UNIVERSITÉ DE NICE-SOPHIA ANTIPOLIS

ÉCOLE DOCTORALE STIC
SCIENCES ET TECHNOLOGIES DE L'INFORMATION ET DE LA COMMUNICATION

THÈSE

pour obtenir le titre de

Docteur en Sciences

de l'Université de Nice-Sophia Antipolis

Mention : Automatique, Traitement du Signal et des Images

présentée et soutenue par

Régis BETTINGER

INVERSION D'UN SYSTÈME PAR KRIGEAGE

Application à la synthèse de catalyseurs à haut débit

Thèse dirigée par Luc PRONZATO

soutenue le 22 octobre 2009

Jury :

M. Pascal DUCHÊNE	Ingénieur de Recherche	Tuteur IFP
M. Fabrice GAMBOA	Professeur	Rapporteur
M. Werner MÜLLER	Professeur	
M. Luc PRONZATO	Directeur de Recherche CNRS	Directeur de thèse
M. Emmanuel VAZQUEZ	Professeur Assistant	
M. Henry WYNN	Professeur	Rapporteur

Remerciements

Je souhaite tout d'abord exprimer ma gratitude à Pascal Duchêne, pour m'avoir engagé en stage à l'IFP auquel a succédé cette thèse. Merci de m'avoir appris le pragmatisme nécessaire à la recherche appliquée combiné à la rigueur scientifique, dans une ambiance de travail à la fois sérieuse et amicale, et de m'avoir fait profiter de tes grandes compétences mathématiques et informatiques. Merci également au personnel de l'IFP pour son accueil chaleureux.

Un grand merci à Luc Pronzato pour avoir mis ses idées géniales et sa gentillesse au service de ce travail de thèse. Malgré la distance, merci d'avoir été là à chaque fois que j'en avais besoin tout au long de ces années de recherche, et pour les bons moments passés aux congrès et à Antibes.

Merci à Éric Thierry pour sa participation amicale à des réunions de travail. Sa connaissance encyclopédique des références de la littérature et sa gentillesse ont été une source d'inspiration très précieuse dans ce travail de recherche.

Je tiens à adresser mes remerciements aux membres du jury qui ont accepté de juger ce travail. Merci aux rapporteurs, à Fabrice Gamboa tout d'abord pour ses remarques très détaillées qui ont grandement contribué à améliorer la qualité du manuscrit, et à Henry Wynn pour leurs très bons commentaires. Merci également à Werner Müller et à Emmanuel Vazquez pour faire partie de ce jury.

Merci à Aurélie, Ana Rita, JP, Matthias, Luc, Olivier, ainsi que l'ensemble du club de Villeurbanne pour avoir été là depuis le début !

Aux collègues de bureau successifs et amis pour avoir rendu le quotidien agréable, un grand merci à Damien, Sébastien et Eugênio.

Une pensée amicale pour Lucile, Laurent, Rachid, Luc, Olivier, Matthieu, Aimad, Mahmoud et, *last but not least*, François !

Une pensée très spéciale aux grands amis de Paris répondant toujours présent, un grand merci à Véronique, Lisandro, Titou, Guillaume, Igor et Fabien. Merci aussi à Hoël, et aux deux Didier pour leur aide.

Skoll! à Signe, Leona, Lena et François, et bon vent, et merci à Guillaume pour les bons conseils avant la soutenance. Merci à l'ensemble du personnel du laboratoire I3S de Sophia Antipolis pour son accueil, sa gentillesse et sa bonne humeur.

Finalement, merci à ma famille à qui j'oublie trop souvent de dire combien elle compte pour moi. Ce travail de thèse lui est dédié.

Résumé

Ce travail concerne la modélisation du processus de synthèse (construction) de supports de catalyseurs obtenus par réaction silice-alumine. La synthèse est caractérisée par 5 variables d'entrée et 2 variables de sortie (la surface spécifique et le volume mésoporeux du support de catalyseur). Chaque combinaison des valeurs de sortie ayant une application potentielle, on voudrait savoir en synthétiser le plus grand nombre, c'est-à-dire connaître les variables d'entrée permettant de construire un catalyseur ayant une surface et un volume donnés quelconques. Les limites atteignables des deux sorties du système sont inconnues.

Ne disposant pas de suffisamment d'essais pour pouvoir espérer construire un modèle fiable sur l'ensemble du domaine de variation des variables d'entrée, nous choisissons une approche par plans d'expérience séquentiels avec modélisation par krigeage, permettant d'éviter une trop grande dispersion des variables d'entrée tout en assurant une exploration du domaine accessible pour les variables de sortie. Les essais sont choisis séquentiellement en se servant de l'information apportée par les essais précédents et traitée par le modèle de krigeage. Cette façon de procéder est *a priori* plus efficace que celle consistant à utiliser un plan d'expériences fixé au départ et comprenant la totalité des essais disponibles.

Des critères d'ajout séquentiel de points d'expérimentation (définissant les valeurs des variables d'entrée) sont proposés, qui favorisent une forte dispersion des sorties correspondantes et prennent en compte les incertitudes associées aux prédictions par krigeage. Enfin, les critères retenus, l'un à base de distance et l'autre à base d'entropie, sont testés sur des données simulées afin de vérifier la bonne répartition finale des valeurs des réponses.

Des rappels sur la modélisation par processus gaussien, la régression/interpolation par krigeage et ses liens avec les méthodes de type splines et SVM, ainsi que la planification d'expériences sont présentés en essayant de concilier rigueur et clarté.

Mots-clés : planification d'expériences, krigeage, entropie de Tsallis.

Abstract

This work deals with the modeling of the synthesis process for catalyst supports obtained by a chemical reaction involving silica and alumina. The process is characterized by 5 inputs and 2 outputs (specific surface and mesoporous volume of the support). Each pair of output values has a potential application and the ultimate objective is to be able to find input values associated with the synthesis of a catalyst having any given output characteristics (surface, volume). The ranges of the two outputs are unknown.

The quantity of runs available is too small to build a satisfactory model over the whole input domain. We thus combine design of experiments and kriging modeling in a way that ensures both a limited dispersion of the input factors and a good exploration of the reachable output domain. The runs are designed sequentially, using the information provided by former runs through their associated kriging model. This sequential construction seems more efficient than the design of a non-sequential experiment containing the total amount of available runs.

Several criteria are proposed for sequential design which favor a high dispersion of the corresponding outputs and take the uncertainties associated with the kriging model into account. The two most appealing are tested on simulated data in order to check the dispersion of outputs; one is based on minimax distance and the other on entropy.

Basic properties of Gaussian processes, regression/interpolation by kriging and links with other methods such as splines and SVMs are reminded, together with standard methods for designing experiments, with the objective of combining rigor and clarity.

Keywords: design of experiments, kriging, Tsallis entropy.

Table des matières

Introduction **1**

1 Méthodes à noyaux **5**
1.1 Cadre général de la théorie de l'apprentissage statistique 5
1.1.1 Un exemple classique 5
1.1.2 Le compromis biais-variance 6
1.1.3 Régression régularisée 7
1.2 Apprentissage par méthodes à noyaux 8
1.2.1 Rappels sur les espaces de Hilbert 8
1.2.2 Espaces de Hilbert à noyau reproduisant 9
1.2.3 RKHS et apprentissage 12
1.3 RKHS construit à partir d'un noyau conditionnellement semi-défini positif 17
1.3.1 Fonctions conditionnellement semi-définies positives 17
1.3.2 Construction de l'espace d'hypothèses 20
1.3.3 Apprentissage 22
1.4 Exemples de méthodes à noyaux 25
1.4.1 Krigeage 25
1.4.2 Splines plaque mince 27
1.4.3 Ondelettes 30
1.4.4 Support Vector Machines 34

2 Krigeage **37**
2.1 Processus aléatoires gaussiens 37
2.1.1 Généralités 37
2.1.2 Moyenne et covariance 40
2.2 Krigeage 49
2.2.1 Prédiction 49
2.2.2 Modèle 50
2.2.3 Estimation des paramètres 55
2.3 Problèmes liés à l'estimation des paramètres 59
2.3.1 Identifiabilité, consistance, efficacité d'un estimateur 59
2.3.2 Propriétés asymptotiques de l'estimateur du maximum de vraisemblance . 65
2.3.3 Conséquences d'une mauvaise estimation des paramètres 68
2.4 Krigeage avec prise en compte d'erreur de mesure 71
2.4.1 Modèle de krigeage avec inclusion de bruit 71
2.4.2 Réinterpolation 74

3 Plans d'expériences **77**
3.1 Plans remplissant l'espace . . . 77
3.1.1 Hypercubes latins, tableaux orthogonaux . . . 78
3.1.2 Plans uniformes . . . 80
3.1.3 Plans construits à partir d'un critère de distance . . . 81
3.2 Méthodes paramétriques de construction de plans d'expériences . . . 84
3.2.1 Modèle de régression à erreurs indépendantes . . . 84
3.2.2 Plans d'expériences avec modèle de krigeage . . . 85
3.3 Optimisation de la réponse avec krigeage . . . 87
3.3.1 Utilisation des bornes de confiance . . . 88
3.3.2 Espérance de gain . . . 89
3.3.3 Utilisation de l'entropie . . . 90

4 Maximisation de la diversité des réponses **92**
4.1 Diversité dans $\mathbb{R}$ . . . 93
4.1.1 Étude de la fonction discrépance . . . 94
4.1.2 Étude d'une fonction de type maximin . . . 99
4.1.3 Étude d'une fonction de type minimax . . . 100
4.1.4 Fonctions utilisant l'entropie de Shannon . . . 102
4.1.5 Entropie de Rényi . . . 112
4.1.6 Entropie de Tsallis . . . 113
4.1.7 Conclusion à l'étude de la diversité dans le cas monodimensionnel . . . 115
4.2 Diversité en dimension 2 (ou plus) . . . 115
4.2.1 Fonction de type maximin . . . 116
4.2.2 Fonction de diversité de Tsallis multidimensionnelle . . . 117
4.2.3 Prise en compte des contraintes pratiques dans les critères de diversité . . . 120

5 Mise en œuvre pratique et résultats **126**
5.1 Structure de l'algorithme . . . 126
5.2 Cas où la fonction inconnue a 2 entrées et 1 sortie . . . 127
5.2.1 Choix des fonctions-test et des plans initiaux . . . 127
5.2.2 Tests . . . 132
5.2.3 Inversion du système . . . 147
5.3 Cas où la fonction inconnue a 5 entrées et 2 sorties . . . 153
5.3.1 Fonction-test et plan initial . . . 153
5.3.2 Tests . . . 158

Conclusion **174**

A Ergodicité **177**

B Régularité en moyenne quadratique **179**

C Krigeage bayésien **182**
C.1 Formalisme bayésien . . . 182
C.2 Application au krigeage . . . 183

D Cokrigeage **186**
D.1 Modèle de cokrigeage 186
D.2 Calcul du prédicteur de cokrigeage 187

E Krigeage intrinsèque **190**
E.1 Processus aléatoires intrinsèques (d'ordre 0) 190
E.2 Processus intrinsèques d'ordre q $(q \in \mathbb{N})$ 192
E.2.1 Définition et propriétés 192
E.2.2 Covariance généralisée 193
E.2.3 Représentation spectrale 194
E.3 Prédicteur de krigeage intrinsèque 195
E.3.1 Calcul de l'erreur quadratique moyenne 195
E.3.2 Équations du krigeage intrinsèque 195

F Premiers moments (tronqués) de la loi normale mono-dimensionnelle **197**

G Formules de discrépance **198**
G.1 Discrépance L^∞ 198
G.1.1 Démonstration de la proposition 4.1.1 199
G.1.2 Preuve de la continuité de $D_\infty^n(\cdot)$ 200
G.2 Discrépance L^1 201
G.3 Discrépance L^2 202

H Distance de Wasserstein **203**
H.1 Définition et propriétés 203
H.2 Critère d'ajout de point utilisant la distance de Wasserstein 203

I Une distance pour mesurer l'éloignement de deux plans **205**

J Triangles de Delaunay et cellules de Voronoi **207**
J.1 Triangles de Delaunay 207
J.2 Cellules de Voronoi 209

Bibliographie **211**

Introduction et motivation

Contexte

L'IFP développe des procédés de raffinage utilisant des catalyseurs solides. Ces catalyseurs sont composés d'un support sur lequel sont déposées des substances actives (métaux) qui permettent d'accélérer une réaction chimique. Selon son utilisation, un support peut exister principalement sous la forme de sphère ou de cylindre dont la taille varie de l'ordre du millimètre au centimètre.

Les catalyseurs sont poreux pour que la surface de contact avec le fluide réactif (liquide ou gaz) soit aussi grande que possible (50 à 500 m^2/g sont des valeurs courantes). On s'intéresse ici aux supports silice-alumine avant leur mise en forme. Le support est synthétisé (construit) par précipitation en faisant réagir une solution de silice et une solution d'alumine.

Un support de catalyseur est caractérisé par deux grandeurs importantes :

- sa surface spécifique (notée S_{bet}), la surface (par gramme) du support qui sera en contact avec le fluide réactif ;
- son volume mésoporeux (noté V_{p}), qui est le volume (par gramme) formé par les « trous » du support.

Le problème posé dans la synthèse de supports de catalyseurs est d'obtenir des supports ayant une porosité et une surface spécifique bien définies. La valeur précise de chacune de ces propriétés dépend de l'application qui est faite du support. Chaque combinaison $(S_{\text{bet}}, V_{\text{p}})$ ayant une application potentielle, il est souhaitable de savoir en synthétiser le plus grand nombre possible. Notons que ces deux grandeurs ne sont généralement pas reliées entre elles : par exemple, il a été observé qu'une multitude de petites cavités dans le support peut donner lieu à une grande surface spécifique, alors que le volume mésoporeux est petit.

La synthèse de support est caractérisée par 5 variables d'entrée et 2 variables de sortie (S_{bet} et V_{p}). Les 5 variables d'entrée ayant une influence sur la réaction chimique produisant le support de catalyseur, ainsi que leur domaine de variation ont été déterminés par des essais préliminaires. Ce sont :

- le pH, maintenu constant (régulé) au cours de la réaction ;
- la durée d'ajout des réactifs ;
- la température (régulée pendant toute la durée de la réaction) ;
- la concentration, notée Si+Al, de silice et d'alumine à la fin de la réaction (ou, à un facteur près, le nombre de moles de réactifs ajoutés au cours de la réaction) ;
- le rapport Si/Al du nombre de moles de silice par le nombre de moles d'alumine ajoutés lors de la synthèse.

Le pH est maintenu constant au cours de la réaction par ajout de solutions d'acide et de base. La température est maintenue constante par circulation d'eau dans une double enveloppe.

La synthèse d'un support de catalyseur est un processus complexe généralement mal connu : il n'existe pas actuellement de modèle complet permettant de prédire les propriétés du catalyseur

à partir des conditions de la synthèse. Les méthodes traditionnelles permettent de construire quelques dizaines de catalyseurs par an. Pour accélérer ce processus, l'IFP met en place des outils d'expérimentation à haut débit (EHD) qui permettront de synthétiser plusieurs centaines de catalyseurs par an (de cinq cent à mille). Le gain en débit est obtenu par une automatisation partielle de la procédure et par la réalisation de plusieurs synthèses en parallèle. Cette augmentation du débit permet d'élargir le domaine des conditions de synthèse habituellement considéré. Elle n'est toutefois pas suffisante pour en faire une exploration systématique en raison du nombre élevé de mesures que cela impliquerait. Il est donc souhaitable d'associer aux nouveaux outils de construction des techniques mathématiques d'aide au choix des conditions expérimentales.

Dans la littérature consacrée à l'EHD, les méthodes employées pour la synthèse de supports ou de matériaux relèvent soit des plans d'expériences, soit de l'optimisation stochastique. Plusieurs articles écrits par des chercheurs du domaine sont regroupés dans [24]. Deux grandes familles de méthodes y sont proposées : plans d'expériences et algorithmes d'optimisation stochastique.

- Les applications proposées des plans d'expériences consistent généralement à couvrir l'espace d'étude de façon aussi uniforme que possible, puis à raffiner les zones intéressantes. Une telle approche n'est toutefois viable que pour des espaces de petite dimension, sauf à disposer d'un nombre d'essais très grand (les exemples illustrant ces méthodes sont de dimension au plus quatre). Les plans d'expériences séquentiels associés à du krigeage sont également cités, ils seront présentés au chapitre 3 avec l'optimisation statistique.
- L'autre grande famille de méthodes proposées est constituée des algorithmes d'optimisation stochastique, comme les algorithmes génétiques et les méthodes de Monte Carlo.
  - Concernant les algorithmes génétiques, seuls des résultats de simulation sont présentés. Les auteurs insistent sur la nécessité de configurer correctement ces algorithmes à l'aide d'essais préalables (codage des variables et choix des paramètres), sans quoi le meilleur catalyseur obtenu peut être loin de l'optimum. Ils ont été testés à l'IFP pour la synthèse de supports de catalyseurs [168]. Les tests ont également été réalisés en simulation. Ils confirment la nécessité de régler correctement les algorithmes, et montrent la variabilité des résultats obtenus (en raison du caractère stochastique de ces algorithmes) : un réglage donné de l'algorithme peut conduire à de plus ou moins bons résultats.
  - Une approche par Monte Carlo est également proposée. Des modifications de l'algorithme de Metropolis standard sont proposées pour diminuer le nombre d'essais nécessaires. Mais ce dernier, de l'ordre de cent mille dans les exemples présentés, reste trop important pour que ces algorithmes puissent être appliqués à l'IFP dans le contexte actuel.

L'objectif de l'étude à la base de ce travail de thèse consiste à pouvoir être capable d'associer à toute valeur « cible » T définie par un couple (surface spécifique $S_{\text{bet}}^{\text{T}}$, volume poreux V_{p}^{T}) réalisable une combinaison x_{T} des facteurs d'entrée permettant cette réalisation (ou du moins telle que les deux réponses en x_{T} soient « proches » de leurs valeurs cibles T). En ce sens, il s'agit d'*inverser le système*. Le problème est donc notablement différent de celui consistant à construire un modèle précis sur l'ensemble du domaine d'étude : en effet, nous souhaitons que le modèle soit précis uniquement pour des valeurs x_{T} permettant de recouvrer l'ensemble des valeurs cibles T. La dimension de l'espace des facteurs (5) étant supérieure à celle de l'espace des sorties (2) on conçoit qu'il est possible (en théorie) de définir un sous-espace de l'espace des facteurs (une variété de dimension 2) qui soit en bijection par le système avec l'ensemble des valeurs cibles, sur lequel il suffirait d'observer. Ainsi, il est souhaitable que les observations ne soient pas trop dispersées dans l'espace des facteurs, mais concentrées en des zones qui permettent de synthétiser l'ensemble des valeurs cibles afin d'obtenir un modèle localement le plus précis possible dans cette zone. Le plan d'expériences doit alors poursuivre les deux objectifs suivants :

1. assurer que les réponses associées soient aussi dispersées que possible dans l'espace des sorties ;
2. assurer que les points demeurent concentrés dans l'espace des entrées.

Une façon de respecter ces deux objectifs consiste à construire un modèle, puis optimiser un critère qui tienne compte de la répartition des sorties d'une part, et d'autre part de l'incertitude sur les prédictions faites par le modèle quand les points du plan d'expérience sont éloignés les uns des autres. C'est cette construction qui est au cœur de ce travail de thèse.

Approche proposée

L'approche que nous proposons est séquentielle, l'ensemble des observations étant enrichi petit à petit. À chaque étape, un modèle est construit par krigeage à l'aide des points du plan courant (qui ont été échantillonnés et pour lesquels les sorties sont connues). Pour choisir le (ou les) nouveau(x) point(s), un critère quantifiant la « diversité » des sorties du nouveau plan (le plan précédent auquel est ajouté le ou les points candidats) est maximisé. Mais comme les valeurs des réponses au point candidat sont inconnues, le modèle de krigeage est utilisé, et l'incertitude de prédiction est prise en compte dans l'évaluation de la « diversité ». Cette façon de procéder est *a priori* plus efficace que de mettre au point dès le départ un plan d'expériences global et de fixer la totalité des essais expérimentaux sans connaissance plus fine du système.

Un algorithme d'optimisation sera présenté dans la suite, prenant en compte les caractéristiques de l'étude résumées ci-dessous.

- L'objectif est de constituer une bibliothèque de supports aussi divers que possible pour un système à 5 entrées et 2 sorties.
- Le budget de mesures disponibles est de l'ordre de 500 à 1000.
- Les essais sont réalisés par séries d'un nombre fixe (*a priori* 6). L'algorithme d'optimisation devra donc travailler par blocs.
- L'erreur de mesure inhérente à tout système physique devra être traitée par le modèle.
- Un retard dans la disponibilité des mesures devra être pris en compte : en effet, la mise en forme du support obtenu demandant du temps il a été décidé, afin d'accélérer le processus, de lancer chaque série d'essais avant d'avoir le résultat des mesures de la série immédiatement précédente.

Les tests des méthodes proposées sont réalisés en simulation à l'aide d'un modèle développé à cet effet. La méthode la plus prometteuse sera mise en œuvre sur l'outil expérimental.

Notons que la planification séquentielle d'expériences avec modélisation par krigeage a déjà été utilisée à l'IFP pour la modélisation de réservoirs pétroliers [148], ainsi qu'au CEA sur un code de calcul modélisant un comportement hydrogéologique [116].

Organisation du manuscrit

Le manuscrit est organisé de la façon suivante.

- Le chapitre 1 est une mise en perspective dans un cadre général des méthodes de modélisation parmi les plus couramment utilisées. Nous écrivons de façon détaillée l'équivalence, sous certaines conditions, des méthodes de splines, krigeage, SVM (Support Vector Machines) et ondelettes. Ces résultats d'équivalence sont bien connus mais n'ont, à notre connaissance, jamais été rassemblés et explicités aussi clairement dans la littérature.
- Nous présentons au chapitre 2 la théorie des processus gaussiens et du krigeage, qui sera utilisé pour modéliser le système. Des propriétés asymptotiques ou à échantillon fini sont

rappelées, qui permettent de se familiariser avec la méthode et de se faire une idée des précautions à prendre en pratique. L'importance du choix de la fonction de covariance apparaît régulièrement au cours de la lecture.

– Le chapitre 3 se veut une introduction très succinte aux plans d'expériences. Des plans remplissant l'espace utilisés au stade initial de la procédure d'ajout sont présentés, ainsi que des éléments de théorie des plans optimaux. Enfin, nous rappelons des critères de construction séquentielle de plans d'expériences permettant d'optimiser une fonction.
– Nous introduisons au chapitre 4 plusieurs façons de mesurer la « diversité » d'un ensemble de points, et examinons sur des cas simples si ces méthodes s'adaptent dans le cas séquentiel où l'on souhaite effectuer les mesures petit à petit. Nous retenons finalement deux critères d'ajout : l'un est basé sur une mesure de distance, l'autre utilise l'entropie de Tsallis. L'approche proposée prend en compte l'ensemble des contraintes du problème rappelées ci-dessus.
– Nous testons de manière plus approfondie au chapitre 5 les deux critères de « diversité » retenus. Nous détaillons la mise en œuvre de l'algorithme d'optimisation, et comparons les critères d'ajout dans le cas simple de 2 facteurs d'entrée et 1 sortie dans un premier temps, puis dans les mêmes dimensions que le problème réel (5 entrées et 2 sorties). Les résultats obtenus semblent montrer que la méthode est applicable au problème pratique. Des soucis techniques n'ont cependant pas permis d'effectuer des phases d'expérimentation sur le système réel pendant la durée de la thèse.

De nombreux points expliqués dans la suite sont longtemps restés obscurs pour moi avant de trouver les bonnes références bibliographiques. J'ai donc tenté, entre autre, d'écrire clairement les liens existant entre les méthodes de modélisation usuelles (splines, krigeage, SVM), de rappeler de façon concise les propriétés théoriques du krigeage, et de présenter le cokrigeage et le krigeage intrinséque de la même manière que le krigeage usuel afin que le rapport entre ces méthodes soit bien mis en relief. La partie théorique de ce document a ainsi été rédigée de façon à rassembler des informations éparses, et à donner les arguments qui expliquent des faits souvent donnés sans justification dans la littérature. En un mot, j'ai voulu écrire le document que j'aurais souhaité avoir entre les mains au moment de débuter cette thèse.

Chapitre 1

Méthodes à noyaux

Les méthodes à noyaux les plus communément utilisées sont présentées dans ce chapitre. Après une brève introduction à l'apprentissage statistique, nous faisons quelques rappels sur les espaces de Hilbert à noyau reproduisant et leur utilisation en apprentissage. Finalement, ce cadre permet d'observer les liens existant entre les méthodes exposées.

1.1 Cadre général de la théorie de l'apprentissage statistique

Le problème de l'*apprentissage* est de construire une fonction (appelée *machine*) qui établit une dépendance entre des quantités x (les entrées) et y (les sorties), à partir d'un nombre fini d'observations $\{(x_i, y_i)\}_{i=1\dots n}$ appelé *ensemble d'entraînement*. La fonction obtenue doit avoir de bonnes capacités de *généralisation*, c'est-à-dire qu'elle doit permettre de bien prédire les valeurs de y pour les x non observés. Les applications potentielles de telles machines sont nombreuses [128, 177].

1.1.1 Un exemple classique

Considérons le cas très classique [55] où les sorties (scalaires) sont supposées reliées aux entrées par la relation

$$y = m(x) + \varepsilon,$$

avec $x \in \mathcal{X} \subset \mathbb{R}^d$, $m : \mathcal{X} \to \mathbb{R}$ une fonction déterministe inconnue et ε une variable aléatoire centrée de carré intégrable, indépendante de x. On suppose que x est déterministe et peut être choisi comme on l'entend. L'ensemble d'entraînement $\mathcal{D} = \{(x_1, y_1), \dots, (x_n, y_n)\}$ étant donné, et pour un x fixé, on recherche un prédicteur de y noté $\widehat{f}_n(x)$. Une façon naturelle de mesurer la qualité du prédicteur en x est donnée par l'erreur quadratique moyenne,

$$\begin{aligned}
\mathbb{E}\Big[y - \widehat{f}_n(x)\Big]^2 &= \mathbb{E}\Big[y - m(x) + m(x) - \widehat{f}_n(x)\Big]^2 \\
&= \mathbb{E}[y - m(x)]^2 + \mathbb{E}\Big[m(x) - \widehat{f}_n(x)\Big]^2 + 2\Big[m(x) - \widehat{f}_n(x)\Big]\underbrace{\mathbb{E}\,[y - m(x)]}_{=0} \\
&= \mathbb{E}\,\varepsilon^2 + \Big[m(x) - \widehat{f}_n(x)\Big]^2.
\end{aligned}$$

On observe ainsi que le terme $[m(x) - \widehat{f}_n(x)]^2$ mesure la qualité du prédicteur (plus il est petit, meilleur est le prédicteur au sens de l'erreur quadratique moyenne).

Remarque 1.1.1 *Le meilleur prédicteur de y en x au sens de l'erreur quadratique moyenne est donc*

$$\widehat{f}_n(x) = m(x) = \mathbb{E}(y),$$

la fonction de régression. Mais, la fonction $m(\cdot)$ étant supposée inconnue, il faut construire un autre prédicteur.

1.1.2 Le compromis biais-variance

On s'intéresse ici à l'influence du choix de l'échantillon sur la qualité du prédicteur : il se peut, d'une part, que le prédicteur $\widehat{f}_n(x)$ de y en x que l'on a construit varie notablement si l'on fait varier les données $\mathcal{D}$ (on parle alors de *variance* du prédicteur) ; d'autre part, il est possible qu'en moyenne (sur tous les $\mathcal{D}$ possibles) $\widehat{f}_n(x)$ soit éloigné de $m(x)$ (on parle alors d'un *biais* du prédicteur) [55].

Supposons qu'une loi de probabilité existe sur l'ensemble des ensembles d'entraînement à n observations $\mathcal{D}$. Par un calcul similaire à celui du paragraphe précédent, on peut alors calculer la moyenne par rapport à l'ensemble d'entraînement (notée $\mathbb{E}_{\mathcal{D}}$) de la quantité mesurant la qualité du prédicteur $\widehat{f}_n(x)$,

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}}\Big[\widehat{f}_n(x) - m(x)\Big]^2 &= \mathbb{E}_{\mathcal{D}}\Big[\widehat{f}_n(x) - \mathbb{E}_{\mathcal{D}}\left[\widehat{f}_n(x)\right] + \mathbb{E}_{\mathcal{D}}\left[\widehat{f}_n(x)\right] - m(x)\Big]^2 \\
&= \mathbb{E}_{\mathcal{D}}\Big(\widehat{f}_n(x) - \mathbb{E}_{\mathcal{D}}\left[\widehat{f}_n(x)\right]\Big)^2 + \mathbb{E}_{\mathcal{D}}\Big(\mathbb{E}_{\mathcal{D}}\left[\widehat{f}_n(x)\right] - m(x)\Big)^2 \\
&\quad +2\Big(\mathbb{E}_{\mathcal{D}}\left[\widehat{f}_n(x)\right] - m(x)\Big)\underbrace{\mathbb{E}_{\mathcal{D}}\Big(\widehat{f}_n(x) - \mathbb{E}_{\mathcal{D}}\left[\widehat{f}_n(x)\right]\Big)}_{=0} \\
&= \underbrace{\Big(\mathbb{E}_{\mathcal{D}}\left[\widehat{f}_n(x)\right] - m(x)\Big)^2}_{biais} + \underbrace{\mathbb{E}_{\mathcal{D}}\Big(\widehat{f}_n(x) - \mathbb{E}_{\mathcal{D}}\left[\widehat{f}_n(x)\right]\Big)^2}_{variance}.
\end{aligned}$$

Si $\widehat{f}_n(x)$ est différent en moyenne de $m(x)$, alors $\widehat{f}_n(x)$ est dit un estimateur biaisé de $m(x)$. Un estimateur sans biais peut cependant avoir une grande erreur quadratique moyenne si la variance (la sensibilité aux données) est grande. Le biais et la variance sont donc deux facteurs pouvant donner lieu à un mauvais prédicteur. En pratique, ne pouvant pas obtenir un estimateur à la fois sans biais et de variance nulle, on cherche un compromis entre la contribution du biais et de la variance : la variance du prédicteur est réduite par un lissage de la fonction qui diminue la sensibilité aux données, mais alors un biais est introduit car on a perdu l'information donnée par les valeurs des observations (voir la figure 2.5 page 71).

Exemple 1.1.2 *Supposons que l'on souhaite estimer $m(x)$ à partir des n observations bruitées $y_i = m(x_i) + \varepsilon_i$, $i = 1, \ldots, n$.*

- *Un premier estimateur possible pourrait être un* interpolateur, i.e. $\widehat{f}_n(x_i) = y_i \ \forall i$. *Cet estimateur est sans biais aux x_i car*

$$\mathbb{E}_{\mathcal{D}}\left[\widehat{f}_n(x_i)\right] = \mathbb{E}\left[m(x_i) + \varepsilon_i\right] = m(x_i).$$

Cependant,

$$\mathbb{E}_{\mathcal{D}}\Big(\widehat{f}_n(x_i) - \mathbb{E}_{\mathcal{D}}\left[\widehat{f}_n(x_i)\right]\Big)^2 = \mathbb{E}_{\mathcal{D}}[m(x_i) + \varepsilon_i - m(x_i)]^2 = \mathbb{E}\,\varepsilon_i^2.$$

Ainsi, la contribution de la variance à l'erreur quadratique moyenne est grande.

– *À l'autre extrême, on pourrait prendre* $\widehat{f}_n(x) = f(x)$*, avec* $f(\cdot)$ *une fonction indépendante des données. La variance est alors nulle aux observations (le prédicteur est insensible aux données), mais la présence d'un biais est très probable car il n'est pas tenu compte des observations dans la construction du prédicteur.*

Une façon de prendre en compte à la fois la contribution du biais et de la variance dans la construction du prédicteur est présentée au paragraphe suivant.

1.1.3 Régression régularisée

Si l'on souhaite construire un prédicteur ayant un biais et une variance réduits, on peut considérer la solution du problème de minimisation suivant. Soit $\mathcal{H}$ un espace abstrait de solutions (pour le moment), appelé *espace d'hypothèses*. On y recherche un prédicteur de la forme

$$\widehat{f}_n(x) = \operatorname*{argmin}_{f \in \mathcal{H}} \sum_{i=1}^{n} c(y_i, f(x_i)) + \mathrm{p}(f),$$

où

– la *fonction de coût*, ou *fonction de perte*, $c(y_i, f(x_i))$ mesure l'adéquation du prédicteur aux donnée (le biais). La plus connue est la fonction de coût quadratique

$$c(y, y') = (y - y')^2,$$

et nous en verrons d'autres au § 1.4.4.

– la *pénalisation* $\mathrm{p}(f)$ contrôle la régularité du prédicteur (la variance). Dans le cas des *splines plaque mince* univariées (§ 1.4.2), la pénalisation utilisée est

$$p(f) = \gamma \int \left(f^{(l)}(x)\right)^2 \mathrm{d}x,$$

pour un entier $l \geq 1$ fixé et $\gamma > 0$ qui peut être choisi *a priori* ou estimé à partir des données. Ce dernier terme détermine le niveau du compromis biais-variance : de petites valeurs de γ privilégieront un petit biais alors que de grandes valeurs de γ donneront une petite variance.

La solution obtenue est à la fois proche des données et pourvue d'une certaine régularité. On dit qu'on fait de la *régression régularisée*.

Exemple 1.1.3 *Quand on fait de la régression classique, l'espace d'hypothèses est souvent formé des fonctions du type* $f(x) = {}^t m(x)\beta$*, avec* $m(x)$ *un vecteur de fonctions connues (en général des polynômes :* $f(x) = \beta_0 + \beta_1 x + \cdots + \beta_{p-1} x^{p-1}$ *si* $d = 1$*), et* $\beta \in \mathbb{R}^p$ *un vecteur de paramètres inconnus que l'on cherche à estimer. La méthode des* moindres carrés *est un cas particulier du problème de minimisation du* risque empirique *(l'erreur commise aux points d'observation) : on cherche*

$$\operatorname*{argmin}_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2$$

(remarquons qu'ici il n'y pas de terme de pénalisation, seul le biais du prédicteur est minimisé). Suivant le choix de l'espace d'hypothèses, ce problème de minimisation est en général mal posé *[63] (notamment, la solution n'est pas unique, ou est numériquement instable). Dans le cas où*

$\mathcal{H} = \{{}^t m(x)\beta, \beta \in \mathbb{R}^p\}$, *et en supposant la matrice* ${}^tMM$ *inversible, on sait que la solution du problème des moindres carrés est*

$$\widehat{\beta} = ({}^tMM)^{-1}\,{}^tM\,Y^n,$$

avec $M = {}^t(m(x_1)\dots m(x_n))$ *et* $Y^n = {}^t(y_1\dots y_n)$ *[86]. Un grand nombre de paramètres p induit des problèmes de* conditionnement *[58, 74] de la matrice* ${}^tMM$ *: le problème est théoriquement bien posé (l'inverse d'une matrice inversible, aussi mal conditionnée soit-elle, est bien défini), mais d'un point de vue algorithmique le problème est considéré comme mal posé car on est confronté à des instabilités numériques lors du calcul de l'inverse.*

Une solution possible est de régulariser *[167] : dans le cas de l'espace d'hypothèses ci-dessus, cela revient à chercher*

$$\operatorname*{argmin}_{\beta\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n}\left(y_i - {}^t m(x_i)\beta\right)^2 + \gamma\|\beta\|^2.$$

La constante de régularisation $\gamma > 0$ *pénalise les grandes valeurs de* β*, et donc les grandes variations de la solution : la variance du prédicteur est ainsi diminuée. Nous verrons en 1.2.3 des exemples de régression régularisée dans une classe d'espaces d'hypothèses plus générale.*

Pour plus de détails sur la théorie de l'apprentissage statistique, on pourra consulter par exemple [36, 70, 136, 176, 177].

1.2 Apprentissage par méthodes à noyaux

Les espaces de Hilbert à noyau reproduisant *(Reproducing Kernel Hilbert Spaces, RKHS)* forment un cadre théorique approprié à la résolution de problèmes de régression régularisée. Parmi les méthodes à noyaux qui en sont issues, on trouve les fonctions à base radiale *(Radial Basis Functions, RBF)*, le krigeage *(kriging)*, les splines, les ondelettes *(wavelets)* et la régression par vecteurs de support *(Support Vector Regression, SVR)*. Nous allons tout d'abord faire quelques rappels sur les espaces de Hilbert, puis présenterons les principales propriétés des RKHS, et finalement leur application à la régression régularisée avec les méthodes à noyaux. Dans toute la suite, nous nous intéressons au cas de fonctions à valeurs réelles.

1.2.1 Rappels sur les espaces de Hilbert

Nous rappelons ici quelques définitions et propriétés essentielles des espaces de Hilbert. On notera $\mathcal{E}$ un espace vectoriel sur $\mathbb{R}$.

Définition 1.2.1 *Une fonction* $\langle\cdot,\cdot\rangle : \mathcal{E}\times\mathcal{E}\to\mathbb{R}$ *est un produit scalaire sur* $\mathcal{E}$ *si* $\langle\cdot,\cdot\rangle$ *est*
- bilinéaire *:* $y \mapsto \langle x, y\rangle$ *et* $x \mapsto \langle x, y\rangle$ *sont linéaires ;*
- symétrique *:* $\langle x, y\rangle = \langle y, x\rangle \ \ \forall x, y \in \mathcal{E}$ *;*
- positive *:* $\langle x, x\rangle \geq 0 \ \ \forall x \in \mathcal{E}$ *;*
- définie *:* $\langle x, x\rangle = 0 \Longleftrightarrow x = 0 \ \ (x \in \mathcal{E})$.

La norme associée au produit scalaire $\langle\cdot,\cdot\rangle$ est $\|\cdot\| = \sqrt{\langle\cdot,\cdot\rangle}$.

Définition 1.2.2 *Un espace* préhilbertien *est un espace vectoriel muni d'un produit scalaire.*

Théorème 1.2.3 *(Cauchy–Schwarz) Soit* $\mathcal{P}$ *un espace préhilbertien. Alors,* $\forall\, x, y \in \mathcal{P}$, $|\langle x, y\rangle| \leq \|x\|\|y\|$*, avec égalité si, et seulement si, x et y sont proportionnels.*

Remarque 1.2.4 *L'inégalité de Cauchy–Schwarz s'applique aussi quand la fonction $\langle\cdot,\cdot\rangle$ n'est pas définie (au sens de la définition 1.2.1) : on dit alors que c'est un* semi produit scalaire*, et la fonction $\|\cdot\|$ associée est appelée* semi norme*.*

Définition 1.2.5 *Un espace préhilbertien $\mathcal{H}$ est un* espace de Hilbert *si l'espace métrique $(\mathcal{H}, \|\cdot\|)$ est complet.*

On rappelle qu'un espace métrique (E, d) est complet si toute suite de Cauchy de E converge dans E pour la distance d. Voici quelques exemples d'espaces de Hilbert classiques.

Exemple 1.2.6

- $\mathbb{R}^d$*, muni du produit scalaire* $\langle x, y\rangle = {}^t xy \ \ (x, y \in \mathbb{R}^d)$;
- $\mathrm{L}^2[0,1] = \left\{ f : [0,1] \to \mathbb{R}, f \ \lambda\text{-mesurable}, \int f^2 \mathrm{d}\lambda < \infty \right\}$ *muni de la relation d'équivalence*

$$\text{« } f \sim g \text{ si } f = g \ \ \lambda\text{-presque partout »},$$

  et du produit scalaire $\langle f, g\rangle = \int fg \, \mathrm{d}\lambda$*, où λ désigne la mesure de Lebesgue* $(f, g \in \mathrm{L}^2[0,1])$;
- $l^2 = \left\{ x \in \mathbb{R}^{\mathbb{N}}, \sum x_i^2 < \infty \right\}$, $\langle x, y\rangle = \sum x_i y_i \ \ (x, y \in l^2)$.

1.2.2 Espaces de Hilbert à noyau reproduisant

Les RKHS ainsi que leur lien avec les fonctions semi-définies positives sont présentés dans ce paragraphe.

Définition 1.2.7 *Une fonction symétrique $k(\cdot,\cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ est dite* semi-définie positive *si*

$$\forall\, l \in \mathbb{N}^*, \forall\, x_1, \ldots, x_l \in \mathcal{X}, \forall\, \lambda_1, \ldots, \lambda_l \in \mathbb{R}, \quad \sum_{i,j=1}^{l} \lambda_i \lambda_j \, k(x_i, x_j) \geq 0.$$

Remarque 1.2.8 *La fonction $k(\cdot,\cdot)$ est dite* définie positive *si l'inégalité est stricte quand au moins un des λ_i est non nul. On prendra garde aux traductions anglo-saxonnes, « positive definite » (semi-défini positif) et « strictly positive definite » (défini positif).*

Définition 1.2.9 *Un RKHS est un espace de Hilbert $\mathcal{H}$ de fonctions définies sur $\mathcal{X}$ à valeurs réelles, tel que*

$$\forall x \in \mathcal{X}, \exists\, M_x \in \mathbb{R}, \quad |f(x)| \leq M_x \|f\|_{\mathcal{H}} \quad \forall f \in \mathcal{H}. \tag{1.1}$$

En d'autres termes, la fonction d'évaluation $L_x : \mathcal{H} \to \mathbb{R}, f \mapsto f(x)$ est une forme linéaire continue. D'après le théorème de Riesz, il existe donc un unique élément k_x de $\mathcal{H}$ vérifiant

$$L_x f = \langle k_x, f\rangle_{\mathcal{H}} = f(x) \ \ \forall f \in \mathcal{H}. \tag{1.2}$$

Exemple 1.2.10 *[181]* $\mathrm{L}^2[0,1]$ *n'est pas un RKHS, car ses éléments ne sont pas définis point par point (ce sont des classes d'équivalence de fonctions).*

Définition 1.2.11 *La fonction $k(\cdot,\cdot)$ définie par $k(x, x') = k_x(x')$ s'appelle* noyau reproduisant *de l'espace hilbertien $\mathcal{H}$.*

Remarque 1.2.12 *Il est aussi possible de définir, de façon équivalente, les noyaux reproduisants en partant de la relation (1.2) : la relation (1.1) en est alors une conséquence.*

Proposition 1.2.13 *Soit $k(\cdot,\cdot)$ un noyau reproduisant d'un espace hilbertien $\mathcal{H}$ de fonctions définies sur $\mathcal{X}$ à valeurs réelles, alors*

- $k(x,x') = k(x',x) \quad \forall\, x, x' \in \mathcal{X}$ *(symétrie) ;*
- *$k(\cdot,\cdot)$ est semi-définie positive ;*
- $\langle k(x,\cdot), k(x',\cdot)\rangle_{\mathcal{H}} = \langle k_x, k_{x'}\rangle_{\mathcal{H}} = k(x,x')$ *(propriété de reproduction).*

Preuve *La démonstration utilise (1.2) en remplaçant f par le noyau :*

- $k(x,x') = k_x(x') = \langle k_{x'}, k_x\rangle_{\mathcal{H}} = \langle k_x, k_{x'}\rangle_{\mathcal{H}} = k_{x'}(x) = k(x',x)$ *;*
- *nous venons de voir que $k(\cdot,\cdot)$ est symétrique. Soit $f = \sum_{i=1}^{l} \lambda_i k_{x_i} \in \mathcal{H}$, avec $\lambda_i \in \mathbb{R}, x_i \in \mathcal{X}$. Alors*
$$\sum_{i,j=1}^{l} \lambda_i\lambda_j k(x_i,x_j) = \sum_{i=1}^{l} \lambda_i\lambda_j \langle k_{x_i}, k_{x_j}\rangle_{\mathcal{H}} = \|f\|^2_{\mathcal{H}} \geq 0\,;$$
- *la troisième assertion a déjà été démontrée dans le premier item.* □

Théorème 1.2.14 *(Moore-Aronszajn)[181] À tout RKHS correspond un unique noyau reproduisant. Inversement, si $k(\cdot,\cdot)$ est une fonction semi-définie positive et symétrique sur $\mathcal{X}\times\mathcal{X}$, on peut construire un unique RKHS de fonctions à valeurs réelles ayant $k(\cdot,\cdot)$ pour noyau reproduisant.*

Preuve *(esquisse) Pour démontrer la première assertion, il suffit d'observer que $k(x,x') = \langle k_x, k_{x'}\rangle_{\mathcal{H}} \quad \forall x, x' \in \mathcal{X}$, ce qui définit le noyau de manière unique. Pour la deuxième, le RKHS est défini comme suit :*

- *on construit l'espace préhilbertien $\mathcal{P}$, formé des combinaisons linéaires finies de la forme $\sum_i \alpha_i k_{x_i}$ avec $k_x = k(x,\cdot)$, et muni du produit scalaire $\langle \sum_i \alpha_i k_{x_i}, \sum_j \beta_j k_{t_j}\rangle_{\mathcal{P}} = \sum_{ij} \alpha_i\beta_j \langle k_{x_i}, k_{t_j}\rangle_{\mathcal{P}} = \sum \alpha_i\beta_j k(x_i,t_j)$. Le seul point délicat est de vérifier que $\langle\cdot,\cdot\rangle_{\mathcal{P}}$ est défini,* i.e. *$\forall f \in \mathcal{P}, \langle f,f\rangle_{\mathcal{P}} = 0 \Longleftrightarrow f = 0$, en observant que $\forall x \in \mathcal{X}, f(x)^2 = \langle k_x, f\rangle^2_{\mathcal{P}} \leq \langle k_x, k_k\rangle_{\mathcal{P}} \langle f,f\rangle_{\mathcal{P}}$ (par l'inégalité de Cauchy–Schwarz, voir la remarque 1.2.4) ;*
- *après avoir remarqué, en écrivant $|f_i(x) - f(x)| = |\langle f_i - f, k_x\rangle_{\mathcal{P}}| \leq \|f_i - f\|_{\mathcal{P}} \|k_x\|_{\mathcal{P}}$ pour $\{(f_i)_{i\in\mathbb{N}}, f\} \in \mathcal{P}$, que la convergence en norme dans $\mathcal{P}$ entraîne la convergence ponctuelle, on complète $\mathcal{P}$ en lui ajoutant toutes les limites de ses suites de Cauchy pour la norme $\|\cdot\|_{\mathcal{P}}$ (définies point par point comme limites de suites de Cauchy de $\mathbb{R}$). L'espace de Hilbert ainsi obtenu est un RKHS, de noyau reproduisant $k(\cdot,\cdot)$.* □

Donnons un exemple de construction explicite d'un noyau reproduisant. Soit $\{\phi_i\}_{i=1,\ldots,N}$ un ensemble de fonctions définies sur $\mathcal{X}$ à valeurs dans $\mathbb{R}$. Pour tout $x \in \mathcal{X}$, soit $\Phi : \mathcal{X} \to \mathbb{R}^N$ définie par

$$\Phi(x) = (\phi_1(x), \ldots, \phi_N(x))\,.$$

Considérons la fonction

$$k(x,x') = \langle \Phi(x), \Phi(x')\rangle = \sum_{i=1}^{N} \phi_i(x)\phi_i(x'),$$

où $\langle\cdot,\cdot\rangle$ désigne le produit scalaire usuel de $\mathbb{R}^N$. Par définition, k est symétrique. De plus, $\sum_{i,j=1}^{l} \lambda_i\lambda_j k(x_i,x_j) = \langle \sum_{i=1}^{l} \lambda_i \Phi(x_i), \sum_{i=1}^{l} \lambda_i \Phi(x_i)\rangle$, ce qui montre que k est semi-définie positive.

Définition 1.2.15 *[129] La fonction Φ est appelée* fonction des caractéristiques (feature map) *et l'ensemble $\mathcal{F} = \{\Phi(x), x \in \mathcal{X}\}$* espace des caractéristiques (feature space) *associés au noyau k.*

Le noyau reproduisant k ainsi construit peut donc être vu comme un produit scalaire dans l'espace des caractéristiques de dimension N.

Exemple 1.2.16 *[129] Soit $\mathcal{X} \subset \mathbb{R}^d, x = (x_1, \ldots, x_d)$.*

- *Noyau polynômial homogène : $k(x, x') = \langle x, x' \rangle^l$ $(l \in \mathbb{N})$. On peut vérifier que*

$$\Phi(x) = \left\{ x^q \sqrt{C_q^l} \right\}_{|q|=l},$$

où pour $q = (q_1, \ldots, q_d)$, $|q| = \sum_{i=1}^d q_i$ et $C_q^l = l!/(q_1! \ldots q_n!)$. L'espace des caractéristiques est donc constitué de l'ensemble des monômes de $\mathbb{R}^d$ de degré l (donc $N = (d-1+l)!/(d-1)!/l!$).

- *Noyau polynômial non-homogène : $k(x, x') = (a + \langle x, x' \rangle)^l$ $(a > 0, l \in \mathbb{N})$. Il s'agit du noyau polynômial homogène appliqué à $(\sqrt{a}, x) \in \mathbb{R}^{d+1}$. L'espace des caractéristiques est formé de l'ensemble des monômes de $\mathbb{R}^d$ de degré au plus l. Par exemple, si $d = l = 2$, on a*

$$\Phi(x) = \left(\sqrt{a}, \sqrt{2a}x_1, \sqrt{2a}x_2, x_1^2, x_2^2, \sqrt{2}x_1x_2 \right)$$

(en faisant $a = 0$, on obtient l'espace des caractéristiques du noyau polynômial homogène).

Nous allons voir dans la suite que pour certains noyaux il y a une infinité de caractéristiques.

Définition 1.2.17 *[129] Une fonction continue $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ est appelée* noyau de Mercer *si $k(\cdot, \cdot)$ est semi-définie positive et symétrique.*

Un noyau $k(\cdot, \cdot)$ étant donné, il est souvent difficile de vérifier si il est semi-défini positif. Un moyen pratique est donné dans la proposition suivante.

Définition 1.2.18 *[36] Une fonction $f : [0, \infty[\to \mathbb{R}$ est* complètement monotone *si f est $\mathcal{C}^\infty$ et si, pour tout $t > 0$ et $l \in \mathbb{N}$, $(-1)^l f^{(l)}(t) \geq 0$.*

Proposition 1.2.19 *[36] Soit $\mathcal{X} \subset \mathbb{R}^d$, $f : [0, \infty[\to \mathbb{R}$ et $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ définis par $k(x, x') = f(\|x - x'\|^2)$. Alors k est semi-définie positive si, et seulement si, f est complètement monotone.*

Exemple 1.2.20 *[36]*

- $k(x, x') = \mathrm{e}^{-\frac{\|x-x'\|^2}{c^2}}$ *(noyau gaussien) ;*
- $k(x, x') = (c^2 + \|x - x'\|^2)^{-\alpha}$ $(\alpha > 0)$.

Le théorème suivant généralise le principe de diagonalisation d'une matrice semi-définie positive aux fonctions semi-définies positives, et permet d'introduire le concept de *représentation*. Nous nous restreignons à un compact $\mathcal{X} \subset \mathbb{R}^d$ pour simplifier les hypothèses, c'est de toute façon le type d'espace qui sera utilisé en pratique.

Théorème 1.2.21 *(Mercer-Hilbert-Schmidt)[36, 129, 181] Soit $\mathcal{X}$ un compact de $\mathbb{R}^d$ et $k(\cdot, \cdot)$ un noyau de Mercer sur $\mathcal{X} \times \mathcal{X}$. Alors il existe une famille orthonormale de fonctions propres*

continues $\{\Phi_i\}_{i\in\mathbb{N}}$ *de carré intégrable sur* $\mathcal{X}$ *par rapport à la mesure de Lebesgue, et de valeurs propres* $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$*, telles que*

$$\begin{aligned}\int_{\mathcal{X}} k(x,x')\Phi_i(x')\,\mathrm{d}x' &= \lambda_i\Phi_i(x), \qquad i=1,2,\ldots;\\ \int_{\mathcal{X}}\int_{\mathcal{X}} k^2(x,x')\,\mathrm{d}x\,\mathrm{d}x' &= \sum_{i=1}^{\infty}\lambda_i^2<\infty;\\ k(x,x') &= \sum_{i=1}^{\infty}\lambda_i\Phi_i(x)\Phi_i(x').\end{aligned}$$

Dans cette dernière égalité, la convergence est absolue ($\forall x,x'\in\mathcal{X}$) *et uniforme (sur* $\mathcal{X}\times\mathcal{X}$*). Le nombre de valeurs propres* λ_i *non nulles s'appelle* dimension *du noyau* $k(\cdot,\cdot)$*. C'est la dimension de l'espace des caractéristiques associé au noyau* k*.*

Sous certaines hypothèses, le noyau reproduisant d'un RKHS $\mathcal{H}$ de fonctions à valeurs réelles définies sur $\mathcal{X}$ admet donc la représentation $k(x,x')=\sum_{i=1}^{\infty}\lambda_i\Phi_i(x)\Phi_i(x')$, où $\{\Phi_i\}_{i\in\mathbb{N}}$ est une famille orthonormale pour le produit scalaire de $\mathcal{L}^2_{\mathrm{cont}}(\mathcal{X})$ (espace des fonctions continues sur $\mathcal{X}$ de carré intégrable). Sous ces conditions, on obtient une caractérisation simple du RKHS.

Proposition 1.2.22 *[181] On se place sous les hypothèses et notations du théorème 1.2.21. Soit* $f:\mathcal{X}\to\mathbb{R}$*, et*

$$a_i=\int_{\mathcal{X}} f(x)\Phi_i(x)\,\mathrm{d}x.$$

Alors $f\in\mathcal{H}$ *si, et seulement si,*

$$\sum_{i=1}^{\infty}\frac{a_i^2}{\lambda_i}<\infty$$

et dans ce cas

$$\|f\|^2_{\mathcal{H}}=\sum_{i=1}^{\infty}\frac{a_i^2}{\lambda_i}.$$

Le RKHS $\mathcal{H}$ admet alors la représentation simple [70, 181]

$$\mathcal{H}=\left\{f=\sum_{i=1}^{\infty}a_i\Phi_i,\ \|f\|^2_{\mathcal{H}}=\sum_{i=1}^{\infty}\frac{a_i^2}{\lambda_i}<\infty\right\}, \tag{1.3}$$

avec le produit scalaire $\langle\sum_{i=1}^{\infty}a_i\Phi_i(x),\sum_{i=1}^{\infty}b_i\Phi_i(x)\rangle_{\mathcal{H}}=\sum_{i=1}^{\infty}a_ib_i/\lambda_i$.

Nous allons maintenant voir comment la théorie des RKHS s'applique aux problèmes d'apprentissage.

1.2.3 RKHS et apprentissage

Les RKHS peuvent être utilisés de façon naturelle comme espaces d'hypothèses quand on fait de l'apprentissage. Reprenons le problème de régression régularisée présenté en 1.1.3, en prenant un RKHS pour espace d'hypothèses ce qui va garantir de bonnes propriétés de la solution. Le théorème suivant, cas particulier du *théorème de réprésentation*, permet de comprendre en quoi l'ajout d'un terme de pénalisation permet de passer à un problème bien posé numériquement.

Théorème 1.2.23 *(du représentant, cas particulier) [87, 128, 182] Soit $\mathcal{H}$ un RKHS de fonctions à valeurs réelles définies sur un ensemble $\mathcal{X}$, de noyau reproduisant $k(\cdot,\cdot)$. Soit $(x_i, y_i)_{i=1\ldots n}$ l'ensemble d'entraînement, avec $x_i \in \mathcal{X}$ et $y_i \in \mathbb{R}$. Alors, pour $\gamma > 0$ donné,*

$$\operatorname*{argmin}_{f\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} (y_i - f(x_i))^2 + \gamma\|f\|_{\mathcal{H}}^2 \tag{1.4}$$

est unique et s'écrit sous la forme

$$f(\cdot) = \sum_{i=1}^{n} \alpha_i k(x_i, \cdot).$$

Le vecteur des coefficients $\alpha = {}^t(\alpha_1, \ldots, \alpha_n)$ est solution du problème numériquement bien posé

$$(n\gamma I_n + K)\alpha = Y^n, \tag{1.5}$$

avec I_n la matrice identité de taille n, K la matrice de taille $n \times n$ $(k(x_i, x_j))_{1\leq i,j\leq n}$ et $Y^n = {}^t(y_1, \ldots, y_n)$.

Preuve *Pour résoudre (1.4), on cherche les zéros de la dérivée de la fonctionnelle*

$$f \mapsto \frac{1}{n}\sum_{i=1}^{n} (y_i - f(x_i))^2 + \gamma\|f\|_{\mathcal{H}}^2,$$

qui s'écrit

$$g \mapsto -\frac{2}{n}\sum_{i=1}^{n} (y_i - f(x_i))g(x_i) + 2\gamma\langle f, g\rangle_{\mathcal{H}}.$$

On cherche donc à résoudre

$$\frac{1}{n}\sum_{i=1}^{n} (y_i - f(x_i))g(x_i) - \gamma\langle f, g\rangle_{\mathcal{H}} = 0. \tag{1.6}$$

L'équation (1.6) doit être verifiée pour tout $g \in \mathcal{H}$, et en particulier pour $g = k_x$. En utilisant la relation $\langle f, k_x\rangle = f(x)$, on obtient

$$\begin{aligned} f(x) &= \frac{1}{n\gamma}\sum_{i=1}^{n} (y_i - f(x_i))\, k_x(x_i) \\ &= \sum_{i=1}^{n} \alpha_i k(x_i, x), \end{aligned}$$

avec $\alpha_i = \dfrac{y_i - f(x_i)}{n\gamma}$. On a donc $n\gamma\alpha_i + f(x_i) = y_i \ \forall i$, c'est-à-dire

$$n\gamma\alpha_i + \sum_{j=1}^{n} \alpha_j k(x_i, x_j) = y_i \quad \forall i,$$

ce qui est exactement (1.5). □

C'est le terme $n\gamma I_n$ qui rend le problème bien posé : plus grand est $n\gamma$, meilleur est le conditionnement de la matrice $n\gamma I_n + K$.

Le terme $(y_i - f(x_i))^2$ dans l'équation (1.4) peut être remplacé par une autre fonction de perte (ou fonction de coût) que la fonction quadratique. L'équation (1.5) est alors non-linéaire en général [128], et doit être résolue numériquement par une méthode itérative de descente [9], alors que dans le cas quadratique on a obtenu un système linéaire qui se résout analytiquement simplement en inversant une matrice.

Voici maintenant une version plus générale du théorème de représentation, dans le cadre de la *régression semi-paramétrique* : la fonction f s'écrit maintenant comme la somme d'une fonction appartenant au RKHS, et d'une combinaison linéaire de fonctions de base qui sont déterminées par notre connaissance *a priori* du problème.

Théorème 1.2.24 *(du représentant)[182] Soit $\mathcal{H}$ un RKHS de fonctions à valeurs réelles définies sur un ensemble $\mathcal{X}$, de noyau reproduisant $k(\cdot,\cdot)$. Soit $\{(x_i, y_i)\}_{i=1\dots n}$ l'ensemble d'entraînement, avec $x_i \in \mathcal{X}$ et $y_i \in \mathbb{R}$. Soient $\{m_j\}_{j=1\dots p}$ des fonctions définies sur $\mathcal{X}$ ayant la propriété que la matrice $M = (m_j(x_i))_{i=1\dots n}^{j=1\dots p}$ est de rang p (on note $\mathcal{M}$ l'espace vectoriel engendré par les m_i). Soit $c_i(y_i, f)$ une fonctionnelle qui ne dépend de f qu'au travers de $f_i = f(x_i)$,* i.e. *$c_i(y_i, f) = c_i(y_i, f_i)$. Alors, pour $\gamma > 0$ donné,*

$$\underset{f=f_{\mathcal{M}}+f_{\mathcal{H}}\in\mathcal{M}+\mathcal{H}}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} c_i(y_i, f) + \gamma\|f_{\mathcal{H}}\|_{\mathcal{H}}^2 \tag{1.7}$$

s'écrit sous la forme

$$f(\cdot) = \sum_{j=1}^{p} \nu_j m_j(\cdot) + \sum_{i=1}^{n} \alpha_i k(x_i, \cdot). \tag{1.8}$$

Une version encore plus générale du théorème de représentation est donnée dans [149].

Remarque 1.2.25 *[182]*

- *Pour que la matrice M soit de rang plein, il faut que le nombre de données d'apprentissage n soit au moins égal à p ;*
- *la norme étant évaluée uniquement pour la fonction $f_{\mathcal{H}}$, la partie paramétrique $f_{\mathcal{M}}$ du modèle n'est pas régularisée, ce qui signifie que les coefficients ν_j peuvent être arbitrairement grands ;*
- *si c_i est dérivable par rapport à f_i et si la matrice $K = (k(x_i, x_j))_{1\le i,j\le n}$ est de rang plein, on peut montrer que le vecteur $\alpha = {}^t(\alpha_1, \dots, \alpha_n)$ vérifie ${}^tM\alpha = 0$ (on retrouvera cette propriété au théorème 1.2.26) ;*
- *l'unicité de la solution est garantie si K est de rang plein et si les c_i sont strictement convexes par rapport à f_i ;*
- *finalement, pour s'assurer que la matrice K est de rang plein, on peut aussi supposer que la fonction $k(\cdot,\cdot)$ est définie positive et que les données d'observation sont distinctes.*

Pour le choix de la constante de régularisation γ, on pourra consulter [35]. Ce sont les fonctions $c_i(\cdot)$, $m_j(\cdot)$ et $k(\cdot,\cdot)$ qui déterminent quel type de méthode à noyaux on utilise.

Voici finalement la forme explicite de la solution quand la fonction de coût est quadratique, dans le cas de la régression tout d'abord (la courbe de f est proche des données d'observation, tout en satisfaisant une certaine condition de régularité), puis de l'interpolation (la courbe de la fonction f passe par les points observés). Ce cadre correspond exactement à la théorie du krigeage.

Théorème 1.2.26 *(du représentant, cas quadratique)[182] Soit $\mathcal{H}$ un RKHS de fonctions à valeurs réelles définies sur un ensemble $\mathcal{X}$, de noyau reproduisant $k(\cdot,\cdot)$. Soit $\{(x_i,y_i)\}_{i=1\dots n}$ l'ensemble d'entraînement, avec $x_i \in \mathcal{X}$ et $Y^n = {}^t(y_1,\dots,y_n) \in \mathbb{R}^n$. Soient $\{m_j\}_{j=1\dots p}$ des fonctions définies sur $\mathcal{X}$ ayant la propriété que la matrice $M = (m_j(x_i))_{i=1\dots n}^{j=1\dots p}$ est de rang p, et soit $\mathcal{M}$ l'espace vectoriel engendré par les m_i. Alors, pour $\gamma > 0$ donné,*

$$\underset{f=f_{\mathcal{M}}+f_{\mathcal{H}}\in\mathcal{M}+\mathcal{H}}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 + \gamma\|f_{\mathcal{H}}\|_{\mathcal{H}}^2 \tag{1.9}$$

est donné par

$$f(\cdot) = \sum_{j=1}^{p}\nu_j m_j(\cdot) + \sum_{i=1}^{n}\alpha_i k(x_i,\cdot),$$

où les vecteurs des coefficients $\nu = {}^t(\nu_1,\dots,\nu_p)$ et $\alpha = {}^t(\alpha_1,\dots,\alpha_n)$ sont solutions du système

$$\begin{pmatrix} K' & M \\ {}^tM & 0 \end{pmatrix}\begin{pmatrix}\alpha\\ \nu\end{pmatrix} = \begin{pmatrix} Y^n \\ 0\end{pmatrix},$$

avec $K' = K + n\gamma I = (k(x_i,x_j) + n\gamma\delta_{ij})_{1\le i,j\le n}$, et sont donnés par

$$\begin{aligned}\nu &= ({}^tMK'^{-1}M)^{-1}{}^tMK'^{-1}Y^n\,;\\ \alpha &= K'^{-1}(I - M({}^tMK'^{-1}M)^{-1}{}^tMK'^{-1})Y^n.\end{aligned}$$

Preuve *On peut écrire la solution de (1.9) sous la forme*

$$f = f_{\mathcal{M}} + f_{\mathcal{H}} = \sum_{j=1}^{p}\nu_j m_j + \sum_{i=1}^{n}\alpha_i k_{x_i} + \xi,$$

avec $\xi \in \mathcal{H}, \xi \perp k_{x_1},\dots,k_{x_n}$ (donc $\xi(x_i) = \langle \xi, k_{x_i}\rangle_{\mathcal{H}} = 0\ \forall i$). On cherche donc le minimum de

$$\frac{1}{n}\|Y^n - (K\alpha + M\nu)\|_2^2 + \gamma\left({}^t\alpha K\alpha + \|\xi\|_{\mathcal{H}}^2\right).$$

Il est alors évident que $\|\xi\|_{\mathcal{H}}^2 = 0$. On recherche donc

$$\underset{\alpha\in\mathbb{R}^n,\nu\in\mathbb{R}^p}{\operatorname{argmin}}\ \Upsilon(\alpha,\nu),$$

avec

$$\Upsilon(\alpha,\nu) = \frac{1}{n}\|Y^n - (K\alpha + M\nu)\|_2^2 + \gamma\,{}^t\alpha K\alpha.$$

On résout le système

$$\begin{aligned}
\begin{cases}\frac{\partial\Upsilon}{\partial\alpha}(\alpha,\nu) = 0\\ \frac{\partial\Upsilon}{\partial\nu}(\alpha,\nu) = 0\end{cases}
&\iff \begin{cases}-\frac{2}{n}{}^tK(Y^n - K\alpha - M\nu) + 2\gamma K\alpha = 0\\ -\frac{2}{n}{}^tM(Y^n - K\alpha - M\nu) = 0\end{cases}\\
&\iff \begin{cases}\alpha = \frac{1}{n\gamma}(Y^n - K\alpha - M\nu)\\ {}^tM(Y^n - K\alpha - M\nu) = 0\end{cases}\\
&\iff \begin{cases}\alpha = \frac{1}{n\gamma}(Y^n - K\alpha - M\nu)\\ {}^tM\alpha = 0\end{cases}\\
&\iff \begin{cases}K'\alpha + M\nu = Y^n\\ {}^tM\alpha = 0\end{cases},
\end{aligned}$$

et on retrouve (1.2.26). On vérifie finalement que

$$\begin{aligned}\nu &= ({}^tMK'^{-1}M)^{-1}{}^tMK'^{-1}Y^n\,;\\ \alpha &= K'^{-1}\left(I - M({}^tMK'^{-1}M)^{-1}{}^tMK'^{-1}\right)Y^n\end{aligned}$$

sont les solutions du système. □

Intéressons-nous finalement au problème de l'*interpolation* : en l'absence d'erreur de mesure, on souhaite trouver une fonction f qui vérifie $f(x_i) = y_i\ \forall i$. Sous certaines hypothèses, on obtient une solution ayant la même forme que dans le cas de la régression.

Théorème 1.2.27 *(du représentant pour l'interpolation, cas quadratique) Soit $\mathcal{H}$ un RKHS de fonctions à valeurs réelles définies sur un ensemble $\mathcal{X}$, de noyau reproduisant $k(\cdot,\cdot)$. Soit $\{(x_i, y_i)\}_{i=1\dots n}$ l'ensemble d'entraînement, avec $x_i \in \mathcal{X}$ et $Y^n = {}^t(y_1,\dots,y_n) \in \mathbb{R}^n$. Soient $\{m_j\}_{j=1\dots p}$ des fonctions définies sur $\mathcal{X}$ ayant la propriété que la matrice $M = (m_j(x_i))_{i=1\dots n}^{j=1\dots p}$ est de rang p, et soit $\mathcal{M}$ l'espace vectoriel engendré par les m_i. On suppose aussi que la matrice $K = (k(x_i,x_j))_{1\le i,j\le n}$ est inversible. Alors,*

$$\begin{cases}\underset{f=f_{\mathcal{M}}+f_{\mathcal{H}}\in\mathcal{M}+\mathcal{H}}{\operatorname{argmin}} \|f_{\mathcal{H}}\|^2_{\mathcal{H}}\,;\\ f(x_i) = y_i \quad \forall i,\end{cases} \tag{1.10}$$

est donné par

$$f(\cdot) = \sum_{j=1}^{p}\nu_j m_j(\cdot) + \sum_{i=1}^{n}\alpha_i k(x_i,\cdot), \tag{1.11}$$

où les vecteurs des coefficients $\nu = {}^t(\nu_1,\dots,\nu_p)$ et $\alpha = {}^t(\alpha_1,\dots,\alpha_n)$ sont solutions du système

$$\begin{pmatrix}K & M\\ {}^tM & 0\end{pmatrix}\begin{pmatrix}\alpha\\ \nu\end{pmatrix} = \begin{pmatrix}Y^n\\ 0\end{pmatrix}, \tag{1.12}$$

et donnés par

$$\begin{aligned}\nu &= ({}^tMK^{-1}M)^{-1}{}^tMK^{-1}Y^n\,;\\ \alpha &= K^{-1}\left(I - M({}^tMK^{-1}M)^{-1}{}^tMK^{-1}\right)Y^n.\end{aligned}$$

Preuve *On peut écrire la solution de (1.10) sous la forme*

$$f = f_{\mathcal{M}} + f_{\mathcal{H}} = \sum_{j=1}^{p}\nu_j m_j + \sum_{i=1}^{n}\alpha_i k_{x_i} + \xi,$$

avec $\xi \in \mathcal{H}, \xi \perp k_{x_1},\dots,k_{x_n}$ (donc $\xi(x_i) = \langle \xi, k_{x_i}\rangle_{\mathcal{H}} = 0\ \forall i$). On cherche donc

$$\begin{cases}\underset{\alpha\in\mathbb{R}^n,\xi\in\mathcal{H}}{\operatorname{argmin}} {}^t\alpha K\alpha + \|\xi\|^2_{\mathcal{H}}\,;\\ f(x_i) = y_i \quad \forall i.\end{cases}$$

Il est alors évident que $\|\xi\|^2_{\mathcal{H}} = 0$, car la fonction ξ n'intervient pas dans la condition $f(x_i) = y_i\ \forall i$ puisque $\xi(x_i) = 0\ \forall i$. On recherche donc

$$\begin{cases}\underset{\alpha\in\mathbb{R}^n,\nu\in\mathbb{R}^p}{\operatorname{argmin}} {}^t\alpha K\alpha\,;\\ K\alpha + M\nu = Y^n.\end{cases}$$

On utilise la technique des multiplicateurs de Lagrange [13] en cherchant $(\alpha, \nu, \lambda) \in \mathbb{R}^{n+p+n}$ *qui minimisent*

$$\Upsilon(\alpha, \nu, \lambda) = {}^t\alpha K\alpha + 2{}^t\lambda(Y^n - K\alpha - M\nu).$$

Le système à résoudre est

$$\begin{cases} \frac{\partial \Upsilon}{\partial \alpha}(\alpha, \nu, \lambda) = 0 \\ \frac{\partial \Upsilon}{\partial \nu}(\alpha, \nu, \lambda) = 0 \\ \frac{\partial \Upsilon}{\partial \lambda}(\alpha, \nu, \lambda) = 0 \end{cases} \iff \begin{cases} 2K\alpha - 2K\lambda = 0 \\ Y^n - K\alpha - M\nu = 0 \\ {}^tM\alpha = 0 \end{cases}$$

$$\iff \begin{cases} \alpha = \lambda \\ Y^n = K\alpha + M\nu \\ {}^tM\lambda = 0 \end{cases}$$

$$\iff \begin{cases} \alpha = \lambda \\ Y^n = K\alpha + M\nu \\ {}^tM\alpha = 0 \end{cases},$$

et on retrouve (1.12). On vérifie finalement que

$$\begin{aligned} \nu &= ({}^tMK^{-1}M)^{-1}{}^tMK^{-1}Y^n\,; \\ \alpha &= K^{-1}\left(I - M({}^tMK^{-1}M)^{-1}{}^tMK^{-1}\right)Y^n \end{aligned}$$

sont les solutions du système. □

1.3 RKHS construit à partir d'un noyau conditionnellement semi-défini positif

Une extension des modèles précédents existe, où l'espace d'hypothèses est construit à partir d'un noyau conditionnellement semi-défini positif. L'analogue du théorème du représentant dans ce nouveau cas est rappelé, et permet d'observer que les solutions s'écrivent sous la même forme que dans le cas semi-défini positif. La théorie présentée dans ce paragraphe est à la base notamment des splines plaque mince et du krigeage intrinsèque. Nous nous restreignons dans toute la suite au cas de fonctions à valeurs réelles (une généralisation au cas complexe est donnée dans [107, 111]).

1.3.1 Fonctions conditionnellement semi-définies positives

Soit $\mathbb{P}_q^d$ l'espace des polynômes de degré au plus q définis sur un ensemble $\mathcal{X} \subset \mathbb{R}^d$, et $\{u_1, \ldots, u_N\}$ un ensemble de points de $\mathcal{X}$.

