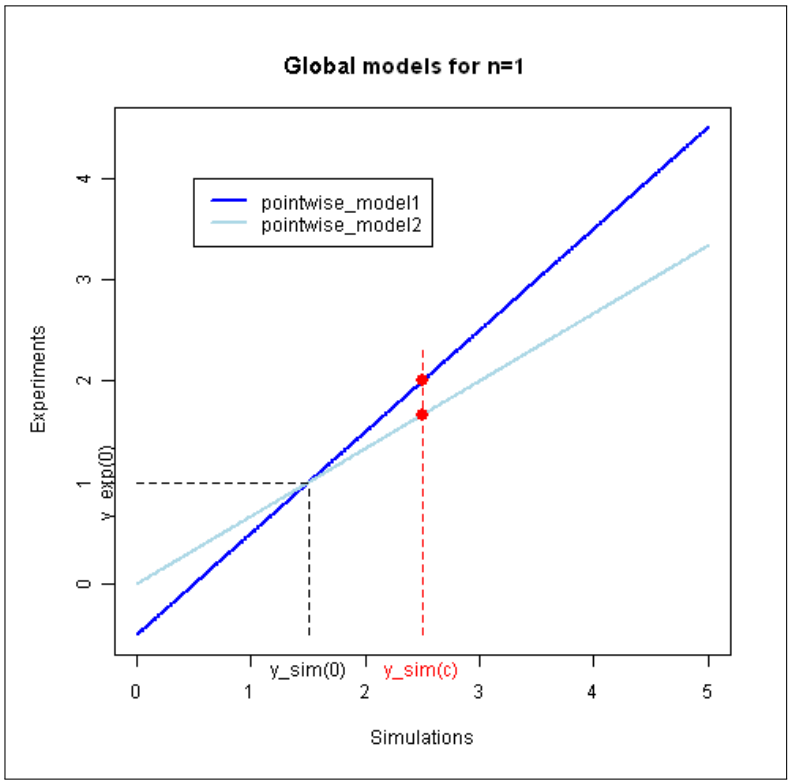


FIGURE 7.1 – Illustration of global models $\widehat{\mathcal{T}}_1$ and $\widehat{\mathcal{T}}_2$ for $n = 1$

Finally, given a simulation data set $\mathbf{Y}^{sim(c)}$ at some configuration $\mathcal{C}$, we can predict the "experimental" data (see Figure 7.1). Afterthat, it makes sense to compute wanted "experimental"

data at the configuration $\mathcal{C}$

$$\widehat{\mathbf{Y}}^{exp(c)} = \widehat{\mathcal{T}}(\mathbf{Y}^{sim(c)})\,.$$

Interpretation of transformations $\widehat{\mathcal{T}}_1$ and $\widehat{\mathcal{T}}_2$.
Interpretation of $\widehat{\mathcal{T}}_1$. For all $i = 1, ..., n$ we have

$$Y_i^{exp(c)} = Y_i^{sim(c)} + Y_i^{exp(c)} - Y_i^{sim(c)}\,.$$

The transformation $\widehat{\mathcal{T}}_1$ provides for all $i = 1, ..., n$

$$\widehat{Y}_i^{exp(c)} = Y_i^{sim(c)} + Y_i^{exp(0)} - Y_i^{sim(0)}\,,$$

hence, the application $\widehat{\mathcal{T}}_1$ would implicitly suppose the **differences conservation**

$$Y_i^{exp(0)} - Y_i^{sim(0)} \approx Y_i^{exp(c)} - Y_i^{sim(c)}\,, \quad i = 1, ..., n\,. \tag{7.8}$$

Interpretation of $\widehat{\mathcal{T}}_2$. For all $i = 1, ..., n$ we also can write

$$Y_i^{exp(c)} = \frac{Y_i^{exp(c)}}{Y_i^{sim(c)}} Y_i^{sim(c)}\,.$$

The transformation $\widehat{\mathcal{T}}_2$ provides that for all $i = 1, ...n$

$$\widehat{Y}_i^{exp(c)} = \frac{Y_i^{exp(0)}}{Y_i^{sim(0)}} Y_i^{sim(c)}\,,$$

hence, the application $\widehat{\mathcal{T}}_2$ would implicitly suppose the **ratios conservation**

$$\frac{Y_i^{exp(0)}}{Y_i^{sim(0)}} \approx \frac{Y_i^{exp(c)}}{Y_i^{sim(c)}}\,, \quad i = 1, ..., n\,. \tag{7.9}$$

This latter interpretation means that simulated data at the configuration $\mathcal{C}$ are obtained by *cross-multiplication* with simulated and experimental data at the configuration $\mathcal{C}_0$. This is exactly the *Transfer Function* approach given in [2]. Besides, both transformations suppose a kind of *conservation property* of the simulation error.
To resume, it seems that global transformations need
– the presence of covariate x making correspondance between simulated and experimental data,
– some conservation property .

In practice, the covariate x allowing to "compare" each data may not be available, i.e one only has at disposal samples $\mathbf{Y}^{exp(0)}$, $\mathbf{Y}^{sim(0)}$, $\mathbf{Y}^{sim(c)}$ and nothing else (think about cloud of points). That's why it would be interesting to investigate other transformations.

7.4 $\mathcal{F}$-transformation

Let us consider the samples $\mathbf{Y}^{exp(0)}$, $\mathbf{Y}^{sim(0)}$ and $\mathbf{Y}^{sim(c)}$. Now, suppose that these samples are drawn from some distributions and denote by $F^{exp(0)}$, $F^{sim(0)}$ and $F^{sim(c)}$ the corresponding empirical distribution functions, respectively. Now, by considering the formalism of

the article [4], for some feature space $\mathcal{F}$ equipped with a metric $\|\cdot\|_{\mathcal{F}}$, let us denote by $\rho_{\mathcal{F}}^{exp(0)}$, $\rho_{\mathcal{F}}^{sim(0)}$ and $\rho_{\mathcal{F}}^{sim(c)}$ the characteristics in $\mathcal{F}$ related to the distributions $F^{exp(0)}$, $F^{sim(0)}$ and $F^{sim(c)}$, respectively. For instance, it can be the means, the density functions etc ... We recall that our approach aims at seeking (generalizing global) transformations by including some trend aspect.

We call **$\mathcal{F}$-transformation** or **$\mathcal{F}$-operator** an application

$$\mathcal{T} : \mathcal{F} \longrightarrow \mathcal{F}.$$

We denote by $\mathbb{T}_{\mathcal{F}}$ a set of $\mathcal{F}$-transformations. For example, in the global transformations case, we have $\mathcal{F} = \mathbb{R}^n$ and the characteristics $\rho_{\mathcal{F}}^{(\cdot)}$ is simply the data sample itself. Here, we propose to deal with other characteristics. In other words, we propose several ways of "linking" simulated data to experimental ones, where both are viewed as samples drawn from some distributions.
A transformation which links $\rho_{\mathcal{F}}^{sim(0)}$ to $\rho_{\mathcal{F}}^{exp(0)}$ can be investigated as

$$\widehat{\mathcal{T}} = \underset{\mathcal{T}\in\mathbb{T}_{\mathcal{F}}}{\text{Argmin}} \left\| \rho_{\mathcal{F}}^{exp(0)} - \mathcal{T}(\rho_{\mathcal{F}}^{sim(0)}) \right\|_{\mathcal{F}}^2 .$$

But, considering only one reference configuration $\mathcal{C}_0$ will inevitably lead to a deep lack of generalization over other configurations, in particular the one of interest. We can see the preceding display as a "least squares regression" estimator computed with "one" observation (x_0, y_0) (think $x_0 = \rho_{\mathcal{F}}^{exp(0)}$ and $y_0 = \rho_{\mathcal{F}}^{sim(0)}$) and which has to predict the output y_c (think $\rho_{\mathcal{F}}^{exp(c)}$) by $\widehat{\mathcal{T}}(x_c)$ ($x_c = \rho_{\mathcal{F}}^{sim(c)}$). Such prediction may have poor performance.
Ideally, let us suppose that one has at disposal several configurations, we call *training configurations*, noted $\mathcal{C}_0 = \{c_1, ..., c_N\}$. For each configuration, let us denote by $(\rho_{\mathcal{F}}^{sim(l)}, \rho_{\mathcal{F}}^{exp(l)})$, $l = 1, ..., N$, the corresponding simulated and experimental characteristics. Then, a $\mathcal{F}$-transformation can be investigated by

$$\widehat{\mathcal{T}} = \underset{\mathcal{T}\in\mathbb{T}_{\mathcal{F}}}{\text{Argmin}} \sum_{l=1}^{N} \left\| \rho_{\mathcal{F}}^{exp(l)} - \mathcal{T}(\rho_{\mathcal{F}}^{sim(l)}) \right\|_{\mathcal{F}}^2 . \tag{7.10}$$

Hence, the transformation $\widehat{\mathcal{T}}$ includes more information (above all when N is large) and it becomes legitim to see such transformation as the (approximation of) "mechanism" linking *Simulation* to *Experiments*. Let us notice that if we consider the mean characteristics, i.e $\rho_{\mathcal{F}}^{(\cdot)} = \frac{1}{n}\sum_{i=1}^{n} Y_i^{(\cdot)}$, the procedure (7.10) turns out to be a classical least squares regression

$$\widehat{\mathcal{T}} = \underset{\mathcal{T}\in\mathbb{T}_{\mathcal{F}}}{\text{Argmin}} \sum_{l=1}^{N} \|y_l - \mathcal{T}(x_l)\|_{\mathcal{F}}^2 , \quad y_l = \frac{1}{n}\sum_{i=1}^{n} Y_i^{exp(l)} , \quad x_l = \frac{1}{n}\sum_{i=1}^{n} Y_i^{sim(l)} .$$

The procedure (7.10) can be viewed as a *functional regression*, i.e the inputs and outputs are functions. This analysis is let to a forthcoming work.
In practice, to have several training configurations is very expensive for some applications. In the next section, we give numerical examples in the case of global transformations (i.e $\mathcal{F} = \mathbb{R}^n$ and $\rho_{\mathcal{F}}^{(\cdot)} = \mathbf{Y}^{(\cdot)}$).

7.5 Numerical example for global transformations

Let us consider $n = 50$ realizations $(x_1, ..., x_n)$ of the uniform distribution on $[-2, 2]$. Suppose that the "true" experimental data at the configuration $\mathcal{C}$ that we want to seek are generated as follows : for all $i = 1, ..., n$,

$$Y_i^{exp(c)} = \exp(x_i)\,.$$

Then, the corresponding simulated data are supposed to be generated as : for all $i = 1, ..., n$,

$$Y_i^{sim(c)} = 1 + x_i\,.$$

In other words, the simulation fact (approximation) is to be understood as a first order Taylor expansion. Let us recall that in practice we only have at disposal simulated data at configuration $\mathcal{C}$.
Now, we need some reference simulated and experimental data. In fact, we will consider three sets of reference data in order to investigate the robustness of the methods. Let us consider these three experimental data sets, for all $i = 1, ..., n$

– $Y_i^{exp(1)} = 3 + \exp(0.2\,x_i)$
– $Y_i^{exp(2)} = 3 + \exp(0.5\,x_i)$
– $Y_i^{exp(3)} = 3 + \exp(0.8\,x_i)\,.$

Since the transition to the simulation is "the first order Taylor expansion", we get the following corresponding simulation data sets, respectively

– $Y_i^{sim(1)} = 4 + 0.2\,x_i$
– $Y_i^{sim(2)} = 4 + 0.5\,x_i$
– $Y_i^{sim(3)} = 4 + 0.8\,x_i\,.$

We plot the experimental data sets versus the x_i's in the Figure 7.2. For clarity, we omit to plot the simulated data which are in fact located on the tangent at 0 of the experimental curves.

In the Figures 7.3, 7.4 and 7.5, we represent the estimation by global transformations of the experimental data, which is in fact $(x_i, \exp(x_i))$ (=configuration $\mathcal{C}$) unknown in reality, for each reference configuration $(\mathbf{Y}^{sim(1)}, \mathbf{Y}^{exp(1)})$, $(\mathbf{Y}^{sim(2)}, \mathbf{Y}^{exp(2)})$ and $(\mathbf{Y}^{sim(3)}, \mathbf{Y}^{exp(3)})$. Let us give an example of calculation (at reference (1)) :
from model $\widehat{\mathcal{T}}_1$,

$$\begin{aligned} \widehat{Y}_i^{exp(c)} &= Y_i^{sim(c)} + Y_i^{exp(1)} - Y_i^{sim(1)} \\ &= 0.8\,x_i + \exp(0.2\,x_i)\,, \end{aligned}$$

from model $\widehat{\mathcal{T}}_2$,

$$\begin{aligned} \widehat{Y}_i^{exp(c)} &= \frac{Y_i^{exp(1)}}{Y_i^{sim(1)}}\, Y_i^{sim(c)} \\ &= \frac{3 + \exp(0.2\,x_i)}{4 + 0.2\,x_i}\,(4 + x_i)\,. \end{aligned}$$

First, we can notice that the transformations $\widehat{\mathcal{T}}_1$ and $\widehat{\mathcal{T}}_2$ provide similar results. Then, we also remark that the choice of the reference configuration affects deeply the predictions provided from the two global models. In other words, it seems that the reference configuration should not be "far" from the configuration of interest (see Figure 7.5). We give the L_2 errors in Table

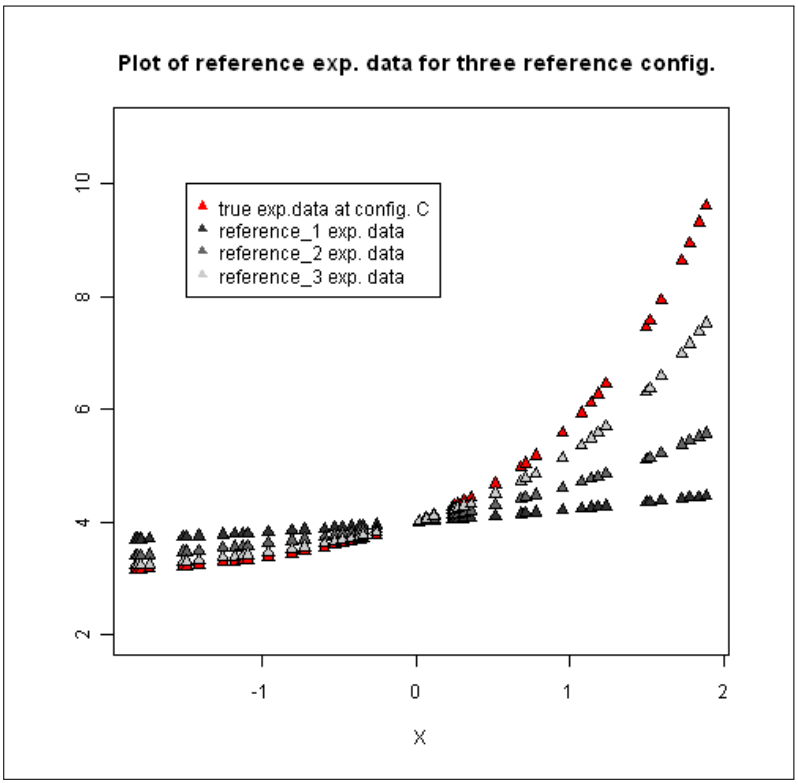


FIGURE 7.2 – Plot of reference experimental data sets

7.1. Finally, these global models may cause *robustness* problem related to the choice of a reference configuration.

	from model $\widehat{\mathcal{T}}_1$	from model $\widehat{\mathcal{T}}_1$
L_2 errors with ref (1)	**5.54**	**5.55**
L_2 errors with ref (2)	**4.49**	**4.48**
L_2 errors with ref (3)	**2.30**	**2.29**

TABLE 7.1 – L_2 errors

As a consequence, it would be interesting to investigate a method less sensitive to the previous ones and, in the same time, which improves the results.

7.6 Statistical learning approach

Let us recall that the goal of our purpose is to construct "experimental" data corresponding to the simulation set $\mathbf{Y}^{sim(c)}$ (at some configuration $\mathcal{C}$), given the reference samples $\mathbf{Y}^{exp(0)}$ and $\mathbf{Y}^{sim(0)}$ (at some reference configuration $\mathcal{C}_0$). In the previous sections, we presented approaches based on data transformations close to heuristics, with some implicit assumptions that can be strong enough. As a matter of fact, statistically speaking, the problem amounts to *predict* experimental data $Y_i^{exp(c)}$ (viewed as "output") corresponding to the simulation $Y_i^{sim(c)}$ (viewed as "input"), for $i = 1, ..., n$, given the training set $(Y_1^{sim(0)}, Y_1^{exp(0)}), ..., (Y_n^{sim(0)}, Y_n^{exp(0)})$, supposed to be independent. In other words, the main task is to **learn** the structure linking experimental and simulated data at the reference configuration $\mathcal{C}_0$ (training set) and then, use

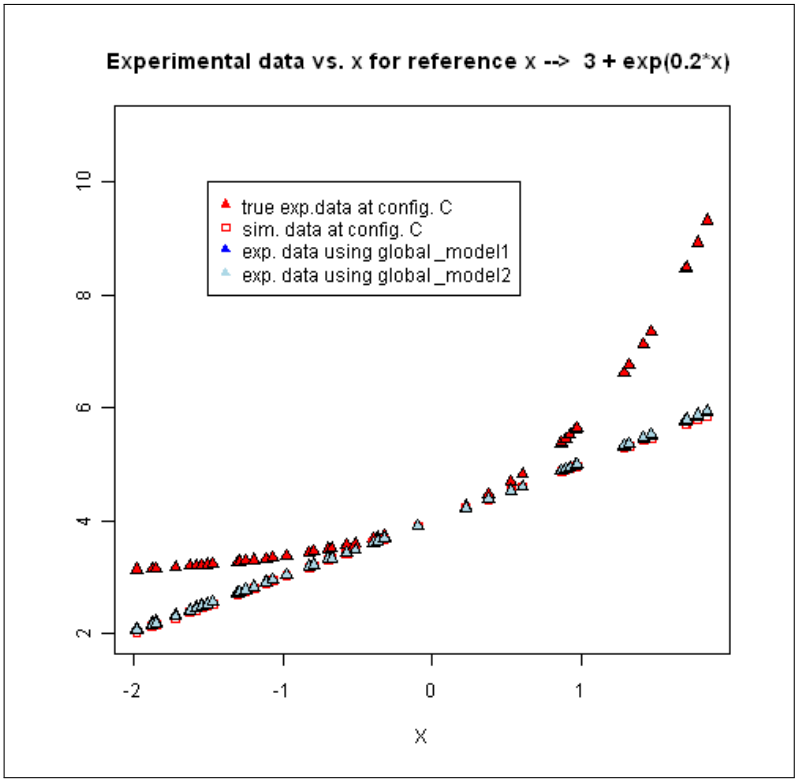


FIGURE 7.3 – Experimental data estimation using the global transformation $\widehat{\mathcal{T}}_1$ (in blue) and $\widehat{\mathcal{T}}_2$ (in lightblue) from reference data $(\mathbf{Y}^{sim(1)}, \mathbf{Y}^{exp(1)})$.

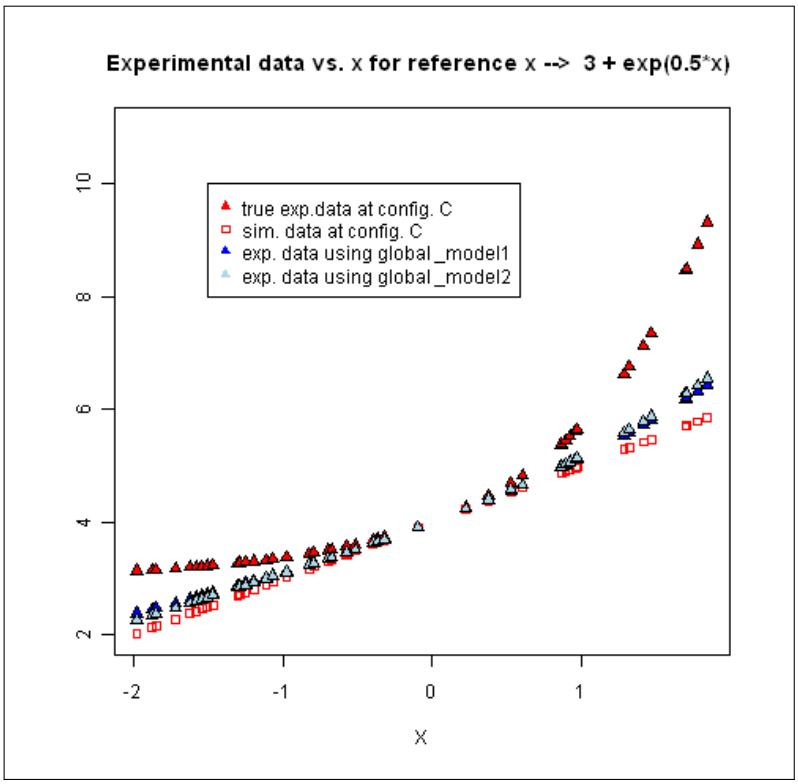


FIGURE 7.4 – Experimental data estimation using the global transformation $\widehat{\mathcal{T}}_1$ (in blue) and $\widehat{\mathcal{T}}_2$ (in lightblue) from reference data $(\mathbf{Y}^{sim(2)}, \mathbf{Y}^{exp(2)})$.