Définition 1.3.1 *[99, 107] L'ensemble* $\{u_1, \ldots, u_N\}$ *est dit* unisolvant *(ou* non-dégénéré*), par rapport à* $\mathbb{P}_q^d$ *si,*

$$\text{pour } m(\cdot) \in \mathbb{P}_q^d,\ \ m(u_i) = 0 \quad \forall i = 1, \ldots, N \Longrightarrow m = 0.$$

Autrement dit, les polynômes de $\mathbb{P}_q^d$ *sont uniquement déterminés par leurs valeurs en* $\{u_1, \ldots, u_N\}$.

Un ensemble unisolvant de $\mathcal{X}$ sera utilisé pour construire l'espace d'hypothèses dans la définition 1.3.13. L'ensemble d'apprentissage $\{x_1, \ldots, x_n\}$ sera supposé unisolvant afin d'assurer l'unicité de la solution du problème de minimisation.

Exemple 1.3.2

- *Si $d = 2$ et $q = 1$: un polynôme de degré 1 sur $\mathbb{R}^2$ s'écrit $m(t) = m(t_1, t_2) = a_1t_1+a_2t_2+a_0$. L'équation $m(t) = 0$ définit une droite de $\mathbb{R}^2$ si $(a_1, a_2) \neq (0, 0)$. Les équations $m(u_i) = 0\ \forall i = 1, \ldots, N$ entraînent donc la nullité des coefficients si, et seulement si, les u_i ne sont pas situés sur une même droite. Autrement dit, tout ensemble de points de $\mathbb{R}^2$ non alignés est un ensemble unisolvant si $q = 1$ (il faut donc $n \geq 3 = \dim \mathbb{P}_1^2$) ;*
- *si $d = 2$ et $q = 2$: un polynôme de degré 2 sur $\mathbb{R}^2$ s'écrit $m(t) = m(t_1, t_2) = a_{11}t_1^2 + a_{22}t_2^2 + a_{12}t_1t_2 + a_1t_1 + a_2t_2 + a_0$. L'équation $m(t) = 0$ définit donc une conique de $\mathbb{R}^2$ si $(a_{11}, a_{12}, a_{22}) \neq (0, 0, 0)$, ou une droite si $(a_1, a_2) \neq (0, 0)$. Comme précédemment, on en déduit que tout ensemble de points de $\mathbb{R}^2$ non situés sur une même conique est un ensemble unisolvant si $q = 2$ (il faut donc $n \geq 6 = \dim \mathbb{P}_2^2$).*

La classe de mesures présentée maintenant va nous permettre de construire l'espace d'hypothèses. Considérons la mesure discrète sur $\mathbb{R}^d$

$$\lambda = \sum_{i=1}^{N} \lambda_i \delta_{z_i},$$

où N est un entier naturel quelconque, $\lambda_i \in \mathbb{R}, z_i \in \mathbb{R}^d$ et δ_x désigne la mesure de Dirac au point $x \in \mathbb{R}^d$. On peut définir une action de λ sur une fonction f de $\mathbb{R}^d$ par

$$\lambda \cdot f = \int f(z)\lambda(\mathrm{d}z) = \sum_{i=1}^{N} \lambda_i f(z_i).$$

Définition 1.3.3 *La mesure discrète λ est* admissible d'ordre q *si*

$$\lambda \cdot m = 0 \quad \forall m \in \mathbb{P}_q^d.$$

L'ensemble des mesures discrètes admissibles d'ordre q est noté Λ_q^d. On appelle combinaison linéaire admissible d'ordre q, *ou* incrément généralisé d'ordre q, *toute expression de la forme*

$$\lambda \cdot f = \sum_{i=1}^{N} \lambda_i f(z_i),$$

avec $\lambda \in \Lambda_q^d$ et $f : \mathbb{R}^d \to \mathbb{R}$.

Exemple 1.3.4

- *Si $q = 0$ [26] : $\lambda \cdot f = f(z_1) + f(z_2) + f(z_3) + f(z_4) - 4f(z_5)$ est une combinaison linéaire admissible d'ordre 0, mais pas d'ordre 1. On peut remarquer qu'il faut imposer la condition $n > 1 = \dim \mathbb{P}_q^0$ pour qu'il existe des mesures admissibles d'ordre 0 non triviales, quel que soit l'espace $\mathcal{X}$ considéré, la plus simple étant obtenue pour $\lambda = \delta_{z_1} - \delta_{z_2} \in \Lambda_0^d$, qui donne la combinaison linéaire admissible (d'ordre 0) $\lambda \cdot f = f(z_1) - f(z_2)$;*
- *si $d = 1$: un polynôme de degré q sur $\mathbb{R}$ s'écrivant $m(x) = a_0 + a_1x + a_2x^2 + \cdots + a_qx^q$, la mesure $\lambda = \sum_{i=1}^{N} \lambda_i \delta_{z_i}$ est admissible d'ordre q si, et seulement si,*

$$(\lambda_1 \ldots \lambda_N) \begin{pmatrix} 1 & z_1 & z_1^2 & \ldots & z_1^q \\ 1 & z_2 & z_2^2 & \ldots & z_2^q \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & z_N & z_N^2 & \ldots & z_N^q \end{pmatrix} = 0 \Longleftrightarrow \lambda = 0.$$

On en déduit que les mesures admissibles d'ordre q non triviales sur $\mathbb{R}$ existent et sont portées par au moins $N > q+1$ points. En effet, si $q \geq N-1$, on reconnaît une matrice de type Vandermonde qui est de rang $\geq N$: le système n'admet donc que la solution triviale $\lambda = 0$; et si $q < N-1$, le système n'est pas de rang plein et admet donc au moins une solution non triviale. Remarquons que la condition d'existence peut s'écrire $N > \dim \mathbb{P}_q^1$;
- *si d est quelconque, on déduit de même l'existence de mesures admissibles non triviales car la condition $N > \dim \mathbb{P}_q^d$ conduit à résoudre un système linéaire singulier.*

Remarque 1.3.5 *[26]*
- *Clairement, $\Lambda_{q+1}^d \subset \Lambda_q^d$;*
- *le choix des fonctions polynômiales dans la définition d'une mesure admissible est le plus classique. Pour une définition utilisant une classe de fonctions plus générale, on pourra consulter [146].*

Il est maintenant possible de définir les fonctions conditionnellement semi-définies positives.

Définition 1.3.6 *Une fonction symétrique $k(\cdot,\cdot)$ est dite* conditionnellement semi-définie positive (c.s-d.p.) *d'ordre q sur $\mathcal{X}$ si*

$$\forall \lambda = \sum_{i=1}^{N} \lambda_i \delta_{z_i} \in \Lambda_q^d, \qquad \sum_{i,j=1}^{N} \lambda_i \lambda_j k(z_i, z_j) \geq 0.$$

Remarque 1.3.7
- *Par convention, on dira qu'une fonction semi-définie positive est d'ordre $q = -1$;*
- *la fonction $k(\cdot,\cdot)$ est dite* conditionnellement définie positive *d'ordre q si l'inégalité est stricte pour tout $\lambda \in \Lambda_q^d \setminus \{0\}$;*
- *on prendra garde au fait que $\mathbb{P}_q^d$ est parfois défini comme l'ensemble des polynômes d'*ordre *au plus q (l'ordre d'un polynôme étant égal à son degré plus 1). Dans ce cas, une fonction c.s-d.p. sera d'ordre plus élevé de un par rapport à la définition 1.3.6, que nous avons choisie par souci de clarté.*

Dans le cas où la fonction $k(\cdot,\cdot)$ est *radiale*, *i.e.* $k(x,x') = k(\|x-x'\|)$, un critère pratique pour déterminer si celle-ci est conditionnellement semi-définie positive est donné par la proposition suivante (qui étend la proposition 1.2.19).

Définition 1.3.8 *[158] Une fonction $f : [0,\infty[\to \mathbb{R}$ est* complètement monotone d'ordre q *si f est $\mathcal{C}^\infty$ et si, pour tout $t > 0$ et $l \in \mathbb{N}, l > q$, $(-1)^l f^{(l)}(t) \geq 0$.*

Remarque 1.3.9 *Nous avons imposé la condition $l > q$ au lieu de $l \geq q$ pour avoir le résultat simple de la proposition suivante. Les fonction complètement monotones d'ordre -1 correspondent aux fonctions complètement monotones de la définition 1.2.18.*

Proposition 1.3.10 *[111, 158] Soit $\mathcal{X} \subset \mathbb{R}^d$, $f : [0,\infty[\to \mathbb{R}$ et $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ définis par $k(x,x') = f(\|x-x'\|^2)$. Alors k est c.s-d.p. d'ordre q si, et seulement si, f est complètement monotone d'ordre q.*

Exemple 1.3.11 *[111, 158, 187]*

$$
\begin{aligned}
k(x,x') &= \mathrm{e}^{-\frac{\|x-x'\|^2}{c^2}} && (\textit{noyau gaussien, } q=-1)\,;\\
k(x,x') &= -\sqrt{\|x-x'\|^2+c^2} && (\textit{multiquadrique, } q=0)\,;\\
k(x,x') &= \frac{1}{\sqrt{\|x-x'\|^2+c^2}} && (\textit{multiquadrique inverse, } q=-1)\,;\\
k(x,x') &= \begin{cases}(-1)^n\|x-x'\|^{2n+1}\\ (-1)^{n+1}\|x-x'\|^{2n}\log\|x-x'\|\end{cases} && \left(\textit{spline plaque mince, } \begin{cases}q=n-1\\ q=n\end{cases}, n\in\mathbb{N}\right).
\end{aligned}
$$

Remarque 1.3.12 *Si $k(\cdot,\cdot)$ est radiale, on entre dans le cadre des* fonctions à base radiale (Radial Basis Functions, RBF).

1.3.2 Construction de l'espace d'hypothèses

Soit $k(\cdot,\cdot)$ une fonction conditionnellement semi-définie positive d'ordre q. Alors l'application

$$
\begin{array}{rccl}
\langle\cdot,\cdot\rangle_{\Lambda_q^d}: & \Lambda_q^d\times\Lambda_q^d & \longrightarrow & \mathbb{R}\\
& (\lambda,\mu) & \longmapsto & \sum_{i=1}^N\sum_{j=1}^M\lambda_i\mu_j k(z_i,t_j),
\end{array}
$$

pour $\lambda=\sum_{i=1}^N\lambda_i\delta_{z_i}$ et $\mu=\sum_{j=1}^M\mu_j\delta_{t_j}$, est un semi produit scalaire. On peut alors définir l'espace préhilbertien

$$\left(\Lambda_q^d/\mathcal{N},\langle\cdot,\cdot\rangle_{\Lambda_q^d}\right),$$

avec $\mathcal{N}=\{\lambda\in\Lambda_q^d,\langle\lambda,\lambda\rangle_{\Lambda_q^d}=0\}$, que l'on complète en un espace de Hilbert $\overline{\Lambda_q^d}$. On note

$$
\begin{array}{rccc}
\Pi: & \Lambda_q^d & \longrightarrow & \overline{\Lambda_q^d}\\
& \lambda & \longmapsto & \lambda+\mathcal{N}
\end{array}
$$

la projection canonique. Cet espace de Hilbert permet de construire le RKHS à l'intérieur duquel on cherchera la solution du problème de minimisation.

L'espace d'hypothèses analogue à l'espace des solutions $\mathcal{M}+\mathcal{H}$ du théorème 1.2.26 est présenté maintenant. La construction est plus abstraite, car on ne peut pas définir ici une fonction f à partir d'une mesure $\lambda\in\Lambda_q^d$ en écrivant simplement $f(x)=\langle\lambda,\delta_x\rangle_{\Lambda_q^d}$, puisque $\delta_x\notin\Lambda_q^d$. On utilise donc une mesure alternative $\delta_{(x)}$ qui appartient à Λ_q^d.

Définition 1.3.13 *[107] Soient $(m_1,\ldots,m_{p'})$ une base de $\mathbb{P}_q^d$ et $\Xi=\{\zeta_1,\ldots,\zeta_{p'}\}$ un ensemble unisolvant de $\mathcal{X}$ tels que $m_i(\zeta_j)=\delta_{ij}\ \forall i,j=1,\ldots,p'$. Introduisons la mesure*

$$\delta_{(x)}=\delta_x-\sum_{i=1}^{p'}m_i(x)\delta_{\zeta_i}.$$

Puisque $\delta_{(x)}\in\Lambda_q^d$ ($\delta_{(x)}\cdot m_j=0\ \forall j=1,\ldots,p'$), on peut définir l'espace de Hilbert

$$\mathcal{F}_k=\left\{f(\cdot)=\left\langle\lambda,\Pi\delta_{(\cdot)}\right\rangle_{\overline{\Lambda_q^d}},\lambda\in\overline{\Lambda_q^d}\right\},$$

muni de la norme issue du produit scalaire

$$\langle C_\lambda,C_\mu\rangle_{\mathcal{F}_k}=\langle\lambda,\mu\rangle_{\overline{\Lambda_q^d}},$$

avec $C_\lambda(\cdot)=\langle\lambda,\Pi\delta_{(\cdot)}\rangle_{\overline{\Lambda_q^d}}$ et $C_\mu(\cdot)=\langle\mu,\Pi\delta_{(\cdot)}\rangle_{\overline{\Lambda_q^d}}$.

Remarque 1.3.14 *[111] On a* $p' = \dim \mathbb{P}_q^d = \dfrac{(q+d)!}{q!d!}$.

On peut vérifier par exemple que $\dim \mathbb{P}_1^2 = 3$ et $\dim \mathbb{P}_2^2 = 6$ (exemple 1.3.2).

Théorème 1.3.15 *[146] L'espace* $\mathcal{F}_k$ *est un RKHS de noyau reproduisant*

$$\Psi(x, x') = \langle \Pi\delta_{(x)}, \Pi\delta_{(x')} \rangle_{\overline{\Lambda_q^d}}.$$

Le noyau $\Psi(\cdot,\cdot)$ *est appelé* normalisation *de* $k(\cdot,\cdot)$*. Il s'écrit*

$$\Psi(x, x') = k(x, x') - \sum_{i=1}^{p'} m_i(x) k(\zeta_i, x') - \sum_{i=1}^{p'} m_i(x') k(x, \zeta_i) + \sum_{i,j=1}^{p'} m_i(x) m_j(x') k(\zeta_i, \zeta_j). \quad (1.13)$$

Preuve *Soit* $f \in \mathcal{F}_k$*, il existe* $\lambda \in \overline{\Lambda_q^d}$ *tel que* $f(\cdot) = \left\langle \lambda, \Pi\delta_{(\cdot)} \right\rangle_{\overline{\Lambda_q^d}}$*. Donc*
$\langle f, \Psi(x,\cdot) \rangle_{\mathcal{F}_k} = \left\langle \langle \lambda, \Pi\delta_{(\cdot)} \rangle_{\overline{\Lambda_q^d}}, \langle \Pi\delta_{(x)}, \Pi\delta_{(\cdot)} \rangle_{\overline{\Lambda_q^d}} \right\rangle_{\mathcal{F}_k} = \left\langle \lambda, \Pi\delta_{(x)} \right\rangle_{\overline{\Lambda_q^d}} = f(x)$*. La fonction* $\Psi(\cdot,\cdot)$ *est alors le RKHS de* $\mathcal{F}_k$ *d'après la remarque 1.2.12, et son expression est*

$$\begin{aligned}
\Psi(x, x') &= \langle \Pi\delta_{(x)}, \Pi\delta_{(x')} \rangle_{\overline{\Lambda_q^d}} \\
&= \langle \delta_{(x)}, \delta_{(x')} \rangle_{\Lambda_q^d} \\
&= \left\langle \delta_x - \sum_{i=1}^{p'} m_i(x) \delta_{\zeta_i}, \delta_{x'} - \sum_{i=1}^{p'} m_i(x') \delta_{\zeta_i} \right\rangle_{\Lambda_q^d} \\
&= k(x, x') - \sum_{i=1}^{p'} m_i(x) k(x, \zeta_i) - \sum_{i=1}^{p'} m_i(x') k(\zeta_i, x') + \sum_{i,j=1}^{p'} m_i(x) m_j(x') k(\zeta_i, \zeta_j). \qquad \square
\end{aligned}$$

On définit alors l'espace d'hypothèses

$$\mathcal{C}_k = \mathbb{P}_q^d \oplus \mathcal{F}_k. \quad (1.14)$$

La somme est directe car si $f \in \mathbb{P}_q^d \cap \mathcal{F}_k, \exists \lambda \in \overline{\Lambda_q^d}, f(\cdot) = \langle \lambda, \Pi\delta_{(\cdot)} \rangle_{\overline{\Lambda_q^d}}$. Or on vérifie facilement que $\delta_{(\zeta_i)} = 0 \ \forall i = 1, \ldots, p'$, donc $f(\zeta_i) = 0 \ \forall i = 1, \ldots, p'$, et puisque $f \in \mathbb{P}_q^d$, $f = 0$.

Remarque 1.3.16 *[111, 146]*
- *Si* $q = -1$*, alors* $k(\cdot,\cdot)$ *est le noyau reproduisant de l'espace* $\mathcal{C}_k$ *;*
- *on peut montrer que l'espace* $\mathcal{C}_k$ *et la norme sur* $\mathcal{F}_k$ *sont indépendants de l'ensemble unisolvant* Ξ *choisi dans la définition 1.3.13.*

En utilisant la base de $\mathbb{P}_q^d$ construite à la définition 1.3.13, on peut définir un opérateur de projection $\Pi_{\mathbb{P}_q^d} : \mathcal{C}_k \to \mathbb{P}_q^d$ en écrivant [146]

$$\Pi_{\mathbb{P}_q^d}(f)(x) = \sum_{j=1}^{p'} m_j(x) f(\zeta_j) \quad \forall f \in \mathcal{C}_k.$$

Puisque $f - \Pi_{\mathbb{P}_q^d}(f)$ s'annule sur Ξ, l'opérateur $\Pi_{\mathcal{F}_k} = \mathrm{Id} - \Pi_{\mathbb{P}_q^d}$ projette les fonctions de $\mathcal{C}_k$ sur $\mathcal{F}_k$. On a donc

$$\forall f \in \mathcal{C}_k, \quad f = f_{\mathbb{P}_q^d} + f_{\mathcal{F}_k} \in \mathbb{P}_q^d \oplus \mathcal{F}_k \iff f_{\mathbb{P}_q^d} = \Pi_{\mathbb{P}_q^d}(f) \text{ et } f_{\mathcal{F}_k} = \Pi_{\mathcal{F}_k}(f). \quad (1.15)$$

1.3.3 Apprentissage

Énonçons maintenant l'analogue du théorème 1.2.26. Dans la suite, l'ensemble d'apprentissage sera supposé unisolvant : cette condition, visant à assurer l'unicité de la solution du problème de minimisation, n'est pas très contraignante en pratique [146] (revoir l'exemple 1.3.2).

Théorème 1.3.17 *(du représentant, cas quadratique, avec un noyau conditionnellement semi-défini positif) Soit $k(\cdot,\cdot)$ une fonction conditionnellement semi-définie positive d'ordre q sur $\mathcal{X}\times\mathcal{X}$. Soit $\{(x_i,y_i)\}_{i=1\dots n}$ l'ensemble d'entraînement, avec $x_i\in\mathcal{X}$ et $Y^n={}^t(y_1,\dots,y_n)\in\mathbb{R}^n$. Soit $\Xi=\{\zeta_1,\dots,\zeta_{p'}\}$ un ensemble $\mathbb{P}_q^d$-unisolvant de $\mathcal{X}$, et $m_1,\dots,m_{p'}$ une base de $\mathbb{P}_q^d$ vérifiant $m_i(\zeta_j)=\delta_{ij}$. On suppose que $\Xi\subset\{x_1,\dots,x_n\}$, et on note $M=(m_j(x_i))_{i=1\dots n}^{j=1\dots p'}$. Soit $\mathcal{C}_k$ l'espace d'hypothèses défini comme en (1.14). Alors, pour $\gamma>0$ donné,*

$$\operatorname*{argmin}_{f\in\mathcal{C}_k}\frac{1}{n}\sum_{i=1}^{n}(y_i-f(x_i))^2+\gamma\|\Pi_{\mathcal{F}_k}f\|^2_{\mathcal{F}_k} \tag{1.16}$$

est donné par

$$f(\cdot)=\sum_{j=1}^{p'}\nu_j m_j(\cdot)+\sum_{i=1}^{n}\alpha_i k(x_i,\cdot),$$

avec les vecteurs des coefficients $\nu={}^t(\nu_1,\dots,\nu_{p'})$ et $\alpha={}^t(\alpha_1,\dots,\alpha_n)$, solutions du système

$$\begin{pmatrix}K' & M\\ {}^tM & 0\end{pmatrix}\begin{pmatrix}\alpha\\ \nu\end{pmatrix}=\begin{pmatrix}Y^n\\ 0\end{pmatrix},$$

qui sont donnés par

$$\begin{aligned}\nu &= ({}^tMK'^{-1}M)^{-1}{}^tMK'^{-1}Y^n\,;\\ \alpha &= K'^{-1}\left(I-M({}^tMK'^{-1}M)^{-1}{}^tMK'^{-1}\right)Y^n\,;\\ K' &= K+n\gamma I=(k(x_i,x_j)+n\gamma\delta_{ij})_{1\le i,j\le n}.\end{aligned}$$

Preuve *Quitte à renuméroter, on peut tout d'abord supposer que $x_i=\zeta_i\ \forall i=1,\dots,p'$. Remarquons ensuite que la matrice M est de rang p' car $\Xi\subset\{x_1,\dots,x_n\}$. Les hypothèses du théorème 1.2.26 étant vérifiées, la solution de (1.16) s'écrit*

$$f=\sum_{j=1}^{p'}\eta_j m_j+\sum_{i=1}^{n}\beta_i\Psi_{x_i},$$

avec les vecteurs des coefficients $\eta={}^t(\eta_1,\dots,\eta_{p'})$ et $\beta={}^t(\beta_1,\dots,\beta_n)$, solutions du système

$$\begin{pmatrix}C' & M\\ {}^tM & 0\end{pmatrix}\begin{pmatrix}\beta\\ \eta\end{pmatrix}=\begin{pmatrix}Y^n\\ 0\end{pmatrix},$$

et $C'=(\Psi(x_i,x_j)+n\gamma\delta_{ij})_{1\le i,j\le n}$. D'après (1.13),

$$C'=K'+\left(\textstyle\sum_{s,t=1}^{p'}m_s(x_i)m_t(x_j)k(\zeta_s,\zeta_t)-\sum_{s=1}^{p'}m_s(x_i)k(\zeta_s,x_j)-\sum_{s=1}^{p'}m_s(x_j)k(x_i,\zeta_s)\right)_{1\le i,j\le n} \tag{1.17}$$

et

$$\begin{aligned}
\sum_{i=1}^{n}\beta_i\Psi_{x_i} &= \sum_{i=1}^{n}\beta_i k_{x_i} - \sum_{i=1}^{n}\beta_i\sum_{j=1}^{p'}m_j(x_i)k_{\zeta_j} - \sum_{i=1}^{n}\beta_i\sum_{j=1}^{p'}k(x_i,\zeta_j)m_j + \sum_{i=1}^{n}\beta_i\sum_{j,l=1}^{p'}m_j(x_i)k(\zeta_j,\zeta_k)m_l \\
&= \sum_{i=1}^{n}\beta_i k_{x_i} - \sum_{j=1}^{p'}\underbrace{\sum_{i=1}^{n}\beta_i m_j(x_i)}_{=0\ car\ {}^tM\beta=0}k_{\zeta_j} - \sum_{j=1}^{p'}\sum_{i=1}^{n}\beta_i k(x_i,\zeta_j)m_j + \sum_{j=1}^{p'}\sum_{l=1}^{p'}k(\zeta_l,\zeta_k)\underbrace{\sum_{i=1}^{n}\beta_i m_l(x_i)}_{=0\ car\ {}^tM\beta=0}m_j \\
&= \sum_{i=1}^{n}\beta_i k_{x_i} - \sum_{j=1}^{p'}\sum_{i=1}^{n}\beta_i k(x_i,\zeta_j)m_j.
\end{aligned}$$

Ainsi,

$$\begin{aligned}
f &= \sum_{j=1}^{p'}\left(\eta_j - \sum_{i=1}^{n}\beta_i k(x_i,\zeta_j)\right)m_j + \sum_{i=1}^{n}\beta_i k_{x_i} \\
&:= \sum_{j=1}^{p'}\nu_j m_j + \sum_{i=1}^{n}\alpha_i k_{x_i}. \qquad (1.18)
\end{aligned}$$

Or, d'après (1.17) et (1.18),

$$\begin{aligned}
\big((C'-K')\beta\big)_i &= \sum_{j=1}^{n}\beta_j\left(\sum_{s,t=1}^{p'}m_s(x_i)m_t(x_j)k(\zeta_s,\zeta_t) - \sum_{s=1}^{p'}m_s(x_i)k(\zeta_s,x_j) - \sum_{s=1}^{p'}m_s(x_j)k(x_i,\zeta_s)\right) \\
&= \sum_{s,t=1}^{p'}m_s(x_i)k(\zeta_s,\zeta_t)\underbrace{\sum_{j=1}^{n}\beta_j m_t(x_j)}_{=0} - \sum_{s=1}^{p'}m_s(x_i)\sum_{j=1}^{n}\beta_j k(\zeta_s,x_j) - \sum_{s=1}^{p'}k(x_i,\zeta_s)\underbrace{\sum_{j=1}^{n}\beta_j m_s(x_j)}_{=0} \\
&= \sum_{s=1}^{p'}(\nu_s-\eta_s)m_s(x_i),
\end{aligned}$$

donc

$$\begin{aligned}
C'\beta &= K'\beta + (C'-K')\beta \\
&= K'\beta + M(\nu-\eta).
\end{aligned}$$

Les vecteurs $\alpha = {}^t(\alpha_1,\ldots,\alpha_n) = \beta$ *et* $\nu = {}^t(\nu_1,\ldots,\nu_{p'})$ *sont donc solution du système*

$$\begin{pmatrix} K' & M \\ {}^tM & 0 \end{pmatrix}\begin{pmatrix}\alpha\\ \nu\end{pmatrix} = \begin{pmatrix} Y^n \\ 0\end{pmatrix}.$$

Les valeurs des coefficients α *et* ν *ont été données dans le théorème 1.2.26.* □

Remarque 1.3.18 *Observons les différences avec le cas semi-défini positif (théorème 1.2.26) :*

- *dans le cas conditionnellement semi-défini positif, la partie paramétrique est fixée, c'est une base de* $\mathbb{P}_q^d$*. On perd donc en généralité par rapport au cas semi-défini positif ;*
- *par contre, la classe de fonctions* $k(\cdot,\cdot)$ *utilisable est plus générale, on y gagne donc en généralité ;*

– *enfin, puisque* $n \geq p' = (q+d)!/(q!d!)$*, le nombre de données d'apprentissage est généralement plus conséquent que dans le cas semi-défini positif, où* $n \geq p$ *et* p *est relativement petit.*

On trouvera parfois dans la littérature [111] que pour un noyau conditionnellement semi-défini positif, la solution est obtenue par minimisation d'une semi-norme (remarque 1.2.4). La norme apparaissant dans (1.16) correspond en effet à une semi-norme sur $\mathcal{C}_k$, $\|f\|_{\mathcal{C}_k} = \|\Pi_{\mathcal{F}_k} f\|_{\mathcal{F}_k}$, dont le noyau est $\mathbb{P}_q^d$.

Voici finalement, sans démonstration, l'analogue du théorème précédent dans le cas de l'interpolation.

Théorème 1.3.19 *(du représentant, cas quadratique, avec un noyau conditionnellement semi-défini positif, pour l'interpolation) Soit* $k(\cdot,\cdot)$ *une fonction conditionnellement semi-définie positive d'ordre* q *sur* $\mathcal{X} \times \mathcal{X}$*. Soit* $\{(x_i, y_i)\}_{i=1\ldots n}$ *l'ensemble d'entraînement, avec* $x_i \in \mathcal{X}$ *et* $Y^n = {}^t(y_1, \ldots, y_n) \in \mathbb{R}^n$*. Soit* $\Xi = \{\zeta_1, \ldots, \zeta_{p'}\}$ *un ensemble* $\mathbb{P}_q^d$*-unisolvant de* $\mathcal{X}$*, et* $m_1, \ldots, m_{p'}$ *une base de* $\mathbb{P}_q^d$ *vérifiant* $m_i(\zeta_j) = \delta_{ij}$*. On suppose que* $\Xi \subset \{x_1, \ldots, x_n\}$*, et on note* $M = (m_j(x_i))_{i=1\ldots n}^{j=1\ldots p'}$*. Soit* $\mathcal{C}_k$ *l'espace d'hypothèses défini comme en (1.14). On suppose que la matrice* $K = (k(x_i, x_j))_{1 \leq i,j \leq n}$ *est inversible. Alors,*

$$\begin{cases} \underset{f \in \mathcal{C}_k}{\operatorname{argmin}} \|\Pi_{\mathcal{F}_k} f\|_{\mathcal{F}_k}^2 \\ f(x_i) = y_i \quad \forall i, \end{cases}$$

est donné par

$$f(\cdot) = \sum_{j=1}^{p'} \nu_j m_j(\cdot) + \sum_{i=1}^{n} \alpha_i k(x_i, \cdot),$$

avec les vecteurs des coefficients $\nu = {}^t(\nu_1, \ldots, \nu_{m'})$ *et* $\alpha = {}^t(\alpha_1, \ldots, \alpha_n)$*, solutions du système*

$$\begin{pmatrix} K & M \\ {}^tM & 0 \end{pmatrix} \begin{pmatrix} \alpha \\ \nu \end{pmatrix} = \begin{pmatrix} Y^n \\ 0 \end{pmatrix},$$

qui sont donnés par

$$\begin{aligned} \nu &= ({}^tMK^{-1}M)^{-1} {}^tMK^{-1}Y^n\,; \\ \alpha &= K^{-1}\left(I - M({}^tMK^{-1}M)^{-1} {}^tMK^{-1}\right) Y^n. \end{aligned}$$

Remarque 1.3.20 *Dans le théorème 1.3.17 et le théorème 1.3.19, la base de polynômes* $\{m_1(\cdot), \ldots, m_{p'}(\cdot)\}$ *de* $\mathbb{P}_q^d$ *peut être choisie quelconque du moment que l'ensemble d'observation* $\{x_1, \ldots, x_n\}$ *est* $\mathbb{P}_q^d$*-unisolvant. En effet, on sait [86] que*

$$M({}^tMK^{-1}M)^{-1} {}^tMK^{-1}$$

est le projecteur orthogonal sur l'espace vectoriel engendré par les colonnes de M*, pour le produit scalaire sur* $\mathbb{R}^n$

$$\langle x, y \rangle_{K^{-1}} = {}^t x K^{-1} y.$$

De plus, l'ensemble $\{x_1, \ldots, x_n\}$ *étant* $\mathbb{P}_q^d$*-unisolvant, le vecteur des coefficients* ν *est uniquement déterminé par les valeurs prises par la fonction* $\sum_{j=1}^{p'} \nu_j m_j(\cdot)$ *en* $x_1, \ldots, x_n$*, rassemblées dans le vecteur*

$$\left(\sum_{j=1}^{p'} \nu_j m_j(x_i) \right)_{i=1,\ldots,n} = M({}^tMK^{-1}M)^{-1}{}^tMK^{-1}Y^n. \tag{1.19}$$

La conclusion s'obtient en remarquant que l'espace vectoriel engendré par les colonnes de M *ne dépend pas de la base de polynômes choisie, le vecteur (1.19) est donc indépendant de la base de* $\mathbb{P}_q^d$ *choisie et il en est également ainsi de la fonction* $\sum_{j=1}^{p'} \nu_j m_j(\cdot)$*. Un argument similaire assure que l'expression du vecteur de coefficients* α *ne dépend pas non plus de* M*.*

Les théorèmes précédents donnent les solutions obtenues par la méthode du krigeage intrinsèque (§ 1.4.1 et annexe E) ou des splines plaque mince (§ 1.4.2), qui sont donc des méthodes équivalentes. La différence est que dans le cas des splines, l'opérateur (de régularisation) $\Pi_{\mathcal{F}_k}$ est défini *a priori* et correspond à des contraintes physiques de régularité de la fonction f (la fonction $k(\cdot,\cdot)$ est alors déterminée par l'opérateur de régularisation). Dans le krigeage, on part de la fonction $k(\cdot,\cdot)$ qui détermine la régularité du processus utilisé pour la modélisation ; le terme de pénalisation $n\gamma$ correspond à la variance du bruit de mesure. Un problème vu d'une des deux perspectives peut être cependant difficile à interpréter de l'autre [26, 31, 158].

Pour beaucoup plus de détails concernant les espaces précités, on pourra consulter [106, 107, 111, 112, 146, 147, 187]. Une borne de l'erreur commise dans le cas de l'interpolation est donnée dans [99, 108, 111, 190].

1.4 Exemples de méthodes à noyaux

1.4.1 Krigeage

Le krigeage est une méthode de modélisation issue de la géostatitique, introduite par l'ingénieur minier sud-africain D.G. Krige dans les années 1950 et formalisée ensuite par G. Matheron à l'École des Mines de Paris [59]. Nous présentons ici brièvement le krigeage et ses liens avec la régression régularisée. Une présentation plus détaillée de la méthode sera faite au chapitre 2.

Le krigeage est une méthode apportant un point de vue probabiliste sur l'apprentissage dans les RKHS. La relation entrées-sortie est modélisée par

$$f(x) = {}^tm(x)\beta + Z(x) \tag{1.20}$$

dans le cas non bruité ($f(x_i) = y_i \ \forall i = 1, \ldots, n$), et

$$f(x) = {}^tm(x)\beta + Z(x) + \varepsilon(x) \tag{1.21}$$

dans le cas bruité, avec $x \in \mathcal{X} \subset \mathbb{R}^d$, $m(x)$ un vecteur de fonctions de base connues, β un vecteur de coefficients inconnus, $Z(\cdot)$ un processus gaussien de moyenne nulle et fonction de covariance $k(\cdot,\cdot)$ connue, et $\varepsilon(\cdot)$ une famille de variables aléatoires i.i.d. de loi $\mathcal{N}(0, \sigma_\varepsilon^2)$, indépendantes de $Z(\cdot)$, de variance σ_ε^2 connue.

Définition 1.4.1 *Le prédicteur de krigeage en* x *est le meilleur prédicteur linéaire sans biais de* $f(x)$ *obtenu à partir des observations* $\{(x_i, y_i), i = 1, \ldots, n\}$*.*

Avant d'établir le lien avec la régression régularisée, observons la relation entre les processus gaussiens de moyenne nulle et les RKHS. En effet, en tant que fonction semi-définie positive, $k(\cdot,\cdot)$ peut être vue comme une fonction de covariance (théorème 2.1.11).

Proposition 1.4.2 *[181] Si $k(\cdot,\cdot)$ est une fonction semi-définie positive, on peut définir une famille $\{Z(x), x \in \mathcal{X}\}$ de variables aléatoires gaussiennes centrées de fonction de covariance $k(\cdot,\cdot)$,* i.e. $\mathbb{E}\left[Z(x)Z(x')\right] = k(x,x')$.

Le théorème suivant complète la proposition 1.4.2.

Théorème 1.4.3 *(Théorème de l'isomorphisme isométrique, Parzen)[182] À tout RKHS $\mathcal{H}$ de noyau reproduisant $k(\cdot,\cdot)$ correspond un processus gaussien $\{Z(x), x \in \mathcal{X}\}$ de moyenne nulle et fonction de covariance $k(\cdot,\cdot)$. Il y a un isomorphisme isométrique entre $\mathcal{Z}$, l'espace de Hilbert engendré par le processus aléatoire $Z(\cdot)$, et $\mathcal{H}$, où la variable aléatoire $Z(x) \in \mathcal{Z}$ correspond au représentant $k_x \in \mathcal{H}$.*

Le produit scalaire de $\mathcal{Z}$ est donc conservé dans $\mathcal{H}$ d'après la proposition 1.4.2. Reprenons les hypothèses et notations du théorème 1.2.26 et du théorème 1.2.27.

Théorème 1.4.4

- *Le prédicteur de krigeage sans bruit (1.20) correspond à*

$$\begin{cases} \underset{f=f_{\mathcal{M}}+f_{\mathcal{H}}\in\mathcal{M}+\mathcal{H}}{\operatorname{argmin}} \|f_{\mathcal{H}}\|_{\mathcal{H}}^2 \\ f(x_i) = y_i \quad \forall i. \end{cases} \tag{1.22}$$

- *Le prédicteur de krigeage avec bruit de mesure (1.21) correspond à*

$$\underset{f=f_{\mathcal{M}}+f_{\mathcal{H}}\in\mathcal{M}+\mathcal{H}}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n}\left(y_i - f(x_i)\right)^2 + \frac{\sigma_\varepsilon^2}{n}\|f\|_{\mathcal{H}}^2. \tag{1.23}$$

Preuve *Les expressions des prédicteurs de krigeage sans et avec bruit de mesure seront rappelées au chapitre 2, où l'on constatera qu'elles correspondent aux solutions respectives de (1.22) et (1.23), données respectivement par le théorème 1.2.27 et le théorème 1.2.26 avec $\gamma = \sigma_\varepsilon^2/n$.* □

On verra plus loin que le krigeage bruité et à moyenne constante $m(x) = 1$ est un cas particulier de la SVR (Support Vector Regression, § 1.4.4), avec une fonction de coût quadratique. La transformation Φ correspond à un processus aléatoire de moyenne nulle $Z(\cdot)$, et l'espace des caractéristiques est $\mathcal{Z}$ muni du produit scalaire $\langle Z_1, Z_2\rangle = \mathbb{E}\left[Z_1 Z_2\right]$. Le krigeage bruité à moyenne quelconque est un cas particulier de la SVR semi-paramétrique évoquée à la remarque 1.4.25.

Remarque 1.4.5 *[181, 182] Soit $Z(\cdot)$ un processus gaussien de moyenne nulle, dont la fonction de covariance $k(x,x') = \mathbb{E}[Z(x)Z(x')]$ vérifie les conditions de Mercer-Hilbert-Schmidt (théorème 1.2.21). Si le noyau $k(\cdot,\cdot)$ est de dimension infinie (théorème 1.2.21), alors les trajectoires de Z n'appartiennent pas, avec probabilité 1, à $\mathcal{H}$. On peut donc dans ce cas se poser la question de la pertinence de la modélisation par krigeage, car le prédicteur obtenu appartient, lui, à $\mathcal{H}$. Pour plus de détails concernant les trajectoires de processus aléatoires à valeurs dans un RKHS, on pourra consulter [109].*

Krigeage intrinsèque

Le krigeage intrinsèque correspond aux problèmes (1.20) et (1.21), à la différence que :

– la fonction $k(\cdot,\cdot)$ est une *covariance généralisée* d'ordre q pour un certain q (ou de façon équivalente, une fonction conditionnellement semi-définie positive d'ordre q), ce qui signifie que

$$\operatorname{var}\left[\sum_{i=1}^{n}\lambda_i f(x_i)\right] = \sum_{i,j=1}^{n}\lambda_i\lambda_j k(x_i,x_j)$$

n'est définie que pour $\sum_{i=1}^{n}\lambda_i\delta_{x_i} \in \Lambda_q^d$ (car une variance est positive) ;
– le vecteur ${}^t m(x)$ est constitué de la base canonique de $\mathbb{P}_q^d$.

Définition 1.4.6 *Le prédicteur de krigeage intrinsèque en x est le meilleur prédicteur linéaire de $f(x)$ obtenu à partir des observations $\{(x_i,y_i), i=1,\ldots,n\}$.*

Du fait que la covariance est généralisée, la combinaison linéaire des observations formant le prédicteur $\sum_{i=1}^{n}\lambda_i f(x_i)$ n'est pas quelconque : en effet, pour pouvoir évaluer

$$\mathbb{E}\left[f(x)-\sum_{i=1}^{n}\lambda_i f(x_i)\right]^2 = \operatorname{var}\left[f(x)-\sum_{i=1}^{n}\lambda_i f(x_i)\right],$$

il faut que $\delta_x - \sum_{i=1}^{n}\lambda_i\delta_{x_i} \in \Lambda_q^d$ afin de garantir une variance positive. En utilisant les résultats du paragraphe 1.3, on obtient le théorème suivant.

Théorème 1.4.7
– *Le prédicteur de krigeage intrinsèque non bruité correspond à*

$$\begin{cases} \underset{f\in\mathcal{C}_k}{\operatorname{argmin}} \|\Pi_{\mathcal{F}_k} f\|_{\mathcal{F}_k}^2 \\ f(x_i) = y_i \quad \forall i. \end{cases} \tag{1.24}$$

– *Le prédicteur de krigeage intrinsèque avec bruit de mesure correspond à*

$$\underset{f\in\mathcal{C}_k}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n}(y_i-f(x_i))^2 + \frac{\sigma_\varepsilon^2}{n}\|\Pi_{\mathcal{F}_k} f\|_{\mathcal{F}_k}^2. \tag{1.25}$$

Preuve *Les expressions des prédicteurs de krigeage intrinsèque sans et avec bruit de mesure seront données dans l'annexe E, où l'on constatera qu'elles correspondent aux solutions respectives de (1.24) et (1.25), données respectivement par le théorème 1.3.19 et le théorème 1.3.17 (en tenant compte de la remarque 1.3.20), avec $\gamma = \sigma_\varepsilon^2/n$.* □

1.4.2 Splines plaque mince

Après avoir rappelé la définition des splines polynômiales naturelles univariées, qui sont tout simplement des fonctions polynômiales par morceaux sur un intervalle de $\mathbb{R}$, le *problème spécial* de régression régularisée dans $\mathbb{R}$ est présenté, dont la solution est une spline polynômiale naturelle univariée.

Définition 1.4.8 *[181] Soit $[a,b]$ un intervalle de $\mathbb{R}$, (éventuellement $]-\infty,b]$, $[a,\infty[$ ou $\mathbb{R}$ tout entier), et $x_1,\ldots,x_n$ des réels appelés* noeuds *vérifiant $a<x_1<\cdots<x_n<b$. On appelle* spline naturelle *de degré l une fonction réelle $s(\cdot)$ sur $[a,b]$ ayant les propriétés suivantes :*
– *$s(\cdot)\in\mathbb{P}_{l-1}$ sur $[a,x_1]$ et $[x_n,b]$;*

– $s(\cdot) \in \mathbb{P}_{2l-1}$ *sur* $[x_i, x_{i+1}]$, $i = 1, \ldots, n-1$;
– $s(\cdot) \in \mathcal{C}^{2l-2}(\mathbb{R})$,
où $\mathbb{P}_l$ *désigne l'ensemble de polynômes de degré au plus* l *sur* $\mathbb{R}$, *et* $\mathcal{C}^l$ *l'ensemble des fonctions* l *fois continûment dérivables.*

On a la propriété d'unicité suivante.

Proposition 1.4.9 *[181] Il existe une unique spline de degré* n *qui est un interpolateur pour l'ensemble d'entraînement* $\{(x_i, y_i), i = 1, \ldots, n\}$.

L'intérêt de l'interpolation par spline par rapport à l'interpolatation polynômiale est que l'on évite le *phénomène de Runge* (l'équivalent du *phénomène de Gibbs* pour les fonctions trigonométriques [204]) : l'utilisation d'un polynôme interpolateur peut donner lieu à de fortes oscillations, il peut même arriver que l'erreur d'approximation tende vers l'infini alors que l'ensemble d'entraînement devient de plus en plus grand (voir la figure 2.5 page 71 pour se faire une idée du phénomène).

Afin de montrer le lien existant entre les splines et les RKHS, nous nous restreignons dans la suite à $\mathcal{X} = [0, 1]$ pour simplifier la présentation. Rappelons tout d'abord une formule classique.

Proposition 1.4.10 *(Théorème de Taylor avec reste intégral)[181] Soit* f *une fonction à valeurs réelles définie sur* $[0, 1]$, q *fois continûment dérivable, telle que* $f^{(q+1)} \in \mathcal{L}^2[0, 1]$. *Alors*

$$f(t) = \sum_{i=0}^{q} \frac{t^i}{i!} f^{(i)}(0) + \int_0^1 \frac{(t-u)_+^q}{q!} f^{(q+1)}(u) \mathrm{d}u, \tag{1.26}$$

avec $(x)_+^q = x^q$ *si* $x \geq 0$, $(x)_+^q = 0$ *sinon.*

On peut alors définir un RKHS de fonctions définies sur $[0, 1]$ comme la somme de deux RKHS, chacun associé à un des termes de la somme dans (1.26).

Pour le premier terme, notons $m_i(t) = t^i/i!$, $i = 0, \ldots, q$, alors les m_i forment une base de l'espace vectoriel $\mathbb{P}_q$.

Proposition 1.4.11 *[181] L'espace* $\mathbb{P}_q$, *muni de la norme*

$$\|f\|_0^2 = \sum_{i=0}^{q} \left[f^{(i)}(0) \right]^2,$$

est un espace de Hilbert dont $m_0, \ldots, m_q$ *est une base orthonormale, et aussi un RKHS de noyau reproduisant*

$$k_0(s, t) = \sum_{i=0}^{q} m_i(s) m_i(t).$$

On notera $\mathcal{H}_0$ *ce RKHS.*

Pour le RKHS associé au deuxième terme de (1.26), notons $\mathcal{B}_{q+1}$ la classe de fonctions satisfaisant les hypothèses de la proposition 1.4.10 et les conditions de bord $f^{(i)}(0) = 0, i = 0, 1, \ldots, q$. Si $f \in \mathcal{B}_{q+1}$ alors, d'après la proposition 1.4.10,

$$\begin{aligned} f(t) &= \int_0^1 \frac{(t-u)_+^q}{q!} f^{(q+1)}(u) \, \mathrm{d}u \\ &= \int_0^1 G_{q+1}(t, u) f^{(q+1)}(u) \, \mathrm{d}u, \end{aligned}$$

avec $G_{q+1}(t, u) = \dfrac{(t-u)_+^q}{q!}$. On peut alors définir un RKHS inclus dans $\mathcal{B}_{q+1}$.

Proposition 1.4.12 *[181] Soit*

$$W_{q+1}^0 = \left\{ f \in \mathcal{B}_{q+1}, f^{(i)} \text{ absolument continue pour } i = 0, 1, \ldots, q, f^{(q+1)} \in \mathcal{L}^2 \right\}.$$

W_{q+1}^0 *est un espace de Hilbert de norme* $\|f\|_1^2 = \int_0^1 \left(f^{(q+1)}(t)\right)^2 \mathrm{d}t$, *et un RKHS de noyau reproduisant*

$$k_1(s,t) = \int_0^1 G_{q+1}(t,u) G_{q+1}(s,u)\, \mathrm{d}u.$$

On notera $\mathcal{H}_1$ *ce RKHS.*

On peut maintenant définir l'espace d'hypothèses W_{q+1} (qui est aussi un *espace de Sobolev* [7], car notamment la norme fait intervenir des dérivées).

Théorème 1.4.13 *[181] Soit*

$$W_{q+1} = \left\{ f \ : [0,1] \to \mathbb{R}, f^{(i)} \text{ absolument continue pour } i = 0, 1, \ldots, q, f^{(q+1)} \in \mathcal{L}^2 \right\}.$$

Muni de la norme $\|\cdot\|_{W_{q+1}}^2 = \|\cdot\|_0^2 + \|\cdot\|_1^2$, W_{q+1} *est un espace de Hilbert et*

$$W_{q+1} = \mathcal{H}_0 \overset{\perp}{\oplus} \mathcal{H}_1.$$

De plus, W_{q+1} *est un RKHS de noyau reproduisant*

$$k(s,t) = k_0(s,t) + k_1(s,t).$$

Remarque 1.4.14 *[181]*

- *On peut montrer en toute généralité que le noyau reproduisant de la somme directe de deux sous-espaces orthogonaux est la somme de leurs noyaux reproduisants ;*
- *la norme sur* $\mathcal{H}_1$ $\|f\|_1^2 = \int_0^1 \left(f^{(q+1)}(t)\right)^2 \mathrm{d}t$ *peut se réécrire* $\|f\|_1^2 = \|P_1 f\|_{W_{q+1}}^2 = \|P_1 f\|_{\mathcal{H}_1}^2$, *où* P_1 *désigne la projection orthogonale sur* $\mathcal{H}_1$ *dans* W_{q+1}. *C'est cette norme (semi-norme sur* W_{q+1}*) qui sera utilisée pour la pénalisation.*

On peut maintenant énoncer le problème spécial de régression régularisée par des splines (il existe aussi un *problème général*, pour lequel nous renvoyons à [181]). On recherche

$$\underset{f \in W_{q+1}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - f(x_i)\right)^2 + \gamma \|P_1 f\|_{\mathcal{H}_1}^2. \tag{1.27}$$

On est ramené dans le cadre du § 1.3. Avec les notations du théorème 1.3.17, cela revient à prendre $\mathcal{X} = [0,1]$, $\mathbb{P}_q^1 = \mathcal{H}_0$, $\mathcal{F}_k = \mathcal{H}_1, \Pi_{\mathcal{F}_k} = P_1$. Si l'on suppose la matrice $M = (m_j(x_i))_{i=1,\ldots,n}^{j=0,\ldots,q}$ de rang $q' = q+1$, alors le théorème du représentant avec la remarque 1.3.20 donne l'expression de la solution.

Remarque 1.4.15 *On peut aussi écrire le problème d'interpolation associé comme dans le théorème 1.3.19, qui donne alors l'expression de la solution.*

La solution du problème (1.27), ou de son équivalent dans le cas de l'interpolation, s'appelle une *spline plaque mince (thin plate spline)*. Physiquement, on peut voir la solution comme une plaque de métal mince et flexible ajustée de façon à être proche des points d'observation tout en étant de courbure minimale (la norme peut être vue comme une énergie de flexion). On a finalement le résultat annoncé.

Théorème 1.4.16 *[181] Si la matrice M est de rang $q+1$, la solution du problème spécial (1.27) est une spline naturelle de degré $q+2$.*

Pour la définition des splines plaque mince sur $\mathbb{R}^d$, ainsi que le choix de la constante de pénalisation γ à l'aide de la validation croisée, on consultera [181, 182]. On pourra aussi voir [136].

La méthode des splines plaque mince est donc équivalente au krigeage intrinsèque. Dans le cas des splines, on commence par se donner un opérateur de régularisation P_1, alors que dans le krigeage on part d'une fonction de covariance généralisée pour construire l'espace d'hypothèses. Passer d'un cadre à l'autre n'est pas toujours évident [158].

1.4.3 Ondelettes

Les ondelettes, introduites par Haar au début du XX$^{\text{e}}$ siècle afin de trouver une base de fonctions pour laquelle la série de Fourier ne diverge jamais, ont été utilisées ensuite pour faire de l'analyse de signal et du codage. Une application célèbre est le format de compression de données bien connu JPEG 2000.

L'idée est de décomposer un signal en fonctions translatées et mises à l'échelle d'une fonction de base $\psi(\cdot)$ (continue, intégrable, de carré intégrable) appelée *ondelette mère*. La courbe de l'ondelette mère peut être vue comme une brêve oscillation, semblable aux enregistrements d'un moniteur cardiaque. Dans le cas où la variable t est réelle, on demande souvent à la fonction ψ de vérifier les conditions de moments

$$\int_{-\infty}^{+\infty} t^i \psi(t)\,\mathrm{d}t = 0, \qquad i = 0, 1, \ldots, l-1,$$

pour un certain l. L'ondelette mère génère ensuite une famille d'ondelettes

$$\psi_{(a,b)}(t) = \frac{1}{\sqrt{a}}\,\psi\left(\frac{t-b}{a}\right), \qquad a > 0, \quad b \in \mathbb{R},$$

par changement d'échelle et translation en temps. La fonction ψ étant en général centrée en 0, les fonctions $\psi_{(a,b)}$ sont centrées en b et d'échelle a par rapport à ψ. On dit alors que l'on fait de l'analyse *temps-échelle*, car en faisant un balayage des grandes aux petites échelles on arrive à une représentation de plus en plus précise d'un signal donné. Des bases orthonormales d'ondelettes seront de la forme

$$\psi_{i,j}(t) = 2^{i/2}\psi(2^i t - j), \qquad i, j \in \mathbb{Z}.$$

Notons qu'il est également possible de faire de l'analyse *temps-fréquence* en utilisant des bases de Fourier (voir [76] pour une mise en perspective historique de la théorie des ondelettes).

La régression à base d'ondelettes consiste à utiliser une famille orthonormale complète, puis à ne garder que les termes dont le coefficient est suffisamment grand, ce que l'on appelle *contraction (shrinkage)*. On obtient alors une *représentation creuse (sparse représentation)* de la solution, ce qui est utile en pratique pour stocker un signal de façon économique (par exemple, les empreintes digitales des américains ont été codées puis compressées sur ce principe, puis mises en bibliothèque par le FBI). L'intérêt des ondelettes est de pouvoir représenter à la fois des fonctions régulières et des fonction ayant des « pics » localisés.

Afin de savoir si la famille de fonctions considérée permet de construire un RKHS, on peut utiliser la théorie des *frames*.

Définition 1.4.17 *[135]. Soit $\mathcal{H}$ un espace de Hilbert. Un ensemble de vecteurs $\{\phi_i\}_{i\in\Gamma}$ est une* frame*, ou* structure oblique *de $\mathcal{H}$ s'il existe deux constantes A et B, $0 < A \leq B < \infty$, telles que*

$$A\|f\|_{\mathcal{H}}^2 \leq \sum_{i\in\Gamma} |\langle f, \phi_i\rangle_{\mathcal{H}}|^2 \leq B\|f\|_{\mathcal{H}}^2 \qquad \forall f \in \mathcal{H}.$$

Si $A = B$, la frame est dite tendue (tight).
L'opérateur de frame U est défini par

$$\begin{array}{rccl} U & : & \mathcal{H} & \longrightarrow & l^2 \\ & & f & \longrightarrow & \{\langle f, \phi_i\rangle_{\mathcal{H}}\}_{i\in\Gamma}\,. \end{array}$$

Pour pouvoir écrire la fonction f comme combinaison linéaire de frames, on définit les frames duales à partir de l'opérateur U^* adjoint de U (qui existe et est unique puisque $\mathcal{H}$ est un espace de Hilbert et U est continu),

$$\begin{array}{rccl} U^* & : & l^2 & \longrightarrow & \mathcal{H} \\ & & \{c_i\}_{i\in\Gamma} & \longrightarrow & \sum_{i\in\Gamma} c_i\phi_i. \end{array}$$

Si $\{\phi_i\}_{i\in\Gamma}$ est une frame sur $\mathcal{H}$, alors l'opérateur U^*U est inversible [38] et permet de définir la frame duale.

Définition 1.4.18 *[135] On appelle* frame duale *de ϕ_i la frame*

$$\overline{\phi_i} = (U^*U)^{-1}\phi_i.$$

On a alors la relation suivante :

$$\frac{1}{B}\|f\|_{\mathcal{H}}^2 \leq \sum_{i\in\Gamma} \left|\langle f, \overline{\phi_i}\rangle_{\mathcal{H}}\right|^2 \leq \frac{1}{A}\|f\|_{\mathcal{H}}^2 \qquad \forall f \in \mathcal{H},$$

avec les coefficients A et B de la définition 1.4.17.
Si la frame est tendue, alors $\overline{\phi_i} = \frac{1}{A}\phi_i$.

Théorème 1.4.19 *[135] Soit ϕ_i une frame de $\mathcal{H}$, et $\overline{\phi_i}$ sa frame duale. Pour tout f de $\mathcal{H}$,*

$$f = \sum_{i\in\Gamma} \langle f, \overline{\phi_i}\rangle_{\mathcal{H}}\phi_i = \sum_{i\in\Gamma} \langle f, \phi_i\rangle_{\mathcal{H}}\overline{\phi_i}. \tag{1.28}$$

Remarque 1.4.20 *[135]*

- *Une base orthonormale d'un espace de Hilbert séparable $\mathcal{H}$ est un cas particulier de frame où $A = B = 1$;*
- *on n'a pas précisé si l'ensemble d'indices Γ est fini ou infini : si Γ est infini, il faut utiliser le théorème 1.4.21 et si Γ est fini, le théorème 1.4.22 pour savoir si l'espace engendré par les $\{\phi_i\}_{i\in\Gamma}$ est un RKHS. Dans ce dernier cas, les hypothèses sont plus simples car un ensemble fini de fonctions $\{\phi_i\}$ est une frame de l'espace qu'il engendre quel que soit le produit scalaire considéré (ce qui est faux en dimension infinie).*

Le théorème suivant permet d'utiliser les frames pour savoir si un espace de Hilbert est un RKHS et obtenir l'expression du noyau reproduisant.

Théorème 1.4.21 *[135] Soit $\mathcal{H}$ un espace de Hilbert de fonctions réelles définies sur un ensemble $\mathcal{X} \subset \mathbb{R}^d$, et $\{\phi_i\}_{i\in\Gamma}$ une frame de $\mathcal{H}$. Si*

$$\forall x \in \mathcal{X}, \qquad \left\| \sum_{i\in\Gamma} \overline{\phi_i}(\cdot)\phi_i(x) \right\|_{\mathcal{H}} < \infty,$$

alors $\mathcal{H}$ est un RKHS, de noyau reproduisant

$$k(x,x') = \sum_{i\in\Gamma} \overline{\phi_i}(x)\phi_i(x').$$

Preuve *Soit $x \in \mathcal{X}$. D'après (1.28), pour tout $f \in \mathcal{H}$,*

$$\begin{aligned} f(x) &= \sum_{i\in\Gamma} \left\langle f(\cdot), \overline{\phi_i}(\cdot) \right\rangle_{\mathcal{H}} \phi_i(x) \\ &= \left\langle f(\cdot), \sum_{i\in\Gamma} \overline{\phi_i}(\cdot)\phi_i(x) \right\rangle_{\mathcal{H}}. \end{aligned} \tag{1.29}$$

Si $\mathcal{H}$ est un RKHS alors, par unicité du noyau reproduisant (théorème de Moore-Aronszajn), on a bien $k(x,x') = \sum_{i\in\Gamma} \overline{\phi_i}(x)\phi_i(x')$. Or, d'après (1.29),

$$\begin{aligned} |f(x)| &= \left| \left\langle f(\cdot), \sum_{i\in\Gamma} \overline{\phi_i}(\cdot)\phi_i(x) \right\rangle_{\mathcal{H}} \right| \\ &\leq \|f\|_{\mathcal{H}} \left\| \sum_{i\in\Gamma} \overline{\phi_i}(\cdot)\phi_i(x) \right\|_{\mathcal{H}} \\ &= M_x \|f\|_{\mathcal{H}}. \end{aligned}$$

Ainsi, puisque $M_x < \infty$ par hypothèse, $\mathcal{H}$ est un RKHS. □

Une conséquence de ce théorème est le résultat suivant, qui permet de construire un RKHS à partir d'un nombre fini de frames.

Théorème 1.4.22 *[135] Soit $\{\phi_i\}_{i=1,\dots,N}$ un ensemble fini de fonctions non identiquement nulles, appartenant à un espace de Hilbert $\mathcal{F}$ de fonctions réelles définies sur un ensemble $\mathcal{X} \subset \mathbb{R}^d$, telles que*

$$\exists M, \sup_{i=1,\dots,N} \sup_{x\in\mathcal{X}} |\phi_i(x)| \leq M.$$

On définit l'ensemble de fonctions

$$\mathcal{H} = \left\{ f = \sum_{i=1}^{N} a_i\phi_i, a_i \in \mathbb{R} \right\}.$$

Alors $(\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{F}})$ est un RKHS de noyau reproduisant

$$k(x,x') = \sum_{i=1}^{N} \overline{\phi_i}(x)\phi_i(x'). \tag{1.30}$$

Remarque 1.4.23 *Ce critère est plus facile à vérifier en pratique que la condition de Mercer (définition 1.2.17). De plus, le théorème 1.2.21 indique qu'un noyau de Mercer de dimension finie N admet l'extension*

$$k(x, x') = \sum_{i=1}^{N} \lambda_i \Phi_i(x) \Phi_i(x'), \tag{1.31}$$

avec $\lambda_i \geq 0$, et $\{\Phi_i(\cdot)\}_i$ une famille orthogonale. Les conditions du théorème 1.4.22 concernant les frames sont moins restrictives. Notons que si les frames sont tendues ou forment une base orthogonale, on obtient le même type de développement du noyau : l'équation (1.30) se ramène à (1.31) car dans ce cas chaque frame est proportionnelle à son dual (définition 1.4.18 et remarque 1.4.20).

Exemple 1.4.24 *[135]*

- *Tout ensemble fini de fonctions bornées de carré intégrable sur $\mathcal{X}$, muni du produit scalaire de $\mathrm{L}^2(\mathcal{X})$, engendre un RKHS. Par exemple, la famille*

$$\left\{\phi_i(t) = t\mathrm{e}^{-(t-i)^2}\right\}_{i=1\ldots,N};$$

- *l'espace engendré par la famille d'ondelettes*

$$\left\{\psi_{ij}(t) = \frac{1}{a^{i/2}}\psi\left(\frac{t - bj}{a^i}\right)\right\}_{j=1\ldots,N}, \qquad a, b \in \mathbb{R}, i \in \mathbb{Z},$$

  (avec ψ l'ondelette mère), muni du produit scalaire de $\mathrm{L}^2(\mathbb{R})$, engendre un RKHS.

Le RKHS $\mathcal{H}$ étant construit, on peut appliquer par exemple le théorème 1.2.26 (du représentant, cas quadratique) et obtenir la solution du problème posé en (1.7)

$$f(\cdot) = \sum_{j=1}^{p} \nu_j m_j(\cdot) + \sum_{i=1}^{n} \alpha_i k(x_i, \cdot).$$

En utilisant la formule (1.30), on obtient

$$\begin{aligned} f(\cdot) &= \sum_{j=1}^{p} \nu_j m_j(\cdot) + \sum_{i=1}^{n} \alpha_i k(x_i, \cdot) \\ &= \sum_{j=1}^{p} \nu_j m_j(\cdot) + \sum_{i=1}^{n} \alpha_i \sum_{j=1}^{N} \overline{\phi_j}(x_i)\phi_j(\cdot) \\ &= \sum_{j=1}^{p} \nu_j m_j(\cdot) + \sum_{j=1}^{N} \left(\sum_{i=1}^{n} \alpha_i \overline{\phi_j}(x_i)\right) \phi_j(\cdot) \\ &= \sum_{j=1}^{p} \nu_j m_j(\cdot) + \sum_{i=1}^{N} d_i \phi_i(\cdot). \end{aligned} \tag{1.32}$$

On remarque donc que la combinaison linéaire de noyaux peut être remplacée par une combinaison linéaire de frames.

L'utilisation de noyaux construits à partir de frames présente plusieurs avantages : la solution peut être réécrite comme une combinaison linéaire de frames comme en (1.32), ce qui augmente

l'interprétabilité du modèle par rapport à un noyau où la base propre est inconnue, et permet également de sélectionner la famille de frames en fonction de la régularité souhaitée pour le modèle. L'inconvénient principal est le coût de calcul algorithmique élevé des frames duales (un algorithme est présenté dans [135]).

Des méthodes de résolution *multi-échelle*, où la régularisation se fait itérativement dans des RKHS emboîtés construits à partir de bases d'ondelettes d'échelle de plus en plus grande, sont présentées dans [9, 61, 135]. L'intérêt de la méthode est de remédier au sur-apprentissage (une fonction f trop oscillante) et au sous-apprentissage (une fonction f trop lisse) dus au fait qu'un noyau donné ne peut s'adapter à toutes les échelles d'un signal. Pour des applications des ondeletes aux statistiques, on pourra consulter [11, 12].

1.4.4 Support Vector Machines

Les *machines à vecteurs de support (Support Vector Machines, SVM)* sont un type de machine d'apprentissage proposé par Vapnik *et al.* [176], pouvant être utilisé pour la classification de données ou la régression. Nous présentons cette méthode en dernier car quand on fait de la régression, il s'agit d'une généralisation de l'ensemble des techniques précédentes (à l'exception du fait que le noyau $k(\cdot,\cdot)$ est ici supposé continu afin de vérifier les hypothèses du théorème 1.2.21, ce qui n'est pas le cas pour les autres méthodes).

L'idée est de supposer que la relation entrée-sortie est linéaire, si l'on prend pour domaine d'entrée l'espace des caractéristiques $\mathcal{F} = \Phi(\mathcal{X})$ (définition 1.2.15). Le produit scalaire dans $\mathcal{F}$ est défini à partir d'un noyau de Mercer $k(\cdot,\cdot)$ (définition 1.2.17) par la relation

$$\langle \Phi(x), \Phi(x') \rangle_{\mathcal{F}} = k(x, x')$$

(le théorème 1.2.21 assure l'existence de l'espace des caractéristiques et du produit scalaire associé pour un noyau de Mercer). Le problème situé à l'origine dans l'espace $\mathcal{X}$ est ainsi déplacé dans l'espace abstrait des caractéristiques $\mathcal{F}$. L'idée de la méthode est que l'on n'a besoin de spécifier ni la transformation Φ, ni l'espace $\mathcal{F}$, ni le produit scalaire $\langle \cdot, \cdot \rangle_{\mathcal{F}}$: tout se fait par l'intermédiaire du noyau $k(\cdot,\cdot)$. C'est *l'astuce du noyau (kernel trick)* : en choisissant un noyau de Mercer, on définit implicitement un espace des caractéristiques et un produit scalaire.

Dans le cas particulier de la *régression par vecteurs de support (Support Vector Regression, SVR)*, le système est modélisé par la fonction

$$f(x) = \langle \omega, \Phi(x) \rangle_{\mathcal{F}} + b, \tag{1.33}$$

avec $\omega \in \mathcal{F}$ [158]. On suppose donc que la fonction de régression est affine dans l'espace des caractéristiques. Avec les notations du théorème 1.2.24 (du représentant), cela revient à prendre $\mathcal{M} = \mathbb{R}$ et $\mathcal{H}$ le RKHS engendré par le noyau $k(\cdot,\cdot)$.

Remarque 1.4.25 *On peut étendre la SVR au cas où le terme paramétrique (ici, b) est plus général qu'une constante, voir [156].*

Afin d'estimer f à partir de l'ensemble d'entraînement $\{(x_i, y_i), i = 1 \dots n, x_i \in \mathbb{R}^p, y_i \in \mathbb{R}\}$, on cherche à minimiser la fonctionnelle de risque régularisée

$$\frac{1}{n} \sum_{i=1}^{n} c\left(f(x_i), y_i\right) + \gamma \|\omega\|_{\mathcal{F}}^2, \tag{1.34}$$

avec $\gamma > 0$, c'est-à-dire le risque empirique avec l'ajout d'un terme de régularisation qui pénalise les fortes variations de la fonction f dans l'espace des caractéristiques. On reconnaît (1.7), avec $\|\cdot\|_{\mathcal{F}}$ la norme de RKHS définie par le noyau $k(\cdot,\cdot)$.

La fonction de coût la plus utilisée est la fonction ε-insensitive [158]

$$c\left(f(x),y\right) = \begin{cases} |f(x)-y| - \varepsilon & \text{si } |f(x)-y| \geq \varepsilon\,; \\ 0 & \text{sinon,} \end{cases} \tag{1.35}$$

où seuls les y_i en dehors de la bande $\{[f(x)-\varepsilon, f(x)+\varepsilon], x \in \mathcal{X}\}$ (le *tube ε-insensitif*) contribuent au coût. Pour d'autres fonctions de coût couramment utilisées et des détails sur la résolution du problème d'optimisation associé, nous renvoyons à [157, 159].

La solution ω de (1.34) s'exprime en fonction de *vecteurs de support (Support Vectors)*

$$\omega = \sum_{i=1}^{n} \alpha_i \Phi(x_i),$$

les x_i pour lesquels $\alpha_i \neq 0$ sont les vecteurs de support. On a donc d'après (1.33)

$$\begin{aligned} f(x) &= \langle \omega, \Phi(x)\rangle_{\mathcal{F}} + b \\ &= \sum_{i=1}^{n} \alpha_i \langle \Phi(x_i), \Phi(x)\rangle_{\mathcal{F}} + b \\ &= \sum_{i=1}^{n} \alpha_i k(x_i, x) + b, \end{aligned}$$

et on retrouve la forme de la solution donnée dans le théorème 1.2.24. Pour une fonction de coût ε-insensitive, les coefficients sont calculés en résolvant le problème d'optimisation convexe équivalent à (1.34) [157].

$$\begin{cases} \underset{\omega, C, \xi_i, \xi_i^*, b}{\operatorname{argmin}} \frac{1}{2}\|\omega\|_{\mathcal{F}}^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*)\,; \\ y_i - \langle \omega, \Phi(x_i)\rangle_{\mathcal{F}} - b \leq \varepsilon + \xi_i\,; \\ \langle \omega, \Phi(x_i)\rangle_{\mathcal{F}} + b - y_i \leq \varepsilon + \xi_i^*\,; \\ \xi_i, \xi_i^* \geq 0, \end{cases}$$

avec $C = \dfrac{1}{2n\gamma}$. On peut montrer qu'il est équivalent de résoudre le problème dual

$$\begin{cases} \underset{\beta, \beta^* \in \mathbb{R}^n}{\operatorname{argmax}} -\frac{1}{2}\sum_{i,j=1}^{n}(\beta_i - \beta_i^*)(\beta_j - \beta_j^*)k(x_i, x_j) - \varepsilon\sum_{i=1}^{n}(\beta_i + \beta_i^*) + \sum_{i=1}^{n} y_i(\beta_i - \beta_i^*)\,; \\ \sum_{i=1}^{n}(\beta_i - \beta_i^*) = 0\,; \\ \beta_i, \beta_i^* \in [0, C], \end{cases}$$

où les variables duales β, β^* vérifient $\omega = \sum_{i=1}^{n}(\beta_i - \beta_i^*)x_i$. On a donc $\alpha_i = \beta_i - \beta_i^*$. Le coefficient b s'obtient à partir des conditions de Karush-Kuhn-Tucker (KKT),

$$\begin{aligned} \beta_i(\varepsilon + \xi_i - y_i + \langle \omega, \Phi(x_i)\rangle_{\mathcal{F}} + b) &= 0\,; \\ \beta_i^*(\varepsilon + \xi_i^* + y_i - \langle \omega, \Phi(x_i)\rangle_{\mathcal{F}} + b) &= 0\,; \\ (C - \beta_i)\xi_i &= 0\,; \\ (C - \beta_i^*)\xi_i^* &= 0. \end{aligned}$$

On en déduit (voir [157] pour les détails) que b peut être calculé de la façon suivante :

$$\begin{aligned} b &= y_i - \langle \omega, \Phi(x_i) \rangle_{\mathcal{F}} - \varepsilon \text{ pour } \beta_i \in]0, C[\, ; \\ b &= y_i - \langle \omega, \Phi(x_i) \rangle_{\mathcal{F}} + \varepsilon \text{ pour } \beta_i^* \in]0, C[. \end{aligned}$$

On remarque que pour tous les y_i à l'intérieur du tube ε-insensitif, $\beta_i = \beta_i^* = 0$, donc $\alpha_i = 0$. Ceci justifie la définition des vecteurs de support.

Une extension de la SVR avec fonction de coût ε-insensitive aux noyaux conditionnellement semi-définis positifs, est présentée dans [158]. Cependant, les contraintes de calcul font que la méthode est applicable seulement pour des noyaux c.s-d.p. d'ordre au plus 1.

La remarque 1.4.25, et le fait que la fonction de coût peut être quelconque (notamment quadratique) font que la SVR est un cadre général qui regroupe (à la continuité du noyau près) l'ensemble des méthodes à noyaux présentées ici (on pourra consulter [10] pour un cas curieux d'équivalence entre krigeage et SVR, avec une fonction de perte qui n'est pas la fonction quadratique). Pour beaucoup plus de détails et d'autres références, nous renvoyons à [157].

Nous avons choisi d'utiliser le krigeage dans notre étude, car cette méthode permet d'obtenir une expression du prédicteur et une estimation de l'erreur commise (l'erreur quadratique moyenne) pour un coût de calcul moindre. Notons cependant qu'il est possible de mettre en perspective bayésienne (voir l'annexe C) chacune des méthodes présentées, et d'obtenir une expression de l'erreur quadratique moyenne en tout point (voir [54, 186] dans le cas de la SVR).