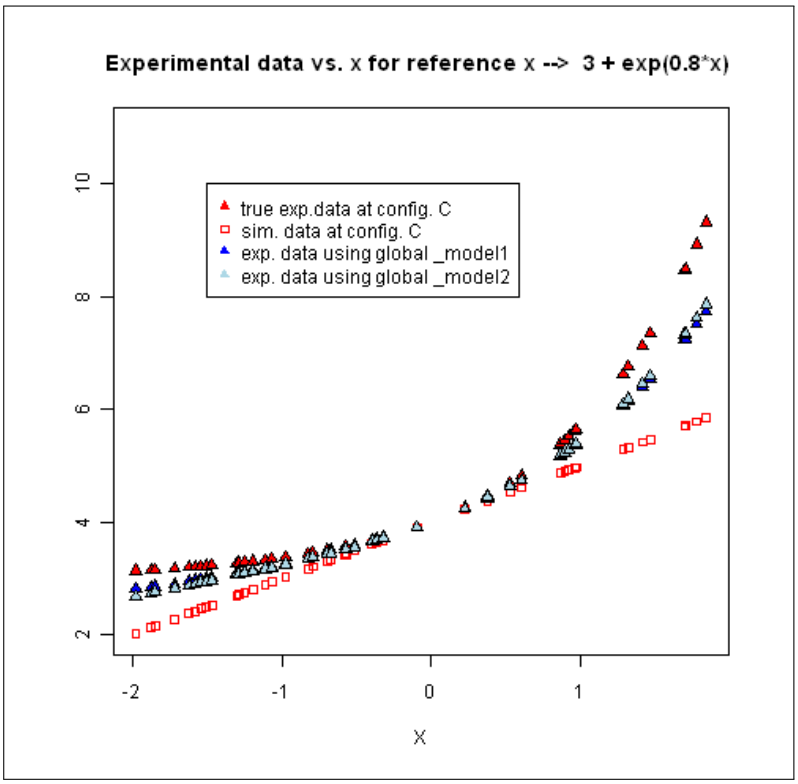


FIGURE 7.5 – Experimental data estimation using the global transformation $\widehat{\mathcal{T}}_1$ (in blue) and $\widehat{\mathcal{T}}_2$ (in lightblue) from reference data $(\mathbf{Y}^{sim(3)}, \mathbf{Y}^{exp(3)})$.

it to **predict** the wanted "experimental" data from the simulation set $\mathbf{Y}^{sim(c)}$ at the configuration $\mathcal{C}$. Similarly, the task can also been understood as *learning* the error between simulation and experiments. Finally, we can adopt the classical framework of statistical learning where the inputs are simulated data : $X = Y^{sim(\cdot)} \in \mathcal{X}$ with $\mathcal{X} \subset \mathbb{R}$, and where the outputs are experimental data $Y = Y^{exp(\cdot)} \in \mathcal{Y}$ with $\mathcal{Y} \subset \mathbb{R}$.
Next, for simplicity we use the notation $Y^{exp(0)} = Y^{exp}$ and $Y^{sim(0)} = Y^{sim}$.

Modeling

Let us suppose the following modeling

$$Y_i^{exp} = h(Y_i^{sim}, \boldsymbol{\theta}) + \varepsilon_i\,, \quad i = 1, ..., n\,, \tag{7.11}$$

where $\{y \mapsto h(y, \boldsymbol{\theta})\,,\ \boldsymbol{\theta} \in \Theta\}$ is some *model* and ε_i are i.i.d centered errors with constant variance σ^2. We denote by $Q^{\mathbf{z}}$ the (unknown) joint distribution of $Z = (Y^{sim}, Y^{exp})$ and by $P^{\mathbf{x}}$ the distribution of Y^{sim} supposed to be known or at least a large sample is available. Moreover, let us notice that the simulated data Y_i^{sim} providing the experimental data Y_i^{exp} may not be observed in that the display (7.11) does not lead to a regression procedure. However, the presence of some covariate x as mentioned in the previous sections allows such procedures. Both of these aspects are taken into account by the formalism given in [4] through the use of contrast functions we briefly recall in the next subsection.
Finally, the method consists in three steps

1. Specification of a model $\{y \mapsto h(y, \boldsymbol{\theta})\,,\ \boldsymbol{\theta} \in \Theta\}$

2. Parameter estimation $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$
3. Predictions computation of experimental data $\widehat{Y}_i^{exp(c)} = h(Y_i^{sim(c)}, \widehat{\boldsymbol{\theta}})$, $i = 1, ..., n$.

Let us pay some attention to the prediction computation. Indeed, the prediction $h(Y_i^{sim(c)}, \widehat{\boldsymbol{\theta}})$ has a statistical sense when the data $Y^{sim(c)}$ is drawn from the **same** distribution as $Y^{sim} = Y^{sim(0)}$, i.e $P^{\mathbf{x}}$. This is not necessarily clear in practice. Such approximation is in the same spirit than "conservation properties" met in Section 7.3.

Parameter estimation

Following the framework in [4], a general estimation procedure can be written

$$\widehat{\boldsymbol{\theta}}_{\Psi} = \underset{\boldsymbol{\theta}\in\Theta}{\text{Argmin}} \frac{1}{n} \sum_{i=1}^{n} \Psi\left(\rho_{\mathcal{F}}(\boldsymbol{\theta}), Z_i\right) , \tag{7.12}$$

for some contrast $\Psi : \mathcal{F} \to L_1(Q^{\mathbf{z}})$ and characteristic $\rho_{\mathcal{F}}(\boldsymbol{\theta})$. In this study, we only present the regression procedure (i.e Y_i^{sim} are observed) where the reg-contrast is given by

$$\text{for}\, \rho \in L_2(P^{\mathbf{x}}) , \quad \Psi(\rho, (y^{sim}, y^{exp})) = (y^{exp} - \rho(y^{sim})) ,$$

and $\rho_{\mathcal{F}}(\boldsymbol{\theta}) = h(\cdot, \boldsymbol{\theta})$. We get the following estimation procedure

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\Theta}{\text{Argmin}} \frac{1}{n} \sum_{i=1}^{n} \left(Y_i^{exp} - h(Y_i^{sim}, \boldsymbol{\theta})\right)^2 . \tag{7.13}$$

For instance, a classical model $h(\cdot, \boldsymbol{\theta})$ is a polynomial class, as for example

$$h(x, \boldsymbol{\theta}) = \sum_{l=1}^{k} \theta_l \, x^l ,$$

which provides a linear model in the parameter $\boldsymbol{\theta}$.

7.7 Numerical example with learning approach

Let us consider the numerical framework at Section 7.5. Next, let $h(\cdot, \boldsymbol{\theta})$ be the polynomial model of degree 3

$$h(y, \boldsymbol{\theta}) = \theta_1 + \theta_2 \, y + \theta_3 \, y^2 + \theta_4 \, y^3 , \quad \boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3, \theta_4)^T .$$

Then, given the three sets of reference data $(\mathbf{Y}^{sim(1)}, \mathbf{Y}^{exp(1)})$, $(\mathbf{Y}^{sim(2)}, \mathbf{Y}^{exp(2)})$ and $(\mathbf{Y}^{sim(3)}, \mathbf{Y}^{exp(3)})$, one can compute the three parameters

$$\widehat{\boldsymbol{\theta}}^{(\nu)} = \underset{\boldsymbol{\theta}\in\Theta}{\text{Argmin}} \frac{1}{n} \sum_{i=1}^{n} \left(Y_i^{exp(\nu)} - h(Y_i^{sim(\nu)}, \boldsymbol{\theta})\right)^2 , \quad \text{for } \nu = 1, 2, 3 . \tag{7.14}$$

Hence, given the simulated data at configuration $\mathcal{C} : Y_i^{sim(c)} = 1 + x_i$, $i = 1, ..., n$, one computes the corresponding experimental data, for each configuration $\nu = 1, 2$ and 3, by

$$\text{for all } i = 1, ..., n , \quad \widehat{Y}_i^{exp(c)}(\nu) = h(Y_i^{sim(c)}, \widehat{\boldsymbol{\theta}}^{(\nu)}) .$$

This is represented in the Figures 7.6, 7.7 and 7.8 where we simply overlay, at each configuration , the curve obtained by the learning approach to the previous graphs of Section 7.5. Contrary to the global cases, the learning procedure seems to be less sensitive to the reference configuration and provide a quite good performance even when the reference is "far" from the configuration of interest, see Figure 7.6. The corresponding L_2 error is 0.98 whereas for the global models the L_2 error is about 5.55 at the same configuration.