Chapitre 2

Krigeage

Parmi toutes les techniques de modélisation présentées au chapitre précédent, notre choix s'est porté sur le krigeage. Le modèle est un processus aléatoire gaussien, de moyenne et fonction de covariance paramétriques, dont on va estimer les paramètres. Si l'on admet la modélisation, un intérêt du krigeage est qu'il fournit, à un coût de calcul faible, une estimation de l'erreur de prédiction en tout point du domaine d'étude, ce qui donne une idée de la précision du modèle aux points non échantillonnés. Un autre point intéressant est que le noyau reproduisant du RKHS est une fonction de covariance, ce qui permet d'adopter une perspective probabiliste et d'utiliser les résultats donnés par la théorie des probabilités.

Des éléments de la théorie des processus aléatoires gaussiens seront rappelés dans un premier temps, qui permettront notamment de comprendre l'importance du choix de la fonction de covariance. Puis la base de la théorie du krigeage sera présentée. Notons que nous avons volontairement insisté sur le point de vue statistique de la théorie : pour une présentation plus axée sur le point de vue de l'analyse fonctionnelle, on consultera [178].

2.1 Processus aléatoires gaussiens

2.1.1 Généralités

Commençons par rappeler ce qu'est un processus aléatoire. On se restreindra dans toute la suite à des processus à valeurs réelles (une généralisation au cas complexe est donnée dans [101, 161]).

Définition 2.1.1 *[43] Soient $(\Omega, \mathcal{T}, \mathcal{P})$ un espace probabilisé et $\mathcal{X}$ un espace de paramètres. Un processus aléatoire est une fonction à valeurs réelles $Y(x, \omega)$ définie sur $\mathcal{X} \times \Omega$ qui, $\forall x \in \mathcal{X}$ fixé, est une fonction mesurable Y_x de $\omega \in \Omega$.*

Dans la suite, $\mathcal{X}$ sera un sous-ensemble de $\mathbb{R}^d$.

Remarque 2.1.2 *[101]. Il est possible de voir un processus aléatoire de deux façons :*

- *comme une collection de variables aléatoires indexée par $\mathcal{X}$, définies sur un même espace d'états (et habituellement de même loi, comme on le verra pour les processus strictement stationnaires), par l'intermédiaire de la fonction $x \in \mathcal{X} \mapsto Y_x$. Ce point de vue sera utilisé pour calculer la loi conditionnelle aux observations en un point non échantillonné.*
- *comme une façon d'associer une mesure de probabilité à un ensemble de fonctions définies sur $\mathcal{X}$, à travers l'application $\omega \in \Omega \mapsto Y_\omega = Y(\cdot, \omega)$, qui à ω associe une* trajectoire *(ou*

réalisation*) du processus. Le modèle de krigeage sera obtenu en prenant une moyenne des trajectoires conditionnelles aux observations.*

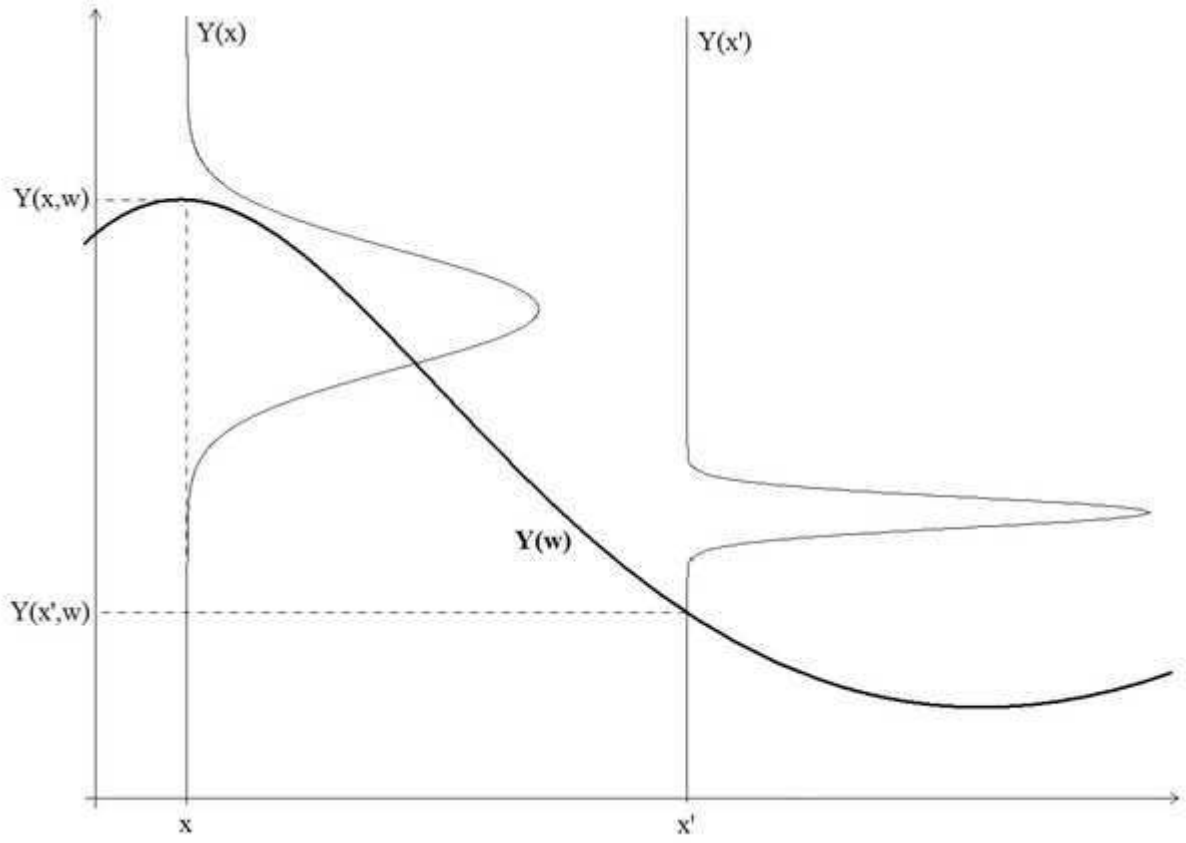

FIGURE 2.1 – Représentations d'un processus aléatoire : en un point $x \in \mathcal{X}$, $Y(x) = Y(x, \cdot)$ est une variable aléatoire définie sur Ω ; en un élément $\omega \in \Omega$, $Y(\omega) = Y(\cdot, \omega)$ est une fonction définie sur $\mathcal{X}$.

La loi d'un processus aléatoire est entièrement déterminée par ses *répartitions finies*,

$$F_{x_1,\ldots,x_i}(y_1,\ldots,y_i) = \mathcal{P}(Y_{x_1} \leq y_1,\ldots,Y_{x_i} \leq y_i).$$

On peut montrer que les répartitions finies doivent satisfaire deux conditions pour être consistantes les unes avec les autres,

$$F_{x_1,\ldots,x_i}(y_1,\ldots,y_i) = F_{x_{\pi(1)},\ldots,x_{\pi(i)}}(y_{\pi(1)},\ldots,y_{\pi(i)}) \quad \forall \pi \in \mathcal{S}_i, \tag{2.1}$$

$$F_{x_1,\ldots,x_{i-1}}(y_1,\ldots,y_{i-1}) = F_{x_1,\ldots,x_{i-1},x_i}(y_1,\ldots,y_{i-1},\infty), \tag{2.2}$$

$\forall i, \forall x_1,\ldots,x_i \in \mathcal{X}$ et $y_1,\ldots,y_i \in \mathbb{R}$, avec $\mathcal{S}_i$ le groupe des permutations de $\{1,\ldots,i\}$. Ces conditions de *symétrie* (2.1) et de *compatibilité* (2.2) viennent simplement du fait que les évènements correspondant aux deux côtés des égalités sont identiques.

L'existence d'un processus aléatoire de répartitions finies symétriques et compatibles données est garantie par le théorème suivant, conséquence du *théorème d'existence (ou d'extension) de Kolmogorov.*

Théorème 2.1.3 *[57] Si un système de répartitions finies $F_{x_1,\ldots,x_i}$ satisfait les conditions (2.1) et (2.2), alors il existe un espace de probabilité $(\Omega, \mathcal{T}, \mathcal{P})$ et un processus aléatoire Y défini sur $(\Omega, \mathcal{T}, \mathcal{P})$, ayant $F_{x_1,\ldots,x_i}$ comme répartitions finies.*

Les répartitions finies d'un processus aléatoire suffisent presque à caractériser l'ensemble des propriétés du processus : on ne s'intéresse donc pas en général à l'espace abstrait $(\Omega, \mathcal{T}, \mathcal{P})$. Afin de pouvoir connaître la régularité des trajectoires du processus à partir de ses répartitions finies, il faut cependant faire l'hypothèse supplémentaire de *séparabilité* du processus aléatoire.

Définition 2.1.4 *[57] On suppose que $\mathcal{X} \subset \mathbb{R}^d$ est un espace métrique. Le processus aléatoire $Y(x, \omega)$ est dit séparable si il existe un sous-ensemble dénombrable dense $\{x_1, x_2, \ldots\}$ de $\mathcal{X}$ et un ensemble $\mathcal{N} \subset \Omega$ de probabilité 0 tels que, pour tout ouvert $O \subset \mathcal{X}$ et tout ensemble fermé $F \subset \mathbb{R}$, les ensembles*

$$\{\omega, Y(x_i, \omega) \in F, x_i \in O\} \text{ et } \{\omega, Y(x, \omega) \in F \ \ \forall x \in O\}$$

diffèrent l'un de l'autre seulement par un sous-ensemble de $\mathcal{N}$.

Observons maintenant l'on peut se restreindre aux processus séparables, en rappelant tout d'abord la notion d'*équivalence*.

Définition 2.1.5 *[57] Les processus aléatoires Y et Z, définis sur le même espace, sont dits* stochastiquement équivalents *(ou une* version *l'un de l'autre), si $\mathcal{P}\{\omega, Y(x, \omega) = Z(x, \omega)\} = 1 \quad \forall x \in \mathcal{X}$.*

Les trajectoires de deux processus équivalents sont donc identiques presque sûrement. Mais deux processus équivalents, bien qu'ayant les mêmes répartitions finies, ne sont pas nécessairement identiques.

Exemple 2.1.6 *[2] Soit $\mathcal{X} = \mathbb{R}$ et $\Omega = [0, 1]$, muni de la loi uniforme. Définissons deux processus aléatoires sur $\mathcal{X} \times \Omega$:*

$$\begin{aligned} Y(x, \omega) &= 0 \quad \forall\, x, \omega\,; \\ Z(x, \omega) &= \begin{cases} 1 & \text{si } x = \omega\,; \\ 0 & \text{sinon.} \end{cases} \end{aligned}$$

Ces deux processus sont équivalents : à x fixé, ils ne diffèrent que sur l'ensemble de mesure nulle $\{\omega = x\}$. Néanmoins,

$$\begin{aligned} \mathcal{P}\{\omega, Y_\omega(\cdot) \text{ est continue sur } [0, 1]\} &= 1\,; \\ \mathcal{P}\{\omega, Z_\omega(\cdot) \text{ est continue sur } [0, 1]\} &= 0. \end{aligned}$$

La proposition suivante permet cependant de se ramener à des processus séparables.

Proposition 2.1.7 *[57] Si $\mathcal{X}$ est un espace métrique, tout processus aléatoire défini sur $\mathcal{X}$ admet une version séparable.*

À partir d'ici, les processus considérés seront donc supposés séparables : ils seront donc entièrement déterminés par leurs répartitions finies.

2.1.2 Moyenne et covariance

2.1.2.1 Généralités

Rappelons maintenant les fonctions classiques associées à un processus aléatoire.

Définition 2.1.8
La moyenne *d'un processus aléatoire $Y(x,\omega)$ est la fonction*

$$x \in \mathcal{X} \longmapsto m(x) = \mathbb{E}(Y_x).$$

Sa fonction de covariance *est la fonction*

$$x, x' \in \mathcal{X} \longmapsto k(x,x') = \mathrm{cov}(Y_x, Y_{x'}) = \mathbb{E}(Y_x Y_{x'}) - m(x)m(x').$$

Sa variance *est la fonction $\sigma^2(x) = k(x,x)$.*
Sa fonction de corrélation *est la fonction*

$$x, x' \in \mathcal{X} \longmapsto \rho(x,x') = \mathrm{corr}(Y_x, Y_{x'}) = \frac{k(x,x')}{\sigma(x)\sigma(x')}.$$

Les fonctions de covariance et de corrélation sont parfois appelées aussi *fonction d'autocovariance* et *fonction d'autocorrélation.*

Une moyenne et une fonction de covariance ne suffisent pas en général à caractériser un processus aléatoire, mais c'est le cas lorsque celui-ci est gaussien.

Définition 2.1.9 *[161] Le vecteur aléatoire X, de taille l, suit la* loi normale multivariée *si la variable aléatoire ${}^taX$ suit une loi normale monodimensionnelle $\forall a \in \mathbb{R}^l$. On peut montrer qu'alors $\mu = \mathbb{E}(X)$ et $\Sigma = \mathrm{cov}(X,X)$ sont bien définies. Si de plus Σ est définie positive, alors X suit la densité*

$$p(x) = \frac{1}{(2\pi)^{\frac{l}{2}}\sqrt{\det(\Sigma)}} \exp\left\{-\frac{{}^t(x-\mu)\Sigma^{-1}(x-\mu)}{2}\right\}. \tag{2.3}$$

On notera cette loi $\mathcal{N}_l(\mu,\Sigma)$.

Définition 2.1.10 *Un* processus (aléatoire) gaussien *Y est un processus aléatoire dont toutes les* distributions finies *$(Y(x_1),\ldots,Y(x_i))$ sont normales multivariées $\forall i$ et $\forall x_1,\ldots,x_i \in \mathcal{X}$.*

Les processus gaussiens ont une place à part pour plusieurs raisons [2]. Ils permettent de modéliser de nombreux phénomènes naturels, et il existe une théorie bien établie à leur sujet : le modèle est entièrement spécifié par sa moyenne et sa fonction de covariance, et la forme simple de leurs distributions finies permet d'obtenir des solutions analytiques pour des problèmes de prédiction ou d'estimation.

Les fonctions de covariance sont étroitement liées aux fonctions semi-définies positives (définition 1.2.7). Rappelons les résultats d'équivalence suivants [2].

Théorème 2.1.11 *La classe des fonctions de covariance coïncide avec la classe des fonctions semi-définies positives (définition 1.2.7).*

Corollaire 2.1.12 *La classe des fonctions de corrélation coïncide avec la classe des fonctions semi-définies positives $\rho(\cdot,\cdot)$ vérifiant $\rho(x,x) = 1$.*

En pratique, on observe un ensemble $\{Y_\omega(x_1)\ldots,Y_\omega(x_n)\}$, qui correspond à de l'information partielle sur une seule réalisation $Y_\omega(\cdot)$ du processus, et on souhaite faire de l'inférence statistique sur la loi du processus. On constate que l'on est très loin du cadre classique, où l'inférence se fait à partir de la connaissance d'un « vrai » échantillon, qui correspondrait ici à un ensemble $\{Y_{\omega_1}(\cdot),\ldots,Y_{\omega_n}(\cdot)\}$ [144] : on dispose donc de beaucoup moins d'information que dans le cadre classique. En restreignant la classe des processus considérés, on peut cependant obtenir un argument théorique qui justifie la validité de l'inférence dans le cas des processus aléatoires : on supposera que la variance et la moyenne sont invariantes selon certaines transformations, ce qui permet de calculer des espérances mathématiques sur l'espace Ω à partir de moyennes dans l'espace $\mathcal{X}$ (ce que l'on appelle *ergodicité*, voir l'annexe A).

Définition 2.1.13 *Un processus aléatoire est* stationnaire au sens strict *si ses répartitions finies sont invariantes par translation,*

$$F_{x_1+t,\ldots,x_i+t}(y_1,\ldots,y_i) = F_{x_1,\ldots,x_i}(y_1,\ldots,y_i) \quad \forall\, i, x_1,\ldots,x_i,\ \textit{et}\ t \in \mathcal{X},$$

autrement dit si les vecteurs aléatoires $(Y_{x_1},\ldots,Y_{x_i})$ *et* $(Y_{x_1+t},\ldots,Y_{x_i+t})$ *suivent la même loi.*

Définition 2.1.14 *Un processus aléatoire est* stationnaire au sens faible *si* $\mathbb{E}[Y(x)^2] < \infty \quad \forall x$ *et si*

$$m(x) = m \ \textit{et}\ k(x,x') = k(\tau),$$

avec $\tau = x - x'$*. La fonction de covariance est alors appelée* fonction de covariance stationnaire.

Un processus aléatoire stationnaire au sens strict ayant des moments d'ordre 2 finis est aussi stationnaire au sens faible. La réciproque est fausse en général, mais vraie pour les processus gaussiens. Puisque nous allons utiliser des processus gaussiens, nous ne ferons plus la distinction dans la suite et parlerons simplement de processus *stationnaires.*

Les fonctions de covariance stationnaire vérifient les propriétés suivantes [161] :
- $k(0) \geq 0$;
- $k(\tau) = k(-\tau)$;
- $|k(\tau)| \leq k(0)$.

En effet, $k(0)$ est la variance du processus, qui est toujours positive. Soient x, x' tels que $\tau = x - x'$. Alors $k(\tau) = \mathrm{cov}(Y_x, Y_{x'}) = \mathrm{cov}(Y_{x'}, Y_x) = k(-\tau)$; enfin,

$$|k(\tau)| = |\mathrm{cov}(Y_x, Y_{x'})| \leq \sqrt{\mathrm{var}(Y_x)}\sqrt{\mathrm{var}(Y_{x'})} = \sqrt{k(0)}^2 = k(0),$$

par application de l'inégalité de Cauchy–Schwarz au semi produit scalaire $k(\cdot,\cdot)$ (remarque 1.2.4).

Remarque 2.1.15 *[161] La régularité d'une fonction de covariance stationnaire* $k(\cdot)$ *est la même que sa régularité en* 0*. Montrons-le pour ce qui concerne la continuité : si* $k(\cdot)$ *est continue en* 0*, alors*

$$\begin{array}{rcl} |k(x) - k(x')| & = & \left|\mathrm{cov}\{Y(x) - Y(x'), Y(0)\}\right| \\ & \leq & \sqrt{\mathrm{var}\{Y(x) - Y(x')\}\mathrm{var}\{Y(0)\}} \\ & = & \sqrt{2\{k(0) - k(x-x')\}k(0)} \\ & \underset{x\to x'}{\longrightarrow} & 0. \end{array}$$

On peut noter que la moyenne et la variance d'un processus aléatoire stationnaire sont constantes, la variance étant égale à $\sigma^2 = k(0)$. Pour un processus aléatoire Y gaussien, les variables aléatoires Y_x sont donc normales et identiquement distribuées (comme évoqué à la remarque 2.1.2).

Remarque 2.1.16 *Une synthèse des résultats disponibles sur les valeurs prises par les trajectoires d'un processus gaussien stationnaire Y est donnée dans [91]. Notons que ces résultats ne s'appliquent pas aux trajectoires du processus conditionnel défini en x par $(Y(x)|Y(x_1) = y_1, \ldots, Y(x_n) = y_n)$, qui n'est pas stationnaire.*

En faisant l'hypothèse d'un processus gaussien stationnaire dont la fonction de covariance tend vers 0 à l'infini, la validité de l'inférence est justifiée par un argument d'ergodicité (voir l'annexe A). Si l'on souhaite encore renforcer les caractéristiques géométriques du processus, on peut ajouter, à l'hypothèse d'invariance par translation, l'invariance par rotation et par réflexion. Comme pour la stationnarité, on dispose d'une formulation stricte et faible pour l'isotropie, qui sont équivalentes dans le cas des processus gaussiens [161]. Nous ne donnons donc que la deuxième.

Définition 2.1.17 *Un processus aléatoire stationnaire est* isotrope (au sens faible) *si*

$$k(x, x') = k(\|\tau\|),$$

avec $\tau = x - x'$. La fonction de covariance est alors appelée fonction de covariance isotrope.

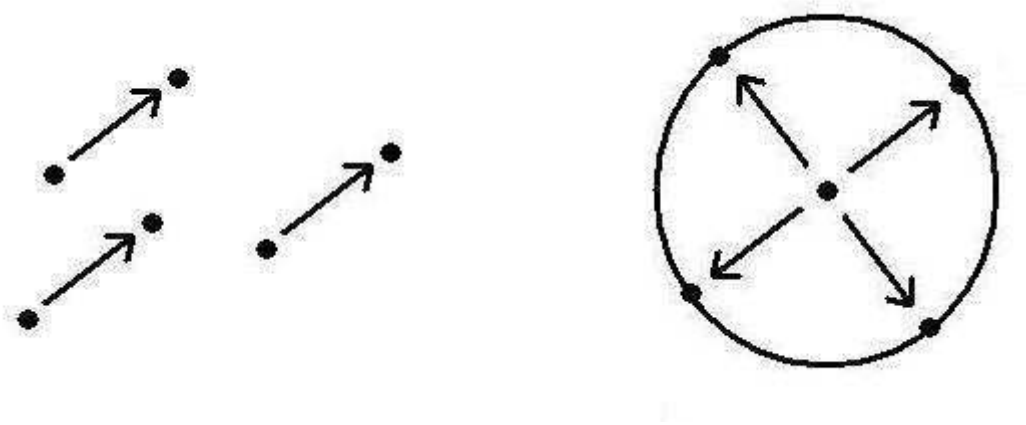

FIGURE 2.2 – À gauche, toute paire de points dont le deuxième est obtenu par la même translation du premier ont la même corrélation pour un processus stationnaire. À droite, tous les points d'un cercle ont la même corrélation avec son centre pour un processus isotrope.

Il existe d'autres types de propriétés géométriques, comme l'*anisotropie*, la *séparabilité* (à ne pas confondre avec la séparabilité du processus), et la *symétrie par quadrant*, pour la définition desquelles nous renvoyons à [2].

Un processus gaussien étant entièrement déterminé par ses deux premiers moments, le choix de la fonction de covariance est d'une importance capitale. Dans le paragraphe suivant, nous allons observer le lien existant entre la fonction de covariance et la régularité d'un processus gaussien stationnaire.

2.1.2.2 Covariance et régularité d'un processus gaussien stationnaire

Il est possible d'établir un lien entre la régularité d'un processus gaussien stationnaire et de sa fonction de covariance en 0. Si on ne suppose pas le processus séparable, on ne peut qu'obtenir des

résultats de régularité *en moyenne quadratique* (voir annexe B), peu intuitifs et peu satisfaisants d'un point de vue pratique. Avec l'hypothèse de séparabilité, on peut établir des résultats plus utiles de *régularité des trajectoires presque sûrement.*

Définition 2.1.18 *Soit $\mathcal{X} \subset \mathbb{R}^d$. Un processus aléatoire Y est à* trajectoires continues presque sûrement *sur $\mathcal{X}$ si*

$$\mathcal{P}\{\omega, Y_\omega(\cdot) \text{ est continue sur } \mathcal{X}\} = 1.$$

On dira par abus de langage que le processus est à trajectoires continues.

La continuité des trajectoires du processus est en général le minimum que l'on attend lorsque l'on fait de la modélisation. Le théorème suivant donne une condition suffisante de continuité des trajectoires.

Théorème 2.1.19 *[30] Soit Y un processus gaussien stationnaire dont la fonction de corrélation $\rho(\cdot)$ est continue en 0. Si il existe $\varepsilon > 0$ tel que*

$$1 - \rho(\tau) = \mathcal{O}\left(\frac{1}{|\log \|\tau\||^{1+\varepsilon}}\right) \quad \text{quand } \|\tau\| \to 0, \tag{2.4}$$

alors il existe une version de Y dont les trajectoires sont continues presque sûrement.

Si $\rho(\cdot)$ est continue en 0, mais ne vérifie pas (2.4), alors le processus Y est seulement continu en moyenne quadratique (voir l'annexe B). Si $\rho(\cdot)$ est dérivable en 0, alors la condition (2.4) est verifiée et Y admet une version dont les trajectoires sont continues presque sûrement.

Remarque 2.1.20 *[2] En pratique, il est difficile de trouver une fonction de corrélation continue ne vérifiant pas (2.4), à tel point qu'il est parfois suggéré de considérer qu'un processus gaussien stationnaire séparable dont la fonction de corrélation est continue est à trajectoires continues.*

Intéressons-nous maintenant à la dérivabilité du processus.

Définition 2.1.21 *Soit $\mathcal{X} \subset \mathbb{R}^d$. Un processus aléatoire Y est* (q fois) différentiable presque sûrement *sur $\mathcal{X}$ si*

$$\mathcal{P}\{\omega, Y_w(\cdot) \text{ est (q fois) différentiable sur } \mathcal{X}\} = 1.$$

On dira par abus de langage que le processus est (q fois) différentiable.

On utilise le fait que le processus est défini sur $\mathbb{R}^d$ pour obtenir des résultats de régularité du processus à partir de la fonction de covariance, en se servant des dérivées partielles.

Proposition 2.1.22 *[30, 161] Soit Y un processus gaussien stationnaire, q fois différentiable sur $\mathcal{X} \subset \mathbb{R}^d$, de fonction de covariance $k(\cdot)$. Alors la fonction de covariance des processus dérivés (qui sont aussi gaussiens et stationnaires),*

$$Y^{(\kappa)}(x, \omega) = \frac{\partial^{|\kappa|} Y}{\partial x_1^{\kappa_1} \dots \partial x_n^{\kappa_n}}(x, \omega) = \frac{\partial^{|\kappa|} Y_\omega}{\partial x_1^{\kappa_1} \dots \partial x_n^{\kappa_n}}(x),$$

avec $\kappa = (\kappa_1 \dots, \kappa_n)$, $|\kappa| = \sum_{i=1}^n \kappa_i \leq q$, est donnée par

$$\mathrm{cov}_\kappa(x, x') = (-1)^{|\kappa|} k^{(2\kappa)}(\tau),$$

avec $\tau = x - x'$, et $k^{(\kappa)}(\tau) = \dfrac{\partial^{|\kappa|} k}{\partial x_1^{\kappa_1} \dots \partial x_n^{\kappa_n}}(\tau)$.

La réciproque est fausse en général : l'existence des fonctions $\text{cov}_\kappa, |\kappa| \leq q$ ne garantit pas que le processus est q fois différentiable (voir cependant la remarque 2.1.23).

Pour montrer que le processus gaussien stationnaire Y est q fois différentiable, on peut appliquer le théorème 2.1.19 aux fonctions $\text{cov}_\kappa, |\kappa| = q$ (si elles sont bien définies) : si le théorème est vrai pour toutes ces fonctions, les dérivées partielles d'ordre q sont toutes continues, et un théorème de calcul différentiel permet de conclure que Y est q fois différentiable [2].

Remarque 2.1.23 (suite de la remarque 2.1.20). *Si l'on considère que le théorème 2.1.19 est toujours vrai en pratique, on peut appliquer le théorème B.0.18 de l'annexe B, et déduire la régularité des trajectoires du processus de la régularité de sa fonction de covariance en* 0.

Le théorème suivant, conséquence d'un résultat donné dans [17], donne une condition suffisante de dérivabilité pour une famille de processus plus générale.

Théorème 2.1.24 *Soit Y un processus stationnaire de moyenne nulle, défini sur $\mathbb{R}^d \times \Omega$, ayant des moments d'ordre* 2 *finis et une fonction de corrélation $(n+1+2q)$ fois continûment dérivable. Alors Y admet une version q fois continûment dérivable.*

Remarque 2.1.25 *Dans le cas d'un processus Y gaussien stationnaire, il suffit que la fonction de covariance soit $(2q+1)$ fois dérivable en* 0 *pour que Y admette une version q fois dérivable, car alors la condition (2.4) est verifiée pour les fonctions de corrélation des processus dérivés $\{Y^{(\kappa)}, |\kappa| = q\}$. Les dérivées partielles d'ordre q des trajectoires de Y étant continues presque sûrement, Y est alors q fois dérivable presque sûrement.*

2.1.2.3 Construction de fonctions de covariance stationnaires

Dans le cas où la fonction de covariance est stationnaire, la fonction de corrélation vérifie $\rho(x, x') = k(x - x')/\sigma^2 = \rho(x - x')$. Définir une fonction de covariance stationnaire revient donc à définir une fonction de corrélation stationnaire.
Il est possible de combiner des fonctions de corrélation existantes.

Proposition 2.1.26 *[30] Soit S_d l'ensemble des fonctions de corrélation stationnaires de $\mathbb{R}^d$.*
- *Si $\rho_1, \rho_2 \in S_d, a_1, a_2 \geq 0$, et $a_1 + a_2 = 1$, alors $a_1\rho_1 + a_2\rho_2 \in S_d$;*
- *si $\rho_1, \rho_2 \in S_d$, alors $\rho_1\rho_2 \in S_d$;*
- *si $\rho_i \in S_d \quad \forall i$, et $\rho(x, x') = \lim_{i\to\infty} \rho_i(x, x')$ existe $\quad \forall x, x' \in \mathbb{R}^d$, alors $\rho \in S_d$.*

Une méthode de construction explicite de fonctions de covariance stationnaires est donnée par le théorème de Bochner.

Théorème 2.1.27 *(de Bochner, cas réel)[2, 144, 161] Une fonction continue à valeurs réelles $k(\cdot)$ est une fonction de covariance stationnaire sur $\mathbb{R}^d$ si, et seulement si, celle-ci peut s'écrire*

$$k(\tau) = \int_{\mathbb{R}^d} \cos({}^t\tau x)\, \mu(\mathrm{d}x),$$

avec μ une mesure positive symétrique bornée sur $\mathbb{R}^d$.

Le résultat suivant s'en déduit immédiatement.

Théorème 2.1.28 *(de Wiener-Khintchine, cas réel)[2, 144, 161] Une fonction continue à valeurs réelles $\rho(\cdot)$ est une fonction de corrélation stationnaire sur $\mathbb{R}^d$ si, et seulement si, celle-ci peut s'écrire*

$$\rho(\tau) = \int_{\mathbb{R}^d} \cos({}^t\tau x)\, \upsilon(\mathrm{d}x), \tag{2.5}$$

avec υ une mesure de probabilité symétrique sur $\mathbb{R}^d$, appelée mesure spectrale *correspondant à $\rho(\cdot)$.*
Si υ admet une densité $f(\cdot)$ par rapport à la mesure de Lebesgue, alors

$$\rho(\tau) = \int_{\mathbb{R}^d} \cos({}^t\tau x) f(x)\, \lambda_d(\mathrm{d}x),$$

avec λ_d la mesure de Lebesgue sur $\mathbb{R}^d$. La fonction $f(\cdot)$ est appelée densité spectrale *correspondant à $\rho(\cdot)$.*

Une fonction de corrélation stationnaire est donc la transformée de Fourier d'une mesure de probabilité symétrique υ. Dans le cas où cette mesure admet une densité spectrale $f(\cdot)$, il existe une formule d'inversion, pour laquelle nous renvoyons à [2, 161]. Le comportement de la mesure spectrale à l'infini est étroitement relié à la régularité de la fonction de covariance en 0 (et donc à la régularité du processus aléatoire) : on pourra consulter les théorèmes abelien et tauberien donnés dans [161].

Pour une fonction de corrélation isotrope, l'intégrale (2.5) prend une forme simple.

Théorème 2.1.29 *[2, 161] Une fonction continue à valeurs réelles $\rho(\cdot)$ est une fonction de corrélation isotrope sur $\mathbb{R}^d$ $(d \geq 2)$ si, et seulement si, elle peut s'écrire*

$$\rho(\|\tau\|) = 2^{\frac{d-2}{2}} \Gamma\left(\frac{d}{2}\right) \int_0^\infty \frac{J_{\frac{d-2}{2}}(\|\tau\| x)}{(\|\tau\| x)^{\frac{d-2}{2}}}\, F(\mathrm{d}x). \tag{2.6}$$

La fonction $F(\cdot)$, appelée fonction de répartition spectrale isotrope *correspondant à $\rho(\cdot)$, est une fonction croissante sur $[0, \infty[$ vérifiant $F(0) = 0$ et $\lim_{x\to\infty} F(x) = 1$, et $J_\nu(\cdot)$ est la fonction de Bessel ordinaire [3].*
La fonction de répartition spectrale isotrope $F(\cdot)$ est liée à la mesure spectrale υ par la formule

$$F(t) = \int_{\|x\|<t} \upsilon(\mathrm{d}x).$$

Dans le cas $d = 1$, on obtient immédiatement à partir de (2.5),

$$\rho(|\tau|) = \int_{\mathbb{R}} \cos(|\tau| x)\, F(\mathrm{d}x), \tag{2.7}$$

qui se réécrit avec la densité spectrale

$$\rho(|\tau|) = 2 \int_0^\infty \cos(|\tau| x) f(x)\, \lambda(\mathrm{d}x). \tag{2.8}$$

Il faut prendre garde au fait qu'une fonction de corrélation isotrope $\rho(\|\cdot\|)$ sur $\mathbb{R}^d$ n'est pas forcément semi-définie positive sur $\mathbb{R}^{d+1}$. Mais la réciproque est vraie.

Proposition 2.1.30 *[30] Une fonction de corrélation isotrope $\rho(\|\cdot\|)$ définie sur $\mathbb{R}^{d+1}$ est aussi une fonction de corrélation isotrope sur $\mathbb{R}^d$.*

Nous rappelons maintenant les fonctions de corrélation stationnaires et isotropes les plus classiques [104, 141, 144, 161].

Exemple 2.1.31 *Construisons des fonctions de corrélation stationnaires ou isotropes sur* $\mathbb{R}^d$ *à partir de densités de probabilité* $f(\cdot)$*, en utilisant les relations (2.6) et (2.8).*

– *si* $f(x) \propto e^{-\frac{x^2}{4\theta}}, \theta > 0$*, on obtient la* fonction de corrélation gaussienne *sur* $\mathbb{R}$*,*

$$\rho(|\tau|) = e^{-\theta\tau^2}, \quad \tau \in \mathbb{R}, \theta > 0. \tag{2.9}$$

Le terme θ *est un paramètre d'échelle qui mesure à quelle vitesse la corrélation décroît à mesure que la distance* $|\tau|$ *devient grande.*
La généralisation à $\mathbb{R}^d$ *s'obtient par multiplication en utilisant la proposition 2.1.26 :*

$$\rho(\tau) = e^{-\sum_{i=1}^{d} \theta_i \tau_i^2}, \quad \tau \in \mathbb{R}^d, \theta_i > 0. \tag{2.10}$$

Si les θ_i *sont tous égaux, on obtient la* fonction de corrélation gaussienne isotrope *sur* $\mathbb{R}^d$

$$\rho(\|\tau\|) = e^{-\theta\|\tau\|^2}, \quad \tau \in \mathbb{R}^d, \theta > 0.$$

C'est la seule fonction de corrélation isotrope qui peut se factoriser en d fonctions de chaque facteur τ_i *[161].*
On remarque que les fonctions de corrélation gaussiennes sont infiniment dérivables à l'origine. Les trajectoires du processus gaussien stationnaire correspondant sont infiniment dérivables presque sûrement.

– *si* $f(x) \propto \dfrac{1}{(\theta^2 + x^2)^{\frac{d+1}{2}}}, \theta > 0$*, on obtient une* fonction de corrélation exponentielle isotrope *sur* $\mathbb{R}^d$*,*

$$\rho(\|\tau\|) = e^{-\theta\|\tau\|}, \quad \tau \in \mathbb{R}^d, \theta > 0. \tag{2.11}$$

Un processus gaussien défini sur $\mathbb{R}$ *de fonction de corrélation (2.11) est appelé* processus d'Ornstein-Uhlenbeck.
En utilisant la proposition 2.1.26 et multipliant d fonctions de corrélation exponentielle isotrope définies sur $\mathbb{R}$*, on obtient une version stationnaire non isotrope de la* fonction de corrélation exponentielle*,*

$$\rho(\tau) = e^{-\sum_{i=1}^{d} \theta_i |\tau_i|}, \quad \tau \in \mathbb{R}^d, \theta_i > 0. \tag{2.12}$$

On remarque que ces fonctions de corrélation exponentielles (2.12) sont continues, mais pas dérivables à l'origine. Les trajectoires du processus gaussien stationnaire correspondant sont seulement continues presque sûrement.
Remarque 2.1.32 *La version la plus générale des fonctions de corrélation exponentielle est donnée par*

$$\rho(\tau) = e^{-\sum_{i=1}^{d} \theta_i |\tau_i|^{p_i}}, \quad \tau \in \mathbb{R}^d, \theta_i > 0, p_i \in]0, 2]. \tag{2.13}$$

On remarque que les fonctions de corrélation gaussienne (2.10) et exponentielle de type (2.12) sont un cas particulier de fonctions de corrélation exponentielle. Les fonctions de corrélation exponentielle sont continues à l'origine, et aucune, à part la fonction de corrélation gaussienne, n'est dérivable à l'origine. Les trajectoires d'un processus gaussien de covariance exponentielle sont continues presque sûrement si $p \in]0, 2[$*, ou infiniment dérivables presque sûrement si* $p = 2$*.*

– *si* $f(x) \propto \dfrac{1}{(\theta^2 + x^2)^{\nu+\frac{d}{2}}}, \theta > 0, \nu > 0$, *on obtient la* fonction de corrélation de Matérn isotrope *sur* $\mathbb{R}^d$ *[161],*

$$\rho(\|\tau\|) = \frac{1}{\Gamma(\nu)2^{\nu-1}}(\theta\|\tau\|)^{\nu} K_{\nu}(\theta\|\tau\|), \quad \tau \in \mathbb{R}^d, \theta > 0, \nu > 0 \tag{2.14}$$

avec $K_{\nu}(\cdot)$ *une fonction de Bessel modifiée [3].*
Le terme θ *est un paramètre d'échelle ; le paramètre* ν *définit la régularité du processus gaussien associé, qui est q fois dérivable presque sûrement si, et seulement si,* $\nu > q$ *[161].*
La forme la plus générale des fonctions de corrélation de Matérn *est obtenue en multipliant des fonctions de type (2.14) définies sur* $\mathbb{R}$ *(c.f. proposition 2.1.26),*

$$\rho(\tau) = \prod_{i=1}^{d} \frac{1}{\Gamma(\nu_i)2^{\nu_i-1}}(\theta_i|\tau_i|)^{\nu_i} K_{\nu_i}(\theta_i|\tau_i|), \quad \tau \in \mathbb{R}^d, \theta_i > 0, \nu_i > 0. \tag{2.15}$$

En observant leurs densités spectrales, on remarque que les fonctions de corrélation exponentielle de type (2.12) et exponentielle isotrope (2.11) sont respectivement un cas particulier des fonctions de corrélation de Matérn (2.15) et Matérn isotrope (2.14).
La fonction de corrélation gaussienne est un cas limite de fonction de corrélation de Matérn. En effet, après reparamétrisation de la fonction de corrélation de Matérn, on obtient [144]

$$\frac{1}{\Gamma(\nu)2^{\nu-1}}\left(2\sqrt{\theta\nu}|\tau|\right)^{\nu} K_{\nu}\left(2\sqrt{\theta\nu}|\tau|\right) \underset{\nu\to\infty}{\longrightarrow} \mathrm{e}^{-\theta\tau^2}.$$

On aurait donc pu définir les fonctions de corrélation gaussienne à partir des fonctions de corrélation de Matérn, en utilisant la proposition 2.1.26.

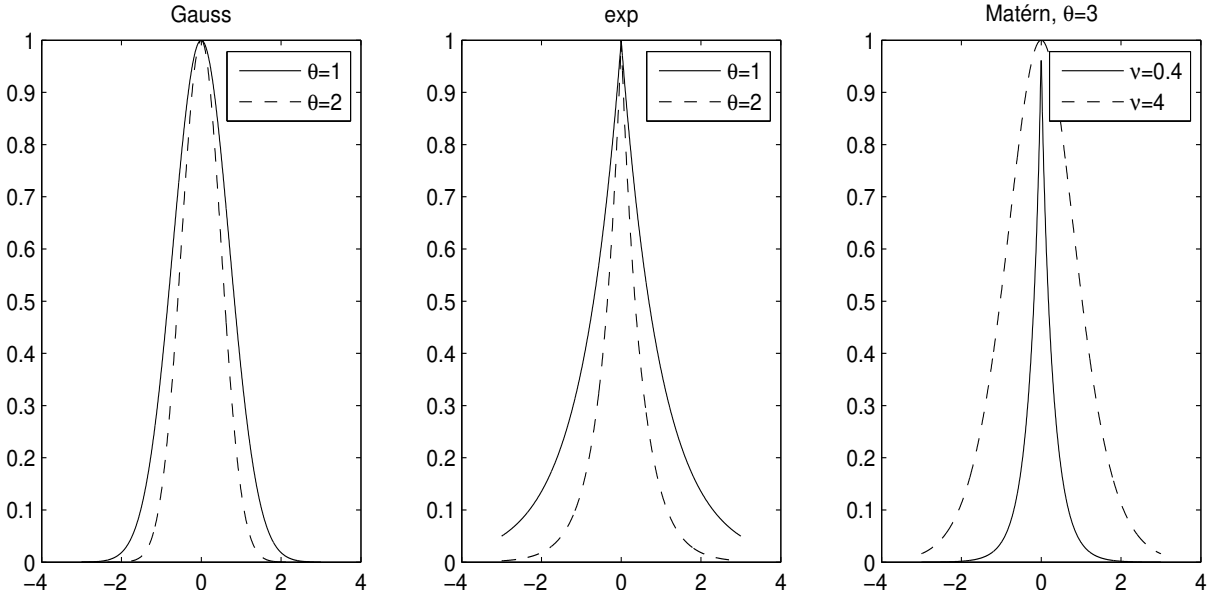

FIGURE 2.3 – Fonctions de covariance gaussienne (gauche), exponentielle (milieu) et Matérn (droite).

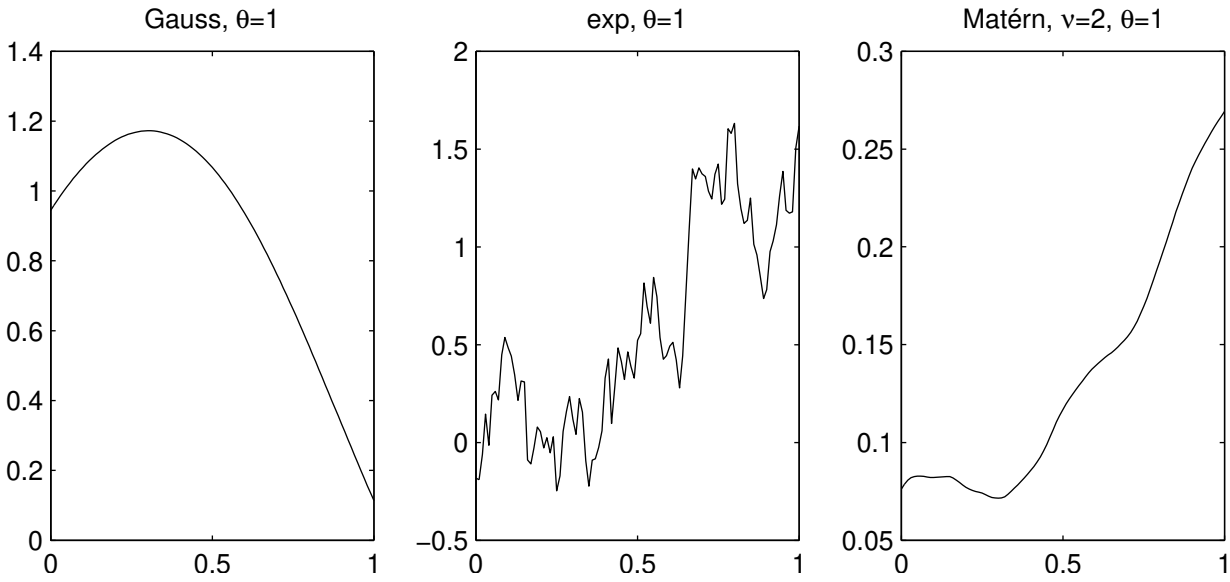

FIGURE 2.4 – Trajectoires d'un processus gaussien de moyenne nulle et fonction de covariance gaussienne (gauche), exponentielle (milieu) et Matérn (droite). Les trajectoires ont été obtenues avec la commande Matlab mvnrnd.

Pour une liste plus exhaustive de fonctions de corrélation stationnaires ou isotropes, nous renvoyons à [2].

Concernant le choix de la fonction de corrélation, il est parfois suggéré d'utiliser les fonctions de corrélation de Matérn, plus flexibles car le paramètre ν permet d'ajuster la régularité du processus que l'on ne connaît pas en général *a priori* [161] ; le nombre de paramètres à estimer est alors plus grand, ce qui rend l'optimisation plus coûteuse en temps de calcul (notons qu'il est aussi possible de fixer la valeur de ν à la régularité souhaitée [178]). De plus, l'évaluation de la fonction $K_\nu(\cdot)$ est sujette à des instabilités numériques pour des grandes valeurs de ν [142]. Nous verrons en 2.3 que la problématique liée à l'estimation des paramètres est différente selon la fonction de covariance utilisée.

2.2 Krigeage

Nous en arrivons à la technique du krigeage, qui consiste à faire de la prédiction optimale d'un processus gaussien (pour une extension à l'estimation d'intégrales de processus, voir [20]). Nous faisons donc tout d'abord quelques rappels concernant la prédiction, puis présentons le formalisme de la méthode et donnons quelques-unes de ses propriétés. Pour un bref aperçu historique de la méthode, revoir le paragraphe 1.4.1.

2.2.1 Prédiction

Faisons tout d'abord quelques rappels concernant la prédiction. On souhaite prédire une variable aléatoire Y_0 à partir d'un échantillon (non i.i.d.) $Y^n = {}^t(Y_1, \ldots, Y_n)$. Dans la suite, toutes les variables considérées sont supposées de carré intégrable.

Définition 2.2.1 *[127, 144] On appelle*
- statistique *: une variable aléatoire $g(Y_1, \ldots, Y_n)$, avec $g : \mathbb{R}^n \to \mathbb{R}$ une fonction borélienne (en pratique, g est une fonction continue).*
- prédicteur *(de Y_0 à partir de l'échantillon Y^n) : une statistique $\widehat{Y}_0 = \widehat{Y}_0(Y^n)$.*
- prédicteur linéaire *: un prédicteur qui s'écrit $\widehat{Y}_0(Y^n) = a_0 + {}^t a Y^n$, avec $a \in \mathbb{R}^n$.*
- prédicteur non biaisé *: un prédicteur sans biais par rapport à une famille donnée $\mathcal{F}$ de lois μ du couple (Y_0, Y^n). Autrement dit, $\mathbb{E}_\mu\{\widehat{Y}_0\} = \mathbb{E}_\mu\{Y_0\} \quad \forall \mu \in \mathcal{F}$, avec $\mathbb{E}_\mu\{\cdot\}$ l'espérance mathématique sous la loi μ.*

On notera LUP (Linear Unbiased Predictor) *tout prédicteur linéaire sans biais.*

La condition d'absence de biais dépend de la famille de lois $\mathcal{F}$ considérée (qui est en général paramétrique) : plus grande sera choisie cette famille de lois, plus restreinte sera la classe des prédicteurs sans biais.

Exemple 2.2.2 *[144] Supposons que $Y_i = \beta + \epsilon_i, 0 \leq i \leq n$, avec $\beta \neq 0$ connu et ϵ_i des variables aléatoires non corrélées, centrées de variance $\sigma^2 > 0$ inconnue, définit la famille de lois $\mathcal{F}$. On cherche un LUP $\widehat{Y}_0 = a_0 + {}^t a Y^n$. La condition de non biais s'écrit $\mathbb{E}\{\widehat{Y}_0\} = \mathbb{E}\{Y_0\} \quad \forall \sigma^2 > 0$, soit $a_0 + \beta \sum_{i=1}^n a_i = \beta$ puisque l'espérance ne dépend pas de σ^2. Des LUPs possibles s'obtiennent en prenant par exemple $a_0 = \beta$ et $\sum_{i=1}^n a_i = 0$, ou $a_0 = 0$ et $\sum_{i=1}^n a_i = 1$.*
Si on agrandit la famille $\mathcal{F}$, en considérant β quelconque et inconnu, on obtient $a_0 = 0$ car la condition de non biais doit être satisfaite pour $\beta = 0$, et $\sum_{i=1}^n a_i = 1$ car elle doit être verifiée aussi pour tout $\beta \neq 0$. On voit donc sur cet exemple simple que l'on a restreint la classe des LUPs admissibles en agrandissant la famille de lois $\mathcal{F}$, ou, inversement, qu'un LUP pour une famille de lois $\mathcal{F}$ est aussi un LUP pour une sous-famille de $\mathcal{F}$.

Afin de pouvoir choisir un prédicteur, on se donne un critère de comparaison. Le critère le plus utilisé est *l'erreur quadratique moyenne, EQM* (ou *mean squared error, MSE*).

Définition 2.2.3 *Supposons que (Y_0, Y^n) suit la loi μ.* L'erreur quadratique moyenne *du prédicteur $\widehat{Y}_0 = \widehat{Y}_0(Y^n)$ est donnée par*

$$\mathrm{EQM}_\mu\left[\widehat{Y}_0\right] = \mathbb{E}_\mu\left[\left(\widehat{Y}_0 - Y_0\right)^2\right]. \tag{2.16}$$

Un prédicteur intéressant aura une petite EQM. Idéalement, on souhaiterait un prédicteur $\widehat{Y}_0^*$ vérifiant

$$\widehat{Y}_0^* \in \underset{\widehat{Y}_0}{\operatorname{argmin}} \operatorname{EQM}_\mu \left[\widehat{Y}_0\right]. \tag{2.17}$$

Les prédicteurs minimisant l'EQM sont appelés *meilleurs prédicteurs (best predictors)*. Ce critère de choix du prédicteur dépend fortement de la loi μ, que l'on ne connait pas en pratique [144]. Il est donc souhaitable que (2.17) soit verifiée pour un grand nombre de lois μ.

Un théorème fondamental de la prédiction donne l'expression du meilleur prédicteur sous forme d'espérance conditionnelle.

Théorème 2.2.4 *[144] Supposons que (Y_0, Y^n) suit la loi μ, pour laquelle l'espérance conditionnelle de Y_0 sachant Y^n existe. Alors*

$$\widehat{Y}_0^* = \mathbb{E}_\mu(Y_0|Y^n)$$

est le meilleur prédicteur de Y_0.

En général, on ne sait pas calculer cette espérance conditionnelle. C'est pourquoi, dans la suite, on fera des hypothèses simplificatrices sur la loi du couple (Y_0, Y^n) qui sera supposée gaussienne : on sait alors que le meilleur prédicteur est linéaire. On restreint la classe des prédicteurs linéaires aux prédicteurs sans biais afin de pouvoir calculer le meilleur prédicteur lorsque les paramètres de la moyenne sont inconnus. On cherche donc à déterminer le meilleur LUP, appelé *BLUP* (Best Linear Unbiased Predictor) [137].

Remarque 2.2.5 *[161] Le BLUP peut être très éloigné du meilleur prédicteur si le processus n'est pas gaussien.*

2.2.2 Modèle

Le krigeage est un cas particulier du problème de prédiction présenté au paragraphe précédent, où les variables aléatoire $\{Y_i, 0 \leq i \leq n\}$ correspondent à des variables aléatoires $\{Y(x_i), 0 \leq i \leq n\}$, avec $Y(x, \omega)$ un processus gaussien.

Dans le cas d'une réponse déterministe, le système est modelisé par

$$Y(x) = {}^t m(x)\beta + Z(x), \tag{2.18}$$

où $m(\cdot)$ est une fonction connue à valeurs dans $\mathbb{R}^p$, β est un vecteur à p coefficients inconnus et $Z(\cdot)$ est un processus gaussien stationnaire, de moyenne nulle et fonction de covariance connue $k(\cdot, \cdot)$. Le modèle contient donc un terme de régression déterministe (la moyenne) et un terme aléatoire.

La fonction $m(\cdot)$ est déterminée par notre connaissance *a priori* du phénomène à modéliser, c'est en général un vecteur de fonctions monomiales. Si par exemple $x \in \mathbb{R}$ et $p = 3$, on pourrait avoir $m(x) = {}^t(1, x, x^2)$ et le modèle s'écrirait alors $Y(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + Z(x)$. Des auteurs préconisent l'utilisation d'une moyenne constante $m(x) = 1$, ayant observé que cela donne un prédicteur satisfaisant en pratique et simplifie les calculs [142] (le processus Y est donc supposé stationnaire).

Il est aussi possible d'utiliser la théorie de la sélection de modèle pour déterminer les termes de la moyenne, voir [73]. On prendra garde au fait que les critères classiques de type R^2 ou C_p de Mallows [86] ne s'appliquent pas ici, car les « erreurs » $Z(x)$ ne sont pas indépendantes.

Remarque 2.2.6

- *En pratique, la fonction de covariance $k(\cdot,\cdot)$ est inconnue mais paramétrique ;*
- *le processus aléatoire $Z(\cdot)$ sera supposé* régulier*, ou* non dégénéré*, c'est-à-dire vérifiant la condition suivante : quel que soit le choix des points d'observation $x_1,\dots,x_n$, la matrice de covariance $K=(k(x_i,x_j))_{1\leq i,j\leq n}$ est inversible. Sous l'hypothèse de stationnarité de $Z(\cdot)$, cela revient à prendre une fonction $k(\cdot)$ définie positive.*

Supposons que l'on observe $Y^n = {}^t(Y(x_1),\dots,Y(x_n))$ et que l'on souhaite prédire $Y_0 = Y(x_0)$ (notons que ce ne sont pas ici des observations i.i.d. mais les valeurs prises par la trajectoire Y_ω aux points $x_1,\dots,x_n$ pour une valeur de ω fixée). Avec un modèle de la forme (2.18), on peut, en utilisant l'hypothèse gaussienne, écrire la loi μ du couple (Y_0, Y^n).

Proposition 2.2.7 *[144] Sous le modèle (2.18),*

$$\begin{pmatrix} Y_0 \\ Y^n \end{pmatrix} \sim \mathcal{N}_{1+n}\left[\begin{pmatrix} {}^t m_0 \\ M \end{pmatrix}\beta, \sigma^2 \begin{pmatrix} 1 & {}^t r_0 \\ r_0 & R \end{pmatrix}\right], \tag{2.19}$$

avec $m_0 = m(x_0), M = {}^t(m(x_1)\dots m(x_n)), R = \mathrm{corr}(Y^n), r_0 = \mathrm{corr}(Y^n, Y_0)$, et σ^2 la variance (constante, par stationnarité) de Z.

Calculons maintenant le BLUP en x_0. Si on note μ la loi (2.19), on cherche un prédicteur linéaire de la forme

$$\widehat{Y}_0 = {}^t c_0 Y^n \tag{2.20}$$

(le terme constant est égal à 0 par le même argument de non biais que dans l'exemple 2.2.2). On va choisir un vecteur de poids c_0 qui minimise la quantité

$$\mathrm{EQM}_\mu[\widehat{Y}_0] = \mathbb{E}_\mu[{}^t c_0 Y^n - Y_0]^2,$$

sous la contrainte (de non biais)

$$\mathbb{E}_\mu[{}^t c_0 Y^n] = \mathbb{E}_\mu[Y_0],$$

qui se réécrit

$${}^t M c_0 = m_0.$$

On a

$$\begin{aligned}
\mathrm{EQM}_\mu[\widehat{Y}_0] &= \mathbb{E}_\mu\left[\left({}^t c_0(M\beta + Y^n) - ({}^t m_0\beta + Y_0)\right)^2\right] \\
&= \mathbb{E}_\mu\left[\left(({}^t c_0 M - {}^t m_0)\beta + {}^t c_0 Y^n - Y_0\right)^2\right] \\
&= \mathbb{E}_\mu\left[{}^t c_0 Y^n {}^t Y^n c_0 - 2\,{}^t c_0 Y^n Y_0 + Y_0^2\right] \\
&= \sigma^2\,{}^t c_0 R c_0 - 2\sigma^2\,{}^t c_0 r_0 + \sigma^2,
\end{aligned}$$

soit

$$\mathrm{EQM}_\mu[\widehat{Y}_0] = \sigma^2\left({}^t c_0 R c_0 - 2\,{}^t c_0 r_0 + 1\right). \tag{2.21}$$

La quantité que l'on doit minimiser est donc ${}^t c_0 R c_0 - 2\,{}^t c_0 r_0$. Le système à résoudre est donc la minimisation d'une fonctionnelle quadratique sous contraintes linéaires,

$$c_0 = \underset{a\in\mathbb{R}^n,\, {}^t M a = m_0}{\mathrm{argmin}} \; {}^t a R a - 2\,{}^t a r_0. \tag{2.22}$$

On utilise la technique des multiplicateurs de Lagrange [13] : on cherche $(c_0, \lambda_0) \in \mathbb{R}^{n+p}$ qui minimisent la quantité

$$ {}^t c_0 R c_0 - 2 {}^t c_0 r_0 + 2 {}^t \lambda_0 ({}^t M c_0 - m_0). $$

Cherchant les zéros du gradient, on obtient alors le système

$$ \begin{pmatrix} R & M \\ {}^t M & 0 \end{pmatrix} \begin{pmatrix} c_0 \\ \lambda_0 \end{pmatrix} = \begin{pmatrix} r_0 \\ m_0 \end{pmatrix}. $$

On montre que la solution de ce système est

$$ \begin{aligned} c_0 &= R^{-1} (r_0 - M\lambda_0) \ ; \\ \lambda_0 &= ({}^t M R^{-1} M)^{-1} \left({}^t M R^{-1} r_0 - m_0 \right). \end{aligned} $$

En injectant la valeur de c_0 dans (2.20) et (2.21), on obtient finalement l'équation condensée du BLUP en x_0.

Proposition 2.2.8 *[144, 161] Le BLUP au point x_0 pour le modèle (2.18) s'écrit*

$$ \widehat{Y}_0 = {}^t m_0 \widehat{\beta} + {}^t r_0 R^{-1} (Y^n - M\widehat{\beta}), \tag{2.23} $$

avec $\widehat{\beta} = ({}^t M R^{-1} M)^{-1} {}^t M R^{-1} Y^n$ l'estimateur des moindres carrés pondérés de $\widehat{\beta}$.
L'EQM correspondant vaut

$$ \mathrm{EQM}_\mu(x_0) = \sigma^2 \left(1 - {}^t r_0 R^{-1} r_0 + {}^t \gamma ({}^t M R^{-1} M)^{-1} \gamma \right), \tag{2.24} $$

avec $\gamma = m_0 - {}^t M R^{-1} r_0$.

Remarque 2.2.9 *[163] Sous certaines conditions, le prédicteur en x_0 dépend principalement des observations situées au voisinage de x_0 : le i^e terme du vecteur c_0 sera d'autant prépondérant que x_i est proche de x_0. Ce phénomène est appelé* effet d'écran (screening effect). *On retrouve un analogue des vecteurs de support dans le cas de la SVR (§ 1.4.4).*

Le BLUP (2.23) s'appelle aussi *prédicteur de krigeage*. Si on a posé $m(x) = 0$, on dit qu'on fait du *krigeage simple*. Si $m(x) = 1$, cela s'appelle du *krigeage ordinaire*, et si $m(x)$ a une forme plus générale c'est du *krigeage universel* ou du *krigeage à dérive externe*.

Remarque 2.2.10 *[31, 144] L'équation (2.23) du BLUP correspond bien à la forme (1.11) donnée par le théorème du représentant. En effet, pour tout $x \in \mathbb{R}^d$,*

$$ \widehat{Y}(x) = \sum_{j=1}^{p} \widehat{\beta}_j m_j(x) + \sum_{i=1}^{n} \alpha_i k(x - x_i), \tag{2.25} $$

avec $\alpha = {}^t(\alpha_1, \ldots, \alpha_n) = K^{-1}(Y^n - M\widehat{\beta})$. La régularité du prédicteur dépend donc de la régularité de la moyenne $m(\cdot)$ et de la covariance $k(\cdot)$, et la régularité aux points d'observation x_i dépend de la régularité de la covariance en 0. On remarque surtout qu'il n'est pas nécessaire de recalculer les coefficients du prédicteur en chaque point : après avoir évalué $\widehat{\beta}$ et α, on dispose en effet de la formule (2.25) qui donne la valeur du prédicteur en tout point x. Les coefficients α et $\widehat{\beta}$ sont solution du système de krigeage dual

$$ \begin{pmatrix} K & M \\ {}^t M & 0 \end{pmatrix} \begin{pmatrix} \alpha \\ \widehat{\beta} \end{pmatrix} = \begin{pmatrix} Y^n \\ 0 \end{pmatrix}, $$

identique à (1.12), qui caractérise le BLUP comme étant l'unique interpolateur de la forme (2.25) vérifiant

$$\begin{cases} \widehat{Y}(x_i) = Y(x_i) & \forall i = 1, \dots, n\,; \\ \sum_{i=1}^{n} \alpha_i m_i(x_j) = 0 & \forall j = 1, \dots, p. \end{cases}$$

L'EQM correspondant au BLUP est souvent appelée *variance de krigeage.*

Corollaire 2.2.11 *[144] Le prédicteur (2.23) est un interpolateur :* $\widehat{Y}(x_i) = Y(x_i) \quad \forall i = 1 \dots, n$. *De plus,* $\mathrm{EQM}_{\mu}(x_i) = 0 \quad \forall i = 1 \dots, n$.

Preuve *Supposons que* $x_0 = x_i$, *pour une valeur de* i *fixée dans* $1, \dots, n$. *Alors* $m_0 = m(x_i)$ *et* ${}^t r_0 = (\rho(x_i - x_1), \dots, \rho(x_i - x_n))$, *qui est la* i^e *ligne de* R. *Donc* $R^{-1} r_0 = {}^t(0, \dots, 0, 1, 0, \dots, 0) = e_i$, *le* i^e *vecteur canonique de* $\mathbb{R}^n$. *Ainsi* $\widehat{Y}(x_0) = {}^t m(x_i)\widehat{\beta} + {}^t e_i(Y^n - M\widehat{\beta}) = {}^t m(x_i)\widehat{\beta} + Y_i - {}^t m(x_i)\widehat{\beta} = Y_i$.
Par le même raisonnement, on obtient $\mathrm{EQM}_{\mu}(x_i) = \sigma^2 \left(1 - \rho(x_i, x_i) + {}^t\gamma({}^t M R^{-1} M)^{-1}\gamma\right)$, *avec* $\gamma = m(x_i) - {}^t M e_i = 0$, *c'est-à-dire* $\mathrm{EQM}_{\mu}(x_i) = 0$. □

Ceci est caractéristique d'un système déterministe (absence d'erreur de mesure d'un code informatique [141, 142]) : observer plusieurs fois en un même point x donne toujours la même valeur $Y(x)$, d'où l'absence d'erreur aux points déjà observés (on pourra cependant consulter [51] pour se convaincre qu'un type de bruit peut affecter la réponse d'un ordinateur si l'on utilise des méthodes itératives : le niveau de bruit est alors déterminé par la précision demandée à la méthode itérative).

Il est en fait facile de calculer la distribution conditionnelle de Y_0 à partir de la loi jointe (2.19), en utilisant le résultat suivant.

Proposition 2.2.12 *[161] Soit* $X = {}^t({}^t X_1, {}^t X_2)$ *un vecteur aléatoire de taille* $q = q_1 + q_2$, *de loi normale multivariée*

$$\mathcal{N}_q\left(\begin{pmatrix}\mu_1 \\ \mu_2\end{pmatrix}, \begin{pmatrix}\Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22}\end{pmatrix}\right).$$

Alors la distribution conditionnelle de X_1 *sachant* $X_2 = x_2$ *est la loi*

$$\mathcal{N}_{q_1}\left(\mu_1 + \Sigma_{12}\Sigma_{22}^{-}(x_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-}\Sigma_{21}\right),$$

avec Σ_{22}^{-} *une inverse généralisée quelconque de* Σ_{22} *(ou son unique inverse si* Σ_{22} *est inversible) [126].*

Remarque 2.2.13 *[144] La loi conditionnelle* $[Y_0 | Y^n]$ *est donc normale lorsque* β *et* $k(\cdot, \cdot)$ *sont connus. En pratique,* β *et* $k(\cdot, \cdot)$ *sont inconnus, et quand on a besoin de la distribution conditionnelle de* Y_0 *on utilise souvent la loi (fausse)*

$$\mathcal{N}(\widehat{Y}_0, \mathrm{EQM}_{\widehat{\mu}}(Y_0)), \tag{2.26}$$

avec $\widehat{Y}_0$ *et* $\mathrm{EQM}_{\widehat{\mu}}(Y_0)$ *donnés en (2.23) et (2.24) (la notation* $\widehat{\mu}$ *signalant que les paramètres inconnus du modèle ont été remplacés par une estimation). La vraie loi conditionnelle de* Y_0 *dépend des lois* a priori *que l'on met sur les paramètres (on entre alors dans le cadre bayésien, voir l'annexe C). La loi (2.26) est cependant généralement une bonne approximation de la distribution de la loi* $[Y_0 | Y^n]$ *lorsque les paramètres du modèle sont inconnus.*

Le corollaire suivant montre que sous l'hypothèse gaussienne, le BLUP (2.23) correspond au meilleur prédicteur si β est connu.

Corollaire 2.2.14 *[144, 161] Si β est connu dans le modèle (2.18), alors*

$$\mathbb{E}\left[Y_0|Y^n\right] = {}^t m_0\beta + {}^t r_0 R^{-1}(Y^n - M\beta). \tag{2.27}$$

Le théorème 2.2.4 nous dit alors que (2.27) est le meilleur prédicteur lorsque β est connu. On remarque que ce prédicteur varie avec β, ce qui explique pourquoi on utilise le BLUP, valable pour toute valeur de β (mais dépendant aussi de la fonction de covariance $k(\cdot,\cdot)$).
Le corollaire 2.2.14 rend donc légitime l'utilisation des prédicteurs linéaires dans le cas gaussien. Rappelons cependant que le meilleur prédicteur linéaire peut être très éloigné du meilleur prédicteur si le processus aléatoire n'est pas gaussien.

Remarque 2.2.15 *[161] On a supposé que seuls les paramètres de la covariance étaient connus pour obtenir l'expression du prédicteur (2.23) et de l'EQM (2.24), mais il aurait été possible de supposer aussi β connu. Dans ce cas, on obtient le BLP (2.27), dont la variance est donnée dans la proposition 2.2.12 et vaut $\sigma^2(1-{}^t r_0 R^{-1} r_0)$ (comparer avec (2.24), on remarque que la variance diminue, ce qui est raisonnable). Un argument de projection dans un espace de Hilbert permet de donner une caractérisation du BLP en terme d'orthogonalité. Soit $\mathcal{H}$ l'espace de Hilbert, completé de l'espace* $\mathrm{Vect}\{Y(x), x \in \mathcal{X}\}$ *pour le produit $\langle h,k\rangle = \mathbb{E}[hk]$. Le BLP est la projection orthogonale dans $\mathcal{H}$ de $Y(x_0)$ sur le sous-espace $\mathcal{G} = \mathrm{Vect}\{Y(x_1),\dots,Y(x_n)\}$: c'est donc l'unique élément g_0 de $\mathcal{G}$ vérifiant*

$$Y(x_0) - g_0 \perp g \qquad \forall g \in \mathcal{G},$$

l'erreur est orthogonale aux observations (voir [178] pour davantage de détails concernant le prédicteur de krigeage vu comme une projection).

En pratique, on ne peut pas utiliser tel quel le prédicteur (2.23) et l'EQM (2.24) car les paramètres de la fonction de covariance sont inconnus. On va donc les estimer puis les injecter dans les équations (2.23) et (2.24) obtenues pour une fonction de covariance connue (ce sont les paramètres de la loi normale utilisés dans la remarque 2.2.13). Ce prédicteur « plug-in » s'appelle *EBLUP (Empirical BLUP, BLUP empirique)* [34].
Nous utilisons donc un prédicteur du type

$$\widehat{Y_0} = {}^t m_0\widehat{\beta} + {}^t\widehat{r_0}\widehat{R}^{-1}(Y^n - M\widehat{\beta}), \tag{2.28}$$

avec $\widehat{\beta} = ({}^t M\widehat{R}^{-1}M)^{-1}{}^t M\widehat{R}^{-1}Y^n$, et $\widehat{R}$ et $\widehat{r}_0$ des estimations de R et r_0. Notons que l'appellation EBLUP est trompeuse, car $\widehat{R} = \widehat{R}(Y^n)$ et $\widehat{r}_0 = \widehat{r}_0(Y^n)$ sont en général non-linéaires en Y^n : les EBLUPs sont donc le plus souvent non linéaires, et mêmes biaisés.
L'EQM empirique « plug-in » correspondant est donné par (voir (2.24))

$$\mathrm{EQM}_{\widehat{\mu}}(x_0) = \widehat{\sigma^2}\left(1 - {}^t\widehat{r}_0\widehat{R}^{-1}\widehat{r}_0 + {}^t\widehat{\gamma}({}^t M\widehat{R}^{-1}M)^{-1}\widehat{\gamma}\right), \tag{2.29}$$

avec $\widehat{\gamma} = m_0 - {}^t M\widehat{R}^{-1}\widehat{r}_0$.

Remarque 2.2.16 *Les calculs présentés ci-dessus impliquent l'inversion de la matrice de corrélation R. Le conditionnement de la matrice de corrélation est donc très important, d'autant plus qu'il devient de plus en plus mauvais quand le nombre d'observations augmente, ce qui peut donner lieu à des résultats numériques très peu fiables. Il est montré dans [1] que le conditionnement de la matrice R est en général mauvais pour une fonction de corrélation gaussienne, et meilleur pour une fonction de corrélation exponentielle. Naturellement, ce conditionnement dépend aussi de la position des observations (le plan d'expériences).*

2.2.3 Estimation des paramètres

Nous voyons maintenant différentes façons d'estimer les paramètres du modèle [144, 161]. On notera ψ le vecteur des paramètres de la fonction de corrélation : par exemple, pour la fonction de corrélation exponentielle (2.13), on aura $\psi = (\theta_1, \ldots, \theta_n, p_1, \ldots, p_n)$.
Pour les méthodes utilisant le maximum de vraisemblance, on fera l'hypothèse que la matrice M est de rang plein p.

– **Maximum de vraisemblance** [8, 122]
La fonction de vraisemblance est la densité jointe des observations, vue comme une fonction des paramètres inconnus. L'estimation par *maximum de vraisemblance (MV, ou Maximum Likelihood, ML)* consiste à chercher une combinaison de valeurs des paramètres qui maximise la fonction de vraisemblance, ou de façon équivalente son logarithme, appelé *log-vraisemblance*.
On peut montrer que la log-vraisemblance correspondant à l'échantillon Y^n, sous le modèle (2.18), s'écrit à une constante près

$$l(\beta, \sigma^2, \psi | Y^n) = -\frac{1}{2}\left[n \log \sigma^2 + \log(\det(R)) + \frac{{}^t(Y^n - M\beta)R^{-1}(Y^n - M\beta)}{\sigma^2}\right]. \quad (2.30)$$

Le maximum est atteint en un point où les dérivées partielles sont nulles. En dérivant par rapport à β et égalant à zéro, on obtient

$$\widehat{\beta} = \widehat{\beta}(\psi) = ({}^tMR^{-1}M)^{-1}{}^tMR^{-1}Y^n. \quad (2.31)$$

On remarque que c'est aussi l'estimateur des moindres carrés pondérés de β quand R est connue (voir la proposition 2.2.8) ; ici, R dépend du paramètre inconnu ψ. Faisant de même avec σ^2, on obtient

$$\widehat{\sigma}^2 = \widehat{\sigma}^2(\psi) = \frac{1}{n}{}^t(Y^n - M\widehat{\beta})R^{-1}(Y^n - M\widehat{\beta}), \quad (2.32)$$

qui est un estimateur biaisé de σ^2 (on a tendance à sous-estimer la valeur du paramètre). En substituant $\widehat{\beta}(\psi)$ et $\widehat{\sigma}^2(\psi)$ dans (2.30), on obtient

$$l(\widehat{\beta}, \widehat{\sigma}^2, \psi | Y^n) = -\frac{1}{2}\left[n \log \widehat{\sigma}^2(\psi) + \log(\det(R(\psi))) + n\right],$$

qui ne dépend que de ψ. L'estimateur du MV de ψ s'écrit de façon compacte

$$\widehat{\psi} = \underset{\psi}{\operatorname{argmin}} \left[n \log \widehat{\sigma}^2(\psi) + \log\left(\det\left(R(\psi)\right)\right)\right], \quad (2.33)$$

avec $\widehat{\sigma}^2(\psi)$ défini en (2.32). Connaissant $\widehat{\psi}$, on peut finalement calculer aussi les estimateurs du MV $\widehat{\beta}$ et $\widehat{\sigma}^2$, et les injecter dans (2.28) pour obtenir l'*EBLUP du maximum de vraisemblance*.
Il faut en général $\mathcal{O}(n^3)$ calculs pour évaluer la vraisemblance, ce qui rend la méthode difficile à appliquer si le nombre d'observations n est grand [161]. Concernant la recherche du minimum de (2.33), si l'optimisation se fait sous contraintes relatives au paramètre ψ (par exemple si la forme de la covariance impose $\psi_1 > 0$, $0 < \psi_2 \leq 2$, ...), on peut effectuer une reparamétrisation qui permet une optimisation sans contrainte [143]. Il est possible que la fonction de vraisemblance soit multimodale pour certaines fonctions de covariance [115], mais nous n'avons pas constaté ce phénomène pour les covariances classiques présentées en 2.1.2.3. L'existence de maxima multiples de la vraisemblance n'est pas forcément un

problème, et peut simplement indiquer que les données ne permettent pas de choisir entre plusieurs valeurs des paramètres [161].

– **Maximum de vraisemblance restreint**

Le *maximum de vraisemblance restreint, MVR (Restricted Maximum Likelihood, RML)*, appelé aussi maximum de vraisemblance marginal, est une méthode visant à construire un estimateur moins biaisé des paramètres de la fonction de covariance : en effet, lors de l'estimation par MV, il y a un biais dû au fait que le paramètre inconnu β est remplacé par l'estimateur $\widehat{\beta}$ dans les équations (2.32) et (2.33). Le prix à payer pour cette diminution du biais est une plus grande variance des estimateurs.

L'idée est de filtrer le terme β des données : on va chercher une application linéaire $L : \mathbb{R}^n \to \mathbb{R}^{n-p}$, de rang plein $n-p$, dont le noyau contient $\mathrm{Im}(M)$. De cette façon, on va envoyer Y^n dans $\mathbb{R}^{n-p}$ orthogonalement à sa moyenne $M\beta$. Puisque les p colonnes de M sont supposées linéairement indépendantes, cette méthode consiste à choisir une matrice L, de taille $(n-p)\times n$ et de rang $n-p$, qui satisfasse l'égalité $LM = 0$ (on peut par exemple partir de la matrice de projection $P = I - M({}^tMM)^{-1}{}^tM$, dont on ne garde que $n-p$ lignes linéairement indépendantes [125] ; notons que l'estimateur du maximum de vraisemblance restreint n'est pas défini si $n \leq p$).

On applique ensuite la méthode du MV au vecteur des données transformées

$$W = LY^n \sim \mathcal{N}\left[LM\beta = 0, \sigma^2 LR(\psi){}^tL\right].$$

Les éléments de W sont appelés *contrastes*, ce sont des combinaisons linéaires des observations dont la loi jointe ne dépend pas de β. La matrice W contient p données de moins que Y^n (d'où l'augmentation de la variance des estimations), mais présente l'avantage de ne pas contenir le paramètre inconnu β (d'où la diminution du biais). Nous allons voir que l'estimation par MVR de ψ est indépendante de la matrice L choisie.

La log-vraisemblance des données transformées s'écrit, à une constante près,

$$l(\sigma^2, \psi | W) = -\frac{1}{2}\left[(n-p)\log\sigma^2 + \log\left(\det(LR{}^tL)\right) + \frac{{}^tW(LR{}^tL)^{-1}W}{\sigma^2}\right].$$

On peut montrer (voir [68]) que cette quantité est égale, à une constante près, à

$$-\frac{1}{2}\left[(n-p)\log\sigma^2 + \log\left(\det(R)\right) + \log\left(\det({}^tMR^{-1}M)\right) + \frac{{}^t(Y^n - M\widehat{\beta})R^{-1}(Y^n - M\widehat{\beta})}{\sigma^2}\right], \tag{2.34}$$

qui ne dépend pas de la matrice L, avec $\widehat{\beta} = \widehat{\beta}(\psi)$ l'estimateur du MV de β défini en (2.31). Le maximum est atteint en un point où les dérivées partielles sont nulles. En dérivant par rapport à σ^2 et égalant à zéro, on obtient l'estimateur du MVR de σ^2,

$$\widetilde{\sigma^2} = \widetilde{\sigma^2}(\psi) = \frac{1}{n-p}{}^t(Y^n - M\widehat{\beta})R^{-1}(Y^n - M\widehat{\beta}). \tag{2.35}$$

Remarque 2.2.17 *On a $\widetilde{\sigma^2} = [n/(n-p)]\widehat{\sigma^2}$, avec $\widehat{\sigma^2}$ l'estimateur du MV (2.32). On retrouve bien le terme correctif $n/(n-p)$ utilisé habituellement pour construire un estimateur non biaisé de σ^2.*

En substituant $\widetilde{\sigma^2}(\psi)$ dans (2.34), on obtient, à une constante multiplicative près,

$$l(\widetilde{\sigma^2}, \psi | Y^n) = -\frac{1}{2}\left[(n-p)\log\widetilde{\sigma^2}(\psi) + \log\left(\det\left(R(\psi)\right)\right) + \log\left(\det\left({}^tMR^{-1}(\psi)M\right)\right) + n - p\right],$$

qui ne dépend que de ψ. L'estimateur du MVR de ψ s'écrit de façon compacte

$$\widetilde{\psi} = \underset{\psi}{\operatorname{argmin}}\left[(n-p)\log\widetilde{\sigma^2}\left(\psi\right) + \log\left(\det\left(R\left(\psi\right)\right)\right) + \log\left(\det\left({}^tMR^{-1}\left(\psi\right)M\right)\right)\right], \quad (2.36)$$

avec $\widetilde{\sigma^2}(\psi)$ défini en (2.35). Connaissant $\widetilde{\psi}$, on peut calculer $\widetilde{\sigma^2}$ et $\widetilde{\beta} = \widehat{\beta}(\widetilde{\psi})$ et les injecter dans (2.28) pour obtenir l'*EBLUP du maximum de vraisemblance resteint.*
L'avantage du MVR est qu'il peut être utilisé aussi pour le krigeage intrinsèque (voir l'annexe E), alors que le MV n'est pas applicable [161] : en effet, une covariance généralisée est seulement définie pour les combinaisons linéaires admissibles des observations, qui correspondent aux contrastes du MVR.

– **Validation croisée**
Notons ϕ le vecteur des paramètres à estimer. La méthode d'estimation de ϕ par *validation croisée ordinaire (ordinary cross-validation, OCV)* consiste à minimiser une estimation de l'erreur de prédiction moyenne du modèle,

$$\mathrm{EPE}(\widehat{Y}) = \int_{\mathcal{X}} \mathbb{E}\left[\left(\widehat{Y}(x) - Y(x)\right)^2\right] \mathrm{d}x.$$

L'estimation de l'EPE s'obtient en faisant la somme des erreurs commises aux x_i quand on les prédit en utilisant les $n-1$ données restantes [181]. On cherche alors

$$\widehat{\phi} = \underset{\phi}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{y}_{-i}(\phi) - y_i\right)^2, \quad (2.37)$$

avec, pour $i = 1, \ldots, n$, $\widehat{y}_{-i}(\phi)$ la prédiction de $Y(x_i)$ obtenue à partir de toutes les données d'apprentissage excepté (x_i, y_i), en utilisant la formule (2.23). Le modèle retenu sera celui ayant la meilleure capacité de prédiction.
Une généralisation utilisant des groupes de données, appelée *l-fold cross-validation*, fonctionne de la façon suivante [70] :

1. partager les n données en l groupes de taille à peu près égale, ce qui revient à se donner une application $\Pi : \{1, \ldots, n\} \longrightarrow \{1, \ldots, l\}$ qui détermine à quel groupe appartient la i^e observation, pour $i = 1, \ldots, n$;
2. calculer l'estimation par validation croisée de l'erreur de prédiction,

$$\mathrm{CV}_l(\phi) = \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{y}_{-\Pi(i)}(\phi) - y_i\right)^2,$$

avec $\widehat{y}_{-\Pi(i)}$ la prédiction de $Y(x_i)$ obtenue à partir de toutes les données d'apprentissage, excepté $\{(x_j, y_j), \Pi(j) = \Pi(i)\}$, les données appartenant au même groupe que l'observation i ;
3. calculer l'estimateur de validation croisée de ϕ,

$$\widehat{\phi} = \underset{\phi}{\operatorname{argmin}}\, \mathrm{CV}_l(\phi).$$

Le cas $l = n$, appelé *leave-one-out cross-validation*, correspond à la formule (2.37), avec $\Pi(i) = i \quad \forall i = 1, \ldots, n$.
Le choix de la constante l n'est pas évident : si $l = n$, CV_l est un estimateur asymptotiquement non biaisé de l'EPE, mais peut avoir une grande variance. Pour de plus petites valeurs de l, CV_l a une variance plus petite, mais peut être très biaisé si l'EPE varie beaucoup pour un nombre de données voisin de $n - n/l$ (le nombre de données utilisées pour la prédiction). En pratique, on utilise souvent $l = 5$ ou $l = 10$.
Pour d'autres extensions de la validation croisée, on pourra consulter [70, 182].
Remarque 2.2.18 *[31, 161] Dans le cas d'un modèle de krigeage, la validation croisée ne permet pas d'estimer le paramètre de variance σ^2, car celui-ci n'intervient pas dans la formule du prédicteur (2.23), et donc pas non plus dans l'équation (2.37). Une façon de prendre en compte l'ensemble des paramètres est d'évaluer*

$$\frac{1}{n}\sum_{i=1}^{n}\frac{(\widehat{y}_{-i}(\phi) - y_i)^2}{\widehat{\mathrm{EQM}}_{-i}(\phi)}, \tag{2.38}$$

avec $\widehat{\mathrm{EQM}}_{-i}$ l'EQM empirique en x_i obtenue à partir de toutes les données d'apprentissage excepté (x_i, y_i), en utilisant la formule (2.24). On cherchera les valeurs de ϕ telles que (2.38) soit proche de 1 (l'idée est que le numérateur est en moyenne égal au dénominateur).

– **EBLUP bayésien**
Par une mise en perspecive bayésienne du krigeage (voir l'annexe C), on obtient un prédicteur qui prend en compte l'information *a priori* dont on dispose sur la répartition des paramètres β, σ^2, ψ.
Après s'être donné une densité de probabilité $f(\beta, \sigma^2, \psi)$ résumant les valeurs *a priori* plus ou moins pertinentes des paramètres, on calcule le *prédicteur du mode a posteriori*,

$$(\widehat{\beta}, \widehat{\sigma^2}, \widehat{\psi}) = \underset{(\beta,\sigma^2,\psi)}{\operatorname{argmax}} \left[l(\beta, \sigma^2, \psi | Y^n) + \log f(\beta, \sigma^2, \psi) \right].$$

On construit ensuite l'*EBLUP du mode a posteriori (posterior mode EBLUP)* en injectant ces valeurs $\widehat{\beta}, \widehat{\sigma^2}, \widehat{\psi}$ dans l'équation (2.28). On note que la complexité algorithmique est du même ordre que pour le maximum de vraisemblance.

Une étude empirique présentée dans [144] tend à montrer qu'il vaut mieux utiliser le MV ou le MVR pour obtenir un bon prédicteur de krigeage (voir aussi [202] pour une comparaison des prédicteurs). De plus, on sait que sous certaines hypothèses de dérivabilité de la vraisemblance, l'estimateur du MV est asymptotiquement efficace (voir le § 2.3.1). On peut aussi utiliser le maximum de vraisemblance pénalisé, voir le § 2.3.2. Un des inconvénients des méthodes utilisant la vraisemblance est le coût de calcul algorithmique élevé quand le nombre de données n est grand, ainsi que l'instabilité liée à l'inversion de la matrice de covariance de taille $n \times n$. Il existe des méthodes d'approximation de la vraisemblance [166] ; on peut aussi utiliser des méthodes d'approximation d'un processus gaussien par un *champ aléatoire markovien gaussien (Gaussian Markov Random Field, GMRF)* [139], où les matrices sont creuses, ce qui permet d'économiser du temps de calcul. Pour une liste exhaustive de méthodes d'approximation, voir [136].

Remarque 2.2.19 *Dans notre présentation, nous avons fait l'hypothèse la plus communément utilisée d'un processus gaussien stationnaire, et nous avons vu au § 2.1.2 que cette hypothèse permet d'obtenir une justification théorique de la validité de l'inférence statistique. Notons cependant qu'il est possible d'utiliser d'autres types de fonctions de covariance : covariance généralisée (ce qui conduit au krigeage intrinsèque, voir l'annexe E), et, plus généralement,*

covariance non stationnaire, pour la construction et l'utilisation desquelles nous renvoyons à [123, 165, 193].

2.3 Problèmes liés à l'estimation des paramètres

Dans ce paragraphe nous commençons par rappeler de « bonnes » propriétés que l'on peut attendre d'un estimateur, puis les propriétés asymptotiques de l'estimateur du maximum de vraisemblance. Finalement, on constatera qu'une mauvaise estimation des paramètres, ou un mauvais choix de la fonction de covariance, n'a pas toujours des conséquences sur la qualité du prédicteur de krigeage. Comme dans le livre de M.L. Stein [161] d'où sont issus la plupart des résultats théoriques concernant le prédicteur de krigeage, les hypothèses « assez générales » des théorèmes ne sont pas détaillées à chaque fois par souci de clarté, car cela alourdirait énormément le propos.

2.3.1 Identifiabilité, consistance, efficacité d'un estimateur

Le problème de la possibilité ou non d'estimer correctement le vecteur des paramètres ϕ d'un processus gaussien

$$\left\{Y(x,\omega), x \in \mathcal{X} \subset \mathbb{R}^d, \omega \in (\Omega, \mathcal{T}, \mathcal{P}_\phi)\right\},$$

ainsi que les propriétés des estimateurs, sont liées à la mesure de probabilité $\mathcal{P}_\phi$ caractérisant le processus, appelée *mesure gaussienne*. Remarquons qu'une mesure gaussienne est caractérisée par une fonction définie sur $\mathcal{X}$ (la moyenne du processus) et une fonction semi-définie positive définie sur $\mathcal{X} \times \mathcal{X}$ (la covariance du processus).

2.3.1.1 Identifiabilité

On voudrait, à partir des observations, être capable de déterminer laquelle des mesures $(\mathcal{P}_\phi)_{\phi\in\Phi}$ est la bonne, avec Φ l'ensemble des valeurs possibles des paramètres.

Définition 2.3.1 *[161] Soient $\mathcal{P}_0$ et $\mathcal{P}_1$ deux mesures de probabilité sur un espace mesurable $(\Omega, \mathcal{T})$. Alors :*

- *$\mathcal{P}_0$ est dite* absolument continue *par rapport à $\mathcal{P}_1$ si*

$$\forall A \in \mathcal{T}, \mathcal{P}_1(A) = 0 \Longrightarrow \mathcal{P}_0(A) = 0.$$

  On note alors $\mathcal{P}_0 \ll \mathcal{P}_1$;
- *$\mathcal{P}_0$ et $\mathcal{P}_1$ sont dites* équivalentes *si $\mathcal{P}_0 \ll \mathcal{P}_1$ et $\mathcal{P}_1 \ll \mathcal{P}_0$. On note alors $\mathcal{P}_0 \equiv \mathcal{P}_1$;*
- *$\mathcal{P}_0$ et $\mathcal{P}_1$ sont dites* orthogonales *(ou* étrangères*), si*

$$\exists A \in \mathcal{T}, \mathcal{P}_0(A) = 1 \text{ et } \mathcal{P}_1(A) = 0.$$

  On note alors $\mathcal{P}_0 \perp \mathcal{P}_1$.

Si l'on sait que la vraie mesure est $\mathcal{P}_0$ ou $\mathcal{P}_1$, alors si $\mathcal{P}_0 \perp \mathcal{P}_1$, il est possible de déterminer laquelle est la bonne avec probabilité 1 par l'observation d'un $\omega \in \Omega$. Si $\mathcal{P}_0 \equiv \mathcal{P}_1$, alors peu importe ce que l'on observe, il est impossible de déterminer quelle mesure est la bonne avec probabilité 1. Si $\mathcal{P}_0$ et $\mathcal{P}_1$ ne sont ni équivalentes, ni orthogonales, alors cela dépend de ce que l'on observe, comme illustré par l'exemple suivant.

Exemple 2.3.2 *[161] Soit* $\Omega = \{-1, 0, 1\}, \mathcal{P}_0 = \frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_0$ *et* $\mathcal{P}_1 = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1$. $\mathcal{P}_0$ *et* $\mathcal{P}_1$ *ne sont ni orthogonales, ni équivalentes. On peut savoir quelle mesure est la bonne si* $\omega = \{-1\}$ *ou* $\omega = \{1\}$, *mais pas si* $\omega = \{0\}$.

Définition 2.3.3 *[175] Soit* ϕ_0 *la vraie valeur du paramètre. La mesure* $\mathcal{P}_{\phi_0}$ *est* identifiable *si*

$$\mathcal{P}_\phi \neq \mathcal{P}_{\phi_0} \quad \forall\, \phi \neq \phi_0.$$

Par extension, le paramètre ϕ_0 *est dit identifiable.*

On peut montrer, sous des hypothèses très générales, que deux mesures gaussiennes sont soit équivalentes, soit orthogonales.

Théorème 2.3.4 *[161] Soient* $m_0(\cdot)$ *et* $m_1(\cdot)$ *des fonctions continues définies sur* $\mathbb{R}^d$, $k_0(\cdot,\cdot)$ *et* $k_1(\cdot,\cdot)$ *des fonctions continues semi-définies positives sur* $\mathbb{R}^d \times \mathbb{R}^d$, *et* $\mathcal{X}$ *un sous-ensemble fermé de* $\mathbb{R}^d$. *Soient* $\mathcal{P}_0$ *et* $\mathcal{P}_1$ *les mesures gaussiennes définies respectivement par les* structures du second ordre (m_0, k_0) *et* (m_1, k_1), *caractérisant un processus gaussien défini sur* $\mathcal{X}$. *Alors*

$$\mathcal{P}_0 \perp \mathcal{P}_1 \text{ ou } \mathcal{P}_0 \equiv \mathcal{P}_1.$$

On peut déterminer l'orthogonalité ou l'équivalence de deux mesures gaussiennes en utilisant la *divergence de Kullback*.

Définition 2.3.5 *[161, 175] Soient* $\mathcal{P}_0$ *et* $\mathcal{P}_1$ *deux mesures de probabilité, de densités respectives* $p_0(\cdot)$ *et* $p_1(\cdot)$ *par rapport à la mesure de Lebesgue. On appelle* divergence de Kullback-Leibler *entre* $\mathcal{P}_0$ *et* $\mathcal{P}_1$ *la quantité*

$$\mathcal{K}(\mathcal{P}_0, \mathcal{P}_1) = \mathcal{K}(p_0, p_1) = \mathbb{E}_{\mathcal{P}_1}\left[\log \frac{p_1}{p_0}\right].$$

La divergence de Kullback est parfois appelée *distance de Kullback*, bien qu'elle n'ait pas les propriétés mathématiques d'une distance.

Proposition 2.3.6 *[161] (mêmes hypothèses que le théorème 2.3.4). Soit* $(x_i)_{i\in\mathbb{N}}$ *une suite d'éléments distincts et denses dans* $\mathcal{X}$, p_0^i *et* p_1^i *les densités respectives du vecteur* $(Y(x_1), \ldots, Y(x_i))$ *sous* $\mathcal{P}_0$ *et* $\mathcal{P}_1$. *Alors*

- *si* $\mathcal{K}(p_0^i, p_1^i) + \mathcal{K}(p_1^i, p_0^i) \underset{i\to\infty}{\longrightarrow} \infty, \quad \mathcal{P}_0 \perp \mathcal{P}_1$;
- *si* $\mathcal{K}(p_0^i, p_1^i) + \mathcal{K}(p_1^i, p_0^i) \underset{i\to\infty}{\longrightarrow} l < \infty, \quad \mathcal{P}_0 \equiv \mathcal{P}_1$.

Preuve du théorème 2.3.4 *Il suffit de montrer que la divergence de Kullback symétrisée*

$$\mathcal{K}_S(p_0^i, p_1^i) = \mathcal{K}(p_0^i, p_1^i) + \mathcal{K}(p_1^i, p_0^i)$$

est une fonction croissante de i, *et a donc une limite finie ou infinie quand* i *tend vers l'infini : on se retrouve alors dans un des deux cas de la proposition 2.3.6 (le résultat est mentionné dans [22], et la démonstration suivante permet de s'en convaincre). Utilisant la densité conditionnelle, on peut écrire*

$$p_l^{i+1}(t_1, \ldots, t_{i+1}) = p_l^i(t_1, \ldots, t_i) q_l^i(t_{i+1}), \quad l = 1, 2,$$

avec $q_l^i(\cdot)$ la densité conditionnelle de $Z(t_{i+1})$ par rapport à $\{Z(t_1),\dots,Z(t_i)\}$ sous la loi $\mathcal{P}_l$. On a donc

$$\begin{aligned}
\mathcal{K}_S(p_0^{i+1},p_1^{i+1}) &= \int_{\mathbb{R}^{i+1}}\left(\log\frac{p_1^{i+1}}{p_0^{i+1}}p_1^{i+1}+\log\frac{p_0^{i+1}}{p_1^{i+1}}p_0^{i+1}\right)\\
&= \int_{\mathbb{R}^{i+1}}\left(\log\frac{p_1^i q_1^i}{p_0^i q_0^i}p_1^i q_1^i+\log\frac{p_0^i q_0^i}{p_1^i q_1^i}p_0^i q_0^i\right)\\
&= \int_{\mathbb{R}^i}p_1^i\int_{\mathbb{R}}\log\frac{p_1^i q_1^i}{p_0^i q_0^i}q_1^i+\int_{\mathbb{R}^i}p_0^i\int_{\mathbb{R}}\log\frac{p_0^i q_0^i}{p_1^i q_1^i}q_0^i\\
&= \int_{\mathbb{R}^i}p_1^i\log\frac{p_1^i}{p_0^i}\underbrace{\int_{\mathbb{R}}q_1^i}_{=1}+\int_{\mathbb{R}^i}p_1^i\int_{\mathbb{R}}\log\frac{q_1^i}{q_0^i}q_1^i+\int_{\mathbb{R}^i}p_0^i\log\frac{p_0^i}{p_1^i}\underbrace{\int_{\mathbb{R}}q_0^i}_{=1}+\int_{\mathbb{R}^i}p_0^i\int_{\mathbb{R}}\log\frac{q_0^i}{q_1^i}q_0^i\\
&= \mathcal{K}_S(p_0^i,p_1^i)+\int_{\mathbb{R}^i}p_1^i\int_{\mathbb{R}}\log\frac{q_1^i}{q_0^i}q_1^i+\int_{\mathbb{R}^i}p_0^i\int_{\mathbb{R}}\log\frac{q_0^i}{q_1^i}q_0^i\\
&\geq \mathcal{K}_S(p_0^i,p_1^i)+\int_{\mathbb{R}^i}\min(p_1^i,p_0^i)\int_{\mathbb{R}}\underbrace{(\log q_1^i-\log q_0^i)(q_1^i-q_0^i)}_{\geq 0}\\
&\geq \mathcal{K}_S(p_0^i,p_1^i).
\end{aligned}$$

□

La proposition 2.3.6 est utilisée dans [6] pour déterminer l'équivalence ou l'orthogonalité des mesures gaussiennes définies par les fonctions de covariance classiques. Le processus est supposé de moyenne nulle, défini sur un domaine rectangulaire, et de fonction de covariance s'écrivant comme produit de fonctions de covariances mono-dimensionnelles. Les points x_i sont placés pour former une grille régulière, ce qui simplifie les calculs.

Définition 2.3.7 *[161] Soit $\{\mathcal{P}_\phi,\phi\in\Phi\}$ une famille de mesures de probabilité pour un processus aléatoire défini sur un domaine borné $\mathcal{X}$. On dit que ϕ est* micro-ergodique *si*

$$\forall\,\phi,\phi'\in\Phi,\phi\neq\phi'\Longrightarrow\mathcal{P}_\phi\perp\mathcal{P}_{\phi'}.$$

Dans le cas de mesures gaussiennes, sous des hypothèses très générales, l'identifiabilité du paramètre est équivalente à sa micro-ergodicité (théorème 2.3.4). La micro-ergodicité permet d'espérer pouvoir déduire les valeurs des paramètres à partir des observations issues d'une seule trajectoire du processus, mais ce n'est pas une condition suffisante.

2.3.1.2 Consistance

La *consistance* d'une suite d'estimateurs, traduction hasardeuse du terme anglais *consistency*, est en fait reliée aux propriétés de *convergence* de cette suite.

Définition 2.3.8 *[175] Soit $\widehat{\phi}_n$ l'estimateur utilisé pour estimer le paramètre ϕ à partir d'un n-échantillon $X_1,\dots,X_n$. La suite des estimateurs $(\widehat{\phi}_n)_n$ est dite :*

– asymptotiquement consistante *si*

$$\forall\,\phi,\quad\widehat{\phi}_n\underset{n\to\infty}{\longrightarrow}\phi\ \textit{en}\ \mathcal{P}_\phi\textit{-probabilité}\ ;$$

– asymptotiquement fortement consistante *si*

$$\forall\,\phi,\quad\widehat{\phi}_n\underset{n\to\infty}{\longrightarrow}\phi\quad\mathcal{P}_\phi\textit{-p.s.}$$

La consistance de l'estimateur d'un paramètre est une propriété importante, qui indique que le paramètre est bien estimé quand le nombre d'observations est grand.

Exemple 2.3.9 *Soit X une variable aléatoire dont la moyenne $\mathbb{E}[X]$ existe. L'estimateur de la* moyenne empirique

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

construit à partir de n réalisations indépendantes $X_1, \ldots, X_n$ de X, est fortement consistant par application de la loi forte des grands nombres.

Si un paramètre n'est pas identifiable, il ne peut exister d'estimateur consistant du paramètre. Cependant, la micro-ergodicité du paramètre n'implique pas l'existence d'une suite d'estimateurs consistante [161]. Nous verrons en 2.3.2 des cas où l'estimation des paramètres est consistante.

2.3.1.3 Efficacité

Définissons finalement la notion d'*efficacité* d'un estimateur, qui traduit le fait que celui-ci est de variance minimale. Supposons que l'on observe un vecteur aléatoire X à valeurs dans $\mathcal{X} \subset \mathbb{R}^d$, dont la distribution est issue d'une famille de lois distinctes $\{\mathcal{P}_\phi, \phi \in \Phi\}$, avec Φ un ouvert de $\mathbb{R}^p$ et $\phi = (\phi_1, \ldots, \phi_p)$. On se propose d'estimer ϕ par maximum de vraisemblance. On note $f_\phi(\cdot)$ la densité de X sous ϕ par rapport à la mesure de Lebesgue, et $l(\phi|x)$ la log-vraisemblance correspondant à l'observation x. Pour simplifier la présentation, on supposera que l'estimateur du MV $\widehat{\phi}$ de ϕ est non biaisé, et que, $\forall(\phi, x) \in \Phi \times \mathcal{X}$, $f_\phi(x) > 0$ et $\partial f_\phi(x)/\partial\phi_i$ existe et est finie pour tout i.

Définition 2.3.10 *[96, 161] La* fonction score *est définie sur $\Phi \times \mathcal{X}$ par*

$$\begin{aligned} S(\phi, x) &= \left(\frac{\partial l(\phi|x)}{\partial \phi_i}\right)_{i=1,\ldots,p} \\ &= \left(\frac{1}{f_\phi(x)}\frac{\partial f_\phi(x)}{\partial \phi_i}\right)_{i=1,\ldots,p}. \end{aligned} \tag{2.39}$$

Il s'agit donc du vecteur des dérivées partielles de la log-vraisemblance.

Puisque Φ est ouvert par hypothèse, tout estimateur du MV $\widehat{\phi}$ situé dans Φ est solution des *équations du score*

$$S(\widehat{\phi}, x) = 0.$$

En observant la deuxième égalité (2.39), on remarque que le score mesure le taux relatif auquel varie la densité f_ϕ en x.

Définition 2.3.11 *[96] Supposons que le score est de carré intégrable pour tout ϕ. On appelle* matrice d'information de Fisher *la matrice*

$$\mathcal{I}(\phi) = \mathbb{E}_{\mathcal{P}_\phi}\left[S(\phi, X)^t S(\phi, X)\right],$$

la moyenne du carré du score.

On peut raisonnablement penser que plus cette moyenne est grande en $\phi_0 \in \Phi$, le plus précisément on peut estimer ϕ_0 si c'est la vraie valeur du paramètre.

Proposition 2.3.12 *[96] Si les dérivées partielles de la fonction $\phi \mapsto \int f_\phi(x)\,\mathrm{d}x$ peuvent s'obtenir en dérivant sous le signe intégral, alors le score est nul en moyenne,* i.e.

$$\mathbb{E}_{\mathcal{P}_\phi}\left[S(\phi, X)\right] = 0.$$

Preuve

$$\begin{aligned}
\mathbb{E}_{\mathcal{P}_\phi}\left[S(\phi, X)\right] &= \left(\int_{\mathcal{X}} \frac{1}{f_\phi(x)} \frac{\partial f_\phi(x)}{\partial \phi_i} f_\phi(x)\,\mathrm{d}x\right)_{i=1,\dots,p} \\
&= \left(\int_{\mathcal{X}} \frac{\partial f_\phi(x)}{\partial \phi_i}\,\mathrm{d}x\right)_{i=1,\dots,p} \\
&= \left(\frac{\partial}{\partial \phi_i}\int_{\mathcal{X}} f_\phi(x)\,\mathrm{d}x\right)_{i=1,\dots,p} \\
&= \left(\frac{\partial}{\partial \phi_i} 1\right)_{i=1,\dots,p} \\
&= 0.
\end{aligned}$$

□

Ceci se traduit par le fait que la valeur du paramètre ϕ maximise en moyenne la log-vraisemblance pour tout ϕ.

Corollaire 2.3.13 *[96] Sous les mêmes hypothèses qu'à la proposition 2.3.12,*

$$\mathcal{I}(\phi) = \mathrm{cov}_{\mathcal{P}_\phi}\left[S(\phi), S(\phi)\right].$$

La matrice d'information est donc la matrice de covariance du score : on en déduit que dans ce cas la matrice d'information est semi-définie positive.

Voyons comment varie l'information quand le nombre d'observations indépendantes augmente.

Proposition 2.3.14 *[96] Reprenons les hypothèses de la proposition 2.3.12. Supposons de plus que les dérivées partielles secondes de la fonction $\phi \mapsto \int f_\phi(x)\,\mathrm{d}x$ peuvent s'obtenir en dérivant sous le signe intégral. Alors,*

$$\mathcal{I}(\phi) = -\mathbb{E}_{\mathcal{P}_\phi}\left[\frac{\partial^2 l(\phi|X)}{\partial \phi_i \partial \phi_j}\right]_{1\le i,j\le p}.$$

Preuve *Soient $i, j \in \{1, \dots, p\}$.*

$$\begin{aligned}
\frac{\partial^2 l(\phi|x)}{\partial \phi_i \partial \phi_j} &= \frac{\partial}{\partial \phi_i}\left(\frac{1}{f_\phi(x)}\frac{\partial f_\phi(x)}{\partial \phi_j}\right) \\
&= \frac{1}{f_\phi(x)}\frac{\partial^2 f_\phi(x)}{\partial \phi_i \partial \phi_j} - \frac{1}{f_\phi(x)^2}\frac{\partial f_\phi(x)}{\partial \phi_i}\frac{\partial f_\phi(x)}{\partial \phi_j} \\
&= \frac{1}{f_\phi(x)}\frac{\partial^2 f_\phi(x)}{\partial \phi_i \partial \phi_j} - \frac{\partial l(\phi|x)}{\partial \phi_i}\frac{\partial l(\phi|x)}{\partial \phi_j}.
\end{aligned}$$

Donc,

$$\begin{aligned}
-\mathbb{E}_{\mathcal{P}_\phi}\frac{\partial^2 l(\phi|X)}{\partial \phi_i \partial \phi_j} &= \int_{\mathcal{X}} \frac{\partial l(\phi|x)}{\partial \phi_i}\frac{\partial l(\phi|x)}{\partial \phi_j} f_\phi(x)\mathrm{d}x - \int_{\mathcal{X}} \frac{\partial^2 f_\phi(x)}{\partial \phi_i \partial \phi_j}\,\mathrm{d}x \\
&= \mathbb{E}_{\mathcal{P}_\phi}\frac{\partial l(\phi|X)}{\partial \phi_i}\frac{\partial l(\phi|X)}{\partial \phi_j} - \frac{\partial^2}{\partial \phi_i \partial \phi_j} 1 \\
&= \mathbb{E}_{\mathcal{P}_\phi}\frac{\partial l(\phi|X)}{\partial \phi_i}\frac{\partial l(\phi|X)}{\partial \phi_j},
\end{aligned}$$

et on retrouve bien le terme (i, j) de la matrice d'information. □

La matrice $\mathcal{I}(\phi)$ mesure l'information apportée par une seule observation de X. Sous les hypothèses de la proposition 2.3.12, l'information de Fisher est additive [96] : si $X_1, \ldots, X_n$ sont i.i.d. de même loi que X, l'information mutuelle apportée par ces observations est égale à n fois l'information individuelle. En effet,

$$\begin{aligned}
\mathcal{I}_n(\phi) &= -\mathbb{E}_{\mathcal{P}_\phi}\left[\frac{\partial^2 l(\phi|X_1,\ldots,X_n)}{\partial\phi_i\partial\phi_j}\right]_{1\le i,j\le p} \\
&= -\mathbb{E}_{\mathcal{P}_\phi}\left[\sum_{k=1}^{n}\frac{\partial^2 l(\phi|X_k)}{\partial\phi_i\partial\phi_j}\right]_{1\le i,j\le p} \\
&= -n\mathbb{E}_{\mathcal{P}_\phi}\left[\frac{\partial^2 l(\phi|X)}{\partial\phi_i\partial\phi_j}\right]_{1\le i,j\le p} \\
&= n\mathcal{I}(\phi).
\end{aligned}$$

On est maintenant proches de pouvoir définir l'efficacité d'un estimateur. Soit $\succeq$ la relation d'ordre partiel sur les matrices symétriques de taille $p \times p$,

$$A \succeq B \Longleftrightarrow A - B \text{ est semi-définie positive.}$$

Le résultat suivant donne une borne inférieure pour la variance d'un estimateur sans biais.

Théorème 2.3.15 *(Inégalité de Fréchet-Cramér-Rao, cas particulier)[62, 96, 127] Soit $X_1, \ldots, X_n$ un échantillon de population de même loi que X. Supposons que Φ est un ouvert de $\mathbb{R}^p$, et $\widehat{\phi}_n$ est un estimateur sans biais de $\phi \in \Phi$ (construit à partir de l'échantillon $X_1, \ldots, X_n$) qui admet une densité par rapport à la mesure de Lebesgue. Supposons également que la matrice d'information de Fisher $\mathcal{I}_n(\phi)$ est inversible. Alors, sous les hypothèses de la proposition 2.3.12, tout estimateur sans biais $\widehat{\phi}_n$ de ϕ vérifie*

$$\mathrm{var}_{\mathcal{P}_\phi}(\widehat{\phi}_n) \succeq \mathcal{I}_n(\phi)^{-1}. \tag{2.40}$$

Définition 2.3.16 *[62, 127] Si la borne de Fréchet-Cramér-Rao est atteinte (l'inégalité (2.40) est une égalité), l'estimateur $\widehat{\phi}_n$ est dit* efficace.

Un estimateur sans biais efficace est donc un estimateur de variance minimale. Voyons maintenant un cas de consistance et d'efficacité de l'estimateur du maximum de vraisemblance.

Proposition 2.3.17 *[96] Soient $X_1, \ldots, X_n$ des v.a.i.i.d. de même loi que X. Reprenons les hypothèses de la proposition 2.3.14. Supposons de plus que, pour tout $\phi \in \Phi$, la matrice d'information $\mathcal{I}(\phi)$ est définie positive, que la densité $f_\phi(x)$ admet toutes les dérivées partielles d'ordre 3 par rapport à ϕ, et qu'il existe des fonction M_{jkl} telles que*

$$\left|\frac{\partial^3 l(\phi|x)}{\partial\phi_j\partial\phi_k\partial\phi_l}\right| \le M_{jkl}(x),$$

avec $\mathbb{E}_{\mathcal{P}_{\phi_0}} M_{jkl}(X) < \infty \quad \forall j,k,l$, où $\phi_0 \in \Phi$ est la vraie valeur du paramètre. Alors, avec probabilité qui tend vers 1 quand $n \to \infty$, il existe une solution $\widehat{\phi}_n = \widehat{\phi}_n(X_1, \ldots, X_n)$ des équations du score telle que

$$\widehat{\phi}_n \xrightarrow{\text{Prob}} \phi \text{ et } \sqrt{n}(\widehat{\phi}_n - \phi) \xrightarrow{\mathcal{L}} \mathcal{N}_p\left(0, \mathcal{I}(\phi)^{-1}\right).$$

Si les équations du score ont une solution unique, l'estimateur du MV correspondant est donc asymptotiquement consistant et asymptotiquement efficace sous les hypothèses de la proposition 2.3.17.

Il s'agit maintenant de transposer la théorie de l'information de Fisher au cas des processus aléatoires [161]. Ici, $\{\mathcal{P}_\phi, \phi \in \Phi\}$ est une famille de mesures définissant chacune un processus aléatoire Y, et $X_1, X_2, \ldots$ correspond à une suite de vecteurs aléatoires d'observations de Y (par exemple, $X_i = (Y(x_1), \ldots, Y(x_i))$ quand les observations s'effectuent une par une séquentiellement). Notons $\mathcal{I}_n(\phi)$ l'information obtenue à partir du vecteur d'observations X_n. Sous certaines hypothèses de régularité, l'estimateur du MV $\widehat{\phi}_n$ vérifie

$$\mathcal{I}_n(\phi)^{\frac{1}{2}}(\widehat{\phi}_n - \phi) \xrightarrow{\mathcal{L}} \mathcal{N}_p(0, I_n). \tag{2.41}$$

Nous verrons au paragraphe suivant quel genre d'hypothèses il faut faire dans le cas d'un processus gaussien pour avoir des résultats de convergence similaires à ceux de la proposition 2.3.17.

2.3.2 Propriétés asymptotiques de l'estimateur du maximum de vraisemblance

Nous nous intéressons ici aux propriétés asymptotiques du maximum de vraisemblance. Nous commençons par rappeler les deux manières classiques de collecter les observations, puis citons des résultats sous le cadre que nous rencontrerons en pratique (un domaine d'étude borné). Nous indiquons finalement une méthode d'estimation des paramètres basée sur une pénalisation de la vraisemblance qui permet de réduire la variance des estimateurs obtenus.

2.3.2.1 Asymptotique par expansion, asymptotique par remplissage

Reprenant l'équation (2.30), et notant $\varsigma = (\sigma^2, \psi)$ le vecteur des paramètres de la fonction de covariance, on peut montrer que la matrice d'information $\mathbb{E}_{\mathcal{P}_{(\beta,\varsigma)}}[-l''(\beta, \varsigma|Y^n)]$ (voir le paragraphe précédent) s'écrit [6, 113]

$$\mathcal{I}(\beta, \varsigma) = \begin{pmatrix} {}^tMK^{-1}M & 0 \\ 0 & A \end{pmatrix} \equiv \mathrm{diag}(\mathcal{I}_\beta, \mathcal{I}_\varsigma), \tag{2.42}$$

avec $A = (a_{ij})_{1 \leq i,j \leq p+1}$ la matrice symétrique définie par

$$\begin{aligned} a_{ij} &= \frac{1}{2}\mathrm{tr}\left(R^{-1}R_i'R^{-1}R_j'\right), \qquad i,j < p+1\,; \\ a_{p+1,i} &= \frac{1}{2\sigma^2}\mathrm{tr}\left(R^{-1}R_i'\right), \qquad i < p+1\,; \\ a_{p+1,p+1} &= \frac{n}{2\sigma^4}, \end{aligned}$$

$R_i' = \partial R/\partial \varsigma_i$ désignant la i^e dérivée partielle de la matrice de corrélation. Sous certaines conditions, on peut montrer que, asymptotiquement, [113, 114]

$$\begin{pmatrix} \widehat{\beta} \\ \widehat{\varsigma} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \beta \\ \varsigma \end{pmatrix}, \begin{pmatrix} {\mathcal{I}_\beta}^{-1} & 0 \\ 0 & \mathcal{I}_\varsigma^{-1} \end{pmatrix}\right), \tag{2.43}$$

donc $\widehat{\beta}$ et $\widehat{\varsigma}$ sont asymptotiquement indépendants. On peut montrer que les estimateurs sont consistants, et aussi que

$$\mathbb{E}\left(\widehat{\beta}-\beta\right)=o\left(\frac{1}{n}\right) \quad \text{et} \quad \mathbb{E}\left(\widehat{\varsigma}-\varsigma\right)=\mathcal{I}_{\varsigma}^{-1}\delta+o\left(\frac{1}{n}\right), \tag{2.44}$$

avec

$$\begin{aligned}
\delta_i &= \frac{1}{2}\mathrm{tr}\left(\mathcal{I}_{\beta}^{-1}\frac{\partial\mathcal{I}_{\beta}}{\partial\varsigma_i}\right)+\frac{1}{2}\mathrm{tr}\left(\mathcal{I}_{\varsigma}^{-1}Q_i\right) ; \\
\left(Q_i\right)_{jk} &= \frac{1}{2}\mathrm{tr}\left(K''_{ij}K^{-1}K'_kK^{-1}-K''_{ik}K^{-1}K'_jK^{-1}-K''_{jk}K^{-1}K'_iK^{-1}\right) ; \\
K''_{ij} &= \frac{\partial^2 K}{\partial\varsigma_i\partial\varsigma_j}, \quad 1\leq i,j,k\leq p+1.
\end{aligned}$$

On retrouve en (2.44) le biais qui avait conduit à définir le maximum de vraisemblance restreint en 2.2.3, et on remarque la valeur typique $1/n$. Des résultats asymptotiques, tels que ceux donnés ci-dessus, sous-entendent que le nombre d'observations augmente d'une certaine façon : pour obtenir (2.43) et (2.44), il faut faire des hypothèses de type *asymptotique par expansion (increasing-domain asymptotics)*, c'est-à-dire augmenter le nombre d'observations tout en agrandissant le domaine d'étude, de façon à aller observer « loin » et obtenir beaucoup d'information sous la forme d'observations *quasi* indépendantes. Des résultats analogues dans le cas du MVR sont donnés dans [33].

Remarque 2.3.18 *[199] Les termes diagonaux de l'inverse de la matrice d'information indiquent la consistance ou non des estimateurs correspondants : on peut souvent remarquer une différence d'échelle entre les termes diagonaux plus petits correspondant aux estimateurs consistants, et les termes plus grands correspondant aux estimateurs non consistants.*

Cependant, nous nous intéressons plutôt au cas où, pour un domaine borné $\mathcal{X}$ donné, le nombre d'observations augmente à l'intérieur de la région d'étude $\mathcal{X}$, ce que l'on appelle *asymptotique par remplissage (fixed-domain asymptotics, infill asymptotics)* [31]. Il s'agit du cadre sous lequel on travaille en pratique, donc les résultats obtenus asymptotiquement donneront une idée de ce que l'on peut au mieux obtenir d'un échantillon de taille finie. Cependant, les résultats sont bien moins forts et généraux que si l'on fait de l'asymptotique par expansion : les conditions générales de consistance des estimateurs données ci-dessus ne sont pas valables sous des hypothèses de type asymptotique par remplissage. Le comportement asymptotique des estimateurs est très différent selon le cadre sous lequel on se place, sauf dans des cas très particuliers (par exemple, pour un processus gaussien stationnaire de moyenne nulle dont la corrélation est connue, la distribution limite de $\widehat{\sigma}^2$ est la même dans les deux cas [197]). Dans le cas où l'on fait de l'asymptotique par remplissage, l'estimateur du MV des paramètres qui ne sont pas micro-ergodiques ne sera généralement pas consistant [196]. On pourra consulter [93] pour un théorème donnant une condition suffisante de non-consistance d'un estimateur sous certaines hypothèses de type asymptotique par remplissage.

2.3.2.2 Résultats dans le cas de l'asymptotique par remplissage

Il existe cependant des résultats intéressants d'identifiabilité et de consistance sous certaines hypothèses de remplissage, lorsque le processus gaussien est muni de l'une des fonctions de covariance stationnaire classique présentées au paragraphe 2.1.2.3. La proposition suivante fait une synthèse de résultats trouvés dans la littérature.

Proposition 2.3.19 *Soit Z un processus gaussien stationnaire de moyenne nulle et fonction de covariance $k(\cdot)$, défini sur un domaine borné $\mathcal{X} \subset \mathbb{R}^d$, dont on estime les paramètres par maximum de vraisemblance. Alors, sous certaines hypothèses de remplissage :*

- *si $k(\tau) = \sigma^2 \prod_{i=1}^d e^{-\theta_i|\tau_i|}$ (produit de fonctions de covariance de type Ornstein-Uhlenbeck), les résultats dépendent de la dimension du domaine d'étude. Si $d = 1$, σ^2 et θ ne sont pas identifiables séparément, mais le produit $\sigma^2\theta$ est identifiable, l'estimateur du MV est fortement consistant et asymptotiquement normal [194]. Si $d > 1$, les paramètres σ^2 et θ_i sont tous identifiables, les estimateurs du MV sont fortement consistants, on a un résultat de normalité asymptotique et les paramètres θ_i sont asymptotiquement indépendants [195]. Si $d = 2$, l'estimateur du MV est asymptotiquement efficace [174].*
- *si $k(\tau) = \sigma^2 \prod_{i=1}^d e^{-\theta_i\tau_i^2}$ (covariance gaussienne), les paramètres σ^2 et θ_i sont tous identifiables. L'estimateur du MV des θ_i est fortement consistant, et on a l'expression de la vitesse de convergence [103]. Des résultats empiriques suggèrent la consistance de l'estimateur du MV de σ^2 [6]. On ne connaît pas de résultat de normalité asymptotique.*
- *si $k(\|\tau\|) = \sigma^2(\theta\|\tau\|)^\nu/(\Gamma(\nu)2^{\nu-1})K_\nu(\theta\|\tau\|)$ (covariance de Matérn isotrope), $\nu > 0$ connu, $d \in \{1, 2, 3\}$, les paramètres σ^2 et θ ne sont pas identifiables séparément, mais le produit $\sigma^2\theta^{2\nu}$ est identifiable et l'estimateur du MV est consistant [196].*
  Si $k(\tau) = \sigma^2 \prod_{i=1}^d 1/(\Gamma(\nu)2^{\nu-1})(\theta_i|\tau_i|)^\nu K_\nu(\theta_i|\tau_i|)$ (covariance de Matérn non isotrope) avec $\nu = \frac{3}{2}$ et $d \geq 3$, les paramètres σ^2 et θ_i sont tous identifiables et les estimateurs du MV sont consistants [102].

Remarque 2.3.20 *[83, 161] Pour que les estimateurs du MV ou du MVR soient asymptotiquement efficaces, il faut que la vraisemblance soit continue et deux fois dérivable sur l'ensemble du domaine d'existence des paramètres. Cette condition n'est pas verifiée pour certaines fonctions de covariance, comme les covariances sphériques dans $\mathbb{R}^3$*

$$K(\|\tau\|) = \begin{cases} \sigma^2\left(1 - \frac{3}{2\theta}\|\tau\| + \frac{1}{2\theta^3}\|\tau\|^3\right) & si\ \|\tau\| \leq \theta \\ 0 & si\ \|\tau\| > \theta \end{cases},$$

où $\theta > 0$ est le paramètre de portée.

Il est plus difficile d'obtenir des résultats quand la moyenne du processus est non nulle. On consultera cependant [194] dans le cas du processus d'Ornstein-Uhlenbeck.
Des résultats empiriques étendant la proposition 2.3.19 sont donnés dans [6], qui utilisent la matrice d'information $\mathcal{I}_\varsigma$. Il est montré tout d'abord que, pour un processus d'Ornstein-Uhlenbeck, la distribution asymptotique du paramètre $\widehat{\sigma^2\theta} = \widehat{\sigma^2}\widehat{\theta}$ (qui est donnée dans [195]) peut s'obtenir à partir de l'inverse de la matrice d'information. L'idée est alors d'utiliser la matrice d'information pour en déduire, de façon analogue, des résultats empiriques de consistance dans le cas d'une fonction de covariance gaussienne, pour laquelle il n'y a pas de résultat théorique concernant le paramètre σ^2. Après avoir verifié expérimentalement dans le cas de 1, puis 2 facteurs, que $\widehat{\sigma^2}$ et $\widehat{\theta}$ semblent non biaisés, et que l'inverse de la matrice d'information fournit une bonne approximation de la variance des paramètres, il est montré que $\mathrm{tr}(\mathcal{I}_\varsigma^{-1}) \to 0$ quand $n \to \infty$, ce qui semble indiquer, empiriquement, que l'estimation des paramètres est consistante dans le cas d'une fonction de covariance gaussienne (la théorie donne le résultat seulement pour $\widehat{\theta}$, proposition 2.3.19).
Nous verrons au paragraphe 2.3.3 que le fait de ne pas pouvoir identifier séparément chacun des paramètres de la fonction de covariance n'est pas un obstacle pour faire de la prédiction optimale : il suffit de bien estimer les quantités identifiables pour obtenir un bon prédicteur.

C'est une illustration de la *loi de Jeffrey* [161] : *les choses que l'on ne peut pas prévoir à partir d'un grand nombre de données ne peuvent pas avoir d'influence sur la prédiction.* À titre d'illustration, on pourra consulter [196], où les figures montrent clairement que les prédictions sont presque identiques pour des mêmes valeurs du produit $\sigma^2\theta$ dans le cas d'un processus d'Ornstein-Uhlenbeck, même si σ^2 et θ sont différents.

Remarque 2.3.21 *[194] Dans le cas où les paramètres ne sont pas tous estimables, il est possible d'en fixer un certain nombre afin de réduire le nombre d'éléments à optimiser. Par exemple, dans le cas d'un processus d'Ornstein-Uhlenbeck de moyenne nulle, il est équivalent de chercher*

$$\begin{aligned} \widehat{\sigma^2\theta} &= \underset{\sigma^2\theta}{\operatorname{argmax}}\, l(\sigma^2,\theta), \\ \widehat{\sigma^2} &= \underset{\sigma^2}{\operatorname{argmax}}\, l(\sigma^2,\theta_0), \qquad \textit{ou} \\ \widehat{\theta} &= \underset{\theta}{\operatorname{argmax}}\, l(\sigma_0^2,\theta), \end{aligned}$$

avec σ_0^2 et θ_0 fixés a priori*, puisqu'au final les produits $\widehat{\sigma^2\theta}, \widehat{\sigma^2}\theta_0, \sigma_0^2\widehat{\theta}$ obtenus seront identiques.*

Intéressons-nous alors à la distribution asymptotique des estimateurs quand une partie des valeurs des paramètres est fixée *a priori*. Si le vecteur des paramètres ϕ d'un processus gaussien n'est pas micro-ergodique, notons $\phi = {}^t({}^t\phi_\mu, {}^t\phi_\nu)$, où ϕ_μ est micro-ergodique et aucune fonction non triviale de ϕ_ν n'est micro-ergodique. Il est raisonnable de conjecturer que si ϕ_ν est fixé *a priori* et ϕ_μ est estimé par une suite consistante $\widehat{\phi_{\mu,n}}$ d'estimateurs du MV, alors le comportement asymptotique des estimateurs $\widehat{\phi_{\mu,n}}$ est le même que si ϕ_ν était connu [161].

2.3.2.3 Maximum de vraisemblance pénalisé

Observant l'équation (2.43), on remarque que lorsque la matrice d'information est mal conditionnée, la variance des estimateurs peut être très élevée. Ceci correspond à une fonction de vraisemblance (ou de vraisemblance restreint) plate autour de l'optimum. Une solution possible pour diminuer la variance des estimateurs est d'ajouter un terme de pénalisation à la log-vraisemblance (2.30),

$$l(\beta,\sigma^2,\psi) = -\frac{1}{2}\left[n\log\sigma^2 + \log(\det(R)) + \frac{{}^t(Y^n - F\beta)R^{-1}(Y^n - F\beta)}{\sigma^2}\right] - n\sum_i p_\lambda(\varsigma_i),$$

avec $\varsigma = (\sigma^2,\psi)$, $p_\lambda(\cdot)$ une fonction de pénalisation à valeurs positives et λ un paramètre contrôlant la régularisation (cette méthode s'applique évidemment aussi au maximum de vraisemblance restreint). Le prix à payer pour la diminution de la variance est un biais supplémentaire sur les paramètres (c'est donc l'idée inverse du MVR). Plusieurs pénalités sont proposées dans [98], ainsi qu'un algorithme pour estimer λ, β, σ^2 et ψ. Sous certaines hypothèses sur la pénalité, on peut retrouver le résultat de normalité asymptotique du maximum de vraisemblance (2.41), montrant l'efficacité asymptotique de l'estimateur ainsi construit. Mais il est suggeré dans [161] que si la vraisemblance est plate, c'est simplement que l'on ne peut rien tirer des données (on retrouve la loi de Jeffrey).

2.3.3 Conséquences d'une mauvaise estimation des paramètres

Les performances du prédicteur lorsque les paramètres du modèle sont incorrects sont rappelées dans un premier paragraphe, de façon heuristique, afin de pouvoir dégager des idées

générales sans trop alourdir la présentation. Ensuite, nous voyons des méthodes alternatives à la technique « plug-in » qui ont été proposées pour l'estimation de l'EQM dans la littérature.

2.3.3.1 Performances du prédicteur

Rappelons ici des idées générales concernant la qualité du prédicteur lorsque la structure du second ordre (m_0, k_0), caractérisant le processus gaussien $\{Y(x,\omega), x \in \mathcal{X}, w \in \Omega\}$, est mal estimée [161]. On constatera qu'il est possible d'obtenir asymptotiquement un prédicteur optimal même si le modèle (m_1, k_1) utilisé est différent de la réalité.
Dans toute la suite, on notera (m_0, k_0) la vraie structure du second ordre de Y et $\mathcal{P}_0$ la mesure gaussienne associée, et (m_1, k_1) la structure avec laquelle on effectue les prédictions et $\mathcal{P}_1$ la mesure gaussienne associée. Les BLPs en un point x obtenus à partir des n observations en $X^n = \{x_1, \ldots, x_n\}$, en utilisant les modèles (m_0, k_0) et (m_1, k_1), seront notés respectivement $\widehat{Y}_0^n(x)$ et $\widehat{Y}_1^n(x)$. Une façon de mesurer la qualité du prédicteur obtenu à partir des termes incorrects (m_1, k_1) est donnée par

$$\frac{\mathbb{E}_{\mathcal{P}_0}\left[\widehat{Y}_1^n(x) - Y(x)\right]^2}{\mathbb{E}_{\mathcal{P}_0}\left[\widehat{Y}_0^n(x) - Y(x)\right]^2} = 1 + \frac{\mathbb{E}_{\mathcal{P}_0}\left[\widehat{Y}_1^n(x) - \widehat{Y}_0^n(x)\right]^2}{\mathbb{E}_{\mathcal{P}_0}\left[\widehat{Y}_0^n(x) - Y(x)\right]^2}, \tag{2.45}$$

qui mesure le *ratio* entre l'EQM du « faux » BLP et du « vrai » BLP (l'égalité est obtenue en utilisant la caractérisation du BLP par orthogonalité donnée dans la remarque 2.2.15). Si l'on souhaite également évaluer de combien on se trompe en utilisant l'EQM $\mathbb{E}_{\mathcal{P}_1}[\widehat{Y}_1^n(x) - Y(x)]^2$ donnée par le mauvais modèle (m_1, k_1), on calcule la quantité

$$\frac{\mathbb{E}_{\mathcal{P}_1}\left[\widehat{Y}_1^n(x) - Y(x)\right]^2}{\mathbb{E}_{\mathcal{P}_0}\left[\widehat{Y}_1^n(x) - Y(x)\right]^2}, \tag{2.46}$$

où l'on a divisé par la vraie EQM du BLP obtenu sous le modèle (m_1, k_1). Enonçons quelques principes généraux :

- utiliser une mauvaise fonction de covariance a un plus grand impact sur l'EQM, à travers (2.46), que sur le prédicteur, à travers (2.45) ;
- le comportement haute fréquence de la mesure spectrale (ou, de façon quelque peu imprécise, le comportement de la fonction de covariance à l'origine) est crucial quand on fait de l'interpolation (de la prédiction dans l'enveloppe convexe des observations) ; le comportement basse fréquence de la mesure spectrale (en particulier, la portée de la covariance) a asymptotiquement peu d'effets sur l'interpolation, surtout pour les processus les plus réguliers (mais pour un échantillon de taille finie, si la portée de la covariance est très petite, la prédiction est égale à la partie déterministe dès qu'on s'écarte un peu des points observés, avec des pics très étroits autour des points observés : on a alors une très mauvaise prédiction). Quand on fait de l'extrapolation (de la prédiction en dehors de l'enveloppe convexe des observations) le comportement basse fréquence de la mesure spectrale joue un rôle plus important que dans le cas de l'interpolation.

À partir de (2.45) et (2.46), il est montré dans [161] qu'utiliser la fonction de covariance gaussienne peut donner un prédicteur acceptable même si le vrai processus n'est pas très régulier ; par contre l'EQM obtenu en utilisant la covariance gaussienne a tendance a être beaucoup trop optimiste. C'est l'une des raisons (en plus des instabilités numériques constatées) pour lesquelles [161] déconseille l'utilisation de la covariance gaussienne.

On peut montrer que, sous certaines hypothèses, des mesures gaussiennes équivalentes génèrent des prédicteurs asymptotiquement équivalents.

Théorème 2.3.22 *[161] Soit $\{Y(x,\omega), x \in \mathcal{X}, \omega \in \Omega\}$ un processus gaussien défini par la structure du second ordre (m_0, k_0), avec $\mathcal{X}$ un sous-ensemble borné de $\mathbb{R}^d$. Soit $\{x_i\}_{i\in\mathbb{N}}$ une suite de points de $\mathcal{X}$. On note $X^n = \{x_1, \ldots, x_n\}$ l'ensemble des n premières observations. Si les mesures gaussiennes $\mathcal{P}_0$ et $\mathcal{P}_1$, associées respectivement aux structures du second ordre (m_0, k_0) et (m_1, k_1), sont équivalentes, alors, sous certaines hypothèses générales,*

$$\lim_{n\to\infty} \sup_{x\in\mathcal{X}\setminus X^n} \left| \frac{\mathbb{E}_{\mathcal{P}_0}\left[\widehat{Y}_1^n(x) - Y(x)\right]^2}{\mathbb{E}_{\mathcal{P}_0}\left[\widehat{Y}_0^n(x) - Y(x)\right]^2} - 1 \right| = 0\,;$$

$$\lim_{n\to\infty} \sup_{x\in\mathcal{X}\setminus X^n} \left| \frac{\mathbb{E}_{\mathcal{P}_1}\left[\widehat{Y}_1^n(x) - Y(x)\right]^2}{\mathbb{E}_{\mathcal{P}_0}\left[\widehat{Y}_1^n(x) - Y(x)\right]^2} - 1 \right| = 0.$$

Autrement dit, le BLP obtenu sous le modèle faux (m_1, k_1), ainsi que son EQM, sont asymptotiquement uniformément optimaux (le supremum est calculé pour les x n'appartenant pas à X^n afin de ne pas diviser par 0).

On retrouve la loi de Jeffrey : sous des conditions générales, deux mesures gaussiennes qui ne peuvent pas être distinguées avec grande probabilité donnent des prédicteurs identiques.

De façon intuitive (et pas entièrement rigoureuse), on retiendra que le prédicteur sera asymptotiquement optimal si la régularité de la fonction de covariance $k_1(\cdot)$ à l'origine est la même que la régularité à l'origine de la fonction $k_0(\cdot)$.

Remarque 2.3.23 *[161] En toute rigueur, ce sont les densités spectrales $f_1(\cdot)$ et $f_0(\cdot)$ (§ 2.1.2.3) qui doivent avoir le même comportement haute fréquence (*i.e.*, à l'infini).*

Nous renvoyons à [162] pour des vitesses de convergence du prédicteur obtenu sous un mauvais modèle, et à [164] pour des résultats d'équivalence dans le cas non-stationnaire. Des expériences numériques illustrant les résultats évoqués dans ce paragraphe sont présentées dans [161, 178].

2.3.3.2 Alternatives pour l'estimation de l'EQM

Dans le cas où les paramètres de la fonction de covariance sont inconnus et estimés, la formule (2.24) de l'EQM n'est plus valide. On peut montrer que la formule « plug-in » (2.29) sous-estime en moyenne l'EQM, car elle ne tient pas compte de la variabilité des estimateurs : on a donc tendance à être trop optimiste sur la qualité du modèle.

Dans [5], un estimateur correctif empirique de l'EQM est proposé, dans le but d'améliorer la qualité d'un autre estimateur correctif proposé dans [69, 82, 201] et utilisé dans [4] dans le cas d'une fonction de corrélation exponentielle. Il est montré par simulation que ce nouvel estimateur peut être plus proche de la réalité dans le cas d'expériences déterministes, mais que les résultats sont comparables en présence de bruit de mesure (§ 2.4). Il est aussi montré que la qualité de ces estimateurs correctifs dépend de la vraie corrélation : l'estimation « plug-in » de l'EQM peut notamment s'avérer meilleure en cas de forte corrélation. C'est pourquoi, en l'absence d'une alternative fiable, il est recommandé dans [201] d'utiliser l'estimateur « plug-in » (2.29).

Dans [40], une méthode d'estimation de l'EQM utilisant le bootstrap [47] est proposée, qui tient compte de la variabilité supplémentaire introduite par les estimateurs, mais cette méthode est coûteuse en temps de calcul.

C'est donc tout de même la formule « plug-in » de l'EQM (2.29) que nous utiliserons en pratique.

2.4 Krigeage avec prise en compte d'erreur de mesure

Dans le cas où l'on souhaite modéliser un phénomène physique, les observations sont entachées d'un bruit de mesure dont il faut tenir compte pour éviter le phénomène de *sur-ajustement (overfitting)*, quand le modèle a des variations trop fortes par rapport à la réalité. De plus, construire un interpolateur basé sur des observations bruitées n'aurait pas de sens, car répéter les observations en un même point donne des valeurs différentes. On va donc faire de la régression des données, et non plus de l'interpolation comme précédemment.
Nous présentons tout d'abord le modèle de krigeage correspondant, puis une méthode permettant de modéliser un système informatique (déterministe) en présence d'un bruit de type « numérique ».

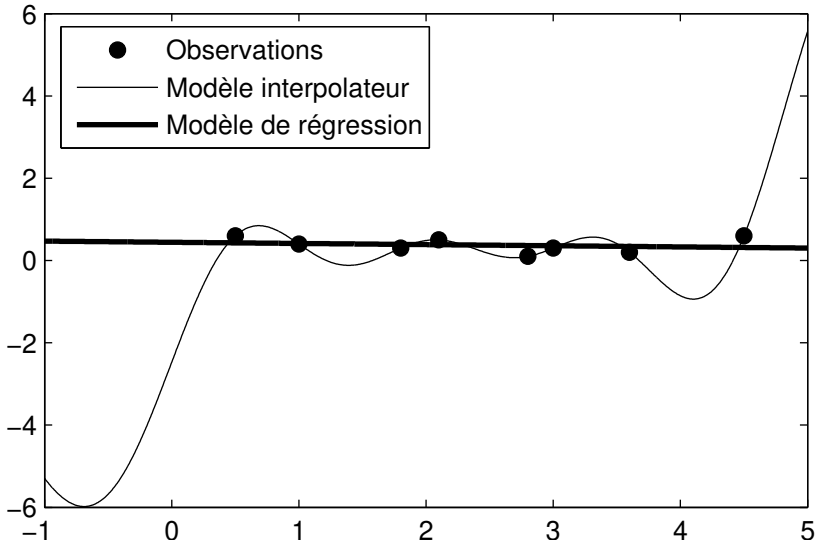

FIGURE 2.5 – Illustration du phénomène de sur-ajustement (overfitting) : les prédictions données par le modèle interpolateur (en trait fin) risquent d'être très éloignées de la réalité quand on s'éloigne des données, alors qu'un modèle de régression (en trait épais) tient compte des erreurs de mesure et présente des variations moindres.

2.4.1 Modèle de krigeage avec inclusion de bruit

Le modèle de krigeage (2.18) est modifié en y ajoutant un terme modélisant le bruit de mesure,

$$Y(x) = {}^t m(x)\beta + Z(x) + \varepsilon(x), \tag{2.47}$$

où $\varepsilon(x) \sim \mathcal{N}(0, \sigma_\varepsilon^2(x))$ est indépendante de $Z(x)$ pour tout x, et de $\varepsilon(x')$ pour tout $x' \neq x$. On supposera la variance de bruit constante et inconnue, *i.e.*

$$\sigma_\varepsilon^2(x) = \sigma_\varepsilon^2.$$

La log-vraisemblance s'écrit alors à une constante près [51, 136]

$$l(\beta,\sigma^2,\psi,\sigma_\epsilon^2|Y^n) = -\frac{1}{2}\left[\log\left(\det\left(\sigma^2 R+\sigma_\varepsilon^2 I_n\right)\right) + {}^t\left(Y^n - M\beta\right)\left(\sigma^2 R+\sigma_\epsilon^2 I_n\right)^{-1}\left(Y^n - M\beta\right)\right]. \tag{2.48}$$

Le paramètre σ_ε^2 peut être vu comme un terme de régularisation, la matrice $\sigma^2 R + \sigma_\varepsilon^2 I_n$ étant mieux conditionnée que la matrice $\sigma^2 R$ apparaissant dans l'équation de la log-vraisemblance (2.30) (*i.e.*, l'inversion est numériquement plus stable). Cette technique est d'ailleurs utilisée dans la routine DACE [104, 105] pour améliorer le conditionnement des matrices de corrélation dans le cas non bruité, en leur ajoutant systématiquement la matrice δI_n, avec $\delta \approx 10^{-15}$.
On remarque que l'on ne peut pas séparer les termes σ^2 et σ_ε^2 dans l'équation (2.48). Après une reparamétrisation en $\lambda = \sigma_\varepsilon^2/\sigma^2$ [113], la log-vraisemblance se réécrit de façon similaire à (2.30)

$$l(\beta,\sigma^2,\psi,\lambda|Y^n) = -\frac{1}{2}\left[n\log\left(\sigma^2\right) + \log\left(\det\left(R+\lambda I_n\right)\right) + \frac{{}^t\left(Y^n - M\beta\right)\left(R+\lambda I_n\right)^{-1}\left(Y^n - M\beta\right)}{\sigma^2}\right]. \tag{2.49}$$

En cherchant les zéros des dérivées partielles par rapport à β et σ^2, on obtient

$$\widehat{\beta} = \widehat{\beta}(\psi,\lambda) = \left({}^t M(R+\lambda I_n)^{-1}M\right)^{-1}{}^t M(R+\lambda I_n)^{-1}Y^n, \tag{2.50}$$

et

$$\widehat{\sigma}^2 = \widehat{\sigma}^2(\psi,\lambda) = \frac{1}{n}{}^t(Y^n - M\widehat{\beta})(R+\lambda I_n)^{-1}(Y^n - M\widehat{\beta}). \tag{2.51}$$

Substituant $\widehat{\beta}(\psi,\lambda)$ et $\widehat{\sigma}^2(\psi,\lambda)$ dans (2.49), on obtient

$$l(\widehat{\beta},\widehat{\sigma}^2,\psi,\lambda|Y^n) = -\frac{1}{2}\left[n\log\widehat{\sigma}^2(\psi,\lambda) + \log\left(\det(R(\psi)+\lambda I_n)\right) + n\right],$$

qui ne dépend que de ψ et λ. L'estimation par maximum de vraisemblance s'écrit de façon compacte

$$(\widehat{\psi},\widehat{\lambda}) = \underset{(\psi,\lambda)}{\operatorname{argmin}}\left[n\log\widehat{\sigma}^2\left(\psi,\lambda\right) + \log\left(\det\left(R(\psi)+\lambda I_n\right)\right)\right], \tag{2.52}$$

avec $\widehat{\sigma}^2(\psi,\lambda)$ défini en (2.51). Connaissant $\widehat{\psi}$ et $\widehat{\lambda}$, on peut finalement calculer aussi $\widehat{\beta}$, $\widehat{\sigma}^2$ et $\widehat{\sigma_\varepsilon}^2$. L'EBLUP du maximum de vraisemblance au point x_0 s'écrit comme en (2.28)

$$\widehat{Y}_0 = {}^t m_0\widehat{\beta} + {}^t\widehat{r}_0(\widehat{R}+\widehat{\lambda}I_n)^{-1}(Y^n - M\widehat{\beta}), \tag{2.53}$$

et l'EQM empirique correspondant vaut

$$\mathrm{EQM}_{\widehat{\mu}}(x_0) = \widehat{\sigma}^2\left[1 - {}^t\widehat{r}_0(\widehat{R}+\widehat{\lambda}I_n)^{-1}\widehat{r}_0 + {}^t\widehat{\gamma}\left({}^t M\left(\widehat{R}+\widehat{\lambda}I_n\right)^{-1}M\right)^{-1}\widehat{\gamma}\right], \tag{2.54}$$

avec $\widehat{\gamma} = m_0 - {}^t M(\widehat{R}+\widehat{\lambda}I_n)^{-1}\widehat{r}_0$. Si on a besoin d'une estimation de la loi *a posteriori* de Y_0, on utilise généralement en pratique la loi $\mathcal{N}(\widehat{Y}_0, \mathrm{EQM}_{\widehat{\mu}}(x_0))$ (revoir cependant la remarque 2.2.13). Le modèle de krigeage ainsi construit n'est plus un interpolateur. De plus, l'EQM empirique ne vaut plus 0 aux x_i, illustrant le fait qu'observer une deuxième fois en x_i donnera une valeur différente.

Remarque 2.4.1 *En faisant $\lambda = 0$ dans (2.53) et (2.54), on retrouve les équations de l'EBLUP (2.28) et de l'EQM empirique (2.29) obtenues dans le cas non bruité.*

En ce qui concerne l'identifiabilité des paramètres (§ 2.3.1), tout dépend de la variance du bruit σ_ε^2.

Proposition 2.4.2 *[161] Soient $(m_0, k_0, \sigma_{\epsilon,0}^2)$ et $(m_1, k_1, \sigma_{\epsilon,1}^2)$ deux mesures gaussiennes pour un processus aléatoire Y défini sur un ensemble $\mathcal{X} \subset \mathbb{R}^d$, et $\{x_1, x_2, \ldots\}$ une suite de points de $\mathcal{X}$ dense dans $\mathcal{X}$. Si $\mathcal{X}$ n'a pas de point isolé, et si Y est continu en moyenne quadratique (annexe B), alors les mesures sont :*

- *orthogonales, si $\sigma_{\varepsilon,0}^2 \neq \sigma_{\varepsilon,1}^2$;*
- *équivalentes, si $\sigma_{\varepsilon,0}^2 = \sigma_{\varepsilon,1}^2$ et les mesures gaussiennes associées à (m_0, k_0) et (m_1, k_1) sont équivalentes.*

Il est donc possible d'identifier correctement le paramètre de bruit, et par suite tout dépend de l'identifiabilité des autres paramètres. En ce qui concerne la qualité du prédicteur sous un mauvais modèle, il existe un analogue du théorème 2.3.22 en présence de bruit.

Théorème 2.4.3 *[161] Soit $\{Y(x,\omega), x \in \mathcal{X}, \omega \in \Omega\}$ un processus gaussien, défini par la structure du second ordre $(m_0, k_0, \sigma_\varepsilon^2)$, avec $k_0(\cdot,\cdot)$ continue, et soit $\mathcal{X}$ un sous-ensemble borné de $\mathbb{R}^d$ n'ayant aucun point isolé. Soit $\{x_i\}_{i\in\mathbb{N}}$ une suite de points de $\mathcal{X}$ dense dans $\mathcal{X}$. On note $X^n = \{x_1, \ldots, x_n\}$ l'ensemble des n premières observations, $\widehat{Y}_0^n(x)$ et $\widehat{Y}_1^n(x)$ les BLPs obtenus respectivement sous les modèles $(m_0, k_0, \sigma_\varepsilon^2)$ et $(m_1, k_1, \sigma_\varepsilon^2)$. Si les mesures gaussiennes $\mathcal{P}_0$ et $\mathcal{P}_1$, associées respectivement aux structures du second ordre $(m_0, k_0, \sigma_\varepsilon^2)$ et $(m_1, k_1, \sigma_\varepsilon^2)$, sont équivalentes, alors,*

$$\lim_{n\to\infty} \sup_{x\in\mathcal{X}\setminus X^n} \left| \frac{\mathbb{E}_{\mathcal{P}_0}\left[\widehat{Y}_1^n(x) - Y(x)\right]^2}{\mathbb{E}_{\mathcal{P}_0}\left[\widehat{Y}_0^n(x) - Y(x)\right]^2} - 1 \right| = 0\,;$$

$$\lim_{n\to\infty} \sup_{x\in\mathcal{X}\setminus X^n} \left| \frac{\mathbb{E}_{\mathcal{P}_1}\left[\widehat{Y}_1^n(x) - Y(x)\right]^2}{\mathbb{E}_{\mathcal{P}_0}\left[\widehat{Y}_1^n(x) - Y(x)\right]^2} - 1 \right| = 0.$$

Autrement dit, le BLP obtenu sous les hypothèses fausses $(m_1, k_1, \sigma_\varepsilon^2)$, ainsi que son EQM, sont asymptotiquement uniformément optimaux.