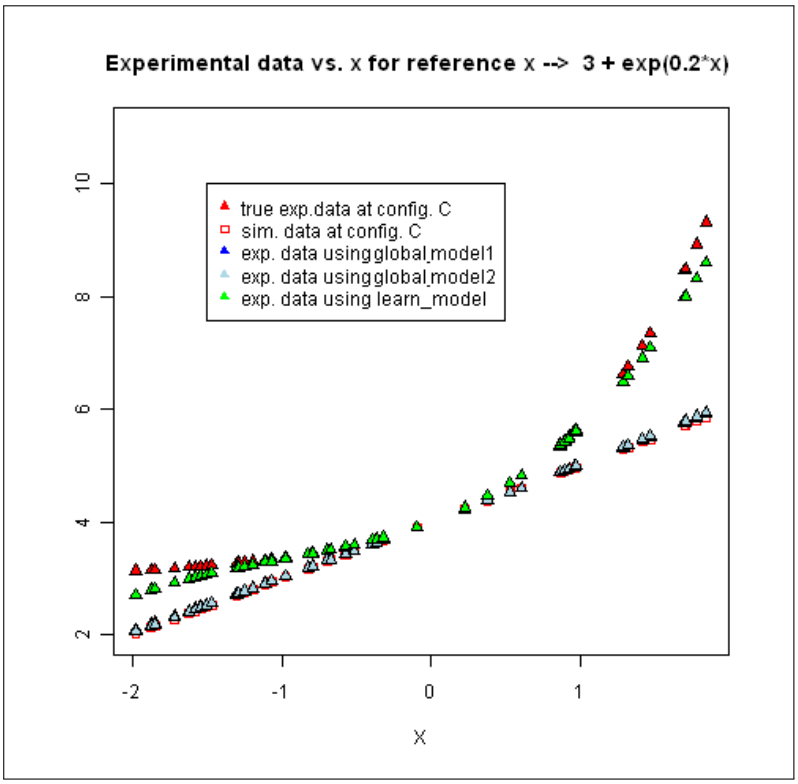


FIGURE 7.6 – Experimental data estimation using the global transformations $\widehat{\mathcal{T}}_1$ (in blue), $\widehat{\mathcal{T}}_2$ (in lightblue) and the learning method (in green), from reference data $(\mathbf{Y}^{sim(1)}, \mathbf{Y}^{exp(1)})$.

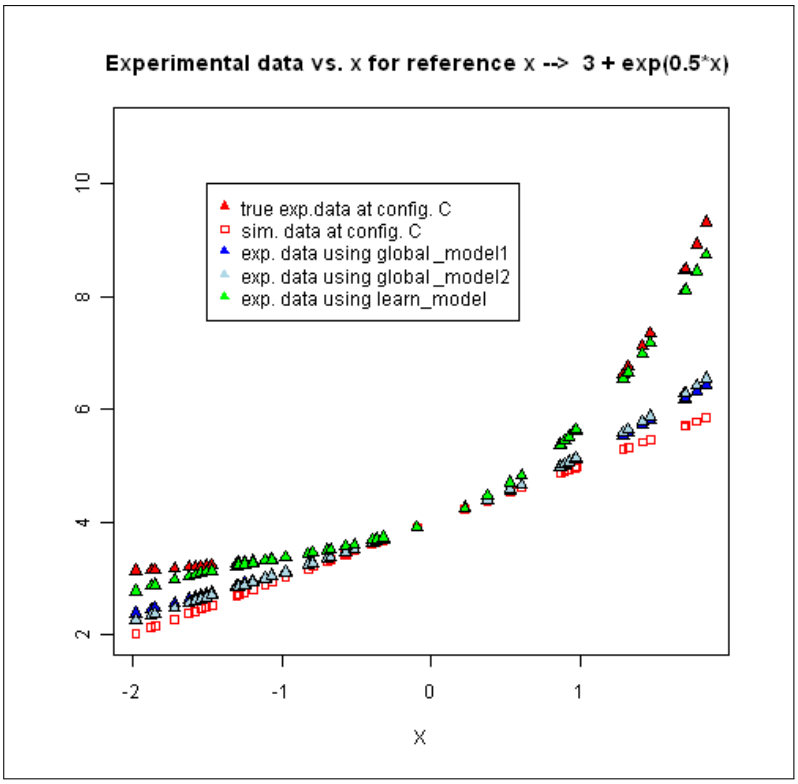


FIGURE 7.7 – Experimental data estimation using the global transformation $\widehat{\mathcal{T}}_1$ (in blue), $\widehat{\mathcal{T}}_2$ (in lightblue) and the learning method (in green), from reference data $(\mathbf{Y}^{sim(2)}, \mathbf{Y}^{exp(2)})$.

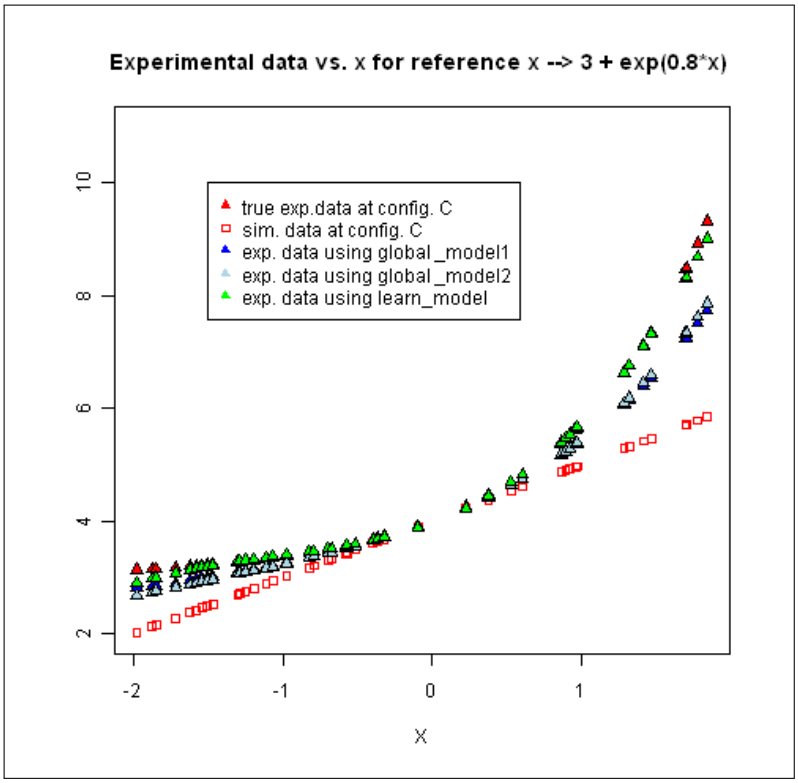


FIGURE 7.8 – Experimental data estimation using the global transformation $\widehat{\mathcal{T}}_1$ (in blue), $\widehat{\mathcal{T}}_2$ (in lightblue) and the learning method (in green), from reference data $(\mathbf{Y}^{sim(3)}, \mathbf{Y}^{exp(3)})$.

	reference (1)	reference (2)	reference (3)
L_2 errors	**0.98**	**0.72**	**0.29**

TABLE 7.2 – L_2 errors

7.8 Conclusions

The global methods presented can hide strong assumptions as we saw in Section 7.3 (difference, ratio conservation...) and can potentially mislead the understanding of the phenomenon of interest if no precautions are taken. We try to propose a more systematic method in the regression framework. The first numerical results obtained seem to be promising. An interesting issue would be, on one hand, to analyse quantitatively the effect of the modeling choice of $h(\cdot, \boldsymbol{\theta})$, and on the other hand, to study the effect of the "reference" choice. Both choices may be related.

Bibliographie

[1] A.P. Berens. Nde reliability data analysis. *ASM Handbook.*, 17 :689–701, 1989.

[2] C.A. Harding, G.R. Hugo, and S.J. Bowles. Application of model assisted pod using a transfer function approach. *Review of Quantitative Nondestructive Evaluation*, 28, 2009.

[3] F. Jenson, E. Iakovleva, and N. Dominguez. Simulated supported pod : methodology and hfet validation case. *Review of Quantitative Nondestructive Evaluation*, 30, 2010.

[4] N. Rachdi, J.C. Fort, and T. Klein. Risk bounds for new M-estimation problems . *hal 00537236*, 2010.

Chapitre 8

Conclusion générale et perspectives

Il existe des conceptions vulgaires tout à fait suffisantes pour la vie pratique ; elles doivent même être la nourriture des hommes. Elles ne suffisent cependant pas à l'intelligence.

Averroès (Ibn Ruchd) ou *Le Commentateur (d'Aristote)*, philosophe arabe du XII ème siècle

Tout au long de ce manuscript de thèse, nous avons développé une méthodologie formelle dans le cadre de l'apprentissage statistique ayant pour but l'étude et l'analyse de modèles numériques de types entrées/sorties en présence de données incertaines. Les avancées de nos travaux ouvrent plusieurs perspectives, nous en présentons quelques unes.

Amélioration des constantes dans le Théorème 3.4.1

Les constantes intervenants dans l'inégalité du Théorème 3.4.1 ne sont pas optimales dans le sens où nous nous sommes davantage concentrés sur la méthode permettant d'obtenir de telles inégalités. Un premier prolongement à donner à ce travail serait de raffiner les hypothèses du Théorème, notamment celles portant sur le contraste. En second lieu, une étude plus fine des processus empiriques mis en jeu mériterait d'être entreprise.

Prédiction et estimation par "contraste régulier"

Les récents travaux de A. Saumard sur l'estimation par minimum de contraste régulier [7] pourraient contribuer à l'étude de la dualité estimation-prédiction abordée au Chapitre 4. En particulier, nous proposions l'étude de la différence donnée à la Section 4.4

$$\mathcal{R}_{\Psi^p}(\widehat{\boldsymbol{\theta}}_{\Psi}) - \mathcal{R}_{\Psi^p}(\widehat{\boldsymbol{\theta}}_{\Psi^p})\,.$$

Etant donné une quantité d'intérêt à prédire, le choix d'un contraste pourrait, outre le calcul de sa "distance" au contraste caractéristique de la quantité d'intérêt, reposer sur sa "régularité". Dans l'exemple de prédiction d'une densité, l'estimation des paramètres peut se faire avec le log-contraste

$$\Psi_{\log}(\rho, y) = -\log(\rho)(y)$$

ou avec le L_2-contraste

$$\Psi_{L_2}(\rho, y) = \|\rho\|_2^2 - 2\,\rho(y)\,.$$

Sous certaines hypothèses de convexité, ces deux contrastes caractérisent la densité (par projection) et il faut donc déterminer sous quelles conditions (sur le modèle etc...) un contraste est "plus performant" qu'un autre. Dans cette évaluation de la performance, la régularité intervient dans la complexité algorithmique. Cet aspect est évidemment important en pratique.