Les résultats disponibles sur les propriétés asymptotiques des estimateurs du MV dans le cas bruité ne sont pas nombreux. La proposition suivante étend un résultat donné dans le cas non bruité.

Proposition 2.4.4 *[25] Soient $\widehat{\sigma^2}, \widehat{\theta}$ et $\widehat{\sigma_\varepsilon^2}$ les estimateurs du MV d'un processus d'Ornstein-Uhlenbeck (§ 2.1.2.3, équation (2.11)), et D un sous-ensemble compact de $\mathbb{R}_+^3$ contenant le vecteur des vraies valeurs des paramètres $(\sigma^2, \theta, \sigma_\varepsilon^2)$. Alors, sous certaines hypothèses infill, l'estimation de $\sigma^2\theta$ est consistante, l'estimation de σ_ε^2 est fortement consistante, et si $(\sigma^2, \theta, \sigma_\varepsilon^2)$ est dans l'intérieur de D, on a*

$$\begin{pmatrix} n^{\frac{1}{4}}(\widehat{\sigma^2\theta} - \sigma^2\theta) \\ n^{\frac{1}{2}}(\widehat{\sigma_\varepsilon^2} - \sigma_\varepsilon^2) \end{pmatrix} \xrightarrow{\mathcal{L}} \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 4\sqrt{2}\sigma_\varepsilon(\sigma^2\theta)^{\frac{3}{2}} & 0 \\ 0 & 2\sigma_\varepsilon^4 \end{pmatrix} \right).$$

La vitesse de convergence de l'estimateur $\widehat{\sigma^2\theta}$ est seulement en $n^{\frac{1}{4}}$, contre $\sqrt{n}$ dans le cas non bruité (voir [194] pour l'expression de la normalité asymptotique dans le cas non bruité).

Remarque 2.4.5 *Il est également possible d'introduire un bruit aléatoire sur les entrées, en supposant que les $x_i, 1 \leq i \leq n$, sont des variables aléatoires [32]. La propagation des incertitudes est étudiée dans [133], où l'on observe que les trajectoires prédites sont aplanies. Dans [50], il est montré, sous certaines hypothèses sur la forme de la moyenne $m(\cdot)$ et de la fonction de covariance $k(\cdot,\cdot)$, que si $k(\cdot,\cdot)$, σ_ε^2 et la distribution des x_i sont connus, alors, sous certaines hypothèses infill, l'EQM tend vers 0 en tout point quand le nombre d'observations tend vers l'infini ; en outre, pour un échantillon de taille finie, prendre en compte le fait que les x_i sont aléatoires diminue l'EQM du BLUP (2.23) obtenu en supposant les x_i fixés.*

L'équivalent du krigeage avec bruit de mesure, en terme de régression régularisée dans un RKHS, est donné dans le théorème 1.4.4, équation (1.21).

2.4.2 Réinterpolation

Certains phénomènes naturels difficiles à observer sont modélisés par un code informatique : le phénomène est alors « observé » par simulation numérique du modèle informatique (voir par exemple [148] pour l'utilisation d'un code informatique modélisant un réservoir pétrolier). Lorsque le phénomène est complexe, les simulations numériques peuvent être très coûteuses en temps de calcul. Il est alors souhaitable de construire un modèle du code informatique (un modèle du modèle, ou *meta-modèle*), en utilisant par exemple le krigeage.
La sortie d'un code informatique étant déterministe, on serait enclin à utiliser le krigeage sans bruit de mesure. Néanmoins, la réponse peut être entachée d'une erreur numérique non-négligeable, dûe par exemple à des problèmes de discrétisation (lorsque le maillage est variable) ou de convergence [51]. Des entrées très proches pourront donner des valeurs très différentes : utiliser un modèle interpolateur produirait alors un phénomène de sur-ajustement. Le modèle de krigeage avec bruit de mesure n'est pas non plus satisfaisant, car l'erreur quadratique est non-nulle aux facteurs d'entrée observés, alors que la réponse du code est déterministe : répéter les observations au même facteur d'entrée donnera toujours la même valeur de sortie. On souhaiterait construire un modèle qui ne soit pas un interpolateur, mais dont l'erreur quadratique soit nulle aux facteurs observés : c'est l'objet de la *réinterpolation*.
Le principe de la méthode est le suivant [51] :

1. construire le prédicteur de krigeage avec bruit de mesure à partir des données d'apprentissage $Y^n = {}^t(y(x_1), \ldots, y(x_n))$,
$$\widehat{Y}(x_0) = {}^t m_0 \widehat{\beta} + {}^t \widehat{r}_0 (\widehat{R} + \widehat{\lambda} I_n)^{-1} (Y^n - M\widehat{\beta}), \tag{2.55}$$
avec $\widehat{\beta}$ donné en (2.50) ;
2. calculer le vecteur des prédictions aux facteurs observés x_i, donné par le modèle de krigeage avec bruit de mesure,
$$\begin{aligned} Y_{\mathrm{r}}^n &= {}^t\Big(\widehat{Y}(x_1), \ldots, \widehat{Y}(x_n)\Big) \\ &= M\widehat{\beta} + \widehat{R}(\widehat{R} + \widehat{\lambda} I_n)^{-1}(Y^n - M\widehat{\beta})\,; \end{aligned} \tag{2.56}$$
3. calculer le prédicteur de krigeage sans bruit, à partir du vecteur des « observations » Y_{r}^n, en utilisant les valeurs des paramètres de corrélation obtenues à l'étape précédente,
$$\widehat{Y}_{\mathrm{r}}(x_0) = {}^t m_0 \widehat{\beta}_{\mathrm{r}} + {}^t \widehat{r}_0 \widehat{R}^{-1} (Y_{\mathrm{r}}^n - M\widehat{\beta}_{\mathrm{r}}),$$
avec
$$\widehat{\beta}_{\mathrm{r}} = \left({}^t M \widehat{R}^{-1} M\right)^{-1} {}^t M \widehat{R}^{-1} Y_{\mathrm{r}}^n. \tag{2.57}$$

Le prédicteur $\widehat{Y}_{\mathrm{r}}(x_0)$ satisfait à nos attentes : ses variations sont plus douces qu'un modèle qui soit un interpolateur pour les données observées avec bruit Y^n, et l'erreur de prédiction est nulle aux facteurs observés. Les valeurs des prédicteurs $\widehat{Y}_{\mathrm{r}}(x_0)$ et $\widehat{Y}(x_0)$ sont en fait identiques. En utilisant (2.56) et (2.57), on montre tout d'abord que $\widehat{\beta}_{\mathrm{r}} = \widehat{\beta}$:

$$\begin{aligned}
\widehat{\beta}_{\mathrm{r}} &= \left({}^t M \widehat{R}^{-1} M\right)^{-1} {}^t M \widehat{R}^{-1} Y_{\mathrm{r}}^n \\
&= \left({}^t M \widehat{R}^{-1} M\right)^{-1} \left({}^t M \widehat{R}^{-1} M \widehat{\beta} + {}^t M \widehat{R}^{-1} \widehat{R} (\widehat{R} + \widehat{\lambda} I_n)^{-1} (Y^n - M\widehat{\beta})\right) \\
&= \widehat{\beta} + \left({}^t M \widehat{R}^{-1} M\right)^{-1} \left({}^t M (\widehat{R} + \widehat{\lambda} I_n)^{-1} Y^n - {}^t M (\widehat{R} + \widehat{\lambda} I_n)^{-1} M \widehat{\beta}\right) \\
&= \widehat{\beta},
\end{aligned}$$

où la dernière égalité vient de l'expression de $\widehat{\beta}$ donnée en (2.50). On a donc, pour tout $x_0 \in \mathcal{X}$,

$$\begin{aligned}
\widehat{Y}_{\mathrm{r}}(x_0) &= {}^t m_0 \widehat{\beta}_{\mathrm{r}} + {}^t \widehat{r}_0 \widehat{R}^{-1} (Y_{\mathrm{r}}^n - M\widehat{\beta}_{\mathrm{r}}) \\
&= {}^t m_0 \widehat{\beta} + {}^t \widehat{r}_0 \widehat{R}^{-1} (Y_{\mathrm{r}}^n - M\widehat{\beta}) \\
&= {}^t m_0 \widehat{\beta} + {}^t \widehat{r}_0 \widehat{R}^{-1} \widehat{R} (\widehat{R} + \widehat{\lambda} I_n)^{-1} (Y^n - M\widehat{\beta}) \\
&= {}^t m_0 \widehat{\beta} + {}^t \widehat{r}_0 (\widehat{R} + \widehat{\lambda} I_n)^{-1} (Y^n - M\widehat{\beta}) \\
&= \widehat{Y}(x_0).
\end{aligned}$$

L'EQM empirique du prédicteur $\widehat{Y}_{\mathrm{r}}(x_0)$ est calculée à l'aide de l'équation (2.29), où seul le terme de variance $\widehat{\sigma_{\mathrm{r}}^2}$ dont l'équation est donnée par (2.32), dépend de Y_{r}^n. Sa valeur en fonction du vecteur des observations Y^n est donnée par

$$\begin{aligned}
\widehat{\sigma_{\mathrm{r}}^2} &= \frac{1}{n} {}^t (Y_{\mathrm{r}}^n - M\widehat{\beta}_{\mathrm{r}}) \widehat{R}^{-1} (Y_{\mathrm{r}}^n - M\widehat{\beta}_{\mathrm{r}}) \\
&= \frac{1}{n} {}^t (Y_{\mathrm{r}}^n - M\widehat{\beta}) \widehat{R}^{-1} (Y_{\mathrm{r}}^n - M\widehat{\beta}) \\
&= \frac{1}{n} {}^t (Y^n - M\widehat{\beta}) (\widehat{R} + \widehat{\lambda} I_n)^{-1} \widehat{R} (\widehat{R} + \widehat{\lambda} I_n)^{-1} (Y^n - M\widehat{\beta}), \qquad (2.58)
\end{aligned}$$

où la dernière égalité a été obtenue en utilisant (2.56).

En résumé, nous avons ainsi construit un prédicteur qui n'est pas un interpolateur, mais dont l'erreur de prédiction est nulle aux facteurs observés (voir la figure 2.6, où les cercles pleins représentent les observations ; le modèle de réinterpolation est tracé en trait continu, les intervalles de confiance à 95% en trait discontinu). Le prédicteur (2.55) est obtenu par krigeage avec bruit de mesure. L'EQM empirique est calculée à partir des valeurs des paramètres de moyenne et de corrélation obtenues avec le modèle bruité, mais en utilisant l'équation (2.29) de l'EQM empirique sans bruit de mesure. La variance est estimée en utilisant la formule (2.58).

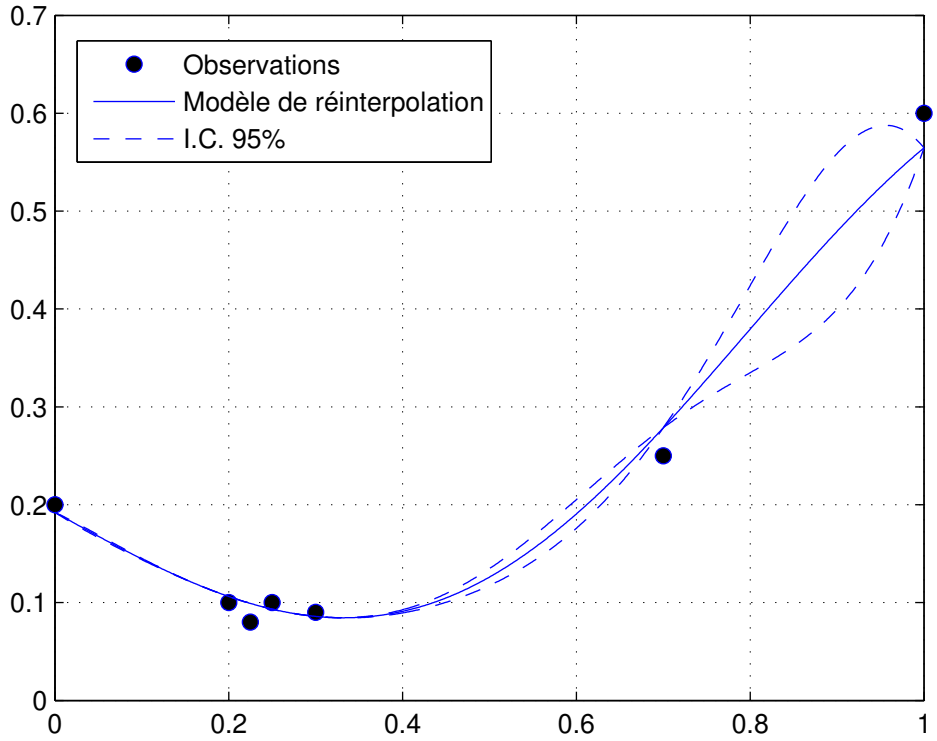

FIGURE 2.6 – Illustration de la technique de réinterpolation. Les observations sont représentées par les cercles pleins, le modèle de réinterpolation en trait continu et les intervalles de confiance à 95% en trait discontinu. On remarque que même si le modèle ne passe pas par les observations, l'erreur de prédiction vaut 0 en ces points.

Chapitre 3

Plans d'expériences

La planification d'expériences correspond au choix des combinaisons des valeurs des facteurs d'entrée, les *points* ou *essais (trials, runs)* $x_1, \ldots, x_n$ du domaine expérimental $\mathcal{X} \subset \mathbb{R}^d$, auxquels on va effectuer les mesures. Un choix judicieux s'avère souvent important en pratique, car le nombre n d'observations est fréquemment assez réduit pour des raisons de coût ou de durée de mise en œuvre des expériences. Le *plan d'expériences (design of experiments, DOE)* sera alors l'ensemble des essais choisis. Il va dépendre du but de l'expérimentation : optimiser la réponse, construire un modèle de la relation entrées/sorties, *etc.* Nous présentons dans ce chapitre différents types de plans d'expériences correspondant à des objectifs variés, appelés *plans optimaux* car optimisant un certain critère. Nous considérons tout d'abord des plans cherchant à répartir les points uniformément dans la région expérimentale, appelés *space-filling designs* ou *exploratory designs*. Ce type de plan est utilisé lorsqu'on ne dispose d'aucune information *a priori* permettant de savoir quelles parties de la région expérimentale $\mathcal{X}$ il est le plus intéressant d'échantillonner. Puis nous verrons des plans optimaux pour des critères liés à la qualité des prédictions du modèle. Enfin, nous examinerons des algorithmes de construction de plans d'expériences séquentiels qui, combinés avec le krigeage, permettent de faire de l'optimisation globale d'une fonction. Pour une introduction plus détaillée aux plans d'expériences, nous renvoyons à [45, 121].

3.1 Plans remplissant l'espace

Les plans remplissant l'espace (« space-filling ») ont pour objectif de répartir les points aussi uniformément que possible dans la région expérimentale. Avec un modèle de krigeage, ils sont utilisés quand la précision de la prédiction sur l'ensemble du domaine d'étude est une priorité : en effet, la variance de krigeage grandit quand on s'éloigne des points, il est donc souhaitable que ceux-ci soient éparpillés dans $\mathcal{X}$. Les plans remplissant l'espace sont construits sans faire d'hypothèse sur le modèle reliant les facteurs et les réponses ; ils peuvent être générés aléatoirement, mais nous évitons cette approche car dans le cas de petits échantillons en grande dimension, les points générés aléatoirement ont tendance à se concentrer à certains endroits (phénomène de *clustering* [144]). Si l'on souhaite générer le plan aléatoirement, une meilleure approche est l'échantillonnage stratifié, qui garantit une bonne équirépartition des points dans le domaine expérimental : nous en verrons un exemple dans le cas des hypercubes latins. Pour des détails sur les plans remplissant l'espace générés aléatoirement, on pourra consulter [49, 144].

3.1.1 Hypercubes latins, tableaux orthogonaux

Si l'on souhaite que les points se projettent uniformément selon certains facteurs, par exemple si l'on pense que la réponse ne dépend que de quelques facteurs ou que le modèle est additif, un plan construit de façon à éparpiller les points dans l'ensemble du domaine expérimental ne sera pas toujours satisfaisant. L'échantillonnage selon un hypercube latin est un moyen de garantir que les projections selon chaque facteur sont uniformes. Dans la suite, un plan généré par hypercube latin sera abusivement appelé hypercube latin.

3.1.1.1 Hypercubes latins

Définition 3.1.1 *[49] Un* hypercube latin (Latin Hypercube Design, LHD) *à n points et d variables d'entrée, noté* $\mathrm{LHD}(n,d)$*, est une matrice de taille $n \times d$ dont chaque colonne est une permutation de $\{1,\dots,n\}$.*

Nous présentons une procédure générale permettant d'obtenir un (plan généré par) hypercube latin particulier de taille n, appelé *midpoint Latin hypercube sampling* ou *centered Latin hypercube sampling*, dans un domaine expérimental de dimension d [49]. On commence par diviser le domaine de variation de chacune des variables d'entrée en n intervalles de même longueur. L'ensemble de tous les produits cartésiens de ces intervalles constitue une partition du domaine expérimental en n^d « cases ». Puis on choisit n cases, de façon à ce que les projections des centres de chaque case soient uniformément réparties sur chacune des coordonnées. Les points du plan sont alors placés aux centres de ces n cases. Plus formellement, voici la procédure à suivre :

- pour $j = 1,\dots,d$, diviser le j^e axe $[a_j, b_j]$ en n parties égales. Les points de la subdivision sont alors
$$a_j, a_j + \frac{b_j - a_j}{n}, \dots, a_j + (n-1)\frac{b_j - a_j}{n}, b_j\,;$$
- se donner une matrice $\Pi = \Pi_{ij}$, de taille $n \times d$, dont les colonnes sont d permutations de $\{1,\dots,n\}$ (la matrice $\mathrm{LHD}(n,d)$ de la définition 3.1.1) ;
- définir les n points du plan comme étant les points de coordonnées

$$\left(a_1 + (\Pi_{i1} - 0,5)\frac{b_1 - a_1}{n}, \dots, a_d + (\Pi_{id} - 0,5)\frac{b_d - a_d}{n}\right), \quad i = 1,\dots,n.$$

Des exemples d'hypercubes latins centrés de taille 6 avec deux facteurs dans $[0,1]$ sont présentés sur la figure 3.1.

Remarque 3.1.2

- *Comme on peut le constater sur la figure 3.1, il est possible d'obtenir, de cette manière, un plan dont les points sont répartis uniformément sur la diagonale du carré. Même si les coordonnées sont réparties uniformément, on ne peut pas dire que les points soient bien répartis dans tout le domaine $[0,1]^2$: le plan à gauche n'est pas « space-filling ». Les plans de gauche et de droite correspondent respectivement aux matrices de permutations*

$$\Pi = \begin{pmatrix} 1 & 1 \\ 2 & 2 \\ 3 & 3 \\ 4 & 4 \\ 5 & 5 \\ 6 & 6 \end{pmatrix} \quad \textit{et} \quad \Pi = \begin{pmatrix} 1 & 5 \\ 2 & 2 \\ 3 & 6 \\ 4 & 4 \\ 5 & 1 \\ 6 & 3 \end{pmatrix} ;$$

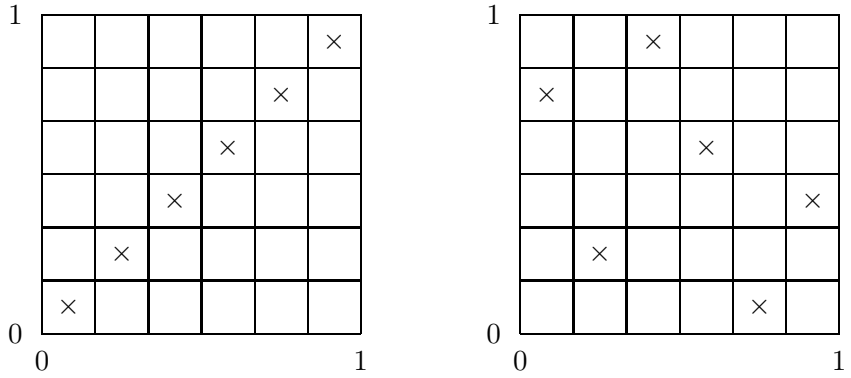

FIGURE 3.1 – Deux hypercubes latins centrés à 6 points dans $[0,1]^2$.

- *une fois les n cases choisies, il est possible de placer les points ailleurs qu'au centre de ces cases. On peut aussi, par exemple, considérer une réalisation de la loi uniforme dans la case (on fait alors de l'échantillonnage stratifié) ;*
- *dans le cas où les facteurs d'entrée sont aléatoires, les n^d « cases » sont déterminées en utilisant les fonctions de répartition marginales des facteurs, de façon à ce que chacune délimite une zone de probabilité $1/n^d$. Le cas déterministe correspond donc à des lois marginales uniformes [144].*

Un des intérêts des hypercubes latins est que, contrairement aux plans factoriels complets, une projection en dimension inférieure ne fait pas apparaître de doublon : en enlevant un ou plusieurs facteurs, les n points ainsi définis sont encore différents.

3.1.1.2 Tableaux orthogonaux

Les tableaux orthogonaux sont une extension des hypercubes latins.

Définition 3.1.3 *[49, 72] On appelle* tableau orthogonal (Orthogonal Array, OA) *de force t, taille n, à d facteurs et s symboles, et on note* $\mathrm{OA}(n,d,s,t)$*, une matrice de taille $n \times d$ dont les entrées sont les s symboles arrangés de telle manière que toute sous-matrice de taille $n \times m$ contienne chacune des s^m lignes possibles le même nombre de fois pour tout $m \leq t$, ou de façon équivalente, de manière à ce que toute sous-matrice de taille $n \times t$ contienne chacune des s^t lignes possibles le même nombre de fois $\lambda = ns^{-t}$. Le paramètre λ est appelé* indice (index) *du tableau.*

Remarque 3.1.4 *[49] Une telle matrice contenant chacune des lignes possibles le même nombre de fois est appelée* plan factoriel complet*, ou* plan complet*.*

En pratique, la taille n du tableau est le nombre de points du plan d'expériences. Le nombre de symboles s est le nombre de niveaux que peuvent prendre chacun des facteurs (chaque facteur prend donc le même nombre de niveaux).

Exemple 3.1.5 *[72] Le tableau suivant est un* $\mathrm{OA}(4,3,2,2)$.

$$\begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}.$$

La construction des tableaux orthogonaux n'est pas évidente. En général, plus la force t est grande, plus la construction est difficile [72]. Une fois que l'on dispose d'un tableau orthogonal, on peut appliquer la procédure présentée dans le paragraphe précédent pour obtenir les points correspondant dans le domaine expérimental $[a_1, b_1] \times \cdots \times [a_n, b_n]$. Les tableaux orthogonaux généralisent la propriété de projection uniforme sur chaque coordonnée des hypercubes latins : si l'on projette les points d'un tableau orthogonal sur t coordonnées ou moins, le résultat est toujours une grille uniforme. Cependant, les tableaux orthogonaux n'existent que pour certaines valeurs des paramètres : il faut en effet $n = \lambda s^t$.

Pour beaucoup plus d'informations sur les hypercubes latins et les tableaux orthogonaux, on pourra consulter [49, 72, 144].

Remarque 3.1.6 *Nous avons vu qu'un hypercube latin (ou un tableau orthogonal) ne suffit pas à définir un plan « space-filling », mais est relativement facile à construire. Nous présentons dans la suite d'autres critères de répartition des points, dont les critères maximin et minimax, permettant d'obtenir des plans space-filling, mais difficiles à générer. C'est pourquoi en pratique on cherche souvent à optimiser un hypercube latin selon un critère « space-filling », de façon à assurer que les points soient bien répartis (un algorithme d'optimisation est présenté dans [77]).*

Une méthode de construction de plans d'expériences dont les points sont répartis uniformément dans des sous-espaces de petite dimension, utilisant une division des sous-espaces en cellules, ainsi qu'un algorithme d'optimisation, sont présentés dans [94].

3.1.2 Plans uniformes

L'idée des plans uniformes est de sélectionner un échantillon dont la répartition est proche de la répartition uniforme. Comme précédemment, supposons que le domaine expérimental s'écrit $\mathcal{X} = [a_1, b_1] \times \cdots \times [a_d, b_d]$. Notons $\mathcal{D} = \{x_1, \ldots, x_n\}$ l'ensemble des n points du plan d'expériences (auxquels on va observer la réponse). Notons

$$F_{\mathcal{U}}(z) = \prod_{i=1}^{d} \left(\frac{z_i - a_i}{b_i - a_i} \right)$$

la fonction de répartition de la loi uniforme sur $\mathcal{X}$. Nous allons évaluer la *discrépance* de l'ensemble $\mathcal{D}$, c'est-à-dire l'écart entre la répartition des points de $\mathcal{D}$ et la répartition uniforme. Notons $F_n(\cdot)$ la fonction de répartition empirique des points de $\mathcal{D}$, *i.e.*

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{[x_{i1}, \infty[\times \cdots \times [x_{id}, \infty[}(x),$$

avec $\mathbb{1}_E(\cdot)$ la fonction indicatrice de l'ensemble E,

$$\mathbb{1}_E(x) = \begin{cases} 1 & \text{si } x \in E\,; \\ 0 & \text{sinon.} \end{cases}$$

La *discrépance* de l'ensemble $\mathcal{D}$, appelée aussi *discrépance* L^∞ ou *discrépance étoile (star discrepancy)* s'écrit

$$D_\infty(\mathcal{D}) = \sup_{x \in \mathcal{X}} |F_n(x) - F_{\mathcal{U}}(x)|\,.$$

Une discrépance plus générale est la *discrépance* L^p,

$$D_p(\mathcal{D}) = \left(\int_{\mathcal{X}} |F_n(x) - F_{\mathcal{U}}(x)|^p \mathrm{d}x \right)^{\frac{1}{p}},$$

qui tend vers la discrépance L^∞ quand $p \to \infty$ [49]. Des plans de discrépance faible seront considérés bien répartis ou *uniformes*.

Définition 3.1.7 *[144] Un* plan uniforme *est un ensemble $\mathcal{D}$ qui minimise $D_p(\cdot)$, pour un certain p.*

Une façon de construire un plan (presque) uniforme est de chercher le plan le plus uniforme dans la famille des hypercubes latins centrés, faciles à générer, dont la construction a été présentée plus haut. En ce qui concerne les performances du critère de discrépance, il est montré dans [144] que le critère D_∞ n'est pas très intuitif si il y a au moins deux facteurs : par exemple, si $d = 2$ et n est pair, un plan similaire au plan à gauche de la figure 3.1 est considéré plus uniforme que le plan dont les points sont situés sur l'autre diagonale. Si $d = 1$, le plan uniforme à n points dans $[0, 1]$ est l'ensemble $\{1/(2n), \ldots, (2n-1)/(2n)\}$, l'hypercube latin centré à n points, ce qui pourrait justifier la méthode de construction donnée ci-dessus. De nombreux tableaux orthogonaux sont aussi des plans uniformes [49].

Les fonctions de discrépance ne sont pas bien adaptées pour construire des plans uniformes : la discrépance L^∞ n'est pas assez sensible, et les discrépances L^p ne prennent pas en compte l'uniformité des points sur les sous-espaces de dimension inférieure de $\mathcal{X}$. De plus, les discrépances L^p ne sont pas invariantes par rotation des coordonnées, en raison du rôle particulier joué par l'origine du repère. Des modifications de la fonction discrépance ont été proposées pour en améliorer les caractéristiques, parmi lesquelles la *discrépance L^2 modifiée*, la *discrépance L^2 centrée* et la *discrépance L^2 wrap-around* ont de bonnes propriétés (les discrépances L^2, L^2 modifiée et L^2 centrée sont cependant équivalentes dans le cas $d = 1$). Pour beaucoup plus d'éléments sur les plans uniformes et les fonctions de discrépance, nous renvoyons à [49]. Nous tenterons au § 4.1.1 d'appliquer la fonction de discrépance à l'ajout séquentiel de points, et constaterons les mêmes défauts que ceux évoqués ici.

Remarque 3.1.8 *Une autre façon de générer un plan uniforme est d'utiliser une* suite à faible discrépance (low-discrepancy sequence) *[53]. Ces suites sont construites en subdivisant le domaine expérimental en « cases » de même volume (comme en 3.1.1), puis en mettant un certain nombre de points dans chacune des cases, avant de recommencer avec des cases plus petites. Les suites de Sobol' par exemple semblent bien résister au phénomène de regroupement (clustering) des points en grande dimension. Un grand avantage des suites de Sobol' est que leur construction peut se faire de façon séquentielle, alors que l'on ne peut pas construire un hypercube latin à $n+1$ points à partir d'un hypercube latin à n points [144]. Les distances inter-points sont en général plus variées, ce qui peut être un avantage pour l'estimation des paramètres de corrélation quand on utilise un modèle de krigeage (§ 2.2.3).*

3.1.3 Plans construits à partir d'un critère de distance

Nous considérons ici des plans d'expériences construits à partir d'un critère utilisant une distance, qui sert à quantifier la plus ou moins bonne répartition des points dans le domaine expérimental. Notons $\mathcal{X} \subset \mathbb{R}^d$ le domaine d'étude, $\mathrm{d}(\cdot, \cdot)$ une distance sur $\mathcal{X}$ et $\mathcal{D} = \{x_1, \ldots, x_n\}$ des points distincts de $\mathcal{X}$. Nous présentons ci-dessous les critères *minimax* et *maximin* introduits dans [79].

3.1.3.1 Maximin distance design

Une première façon de voir les choses est de dire que les points de $\mathcal{D}$ sont bien répartis si ils ne sont pas trop proches les uns des autres, autrement dit si

$$\min_{x_i,x_j\in\mathcal{D}} \mathrm{d}(x_i,x_j) \tag{3.1}$$

est grand.

Définition 3.1.9 *[49, 144] Un plan d'expériences $\mathcal{D}=\{x_1,\ldots,x_n\}$ qui maximise (3.1) est appelé* plan maximin *et noté $\mathcal{D}_{\mathrm{Mm}}$.*

On a donc

$$\max_{\mathcal{D}\subset\mathcal{X}} \min_{x_i,x_j\in\mathcal{D}} \mathrm{d}(x_i,x_j) = \min_{x_i,x_j\in\mathcal{D}_{\mathrm{Mm}}} \mathrm{d}(x_i,x_j) = \mathrm{d}_{\mathrm{Mm}}. \tag{3.2}$$

Il n'y a en général pas unicité dans le choix d'un plan maximin à n points. On peut considérer les plans maximin dont *l'indice (index)*, c'est-à-dire le nombre de paires de points séparées d'une distance d_{Mm}, est minimal (dans le cas où il n'y aurait pas unicité d'un tel plan, il faut recommencer avec la seconde plus proche distance, *etc.* La méthode générale est présentée ci-dessous). L'utilisation d'un plan maximin garantit que deux points quelconques ne sont pas trop proches, et donc que les points sont répartis dans le domaine d'étude. Les points d'un plan maximin ont cependant tendance à être concentrés sur les bords du domaine [79].

La construction d'un plan maximin à n points est un problème de programmation non-linéaire. On pourra consulter [42, 170], qui proposent une méthode de construction d'un plan maximin approché, économe en temps de calcul et efficace aussi dans le cas d'un domaine d'étude non rectangulaire (mais convexe). Les plans maximin factoriels à 2 niveaux ($\mathcal{X}=\{0,1\}^d$) sont étudiés dans [46], avec un lien vers la théorie des codes.

Une façon pratique de construire un plan maximin approché est de se focaliser sur les hypercubes latins : on part de l'ensemble des hypercubes latins à n points et on choisit celui (ou un de ceux) qui satisfait le critère maximin (3.2), appelé *hypercube latin maximin*. Une façon très simple de construire un hypercube latin maximin approché est la suivante : dans un premier temps, générer un certain nombre N_D d'hypercubes latins de taille n, puis choisir parmi les N_D plans obtenus celui qui est maximin. Pour comparer la « maximinité » de deux plans $\mathcal{D}$ et $\mathcal{D}'$, on peut opérer comme suit [120] :

- dresser la liste de toutes les distances entre les points de chaque plan. Puisque les plans contiennent n points, il y a $n(n-1)/2$ distances à calculer pour chacun d'eux, $(d_i, 1\le i\le n(n-1)/2)$ pour $\mathcal{D}$ et $(d'_i, 1\le i\le n(n-1)/2)$ pour $\mathcal{D}'$;
- pour chaque plan, trier la liste des distances obtenues dans l'ordre croissant

$$d_1\le d_2\le\cdots\le d_{\frac{n(n-1)}{2}} \text{ et } d'_1\le d'_2\le\cdots\le d'_{\frac{n(n-1)}{2}} ;$$

- considérer l'ordre suivant

$$(d_1,\ldots,d_m)\prec(d'_1,\ldots,d'_m) \text{ si } \exists j\in 1,\ldots,m \text{ tel que } d_i=d'_i \quad \forall i<j, \text{ et } d_{j+1}<d'_{j+1} ;$$

- poser que

$$\text{« } \mathcal{D}' \text{ est plus maximin que } \mathcal{D} \text{ » si } (d_1,\ldots,d_{\frac{n(n-1)}{2}})\prec(d'_1,\ldots,d'_{\frac{n(n-1)}{2}}).$$

La procédure sera d'autant plus efficace que N_D est grand, ce qui dépendra en pratique des contraintes de temps de calcul.

Remarque 3.1.10 *Le tri les distances inter-points est une étape très coûteuse en temps de calcul. C'est pourquoi, pour un plan $\mathcal{D}$ avec distances inter-points $d_1 \ldots, d_{\frac{n(n-1)}{2}}$, on peut utiliser le critère approché*

$$\phi_p(\mathcal{D}) = \left[\sum_{i=1}^{n(n-1)/2} \frac{1}{d_i^p} \right]^{\frac{1}{p}},$$

pour une valeur de $p \in \mathbb{N}^*$ fixée, et garder l'hypercube latin qui minimise le critère $\phi_p(\cdot)$. Un algorithme stochastique de construction d'un hypercube latin maximin approché à l'aide de ce critère est présenté dans [120], utilisant le recuit simulé (simulated annealing).

Les hypercubes latins maximin à deux facteurs sont étudiés extensivement dans [173] : il y est montré que se restreindre aux hypercubes latins maximin permet d'obtenir des plans proches d'un plan maximin optimal.

Remarque 3.1.11 *[27] Le problème de la construction d'un plan maximin à n points dans $[0,1]^d$ est équivalent au problème de placer dans $[0,1]^d$ les centres de n sphères, disjointes ou d'intersection égale à un singleton, de même rayon r, de façon à ce que r soit le plus grand possible.*

3.1.3.2 Minimax distance design

Une seconde façon d'utiliser la distance $\mathrm{d}(\cdot,\cdot)$ est de considérer que le plan $\mathcal{D}$ est bien réparti dans $\mathcal{X}$ si tout point de $\mathcal{X}$ est assez proche d'un point de $\mathcal{D}$, autrement dit si

$$\max_{x \in \mathcal{X}} \mathrm{d}(x, \mathcal{D}) = \max_{x \in \mathcal{X}} \min_{x_i \in \mathcal{D}} \mathrm{d}(x, x_i) \tag{3.3}$$

est petit.

Définition 3.1.12 *[49, 144] Un plan d'expériences $\mathcal{D} = \{x_1, \ldots, x_n\}$ qui minimise (3.3) est appelé* plan minimax *et noté $\mathcal{D}_{\mathrm{mM}}$.*

On a donc

$$\min_{\mathcal{D} \subset \mathcal{X}} \max_{x \in \mathcal{X}} \mathrm{d}(x, \mathcal{D}) = \max_{x \in \mathcal{X}} \mathrm{d}(x, \mathcal{D}_{\mathrm{mM}}) = \mathrm{d}_{\mathrm{mM}}.$$

Il n'y a en général pas unicité dans le choix d'un ensemble minimax. On peut considérer les plans minimax d'indice maximal (c'est-à-dire dont le plus grand nombre de points de $\mathcal{X}$ est à une distance d_{mM} de $\mathcal{D}$), mais il n'y a toujours pas unicité d'un tel plan, auquel cas on pourrait recommencer avec une distance plus petite, mais la méthode est plus délicate à mettre en œuvre que dans le cas maximin du fait que l'ensemble des distances entre les points de $\mathcal{X}$ et $\mathcal{D}$ est continu. Les points sont moins situés au bord que dans le cas d'un plan maximin [79]. Les plans minimax sont cependant beaucoup plus difficiles à générer que les plans maximin, car la distance doit (en théorie) être évaluée pour tous les points du domaine.

Une étude des plans minimax factoriels à 2 niveaux ($\mathcal{X} = \{0,1\}^d$) est effectuée dans [78], où les performances de ces plans sont évaluées par rapport aux critères classiques des plans factoriels. Les hypercubes latins minimax à 2 facteurs sont étudiés dans [172].

Remarque 3.1.13 *[27] Le problème de la construction d'un plan minimax à n points dans $[0,1]^d$ est équivalent au problème de placer dans $[0,1]^d$ les centres de n sphères de même rayon r qui recouvrent $\mathcal{X}$, de façon à ce que r soit le plus petit possible.*

D'autres critères de distance existent, notamment le *critère de distance moyenne*, pour la présentation duquel nous renvoyons à [144].

3.2 Méthodes paramétriques de construction de plans d'expériences

Nous abordons ici le choix du positionnement des points en vue de construire un bon modèle. Une fois la forme du modèle choisie, les plans sont construits de façon à minimiser une certaine variance : minimisation de la variance des paramètres du modèle, ou minimisation de la variance du prédicteur. Nous présentons dans un premier temps les critères communs utilisés dans le cas d'un modèle de régression classique, puis leur extension au cas d'un modèle de krigeage.

3.2.1 Modèle de régression à erreurs indépendantes

Dans le cadre de la régression classique, on souhaite construire un modèle simple de la relation entrées-sorties, et construire la courbe correspondante appelée *surface de réponse (response surface)* [86]. Quand on dispose de peu d'information sur cette relation, on cherche habituellement à ajuster un *modèle empirique* qui est souvent choisi comme un polynôme de degré 1 ou 2. Plaçons-nous dans le cas d'une seule réponse pour simplifier la présentation. Dans ce cas, les données sont supposées avoir été générées suivant la loi

$$y_i = {}^t m(x_i)\beta + \varepsilon_i, \quad i = 1, \dots, n, \tag{3.4}$$

avec $\beta \in \mathbb{R}^p$ inconnu que l'on va chercher à estimer, et les ε_i indépendantes de moyenne nulle et variance σ^2 inconnue. Dans le cas d'un modèle polynômial de degré 1, on aura $m(x_i) = {}^t(1, x_{i,1}, \dots, x_{i,d})$ et dans le cas d'un modèle de degré 2 on pourra avoir $m(x_i) = {}^t(1, x_{i,1}, \dots, x_{i,d}, x_{i,1}^2, \dots, x_{i,d}^2, x_{i,1}x_{i,2}, \dots, x_{i,d-1}x_{i,d})$ [44]. Sous forme matricielle, le modèle (3.4) se réécrit

$$Y^n = M\beta + \varepsilon,$$

avec $Y^n = {}^t(y_1, \dots, y_n)$, $M = {}^t(m(x_1), \dots, m(x_n))$ et $\varepsilon = {}^t(\varepsilon_1, \dots, \varepsilon_n)$. L'estimateur des moindres carrés de β est donné par

$$\widehat{\beta} = ({}^tMM)^{-1}{}^tMY^n,$$

et sa matrice de variance-covariance par

$$\mathrm{var}(\widehat{\beta}) = \sigma^2({}^tMM)^{-1}. \tag{3.5}$$

La variance de l'estimateur $\widehat{\beta}_i$ est donc le i^e terme diagonal de la matrice ci-dessus. En pratique, on estime σ par l'estimateur sans biais

$$\widehat{\sigma^2} = \frac{\|Y^n - M\widehat{\beta}\|_2^2}{(n-p)}.$$

On part de l'équation (3.5) pour construire le plan d'expériences, en remarquant que les variances des estimateurs des moindres carrés des paramètres du modèle sont proportionnelles aux éléments de l'inverse de la *matrice du plan (design matrix)* $\mathcal{I}_\mathcal{D} = {}^tMM$. On va choisir la position des n points du plan $\mathcal{D} = \{x_1, \dots, x_n\}$ de façon à minimiser ces variances. Il existe plusieurs critères d'optimalité pour choisir les n points du plan, parmi lesquels [66] :

- le critère le plus courant, appelé *D-optimalité* (D pour « déterminant »), évalue le déterminant de la matrice du plan, que l'on va chercher à maximiser. On obtient un *plan D-optimal* en cherchant

$$\mathcal{D}_D^* \in \underset{x_1,\dots,x_n \in \mathcal{X}}{\mathrm{argmax}} \det\left(\mathcal{I}_\mathcal{D}\right).$$

– le critère de *A-optimalité* (A pour « average », moyenne), qui évalue la somme des variances des estimateurs $\widehat{\beta}_1, \ldots, \widehat{\beta}_p$, que l'on souhaite minimiser. On obtient un *plan A-optimal* en cherchant

$$\mathcal{D}_A^* \in \underset{x_1,\ldots,x_n \in \mathcal{X}}{\operatorname{argmin}} \operatorname{tr}\left(\mathcal{I}_{\mathcal{D}}^{-1}\right).$$

– un autre critère moins connu est la *E-optimalité* (E pour « eigenvalue », valeur propre), qui évalue la plus petite valeur propre de la matrice du plan, que l'on cherchera à maximiser.

Le critère de D-optimalité est très populaire du fait qu'il est invariant par transformation linéaire des facteurs de régression ou par reparamétrisation du modèle, alors que les critères de A-optimalité et E-optimalité ne sont pas invariants par changement d'échelle.

D'autres critères prennent en compte la variance des prédictions, donnée par [66]

$$\operatorname{var}\left[\widehat{y}(x)\right] = \sigma^2\, {}^t m(x)\, \mathcal{I}_{\mathcal{D}}^{-1}\, m(x).$$

Parmi ceux-ci,

– le critère populaire de *G-optimalité* (G pour « greatest »), qui évalue la plus grande valeur de la variance de prédiction sur le domaine d'étude,

$$\max_{x \in \mathcal{X}} \operatorname{var}\left[\widehat{y}(x)\right],$$

que l'on va minimiser.

– le critère de *I-optimalité* (I pour « intégrale »), considère la moyenne de la variance de prédiction (que l'on souhaite minimiser)

$$\frac{1}{\sigma^2} \int_{\mathcal{X}} \operatorname{var}\left[\widehat{y}(x)\right] \mu(\mathrm{d}x),$$

où μ désigne une mesure d'intérêt sur $\mathcal{X}$. Après simplification, on obtient un *plan I-optimal*

$$\mathcal{D}_I^* \in \underset{x_1,\ldots,x_n \in \mathcal{X}}{\operatorname{argmin}} \operatorname{tr}\left\{\left(\int_{\mathcal{X}} m(x)\, {}^t m(x) \mu(\mathrm{d}x)\right) \mathcal{I}_{\mathcal{D}}^{-1}\right\}.$$

On remarque que dans tous les cas, l'optimalité d'un plan dépend du modèle choisi *a priori*. De plus, il n'y a le plus souvent pas unicité de l'optimum. De manière générale, le critère de D-optimalité a tendance à placer les points au bord du domaine, alors que les critères de A-optimalité et G-optimalité ont plus tendance à placer les points au centre du domaine (voir [15, 66]). Il existe d'autres critères d'optimalité, pour lesquels on pourra consulter [100, 134]. Sous certaines hypothèses, on peut montrer l'équivalence des critères de D-optimalité et G-optimalité (théorème d'équivalence de Kiefer-Wolfowitz [15, 100, 134]).

En pratique, un plan optimal sera construit en discrétisant le domaine d'étude et en utilisant un algorithme d'échange (voir [15, 66, 191]). Pour la construction séquentielle de plans D-optimaux, on pourra consulter [131].

Pour beaucoup plus de détails sur les surfaces de réponses et les plans optimaux, nous renvoyons à [86, 134].

3.2.2 Plans d'expériences avec modèle de krigeage

Nous présentons maintenant des critères utilisés pour construire des plans optimaux avec un modèle de krigeage. On pourra voir que certains critères étendent ceux présentés au paragraphe précédent. Leur implémentation pratique est cependant plus difficile du fait que les paramètres

de corrélation sont inconnus, et il faut utiliser une partie des points pour les estimer (voir la remarque 3.2.1).

Des critères de construction de plans d'expériences dans le but de construire un modèle de krigeage donnant de bonnes prédictions aux points non échantillonnés sont présentés dans [118, 140, 141, 142]. Soit $Y(x), x \in \mathcal{X} \subset \mathbb{R}^d$ un processus gaussien stationnaire de fonction de covariance $k(\cdot,\cdot) = \sigma^2\rho(\cdot,\cdot)$, et $\mathcal{D} = \{x_1, \ldots, x_n\}$ l'ensemble des points que l'on cherche à placer dans $\mathcal{X}$. Après avoir choisi la fonction de corrélation $\rho(\cdot,\cdot)$ du processus Y, on peut construire le plan en utilisant l'un des critères suivants (notons qu'ici un plan optimal va dépendre des paramètres de corrélation) :

- échantillonnage à maximum d'entropie *(Maximum Entropy sampling)* [154]. On évalue l'entropie de la loi conditionnelle des réponses auxquelles on n'observe pas, que l'on cherchera à minimiser de façon à maximiser l'information apportée par les mesures. Sous certaines hypothèses, on peut montrer [49, 144] qu'un tel plan s'obtient par

$$\mathcal{D}_{\mathrm{ME}} \in \operatorname*{argmax}_{x_1,\ldots,x_n\in\mathcal{X}} \det(R),$$

  avec $R = (\rho(x_i, x_j))_{1\leq i,j\leq n}$ la matrice de corrélation des observations. Comme pour le critère de D-optimalité, ce critère a tendance à placer les points au bord du domaine expérimental. Pour une généralisation de cette idée, on pourra consulter [192] ;
- *Integrated Mean Square Error (IMSE)*,

$$\mathcal{D}_{\mathrm{IMSE}} \in \operatorname*{argmin}_{x_1,\ldots,x_n\in\mathcal{X}} \int_{\mathcal{X}} \frac{\mathrm{EQM}[\widehat{Y}(x)]}{\sigma^2}\mu(\mathrm{d}x),$$

  avec μ une mesure d'intérêt sur $\mathcal{X}$ et l'EQM donnée en (2.16). Ce critère généralise la I-optimalité, et ne place pas les points au bord du domaine ;
- *Maximum Mean Square Error (MMSE)*,

$$\mathcal{D}_{\mathrm{MMSE}} \in \operatorname*{argmin}_{x_1,\ldots,x_n\in\mathcal{X}} \max_{x\in\mathcal{X}} \frac{\mathrm{EQM}[\widehat{Y}(x)]}{\sigma^2},$$

  qui est une généralisation de la G-optimalité. Ce critère ne place pas les points au bord du domaine, et est plus coûteux en temps de calcul que l'IMSE.

En pratique le domaine $\mathcal{X}$ est discrétisé, car l'optimisation sur un domaine continu est délicate. Des algorithmes de construction sont présentés dans [142], et un algorithme d'échange utilisant le *recuit simulé (Simulated Annealing, SA)* [88] est proposé dans [140] (voir aussi [160] pour des propriétés des algorithmes d'échange). Le lien existant entre critères ME, MMSE et minimax et maximin, pour certaines fonctions de corrélation, est établi dans [79, 118].

La grande difficulté dans la construction des plans présentés ici est que les paramètres du modèle sont inconnus, et un plan optimal pour certaines valeurs des paramètres pourra s'avérer médiocre pour d'autres valeurs. Afin de contourner cette difficulté, [141] propose d'utiliser une valeur arbitraire mais robuste des paramètres inconnus, c'est-à-dire telle que le plan optimal ainsi construit soit performant pour un grand nombre d'autres valeurs des paramètres. L'efficacité de la méthode est discutable, car les plans optimaux sont en général peu robustes aux changements de valeurs des paramètres. Il est cependant montré dans [140] que les critères MMSE et IMSE sont assez robustes l'un par rapport à l'autre : un plan optimal pour le critère MMSE est quasi-optimal pour le critère IMSE, et *vice-versa*.

Remarque 3.2.1 *L'inconvénient des critères de construction ci-dessus est qu'ils supposent les paramètres de corrélation connus. Pour faire face à cette difficulté, on peut décider d'utiliser une*

partie des données pour estimer les paramètres dans un premier temps, puis placer les points restants de façon à obtenir de bonnes prédictions. Cependant, un plan adapté pour une bonne estimation des paramètres contiendra des points proches, et n'est donc pas pertinent pour la prédiction car il faudrait des points bien répartis dans l'espace. C'est pourquoi d'autres critères de construction ont été mis au point, avec pour objectif de trouver un plan permettant de construire un bon modèle de krigeage lorsque les paramètres sont inconnus, en utilisant par exemple la matrice d'information de Fisher (§ 2.3.2). Différents critères sont proposés et comparés dans [198, 199, 200], où il est observé qu'un bon plan pour la prédiction et l'estimation des paramètres de corrélation contient des points bien répartis dans le domaine expérimental (pour la prédiction), mais aussi quelques groupes de points proches (pour l'estimation de la corrélation). On pourra consulter [75] pour observer l'influence du choix du plan sur la qualité de l'estimation des paramètres.

Une procédure séquentielle pour d'un bon modèle de krigeage est présentée dans [143]. Un premier modèle est construit à partir d'un plan initial remplissant l'espace (pour que le modèle soit assez précis sur l'ensemble du domaine), puis le plan d'expériences est enrichi du nouveau point

$$x^* = \underset{x\in\mathcal{X}}{\operatorname{argmax}}\, \mathrm{EQM}[\widehat{Y}(x)],$$

là où l'incertitude sur la prédiction est la plus grande. La procédure continue ainsi jusqu'à épuisement des essais à disposition. Dans le cas où l'on souhaite ajouter l points à la fois, le critère proposé est

$$\{x_1^*, \ldots, x_l^*\} = \underset{x_1,\ldots,x_l\in\mathcal{X}}{\operatorname{argmax}} \det(\Sigma^{\mathrm{ap}}),$$

où Σ^{ap} est la matrice de covariance *a posteriori*, $\Sigma^{\mathrm{ap}}_{ii} = \mathrm{EQM}[\widehat{y}(x_i)]$ et $\Sigma^{\mathrm{ap}}_{ij} = \mathrm{cov}[\widehat{y}(x_i), \widehat{y}(x_j)]$. La variance de prédiction s'obtient à partir de (2.29), qui peut se réécrire [142]

$$\mathrm{EQM}[\widehat{Y}(x)] = \sigma^2 \left[1 - \begin{pmatrix} {}^t m(x) & {}^t r(x) \end{pmatrix} \begin{pmatrix} 0 & {}^t M \\ M & R \end{pmatrix}^{-1} \begin{pmatrix} m(x) \\ r(x) \end{pmatrix}\right],$$

et, plus généralement, la covariance entre deux prédicteurs $\widehat{y}(x_i)$ et $\widehat{y}(x_j)$ s'écrit [143]

$$\mathrm{cov}[\widehat{Y}(x_i), \widehat{Y}(x_j)] = \sigma^2 \left[1 - \begin{pmatrix} {}^t m(x_i) & {}^t r(x_i) \end{pmatrix} \begin{pmatrix} 0 & {}^t M \\ M & R \end{pmatrix}^{-1} \begin{pmatrix} m(x_j) \\ r(x_j) \end{pmatrix}\right].$$

3.3 Optimisation de la réponse avec krigeage

L'*optimisation globale* d'une fonction (*i.e.*, la recherche de l'optimum sur l'ensemble du domaine d'étude) est un problème difficile, en raison notamment de la présence éventuelle d'optima locaux multiples de la fonction. Une recherche approfondie dans le domaine d'étude impliquant de nombreuses évaluations de la fonction est parfois impossible pour des raisons de temps de calcul ou de coût des mesures. Il faut alors trouver une stratégie permettant de localiser l'optimum en un nombre limité d'essais. Plusieurs approches, dans deux cadres théoriques différents, essaient d'y remédier : des algorithmes d'optimisation déterministe n'utilisant pas les dérivées de la fonction objectif, et des approches statistiques par des techniques s'apparentant aux plans d'expériences.

Parmi les méthodes d'optimisation déterministes, certaines n'utilisent que les valeurs de la fonction objectif et pas ses dérivées (que celles-ci soient calculées analytiquement ou numériquement), ce qui les rend utilisables pour optimiser des fonctions bruitées. Deux grandes familles peuvent être distinguées.

- La première est constituée des méthodes d'optimisation de type *Pattern Search* ou *Direct Search* [21, 89, 169], qui utilisent simplement des comparaisons des valeurs des réponses, comme la méthode du simplexe dont la plus populaire est celle de Nelder et Mead.
- La seconde est formée des méthodes du type régions de confiance avec construction d'approximations quadratiques de la fonction objectif [130]. À chaque itération, la mesure suivante est déterminée en minimisant une approximation quadratique locale de la fonction objectif, et le modèle quadratique est mis à jour (cette idée a été adaptée pour qu'il soit possible de se passer des dérivées de la fonction objectif).

Mais il existe aujourd'hui des méthodes plus performantes qui sont présentées maintenant.

Une autre idée mise en œuvre pour limiter le nombre d'essais consiste à construire des approximations de la fonction objectif. Ce type de méthode est appelé en anglais *optimization by surrogates*. Nous présentons dans la suite des méthodes d'optimisation globale utilisant un modèle de krigeage, dont l'idée est de trouver un compromis entre l'exploration du domaine et l'exploitation des valeurs données par le modèle [28, 119, 152] (un modèle de régression simple est utilisé dans [132]). On dispose ainsi d'un modèle statistique dont l'incertitude peut être quantifiée. À chaque itération, la mesure suivante est déterminée en optimisant un critère réalisant un compromis entre la recherche de l'optimum du modèle et la recherche en des points où l'incertitude sur le modèle est la plus grande. Le plan initial est choisi de petite taille, de façon à échantillonner le maximum de points restant en se servant de l'information apportée par le modèle de krigeage, et remplit l'espace pour avoir un modèle initial relativement précis sur l'ensemble du domaine d'étude. On s'intéressera dans la suite à la détermination du minimum global d'une fonction déterministe $f(\cdot)$, définie sur un ensemble compact $\mathcal{X} \subset \mathbb{R}^d$ (la recherche du maximum s'effectuerait de façon identique en considérant la fonction $-f(\cdot)$). Plusieurs critères ont été proposés dans la littérature, nous en présentons trois dans la suite (pour une introduction plus détaillée à l'optimisation globale utilisant les surfaces de réponse, nous renvoyons à [80]).

3.3.1 Utilisation des bornes de confiance

L'idée de l'algorithme *SDO (Sequential Design Optimization)* proposé dans [28, 29] est d'utiliser la prédiction et la variance de krigeage pour construire des bornes de confiance, afin d'estimer en tout point du domaine la probabilité que la valeur de la fonction $f(\cdot)$ soit inférieure à la valeur minimale observée jusqu'à présent. Plus précisément, après avoir observé $\{y(x_1), \ldots, y(x_n)\}$, dont le minimum est noté $y_{\min}$, on calcule la prédiction de krigeage $\widehat{y}(x)$ et la variance associée $\mathrm{EQM}_{\widehat{\mu}}(x)$. Puis on estime la *borne de confiance inférieure (Lower Confidence Bound, LCB)* au point x,

$$\mathrm{LCB}(x) = \widehat{y}(x) - b\sqrt{\mathrm{EQM}_{\widehat{\mu}}(x)},$$

où b est un paramètre de contrôle. L'idée est que si les paramètres du modèle sont connus, alors la loi de $Y(x)$ est gaussienne, et pour $b = 1.96$ par exemple on a le résultat classique

$$\mathcal{P}\left(Y(x) < \widehat{y}(x) - 1.96\sqrt{\mathrm{EQM}_{\mu}(x)}\right) = 0.025,$$

ce qui signifie que la valeur $\mathrm{LCB}(x)$ est en-dessous de la vraie valeur de la fonction $Y(x)$ avec grande probabilité (0.975). On n'a donc pas intérêt à choisir un point x du domaine où la valeur $\mathrm{LCB}(x)$ est grande. Ainsi, l'algorithme choisit à chaque itération

$$x^* = \operatorname*{argmin}_{x \in \mathcal{G}} \mathrm{LCB}(x),$$

où $\mathcal{G} \subset \mathcal{X}$ est une grille de points candidats (à laquelle on a retiré les observations). Ce nouveau point est ajouté à l'ensemble des observations, et le minimum observé $y_{\min}$ est mis à jour ainsi que

l'ensemble des points candidats $\mathcal{G}$. L'algorithme continue jusqu'à ce que le nombre d'évaluations spécifié par l'utilisateur soit atteint, ou bien lorsque le critère de convergence

$$y_{\min} < \min_{x \in \mathcal{G}} \mathrm{LCB}(x)$$

est atteint, auquel cas, puisque $\mathrm{LCB}(x)$ est très probablement inférieur à $y(x)$ aux points de $\mathcal{G}$, on a certainement trouvé le minimum global.

Le choix du facteur b n'est pas évident. Indiquons seulement qu'une grande valeur de b privilégie une recherche globale (une grande valeur de la prédiction $\widehat{y}(x)$ peut être compensée par une grande amplitude de l'intervalle de confiance si l'incertitude est grande), alors qu'une petite valeur de b privilégie une recherche locale, où l'on affine le modèle seulement au voisinage du minimum du prédicteur (l'intervalle de confiance est peu étendu autour de la prédiction). L'idée de l'algorithme perd cependant de son sens quand b devient petit : en effet, si les paramètres du modèle sont connus et $b = 0$, alors la valeur $\mathrm{LCB}(x)$ est en-dessous de la vraie valeur de la fonction $Y(x)$ avec seulement une probabilité de 50%.

3.3.2 Espérance de gain

L'*espérance de gain (Expected Improvement, EI)* [150, 151] est également un critère combinant l'information apportée par la surface de réponse (il est souhaitable d'échantillonner au minimum des prédictions) et par l'erreur de prédiction (il est souhaitable d'échantilloner là où la prédiction est la plus mauvaise, afin d'améliorer la qualité de l'approximation). À chaque itération, le point qui maximise l'espérance de gain est ajouté au plan d'expériences. Cet algorithme est appellé *EGO*, pour *Efficient Global Optimization*.

Notons $\{y(x_1), \ldots, y(x_n)\}$ l'ensemble des observations déjà faites, et $y_{\min} = \min(y_1, \ldots, y_n)$. On calcule la prédiction de krigeage $\widehat{y}(x)$ et la variance associée $\mathrm{EQM}_{\widehat{\mu}}(x)$. On estime la loi *a posteriori* en un nouveau point x par $Y(x) \sim \mathcal{N}(\widehat{y}(x), \mathrm{EQM}_{\widehat{\mu}}(x))$ (revoir cependant la remarque 2.2.13). On définit ensuite le *gain (improvement)* en x par rapport à $y_{\min}$ comme étant

$$\mathrm{I}(x) = \begin{cases} y_{\min} - y(x) & \text{si } y(x) < y_{\min}\,; \\ 0 & \text{sinon.} \end{cases}$$

L'espérance de gain $\mathrm{EI}(x)$ est l'espérance de $\mathrm{I}(x)$ par rapport à la loi *a posteriori* de $Y(x)$: $\mathrm{EI}(x) = \mathbb{E}_{Y(x)}[\mathrm{I}(x)]$. Cette quantité est estimée en utilisant l'estimation de la loi *a posteriori* de $Y(x)$ donnée par le krigeage. Un calcul immédiat montre que

$$\widehat{\mathrm{EI}}(x) = \begin{cases} (y_{\min} - \widehat{y}(x))\, \Phi\left(\frac{y_{\min} - \widehat{y}(x)}{\sqrt{\mathrm{EQM}_{\widehat{\mu}}(x)}}\right) + \sqrt{\mathrm{EQM}_{\widehat{\mu}}(x)}\, \phi\left(\frac{y_{\min} - \widehat{y}(x)}{\sqrt{\mathrm{EQM}_{\widehat{\mu}}(x)}}\right) & \text{si } \mathrm{EQM}_{\widehat{\mu}}(x) > 0\,; \\ 0 & \text{si } \mathrm{EQM}_{\widehat{\mu}}(x) = 0, \end{cases} \tag{3.6}$$

où $\phi(\cdot)$ et $\Phi(\cdot)$ désignent respectivement la densité et la fonction de répartition de la loi $\mathcal{N}(0, 1)$. L'espérance de gain est grande en un point où la valeur de la prédiction est beaucoup plus petite que la plus petite valeur observée jusqu'à présent, ou en un point où l'incertitude, mesurée par la variance de krigeage, est grande.

À chaque itération, on ajoute le point où l'espérance de gain est la plus grande,

$$x^* = \operatorname*{argmax}_{x \in \mathcal{X}} \widehat{\mathrm{EI}}(x).$$

En l'absence d'erreur de mesure, on n'a donc pas intérêt à rééchantilloner en un x_i car l'espérance de gain vaut 0, ce qui est conforme à l'intuition puisqu'on connait déjà $y(x_i)$. L'optimisation prend

fin quand le budget de mesures disponibles est atteint, ou quand un critère d'arrêt est satisfait. Le critère d'arrêt proposé mesure la valeur relative de l'espérance de gain par rapport au minimum observé : si celle-ci est inférieure à un certain seuil

$$\frac{\max_{x\in\mathcal{X}}\widehat{\mathrm{EI}}(x)}{|y_{\min}|} < \varepsilon_r,$$

l'algorithme s'arrête. Cet algorithme est particulièrement efficace en petite dimension, où l'optimum est en général trouvé en un nombre très restreint d'essais.

Une version plus générale de l'espérance de gain (*generalized expected improvement*) est présentée dans [152]. Un entier $g \in \mathbb{N}$ étant donné, on évalue la quantité

$$\mathrm{I}^{\mathrm{g}}(x) = \begin{cases} [y_{\min} - y(x)]^g & \text{si } y(x) < y_{\min}\,; \\ 0 & \text{sinon.} \end{cases}$$

Le coefficient g permet de régler le compromis entre recherche globale et locale : une grande valeur de g privilégie une recherche globale, où le gain est grand mais avec une probabilité faible (les points ont tendance à être ajoutés au maximum de la variance de krigeage), et peut être utilisé dans la première partie de l'optimisation ; au contraire, une petite valeur de g privilégie une recherche locale, où le gain est petit mais avec une grande probabilité (le minimum est alors approché très finement). Comme dans le cas $g = 1$, l'estimation de l'espérance de gain généralisée à l'aide du krigeage s'obtient de façon analytique.

Pour plus de détails sur l'algorithme EGO, on pourra consulter [81, 144]. Dans [145] les performances de l'algorithme EGO sont comparées avec celles de critères alternatifs proches présentés dans [185], et des indications sont données pour l'implémentation pratique de tels critères.

La convergence de l'algorithme EGO est démontrée dans [179] pour un processus centré dont la fonction de covariance $k(\cdot,\cdot)$ vérifie certaines propriétés (en particulier, le RKHS $\mathcal{H}$ engendré par $k(\cdot,\cdot)$ est formé de fonctions continues). La fonction à optimiser $f(\cdot)$ est supposée appartenir à $\mathcal{H}$, et les observations se font sans bruit de mesure. Il est montré qu'alors la suite de points construite par l'algorithme utilisant l'espérance de gain est dense dans $\mathcal{X}$, ce qui prouve le résultat.

Remarque 3.3.1 *L'inclusion de bruit de mesure dans l'algorithme EGO n'est pas évidente. Des essais ont montré que la formule (3.6) ne donne pas de résultats satisfaisants dans le cas d'un modèle bruité. En pratique, nous avons remarqué que les résultats sont meilleurs en remplaçant $y_{\min}$ par la plus petite prédiction donnée par le modèle de krigeage bruité, $y'_{\min} = \min_{i=1,\dots,n} \widehat{y}(x_i)$, et en prenant pour variance de prédiction celle donnée par le modèle ne prenant pas en compte le bruit (garantissant ainsi que les observations ne se feront jamais deux fois au même endroit, revoir le paragraphe 2.4.2 consacré à la réinterpolation).*

3.3.3 Utilisation de l'entropie

L'idée de l'algorithme *IAGO (Informal Approach to Global Optimization)* [180] est d'observer au point où l'information apportée sur la loi du minimum est maximale. Cette méthode utilise l'entropie (§ 4.1.4). Comme précédemment, on note $Y^n = \{y(x_1), \dots, y(x_n)\}$ les observations dont on dispose, et $y_{\min} = \min(y_1, \dots, y_n)$. On estime dans un premier temps les paramètres du modèle de krigeage (§ 2.2.3). Notons

$$\Omega_{\mathcal{X}} = \{x \in \mathcal{X}, Y(x) = \min_{t\in\mathcal{X}} Y(t)\}$$

l'ensemble (aléatoire) des minima globaux de $Y(\cdot)$ sur $\mathcal{X}$. Cet ensemble existe si les trajectoires de $Y(\cdot)$ sont continues presque sûrement, ce qui est garanti par des hypothèses convenables sur la fonction de covariance (voir le § 2.1.2). On aimerait évaluer la distribution conditionnelle $p_{U|Y^n}(\cdot)$ de U, une variable aléatoire uniformément distribuée sur $\Omega_{\mathcal{X}}$. Comme il n'existe pas de formule analytique, la solution proposée est de discrétiser le domaine d'étude $\mathcal{X}$ et d'utiliser des réalisations du processus aléatoire avec une méthode de Monte Carlo. Soit donc $\mathcal{G} = \{t_1, \ldots, t_g\} \subset \mathcal{X}$ une discrétisation du domaine d'étude, $\Omega_{\mathcal{G}}$ l'ensemble aléatoire des minima globaux de $Y(\cdot)$ sur $\mathcal{G}$ et $U_{\mathcal{G}}$ une variable aléatoire uniformément distribuée sur $\Omega_{\mathcal{G}}$. On estime la densité conditionnelle de $U_{\mathcal{G}}$,

$$p_{U_{\mathcal{G}}|Y^n}(x) = \mathcal{P}(U_{\mathcal{G}} = x|Y^n), \quad x \in \mathcal{G},$$

de la façon suivante : supposons que l'on a simulé (sur $\mathcal{G}$) l réalisations conditionnelles du processus notées $z_1(\cdot), \ldots, z_l(\cdot)$, et notons, pour $i = 1, \ldots, l$, $\xi_i = \operatorname{argmin}_{x \in \mathcal{G}} z_i(x)$ (si le minimum n'est pas unique, en choisir un au hasard). Alors, on peut montrer que, $\forall x \in \mathcal{G}$, la masse de probabilité empirique

$$\widehat{p}_{U_{\mathcal{G}}|Y^n}(x) = \frac{1}{l} \sum_{i=1}^{l} \delta_{\xi_i}(x)$$

tend presque sûrement vers $p_{U_{\mathcal{G}}|Y^n}(x)$ quand l tend vers l'infini. De plus, $[U_{\mathcal{G}}|Y^n]$ converge en loi vers $[U|Y^n]$ quand $\mathcal{G}$ devient dense dans $\mathcal{X}$. On peut alors estimer l'entropie conditionnelle de la loi du minimum apportée par une nouvelle observation au point x,

$$H_n(x) = H(U|Y^n, Y(x)),$$

par une quantité $\widehat{H}_n(x)$ (voir [180] pour la définition de l'entropie conditionnelle et les détails des calculs). Le choix du nouveau point se fera en

$$x^* = \operatorname*{argmin}_{x \in \mathcal{X}} \widehat{H}_n(x).$$

La méthode permet la prise en compte du bruit de mesure de façon naturelle. L'estimation de la densité conditionnelle par Monte Carlo fait que la méthode est cependant plus coûteuse en temps de calcul que l'espérance de gain.

Chapitre 4

Maximisation de la diversité des réponses

Nous entrons à présent au cœur de la problématique présentée dans l'introduction. On s'intéresse à un système inconnu $f : x \in \mathcal{X} \subset \mathbb{R}^d \to \mathcal{Y} \subset \mathbb{R}^q$, avec un domaine atteignable en sortie $\mathcal{Y}$ inconnu.

Approche théorique souhaitable, mais impossible à mettre en œuvre. Nous sommes à la recherche d'un critère d'ajout de points permettant de répondre au problème inverse posé par l'expérimentateur : une fois que l'ensemble des essais $x_1, \ldots, x_N$ aura été effectué et les réponses $Y^N = {}^t(y_1, \ldots, y_N)$ observées, il faudra, pour une valeur de la réponse $y \in \mathcal{Y}$ donnée, être capable de déterminer une entrée $x_y \in \mathcal{X}$ telle que $f(x_y) \approx y$. Si l'on modélise la réponse par un processus gaussien $Y(\cdot)$ en utilisant le krigeage (chapitre 2), une façon de mesurer *a posteriori* l'éloignement de $Y(x)$ à y est donnée par

$$\begin{aligned} \zeta_N(x, y) &= \mathbb{E}\left[(Y(x) - y)^2 \middle| Y^N\right] \\ &= \mathrm{EQM}(x) + (\widehat{y}(x) - y)^2, \end{aligned}$$

avec $\widehat{y}(x)$ la prédiction en x donnée par le krigeage et $\mathrm{EQM}(x)$ l'erreur quadratique moyenne associée. La *prédiction inverse* de y est alors définie par

$$x_y = \operatorname*{argmin}_{x \in \mathcal{X}} \zeta_N(x, y),$$

qui est le point où l'on observera si l'on souhaite obtenir une réponse proche de y, au sens de la « distance » $\zeta_N(\cdot, \cdot)$. Un critère de construction du plan d'expériences répondant au problème posé est donc

$$\{x_1^*, \ldots, x_N^*\} = \operatorname*{argmin}_{x_1, \ldots, x_N \in \mathcal{X}} \mathbb{E}\left[\int_{\mathcal{Y}} \zeta_N(x_y, y)\, \mu(\mathrm{d}y)\right],$$

avec μ une mesure sur $\mathcal{Y}$ choisie de façon à indiquer l'intérêt de l'utilisateur pour les différentes valeurs des réponses. Ce critère est malheureusement trop compliqué à implémenter en pratique.

Approche retenue : diversité des sorties. Le problème précédent est reformulé de la façon suivante : on souhaite, une fois le plan construit, que les réponses obtenues soient « bien réparties » dans le domaine atteignable $\mathcal{Y}$. L'idée est que si l'on sait atteindre un ensemble de réponses bien réparti dans $\mathcal{Y}$, on devrait être capable d'atteindre rapidement n'importe quelle

réponse en utilisant les observations adjacentes et les entrées correspondantes. La notion de « bonne répartition » d'un ensemble de points sera appelée *diversité* dans la suite. Nous nous intéressons à la manière de quantifier la diversité d'un ensemble de points, afin de pouvoir évaluer la diversité des réponses que l'on attend d'une série de mesures : la série de mesures qu'effectuera l'expérimentateur sera celle dont il espère qu'elle résultera en une diversité maximale des réponses. Le terme « diversité » a été utilisé dans [183], et une étude sur le sujet a été effectuée à l'IFP dans [168].

Rappelons que l'on dispose d'un budget fini de N mesures. Il serait peu judicieux de planifier les N mesures en même temps, car on ne dispose que de très peu d'information sur la relation entre les facteurs et les réponses. Nous allons donc supposer que l'on a déjà effectué $n < N$ mesures $\{(x_1, y_1) \dots, (x_n, y_n)\}$, et allons étudier la diversité globale des $n+1$ réponses $y_1, \dots, y_{n+1}$ que l'on attend d'une mesure supplémentaire en x_{n+1}. Il nous faut donc dans un premier temps construire une fonction $\mathrm{div}_n(y_{n+1})$ qui mesure la diversité de l'ensemble $\{y_1, \dots, y_{n+1}\}$. Mais après avoir trouvé la réponse y^*_{n+1} maximisant la fonction $\mathrm{div}_n(\cdot)$, encore faut-il être capable de trouver le facteur (ou un des facteurs) x^*_{n+1} vérifiant $y^*_{n+1} = f(x^*_{n+1})$! C'est pourquoi nous allons construire une fonction des facteurs d'entrée, qui va évaluer en x_{n+1} la diversité *attendue* des $n+1$ réponses $\{y_1, \dots, y_{n+1}\}$ (l'idée est inspirée par l'espérance de gain, présentée au chapitre précédent). Le krigeage permet d'obtenir la loi de probabilité conditionnelle de la sortie $[Y(x)|Y^n]$ pour chaque facteur d'entrée x (remarque 2.2.13), nous allons donc prendre la moyenne de la fonction $\mathrm{div}_n(\cdot)$ pondérée par la densité de probabilité de la loi $[Y(x)|Y^n]$. Au moment de choisir le nouveau facteur d'entrée pour la prochaine mesure, c'est donc un point

$$x^* \in \operatorname*{argmax}_{x \in \mathcal{X}} \mathbb{E}\, \mathrm{div}_n(Y_n(x)), \tag{4.1}$$

avec $Y_n(x)$ distribuée selon la loi de $Y(x)$ conditionnelle à Y^n, qui sera choisi. Il est donc souhaitable de trouver une fonction de diversité telle que l'espérance ci-dessus se calcule aisément quand la loi de $Y_n(x)$ est normale.

Plan du chapitre. Nous étudions dans un premier temps des fonctions permettant de mesurer la diversité apportée par un point y_{n+1} à un ensemble de points $y_1, \dots, y_n$, tout d'abord en dimension 1 pour fixer les idées ($y_i \in \mathbb{R}$, le système a une seule sortie), puis nous passons aux dimensions du problème pratique ($y_i \in \mathbb{R}^2$, il y a 2 sorties : voir l'introduction). Nous vérifions à chaque fois si la fonction $\mathrm{div}_n(y_{n+1})$ mesure correctement la diversité de l'ensemble $\{y_1, \dots, y_{n+1}\}$, puis, si c'est le cas, nous nous assurons que l'espérance (4.1) se calcule sans trop de difficulté. Nous voyons finalement si les fonctions de diversité ayant passé cet examen préliminaire peuvent être modifiées pour prendre en compte les contraintes particulières à l'étude, à savoir la présence d'erreurs de mesure, les mesures qui se font par séries de 6 et le retard d'arrivée des résultats.

4.1 Diversité dans $\mathbb{R}$

L'objectif de l'étude étant de maximiser la diversité des réponses, il s'agit dans un premier temps de quantifier ce que l'on appelle diversité d'un ensemble de points. Nous nous sommes intéressés à une fonction de diversité pouvant être utilisée dans un cadre séquentiel, c'est-à-dire qui, partant d'un ensemble de points existants $y_1, \dots, y_n$, donne un critère pour choisir le nouveau point y_{n+1} à ajouter de façon à maximiser la diversité de l'ensemble $\{y_1, \dots, y_n, y_{n+1}\}$. Afin de fixer les idées, nous commençons par le cas le plus simple, où l'ensemble des points est défini dans $\mathbb{R}$.

En pratique y n'est pas une variable indépendante que l'on peut choisir : en effet, chaque valeur de sortie y_i correspond à une observation en un facteur d'entrée x_i, et le nouveau point y_{n+1} s'écrit en fait $y_{n+1} = y(x_{n+1})$, avec x_{n+1} une nouvelle entrée. Par une approche bayésienne, le krigeage fournit cependant une approximation de la loi conditionnelle de la sortie en un point x, $Y_n(x) \sim \mathcal{N}(\mu_n(x), \sigma_n^2(x))$. Il est donc souhaitable de disposer de fonctions de diversité permettant la prise en compte de cette incertitude (par exemple une fonction polynomiale $P(y_{n+1})$, de sorte que l'espérance $\mathbb{E}P(Y(x_{n+1}))$ puisse être évaluée analytiquement si la loi de $Y(x_{n+1})$ est normale).

Dans la suite, nous allons tout d'abord tester la minimisation de la discrépance, qui est une mesure du caractère uniforme d'une distribution de points. Nous testons également la maximisation de deux fonctions du type « plus proche voisin », car ces fonctions sont généralement simples à évaluer. Finalement, nous nous intéressons à la maximisation de l'entropie, qui est également un critère mesurant l'uniformité de la répartition d'un ensemble de points.

4.1.1 Étude de la fonction discrépance

Nous avons vu au § 3.1.2 que la discrépance, qui est une mesure de l'éloignement de la répartition d'un ensemble de points à la répartition uniforme, permet de construire des plans d'expériences remplissant l'espace (plus petite est la discrépance, meilleure est la diversité). Il est donc intéressant de tester l'efficacité de ce critère dans le cas séquentiel. C'est l'objet de ce premier paragraphe.

Plaçons-nous dans le cas où les points sont dans $[0,1]$ (on se ramène aisément à ce cas par une normalisation). Soient donc $y_1 \leq y_2 \leq \cdots \leq y_n \in [0,1]$. On note $F_{\mathcal{U}}$ la fonction de répartition de la loi uniforme sur $[0,1]$, et F_n la fonction de répartition empirique de l'échantillon $Y^n = \{y_1, \ldots, y_n\}$. La mesure de la discrépance L^p est donnée par

$$D_p(Y^n) = \|F_n - F_{\mathcal{U}}\|_p.$$

Souhaitant ajouter un point $y_{n+1} \in [0,1]$, nous allons évaluer la discrépance de l'échantillon $\{y_1, \ldots, y_n, y_{n+1}\}$. Notons $D_p^n(y) = D_p(Y^n \cup \{y\})$, qui mesure la discrépance de l'ensemble de points $\{y_1, \ldots, y_n, y\}$ dans $[0,1]$. Nous avons restreint notre attention aux fonctions de discrépance L^∞, L^1 et L^2 qui sont faciles à évaluer.

Proposition 4.1.1 *La fonction $D_\infty^n(\cdot)$ est continue et affine par morceaux. Les pentes prennent uniquement les valeurs $-1, 0$ et 1. Si l'on suppose $0 \leq y_1 < y_2 \cdots < y_n \leq 1$, $D_\infty^n(\cdot)$ a pour expression analytique*

$$D_\infty^n(y) = \begin{cases} -y + \frac{1}{n+1}, & si\ y \in [0, y_1] \cap [0, \frac{1}{n+1} - R_{0,1}] \cap [0, \frac{1}{2(n+1)}]\,; \\ R_{0,1}, & si\ y \in [0, y_1] \cap [0, \frac{1}{n+1}] \cap [0, R_{0,1}] \cap [\frac{1}{n+1} - R_{0,1}, 1]\,; \\ y, & si\ y \in [0, y_1] \cap [0, \frac{1}{n+1}] \cap [R_{0,1}, 1] \cap [\frac{1}{2(n+1)}, 1]\,; \\ R_{0,1}, & si\ y \in [0, y_1] \cap [\frac{1}{n+1}, 1] \cap [0, R_{0,1}]\,; \\ y, & si\ y \in [0, y_1] \cap [\frac{1}{n+1}, 1] \cap [R_{0,1}, 1]\,; \\ -y + \frac{i+1}{n+1}, & si\ y \in [y_i, y_{i+1}] \cap [0, \frac{i}{n+1}] \cap [0, \frac{i+1}{n+1} - R_{i,i+1}]\,; \\ R_{i,i+1}, & si\ y \in [y_i, y_{i+1}] \cap [0, \frac{i}{n+1}] \cap [\frac{i+1}{n+1} - R_{i,i+1}, 1]\,; \\ -y + \frac{i+1}{n+1}, & si\ y \in [y_i, y_{i+1}] \cap [\frac{i}{n+1}, \frac{i+1}{n+1}] \cap [0, \frac{2i+1}{2(n+1)}] \cap [0, \frac{i+1}{n+1} - R_{i,i+1}]\,; \\ R_{i,i+1}, & si\ y \in [y_i, y_{i+1}] \cap [\frac{i}{n+1}, \frac{i+1}{n+1}] \cap [0, R_{i,i+1} + \frac{i}{n+1}] \cap [\frac{i+1}{n+1} - R_{i,i+1}, 1]\,; \\ y - \frac{i}{n+1}, & si\ y \in [y_i, y_{i+1}] \cap [\frac{i}{n+1}, \frac{i+1}{n+1}] \cap [\frac{2i+1}{2(n+1)}, 1] \cap [\frac{i}{n+1} + R_{i,i+1}, 1]\,; \\ R_{i,i+1}, & si\ y \in [y_i, y_{i+1}] \cap [\frac{i+1}{n+1}, 1] \cap [0, R_{i,i+1} + \frac{i}{n+1}]\,; \\ y - \frac{i}{n+1}, & si\ y \in [y_i, y_{i+1}] \cap [\frac{i+1}{n+1}, 1] \cap [\frac{i}{n+1} + R_{i,i+1}, 1]\,; \\ -y + 1, & si\ y \in [y_n, 1] \cap [0, \frac{n}{n+1}] \cap [0, 1 - R_{n,n+1}]\,; \\ R_{n,n+1}, & si\ y \in [y_n, 1] \cap [0, \frac{n}{n+1}] \cap [1 - R_{n,n+1}, 1]\,; \\ -y + 1, & si\ y \in [y_n, 1] \cap [\frac{n}{n+1}, 1] \cap [0, \frac{2n+1}{2(n+1)}] \cap [0, 1 - R_{n,n+1}]\,; \\ R_{n,n+1}, & si\ y \in [y_n, 1] \cap [\frac{n}{n+1}, 1] \cap [0, R_{n,n+1} + \frac{n}{n+1}] \cap [1 - R_{n,n+1}, 1]\,; \\ y - \frac{n}{n+1}, & si\ y \in [y_n, 1] \cap [\frac{2n+1}{2(n+1)}, 1] \cap [R_{n,n+1} + \frac{n}{n+1}, 1], \end{cases}$$

avec

$$\begin{aligned} R_{0,1} &= \sup\left(\left|y_1 - \frac{1}{n+1}\right|, \left|y_1 - \frac{2}{n+1}\right|, \ldots, \left|y_i - \frac{i}{n+1}\right|, \left|y_i - \frac{i+1}{n+1}\right|, \ldots, \left|y_n - \frac{n}{n+1}\right|, 1 - y_n\right)\,; \\ R_{i,i+1} &= \sup\left(y_1, \left|y_1 - \frac{1}{n+1}\right|, \ldots, \left|y_i - \frac{i-1}{n+1}\right|, \left|y_i - \frac{i}{n+1}\right|, \left|y_{i+1} - \frac{i+1}{n+1}\right|, \left|y_{i+1} - \frac{i+2}{n+1}\right|, \ldots,\right. \\ &\quad \left.\left|y_n - \frac{n}{n+1}\right|, 1 - y_n\right) \quad pour\ i = 1, \ldots, n-1\,; \\ R_{n,n+1} &= \sup\left(y_1, \left|y_1 - \frac{1}{n+1}\right|, \ldots, \left|y_i - \frac{i-1}{n+1}\right|, \left|y_i - \frac{i}{n+1}\right|, \ldots, \left|y_n - \frac{n-1}{n+1}\right|, \left|y_n - \frac{n}{n+1}\right|\right). \end{aligned}$$

Preuve *Voir l'annexe G.* □

Sur la figure 4.1 nous avons tracé, pour un échantillon $\{y_1, \ldots, y_n\}$ donné représenté par des cercles pleins, la fonction de répartition de la loi uniforme sur $[0, 1]$ en trait discontinu, la fonction de répartition empirique de l'échantillon en trait continu, et la fonction discrépance L^∞ en trait continu fort. On souhaite ajouter l'argument minimum y^* de la fonction discrépance L^∞, de façon à ce que la distribution des $n+1$ points $\{y_1, \ldots, y_n, y^*\}$ soit la plus proche possible de la loi uniforme. Comme on peut le constater sur la figure 4.1, les minima forment un palier, ce qui pose un problème de détermination de l'argument minimum. De plus, les points d'abcisse $0.17, 0.26, 0.28, 0.52$, déjà présents dans l'échantillon, minimisent la fonction $D_\infty^n(\cdot)$: on aura donc tendance à ajouter des points aux mêmes endroits, ce qui n'est pas bon en terme de diversité. Cela confirme la remarque sur le manque de sensibilité de la discrépance L^∞ donnée dans [49] page 70.

Remarque 4.1.2 *Nous avons aussi considéré la* distance de Lévy *[203] entre deux fonctions de répartition F et G sur $\mathbb{R}$, définie par*

$$\mathrm{L}(F, G) = \inf\left\{\varepsilon > 0 | F(x - \varepsilon) - \varepsilon \leq G(x) \leq F(x + \varepsilon) + \varepsilon \ \ \forall x \in \mathbb{R}\right\},$$

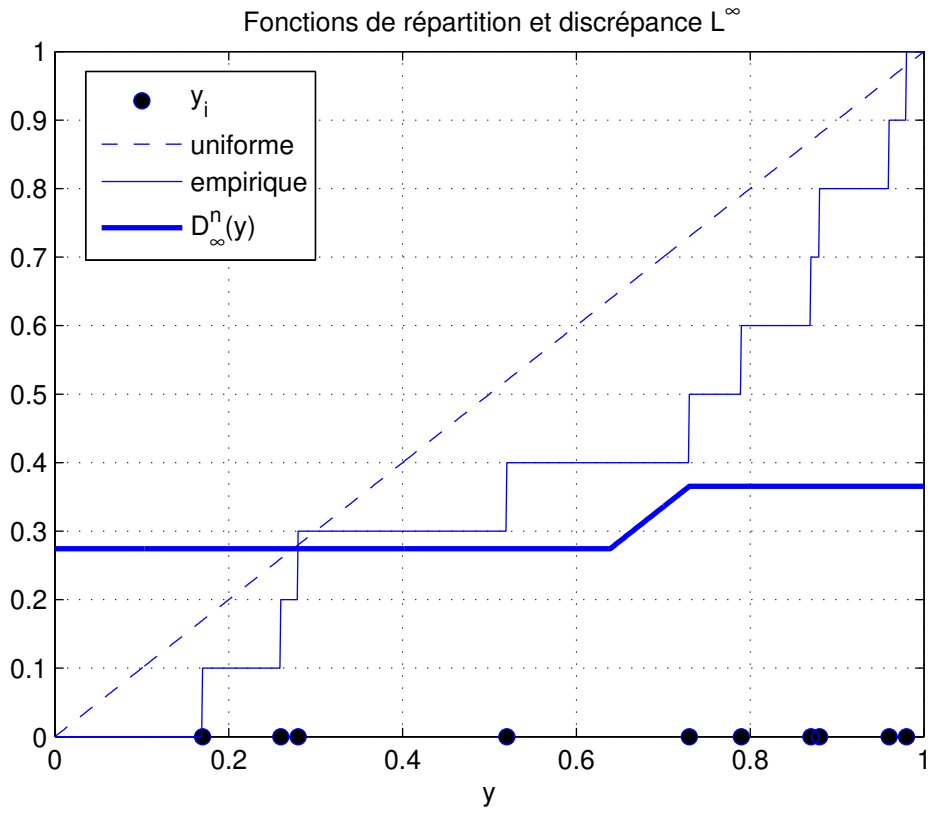

FIGURE 4.1 – Fonctions de répartition uniforme (en trait discontinu) et empirique (en trait continu fin), et fonction D_∞^n (en trait continu fort). Les cercles pleins désignent les points du plan courant.

et plus précisément la quantité $\mathrm{L}(Y^n) = \mathrm{L}(F_n, F_{\mathcal{U}})$ *qui mesure la distance de Lévy entre la fonction de répartition empirique et la fonction de répartition de la loi uniforme sur* $[0,1]$. *On a*

$$\begin{aligned}
\mathrm{L}(Y^n) &= \inf\{\varepsilon > 0 | F_{\mathcal{U}}(x-\varepsilon) - \varepsilon \leq F_n(x) \leq F_{\mathcal{U}}(x+\varepsilon) + \varepsilon \ \ \forall x \in \mathbb{R}\} \\
&= \inf\{\varepsilon > 0 | x - 2\varepsilon \leq F_n(x) \leq x + 2\varepsilon \ \ \forall x \in [0,1]\} \\
&= \inf\{\varepsilon > 0, |F_n(x) - x| \leq 2\varepsilon \ \ \forall x \in [0,1]\} \\
&= \frac{1}{2} D_\infty(Y^n),
\end{aligned}$$

et on retrouve donc, à un facteur multiplicatif près, la quantité $D_\infty(Y^n)$.

La distance de Wasserstein *[37, 138] entre deux mesures de probabilité* μ *et* ν *définies sur un même espace probabilisé, ou plus précisément sa version* L^2

$$W(\mu,\nu) = \inf_{X\sim\mu, Y\sim\nu} \sqrt{\mathbb{E}(X-Y)^2} = \inf_{X\sim\mu, Y\sim\nu} \sqrt{\mathbb{E}X^2 + \mathbb{E}Y^2 - 2\mathbb{E}XY},$$

fait aussi intervenir leurs fonctions de répartition et peut être envisagée dans ce paragraphe. On peut montrer (voir l'annexe H) que si $\mu \sim \mathcal{U}([0,1])$ *et* $\nu_n(y) \sim 1/(n+1)(\sum_{i=1}^n \delta_{y_i} + \delta_y)$, *alors*

$$W_n(y) = W(\mu, \nu_n(y)) = \sqrt{\frac{1}{3} + \frac{1}{n+1}\sum_{i=1}^{n+1} y_i^2 - \frac{1}{(n+1)^2}\sum_{i=1}^{n+1}(2i-1)y_{(i)}}\,,$$

où $y_{n+1} = y$. Le critère $W_n(y)$, que l'on cherche à minimiser, est tracé sur la figure 4.2 pour le même ensemble d'observations que précédemment. On remarque que ce critère est bien discriminant (le minimum est unique), mais peu intuitif. De plus, les calculs seront très délicats dans le cas de deux sorties. Nous ne retenons donc pas ce critère pour la suite de notre étude.

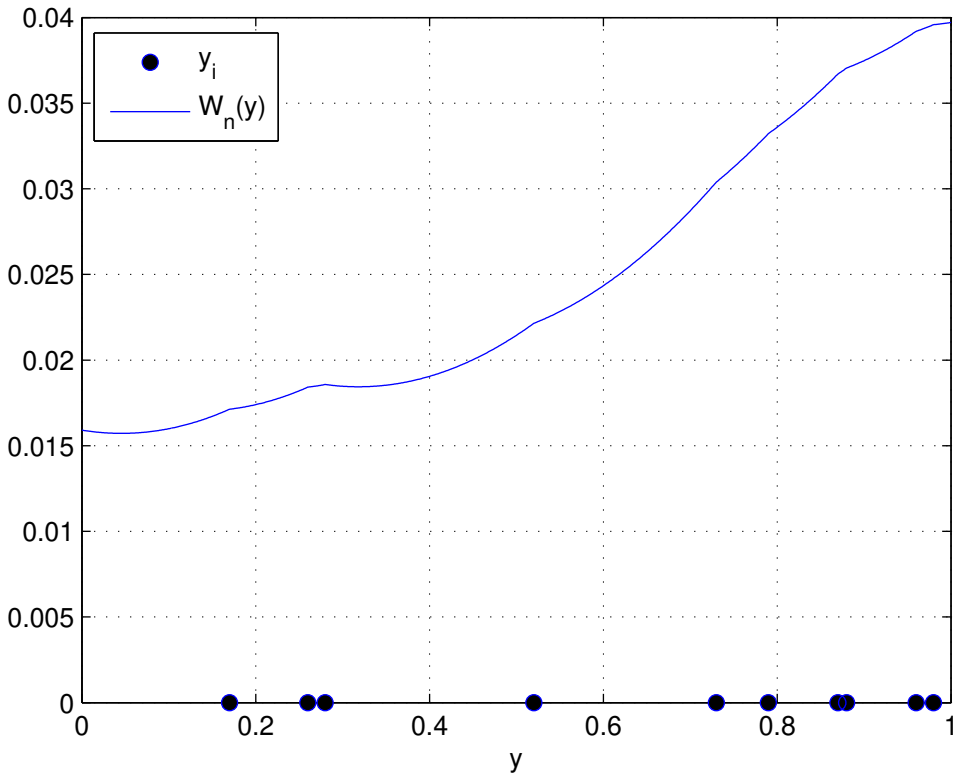

FIGURE 4.2 – Critère de distance de Wasserstein. Les cercles pleins désignent les points du plan courant.

Étudions maintenant les fonctions de discrépance L^1 et L^2.

Proposition 4.1.3 *La fonction $D_1^n(\cdot)$ est continue et polynomiale de degré 2 par morceaux. En réarrangant $\{y_1, \ldots, y_n, y_{n+1} = y\}$ dans l'ordre croissant $y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(n+1)}$, $D_1^n(\cdot)$ a pour expression analytique*

$$D_1^n(y) = \frac{1}{2} \sum_{i=1}^{n+1} \left[(-1)^{\mathbb{1}_{]\frac{i}{n+1}, \infty[}(y_{(i)})} \left(y_{(i)} - \frac{i}{n+1} \right)^2 - (-1)^{\mathbb{1}_{]\frac{i-1}{n+1}, \infty[}(y_{(i)})} \left(y_{(i)} - \frac{i-1}{n+1} \right)^2 \right].$$

Preuve *Voir l'annexe G.* □

Proposition 4.1.4 *La fonction $D_2^n(\cdot)$ est continue. En réarrangant $\{y_1, \ldots, y_n, y_{n+1} = y\}$ dans l'ordre croissant $y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(n+1)}$, $D_2^n(\cdot)$ a pour expression analytique*

$$D_2^n(y) = \sqrt{\frac{1}{3} \sum_{i=1}^{n+1} \left[\left(y_{(i)} - \frac{i-1}{n+1} \right)^3 - \left(y_{(i)} - \frac{i}{n+1} \right)^3 \right]}. \tag{4.2}$$

Preuve *Voir l'annexe G.* □

Remarque 4.1.5 *On peut vérifier que la formule (4.2), valable pour $q = 1$, coïncide avec la formule (3.5) de [49] pour $\mathbb{R}^q$,*

$$D_2(Y^n) = \sqrt{\frac{1}{3^q} - \frac{1}{2^{q-1}n}\sum_{k=1}^{n}\prod_{l=1}^{q}(1-y_{kl}^2) + \frac{1}{n^2}\sum_{k=1}^{n}\sum_{j=1}^{n}\prod_{i=1}^{q}[1-\max(y_{ki},y_{ji})]},$$

avec $y_i = (y_{i1}, \ldots, y_{iq})$ pour $i = 1, \ldots, n$.

Nous avons évoqué en 3.1.2 des modifications ayant été apportées à la discrépance L^2 pour en améliorer les caractéristiques (le cadre L^2 permettant des calculs simples). Dans le cas de la dimension 1, les discrépances modifiée et centrée étant équivalentes à la discrépance L^2, il reste la discrépance wrap-around.

Proposition 4.1.6 *La fonction de discrépance wrap-around $D_{\text{wa}}^n(\cdot)$ s'écrit, en notant $y = y_{n+1}$,*

$$D_{\text{wa}}^n(y) = \frac{4}{3} + \frac{1}{(n+1)^2}\sum_{i,j=1}^{n+1}\left[\frac{3}{2} - |y_i - y_j|\,(1 - |y_i - y_j|)\right].$$

Preuve *Voir la formule (3.8) de [49], qui donne le résultat pour q quelconque.* □

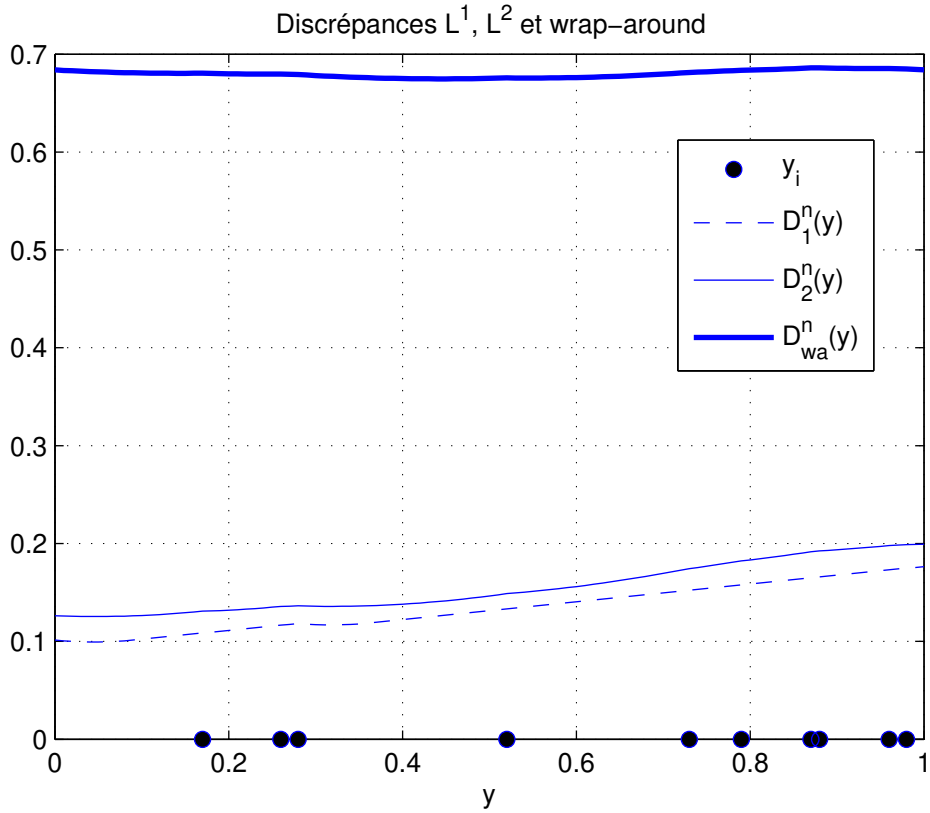

FIGURE 4.3 – Discrépance L^1 (en trait discontinu) et L^2 (en trait continu fin), et discrépance wrap-around (en trait continu fort). Les cercles pleins désignent les points du plan courant.

Observons la figure 4.3 où sont représentées les fonctions de discrépance L^1, L^2 et wrap-around. On remarque que les allures des fonctions $D_1^n(\cdot)$ et $D_2^n(\cdot)$ sont similaires, le minimum est situé proche de 0. L'allure de la fonction $D_{\text{wa}}^n(\cdot)$ est en revanche complètement différente, le minimum est situé proche de 0.44. Dans tous les cas, le critère est peu intuitif, ce qui confirme les remarques de [49] sur l'utilisation des fonctions de discrépance dans la construction de plans uniformes.

Nous ne retenons pas les fonctions de discrépance pour la suite de l'étude, en raison du caractère peu intuitif du critère d'ajout de point.

4.1.2 Étude d'une fonction de type maximin

Nous avons testé une fonction simple, inspiré du critère maximin utilisé pour construire des plans d'expériences (§ 3.1.3.1), dont l'idée intuitive est que le nouveau point y_{n+1} doit être le plus éloigné possible des points déjà présents $\{y_1, \dots, y_n\}$. L'échantillon $Y^n = \{y_1, \dots, y_n\} \in [0,1]$ étant donné, on considère, pour $y \in [0,1]$, la fonction de diversité

$$\mathrm{d}_{\text{Mm}}^n(y) = \min_{i=1\dots n} |y - y_i|.$$

Remarque 4.1.7 *On pourrait considérer également la fonction* $(\mathrm{d}_{\text{Mm}}^n(y))^l$*, pour un certain* l*.*

Proposition 4.1.8 *La fonction* $\mathrm{d}_{\text{Mm}}^n(\cdot)$ *est continue, et affine par morceaux de pentes* -1 *et* 1*. Son expression analytique s'obtient immédiatement (en supposant* $y_1 \leq y_2 \leq \dots \leq y_n$*),*

$$\mathrm{d}_{\text{Mm}}^n(y) = \begin{cases} -y + y_1, & \text{si } y \leq y_1\,; \\ y - y_i, & \text{si } y_i \leq y \leq \frac{y_i + y_{i+1}}{2}\,; \\ -y + y_{i+1}, & \text{si } \frac{y_i + y_{i+1}}{2} \leq y \leq y_{i+1}\,; \\ y - y_n, & \text{si } y \geq y_n. \end{cases}$$

Preuve *Triviale.* □

Sur la figure 4.4, où nous avons tracé la fonction de type maximin $\mathrm{d}_{\text{Mm}}^n(\cdot)$ pour un échantillon $\{y_1, \dots, y_n\}$ donné (représenté par les cercles pleins), nous pouvons constater que les maxima locaux de la fonction de diversité sont situés au milieu de deux points de l'échantillon, et au bord si aucun point de l'échantillon n'y est situé. Cette fonction de diversité invite donc à ajouter un point soit au milieu des deux points de l'échantillon les plus éloignés, soit au bord du domaine si aucun point de l'échantillon n'en est assez près.

Prise en compte de l'incertitude. Le critère de choix du nouveau point x est la moyenne de la fonction $\mathrm{d}_{\text{Mm}}^n(\cdot)$ relativement à la densité de la loi de la sortie $Y_n(x)$ donnée par le krigeage, que l'on cherche à maximiser, c'est-à-dire que la prochaine mesure se fera en

$$x^* = \operatorname*{argmax}_{x \in \mathcal{X}} \mathbb{E}\mathrm{d}_{\text{Mm}}^n(Y_n(x)).$$

L'espérance ci-dessus se calcule analytiquement lorsque la loi de la v.a. $Y_n(x)$ est normale de moyenne $\mu_n(x)$ et variance $\sigma_n^2(x)$. Elle s'obtient directement à partir de l'expression

$$\mathbb{E}\mathrm{d}_{\text{Mm}}^n(Y_n(x)) = \frac{1}{\sigma_n(x)\sqrt{2\pi}} \int_{-\infty}^{\infty} \mathrm{d}_{\text{Mm}}^n(y) \exp\left\{ -\frac{(y - \mu_n(x))^2}{2\sigma_n(x)^2} \right\} \mathrm{d}y, \tag{4.3}$$

en utilisant les formules donnés dans l'annexe F (la fonction d_{Mm}^n est un polynôme de degré 1). En pratique, $\mu_n(x)$ et $\sigma_n^2(x)$ sont respectivement la prédiction et la variance de krigeage au point x.

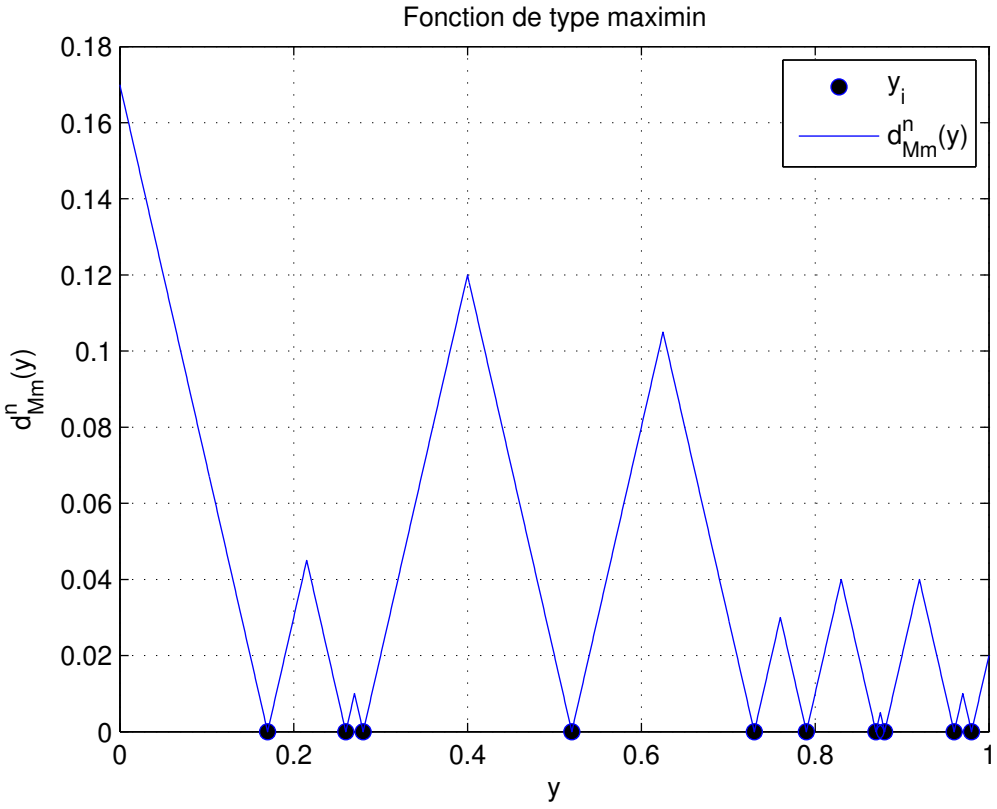

FIGURE 4.4 – Fonction de type maximin. Les cercles pleins désignent les points du plan courant.

Remarque 4.1.9 *En prenant $l = 2$ dans la remarque 4.1.7, on retrouve la* fonction de distorsion *utilisée pour la* quantification *de la loi normale [124]. Il pourrait être fructueux de s'intéresser aux méthodes d'optimisation de la fonction de distorsion afin d'essayer de les appliquer à notre problème.*

La fonction de diversité maximin, malgré son caractère « local » (sa valeur en un point est déterminée par seulement 1 ou 2 points de l'échantillon $\{y_1, \ldots, y_n\}$), semble tout-à-fait satisfaisante dans le cadre de notre étude.

4.1.3 Étude d'une fonction de type minimax

Reprenons les hypothèses et notations du paragraphe précédent. À partir de l'échantillon de points $Y^n = \{y_1, \ldots, y_n\}, 0 \leq y_1 \leq y_2 \leq \cdots \leq y_n \leq 1$, on cherche à ajouter un point $y_{n+1} \in [0, 1]$ de façon à maximiser la diversité de l'ensemble à $n + 1$ points $\{y_1, \ldots, y_n, y_{n+1}\}$. Une fonction inspirée directement du critère minimax (§ 3.1.3.2) serait

$$\mathrm{d}^{n,0}_{\mathrm{mM}}(y) = \max_{t \in [0,1]} \min_{z \in Y^n \cup \{y\}} |t - z|.$$

Comme on le voit sur la figure 4.5, la fonction n'est pas très intuitive ; de plus, les minima forment un palier, ce qui rend l'optimisation délicate. Considérons alors une version modifiée de la fonction précédente,

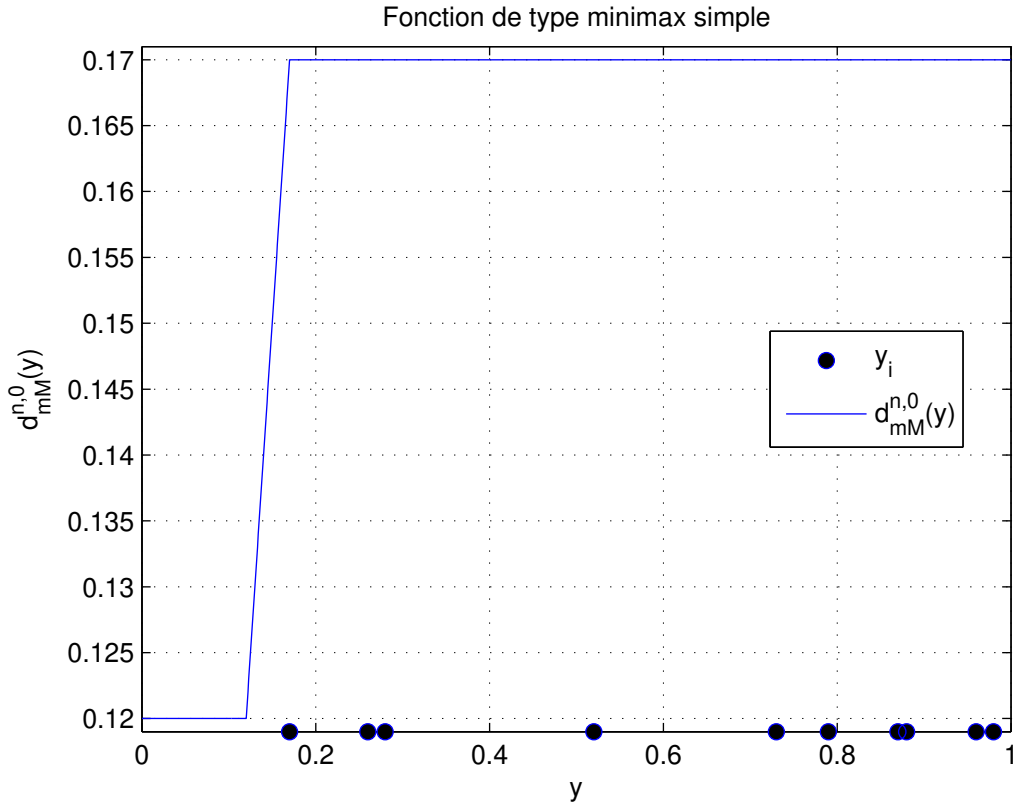

FIGURE 4.5 – Fonction diversité de type minimax simple. Les cercles pleins désignent les points du plan courant.

$$\mathrm{d}^n_{\mathrm{mM}}(y) = \int_0^1 \min_{z \in Y^n \cup \{y\}} |t - z| \mathrm{d}t,$$

que l'on cherche à minimiser de façon à ce que la moyenne des distances entre l'ensemble $[0,1] \setminus \{y_1, \ldots, y_n, y\}$ et l'ensemble $\{y_1, \ldots, y_n, y\}$ soit la plus petite possible.

Proposition 4.1.10 *La fonction* $\mathrm{d}^n_{\mathrm{mM}}(\cdot)$ *est un polynôme de degré* 2 *par morceaux, continu. Son expression analytique s'obtient à partir des formules suivantes,*

$$\begin{aligned}
\int_0^{y_1} \min_{z \in Y^n \cup \{y\}} |t - z| \mathrm{d}t &= \begin{cases} \frac{3}{4}y^2 - \frac{y_1}{2}y + \frac{{y_1}^2}{4}, & \textit{si } y \in [0, y_1]\,; \\ \frac{{y_1}^2}{2}, & \textit{sinon.} \end{cases} \\
\int_{y_i}^{y_{i+1}} \min_{z \in Y^n \cup \{y\}} |t - z| \mathrm{d}t &= \begin{cases} \frac{1}{2}y^2 - \frac{y_i + y_{i+1}}{2}y + \frac{{y_i}^2 + {y_{i+1}}^2}{4}, & \textit{si } y \in [y_i, y_{i+1}]\,; \\ \frac{(y_{i+1} - y_i)^2}{4}, & \textit{sinon.} \end{cases} \\
\int_{y_n}^1 \min_{z \in Y^n \cup \{y\}} |t - z| \mathrm{d}t &= \begin{cases} \frac{3}{4}y^2 - \frac{y_n + 2}{2}y + \frac{{y_n}^2 + 2}{4}, & \textit{si } y \in [y_n, 1]\,; \\ \frac{(1 - y_n)^2}{2}, & \textit{sinon.} \end{cases}
\end{aligned}$$

Preuve *Immédiate.* □

Voyons la figure 4.6, où nous avons représenté le critère de type minimax pour le même échantillon que précédemment représenté en cercles pleins. L'allure générale de cette fonction est (au signe

près) la même que la fonction de type maximin du § 4.1.2, sauf au bord du domaine, auquel il est donné ici moins d'importance que dans le cas de l'autre critère (comparer la figure 4.4 et la figure 4.6). Le nouveau point y_{n+1} choisi par le critère de diversité minimax est ici proche de 0.1, alors que c'était le point 0 pour le critère de type maximin.

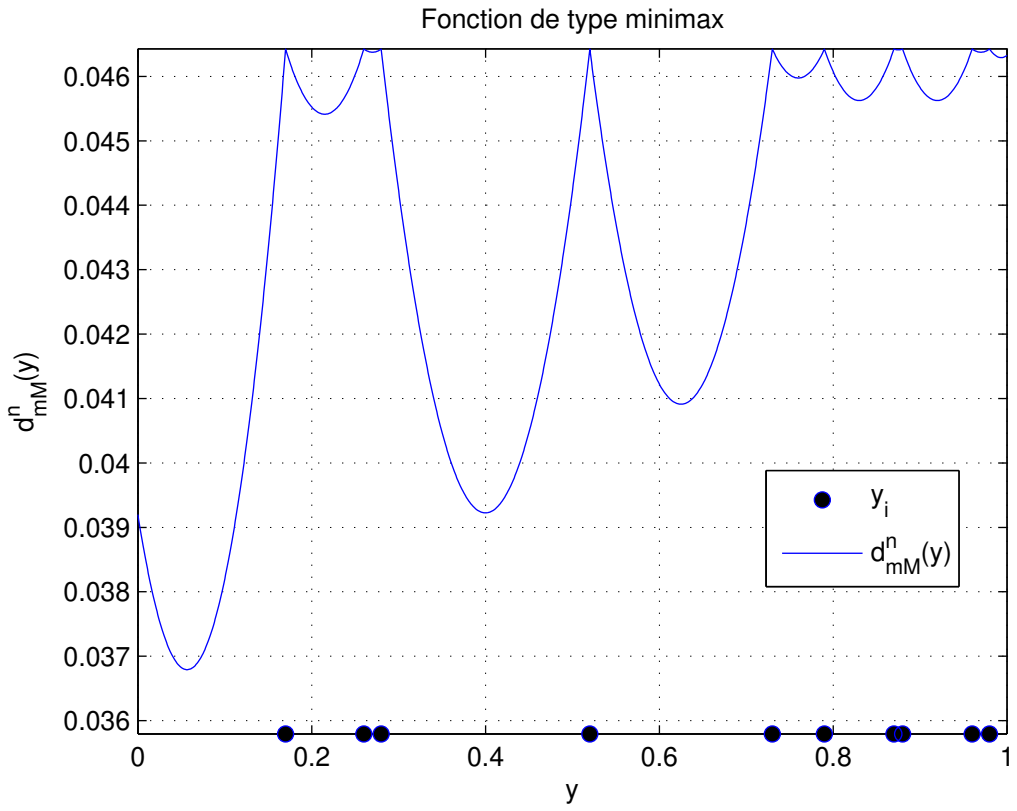

FIGURE 4.6 – Fonction diversité de type minimax. Les cercles pleins désignent les points du plan courant.

L'expression analytique de l'espérance $\mathbb{E}\mathrm{d}^n_{\mathrm{mM}}(Y_n(x))$ se calcule très facilement quand $Y_n(x) \sim \mathcal{N}(\mu_n(x), \sigma^2_n(x))$, en utilisant les formules données dans l'annexe F.

La fonction minimax est légèrement plus compliquée que la fonction maximin. De plus, elle donne moins d'importance aux points du bord. Or nous sommes très intéressés par la position (inconnue) des bords du domaine atteignable par les réponses. Nous ne retenons donc pas le critère de type minimax pour la suite.

Nous passons maintenant à l'étude de fonctions de diversité utilisant l'entropie.

4.1.4 Fonctions utilisant l'entropie de Shannon

Le concept d'*entropie* a été introduit à l'origine en thermodynamique par Clausius (1865) et Boltzmann (1870). Il s'agit d'une grandeur mesurant le désordre d'un système. Un système isolé tend vers un état d'équilibre, et son entropie augmente au fur et à mesure. Quand le système atteint l'état d'équilibre, l'entropie est maximale et le système est complètement désordonné, c'est-à-dire qu'il est dans un des états qu'il peut atteindre avec équiprobabilité : l'expérimentateur est donc dans la plus grande incertitude sur l'état final du système. L'entropie de Clausius-

Boltzmann est *extensive* (ou *additive*) : si deux systèmes A et B sont indépendants, alors $\text{ent}(A \cup B) = \text{ent}(A) + \text{ent}(B)$; dans le cas général, $\text{ent}(A \cup B) \leq \text{ent}(A) + \text{ent}(B)$.

En théorie de l'information, l'entropie a été introduite par Shannon en 1948 [153]. Intuitivement, l'entropie de Shannon mesure la quantité d'incertitude liée à un évènement aléatoire : un évènement dont la loi est uniforme présente la plus grande incertitude, son entropie est donc maximale. À l'inverse, si la masse de la densité de probabilité est très concentrée, l'incertitude est très petite et l'entropie est minimale. Nous allons appliquer cette mesure de la quantité de hasard à notre problème de façon séquentielle : nous voulons ajouter un point y_{n+1} à un ensemble existant $Y^n = \{y_1, \ldots, y_n\}$, de sorte que l'ensemble obtenu à $n+1$ points soit de répartition la plus uniforme possible.

Définition 4.1.11 *Soit $\phi(\cdot)$ une densité de probabilité. On appelle* entropie de Shannon *de ϕ, notée $H(\phi)$, la quantité*

$$H(\phi) = -\int \phi \log \phi.$$

Par analogie, l'entropie d'une variable aléatoire X dont la loi a pour densité $\phi(\cdot)$ est $H(X) = H(\phi)$. L'entropie de Shannon est extensive : si X et Y sont deux v.a. indépendantes, alors $H(X,Y) = H(X) + H(Y)$.

Proposition 4.1.12 *Si $\phi(\cdot)$ est à support compact $\mathbb{K}$, le maximum de la fonctionnelle d'entropie $H(\cdot)$ est obtenu pour la loi uniforme sur $\mathbb{K}$.*

L'idée reposant sur cette propriété est de placer le nouveau point y_{n+1} de façon à maximiser l'entropie de la loi ayant généré l'échantillon $\{y_1, \ldots, y_n, y_{n+1}\}$. La solution proposée est d'utiliser un estimateur de la densité de cette loi. Notons que cette façon de procéder est en contradiction avec les hypothèses du krigeage, où l'on suppose que le vecteur aléatoire $(Y(x_1), \ldots, Y(x_n), Y(x_{n+1}))$ suit une loi normale multivariée (l'ensemble $\{y_1, \ldots, y_n, y_{n+1}\}$ n'est donc pas un échantillon i.i.d.), mais elle permet de construire un critère de diversité cohérent.

Il existe de nombreuses méthodes permettant d'estimer l'entropie d'une loi de probabilité à partir d'un échantillon, voir [19, 64] (et aussi [95], si l'entropie fait intervenir le carré de la densité). Nous en avons testé deux : une méthode par substitution avec estimation de la densité par une méthode à noyaux [48, 155, 184] et la méthode des plus proches voisins [90, 97]. Enfin, nous avons utilisé une méthode de calcul de divergence proposée par M. Broniatowski [23] pouvant servir à maximiser l'entropie.

4.1.4.1 Estimateurs à noyaux

Commençons par présenter la méthode d'estimation d'une densité par des noyaux, qui sera utilisée aussi pour les autres fonctions d'entropie. Supposons que l'on dispose d'un échantillon aléatoire $Y_1, \ldots, Y_n$, issu d'une densité $\phi(\cdot)$, continue et monodimensionnelle.

Définition 4.1.13 *L'estimateur à noyaux de la densité $\phi(\cdot)$ s'écrit*

$$\widehat{\phi}_n(y,h) = \frac{1}{nh} \sum_{i=1}^{n} k\left(\frac{y - Y_i}{h}\right).$$

La fonction $k(\cdot)$, appelée noyau (kernel)*, vérifie $\int_{\mathbb{R}} k(x)\mathrm{d}x = 1$, et h est un nombre positif, appelé* pas *ou* largeur de fenêtre (bandwidth)*.*

Un exemple de noyau d'utilisation commune est le noyau gaussien

$$k(x) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{x^2}{2}}, \tag{4.4}$$

la densité de la loi normale $\mathcal{N}(0,1)$ (les noyaux utilisés en pratique sont généralement des densités de probabilité unimodales et symétriques autour de 0). Les centres des noyaux sont donc situés aux Y_i, qui ont chacun une influence sur leur voisinage, puis la contribution de chaque Y_i est sommée pour obtenir l'estimation globale.

Remarque 4.1.14 *Le domaine atteignable en sortie $\mathcal{Y}$ étant borné, le choix du noyau gaussien, qui n'est pas à support compact, peut sembler surprenant. Cependant, lorsque le pas est relativement petit, la densité gaussienne tend rapidement vers* 0 *quand on s'éloigne des points de l'échantillon. De plus, les calculs d'entropie sont facilités, comme nous le verrons dans la suite.*

La forme du noyau n'est pas tellement importante (du moment qu'il vérifie certaines conditions de régularité), c'est surtout le choix du pas qui est déterminant : un pas trop petit donnera une estimation peu biaisée, mais avec une grande variance (*i.e.* une estimation assez fidèle, mais très bruitée), alors qu'une grande valeur de pas donnera une estimation trop lisse qui ne tiendra pas suffisamment compte de la forme de la fonction que l'on cherche à estimer (une petite variance, mais un grand biais). Afin d'établir un compromis biais-variance, on considère le critère *erreur quadratique intégrale moyenne (Mean Integrated Squared Error, MISE)*,

$$\mathrm{MISE}\left\{\widehat{\phi}(\cdot,h)\right\} = \mathbb{E}\int \left\{\widehat{\phi}(x,h) - \phi(x)\right\}^2 \mathrm{d}x.$$

Sous des hypothèses raisonnables, on peut obtenir une expression asymptotique de l'MISE.

Proposition 4.1.15 *[184] Supposons que :*

- *la densité $\phi(\cdot)$ est telle que sa dérivée seconde ϕ'' est continue, de carré intégrable, et monotone à l'infini (*i.e. *monotone sur $]-\infty,-M[\cup]M,\infty[$, pour un certain $M>0$) ;*
- *la largeur de fenêtre dépend de n, et la suite des pas notée h_n est une suite de nombres non aléatoire. De plus, $\lim_{n\to\infty} h_n = 0$ et $\lim_{n\to\infty} nh_n = \infty$;*
- *le noyau $k(\cdot)$ est une densité de probabilité symétrique par rapport à l'origine, et ayant un moment d'ordre 4 fini.*

Alors,

$$\mathrm{MISE}\left\{\widehat{\phi}(\cdot,h_n)\right\} = \frac{1}{nh_n}R(k) + \frac{1}{4}h_n^4\mu_2(k)R(\phi'') + o\left(\frac{1}{nh_n} + h_n^4\right),$$

avec $R(g) = \int g(x)^2\mathrm{d}x$ pour toute fonction g de carré intégrable, et $\mu_2(k) = \int x^2k(x)\mathrm{d}x$.

Définition 4.1.16 *La quantité*

$$\mathrm{AMISE}\left\{\widehat{\phi}(\cdot,h)\right\} = \frac{1}{nh}R(k) + \frac{1}{4}h^4\mu_2(k)R(\phi'') \tag{4.5}$$

est appelée MISE asymptotique (Asymptotic MISE).

Le pas optimal, relativement au critère AMISE, est obtenu en minimisant (4.5). En cherchant les zéros de la dérivée, on obtient

$$h_{\mathrm{AMISE}} = \left[\frac{R(k)}{\mu_2(k)^2R(\phi'')n}\right]^{\frac{1}{5}}.$$

La valeur correspondante de l'AMISE est

$$\inf_{h>0} \text{AMISE}\left\{\widehat{\phi}(\cdot,h)\right\} = \frac{5}{4n^{\frac{4}{5}}}\left\{\mu_2(k)^2 R(k)^4 R(\phi'')\right\}^{\frac{1}{5}}.$$

On peut donc écrire

$$h_{\text{MISE}} \sim \left[\frac{R(k)}{\mu_2(k)^2 R(\phi'')n}\right]^{\frac{1}{5}}$$

et

$$\inf_{h>0} \text{MISE}\left\{\widehat{\phi}(\cdot,h)\right\} \sim \frac{5}{4n^{\frac{4}{5}}}\left\{\mu_2(k)^2 R(k)^4 R(\phi'')\right\}^{\frac{1}{5}},$$

ce qui donne une vitesse de convergence de l'ordre de $n^{-\frac{4}{5}}$ pour les estimateurs à noyaux. En pratique $R(\phi'')$ est inconnu, mais il existe des méthodes permettant d'obtenir h à partir d'une estimation de $R(\phi'')$. Ne disposant que de petits échantillons, nous nous contentons de donner deux méthodes simples, utilisant un pas constant (*constant bandwidth*).

– *La règle de mise à l'échelle normale (Normal Scale Rule) [184].* Dans le cas où $\phi(\cdot)$ est la densité d'une loi normale de variance σ^2, on peut montrer que

$$h_{\text{AMISE}} = \sigma\left[\frac{8\pi^{\frac{1}{2}}R(k)}{3\mu_2(k)^2 n}\right]^{\frac{1}{5}}.$$

On peut donc prendre h de la forme

$$\widehat{h}_{\text{NS}} = \widehat{\sigma}\left[\frac{8\pi^{\frac{1}{2}}R(k)}{3\mu_2(k)^2 n}\right]^{\frac{1}{5}},$$

où $\widehat{\sigma}$ est une estimation de σ. L'utilisation de la loi gaussienne (alors que la loi est inconnue) permet d'obtenir un ordre de grandeur du pas.

– *Le principe de lissage maximal (Maximal Smoothing Principle) [184].* Cette méthode se base sur une majoration de h_{AMISE}. En effet, pour toute densité ayant un écart-type σ,

$$h_{\text{AMISE}} \leq \sigma\left[\frac{243R(k)}{35\mu_2(k)^2 n}\right]^{\frac{1}{5}}.$$

Une façon de choisir h qui surestime la régularité de la densité est donc donnée par

$$\widehat{h}_{\text{OS}} = \widehat{\sigma}\left[\frac{243R(k)}{35\mu_2(k)^2 n}\right]^{\frac{1}{5}}.$$

Le pas $\widehat{h}_{\text{OS}}$ est trop grand pour une estimation optimale de la densité, mais permet une détermination empirique de la bonne valeur du pas : on trace tout d'abord l'estimation de la densité obtenue pour $\widehat{h}_{\text{OS}}$, puis successivement celles obtenues pour des fractions de $\widehat{h}_{OS}$, afin de se faire une idée des caractéristiques de la densité, comme, par exemple, le nombre de modes.

Il existe une très vaste littérature sur les méthodes permettant d'estimer le pas h voir par exemple [41, 48, 155, 184]. Nous avons testé la règle de mise à l'échelle normale. Si le noyau $k(\cdot)$ est gaussien (4.4), on montre facilement que $R(k) = 1/(2\sqrt{\pi})$ et $\mu_2(k) = 1$, donc

$$\widehat{h}_{\mathrm{NS}} = \widehat{\sigma}\left(\frac{4}{3n}\right)^{\frac{1}{5}}, \tag{4.6}$$

avec $\widehat{\sigma}$ une estimation du paramètre σ.

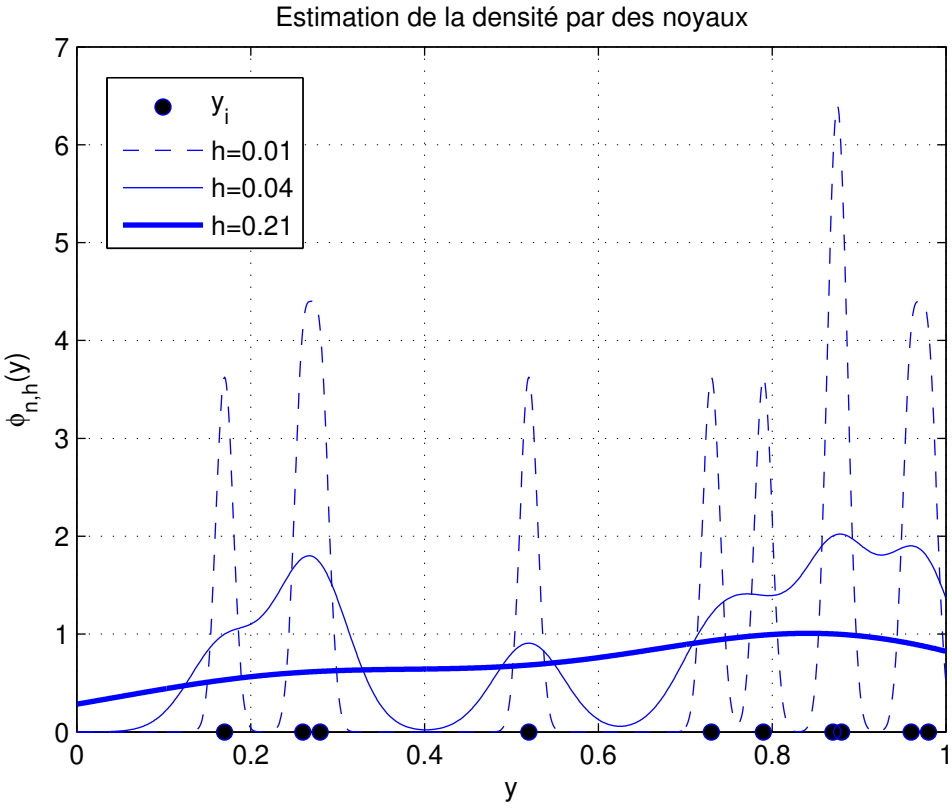

FIGURE 4.7 – Estimation de la densité par des noyaux, pour trois valeurs du pas h. Les cercles pleins désignent les points du plan courant.

La figure 4.7 présente l'estimation par noyau gaussien de la densité de la loi ayant généré les y_i (représentés en cercles pleins), pour trois valeurs du pas h. Une petite valeur du pas $h = 0.01$ donne une densité dont les pics sont situés aux y_i. Une grande valeur du pas $h = \widehat{h}_{\mathrm{NS}} \approx 0.21$ (σ étant estimé par l'écart-type empirique des y_i) donne une densité très plate. Nous avons testé une valeur intermédiaire définie de façon empirique à partir des *espacements* [39],

$$h_{\mathrm{emp}} = \max_{i=1,\ldots,n-1} \frac{y_{(i+1)} - y_{(i)}}{6}, \tag{4.7}$$

où les $y_{(i)}$ sont rangés selon les valeurs croissantes des y_i. L'idée est qu'en choisissant h de cette façon, la densité tombe proche de 0 entre les deux points $y_{(i)}$ les plus éloignés, à environ 3 écarts-types du mode (ici $h_{\mathrm{emp}} = 0.04$).

Une fois la densité estimée, on cherche à placer le nouveau point y_{n+1} de façon à maximiser l'estimation de l'entropie

$$\widehat{H}_n(y_{n+1}) = -\int_{\mathbb{R}} \widehat{\phi}_{n+1}(t) \log \widehat{\phi}_{n+1}(t) \mathrm{d}t, \tag{4.8}$$

où l'estimation $\widehat{\phi}_{n+1}(\cdot)$ est construite avec l'échantillon $\{y_1, \ldots, y_n, y_{n+1}\}$.

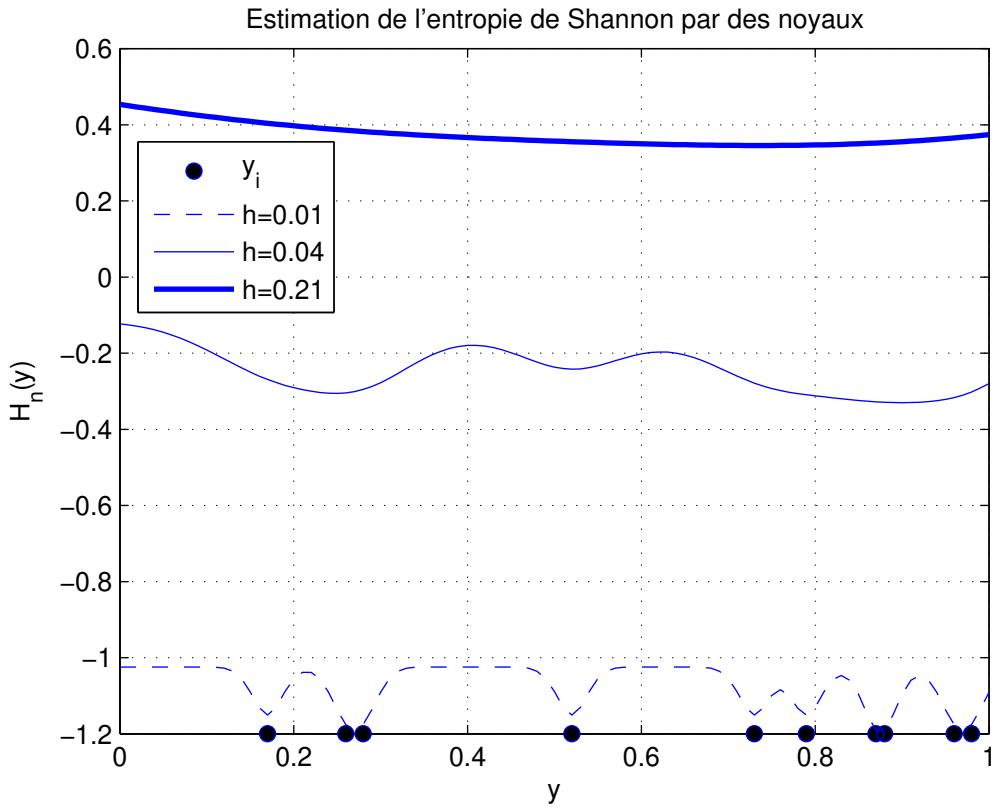

FIGURE 4.8 – Estimation de l'entropie de Shannon par des noyaux, selon la largeur du pas. Les cercles pleins désignent les points du plan courant.

Comparons les estimations des valeurs d'entropie obtenues en prenant les mêmes pas h que sur la figure 4.7, en faisant varier le nouveau point $y = y_{n+1}$. Dans tous les cas, le maximum est atteint au nouveau point $y = 0$, ce qui est raisonnable car le y_i le plus proche est éloigné : on observe le même résultat que sur la figure 4.4 pour la fonction de type maximin. Une valeur de pas h trop petite pourra avoir tendance à résulter en un critère de diversité plat quant on s'éloigne des y_i (courbe en pointillés), de plus certaines valeurs de $\widehat{\phi}_{n+1}$ proches de 0 poseront problème dans l'évaluation du logarithme (ceci étant dû au fait que la gaussienne tend très vite vers 0 quand on s'éloigne de la moyenne). La grande valeur de pas $\widehat{h}_{\mathrm{NS}} \approx 0.21$, donnée par la règle de mise à l'échelle normale, donne un critère peu intuitif (courbe en trait fort). La valeur empirique intermédiaire $h = 0.04$, calculée avec la formule (4.7), semble donner le critère le plus réaliste, où les maxima locaux sont situés au milieu des y_i les plus éloignés, et le critère indique qu'il n'est pas intéressant de placer le nouveau point y_{n+1} aux endroits de forte présence des y_i.

Cette méthode requiert l'évaluation numérique des intégrales du type (4.8), afin de calculer l'estimation « plug-in » de l'entropie pour toutes les valeurs y que l'on veut tester, ce qui peut

s'avérer contraignant en temps de calcul.

Le critère de sélection du nouveau point x est donc la moyenne de (4.8) par rapport à la densité de la loi de la sortie $Y_n(x)$, $\mathbb{E}\widehat{H}_n(Y_n(x))$, qui doit être évaluée numériquement, en simulant des réalisations $z_1, \ldots, z_l$ de la v.a. $Y_n(x)$ et prenant la moyenne

$$\widehat{\mathbb{E}}\widehat{H}_n(Y_n(x)) = \frac{1}{l}\sum_{i=1}^{l} \widehat{H}_n(z_i).$$

La valeur sélectionnée pour la prochaine mesure est finalement

$$x^* = \underset{x\in\mathcal{X}}{\operatorname{argmax}}\, \widehat{\mathbb{E}}\widehat{H}_n(Y_n(x)).$$

Nous retenons la méthode d'estimation de l'entropie par des noyaux, mais nous allons l'appliquer (au paragraphe 4.1.6) à une fonction d'entropie permettant d'obtenir une formule analytique, car pour l'entropie de Shannon l'évaluation de la fonction diversité $H_n(Y_n(x))$, puis de sa moyenne, sont coûteuses en temps de calcul.

4.1.4.2 Estimation de l'entropie par plus proches voisins

Cette méthode consiste à estimer l'entropie $H(\phi_{n+1})$ où $\phi_{n+1}(\cdot)$ est la densité construite à partir de l'échantillon $\{y_1, \ldots, y_n, y_{n+1}\}$ par l'expression [90, 97]

$$\widehat{H}_n(y) = \frac{1}{n+1}\sum_{i=1}^{n+1} \log \zeta_{n+1,i,k}, \tag{4.9}$$

où

- $\zeta_{n+1,i,k} = n\mathrm{e}^{-\psi(k)} V_q \left[\rho_{k,n}^{(i)}\right]^q$;
- q est la dimension de l'espace des sorties (ici $q = 1$) ;
- V_q est le volume de la boule unité dans l'espace des sorties ($V_1 = 2$) ;
- $\rho_{k,n}^{(i)}$ est la distance de y_i à son k^e plus proche voisin ;
- $\psi(x) = \dfrac{\Gamma'(x)}{\Gamma(x)}$, où $\Gamma(\cdot)$ est la fonction gamma d'Euler $\Gamma(x) = \int_0^\infty t^x e^{-t}\mathrm{d}t$.

Proposition 4.1.17 *[97] Si la densité ϕ est bornée, et si $I_l = \int \phi^l(x)\mathrm{d}x$ existe pour un certain $l < 1$, alors l'entropie $H(\phi)$ existe, $\mathbb{E}\widehat{H}_n \underset{n\to\infty}{\longrightarrow} H$ et $\widehat{H}_n \underset{n\to\infty}{\overset{L^2}{\longrightarrow}} H$.*

Remarque 4.1.18

- *Pour n fixé, lorsque $y = y_{n+1}$ varie, l'estimateur de l'entropie est de la forme $a + b\sum \log \rho_{k,n}^{(i)}$, avec $a \geq 0$ et $b \geq 0$. Son argument maximum sera le même que celui de la fonction $\prod \rho_{k,n}^{(i)}$, ce qui suggère que les points sont bien répartis si ils sont le plus loin possible de leur k^e plus proche voisin (car de cette façon $\widehat{H}_n$ sera maximal) ;*
- *l'intérêt de la méthode des plus proches voisins, par rapport aux estimateurs à noyaux, est qu'il n'y a pas à faire le choix du pas h. La méthode est aussi utilisable directement en dimension supérieure à 1.*

Observons la figure 4.9, où est représentée l'estimation de l'entropie de Shannon par la méthode du 1$^{\text{er}}$ plus proche voisin en fonction des valeurs de $y_{n+1} = y$. On remarque que prendre $k = 1$ (1er plus proche voisin) donne des résultats satisfaisants, car le comportement du critère d'entropie

en fonction de y est similaire à la fonction de type maximin du § 4.1.2 : les maxima locaux sont situés entre les y_i, et le maximum global est en 0. La seule précaution à prendre en pratique est de ne pas évaluer $\widehat{H}_n$ trop près des y_i, à cause des termes en $\log \rho_{1,n}^{(i)}$ dans la formule (4.9) qui tendent vers $-\infty$.

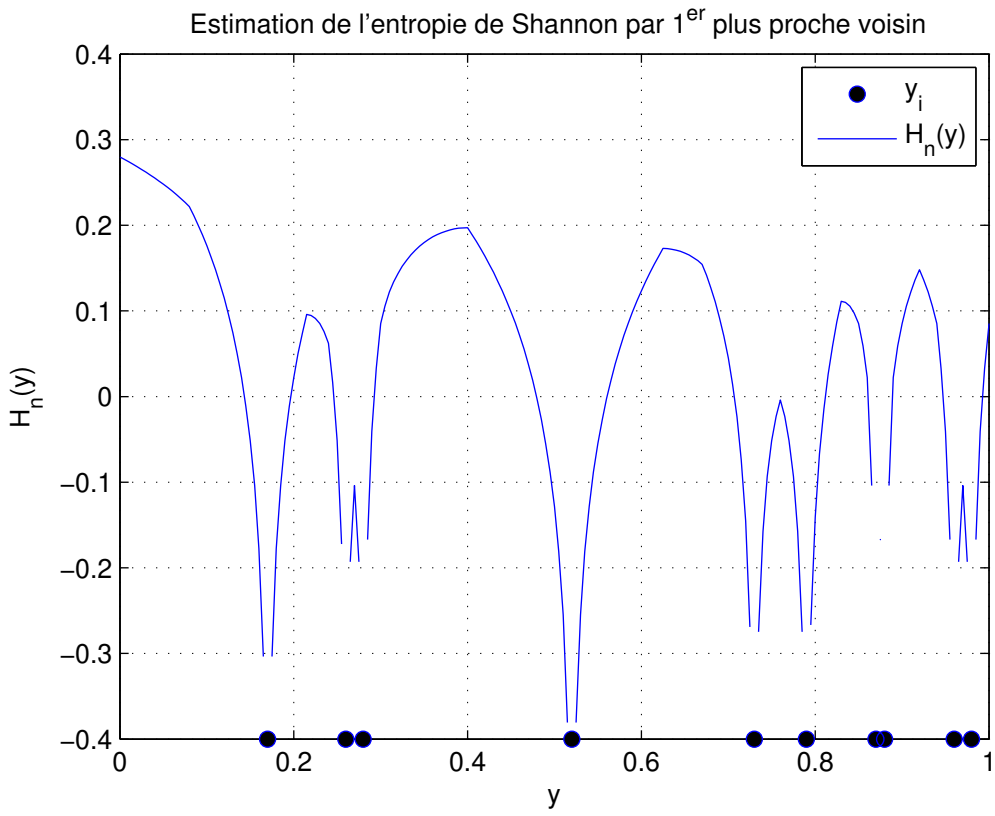

FIGURE 4.9 – Estimation de l'entropie de Shannon par la méthode du 1er plus proche voisin. Les cercles pleins désignent les points du plan courant.

Nous venons de voir que prendre $k = 1$ peut poser un problème si on évalue $\widehat{H}_n$ en l'un des y_i, à cause du terme $\log 0$. Nous avons donc envisagé la possibilité de prendre $k = 2$ (2e plus proche voisin). Sur la figure 4.10, nous constatons que le maximum global est toujours en 0, mais le critère est moins intuitif que pour $k = 1$, notamment en raison du maximum local en 0.51 qui est situé très près d'un y_i. Ce phénomène est compréhensible, car exiger qu'un point soit le plus éloigné possible de son 2e plus proche voisin ne garantit pas qu'il soit éloigné de son 1^{er} plus proche voisin. Le critère ne convient donc pas.

Pour estimer l'espérance par rapport à la loi de la sortie $Y_n(x)$, $\mathbb{E}\,\widehat{H}_n(Y_n(x))$, on simule des réalisations de la variable aléatoire $\widehat{H}_n(Y_n(x))$, avec $Y_n(x) \sim \mathcal{N}(\mu_n(x), \sigma_n^2(x))$ (les paramètres de la loi normale donnés par le krigeage), puis on prend la moyenne de ces réalisations comme au paragraphe précédent. Ceci est coûteux en temps de calcul, et donne lieu à une fonction bruitée.

L'estimation de l'entropie par 1er plus proche voisin donne des résultats satisfaisants, mais est coûteuse en temps de calcul lorsqu'on évalue l'espérance du critère par rapport à la loi de la v.a. $Y_n(x)$, ce qui rend la méthode inutilisable pour notre étude. Nous ne retenons pas non plus la méthode du 2e plus proche voisin, en raison du caractère peu intuitif du critère.

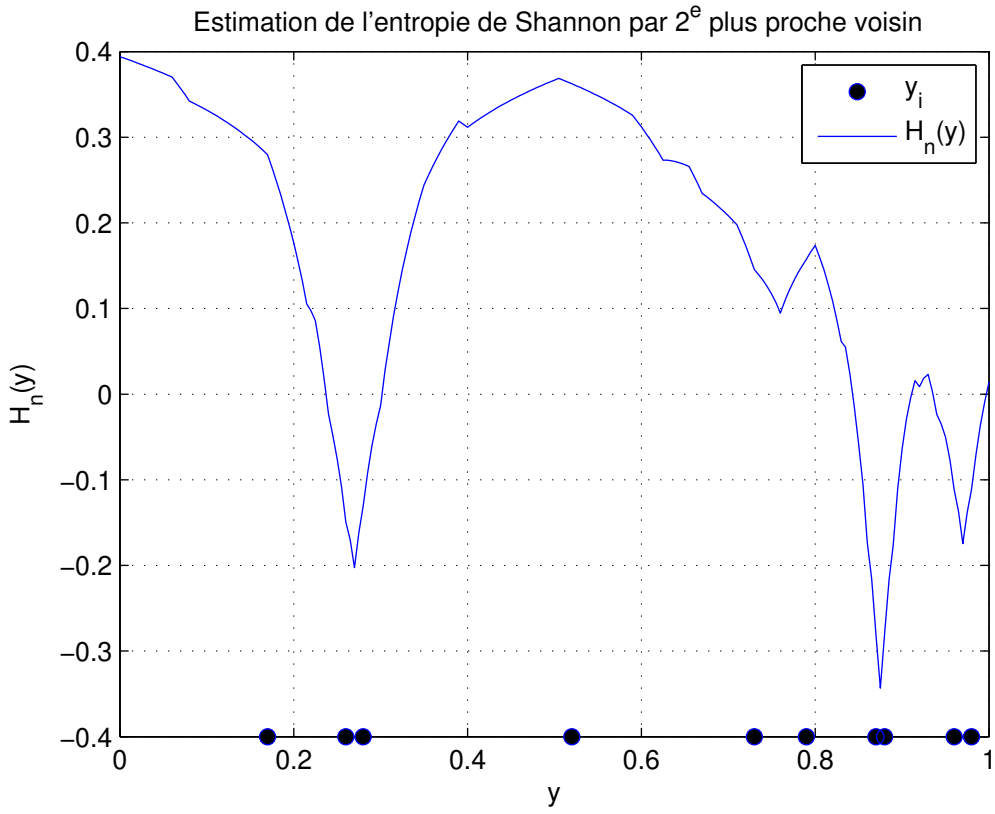

FIGURE 4.10 – Estimation de l'entropie de Shannon par la méthode du 2^{e} plus proche voisin. Les cercles pleins désignent les points du plan courant.