Apprentissage "actif" et Apprentissage "semi-supervisé"

Dans le Chapitre 3, nous avons proposé une méthode d'estimation paramétrique basée sur les données $Y_1, ..., Y_n$ et $\mathbf{X}'_1, ..., \mathbf{X}'_m$

$$\widehat{\boldsymbol{\theta}}_{\Psi} = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{Argmin}} - \sum_{i=1}^{n} \Psi\left(\sum_{j=1}^{m} \widetilde{\rho}_{\mathcal{F}}(h(\mathbf{X}'_j, \boldsymbol{\theta})),\, Y_i\right)\,.$$

Ce paramètre peut alors servir à estimer la quantité $\rho_{\mathcal{F}}(\boldsymbol{\theta})$ par $\rho_{\mathcal{F}}(\widehat{\boldsymbol{\theta}}_{\Psi})$.
Il serait envisageable de mettre en perspective le cadre précédent avec l'"apprentissage actif" et l'"apprentissage semi-supervisé" [2], [1], [3]. Par des points de vue différents (que nous ne détaillerons pas ici), ces deux disciplines traitent de l'apprentissage en présence de données "labellées" $(\mathbf{X}_i^{lab}, Y_i^{lab})_{1\leq i\leq n}$ et de données "non labellées" $\mathbf{X}_{n+1}^{nlab}, ..., \mathbf{X}_N^{nlab}$. Dans le cadre cité plus haut, les données "non labellées" correspondent aux données $\mathbf{X}'_1, ..., \mathbf{X}'_m$ et les données $\mathbf{X}_i^{lab}$ ne sont pas observées. Une question pourrait se poser : si la $P^{\mathbf{x}}$ (qui génère les $\mathbf{X}'_j$) est connue, peut-on améliorer l'estimation de $\rho_{\mathcal{F}}(\boldsymbol{\theta})$ en choisissant convenablement les $\mathbf{X}'_j$?

Utilisation de métamodèles dans les algorithmes d'apprentissage

Dans ces travaux de thèse, nos développements ont porté sur un modèle h sensé être assez "maniable" (cf. Introduction). Or, en pratique même si le *run* est de l'ordre de la seconde, le temps de calcul d'un Ψ-estimateur du type $\widehat{\boldsymbol{\theta}}_{\Psi} = \text{Argmin}_{\boldsymbol{\theta}\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \Psi\left(\rho_{\mathcal{F}}^{m}(\boldsymbol{\theta}), Z_i\right)$, par exemple

$$\widehat{\boldsymbol{\theta}}_{\log} = \underset{\boldsymbol{\theta}\in\Theta}{\text{Argmin}} \; -\sum_{i=1}^{n} \log\left(\sum_{j=1}^{m} K_b(Y_i - h(\mathbf{X}'_j, \boldsymbol{\theta}))\right) , \quad \text{pour un noyau } K_b$$

peut ne pas être négligeable car il nécessite m runs à chaque itération. Une alternative serait d'utiliser une approximation du modèle h, généralement appelée métamodèle (par exemple Krigeage, approximation polynômiale, développement dans une base etc...). Il serait intéressant d'étudier quantitativement l'effet de l'approximation sur la complexité des algorithmes d'apprentissage.

Plans d'expériences et quantité d'intérêt

Dans de nombreux cas, l'échantillon $\mathbf{X}'_1, ..., \mathbf{X}'_m$ n'est pas directement simulé, mais provient d'une base de données obtenue à partir d'un plan d'expérience. Dans ce cas, il faudrait tenir compte de la nature de cet "échantillon" dans l'étude qualitative de la procédure d'estimation. On renvoie aux travaux de WG. Müller [5], L. Pronzato et A.A. Zhigljavsky [6], ainsi qu'aux références qui y figurent.

Étude numérique et algorithmique sur des cas réels

Une perspective de ces travaux de thèse est l'étude de cas réels avec le formalisme décrit dans ce manuscript. Bien entendu, les aspects algorithmiques et informatiques (choix du langage, optimisation du temps de calcul, mise en place de wrappers etc...) seront essentiels pour cette étude.

Extension de la méthodologie au contexte de modèles hiérarchisés (granularité) et/ou multi-disciplinaire

La question du choix de la *granularité*[1] de modèles en tenant compte des incertitudes en ingénierie de conception est un véritable enjeu industriel. En particulier, une problématique importante est le choix du *niveau de modélisation* en vue d'un certain post-traitement tout en

1. terme pour désigner un certain niveau (raffinement) de modélisation

quantifiant les erreurs induites.
Par ailleurs, le modèle h peut soit être *mono-physique* (par exemple, un modèle en acoustique, en électromagnétisme etc...) ou bien être *multi-physique* c'est à dire intégrant plusieurs physiques (par exemple, électromagnétisme et acoustique), ce qui est le cas en conception où l'architecture globale du produit est étudiée.
Dans le cas mono-physique, le problème s'inscrit clairement dans le cadre de la sélection de modèles, voir les travaux de P. Massart [4]. Il s'agit d'une prolongation naturelle de ces travaux.
Dans le cas multi-physique, un point important est l'étude de la contribution d'une discipline particulière vis-à-vis d'une certaine performance considérée sur le produit. Il s'agit donc d'une analyse de sensibilité relative à chaque discipline. De même, l'analyse de la *dépendance entre disciplines* est importante puisqu'elle peut avoir un impact non négligeable sur le résultat global.

Algorithme Stochastique Dynamique

Le Chapitre 5 propose un algorithme original S^2dyn, basé à la fois sur une méthode de régularisation et sur des algorithmes stochastiques. En quelque sorte, il s'agit d'*Algorithmes Stochastiques Régularisants*. Une perspective naturelle serait l'étude de l'algorithme proposé, en étudiant notamment l'effet de la "vitesse" de lissage $t \mapsto \sigma_t$ sur la convergence de l'algorithme. On pourra s'inspirer des travaux de A. Zhigljavsky et A. and Zilinskas [8] sur l'optimisation globale stochastique.

Bibliographie

[1] A. Beygelzimer, S. Dasgupta, and J. Langford. Importance weighted active learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 49–56. ACM, 2009.

[2] O. Chapelle, B. Schölkopf, and A. Zien. *Semi-supervised learning*, volume 2. MIT press Cambridge, MA, 2006.

[3] S. Dasgupta and J. Langford. Active learning tutorial, icml 2009.

[4] P. Massart. *Concentration inequalities and model selection : Ecole d'Eté de Probabilités de Saint-Flour XXXIII-2003*. Springer Verlag, 2007.

[5] W.G. Müller. *Collecting spatial data : optimum design of experiments for random fields*. Springer Verlag, 2007.

[6] L. Pronzato and A. Zhigljavsky. *Optimal design and related areas in optimization and statistics*. Springer Verlag, 2008.

[7] A. Saumard. (phd thesis) estimation par minimum de contraste régulier et heuristique de pente en sélection de modèles. 2010.

[8] A. Zhigljavsky and A. Zilinskas. *Stochastic global optimization*. Springer New York, 2007.

Annexe

Éléments sur les processus empiriques

8.1 Introduction

Nous donnons une brève introduction de la théorie des *Processus Empiriques* inspirée du livre de M.R. Kosorok [4], qui prépare au remarquable travail de A.W. van der Vaart et J.A. Wellner [7] qui est l'une des références principales.
Cette théorie a donné de puissants outils permettant une étude fine et rigoureuse de procédures statistiques, des plus classiques aux plus exotiques, dans un cadre asymptotique et non asymptotique. Les travaux précurseurs ont été menés par M.D. Donsker dans les années 1950 [1], suivis de ceux de R.M. Dudley [2], D. Pollard [5], P. Gaenssler [3], Galen R. Shorack et Jon A. Wellner [6] parmi d'autres.
Tout au long de cette introduction, nous illustrerons les différentes notions abordées à travers un exemple classique : celui de la fonction de répartition empirique d'un échantillon.
Par ailleurs, nous ne discuterons pas des arguments de mesurabilité relatifs aux processus empiriques qui mènent à des sujets hors de notre propos.

8.2 Définitions et notations

Comme son nom l'indique, un processus empirique est un processus stochastique particulier dont l'aléa provient d'un échantillon. On le définit comme suit.
Soit $\xi_{1..n} := \xi_1, ..., \xi_n$ un n-échantillon i.i.d d'une variable aléatoire ξ définie sur $\mathcal{W}$ et de loi P. La *mesure empirique* associée à cet échantillon est définie par

$$P_n := \frac{1}{n} \sum_{i=1}^{n} \delta_{\xi_i} ,$$

où δ_w est la masse de Dirac en w.
Pour toute mesure P sur $\mathcal{W}$ et toutes fonctions $g : \mathcal{W} \to \mathbb{R}$, on notera

$$P\,g := \mathbb{E}_{\xi \sim P}(g(\xi)) = \int_{\mathcal{W}} g(w)\, P(dw) .$$

De plus, si $\mathcal{G}$ est une classe (ensemble) de fonctions de $\mathcal{W}$ dans $\mathbb{R}$, on notera

$$\|P\|_{\mathcal{G}} := \sup_{g \in \mathcal{G}} |P\,g| .$$

Définition 8.2.1. Processus empirique .
Soit $\mathcal{G}$ une classe de fonctions mesurables $g : \mathcal{W} \to \mathbb{R}$.
Le *processus empirique* associé à $\xi_{1..n}$ et indexé par $\mathcal{G}$ est défini par

$$\begin{aligned} \mathbb{G}_n : \mathcal{G} &\longrightarrow \mathbb{R} \\ g &\longmapsto \mathbb{G}_n g := \sqrt{n}(P_n - P)g \,. \end{aligned}$$

Remarque 8.2.1. Par abus de notation, la mesure $\mathbb{G}_n$ sera identifiée au processus empirique.

Selon l'usage, on utilisera les différentes écritures

$$\mathbb{G}_n g = \int_{\mathcal{W}} g(w)\mathbb{G}_n(dw) = \sqrt{n} \int_{\mathcal{W}} g(w)(P_n - P)(dw) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (g(\xi_i) - \mathbb{E}(g(\xi))) \,.$$

Soit $g \in \mathcal{G}$. On sait, par la loi forte des grands nombres, que si $\mathbb{E}(g(\xi))$ existe, alors

$$\frac{1}{\sqrt{n}} \mathbb{G}_n g \longrightarrow 0 \quad \text{presque sûrement} \,.$$

De plus, si $\mathbb{E}(g^2(\xi)) < +\infty$ alors

$$\mathbb{G}_n g \rightsquigarrow \mathcal{N}(0, P(g - P g)^2) \,,$$

où $\rightsquigarrow$ désigne la convergence en loi et $\mathcal{N}$ la loi normale.
Un des buts premiers de la théorie des processus empiriques est de donner des conditions, notamment sur $\mathcal{G}$, afin de rendre ces convergences **uniforme sur** $\mathcal{G}$. Mais il existait déjà deux résultats "uniformes" bien connus.

8.3 Classes de Glivenko-Cantelli et de Donsker

Dans le cas où $\mathcal{W} = \mathbb{R}$, un processus empirique classique est celui donné par la classe de fonctions des indicatrices

$$\mathcal{G}_{ind} = \{w \mapsto \mathbb{1}_{]-\infty,t]}(w) \,,\, t \in \mathbb{R}\} \,.$$

En notant $\mathbb{F}_n$ la fonction de répartition empirique associée à l'échantillon $\xi_{1..n}$ et F celle donnée par la mesure P, on a

$$\forall\, g \in \mathcal{G}_{ind} \,, \quad \mathbb{G}_n g = \mathbb{G}_n \mathbb{1}_{]-\infty,t]} = \sqrt{n}(\mathbb{F}_n(t) - F(t)) \,.$$

Dans ce cas, le processus empirique pourra être vu comme un processus indexé par $t \in \mathbb{R}$, on notera abusivement $\mathbb{G}_n(t)$.
Cette classe de fonctions particulière présente un grand intérêt puisqu'elle permet de traiter les propriétés de la fonction de répartition empirique, très utile en pratique.
Un premier résultat (connu avant la théorie des processus empiriques) dû a Glivenko et Cantelli qui ont démontré en 1933 la convergence uniforme

$$\sup_{t \in \mathbb{R}} |\mathbb{F}_n(t) - F(t)| \longrightarrow 0 \quad \text{presque sûrement.}$$

Autrement dit que

$$\sup_{g\in\mathcal{G}_{ind}}\left(\frac{1}{\sqrt{n}}\,|\mathbb{G}_n\,g|\right)=\frac{1}{\sqrt{n}}\,\|\mathbb{G}_n\|_{\mathcal{G}_{ind}}\longrightarrow 0\quad\text{presque sûrement}\,.$$

On montrera ce résultat plus tard en utilisant des outils relatifs aux processus empiriques. Cela a motivé la définition plus générale suivante :

Une classe $\mathcal{G}$ est dite P**-Glivenko-Cantelli** si

$$\|P_n-P\|_{\mathcal{G}}=\frac{1}{\sqrt{n}}\,\|\mathbb{G}_n\|_{\mathcal{G}}\longrightarrow 0\quad\text{presque sûrement}\,. \tag{8.1}$$

Remarque 8.3.1. Une classe de Glivenko-Cantelli est donc une classe telle que le supremum du processus empirique $\mathbb{G}_n$ est négligeable devant $\sqrt{n}$, presque sûrement :

$$\mathcal{G}\text{ est }P\text{-Glivenko-Cantelli}\Longleftrightarrow\|\mathbb{G}_n\|_{\mathcal{G}}=o(\sqrt{n})\quad\text{presque sûrement}\,.$$

Ensuite, en reconsidérant la classe des indicatrices $\mathcal{G}_{ind}$ et en notant $l^\infty(\mathbb{R})$ l'espace des fonctions bornées de $\mathbb{R}$ dans $\mathbb{R}$, Donsker montre en 1952 que la suite de processus $(\mathbb{G}_n(t)=\sqrt{n}(\mathbb{F}_n(t)-F(t)),\,t\in\mathbb{R})$ converge en distribution, dans $l^\infty(\mathbb{R})$, vers un processus gaussien $\mathbb{G}=\{\mathbb{G}(t),\,t\in\mathbb{R}\}$. Le processus limite G s'écrit $G(t)=B(F(t))$ où B est un *pont brownien* standard. Et cela pour une mesure P quelconque. D'où la définition :

Une classe $\mathcal{G}$ est dite P**-Donsker** si

$$\|\mathbb{G}_n\|_{\mathcal{G}}\rightsquigarrow\mathbb{G}\,,\quad\text{dans }l^\infty(\mathcal{G})\,, \tag{8.2}$$

où $l^\infty(\mathcal{G})$ est l'espace des fonctions bornées $\mathcal{G}\to\mathbb{R}$ et $\mathbb{G}$ le P-pont brownien, i.e le processus gaussien centré de covariance $\mathrm{Cov}(g_1,g_2)=P\,g_1g_2-P\,g_1\,P\,g_2$.

Remarque 8.3.2. La classe $\mathcal{G}_{ind}$ est *universellement* Donsker, i.e pour toute mesure P sur $\mathcal{W}$.

Par ailleurs, on peut montrer qu'une classe de Donsker est nécessairement de Glivenko-Cantelli (voir [7] page 82).

En outre, ce qui caractérise un processus empirique, au delà de la mesure P qui génère les données, est la classe de fonctions $\mathcal{G}$ qui l'indexe. En effet, elle va jouer un rôle crucial quant aux propriétés du processus empirique. Il serait donc intéressant de savoir si une classe est de Glivenko-Cantelli ou encore Donsker. Pour cela, la théorie des processus empiriques utilise la notion d'*entropie* dans le sens que nous décrivons ci-après.

8.4 Notion d'entropie

On se doute bien que la "complexité" de la classe $\mathcal{G}$ est responsable du comportement du processus empirique. En effet, le processus empirique est un processus stochastique particulier, et le théorème de Prohorov (voir par exemple Theorem 1.3.9 p.21 dans [7]) laisse présager de l'importance d'un argument de compacité. La notion d'entropie que nous allons introduire quantifie la compacité d'une classe de fonctions.

Soit $\mathcal{G}$ une classe de fonctions $g : \mathcal{W} \to \mathbb{R}$ et P une mesure sur $\mathcal{W}$. Pour $1 \leq r < +\infty$, on note

$$\|g\|_{r,P} = \left(\int_{\mathcal{W}} |g(w)|^r \, P(dw) \right)^{1/r} .$$

Pour introduire la notion de compacité dans les classes de fonctions, on donne la définition intuitive suivante.

Définition 8.4.1. $L_r(P)$ **Nombre de recouvrements et entropie.**
On définit le **nombre de recouvrements** $N(\epsilon, \mathcal{G}, L_r(P))$ comme le nombre de boules $\{g, \|g - g_0\|_{r,P}\}$ de rayon ϵ nécessaire pour recouvrir l'ensemble $\mathcal{G}$.
L'**entropie** est définie comme le logarithme du nombre de recouvrements

$$H(\epsilon, \mathcal{G}, L_r(P)) = \log N(\epsilon, \mathcal{G}, L_r(P)) .$$

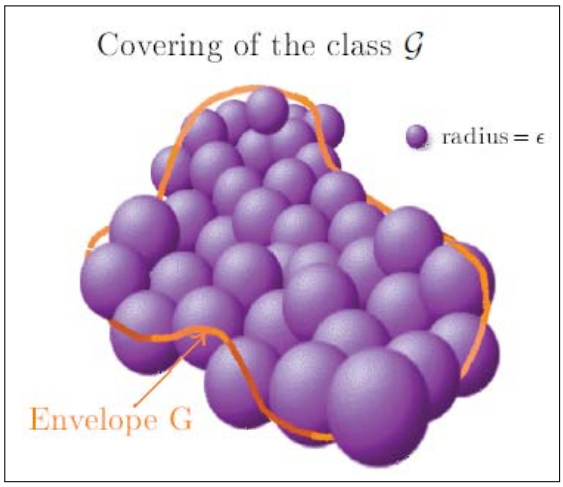


FIGURE 8.1 – Illustration de recouvrements (*empirical process theory and applications* par Sara van de Geer)

Notons que si une classe $\mathcal{G}$ est une partie compacte d'un espace métrique $(E, \|\cdot\|_{r,P})$ alors pour tout $\epsilon > 0$, le nombre $N(\epsilon, \mathcal{G}, L_r(P))$ est fini.
Toutefois, pour éviter quelques difficultés que l'on peut rencontrer avec cette notion d'entropie, nous allons plutôt considerer une autre notion : l'entropie à crochets, un peu moins intuitive peut être, mais qui a le mérite de présenter des résultats simples à comprendre.

Définition 8.4.2. $L_r(P)$ **Nombre de crochets et entropie à crochets.**
Soient deux fonctions l et u, le crochet $[l, u]$ est l'ensemble des fonctions g tel que $l \leq g \leq u$.
Un ϵ-crochet est un crochet $[l, u]$ tel que $||u - l||_{r,P} < \epsilon$.
Un **nombre de crochets**, noté $N_{[]}(\epsilon, \mathcal{G}, L_r(P))$, est le nombre minimal d'ϵ-crochets nécessaires pour recouvrir la classe de fonctions $\mathcal{G}$ dans $L_r(P)$.
L'**entropie à crochet** est le logarithme du nombre à crochets, notée

$$H_{[]}(\epsilon, \mathcal{G}, L_r(P)) = \log N_{[]}(\epsilon, \mathcal{G}, L_r(P)) .$$

L'entropie à crochets nécessite un ordre sur les classes de fonctions et rend donc un peu plus délicate son interprétation. Cependant, on montre sans difficulté que le nombre de recouvrements et le nombre de crochets sont liés par la relation

$$N(\epsilon, \mathcal{G}, L_r(P)) \leq N_{[]}(2\epsilon, \mathcal{G}, L_r(P)) .$$

8.5 Théorème de Glivenko-Cantelli et Théorème de Donsker

Avec la définition précédente de l'entropie à crochet, on a une condition suffisante pour qu'une classe de fonctions soit de Glivenko-Cantelli.