4.1.4.3 Variation d'entropie et divergence

Rappelons que l'on dispose d'un échantillon $Y^n = \{y_1, \ldots, y_n\}$ issu d'une loi inconnue de densité ψ, et d'un point candidat y_{n+1} qui est en fait une variable aléatoire $Y_n(x) \sim \mathcal{N}(\mu_n(x), \sigma_n^2(x))$ dont la densité est notée $\phi_{n,x}$. Faisons l'hypothèse que l'échantillon (Y^n, y_{n+1}) est issu de la loi de mélange de densité

$$(1-\alpha)\psi + \alpha\phi_{n,x}, \tag{4.10}$$

avec $\alpha = 1/(n+1)$. On cherchera à maximiser l'entropie de cette loi de mélange. Soit la fonction

$$h(\alpha) = H[(1-\alpha)\psi + \alpha\phi_{n,x}],$$

avec $H(\phi) = -\int \phi \log \phi$ l'entropie de Shannon. On a :

$$\begin{aligned}
\frac{\mathrm{d}h(\alpha)}{\mathrm{d}\alpha} &= -\int (\phi_{n,x} - \psi) \log\left[(1-\alpha)\psi + \alpha\phi_{n,x}\right]; \\
\frac{\mathrm{d}^2 h(\alpha)}{\mathrm{d}\alpha^2} &= -\int \frac{(\phi_{n,x} - \psi)^2}{(1-\alpha)\psi + \alpha\phi_{n,x}} < 0 \quad (\text{donc } h(\cdot) \text{ est concave}); \\
\frac{\mathrm{d}h(\alpha)}{\mathrm{d}\alpha}\Big|_0 &= -\int (\phi_{n,x} - \psi) \log \psi.
\end{aligned}$$

L'idée est de maximiser le terme $-\int (\phi_{n,x} - \psi) \log \psi$, car α étant petit, une grande valeur de $\mathrm{d}h(\alpha)/\mathrm{d}\alpha\big|_0$ devrait donner une grande valeur de $h(\alpha)$ (par un développement limité à l'ordre 1). On peut alors faire intervenir la *divergence* des lois $\phi_{n,x}$ et ψ, $\mathcal{K}(\phi_{n,x}, \psi)$ (définition 2.3.5). La dérivée de $h(\cdot)$ en 0 se réécrit

$$\frac{\mathrm{d}h(\alpha)}{\mathrm{d}\alpha}\Big|_0 = \mathcal{K}(\phi_{n,x}, \psi) + H(\phi_{n,x}) - H(\psi),$$

que l'on cherche à maximiser par rapport à x. L'entropie d'une loi gaussienne est connue et vaut [161]

$$H(\phi_{n,x}) = \frac{1}{2} + \frac{1}{2}\log(2\pi) + \log \sigma_n(x).$$

L'entropie $H(\psi)$ intervient comme une constante dans la fonction que l'on souhaite maximiser par rapport à x. Afin d'estimer $\mathcal{K}(\phi_{n,x}, \psi)$, on utilise une méthode proposée par M. Broniatowski [23] (un calcul direct est impossible, car la densité ψ est inconnue : seule est connue la mesure empirique P_n liée à l'échantillon $\{y_1, \ldots, y_n\}$, et $\mathcal{K}(\phi_{n,x}, P_n) = \infty \ \forall x$). M. Broniatowski propose une méthode pour estimer $\mathcal{K}(\phi_{n,x}, \psi)$ à partir de la mesure empirique, que nous présentons maintenant. Soit

$$\Omega_x = \left\{ Q \text{ mesure de probabilité}, \mathbb{E}\, Q = \mu_n(x), \operatorname{var} Q = \sigma_n^2(x) \right\}.$$

On sait calculer

$$Q^* = \operatorname*{arginf}_{Q \in \Omega_x} \mathcal{K}(Q, P_n).$$

En effet, Q^* est une mesure discrète de même support que P_n, c'est-à-dire que ses poids sont concentrés en $y_1, \ldots, y_n$. Pour $i = 1, \ldots, n$, les poids de la mesure Q^* sont donnés par

$$q_i^* = Q^*(y_i) = p_i \exp\left\{ a_0 + a_1(y_i - \mu_n(x)) + a_2\left[(y_i - \mu_n(x))^2 - \sigma_n^2(x)\right] \right\},$$

où $p_i = 1/n, i = 1, \ldots, n$, sont les poids de la mesure empirique P_n. Les coefficients a_0, a_1 et a_2 sont obtenus à partir des contraintes suivantes, conséquences de la définition de Ω_x.

- $\displaystyle\sum_{i=1}^{n} q_i^* = 1\,;$
- $\displaystyle\sum_{i=1}^{n} q_i^*(y_i - \mu_n(x)) = 0\,;$
- $\displaystyle\sum_{i=1}^{n} q_i^*(y_i - \mu_n(x))^2 = \sigma_n^2(x).$

Finalement, $\mathcal{K}(Q^*, P_n) = \sum_i q_i^* \log(q_i^*/p_i)$ estime $\mathcal{K}(\phi_{n,x}, \psi)$. Le problème de cette méthode est que le système d'équations ci-dessus n'a pas toujours de solution (imaginer le cas où la moyenne prédite $\mu_n(x)$ serait plus petite que tous les y_i). Pour les cas que nous avons testés, le système avait en fait très peu souvent de solution. La fonction à maximiser n'était alors pratiquement définie nulle part.

Remarque 4.1.19 *Il est possible d'utiliser la divergence du χ^2 [85] à la place de la divergence de Kullback-Leibler. On obtient alors un système d'équations linéaires pour a_0, a_1, a_2. Cependant dans de nombreux cas testés, les poids q_i^* obtenus étaient négatifs.*

Ces méthodes d'estimation de la divergence semblent donc inutilisables pour notre problème.

Les méthodes utilisant l'entropie de Shannon, estimée par des noyaux ou par la méthode du premier plus proche voisin, donnent des résultats satisfaisants, excepté en ce qui concerne le temps de calcul. C'est pourquoi nous décidons de tester d'autres fonctions d'entropie.

4.1.5 Entropie de Rényi

Les résultats obtenus par l'entropie de Shannon ne s'étant pas révélés tout-à-fait satisfaisants pour notre problème, en raison du coût de calcul élevé, nous avons testé une autre fonction d'entropie, appelée *entropie de Rényi.*

Définition 4.1.20 *[97] Soit $\phi(\cdot)$ une densité de probabilité. On appelle* entropie de Rényi d'ordre *l ($l \neq 1$) de ϕ, et on note $H_l^*(\phi)$, la quantité*

$$\begin{aligned} H_l^*(\phi) &= \frac{1}{1-l} \log \int \phi^l \\ &= \frac{1}{1-l} \log \|\phi\|_l^l. \end{aligned}$$

Proposition 4.1.21 *[97] Lorsque $l \to 1$, l'entropie de Rényi $H_l^*(\phi)$ tend vers l'entropie de Shannon $H_1(\phi) = -\int \phi \log \phi$.*

Nous nous sommes en particulier intéressés à l'entropie de Rényi d'ordre 2,

$$\begin{aligned} H_2^*(\phi) &= -\log \int \phi^2 \\ &= -\log \|\phi\|_2^2, \end{aligned}$$

qui permet de simplifier les calculs lorsque $\phi_{n,y_{n+1}}(\cdot)$ est estimée par des noyaux gaussiens, suivant

$$\phi_{n,y_{n+1}}(y) = \frac{1}{n+1} \sum_{i=1}^{n+1} \frac{1}{h\sqrt{2\pi}} \exp\left\{ -\frac{(y-y_i)^2}{2h^2} \right\}. \tag{4.11}$$

Proposition 4.1.22 *Si $\phi_{n,y_{n+1}}(\cdot)$ est la densité (4.11), alors*

$$\left\|\phi_{n,y_{n+1}}\right\|_2^2 = \frac{1}{\sqrt{2\pi}\sqrt{2h^2}(n+1)^2} \sum_{i,j=1}^{n+1} \exp\left\{ -\frac{(y_i-y_j)^2}{4h^2} \right\}.$$

Preuve *Donnée dans le cas général multidimensionnel, voir la proposition 4.2.2 page 118.* □

On cherchera y_{n+1} minimisant cette quantité, de façon à maximiser l'entropie de Rényi. On remarque que par rapport à l'entropie de Shannon, l'entropie de Rényi d'ordre 2 a une expression analytique (mais nous verrons que sa moyenne par rapport à la loi de la sortie $Y_n(x)$ ne s'écrit pas sous forme analytique).

Comparons la figure 4.11, où est tracée l'entropie de Rényi de la densité de l'échantillon $\{y_1, \ldots, y_n, y\}$ (les y_i sont les cercles pleins), pour différentes valeurs du pas h, avec la figure 4.8 où est tracée l'entropie de Shannon estimée par des noyaux. Les résultats sont très similaires, mais les calculs sont ici beaucoup plus faciles. Nous remarquons également l'influence du pas h, et les conclusions sont identiques à celles données pour l'entropie de Shannon estimée par des noyaux.

La fonction de x construite en prenant l'espérance de la v.a. $H_2^*(\phi_{n,Y_n(x)})$, ne s'exprime pas sous forme analytique en raison de la présence du logarithme. L'inégalité de Jensen [60] indique cependant que

$$\mathbb{E}\left(-\log \left\|\phi_{n,Y_n(x)}\right\|_2^2 \right) \geq -\log \mathbb{E}\left\|\phi_{n,Y_n(x)}\right\|_2^2. \tag{4.12}$$

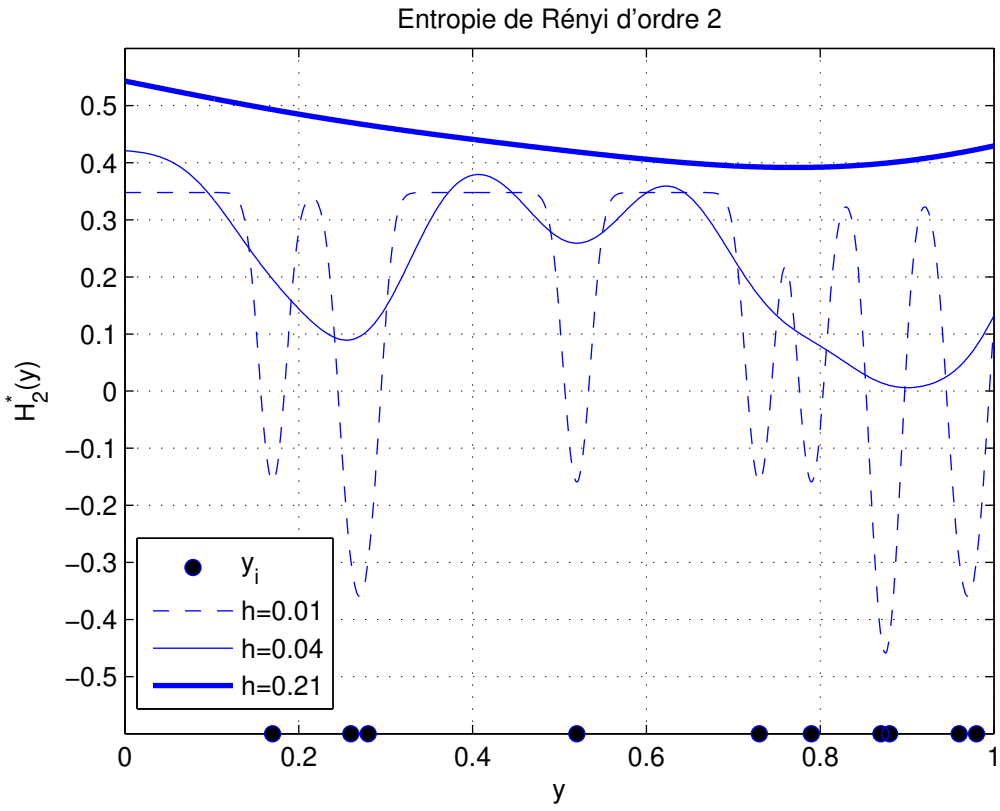

FIGURE 4.11 – Entropie de Rényi d'ordre 2. Les cercles pleins désignent les points du plan courant.

Minimiser $\mathbb{E}\left\|\phi_{n,Y_n(x)}\right\|_2^2$ (qui, comme ce sera démontré, a une forme analytique quand $Y_n(x)$ suit une loi normale) devrait donc rendre grand le terme $\mathbb{E}H_2^*(\phi_{n,Y_n(x)})$. Mais cela revient à utiliser l'entropie de Tsallis d'ordre 2, considérée au paragraphe suivant.

Nous ne retenons donc pas le critère d'entropie de Rényi, car la fonction de x obtenue n'a pas d'expression analytique.

4.1.6 Entropie de Tsallis

Le critère d'entropie de Rényi ne s'est pas révélé satisfaisant pour notre problème, car tout comme le critère de Shannon il est coûteux en temps de calcul. Cependant, la formule (4.12) donne un argument pour utiliser une autre fonction d'entropie, appelée *entropie de Havrda-Charvát* ou *entropie de Tsallis* [71, 97, 171].

Définition 4.1.23 *Soit $\phi(\cdot)$ une densité de probabilité. On appelle* entropie de Tsallis d'ordre l *($l \neq 1$) de la densité ϕ, et on note $H_l(\phi)$, la quantité*

$$\begin{aligned} H_l(\phi) &= \frac{1}{l-1}\left(1 - \int \phi^l\right) \\ &= \frac{1}{l-1}\left(1 - \|\phi\|_l^l\right). \end{aligned}$$

Contrairement à l'entropie de Shannon et de Rényi, l'entropie de Tsallis n'est pas additive : pour deux v.a. indépendantes X et Y, $H_l(X,Y) = H_l(X) + H_l(Y) + (1-l)H_l(X)H_l(Y)$. Le coefficient l mesure l'écart à l'additivité. Pour l'ensemble des densités $\phi(\cdot)$ à support compact $\mathbb{K}$, le maximum d'entropie de Tsallis est atteint pour la loi uniforme sur $\mathbb{K}$ si $l > 0$. On a le même résultat de convergence vers l'entropie de Shannon que pour l'entropie de Rényi.

Proposition 4.1.24 *[97] Lorsque $l \to 1$, l'entropie de Tsallis $H_l(\phi)$ tend vers l'entropie de Shannon $H_1(\phi) = -\int \phi \log \phi$.*

On s'intéresse en particulier à l'entropie de Tsallis d'ordre 2,

$$\begin{aligned} H_2(\phi) &= 1 - \int \phi^2 \\ &= 1 - \|\phi\|_2^2. \end{aligned}$$

Nous faisons les mêmes hypothèses qu'au paragraphe précédent pour modéliser la densité $\phi_{n,y_{n+1}}(\cdot)$. La proposition 4.1.22 donne la fonction de y_{n+1} à minimiser de façon à maximiser l'entropie de Tsallis d'ordre 2 de la densité correspondant à l'échantillon $\{Y^n, y_{n+1}\}$.

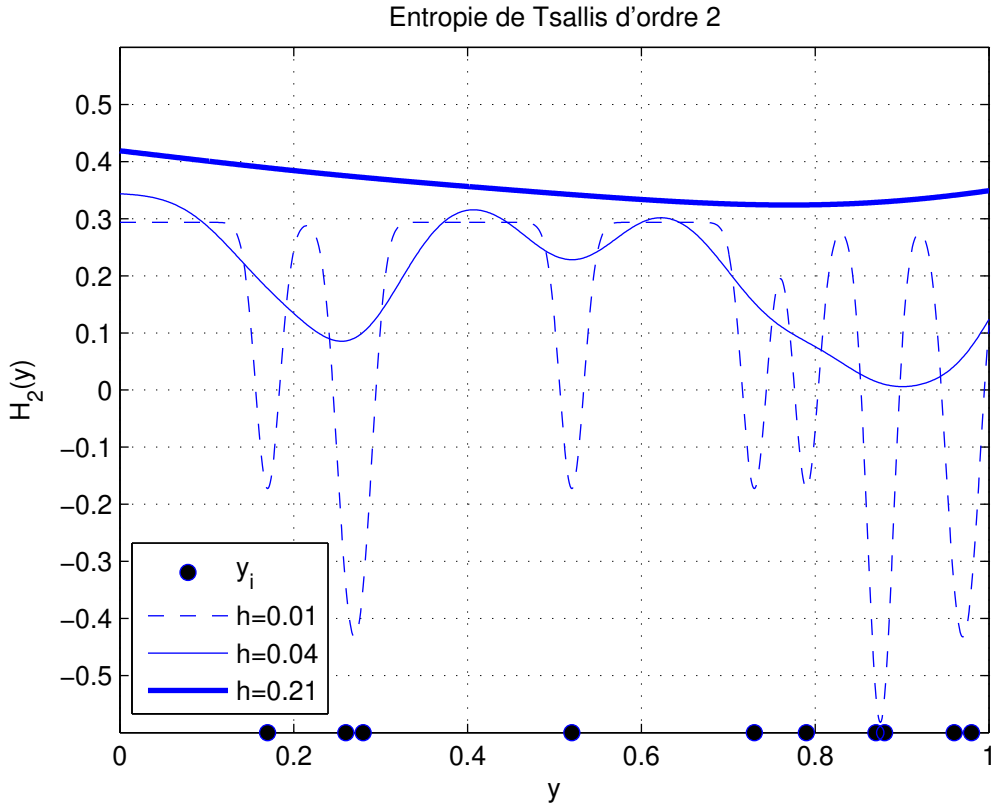

FIGURE 4.12 – Entropie de Tsallis d'ordre 2. Les cercles pleins désignent les points du plan courant.

L'allure du critère est tracé sur la figure 4.12. Nous constatons que celui-ci est pratiquement équivalent au critère de Rényi de la figure 4.11.

Remarque 4.1.25 *Quand h n'est pas trop grand, le maximum est situé entre les points ; si h est trop petit, le critère est plat autour du maximum. Il faut donc trouver une bonne valeur de h.*

La fonction de x correspondante

$$\mathbb{E}H_2\left(\phi_{n,Y_n(x)}\right)$$

se calcule analytiquement si la loi de $Y_n(x)$ est normale.

Proposition 4.1.26 *Si $\phi_{n,Y_n(x)}(\cdot)$ est la densité (4.11) et si $Y_n(x) \sim \mathcal{N}(\mu_n(x), \sigma_n^2(x))$, alors*

$$\mathbb{E}\left\|\phi_{n,Y_n(x)}\right\|_2^2 \propto \frac{1}{\sqrt{\sigma_n^2(x)+2h^2}} \sum_{i=1}^{n} \exp\left\{-\frac{(y_i-\mu_n(x))^2}{2(\sigma_n^2(x)+2h^2)}\right\}. \tag{4.13}$$

Preuve *C'est un cas particulier de la proposition 4.2.4 et du corollaire 4.2.5, qui donnent le résultat dans le cas multidimensionnel.* □

Le point x ajouté à chaque itération est donc

$$x^* = \underset{x\in\mathcal{X}}{\operatorname{argmin}} \frac{1}{\sqrt{\sigma_n^2(x)+2h^2}} \sum_{i=1}^{n} \exp\left\{-\frac{(y_i-\mu_n(x))^2}{2(\sigma_n^2(x)+2h^2)}\right\}.$$

Nous retenons le critère d'entropie de Tsallis pour la suite de l'étude.

4.1.7 Conclusion à l'étude de la diversité dans le cas monodimensionnel

De toutes les fonctions étudiées dans le cas monodimensionnel, nous retenons les deux suivantes :

- la fonction de type maximin (§ 4.1.2), très simple à calculer dans le cas d'une seule sortie, et qui donne des résultats très satisfaisants ;
- la méthode issue de l'entropie de Tsallis (§ 4.1.6), qui, si elle donne des résultats moins intuitifs que la fonction précédente, est cependant facile à calculer et très économe en temps de calcul.

Il faut maintenant tester ces fonctions dans le cas de la dimension 2 (le système réel a 2 sorties, voir l'introduction). L'entropie de Tsallis s'adapte très bien aux dimensions supérieures, nous disposerons d'une forme analytique. L'intégrale double de la fonction maximin ne peut cependant plus être évaluée analytiquement, et nous aurons alors recours à une intégration numérique.

4.2 Diversité en dimension 2 (ou plus)

Passons maintenant à l'étude de la diversité en plus grande dimension. De l'étude préliminaire en dimension 1, nous avons retenu deux fonctions pouvant servir à évaluer la diversité qui semblent satisfaisantes pour notre problème : la fonction distance de type maximin (§ 4.1.2) et la fonction d'entropie de Tsallis (§ 4.1.6). Il s'agit de voir si ces deux critères s'adaptent bien en dimension supérieure.

4.2.1 Fonction de type maximin

Supposons que l'on ait observé l'échantillon $Y^n = \{y_1, \ldots, y_n\}$ ($y_i \in \mathbb{R}^q$), correspondant aux entrées $x_1, \ldots, x_n$ de dimension quelconque, et que l'on souhaite évaluer la diversité en un nouveau point x. Le cokrigeage (annexe D) nous donne, au point x, une moyenne $m_n(x) \in \mathbb{R}^q$ et une matrice de covariance $\Sigma_n(x)$ de taille $q \times q$ (en pratique, les sorties sont supposées indépendantes et chaque sortie est modélisée indépendamment : la matrice $\Sigma_n(x)$ est donc diagonale). L'intégrale à évaluer s'écrit ici comme en (4.3)

$$\mathrm{div}^n_{\mathrm{Mm}}(x) = \int_{\mathbb{R}^q} \min_i ||z - y_i||\, \phi_{(m_n(x),\Sigma_n(x))}(z)\, \mathrm{d}z, \tag{4.14}$$

avec $\phi_{(m_n(x),\Sigma_n(x))}(\cdot)$ la densité de la loi normale de dimension q, de moyenne $m_n(x)$ et variance $\Sigma_n(x)$. Après avoir déterminé les n cellules de Voronoi $C_1, \ldots, C_n$ pour l'ensemble de points $y_1, \ldots, y_n$ (voir annexe J), l'intégrale (4.14) se ramène, à un facteur multiplicatif près, à

$$\mathrm{div}^n_{\mathrm{Mm}}(x) \propto \frac{1}{\sqrt{\det(\Sigma_n(x))}} \sum_{i=1}^{n} \int_{\mathbb{R}^q \cap C_i} \|z - y_i\| \exp\left\{ -\frac{{}^t(z - m_n(x))\Sigma_n^{-1}(x)(z - m_n(x))}{2} \right\} \mathrm{d}z. \tag{4.15}$$

Cette intégrale ne s'exprime malheureusement pas sous forme analytique. Si l'on veut déterminer

$$x^* = \operatorname*{argmax}_{x \in \mathcal{X}} \mathrm{div}^n_{\mathrm{Mm}}(x),$$

il faut donc procéder à une évaluation numérique de la fonction de diversité. Nous allons directement nous placer dans le cas de $q = 2$ sorties, car il est coûteux en temps de calcul d'évaluer numériquement des intégrales en grande dimension. Supposons également que les deux sorties sont indépendantes pour simplifier les choses (c'est de toute façon ce qui sera fait en pratique) : $\Sigma_n(x) = \mathrm{diag}(\sigma^2_{n1}(x), \sigma^2_{n2}(x))$. On notera $m_n(x) = {}^t(m_{n1}(x), m_{n2}(x))$ les moyennes données par le krigeage appliqué séparément à chacune des deux sorties. La formule (4.15) se réécrit alors

$$\mathrm{div}^n_{\mathrm{Mm}}(x) \propto \frac{1}{\sigma_{n1}(x)\sigma_{n2}(x)} \sum_{i=1}^{n} \int_{\mathbb{R}^2 \cap C_i} ||z - y_i|| \exp\left[-\frac{(z_1 - m_{n1}(x))^2}{2\sigma^2_{n1}(x)} \right] \exp\left[-\frac{(z_2 - m_{n2}(x))^2}{2\sigma^2_{n2}(x)} \right] \mathrm{d}z_1 \mathrm{d}z_2. \tag{4.16}$$

Afin de pouvoir évaluer numériquement cette intégrale, il faut se placer sur un domaine borné. Notons donc $\mathbb{K}_n$ un compact de $\mathbb{R}^2$ construit à partir de $y_1, \ldots, y_n$. L'intégrale évaluée sur $\mathbb{K}_n$ se ramène à

$$\mathrm{div}^n_{\mathrm{Mm}}(x) \propto \frac{1}{\sigma_{n1}(x)\sigma_{n2}(x)} \sum_{i=1}^{n} \int_{\mathbb{K}_n \cap C_i} ||z - y_i|| \exp\left[-\frac{(z_1 - m_{n1}(x))^2}{2\sigma^2_{n1}(x)} \right] \exp\left[-\frac{(z_2 - m_{n2}(x))^2}{2\sigma^2_{n2}(x)} \right] \mathrm{d}z_1 \mathrm{d}z_2.$$

Remarque 4.2.1 *En pratique, le compact $\mathbb{K}_n$ peut être déterminé de sorte que l'intégrale soit inférieure à un seuil $\varepsilon > 0$ donné sur $\mathbb{R}^2 \setminus \mathbb{K}_n$.*

Après changement de variable $w = {}^t(w_1, w_2) = \zeta\left({}^t(z_1, z_2)\right) = {}^t((z_1 - m_{n1}(x))/\sigma_{n1}(x), (z_2 - m_{n2}(x))/\sigma_{n2}(x))$, on obtient

$$\mathrm{div}^{n}_{\mathrm{Mm}}(x) \propto \sum_{i=1}^{n} \int_{D_i} \sqrt{\sigma^2_{n1}(x)(w_1 - y'_{i1})^2 + \sigma^2_{n2}(x)(w_2 - y'_{i2})^2}\, \exp\left[-\frac{w_1^2}{2}\right] \exp\left[-\frac{w_2^2}{2}\right] \mathrm{d}w_1 \mathrm{d}w_2,$$

où $D_i = \zeta(\mathbb{K}_n \cap C_i)$ et $y'_i = {}^t(y'_{i1}, y'_{i2}) = \zeta(y_i)$. En séparant chaque cellule image D_i en domaines de la forme $D_i^{uv} = \{a_i^{uv} w_1 + b_i^{uv} \leq w_2 \leq c_i^{uv} w_1 + d_i^{uv}, u \leq w_1 \leq v\}$ comme sur la figure 4.13, l'intégrale sur chacune de ces zones devient

$$\mathrm{div}^{n,D_i^{uv}}_{\mathrm{Mm}}(x) = \int_u^v \int_{a_i^{uv} w_1 + b_i^{uv}}^{c_i^{uv} w_1 + d_i^{uv}} \sqrt{\sigma^2_{n1}(x)(w_1 - y'_{i1})^2 + \sigma^2_{n2}(x)(w_2 - y'_{i2})^2}\, \exp\left[-\frac{w_1^2}{2}\right] \exp\left[-\frac{w_2^2}{2}\right] \mathrm{d}w_1 \mathrm{d}w_2.$$

Cette intégrale n'est pas calculable analytiquement, il faut donc se résoudre à utiliser une intégration numérique. Afin de diminuer l'influence des erreurs numériques issues du calcul des coefficients $a_i^{uv}, b_i^{uv}, c_i^{uv}, d_i^{uv}$ (qui peuvent être très grands), nous faisons finalement le changement de variable $t_1 = (w_1 - u)/(v - u)$, qui fait apparaître directement les coordonnées des sommets du domaine D_i^{uv}, pour obtenir

$$\mathrm{div}^{n,D_i^{uv}}_{\mathrm{Mm}}(x) = \int_0^1 \int_{p_i^{uv} t_1 + l_i^{uv}}^{q_i^{uv} t_1 + h_i^{uv}} \sqrt{\sigma^2_{n1}(x)(w_1 - y'_{i1})^2 + \sigma^2_{n2}(x)(t_2 - y'_{i2})^2}\, \exp\left[-\frac{w_1^2}{2}\right] \exp\left[-\frac{t_2^2}{2}\right] \mathrm{d}t_1 \mathrm{d}t_2,$$

avec $w_1 = (v - u)t_1 + u$, et les nouveaux coeficients $p_i^{uv} = a_i^{uv}(v - u), l_i^{uv} = a_i^{uv} u + b_i^{uv}, q_i^{uv} = c_i^{uv}(v - u), h_i^{uv} = c_i^{uv} u + d_i^{uv}$, qui n'ont pas besoin d'être calculés à partir des anciens coefficients $a_i^{uv}, b_i^{uv}, c_i^{uv}, d_i^{uv}$ puisqu'ils s'obtiennent à partir des ordonnées des sommets de D_i^{uv} (voir la figure 4.13).

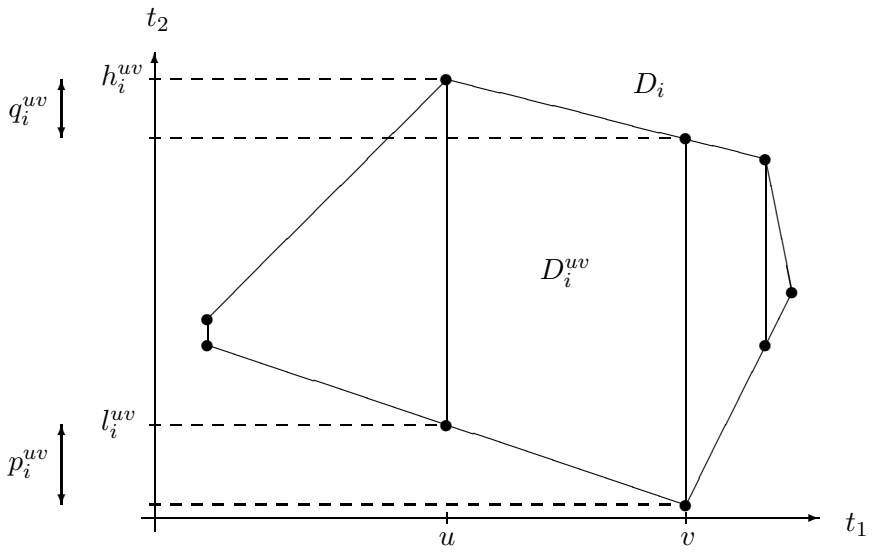

FIGURE 4.13 – Découpage d'une cellule image D_i en domaines d'intégration.

4.2.2 Fonction de diversité de Tsallis multidimensionnelle

Rappelons tout d'abord la modélisation choisie. On se place dans le cas général où les sorties sont q–dimensionnelles (il est ensuite facile de se ramener au cas pratique $q = 2$). On suppose que

l'on a déjà observé l'échantillon $y_1, \ldots, y_n$ et que l'on s'intéresse au point y_{n+1} qui apportera la plus grande diversité à l'ensemble $\{y_1, \ldots, y_n, y_{n+1}\}$. La densité de l'échantillon $\{y_1, \ldots, y_n, y_{n+1}\}$ est modélisée par des noyaux

$$\phi_{n,y_{n+1}}(y) = \frac{1}{n+1} \sum_{i=1}^{n+1} \frac{1}{(2\pi)^{\frac{q}{2}} \sqrt{\det \Sigma}} \exp\left\{ -\frac{{}^t(y-y_i)\Sigma^{-1}(y-y_i)}{2} \right\},$$

où Σ est une matrice symétrique définie positive choisie.

Proposition 4.2.2 *Si $\phi_{n,y_{n+1}}(\cdot)$ est la densité ci-dessus, alors*

$$\left\|\phi_{n,y_{n+1}}\right\|_2^2 = \frac{1}{(2\pi)^{\frac{q}{2}} \sqrt{\det(2\Sigma)} (n+1)^2} \sum_{i,j=1}^{n+1} \exp\left\{ -\frac{{}^t(y_i-y_j)\Sigma^{-1}(y_i-y_j)}{4} \right\}. \tag{4.17}$$

Preuve *Afin de se débarrasser définitivement du terme constant, notons*

$$\alpha_q = (2\pi)^{-\frac{q}{2}} (\det(2\Sigma))^{-\frac{1}{2}} (n+1)^{-2}.$$

$$\begin{aligned}
\int_{\mathbb{R}^q} \phi^2_{n,y_{n+1}}(y)\mathrm{d}y &= \alpha_q \frac{\sqrt{\det(2\Sigma)}}{(2\pi)^{\frac{q}{2}} \det \Sigma} \sum_{i,j=1}^{n+1} \int_{\mathbb{R}^q} \exp\left\{ -\frac{{}^t(y-y_i)\Sigma^{-1}(y-y_i) + {}^t(y-y_j)\Sigma^{-1}(y-y_j)}{2} \right\} \mathrm{d}y \\
&= \frac{\alpha_q}{(2\pi)^{\frac{q}{2}} \sqrt{\det(\frac{\Sigma}{2})}} \sum_{i,j=1}^{n+1} \int_{\mathbb{R}^q} \exp\left\{ -\frac{1}{2}\left[{}^t\left(y - \frac{y_i+y_j}{2}\right) \left(\frac{\Sigma}{2}\right)^{-1} \left(y - \frac{y_i+y_j}{2}\right) \right.\right. \\
&\quad \left.\left. -{}^t(y_i+y_j)(2\Sigma)^{-1}(y_i+y_j) + {}^t y_i \Sigma^{-1} y_i + {}^t y_j \Sigma^{-1} y_j \right] \right\} \mathrm{d}y \\
&= \alpha_q \sum_{i,j=1}^{n+1} \exp\left\{ -\frac{-{}^t(y_i+y_j)(2\Sigma)^{-1}(y_i+y_j) + {}^t y_i \Sigma^{-1} y_i + {}^t y_j \Sigma^{-1} y_j}{2} \right\} \\
&= \alpha_q \sum_{i,j=1}^{n+1} \exp\left\{ -\frac{{}^t y_i \Sigma^{-1} y_i + {}^t y_j \Sigma^{-1} y_j - 2{}^t y_i \Sigma^{-1} y_j}{4} \right\} \\
&= \alpha_q \sum_{i,j=1}^{n+1} \exp\left\{ -\frac{{}^t(y_i-y_j)\Sigma^{-1}(y_i-y_j)}{4} \right\}.
\end{aligned}$$

□

Remarque 4.2.3 *En pratique, on peut s'inspirer des résultats observés dans le cas d'une seule sortie (§ 4.1.5) et prendre par exemple, après avoir normalisé chaque sortie par la transformation linéaire*

$$y \mapsto \frac{y}{y_{\max} - y_{\min}},$$

(avec $y_{\min}$ et $y_{\max}$ respectivement le minimum et le maximum observé pour cette sortie), la matrice $\Sigma = h^2 I_2$, avec I_2 la matrice unité d'ordre 2 et h une constante à choisir (dans ce cas, les sorties sont donc supposées indépendantes).

On cherche à minimiser cette quantité, de façon à maximiser l'entropie de Tsallis d'ordre 2, $H_2(\phi_{n,y_{n+1}}) = 1 - \left\|\phi_{n,y_{n+1}}\right\|_2^2$. Puisque $y_{n+1} = Y_n(x)$ est inconnu, mais suit la loi $\mathcal{N}(m_n(x), \Sigma_n(x))$

donnée par le krigeage, le critère de choix du nouveau point est (en se débarassant de la constante multiplicative)

$$x^* = \underset{x\in\mathcal{X}}{\operatorname{argmin}}\, \mathbb{E}_{y_{n+1}\sim Y_n(x)} \sum_{i,j=1}^{n+1} \exp\left\{-\frac{{}^t(y_i-y_j)\Sigma^{-1}(y_i-y_j)}{4}\right\}.$$

Ne gardant que les termes en y_{n+1}, on obtient

$$x^* = \underset{x\in\mathcal{X}}{\operatorname{argmin}}\, \mathbb{E} \sum_{i=1}^{n} \exp\left\{-\frac{{}^t(y_i-Y_n(x))\Sigma^{-1}(y_i-Y_n(x))}{4}\right\}.$$

Proposition 4.2.4 *Le point ajouté par le critère de diversité de Tsallis est*

$$x^* = \underset{x\in\mathcal{X}}{\operatorname{argmin}}\, \frac{1}{\sqrt{\det(\Sigma_n(x)+2\Sigma)}} \sum_{i=1}^{n} \exp\left\{\frac{{}^tU_{xi}C_x^{-1}U_{xi}-R_{xi}}{2}\right\}, \tag{4.18}$$

avec $C_x = \Sigma_n^{-1}(x) + (2\Sigma)^{-1}$, $U_{xi} = \Sigma_n^{-1}(x)m_n(x) + (2\Sigma)^{-1}y_i$ *et* $R_{xi} = {}^ty_i(2\Sigma)^{-1}y_i + {}^tm_n(x)\Sigma_n^{-1}(x)m_n(x)$.

Preuve *Définissons* C_x, U_{xi} *et* R_{xi} *comme ci-dessus. Alors,*

$$\begin{aligned}
&\mathbb{E}\sum_{i=1}^{n} \exp\left\{-\frac{{}^t(y_i-Y_n(x))\Sigma^{-1}(y_i-Y_n(x))}{4}\right\}\\
&\propto \frac{1}{\sqrt{\det(\Sigma_n(x))}} \sum_{i=1}^{n} \int_{\mathbb{R}^q} \exp\left\{-\frac{{}^t(y_i-y)\Sigma^{-1}(y_i-y)}{4}\right\} \exp\left\{-\frac{{}^t(y-m_n(x))\Sigma_n^{-1}(x)(y-m_n(x))}{2}\right\} \mathrm{d}y\\
&= \frac{1}{\sqrt{\det(\Sigma_n(x))}} \sum_{i=1}^{n} \int_{\mathbb{R}^q} \exp\left\{-\frac{{}^tyC_xy - 2{}^tU_{xi}y + R_{xi}}{2}\right\} \mathrm{d}y\\
&= \frac{1}{\sqrt{\det(\Sigma_n(x))}} \sum_{i=1}^{n} \int_{\mathbb{R}^q} \exp\left\{-\frac{1}{2}\left[{}^t(y-C_x^{-1}U_{xi})C_x(y-C_x^{-1}U_{xi}) - {}^tU_{xi}C_x^{-1}U_{xi} + R_{xi}\right]\right\}\mathrm{d}y\\
&\propto \frac{\sqrt{\det(C_x^{-1})}}{\sqrt{\det(\Sigma_n(x))}} \sum_{i=1}^{n} \exp\left\{\frac{{}^tU_{xi}C_x^{-1}U_{xi}-R_{xi}}{2}\right\}\\
&= \frac{1}{\sqrt{\det\left(I_2 + (2\Sigma)^{-1}\Sigma_n(x)\right)}} \sum_{i=1}^{n} \exp\left\{\frac{{}^tU_{xi}C_x^{-1}U_{xi}-R_{xi}}{2}\right\}\\
&\propto \frac{1}{\sqrt{\det(2\Sigma+\Sigma_n(x))}} \sum_{i=1}^{n} \exp\left\{\frac{{}^tU_{xi}C_x^{-1}U_{xi}-R_{xi}}{2}\right\}
\end{aligned}$$

et on retrouve bien (4.18). □

La formule (4.18) ne se factorise pas dans le cas général $q > 1$ sous une forme similaire au cas $q = 1$, sauf dans le cas où les matrices de covariance sont diagonales.

Corollaire 4.2.5 *Si les matrices de covariance* Σ *et* $\Sigma_n(x)$ *sont diagonales (les sorties sont supposées indépendantes), alors la formule (4.18) se factorise. En effet, si l'on pose* $\Sigma = \operatorname{diag}(\sigma_1^2,\ldots,\sigma_q^2)$, $\Sigma_n(x) = \operatorname{diag}(\sigma_{n1}^2(x),\ldots,\sigma_{nq}^2(x))$, $y_i = {}^t(y_{i1},\ldots,y_{iq})$ *et* $m_n(x) = {}^t(m_{n1}(x),\ldots,m_{nq}(x))$, *le critère se réécrit*

$$x^* = \underset{x\in\mathcal{X}}{\operatorname{argmin}}\, \sum_{i=1}^{n}\prod_{j=1}^{q} \frac{1}{\sqrt{\sigma_{nj}^2(x)+2\sigma_j^2}} \exp\left\{-\frac{(y_{ij}-m_{nj}(x))^2}{2(\sigma_{nj}^2(x)+2\sigma_j^2)}\right\}. \tag{4.19}$$

Remarque 4.2.6 *En faisant $q = 1$, on retrouve bien la formule (4.13) obtenue dans le cas d'une seule sortie.*

Le grand avantage du critère de Tsallis est que la formule analytique (4.18) est valable quelles que soient les dimensions du problème.

4.2.3 Prise en compte des contraintes pratiques dans les critères de diversité

Il s'agit maintenant d'essayer d'adapter les deux fonctions de diversité aux contraintes de l'étude. Rappelons-les brièvement :

- le système étudié a 5 entrées et 2 sorties ;
- les valeurs des sorties sont entachées d'un bruit de mesure ;
- les essais sont effectuée par séries, c'est-à-dire que le critère de diversité doit être capable de sélectionner plusieurs points à chaque itération de l'algorithme ;
- les résultats des mesures arrivent avec un temps de retard. Au moment de planifier les mesures de la série $n + 1$, on ne disposera pas des résultats des mesures de la série n planifiée précédemment.

Les fonctions de diversité ont passé le test de la dimension 2. La présence de bruit peut être prise en compte par le krigeage (voir le § 2.4). Il reste donc à adapter les deux dernières conditions. Nous allons voir que la fonction de type distance maximin ne s'adapte pas bien à ces contraintes. Puis nous constaterons que la fonction d'entropie de Tsallis prend quant à elle naturellement en compte l'ensemble des contraintes. Enfin, nous présenterons une approche simplifiée pouvant être utilisée dans les deux cas.

4.2.3.1 Notations

Nous noterons dans la suite $((x_1, y_1), \dots, (x_n, y_n))$ les observations déjà effectuées, avec $x_i = {}^t(x_i^1, \dots, x_i^p) \in \mathbb{R}^p$ et $y_i = {}^t(y_i^1, \dots, y_i^q) \in \mathbb{R}^q$ pour $i = 1, \dots, n$, et $Y^j = {}^t(y_1^j, \dots, y_n^j)$ le vecteur des observations correspondant à la j^e sortie observée pour $j = 1, \dots, q$. Les q sorties observées sont supposées indépendantes.

On suppose que l'on souhaite ajouter k points en même temps. Dans la suite, les nouveaux points seront notés $((x_{n+1}, y_{n+1}), \dots, (x_{n+k}, y_{n+k}))$, avec $X_0 = (x_{n+1}, \dots, x_{n+k}) \in (\mathbb{R}^p)^k$ la variable, $Y_0 = (y_{n+1}, \dots, y_{n+k}) \in (\mathbb{R}^q)^k$ les sorties correspondantes et $Y_0^j = {}^t(y_{n+1}^j, \dots, y_{n+k}^j)$ le vecteur de dimension k des sorties numéro j, pour $j = 1, \dots, q$.

On rappelle que l'on a modelisé les sorties par krigeage (chapitre 2),

$$y^j(x) = {}^t m_j(x)\beta^j + Z_j(x),$$

où $m_j(\cdot)$ est une fonction connue à valeurs dans $\mathbb{R}^{p_j}$, β^j est un vecteur à p_j coefficients inconnus et $Z_j(\cdot)$ est un processus gaussien de moyenne nulle et fonction de covariance connue $k_j(\cdot, \cdot)$. L'hypothèse du krigeage est

$$\begin{pmatrix} Y^j \\ Y_0^j \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} M^j\beta^j \\ M_0^j\beta^j \end{pmatrix}, \begin{pmatrix} \Sigma_{nn}^j & \Sigma_{n0}^j \\ \Sigma_{0n}^j & \Sigma_{00}^j \end{pmatrix} \right),$$

avec $M^j = {}^t(m_j(x_1), \dots, m_j(x_n)), M_0^j = {}^t(m_j(x_{n+1}), \dots, m_j(x_{n+k})), \Sigma_{nn}^j = \mathrm{cov}(Y^j), \Sigma_{n0}^j = \mathrm{cov}(Y^j, Y_0^j), \Sigma_{0n}^j = \mathrm{cov}(Y_0^j, Y^j)$, et $\Sigma_{00}^j = \mathrm{cov}(Y_0^j)$. Sous certaines hypothèses, la loi *a posteriori* des vecteurs aléatoires Y_0^j, pour $j = 1, \dots, q$, est alors donnée par (voir [133])

$$\left(Y_0^j | Y^j\right) \sim \mathcal{N}(\mu_0^j, \Sigma_0^j)\,,$$

avec

$$\begin{aligned}\mu_0^j &= \left\{\Sigma_{0n}^j(\Sigma_{nn}^j)^{-1} + {}^tV^j({}^tM^j(\Sigma_{nn}^j)^{-1}M^j)^{-1}{}^tM^j(\Sigma_{nn}^j)^{-1}\right\}Y^j\,,\\ \Sigma_0^j &= \Sigma_{00}^j - \Sigma_{0n}^j(\Sigma_{nn}^j)^{-1}\Sigma_{n0}^j + {}^tV^j({}^tM^j(\Sigma_{nn}^j)^{-1}M^j)^{-1}V^j\,,\end{aligned}$$

où ${}^tV^j = M_0^j - \Sigma_{0n}^j(\Sigma_{nn}^j)^{-1}M^j$. Pour $k=1$, on retrouve le prédicteur et la variance de krigeage (2.23) et (2.24).

Remarque 4.2.7 *La loi* $\mathcal{N}\left(\mu_0^j, \Sigma_0^j\right)$ *n'a une densité que si* Σ_0^j *est définie positive. On supposera donc que la fonction de covariance utilisée est définie positive.*

Par indépendance des sorties, si l'on note $f(T)$ la densité d'une variable aléatoire T, alors

$$f(Y_0) = f(Y_0^1)\dots f(Y_0^k).$$

4.2.3.2 Fonction de type maximin

Reprenons les hypothèses et notations utilisés précédemment, et cherchons à optimiser plusieurs points en même temps en utilisant la fonction de type maximin. Le critère à évaluer est

$$\mathbb{E}_{Y_0}\left\{\min_{i=1,\dots,n+k}\ \min_{\substack{j=n+1,\dots,n+k\\ j\neq i}} \|y_i - y_j\|\right\}$$

et on doit déterminer le point X_0^* qui maximise le critère. Cela semble cependant très difficile à implémenter en pratique.

La fonction de type maximin ne s'adapte pas telle quelle pour prendre en compte les contraites liées au problème pratique. Néanmoins, nous présentons en fin de paragraphe une approche simplifiée qui sera utilisée. Nous gardons donc la fonction de type maximin pour des tests comparatifs sur des données simulées (voir le chapitre 5).

4.2.3.3 Entropie de Tsallis

On rappelle tout d'abord le formalisme de la méthode utilisant l'entropie de Tsallis. Soit $y \in \mathbb{R}^q$, on définit la densité de l'échantillon $\{y_1, \dots, y_n, y_{n+1}, \dots, y_{n+k}\} \in \mathbb{R}^q$ comme étant

$$\phi_{n,Y_0}(y) \propto \sum_{i=1}^{n+k} \exp\left\{-\frac{{}^t(y-y_i)\Sigma^{-1}(y-y_i)}{2}\right\}$$

(avec $\Sigma = \sigma^2 I_q$ par indépendance des sorties ; σ^2 étant fixé *a priori*, les sorties ont été préalablement normalisées comme expliqué à la remarque 4.2.3). On a alors

$$\begin{aligned}\int_{\mathbb{R}^q}\phi_{n,Y_0}^2 &\propto \sum_{i,j=1}^{n+k}\exp\left\{-\frac{{}^t(y_i-y_j)\Sigma^{-1}(y_i-y_j)}{4}\right\} \quad (\text{vrai } \forall\Sigma,\ \text{proposition 4.2.2})\\ &= \sum_{i,j=1}^{n+k}\exp\left\{-\frac{(y_i^1-y_j^1)^2}{4\sigma^2}\right\}\dots\exp\left\{-\frac{(y_i^q-y_j^q)^2}{4\sigma^2}\right\} \quad (\text{car } \Sigma=\sigma^2 I_q,\ \text{corollaire 4.2.5}),\end{aligned}$$

et on recherche

$$X_0^* = \underset{X_0}{\operatorname{argmin}}\, \mathbb{E}_{Y_0} \int_{\mathbb{R}^q} \phi_{n,Y_0}^2. \tag{4.20}$$

Évaluons l'espérance

$$\begin{aligned}
\mathbb{E}_{Y_0} \int_{\mathbb{R}^q} \phi_{n,Y_0}^2 &= \int_{(\mathbb{R}^q)^k} \left(\int_{\mathbb{R}^q} \phi_{n,Y_0}^2 \right) f(Y_0)\, \mathrm{d}Y_0 \\
&\propto \int_{(\mathbb{R}^k)^q} \left(\sum_{i,j=1}^{n+k} \prod_{l=1}^{q} \exp\left\{ -\frac{(y_i^l - y_j^l)^2}{4\sigma^2} \right\} \right) \prod_{l=1}^{q} f(Y_0^l)\, \mathrm{d}Y_0^l \\
&= \int_{(\mathbb{R}^k)^q} \sum_{i,j=1}^{n+k} \prod_{l=1}^{q} \exp\left\{ -\frac{(y_i^l - y_j^l)^2}{4\sigma^2} \right\} f(Y_0^l)\, \mathrm{d}Y_0^l \\
&= \sum_{i,j=1}^{n+k} \prod_{l=1}^{q} \int_{\mathbb{R}^k} \exp\left\{ -\frac{(y_i^l - y_j^l)^2}{4\sigma^2} \right\} \frac{1}{(2\pi)^{\frac{k}{2}} \sqrt{\det(\Sigma_0^l)}} \exp\left\{ -\frac{{}^t(Y_0^l - \mu_0^l)(\Sigma_0^l)^{-1}(Y_0^l - \mu_0^l)}{2} \right\} \mathrm{d}Y_0^l \\
&\propto \sum_{i,j=1}^{n+k} \left\{ \prod_{l=1}^{q} \frac{1}{\sqrt{\det(\Sigma_0^l)}} \int_{\mathbb{R}^k} \exp\left\{ -\frac{(y_i^l - y_j^l)^2}{4\sigma^2} - \frac{{}^t(Y_0^l - \mu_0^l)(\Sigma_0^l)^{-1}(Y_0^l - \mu_0^l)}{2} \right\} \mathrm{d}Y_0^l \right\}.
\end{aligned}$$

En décomposant la somme

$$\sum_{i,j=1}^{n+k} = \sum_{i=1}^{n}\sum_{j=1}^{n} + \sum_{i=1}^{n}\sum_{j=n+1}^{n+k} + \sum_{i=n+1}^{n+k}\sum_{j=1}^{n} + \sum_{i=n+1}^{n+k}\sum_{j=n+1}^{n+k}$$

et remarquant que

$$\sum_{i=1}^{n}\sum_{j=1}^{n} = \text{ constante},$$

$$\sum_{i=1}^{n}\sum_{j=n+1}^{n+k} + \sum_{i=n+1}^{n+k}\sum_{j=1}^{n} = 2\sum_{i=1}^{n}\sum_{j=n+1}^{n+k}$$

par symétrie des termes sommés, et

$$\begin{aligned}
\sum_{i=n+1}^{n+k}\sum_{j=n+1}^{n+k} &= \sum_{i=n+1}^{n+k-1}\sum_{j=i+1}^{n+k} + \sum_{j=n+1}^{n+k-1}\sum_{i=j+1}^{n+k} + \sum_{i=j=n+1}^{n+k} \\
&= 2\sum_{i=n+1}^{n+k-1}\sum_{j=i+1}^{n+k} + \text{constante},
\end{aligned}$$

l'expression à évaluer se ramène à

$$\boxed{\left(\sum_{i=1}^{n}\sum_{j=n+1}^{n+k} + \sum_{i=n+1}^{n+k-1}\sum_{j=i+1}^{n+k} \right) \left\{ \prod_{l=1}^{q} \frac{1}{\sqrt{\det(\Sigma_0^l)}} \int_{\mathbb{R}^k} \exp\left\{ -\frac{(y_i^l - y_j^l)^2}{4\sigma^2} - \frac{{}^t(Y_0^l - \mu_0^l)(\Sigma_0^l)^{-1}(Y_0^l - \mu_0^l)}{2} \right\} \mathrm{d}Y_0^l \right\}.}$$

Nous allons évaluer séparément les termes de la somme. On notera I_{ij} la matrice carrée nulle partout sauf en (i,j) où elle vaut 1, $\mathbf{1}_i$ le vecteur colonne dont le i^e terme vaut 1 et tous les autres 0, et $\bar{i} = i - n$.

– $\underline{1 \leq i \leq n, n+1 \leq j \leq n+k}$: dans ce cas $y_i^l \notin Y_0^l$ et $y_j^l \in Y_0^l$, y_i^l est donc une constante et y_j^l une variable d'intégration.

$$\begin{aligned}
&\exp\left\{-\frac{(y_i^l - y_j^l)^2}{4\sigma^2} - \frac{{}^t(Y_0^l - \mu_0^l)(\Sigma_0^l)^{-1}(Y_0^l - \mu_0^l)}{2}\right\} \\
&= \exp\left\{-\frac{1}{2}\left({}^tY_0^l\left[(\Sigma_0^l)^{-1} + \frac{1}{2\sigma^2}I_{\bar{j}\bar{j}}\right]Y_0^l - 2{}^tY_0^l\left[(\Sigma_0^l)^{-1}\mu_0^l + \frac{y_i^l}{2\sigma^2}\mathbf{1}_{\bar{j}}\right] + {}^t\mu_0^l(\Sigma_0^l)^{-1}\mu_0^l + \frac{(y_i^l)^2}{2\sigma^2}\right)\right\} \\
&= \exp\left\{-\frac{1}{2}\left({}^tY_0^lW_0^lY_0^l - 2{}^tY_0^lt_0^l + r_0^l\right)\right\} \quad \text{(pour alléger la formule)} \\
&= \exp\left\{-\frac{1}{2}\left[{}^t\left(Y_0^l - (W_0^l)^{-1}t_0^l\right)W_0^l\left(Y_0^l - (W_0^l)^{-1}t_0^l\right) - {}^tt_0^l(W_0^l)^{-1}t_0^l + r_0^l\right]\right\}.
\end{aligned}$$

En intégrant par rapport à Y_0^l, on obtient

$$\int_{\mathbb{R}^k} \exp\left\{-\frac{(y_i^l - y_j^l)^2}{4\sigma^2} - \frac{{}^t(Y_0^l - \mu_0^l)(\Sigma_0^l)^{-1}(Y_0^l - \mu_0^l)}{2}\right\} \mathrm{d}Y_0^l = \frac{(2\pi)^{\frac{k}{2}}}{\sqrt{\det(W_0^l)}} \exp\left\{\frac{{}^tt_0^l(W_0^l)^{-1}t_0^l - r_0^l}{2}\right\},$$

avec $W_0^l = (\Sigma_0^l)^{-1} + (1/(2\sigma^2))I_{\bar{j}\bar{j}}$, $t_0^l = (\Sigma_0^l)^{-1}\mu_0^l + (y_i^l/(2\sigma^2))\mathbf{1}_{\bar{j}}$ et $r_0^l = {}^t\mu_0^l(\Sigma_0^l)^{-1}\mu_0^l + (y_i^l)^2/(2\sigma^2)$ (la matrice W_0^l est inversible car c'est la somme d'une forme quadratique définie positive et d'une forme quadratique positive).
Il reste à évaluer le produit des déterminants, dont le l^e terme est

$$\begin{aligned}
\frac{1}{\sqrt{\det\left(\Sigma_0^lW_0\right)}} &= \frac{1}{\sqrt{\det\left(I_k + \frac{1}{2\sigma^2}\left(0_{k\times(\bar{j}-1)}\ \Sigma_0^l(:,\bar{j})\ 0_{k\times(k-\bar{j})}\right)\right)}} \\
&= \frac{1}{\sqrt{1 + \frac{\Sigma_0^l(\bar{j},\bar{j})}{2\sigma^2}}} \\
&= \sqrt{\frac{2\sigma^2}{2\sigma^2 + \Sigma_0^l(\bar{j},\bar{j})}}.
\end{aligned}$$

Le terme (i,j) de la somme dans le cas $1 \leq i \leq n, n+1 \leq j \leq n+k$ est donc

$$\boxed{\frac{\sqrt{2\sigma^2}(2\pi)^{\frac{k}{2}}}{\sqrt{2\sigma^2 + \Sigma_0^l(\bar{j},\bar{j})}} \exp\left\{\frac{{}^tt_0^l(W_0^l)^{-1}t_0^l - r_0^l}{2}\right\}.}$$

Remarque 4.2.8 *En faisant $k = 1$, on retrouve (à une constante multiplicative près) la formule de l'ajout de points* 1 *par* 1 *(4.19).*

– $\underline{n+1 \leq i < j \leq n+k}$: dans ce cas $y_i^l \in Y_0^l$ et $y_j^l \in Y_0^l$, y_i^l et y_j^l sont des variables d'intégration.

$$\exp\left\{-\frac{(y_i^l-y_j^l)^2}{4\sigma^2}-\frac{{}^t(Y_0^l-\mu_0^l)(\Sigma_0^l)^{-1}(Y_0^l-\mu_0^l)}{2}\right\}$$
$$=\exp\left\{-\frac{1}{2}\left({}^tY_0^l\left[(\Sigma_0^l)^{-1}+\frac{1}{2\sigma^2}(I_{\bar{i}\bar{i}}+I_{\bar{j}\bar{j}}-I_{\bar{i}\bar{j}}-I_{\bar{j}\bar{i}})\right]Y_0^l-2{}^tY_0^l(\Sigma_0^l)^{-1}\mu_0^l+{}^t\mu_0^l(\Sigma_0^l)^{-1}\mu_0^l\right)\right\}$$
$$=\exp\left\{-\frac{1}{2}\left({}^tY_0^lZ_0^lY_0^l-2{}^tY_0^l\tau_0^l+\rho_0^l\right)\right\}\quad\text{(pour alléger la formule)}$$
$$=\exp\left\{-\frac{1}{2}\left[{}^t(Y_0^l-(Z_0^l)^{-1}\tau_0^l)Z_0^l(Y_0^l-(Z_0^l)^{-1}\tau_0^l)-{}^t\tau_0^l(Z_0^l)^{-1}\tau_0^l+\rho_0^l\right]\right\}.$$

En intégrant par rapport à Y_0^l, on obtient

$$\int_{\mathbb{R}^k}\exp\left\{-\frac{(y_i^l-y_j^l)^2}{4\sigma^2}-\frac{{}^t(Y_0^l-\mu_0^l)(\Sigma_0^l)^{-1}(Y_0^l-\mu_0^l)}{2}\right\}\mathrm{d}Y_0^l=\frac{(2\pi)^{\frac{k}{2}}}{\sqrt{\det(Z_0^l)}}\exp\left\{\frac{{}^t\tau_0^l(Z_0^l)^{-1}\tau_0^l-\rho_0^l}{2}\right\},$$

avec $Z_0^l=(\Sigma_0^l)^{-1}+(1/(2\sigma^2))(I_{\bar{i}\bar{i}}+I_{\bar{j}\bar{j}}-I_{\bar{i}\bar{j}}-I_{\bar{j}\bar{i}})$, $\tau_0^l=(\Sigma_0^l)^{-1}\mu_0^l$ et $\rho_0^l={}^t\mu_0^l(\Sigma_0^l)^{-1}\mu_0^l$ (la matrice Z_0^l est inversible car c'est la somme d'une forme quadratique définie positive et d'une forme quadratique positive).
Il reste à évaluer le produit des déterminants. Calculons

$$\begin{aligned}\det\left(\Sigma_0^lZ_0\right) &= \det\left\{I_k+\frac{1}{2\sigma^2}\left(0_{k\times(\bar{i}-1)}\,;\Sigma_0^l(:,\bar{i})-\Sigma_0^l(:,\bar{j})\,;0_{k\times(\bar{j}-\bar{i}-1)}\,;\Sigma_0^l(:,\bar{j})-\Sigma_0^l(:,\bar{i})\,;0_{k\times(k-\bar{j})}\right)\right\}\\
&= \left(1+\frac{\Sigma_0^l(\bar{i},\bar{i})-\Sigma_0^l(\bar{i},\bar{j})}{2\sigma^2}\right)\left(1+\frac{\Sigma_0^l(\bar{j},\bar{j})-\Sigma_0^l(\bar{j},\bar{i})}{2\sigma^2}\right)\\
&\quad-\left(\frac{\Sigma_0^l(\bar{j},\bar{i})-\Sigma_0^l(\bar{j},\bar{j})}{2\sigma^2}\right)\left(\frac{\Sigma_0^l(\bar{i},\bar{j})-\Sigma_0^l(\bar{i},\bar{i})}{2\sigma^2}\right)\\
&= 1+\frac{\Sigma_0^l(\bar{i},\bar{i})+\Sigma_0^l(\bar{j},\bar{j})-\Sigma_0^l(\bar{i},\bar{j})-\Sigma_0^l(\bar{j},\bar{i})}{2\sigma^2}\\
&= \frac{2\sigma^2+\Sigma_0^l(\bar{i},\bar{i})+\Sigma_0^l(\bar{j},\bar{j})-\Sigma_0^l(\bar{i},\bar{j})-\Sigma_0^l(\bar{j},\bar{i})}{2\sigma^2}.\end{aligned}$$

Le terme (i,j) de la somme dans le cas $n+1\leq i<j\leq n+k$ est donc

$$\boxed{\frac{\sqrt{2\sigma^2}(2\pi)^{\frac{k}{2}}}{\sqrt{2\sigma^2+\Sigma_0^l(\bar{i},\bar{i})+\Sigma_0^l(\bar{j},\bar{j})-2\Sigma_0^l(\bar{i},\bar{j})}}\exp\left\{\frac{{}^t\tau_0^l(Z_0^l)^{-1}\tau_0^l-\rho_0^l}{2}\right\}.}$$

Remarque 4.2.9 *On peut simplifier dans les deux cas par* $\sqrt{2\sigma^2}(2\pi)^{\frac{k}{2}}$.

Le critère de diversité de Tsallis s'adapte donc naturellement au cas où les mesures s'effectuent plusieurs à la fois (d'un point de vue théorique en tout cas).

Mesures avec retard Il nous reste à adapter le critère de Tsallis au cas où les résultats des mesures arrivent avec un temps de retard : au moment de planifier la série $\{x_{n+k+1},\ldots,x_{n+2k}\}$, on ne connaît pas encore les résultats correspondant à la série de mesures en cours aux points $\{x_{n+1}^*,\ldots,x_{n+k}^*\}$. Nous proposons d'utiliser le critère (4.20) appliqué à $X_0=(x_{n+1},\ldots,x_{n+k},x_{n+k+1},\ldots,x_{n+2k})\in(\mathbb{R}^p)^{2k}$, mais en fixant les valeurs de $\{x_{n+1},\ldots,x_{n+k}\}$, que l'on connaît. Notons maintenant $Y_0=(y_{n+1},\ldots,y_{n+k},y_{n+k+1},\ldots,y_{n+2k})\in(\mathbb{R}^q)^{2k}$, on recherche

$$X_0^* = \underset{\substack{X_0 \\ x_{n+1}=x^*_{n+1},\dots,x_{n+k}=x^*_{n+k}}}{\operatorname{argmin}} \mathbb{E}_{Y_0} \int_{\mathbb{R}^q} \phi^2_{n,Y_0}.$$

Les k derniers éléments de X_0^*

$$\{x^*_{n+k+1}, \dots, x^*_{n+2k}\}$$

forment alors la prochaine série de mesures à effectuer.

La fonction de diversité de Tsallis prend naturellement en compte l'ensemble des contraintes de l'étude, ce qui n'est pas le cas de la fonction de type maximin. Au chapitre suivant, nous verrons que le critère de Tsallis présente aussi des difficultés pratiques d'implémentation, en raison notamment du mauvais conditionnement des matrices de covariance *a posteriori* Σ_0^j. C'est pourquoi nous utiliserons aussi la méthode simplifiée présentée au paragraphe suivant.

4.2.3.4 Une approche simplifiée

Nous présentons ici une approche pouvant être utilisée pour faire face aux problème d'implémentation des critères de diversité avec prise en compte des contraintes pratiques. Cette approche est nettement moins optimale que la méthode présentée au paragraphe précédent, mais permet d'utiliser la fonction de type maximin (analytiquement pour $q = 1$, avec une intégration numérique pour $q = 2$, voir le § 4.1.2) et d'utiliser le critère de Tsallis sans devoir faire face aux difficultés exposées à la remarque 2.2.16.

- **Prise en compte du bruit de mesure :** les réponses sont modélisées par krigeage avec bruit de mesure, mais afin d'éviter les répétitions, nous proposons d'utiliser la réinterpolation (§ 2.4). Le modèle est donc un interpolateur et la variance de krigeage est nulle aux entrées observées, ce qui garantit la non-répétition des mesures (cette approche est utilisée en pratique en raison du nombre d'essais relativement restreint à notre disposition).
- **Ajout de k points à la fois :** le critère utilisé est celui de l'ajout de points 1 par 1, (4.14) ou (4.18) selon que l'on utilise la fonction de type maximim ou l'entropie de Tsallis. Les k points sélectionnés pour la prochaine série de mesures sont ceux réalisant les k maxima locaux les plus grands du critère choisi.
- **Retard d'arrivée des mesures :** on suppose connaître la position $x_{n+1}, \dots, x_{n+k}$ des mesures en cours, dont on ne sait pas encore la valeur des réponses. On souhaite appliquer le critère ci-dessus pour choisir la prochaine série $x_{n+k+1}, \dots, x_{n+2k}$, mais en tenant compte de l'emplacement de la série de mesures en cours. Nous proposons d'attribuer une valeur fictive aux réponses, donnée par les prédictions du modèle de krigeage $\widehat{y}_n(x_{n+1}), \dots, \widehat{y}_n(x_{n+k})$. Au moment de choisir la prochaine série $x_{n+k+1}, \dots, x_{n+2k}$, on applique donc la méthode ci-dessus aux $n + k$ « observations » $\{(x_1, y_1), \dots, (x_n, y_n), (x_{n+1}, \widehat{y}_n(x_{n+1})), \dots, (x_{n+k}, \widehat{y}_n(x_{n+k}))\}$.

La méthode heuristique proposée dans ce paragraphe permet d'utiliser chacun des deux critères de diversité retenus, en respectant les contraintes pratiques. Son efficacité sera testée au chapitre suivant sur des données simulées.

Chapitre 5

Mise en œuvre pratique et résultats

Il s'agit maintenant de mettre au point une procédure permettant de tester les deux fonctions de diversité retenues au chapitre précédent dans des conditions proches de ce qui sera fait en pratique. Pour cela, nous nous sommes donnés une famille de fonctions-test modélisant chacune un système imaginaire, et avons appliqué pour chacune d'elle le même algorithme séquentiel : un plan initial est enrichi à chaque itération du point optimisant le critère de diversité, jusqu'à atteindre un nombre donné d'observations. Afin de vérifier que la qualité des résultats ne dépend pas du plan initial, l'algorithme a été testé pour plusieurs configurations initiales des points à chaque fois. Les premiers essais ont été effectués sur une famille de 6 fonctions-test définies sur $[-1, 1]^2$ et à valeurs dans $\mathbb{R}$ afin de pouvoir observer graphiquement la façon dont sont placés les points en entrée et en sortie. Après avoir constaté que l'approche utilisée donne des résultats satisfaisants et similaires pour les deux critères de diversité, la procédure d'ajout a été testée dans les dimensions du problème pratique (5 entrées et 2 sorties), en utilisant une fonction-test modélisant grossièrement le système réel construite à partir de données physiques.

Nous présentons dans un premier temps la structure générale de l'algorithme, que l'on retrouvera quelle que soit la configuration. Puis chacun des deux critères de diversité est testé dans le cas de 2 entrées et une sortie, et la répartition finale des points est observée graphiquement. Afin de s'assurer de la pertinence de l'approche par maximisation de la diversité, nous comparons les résultats obtenus avec le critère classique d'ajout de point au maximum de la variance de krigeage (§ 3.2.2). L'inversion du système est présentée. Finalement, l'algorithme est modifié pour prendre en compte les spécificités du problème (délai des réponses et traitement par blocs, voir l'introduction) et testé sur une fonction à 5 entrées et 2 sorties.

5.1 Structure de l'algorithme

Le principe général de l'algorithme est présenté ici. Dans tous les cas, celui-ci est construit autour du noyau suivant :

1. discrétiser le domaine d'entrée $\mathcal{X}$ en une grille notée $\mathcal{G}_\mathcal{X}$;
2. choisir la fonction-test f et le plan initial $P_{n_0} \subset \mathcal{G}_\mathcal{X}$ à n_0 points, poser $n = n_0$;
3. observer les valeurs de sortie $Y^n = f(P_n)$;
4. construire un modèle de krigeage $\widehat{f}_n$ de f à partir des points du plan courant P_n et des observations Y^n, ce qui donne la loi conditionnelle $Y_n(x)$ de la sortie en tout point $x \in \mathcal{X}$;
5. optimiser une fonction de diversité $\mathbb{E}\,\mathrm{div}_n(Y_n(\cdot))$ sur la grille des entrées, ce qui fournit un nouveau point $x^* = \mathrm{argmax}_{x \in \mathcal{G}_\mathcal{X}} \mathbb{E}\,\mathrm{div}_n(Y_n(x))$, et un nouveau plan courant $P_{n+1} := P_n \cup \{x^*\}$, faire $n \leftarrow n + 1$;

6. répéter les étapes $3. - 4. - 5.$, jusqu'à ce que le plan contienne le nombre de points voulus.

L'algorithme est écrit en langage Matlab, et le modèle de krigeage est construit à l'aide de la boîte à outils *DACE* [104, 105]. Ayant observé qu'en pratique la modélisation par krigeage avec une moyenne constante donne des résultats satisfaisants, nous modélisons chacune des sorties indépendamment sous la forme

$$Y(x) = \beta + Z(x) + \varepsilon(x),$$

avec β inconnu, $Z(\cdot)$ un processus gaussien centré stationnaire, et $\varepsilon(x) \sim \mathcal{N}(0, \sigma_\varepsilon^2(x))$ comme au § 2.4 (mais la variance de bruit σ_ε^2 est ici supposée être une fonction connue de x). Les paramètres du modèle sont estimés par maximum de vraisemblance.

Remarque 5.1.1 *Dans les conditions du problème réel (5 facteurs d'entrée), le domaine d'entrée est discrétisé car d'une part l'optimisation globale d'une fonction est très coûteuse en temps de calcul pour de grandes dimensions (la recherche de l'optimum se ramène donc à une maximisation sur les points d'une grille), et d'autre part les mesures étant bruitées il est inutile de vouloir être très précis sur les entrées. Par souci de cohérence, nous décidons de discrétiser le domaine d'entrée quelle que soit sa dimension.*

Les particularités de l'algorithme seront rapportées dans les paragraphes suivants selon le critère utilisé.

5.2 Cas où la fonction inconnue a 2 entrées et 1 sortie

Nous présentons ici l'algorithme tel qu'il est mis en œuvre pour des systèmes déterministes à 2 entrées et 1 sortie (pas de bruit de mesure, $\varepsilon(x) = 0$), puis comparons par simulation les résultats obtenus avec chacun des deux critères de diversité (maximin et Tsallis), ainsi qu'avec l'ajout de point au maximum de la variance de krigeage. Les fonctions-test choisies ainsi que la construction des plans initiaux sont détaillées dans un premier temps, puis nous présentons la mise en œuvre ainsi que les résultats des tests. Finalement, l'inversion du système telle qu'introduite au début du chapitre 4 est présentée.

5.2.1 Choix des fonctions-test et des plans initiaux

Nous dressons ici la liste des fonctions-test choisies comme systèmes fictifs sur lequels appliquer l'algorithme, ainsi que des plans initiaux qui permettront de mesurer la robustesse de la procédure vis-à-vis des conditions initiales.