Théorème 8.5.1. (Glivenko-Cantelli) *Soit $\mathcal{G}$ une classe de fonctions mesurables et P une mesure sur $\mathcal{W}$. Si pour tout $\epsilon > 0$ on a $N_{[]}(\epsilon, \mathcal{G}, L_1(P)) < +\infty$, alors $\mathcal{G}$ est P-Glivenko-Cantelli.*

Revenons à la classe des indicatrices $\mathcal{G}_{ind} = \{w \mapsto \mathbb{1}_{]-\infty,t]}(w)\,,\, t \in \mathbb{R}\}$. On a vu que cette classe est bien de Glivenko-Cantelli (pour toute mesure sur $\mathcal{W}$). Vérifions que cette classe satisfait l'hypothèse du théorème précédent.
Soit $\epsilon > 0$, et rappelons que F est la fonction empirique associée à la mesure P. On peut trouver un entier k_ϵ ensemble de points $t_0, ..., t_{k_\epsilon}$ tel que $-\infty = t_0 < ... < t_{k_\epsilon} = +\infty$ et satisfaisant

$$F(t_j-) - F(t_{j-1}) \leq \epsilon\,, \quad \forall\, 1 \leq j \leq k_\epsilon \quad (\text{avec } F(t_0) = 0 \text{ et } F(t_{k_\epsilon}) = 1\,)\,.$$

Comme $0 \leq F \leq 1$, alors $k_\epsilon \leq \frac{1}{\epsilon}$. Considérons l'ensemble de crochets $\{[l_j, u_j]\,,\, 1 \leq j \leq k_\epsilon)\}$ (k_ϵ au total) où $l_j(w) = \mathbb{1}_{]-\infty,t_{j-1}]}(w)$ et $u_j(w) = \mathbb{1}_{]-\infty,t_j[}(w)$ (notons que u_j n'est pas dans $\mathcal{G}_{ind}$). Ainsi, chaque crochet est de taille $\|l_j - u_j\|_{1,P} = F(t_j-) - F(t_{j-1}) \leq \epsilon$ dans $L_1(P)$ et il est clair que toute fonction $g \in \mathcal{G}_{ind}$ est au moins dans un de ces crochets.
Par conséquent, il vient que

$$N_{[]}(\epsilon, \mathcal{G}_{ind}, L_1(P)) \leq k_\epsilon < +\infty\,.$$

Ceci pour tout $\epsilon > 0$, donc la classe $\mathcal{G}_{ind}$ satisfait bien l'hypothèse du théorème précédent.

En outre, pour établir un résultat similaire pour les classes de Donsker, on a besoin d'une condition un peu plus forte concernant l'entropie à crochets. En effet, le nombre de crochets $N_{[]}(\epsilon, \mathcal{G}, L_r(P))$ de beaucoup de classes de fonctions tend vers l'infini quand ϵ tend vers 0. Naturellement, ce qui importe est que la "vitesse" à laquelle le nombre de crochets tend vers l'infini soit "raisonnable".
On quantifie cela grâce à l'**intégrale à crochets** définie par

$$J_{[]}(\delta, \mathcal{G}, L_r(P)) = \int_0^\delta \sqrt{\log N_{[]}(\epsilon, \mathcal{G}, L_r(P))}\, d\epsilon\,.$$

On a le théorème suivant.

Théorème 8.5.2. (Donsker) *Soit $\mathcal{G}$ une classe de fonctions mesurables et P une mesure sur $\mathcal{W}$. Si $J_{[]}(+\infty, \mathcal{G}, L_2(P)) < +\infty$, alors $\mathcal{G}$ est P-Donsker.*

Vérifions que la classe des indicatrices $\mathcal{G}_{ind}$ satisfait l'hypothèse de ce théorème.
Le premier travail est de calculer (borner) le nombre de crochets $N_{[]}(\epsilon, \mathcal{G}_{ind}, L_2(P))$. En reprenant les notations précédentes, remarquons que

$$\|l_j - u_j\|_{2,P} = \left(\|l_j - u_j\|_{1,P}\right)^{1/2}\,.$$

Puisque $\|l_j - u_j\|_{1,P} \leq \epsilon$, on a $\|l_j - u_j\|_{2,P} \leq \epsilon^{1/2}$, et cela pour tout $1 \leq j \leq k_\epsilon$. Ainsi, un ϵ-crochet dans $L_1(P)$ est un $\sqrt{\epsilon}$-crochet dans $L_2(P)$, ce qui implique

$$N_{[]}(\sqrt{\epsilon}, \mathcal{G}_{ind}, L_2(P)) \leq N_{[]}(\epsilon, \mathcal{G}_{ind}, L_1(P)) \leq k_\epsilon\,.$$

Comme $k_\epsilon \leq \frac{1}{\epsilon}$, après changement de variable $\epsilon' = \sqrt{\epsilon}$ on obtient finalement que pour tout $\epsilon > 0$, $N_{[]}(\epsilon, \mathcal{G}_{ind}, L_2(P)) \leq \frac{1}{\epsilon^2}$. De plus, si $\epsilon > 1$ alors le nombre de ϵ-crochets dans $L_2(P)$ nécessaires pour recouvrir $\mathcal{G}_{ind}$ vaut 1. Donc $J_{[]}(+\infty, \mathcal{G}_{ind}, L_2(P)) = J_{[]}(1, \mathcal{G}_{ind}, L_2(P))$.
On en déduit que

$$J_{[]}(1, \mathcal{G}_{ind}, L_2(P)) \leq \sqrt{2} \int_0^1 \sqrt{\log\left(\frac{1}{\epsilon}\right)}\, d\epsilon\,.$$

En posant le changement de variable $u = \log\left(\frac{1}{\epsilon}\right)$, le membre de droite vaut $\sqrt{\frac{\pi}{2}}$ donc la classe $\mathcal{G}_{ind}$ satisfait bien le théorème de Donsker précédent.
Les processus empiriques trouvent des applications dans un cadre bien plus général.

8.6 Applications statistiques

Après avoir défini (assez succinctement) le processus empirique et quelques unes de ses caractéristiques, nous illustrons comment il intervient de manière assez naturelle.

Tests statistiques.
Considérons les tests basés sur les fonctions de répartition. Le théorème de Glivenko-Cantelli justifie la statistique de test $\|\mathbb{F}_n - F\|_\infty = \sup_{t\in\mathbb{R}} |\mathbb{F}_n(t) - F(t)|$, appelée statistique de Kolmogorov, et garantit que la zone de rejet $\{\|\mathbb{F}_n - F\|_\infty \geq c\}$ est bien consistante. Par ailleurs, le théorème de Donsker, qui assure la convergence vers un pont brownien, nous permet de calculer une approximation des quantiles de la loi de $\|\mathbb{F}_n - F\|_\infty$ afin d'établir le *niveau* du test.

Propriétés des estimateurs.
En reprenant les notation données dans le Chapitre 2, considérons l'estimateur du maximum de vraisemblance ($\Psi_{\log}$-estimateur)

$$\widehat{\boldsymbol{\theta}}_{\Psi_{log}} = \operatorname*{Argmin}_{\boldsymbol{\theta}\in\Theta} \widehat{\mathcal{R}}_{\Psi_{log}}(\boldsymbol{\theta})\,, \quad \widehat{\mathcal{R}}_{\Psi_{log}}(\boldsymbol{\theta}) = -\frac{1}{n}\sum_{i=1}^{n} \log(\rho_{\mathcal{F}}(\boldsymbol{\theta}))(Y_i)\,,$$

où $\{\rho_{\mathcal{F}}(\boldsymbol{\theta}),\, \boldsymbol{\theta}\in\Theta\}$ est un ensemble de densités. Puis notons la cible ($\Psi_{\log}$-minimiseur)

$$\boldsymbol{\theta}_{\Psi_{log}} = \operatorname*{Argmin}_{\boldsymbol{\theta}\in\Theta} \mathcal{R}_{\Psi_{log}}(\boldsymbol{\theta})\,, \quad \mathcal{R}_{\Psi_{log}}(\boldsymbol{\theta}) = -\mathbb{E}_Y\left(\log(\rho_{\mathcal{F}}(\boldsymbol{\theta}))(Y)\right)\,.$$

Sous des conditions raisonnables, le comportement de la différence $\widehat{\boldsymbol{\theta}}_{\Psi_{log}} - \boldsymbol{\theta}_{\Psi_{log}}$ est donné par le comportement de la différence $\widehat{\mathcal{R}}_{\Psi_{log}}(\boldsymbol{\theta}) - \mathcal{R}_{\Psi_{log}}(\boldsymbol{\theta})$ lorsque θ parcourt Θ. Par exemple, si $\boldsymbol{\theta} \mapsto \mathcal{R}_{\Psi_{log}}(\boldsymbol{\theta})$ est de classe $\mathcal{C}^2$ avec une Hessienne définie positive, on montre qu'il existe une constante $C > 0$ tel que

$$d^2(\widehat{\boldsymbol{\theta}}_{\Psi_{log}}, \boldsymbol{\theta}_{\Psi_{log}}) \leq C \sup_{\boldsymbol{\theta}\in\Theta} |\widehat{\mathcal{R}}_{\Psi_{log}}(\boldsymbol{\theta}) - \mathcal{R}_{\Psi_{log}}(\boldsymbol{\theta})|\,, \tag{8.3}$$

où $d(\cdot,\cdot)$ est une certaine distance sur $\mathbb{R}^k$. Notons que cette dernière différence s'écrit

$$\widehat{\mathcal{R}}_{\Psi_{log}}(\boldsymbol{\theta}) - \mathcal{R}_{\Psi_{log}}(\boldsymbol{\theta}) = \frac{1}{\sqrt{n}}\,\mathbb{G}_n(g_{\boldsymbol{\theta}})\,,$$

où $\mathbb{G}_n$ est le processus empirique et $g_{\boldsymbol{\theta}}$ appartient à la classe de fonctions $\mathcal{G} = \{g_{\boldsymbol{\theta}} = -\log(\rho_{\mathcal{F}}(\boldsymbol{\theta}))\,,\, \boldsymbol{\theta}\in\Theta\}$. Par conséquent, si $\mathcal{G}$ est de Glivenko-Cantelli, c'est à dire

$$\frac{1}{\sqrt{n}}\,\|\mathbb{G}_n\|_{\mathcal{G}} = \sup_{\boldsymbol{\theta}\in\Theta} |\widehat{\mathcal{R}}_{\Psi_{log}}(\boldsymbol{\theta}) - \mathcal{R}_{\Psi_{log}}(\boldsymbol{\theta})| \longrightarrow 0\,,$$

alors (d'après (8.3)) on peut envisager que $\widehat{\boldsymbol{\theta}}_{\Psi_{log}}$ converge vers $\boldsymbol{\theta}_{\Psi_{log}}$.
Avec le même raisonnement, si $\mathcal{G}$ est Donsker, on peut espérer montrer une convergence en loi et exhiber une vitesse de convergence.

Bibliographie

[1] M.D. Donsker. Justification and extension of Doob's heuristic approach to the Kolmogorov-Smirnov theorems. *The Annals of mathematical statistics*, pages 277–281, 1952.

[2] R.M. Dudley. Weak convergence of measures on nonseparable metric spaces and empirical measures on euclidian spaces. *Illinois Journal of Mathematics*, 11 :109–126, 1966.

[3] P. Gaenssler. *Empirical Processes*. Institute of Mathematical Statistics, Hayward, CA, 1983.

[4] M.R. Kosorok. *Introduction to empirical processes and semiparametric inference*. Springer series in statistics, 2008.

[5] D. Pollard. Empirical processes : theory and applications. *Regional Conference Series in Probability and Statistics Hayward*, 1990.

[6] G.R Shorack and J.A Wellner. *Empirical processes with applications to statistics*. Wiley Series in Probability and Statistics, 1986.

[7] A.W. van der Vaart and J.A. Wellner. *Weak Convergence and Empirical Processes*. Springer Series in Statistics, 1996.

Bibliographie Générale

S. Arlot and P. Massart. Data-driven calibration of penalties for least-squares regression. *The Journal of Machine Learning Research*, 10 :245–279, 2009.

P. Barbillon, G. Celeux, A. Grimaud, Y. Lefebvre, and E. De Rocquigny. Nonlinear methods for inverse statistical problems. *Computational Statistics & Data Analysis*, 55(1) :132–142, 2011.

P.L. Bartlett, M.I. Jordan, and J.D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473) :138–156, 2006.

M. Benaïm. Dynamics of stochastic approximation algorithms. *Seminaire de probabilites XXXIII*, pages 1–68, 1999.

A.P. Berens. Nde reliability data analysis. *ASM Handbook.*, 17 :689–701, 1989.

A. Beygelzimer, S. Dasgupta, and J. Langford. Importance weighted active learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 49–56. ACM, 2009.

P. Billingsley. *Convergence of probability measures*. Wiley New York, 1968.

S. Boucheron, O. Bousquet, and G. Lugosi. (chapter) concentration inequalities. *Machine Learning Summer School 2003*, 3176 :169–207, 2004.

O. Bousquet, S. Boucheron, and G. Lugosi. (chapter) introduction to statistical learning theory. *Machine Learning Summer School 2003*, 3176 :208–240, 2004.

W. Cauer. Die verwirklichung von wechselstromwiderständen vorgeschriebener frequenzabhängigkeit. *Electrical Engineering (Archiv fur Elektrotechnik)*, 17(4) :355–388, 1926.

O. Chapelle, B. Schölkopf, and A. Zien. *Semi-supervised learning*, volume 2. MIT press Cambridge, MA, 2006.

S. Dasgupta and J. Langford. Active learning tutorial, icml 2009.

E. de Rocquigny, N. Devictor, and S. Tarantola, editors. *Uncertainty in industrial practice*. John Wiley.

D.L. Donoho and J.M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3) :425–455, 1994.

M.D. Donsker. Justification and extension of Doob's heuristic approach to the Kolmogorov-Smirnov theorems. *The Annals of mathematical statistics*, pages 277–281, 1952.

R.M. Dudley. Weak convergence of measures on nonseparable metric spaces and empirical measures on euclidian spaces. *Illinois Journal of Mathematics*, 11 :109–126, 1966.

M. Duflo. *Algorithmes stochastiques*. Springer, 1996.

M. Duflo. *Random iterative models*, volume 34. Springer Verlag, 1997.

J.C. Fort and G. Pages. Asymptotic behavior of a markovian stochastic algorithm with constant step. *SIAM journal on control and optimization*, 37 :1456, 1999.

J.C. Fort and G. Pagès. Decreasing step stochastic algorithms : As behaviour of weighted empirical measures. *Monte Carlo Methods and Applications*, 8(3) :237–270, 2002.

P. Gaenssler. *Empirical Processes*. Institute of Mathematical Statistics, Hayward, CA, 1983.

A. Goldenshluger and O. Lepski. Uniform bounds for norms of sums of independent random functions. *Arxiv preprint arXiv :0904.1950*, 2009.

C.A. Harding, G.R. Hugo, and S.J. Bowles. Application of model assisted pod using a transfer function approach. *Review of Quantitative Nondestructive Evaluation*, 28, 2009.

T. Hastie, R. Tibshirani, and J.H. Friedman. *The elements of statistical learning : data mining, inference, and prediction*. Springer Verlag, 2009.

P.J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, pages 73–101, 1964.

P.J. Huber. *Robust statistics*. Wiley-Interscience, 1981.

F. Jenson, E. Iakovleva, and N. Dominguez. Simulated supported pod : methodology and hfet validation case. *Review of Quantitative Nondestructive Evaluation*, 30, 2010.

M.C. Kennedy and A. O'Hagan. Bayesian calibration of computer models. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 63(3) :425–464, 2001.

J.P.C. Kleijnen. *Design and analysis of simulation experiments*. Springer Verlag, 2007.

T. Klein and E. Rio. Concentration around the mean for maxima of empirical processes. *Annals of probability*, 33(3) :1060–1077, 2005.

M.R. Kosorok. *Introduction to empirical processes and semiparametric inference*. Springer series in statistics, 2008.

M. Ledoux. *The concentration of measure phenomenon*. AMS, 2001.

P. Massart. *Concentration inequalities and model selection : Ecole d'Eté de Probabilités de Saint-Flour XXXIII-2003*. Springer Verlag, 2007.

P. Massart and É. Nédélec. Risk bounds for statistical learning. *The Annals of Statistics*, 34(5) :2326–2366, 2006.

J.J. Moré and Z. Wu. Global continuation for distance geometry problems. *Preprint Mcs-p, SIAM J. Optimization*, 7(7) :814–836, 1995.

W.G. Müller. *Collecting spatial data : optimum design of experiments for random fields*. Springer Verlag, 2007.

D. Pollard. Empirical processes : theory and applications. *Regional Conference Series in Probability and Statistics Hayward*, 1990.

L. Pronzato and A. Zhigljavsky. *Optimal design and related areas in optimization and statistics*. Springer Verlag, 2008.

N. Rachdi, J.C. Fort, and T. Klein. Risk bounds for new M-estimation problems . *hal 00537236*, 2010.

H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, pages 400–407, 1951.

T.J. Santner, B.J. Williams, and W. Notz. *The design and analysis of computer experiments*. Springer Verlag, 2003.

A. Saumard. (phd thesis) estimation par minimum de contraste régulier et heuristique de pente en sélection de modèles. 2010.

G.R Shorack and J.A Wellner. *Empirical processes with applications to statistics*. Wiley Series in Probability and Statistics, 1986.

B.W. Silverman. *Density estimation for statistics and data analysis*. Chapman & Hall/CRC, 1986.

C. Soize and R. Ghanem. Physical systems with random uncertainties : chaos representations with arbitrary probability measure. *SIAM Journal on Scientific Computing*, 26 :395–410, 2004.

M. Talagrand. Sharper bounds for Gaussian and empirical processes. *The Annals of Probability*, 22(1) :28–76, 1994.

A.N. Tikhonov. *On the stability of inverse problems*, volume 39. 1943.

S. van de Geer. *Empirical processes in M-estimation*. Cambridge University Press, 2000.

A.W. van der Vaart. *Asymptotic statistics*. Cambridge University Press, 2000.

A.W. van der Vaart and J.A. Wellner. *Weak Convergence and Empirical Processes*. Springer Series in Statistics, 1996.

V. Vapnik. *Statistical learning theory*. Wiley New York, 1998.

V. Vapnik. *The nature of statistical learning theory*. Springer Verlag, 2000.

E. Vazquez. (PhD thesis) Modélisation comportementale de systèmes non-linéaires multivariables par méthodes à noyaux et applications. 2005.

N. Wiener. The homogeneous chaos. *American Journal of Mathematics*, 60(4) :897–936, 1938.

A. Zhigljavsky and A. Zilinskas. *Stochastic global optimization*. Springer New York, 2007.

APPRENTISSAGE STATISTIQUE ET COMPUTER EXPERIMENTS

Résumé – *Cette thèse s'inscrit dans le domaine de l'apprentissage statistique et dans celui des expériences simulées (computer experiments). Son objet est de proposer un cadre général permettant d'estimer les paramètres d'un code de simulation numérique de façon à reproduire au mieux certaines caractéristiques d'intérêt extraites de données observées. Ce travail recouvre le cadre classique de l'estimation paramétrique dans un modèle de regression et également la calibration de la densité de probabilité des variables d'entrée d'un code numérique afin de reproduire une loi de probabilité donnée en sortie.*

Une partie importante de ce travail consiste dans l'estimation paramétrique d'un code numérique à partir d'observations. Nous proposons une classe de méthode originale nécessitant une simulation intensive du code numérique, que l'on remplacera par un méta-modèle s'il est trop coûteux. Nous validons théoriquement les algorithmes proposés du point de vue non-asymptotique, en prouvant des bornes sur l'excès de risque. Ces résultats reposent entres autres sur des inégalités de concentration.

Un second problème que nous abordons est celui de l'étude d'une dualité entre procédure d'estimation et nature de la prédiction recherchée. Il s'agit ici de mieux comprendre l'effet d'une procédure d'estimation des paramètres d'un code numérique sur une caractéristique d'intérêt donnée.

Enfin, en pratique la détermination des paramètres optimaux au sens du critère donné par le risque empirique nécessite la recherche du minimum d'une fonction généralement non convexe et possédant plusieurs minima locaux. Nous proposons un algorithme stochastique consistant à combiner une régularisation du critère par convolution avec un noyau gaussien, de variance décroissante au fil des itérations, avec une méthode d'approximation stochastique du type Kiefer-Wolfowitz.

STATISTICAL LEARNING AND COMPUTER EXPERIMENTS

Abstract – *This thesis work consists in gathering statistical learning theory with the field of computer experiments. As the considered computer codes are only known through simulations, we propose an original statistical framework, including the classical ones, which takes into account the simulation aspect.*

We investigate learning algorithms for parameter estimation in computer codes which depend on both observed and simulation data. We validate these algorithms by proving excess risk bounds using concentration inequalities.

We also study the duality between the estimation procedure and the wanted feature prediction. Here, we try to understand the impact of an estimation procedure on some given characteristic of the phenomenon of interest.

Finally, the computation of optimal parameters in practice involves the minimization of a criterion which is generally highly non convex and with irregularities. We propose a stochastic algorithm which consists in combining regularization methods with a stochastic approximation method like the Kiefer-Wolfowitz one.