5.2.1.1 Fonctions-test

Nous avons testé la procédure d'ajout sur 6 fonctions-test, dont les expressions analytiques sont données ci-dessous, toutes définies sur $[-1, 1]^2$. Les surfaces représentatives de ces fonctions sont tracées sur la figure 5.1. Les fonctions f_1, f_2, f_3 et f_6 ont été utilisées dans [148] pour tester une autre procédure d'ajout, et les fonctions f_4 et f_5 dans [168]. Chacune présente des caractéristiques particulières : la courbe de la fonction f_1 a de nombreux « creux » et « bosses » difficiles à modéliser ; la courbe de f_2 a une forme sinusoïdale ; la courbe de f_3 a une forme assez similaire à la courbe de f_1 sur le domaine agrandi $[-3, 3]$; les courbes des fonctions f_4 et f_5 sont plates sur une grande partie du domaine, et à variation assez rapide ailleurs ; la courbe de f_6 présente des similitudes avec celle de la fonction f_1, mais avec la présence d'un « pic » central

très difficile à modéliser. Tester l'algorithme sur ces 6 fonctions permettra de voir si il peut faire face à des systèmes relativement compliqués.

$$
\begin{aligned}
f_1(x_1,x_2) &= 4x_1^2 - 2.1x_1^4 + \frac{x_1^6}{3} + x_1x_2 - 4x_2^2 + 4x_2^4\,;\\
f_2(x_1,x_2) &= -\frac{\sin(3x_1)}{x_1-3} - \frac{\sin(5x_2)}{x_2+5}\,;\\
f_3(x_1,x_2) &= 81x_1^2 - 2,1(3x_1)^4 + \frac{(3x_1)^6}{3} + 9x_1x_2 - 36x_2^2 + 81x_2^4\,;\\
f_4(x_1,x_2) &= \frac{2}{1-\mathrm{e}^{-9}}\exp\left\{-\frac{(x_1+0.1)^2+(x_2+0.1)^2}{0.18}\right\}\,;\\
f_5(x_1,x_2) &= 1 + \frac{2}{e^6+\frac{1}{4}-e^{-6}}\left(e^{6x_1+\frac{x_2}{8}} - e^6 - \frac{1}{8}\right)\,;\\
f_6(x_1,x_2) &= 0.2\mathrm{e}^{x_1-3} + 2.2|x_2| + 1.3x_2^6 - 2x_2^2 - 0.5x_2^4 - 0.5x_1^4 + 2.5x_1^2 +\\
&\quad 0.7x_1^3 + \frac{3}{(8x_1-2)^2+(5x_2-3)^2+1} + \sin(5x_1)\cos(3x_1^2).
\end{aligned}
$$

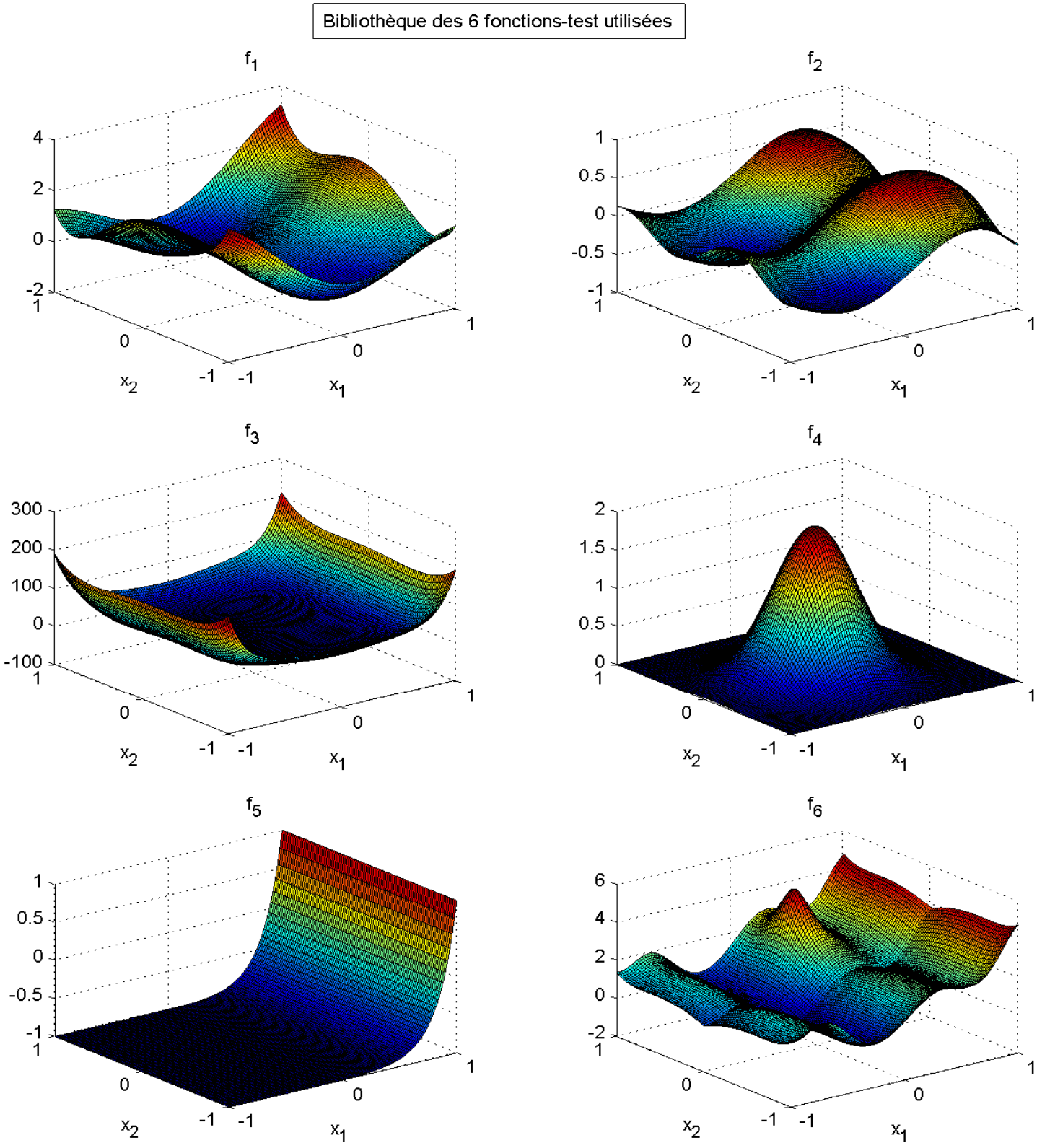

FIGURE 5.1 – Les 6 fonctions-test utilisées dans cette étude.

5.2.1.2 Plans initiaux

Ne disposant pas d'information *a priori* sur le système au stade initial de l'expérimentation, nous sommes intéressés par un plan initial remplissant l'espace, et nous choisissons un hypercube latin maximin (§ 3.1.3.1) en raison de sa simplicité de construction. Nous allons donc générer une bibliothèque d'hypercubes latins ayant de bonnes propriétés maximin et suffisamment différents les uns des autres afin de pouvoir tester l'influence du plan initial sur les approches séquentielles proposées.

Afin de quantifier la différence entre deux plans $A = \{a_1, \ldots, a_n\}$ et $B = \{b_1, \ldots, b_n\}$ de $\mathbb{R}^2$ (mais la méthode fonctionne aussi pour $\mathbb{R}^d$), nous avons utilisé la fonction

$$\mathrm{d}_P(A, B) = \frac{1}{2n} \sum_{i=1}^{n} \left(\min_{j=1,\ldots,n} \|a_i - b_j\|_2 + \min_{j=1,\ldots,n} \|b_i - a_j\|_2 \right),$$

où $\|\cdot\|_2$ désigne la norme euclidienne sur $\mathbb{R}^2$.

Remarque 5.2.1 *La fonction $\mathrm{d}_P(\cdot,\cdot)$ n'est pas une distance, car elle ne vérifie pas l'inégalité triangulaire. Une autre façon de quantifier l'éloignement de deux plans A et B, qui est une vraie distance (voir l'annexe I), serait la fonction*

$$\mathrm{d}_I(A, B) = \frac{1}{n} \min_{\tau \in \mathcal{S}_n} \sum_{i=1}^{n} \left\| a_i - b_{\tau(i)} \right\|_2,$$

où $\mathcal{S}_n$ désigne l'ensemble des permutations de $\{1, \ldots, n\}$. Malheureusement, celle-ci exige l'évaluation de la somme pour les $n!$ permutations τ possibles, ce qui est très coûteux en temps de calcul.

Nous avons décidé de prendre des plans initiaux à 9 points afin de disposer d'un modèle initial relativement précis. La construction des 20 plans initiaux de la bibliothèque est détaillé ci-dessous.

1. Construction du 1$^{\text{er}}$ plan : générer 1000 hypercubes latins de façon aléatoire, et garder parmi les 1000 un de ceux qui sont maximin ($\rightarrow P_1$).
2. Pour $i = 2, \ldots, 20$:
   - générer 1000 hypercubes latins ;
   - parmi ces 1000 plans, choisir aléatoirement un de ceux qui sont maximin, sous la contrainte qu'il soit à une distance $\mathrm{d}_P > 0,5$ de chacun des plans déjà obtenus $P_1, \ldots, P_{i-1}$ ($\rightarrow P_i$).

L'ensemble des 20 plans obtenus par cette méthode est représenté sur la figure 5.2. Nous avons observé empiriquement que la valeur seuil $\mathrm{d}_P > 0,5$ permet d'obtenir 20 plans suffisamment différents.

Remarque 5.2.2 *Il aurait été possible d'utiliser un algorithme de construction d'hypercubes latins maximin, comme par exemple celui présenté dans [77], mais les 20 plans obtenus par notre méthode sont satisfaisants, voir la figure 5.2.*

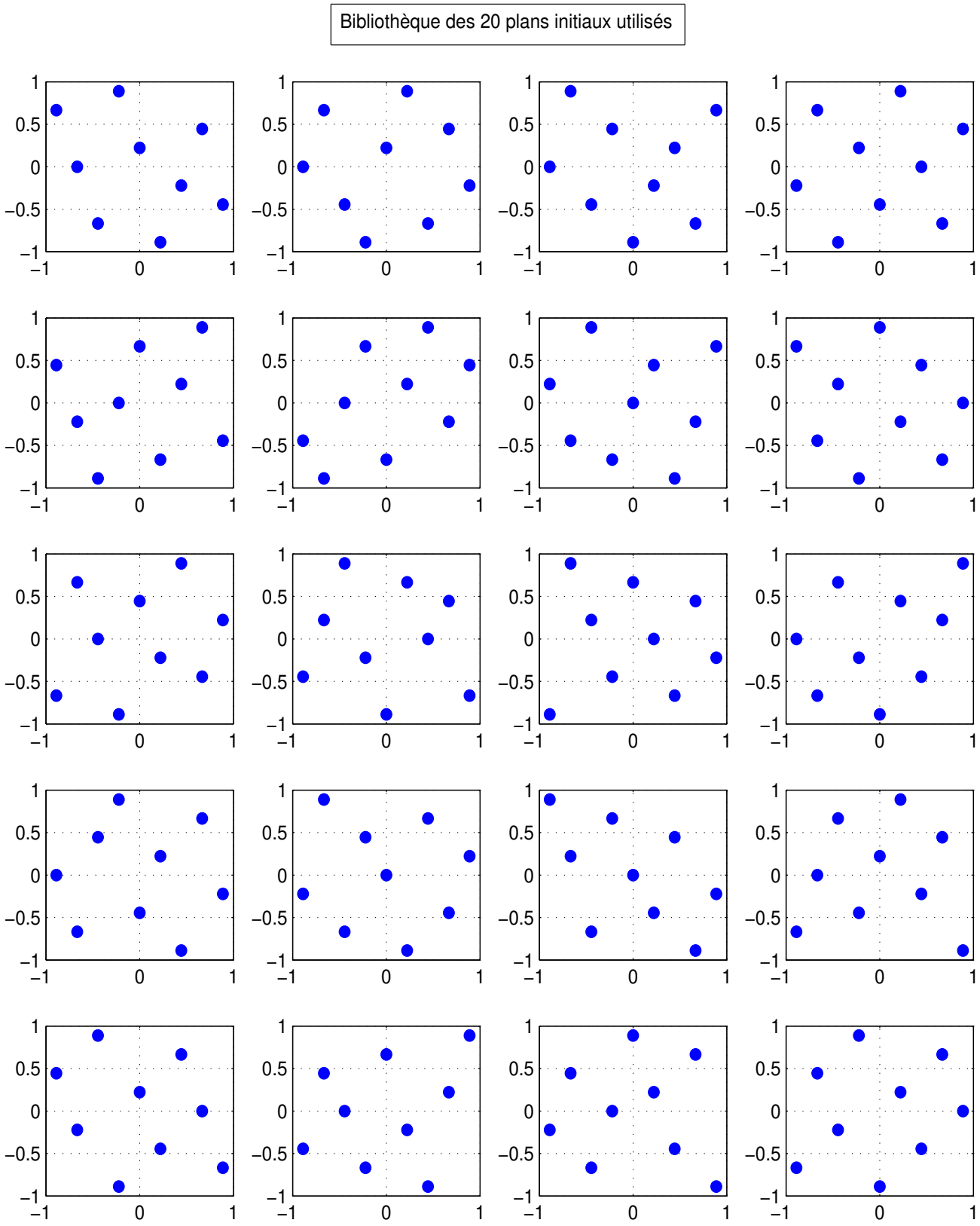

FIGURE 5.2 – Les 20 plans initiaux utilisés dans cette étude.

5.2.2 Tests

Nous présentons maintenant les résultats des tests effectués en utilisant les fonctions-test et les plans initiaux construits précédemment. L'algorithme est paramétré de la façon suivante :

- le domaine d'entrée $[-1,1]^2$ est discrétisé en une grille régulière de taille 201×201. Les calculs sont ensuite effectués « vectoriellement » sur la matrice des entrées toute entière et non pas point par point afin d'accélérer le temps de calcul ;
- la fonction de covariance utilisée dans le modèle de krigeage est la covariance gaussienne isotrope
$$k(x,x') = \sigma^2 \mathrm{e}^{-\theta\|x-x'\|^2},$$
car nous avons observé qu'elle donne en général un bon modèle même quand le nombre de données est restreint (il faut cependant prendre garde aux problèmes de conditionnement des matrices de covariance [1]) ;
- pour l'estimation du paramètre de corrélation θ par maximum de vraisemblance, nous limitons le domaine de recherche de l'algorithme d'optimisation de DACE à $\theta \in [0.1, 20]$ (ce qui permet déjà de modéliser une large gamme de processus), et la valeur initiale est fixée à $\theta_0 = 1$ (ces valeurs ne dépendent pas de la plage de variation des entrées, qui sont systématiquement normalisées par DACE, mais dépendent de la fonction de corrélation choisie).

  Remarque 5.2.3 *[142] Nous avons pu observer qu'en pratique, fixer le paramètre de corrélation à $\theta = 1$ donne souvent lieu à un modèle satisfaisant.*

Nous allons exécuter l'algorithme pour chaque fonction-test et pour chaque plan initial, ce qui va nous permettre d'observer le comportement des deux critères de diversité et d'ajuster leurs paramètres. Puis nous allons comparer la répartition finale des points obtenue par chaque méthode (nous décidons de nous arrêter lorsque le plan contient 20 points). Finalement, nous testons le critère de Tsallis modifié de façon à s'approcher des conditions du problème réel.

5.2.2.1 Critère de diversité maximin

Détaillons tout d'abord le calcul du critère de diversité. Arrivés à l'étape 4 de l'algorithme, on dispose des observations $(x_1, y_1), \ldots, (x_n, y_n)$, où les x_i sont les points du plan courant (des vecteurs de taille 2), les y_i sont les sorties, scalaires, correspondantes. On souhaite évaluer le critère de diversité maximin en un point x de la grille des entrées. Le krigeage donne une approximation de la loi *a posteriori* de la sortie $Y_n(x) \sim \mathcal{N}(\mu_n(x), \sigma_n^2(x))$, dont on va se servir pour évaluer le critère de diversité.

Nous avons vu au § 4.1.2, proposition 4.1.8, que le critère de diversité maximin est affine par morceaux. Il s'agit de calculer l'intégrale (4.3). Pour cela, on partage le domaine d'intégration $\mathbb{R}$ en intervalles (donnés par la proposition 4.1.8) où le critère de diversité est affine. On calcule ensuite les intégrales séparément sur chaque intervalle en utilisant les formules de l'annexe F, puis on fait la somme des valeurs obtenues.

Anticipant les difficultés qui nous attendent dans le cas de 2 sorties (le critère ne s'écrivant pas sous forme analytique, il faudra faire une intégration numérique sur un domaine borné), nous avons voulu observer le comportement du critère dans le cas où l'intégrale (4.3) est tronquée et s'écrit

$$\frac{1}{\sigma_n(x)\sqrt{2\pi}} \int_a^b \mathrm{d}_{\mathrm{Mm}}^n(y) \exp\left\{-\frac{(y-\mu_n(x))^2}{2\sigma_n^2(x)}\right\} \mathrm{d}y$$

(ce qui revient à travailler avec la sortie $Y_n \mathbb{1}_{[a,b]}(Y_n)$). On souhaiterait que a et b correspondent aux bornes du domaine atteignable par la sortie ($[a, b] = f(\mathcal{X})$) : en effet, prendre des bornes infinies conduit à tenir compte de prédictions situées hors du domaine atteignable, et ainsi à sur-échantillonner les bords du domaine ; au contraire, prendre des bornes trop resserrées empêche l'algorithme d'explorer les bornes du domaine atteignable par la sortie. Si ces bornes sont connues *a priori*, il n'y a pas de problème, mais malheureusement les limites du domaine de sortie seront inconnues en pratique. Nous avons ainsi testé, en plus du cas où l'intégrale (4.3) n'est pas tronquée, le cas où elle est tronquée aux vraies bornes (supposées connues) du domaine, et le cas où elle est tronquée au minimum et au maximum des valeurs observées ($a = \min_{i=1,\dots,n} y_i, b = \max_{i=1,\dots,n} y_i$).

Observons sur la figure 5.3 l'allure du critère maximin, que l'on cherche à maximiser. Les points du plan courant P_1 sont localisés aux cercles pleins bleus. On remarque, au vu de la forme de la surface, que le critère ne se résume pas à une distance mais que l'incertitude est aussi prise en compte.

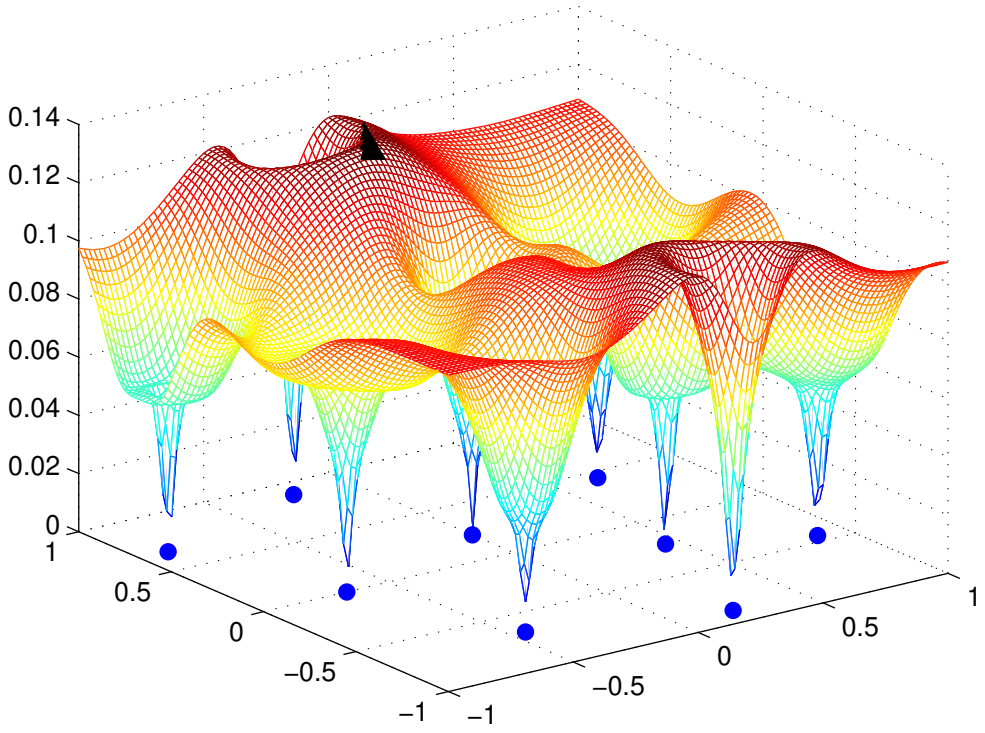

FIGURE 5.3 – Critère maximin, avec la fonction-test f_1 et le plan initial P_1. Les points du plan courant P_1 sont les cercles pleins bleus, le maximum du critère est situé au triangle noir.

Nous pouvons comparer sur la figure 5.4 les plans finaux obtenus avec le critère de diversité maximin pour la fonction f_6 et le plan initial n°10 dans les trois configurations : non tronqué (en haut), tronqué aux vraies bornes du domaine de sortie (au milieu) et tronqué à chaque itération aux bornes observées (en bas). Cet exemple est représentatif de ce que l'on a pu observer pour l'ensemble des 6 fonctions et des 20 plans initiaux. La répartition des entrées est affichée dans la partie gauche, où les pointillés représentent les lignes de niveau de la fonction f_6, les points du plan

initial sont représentés par des cercles pleins et les points ajoutés par leur numéro d'apparition (de 10 à 20). Les valeurs des sorties en fonction de leur ordre d'apparition sont représentées dans la partie droite de la figure, où les points initiaux sont désignés par des cercles pleins et les points ajoutés par des étoiles. L'axe des ordonnées est limité au domaine atteignable par les sorties. On peut observer que le critère maximin non tronqué explore davantage le domaine atteignable en sortie que les critères tronqués (l'ensemble du domaine est décrit par les étoiles en haut à droite), alors que le critère tronqué aux bornes observées (noté O sur la figure, en bas) n'explore pas au-delà des valeurs de sorties observées, ce qui fait qu'on ne dépasse jamais la plage des valeurs décrites par les sorties du plan initial. En ce qui concerne la position des facteurs d'entrée, on constate qu'ils sont d'autant plus concentrés que le critère est tronqué.

Cet exemple pourrait laisser penser que le critère non tronqué est très efficace. Cependant, nous avons observé des cas où il a tendance à chercher les bornes du domaine de sortie de façon très fine, ce qui fait qu'un nombre conséquent d'observations peuvent être situées proches des bornes du domaine atteignable, alors que ces essais auraient pu être utilisés pour échantillonner le domaine des sorties plus uniformément. Observons la figure 5.5, obtenue pour la fonction f_5 et le plan initial n°13, avec le plan final obtenu par le critère non tronqué en haut et par le critère tronqué aux vraies bornes du domaine de sortie en bas. On peut voir en haut à droite qu'une grande partie des observations est proche du minimum atteignable par les sorties. Ce comportement s'explique par le fait suivant : supposons qu'au point x, la prédiction soit située en-deça du domaine atteignable par les sorties. La diversité en x est alors inférieure à $(\min_{i=1\ldots,n} y_i - \hat{y}(x))/2$, qui sera d'autant plus grand que la prédiction est mauvaise. Le critère non tronqué perd donc une grande partie de son efficacité quand le modèle de krigeage est imprécis. Nous remarquons en haut à gauche que le critère non tronqué explore alors le domaine d'entrée.

Le critère qui s'est révélé satisfaisant dans tous les cas est le critère de diversité maximin tronqué aux vraies valeurs (noté V sur les figures). Celui-ci nécessite cependant la connaissance préalable des bornes atteignables par les réponses, ou au moins une idée de celles-ci.

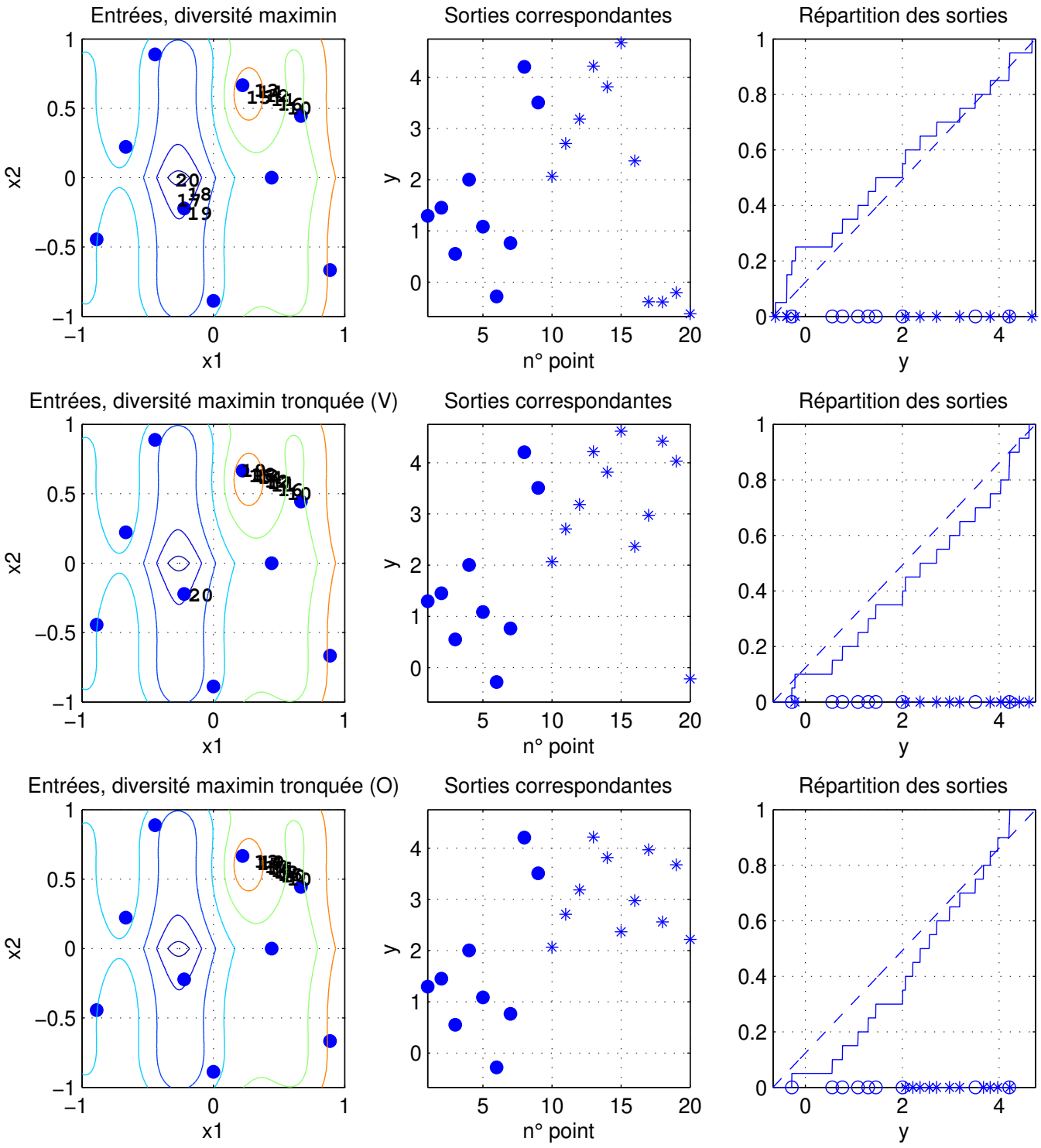

FIGURE 5.4 – Comparaison des critères de diversité maximin : sans troncation (haut), tronqué aux vraies bornes des sorties (milieu), tronqué aux bornes observées à chaque itération (bas), pour la fonction-test f_6 et le plan initial P_{10}.

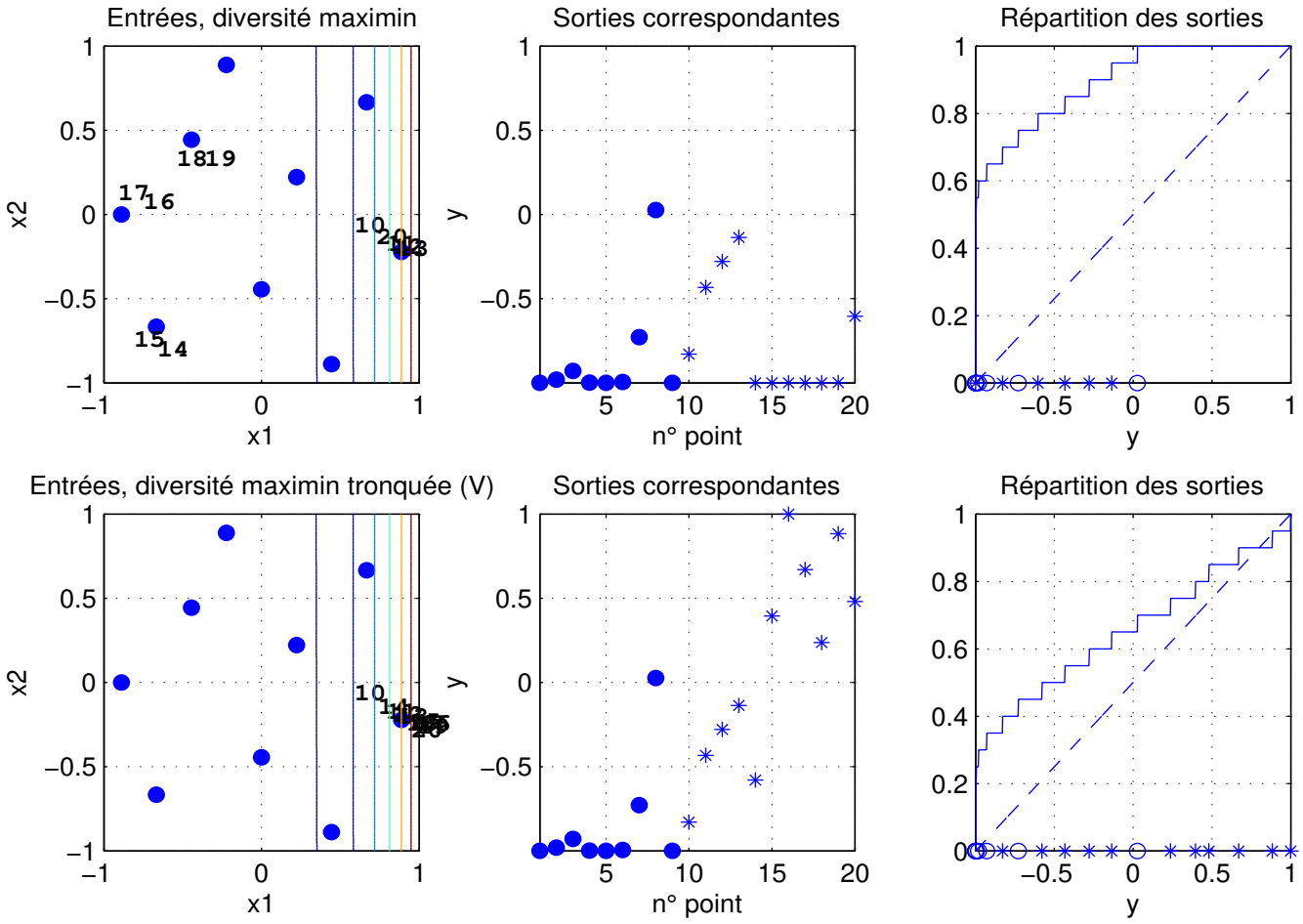

FIGURE 5.5 – Un cas où le critère de diversité maximin non tronqué n'est pas satisfaisant (haut), le cas correspondant obtenu en utilisant le critère tronqué aux vraies bornes des sorties (bas), avec la fonction-test f_5 et le plan initial P_{13}.

5.2.2.2 Critère de diversité de Tsallis

Dans le cas du critère de diversité utilisant l'entropie de Tsallis (§ 4.1.6), nous avons vu en (4.13) que le nouveau point s'obtient par

$$x^* = \underset{x\in\mathcal{G}_\mathcal{X}}{\operatorname{argmin}} \frac{1}{\sqrt{\sigma_n^2(x)+2h^2}} \sum_{i=1}^{n} \exp\left\{-\frac{(y_i-\mu_n(x))^2}{2(\sigma_n^2(x)+2h^2)}\right\},$$

où $y_1,\ldots,y_n$ sont les sorties observées, $\mu_n(x)$ et $\sigma_n^2(x)$ sont respectivement la moyenne et la variance de krigeage au point x, et la valeur du pas h est à choisir. On peut le déterminer *a priori* ou le calculer en fonction des données, en utilisant par exemple la formule empirique (4.7).

Observons sur la figure 5.6 l'allure de l'opposé du critère de Tsallis quand $h = 10^{-2}$ (haut) et $h = 0.5$ (bas). Les points du plan courant sont les cercles pleins bleus, localisés aux « pics renversés » de la surface au-dessus, où le critère de Tsallis a été tronqué car il prend des valeurs très élevées. On peut remarquer que le critère est assez similaire au critère maximin de la figure 5.3 quand $h = 10^{-2}$, dans le sens où ses pics sont aussi situés aux points du plan courant et où le critère n'est pas une simple distance aux observations. Par contre, pour une valeur du pas grande (ici $h = 0.5$, figure du bas), le critère n'est pas satisfaisant car ici le maximum est localisé en un point du plan.

Les plans finaux obtenus avec le critère de Tsallis pour trois valeurs du paramètre h sont représentés sur la figure 5.7. En haut, les sorties ont été ramenées à $[0,1]$ à chaque itération et le paramètre a été fixé à $h = 10^{-2}$; au milieu, h a été recalculé à chaque itération par la formule empirique (4.7)

$$h_{\text{emp}} = \max_{i=1,\ldots,n-1} \frac{y_{(i+1)} - y_{(i)}}{6},$$

et en bas h est recalculé à chaque itération par la règle de mise à l'échelle normale (4.6)

$$\widehat{h}_{\text{NS}} = \widehat{\sigma}\left(\frac{4}{3n}\right)^{\frac{1}{5}},$$

avec $\widehat{\sigma}$ l'écart-type empirique de l'échantillon $\{y_1,\ldots,y_n\}$. Les valeurs de h utilisées grandissent du haut au bas de la figure et leurs évolutions sont illustrées sur la figure 5.8. La fonction et le plan initial utilisés sont les mêmes que pour la figure 5.4, respectivement f_6 et le n°10. Cet exemple particulier reflète ce que l'on a pu observer pour l'ensemble des 6 fonctions et des 20 plans initiaux. La répartition des entrées du plan final est à gauche sur la figure, les pointillés représentant les lignes de niveau de la fonction f_6, les points du plan initial étant représentés par des cercles pleins et les points ajoutés par leur numéro d'apparition, de 10 à 20. À droite sur la figure sont représentées les valeurs des sorties observées en fonction de leur ordre d'apparition, les points initiaux sont désignés par des cercles pleins et les points ajoutés par des étoiles. L'axe des ordonnées est limité au domaine atteignable par les sorties. On remarque qu'une petite largeur de fenêtre $h = 10^{-2}$ a tendance à concentrer les points dans le domaine d'entrée (en haut à gauche), alors qu'une grande valeur du pas à tendance à explorer davantage le domaine d'entrée (au milieu et en bas à gauche). Exceptés les points 16 et 18 pour h_{emp}, et 17 et 20 pour h_{NS}, situés en-dehors du groupe, les trois valeurs de pas h donnent à peu près les mêmes valeurs en entrée. Cependant, cette exploration du domaine d'entrée se révèle ici inutile car en ces 2 points les valeurs de sortie sont proches d'une valeur déjà observée. Nous avons généralement pu observer qu'une petite valeur de pas garantit que les points sont répartis « assez uniformément » en sortie, mais ceux-ci recouvrent parfois une zone plus restreinte que lorsque la valeur de pas est grande, où les sorties occupent un plus grand espace mais de façon moins uniforme.

Si l'on compare avec les critères maximin de la figure 5.4, où la fonction-test et le plan initial sont identiques, on remarque que le critère de diversité maximin tronqué aux vraies bornes des sorties place les points de façon assez similaire aux critères de Tsallis (qui ne nécessitent pas de connaissance des bornes atteignables en sortie). Nous avons observé ce type d'analogie entre les deux critères sur l'ensemble des cas testés.

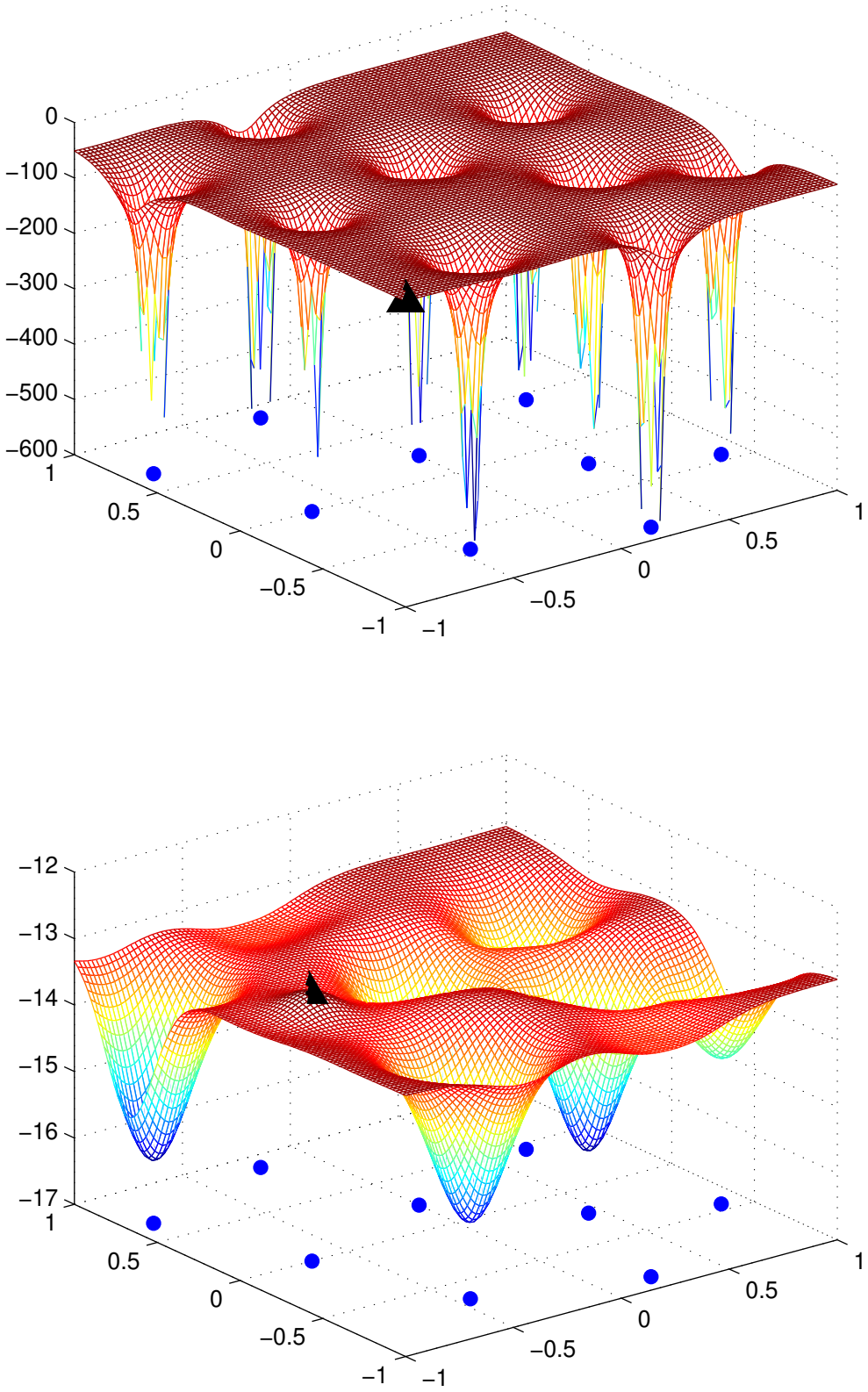

FIGURE 5.6 – Opposé du critère de Tsallis, avec la fonction f_1 et le plan initial P_1, selon la largeur de fenêtre : h=0.01 (haut) et h=0.5 (bas). Les points du plan courant P_1 sont les cercles pleins bleus, le maximum du critère est situé au triangle plein noir dans les deux cas.

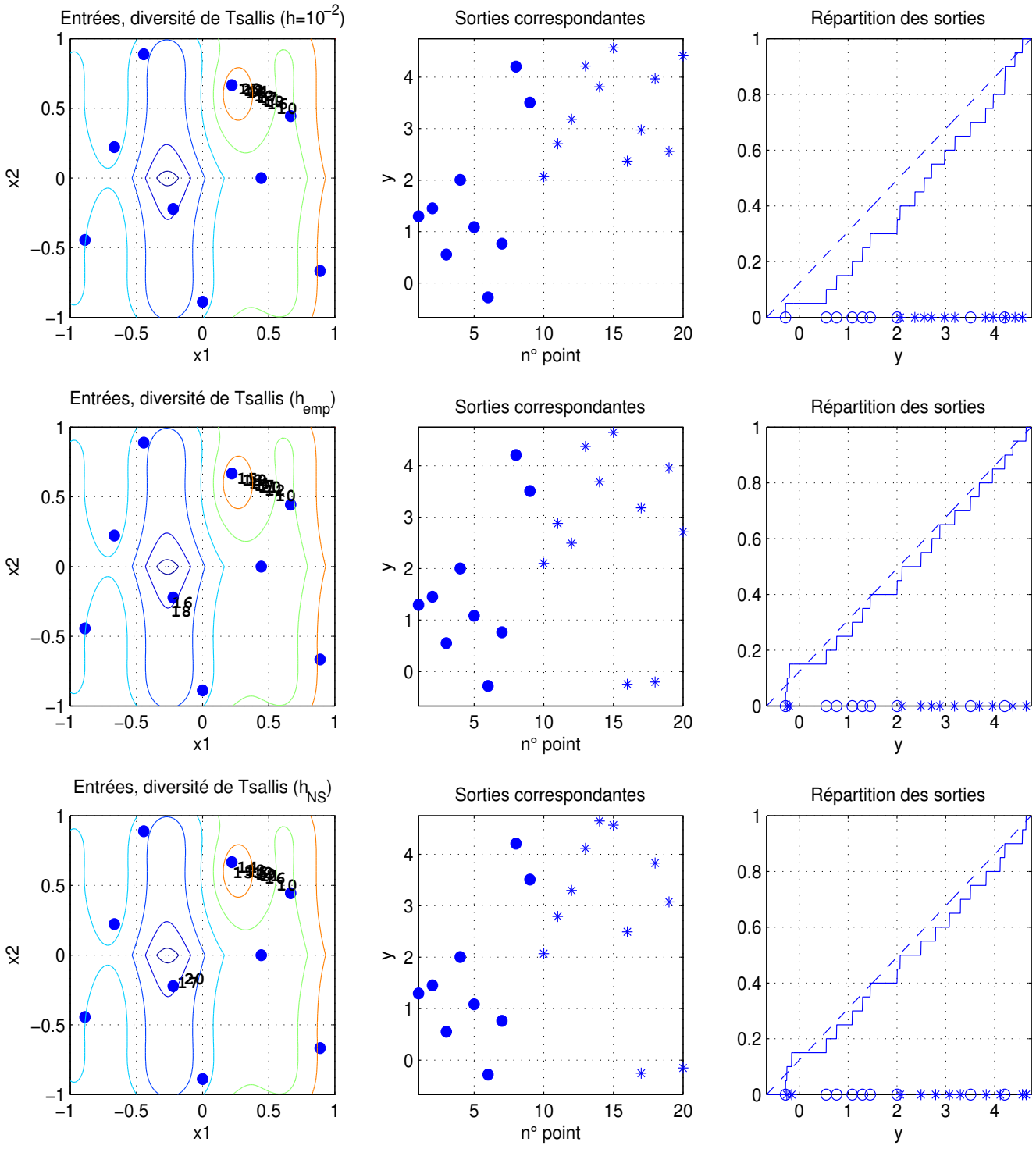

FIGURE 5.7 – Comparaison des critères de diversité de Tsallis : $h = 10^{-2}$ (haut), h_{emp} calculé à chaque itération (milieu), h_{NS} calculé à chaque itération (bas), pour la fonction-test f_6 et le plan initial P_{10}.

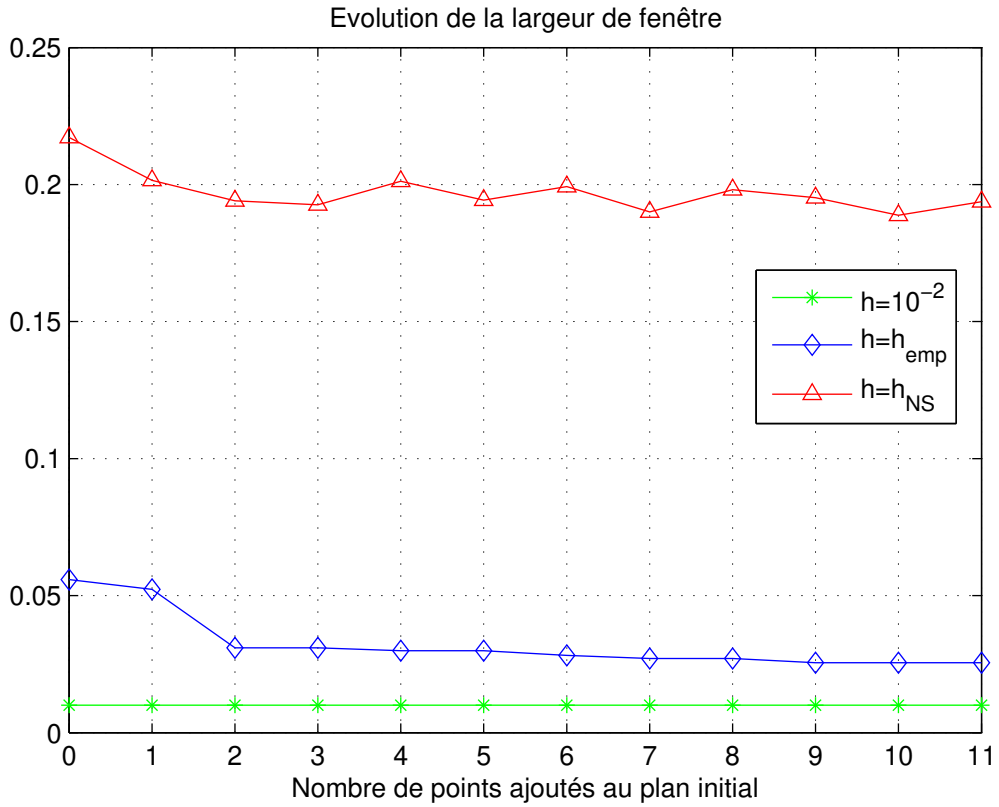

FIGURE 5.8 – Evolution des largeurs de fenêtres h en fonction du nombre de points ajoutés au plan initial.

5.2.2.3 Variance de krigeage

Afin de s'assurer qu'il était bien utile de définir les critères de diversité précédents, nous avons testé l'algorithme séquentiel en utilisant le critère classique de variance de krigeage. Le point ajouté à chaque itération est

$$x^* = \underset{x \in \mathcal{G}_\mathcal{X}}{\operatorname{argmax}}\, \sigma_n^2(x).$$

Sur la figure 5.9, en haut, sont tracées les entrées et sorties du plan final correspondant à la même fonction et au même plan initial que pour la figure 5.4 et la figure 5.7. Nous pouvons remarquer qu'à l'inverse des critères précédents, les entrées sont ici très dispersées : on a construit un plan remplissant l'espace. Les sorties sont dispersées, mais sont souvent proches de valeurs déjà obtenues. Un cas extrême est présenté au bas de la figure 5.9, où sont utilisées la fonction f_4 et le plan initial n°10, et l'on constate que des valeurs d'entrée éloignées les unes des autres (à gauche) ne garantissent pas d'obtenir des valeurs de sortie bien réparties. Le critère de variance de krigeage n'est donc pas satisfaisant pour notre étude.

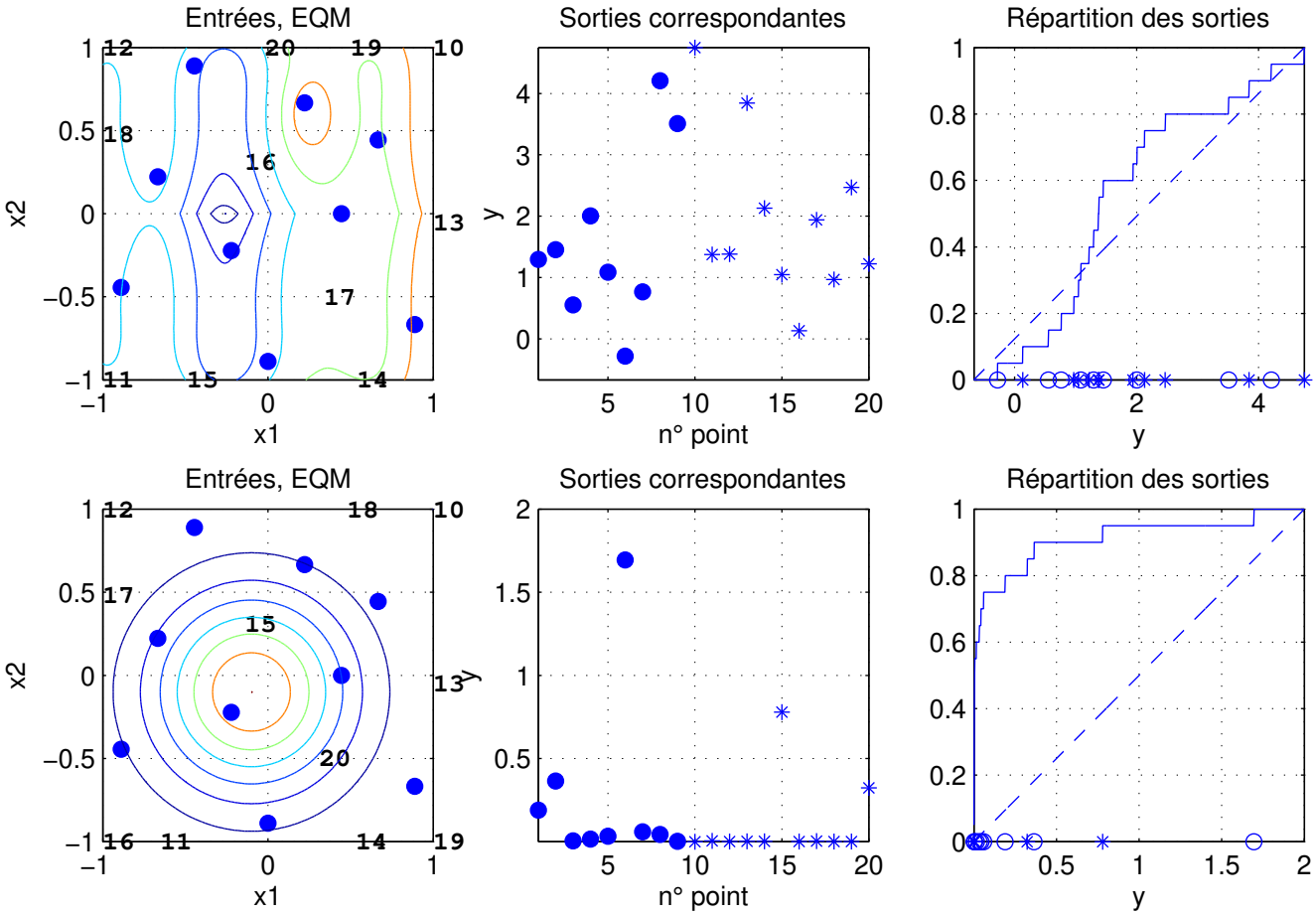

FIGURE 5.9 – Ajout du nouveau point au maximum de la variance de krigeage, pour la fonction-test f_6 et le plan initial P_{10} (haut), et pour la fonction-test f_4 et le plan initial P_{10} (bas).

5.2.2.4 Critère de Tsallis modifié

Nous testons finalement le critère de Tsallis avec prise en compte des contraintes de mesures groupées et de retard, présentées au § 4.2.3.3. Dans un premier temps, nous ajoutons 2 points à la fois au plan courant. Puis nous introduisons un retard d'arrivée des mesures. Les résultats présentés sont assez sommaires et incomplets car nous n'avons pas résolu le problème lié au coût de calcul de la méthode : dans le cas d'une mesure à la fois, l'optimisation consistait à chercher l'optimum sur une grille, ce qui est un problème simple de tri ; dans le cas de plusieurs mesures à la fois, avec l'approche exacte de Tsallis, il faut chercher la série de points de la grille réalisant l'optimum du critère, ce qui est un problème d'optimisation multi-entrées beaucoup plus coûteux en temps de calcul. L'approche simplifiée présentée au § 4.2.3.4 conduit à chercher les meilleurs optima locaux sur la grille, ce qui est aussi un problème simple de tri. Cette approche s'est heureusement révélée satisfaisante dans le cas de 5 entrées et 2 sorties.

Ajout de 2 points à la fois. Nous ajoutons à chaque itération le meilleur couple de points pour le critère de Tsallis. L'optimisation se fait maintenant sur le domaine d'entrée discrétisé en une grille régulière de taille 41×41 points, car la recherche de l'optimum du critère pour deux entrées en même temps est très coûteuse en temps de calcul. Nous n'avons pu mener l'algorithme jusqu'à son terme que pour un nombre restreint de plans initiaux et de fonctions-test. Les remarques que nous faisons ici sont représentatives de ce qui a été observé.

Sur la figure 5.10 nous avons représenté le plan final obtenu par ajout de points par paquets de 2, pour la fonction f_1 et le plan initial n°1. Nous pouvons remarquer à gauche que les points sont assez proches dans le domaine d'entrée, et plutôt bien répartis dans le domaine de sortie.

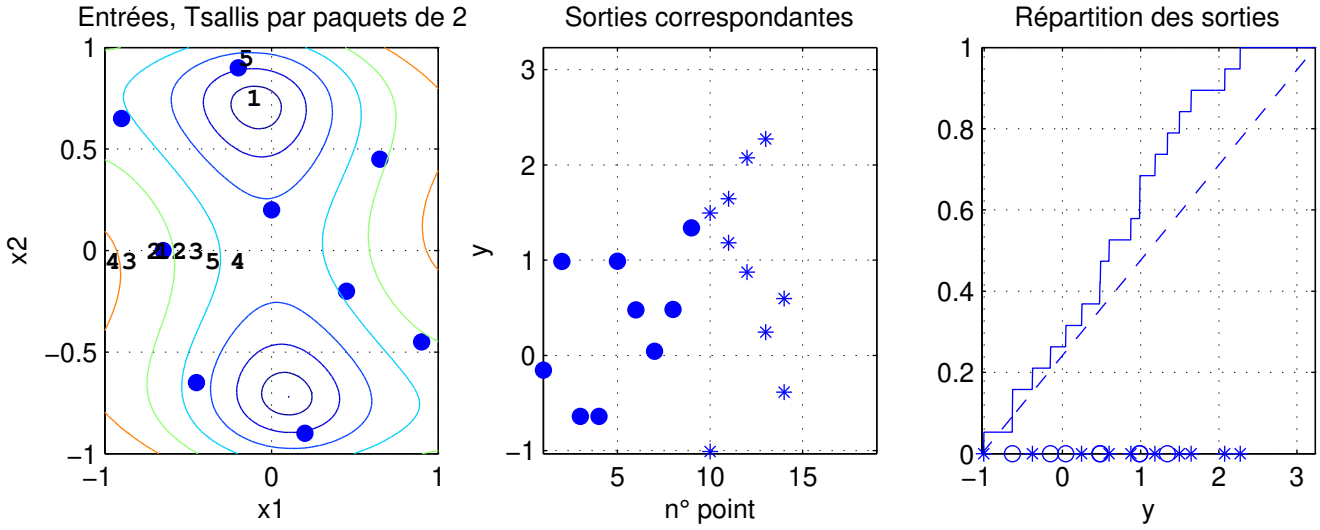

FIGURE 5.10 – Critère de Tsallis avec prise en compte de mesures effectuées 2 par 2, pour la fonction-test f_1 et le plan initial P_1.

Comparons avec la figure 5.11, où les points ont été ajoutés 1 par 1 (avec le domaine d'entrée discrétisé en une grille de taille 201×201). On constate que le positionnement des sorties n'est pas beaucoup dégradé par rapport à l'ajout de points 1 par 1.

Si l'on compare finalement avec l'approche simplifiée présentée en 4.2.3.4 (figure 5.12), on constate que la répartition des sorties est relativement satisfaisante, même si la partie inférieure du domaine atteignable par les sorties n'est pas explorée.

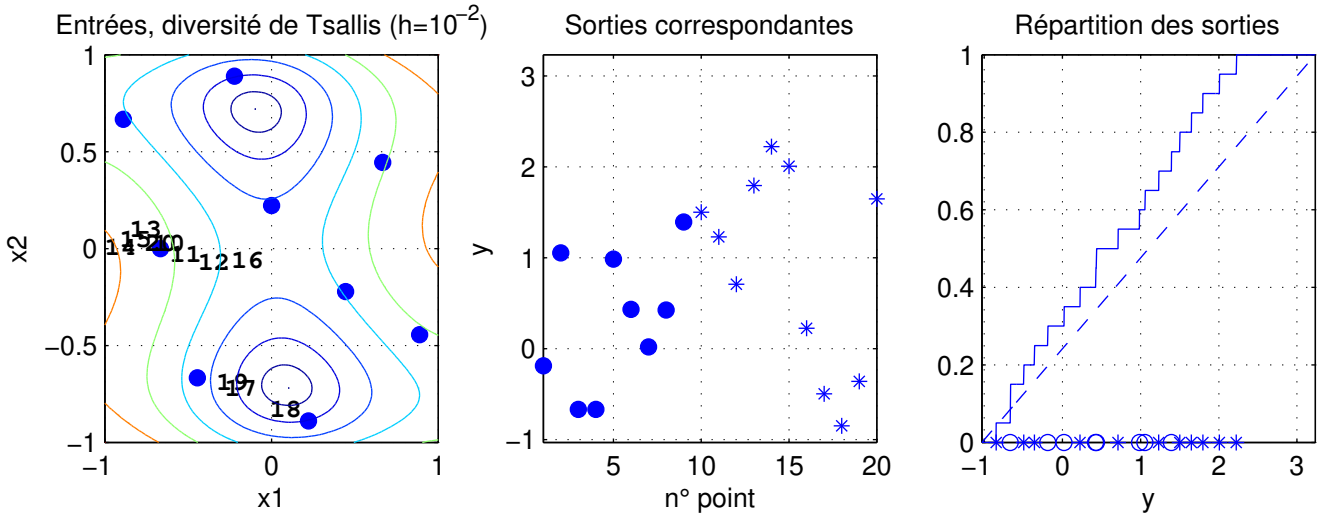

FIGURE 5.11 – Plan final obtenu par ajout de points 1 par 1 avec critère de diversité de Tsallis, pour la fonction-test f_1 et le plan initial P_1.

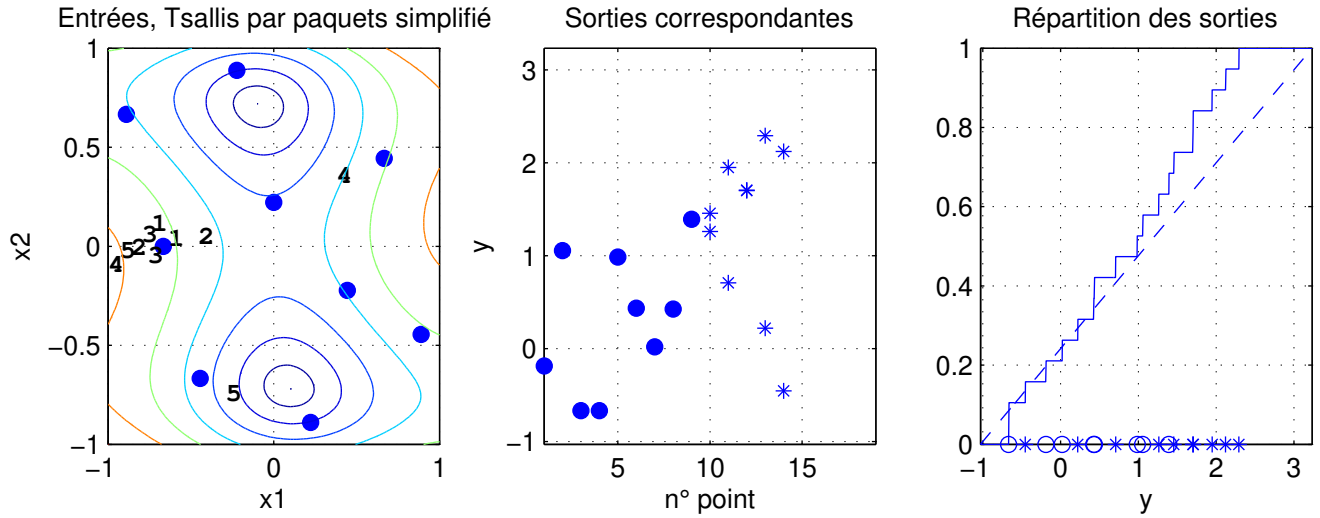

FIGURE 5.12 – Plan final obtenu par ajout de 2 points à la fois avec critère de diversité de Tsallis simplfié, pour la fonction-test f_1 et le plan initial P_1.

Ajout de 2 points à la fois avec retard. Voyons enfin comment sont répartis les points quand les points sont ajoutés par paquets de 2 avec retard d'arrivée des mesures. Le cas correspondant à la fonction f_1 et au plan initial n°1 est représenté sur la figure 5.13. On peut remarquer que la position des points en sortie est nettement dégradée par rapport au cas sans retard de la figure 5.10 (la première série de points est identique à celle de la figure 5.10, ce qui est normal puisqu'elle est déterminée de la même manière). Nous avons observé que les résultats sont sensibles à la taille de la grille, il serait donc intéressant de pouvoir tester l'approche avec une grille plus fine une fois le problème de temps de calcul résolu. Pour une comparaison avec l'approche simplifiée du paragraphe 4.2.3.4, voir la figure 5.14.

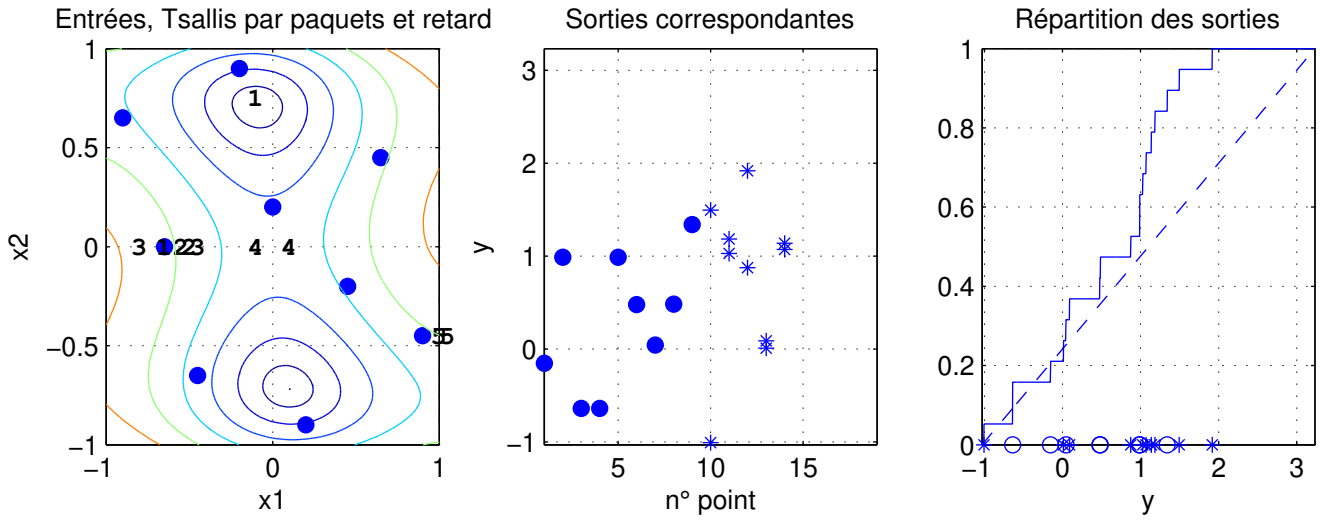

FIGURE 5.13 – Critère de Tsallis avec prise en compte de mesures effectuées 2 par 2 et retard d'arrivée des résultats, pour la fonction-test f_1 et le plan initial P_1.

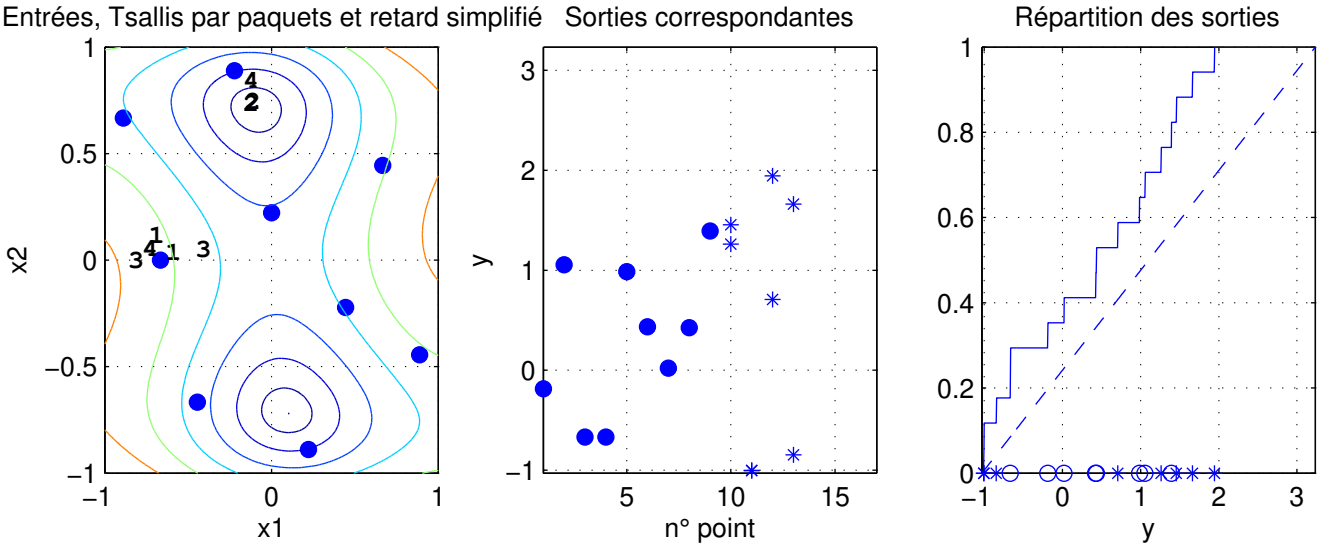

FIGURE 5.14 – Plan final obtenu par ajout de 2 points à la fois et retard, avec critère de diversité de Tsallis simplfié, pour la fonction-test f_1 et le plan initial P_1.

Une recherche plus poussée est nécessaire pour résoudre les problèmes pratiques que nous

avons rencontré lors de l'optimisation, concernant notamment les problème de conditionnement des matrices de covariance, plus importants que dans le cas sans retard : un changement de la fonction de covariance est sans doute nécessaire (la covariance de Matérn, donnant des matrices de covariance numériquement plus stables, est à considérer). Un algorithme d'optimisation plus efficace qui n'évaluerait pas le critère de diversité pour tous les points de la grille pourrait diminuer le temps de calcul.

5.2.3 Inversion du système

Nous nous intéressons maintenant aux prédictions inverses. Rappelons le but de la modélisation tel qu'il avait été défini dans l'introduction du chapitre 4 : après avoir effectué l'ensemble des essais $x_1, \dots, x_N$ et observé le vecteur de réponses $Y^N = {}^t(y_1, \dots, y_N)$, on souhaite, pour une valeur de la réponse $y \in \mathcal{Y}$ donnée, être capable de déterminer une entrée $x_y \in \mathcal{X}$ telle que $f(x_y) \approx y$. La fonction suivante avait été introduite pour mesurer la distance entre la valeur (inconnue) de la réponse en un point x et la valeur y,

$$\begin{aligned} \zeta_N(x,y) &= \mathbb{E}\left[(Y(x)-y)^2 \middle| Y^N\right] \\ &= \mathrm{EQM}(x) + (\widehat{y}(x)-y)^2, \end{aligned}$$

avec $\widehat{y}(x)$ la prédiction en x et $\mathrm{EQM}(x)$ l'erreur quadratique moyenne associée. Nous avions défini la *prédiction inverse* de y par

$$x_y = \operatorname*{argmin}_{x \in \mathcal{X}} \zeta_N(x,y),$$

qui est le point où l'on va observer si l'on souhaite obtenir une réponse proche de y, au sens de la « distance » $\zeta_N(\cdot,\cdot)$. On notera désormais $\zeta_N(y) = \zeta_N(x_y, y)$. Nous allons comparer les plans finaux obtenus par les critères maximin et de Tsallis en utilisant la quantité

$$I_\zeta(x_1, \dots, x_N) = \int_{\mathcal{Y}} \zeta_N(y)\,\mathrm{d}y,$$

qui est une version *a posteriori* du critère de construction de plans d'expériences (impossible à mettre en œuvre) présenté dans l'introduction du chapitre 4.

Dès lors que les mesures ont été effectuées, il est cependant possible de calculer l'écart effectif entre la valeur de la réponse en x_y et y,

$$\Delta_N(y) = (f(x_y) - y)^2.$$

Une mesure de la précision effective des réponses aux prédictions inverses est alors donnée par la quantité

$$I_\Delta(x_1, \dots, x_N) = \int_{\mathcal{Y}} \Delta_N(y)\,\mathrm{d}y.$$

Les valeurs de I_ζ et I_Δ sont tracées sur la figure 5.15 pour chacune des 6 fonctions-test et des 20 plans initiaux, et pour chacun des critères d'ajout, maximin (bleu), Tsallis (vert) et variance de krigeage (rouge). Ces valeurs ont été calculées par discrétisation du domaine de sortie en 201 points espacés régulièrement et du domaine d'entrée en une grille de taille 801×801 (afin d'avoir une bonne précision de la prédiction inverse pour un temps de calcul raisonnable), puis en approximant I_ζ et I_Δ par la moyenne de $\zeta_N(y)$ et $\Delta_N(y)$ en ces 201 points. Les valeurs approchées de I_ζ et I_Δ ainsi obtenues ont ensuite été normalisées en les divisant, pour chaque fonction-test f_i, par $(\max(f_i) - \min(f_i))^2$. Observons que l'ordre de grandeur de I_ζ et I_Δ est le même. On remarque que le critère de variance de krigeage est globalement bien meilleur pour les fonctions-test f_1 et f_3, en raison du fait que le maximum atteignable par les sorties a été observé. Ceci s'explique par le fait que le critère du maximum de variance a tendance à explorer aux bords du domaine d'entrée, or le point où les fonctions-test f_1 et f_3 atteignent leur maximum est situé au bord du domaine d'entrée (figure 5.1). C'est pour cette seule raison que le critère

de variance est ici le plus performant pour ces deux fonctions-test. Les critères maximin et de Tsallis sont plus performants (environ dans les mêmes proportions) pour la fonction-test f_2. Pour les fonctions-test f_4 et f_5, très difficiles à modéliser car présentant des variations brusques, les critères maximin et de Tsallis sont bien plus performants que le critère de variance. Pour la fonction f_6 qui présente un pic très difficile à modéliser, les performances des critères d'ajout sont difficiles à comparer, mais à peu près équivalentes. Un fait marquant est observé pour le critère de variance de krigeage appliqué aux fonctions-test f_4 et f_5, où les valeurs de I_Δ sont plus grandes que celles de I_ζ pour f_4, et plus petites pour f_5. Dans le premier cas, le modèle de krigeage est plus fiable que ce qu'il semble pour la recherche des prédictions inverses, alors que dans le deuxième cas les prédictions inverses sont bien moins précises que prévu par le modèle.

Sur les figures 5.16 et 5.17, la position des prédictions inverses ainsi que les fonctions ζ_N et Δ_N sont tracées, pour la fonction-test f_1 et le plan initial P_1 et la fonction-test f_4 et le plan initial P_{10} respectivement. Pour faciliter la recherche des prédictions inverses, le domaine d'entrée a été discrétisé en une grille régulière de taille 801×801 et le domaine de sortie en 201 points espacés régulièrement. Les plans finaux utilisés correspondent au critère maximin tronqué aux vraies bornes de sorties (en haut), au critère de Tsallis avec $h = 10^{-2}$ (milieu) et au critère de variance de krigeage (bas). Nous pouvons remarquer, en observant la partie gauche, que les prédictions inverses (les points) sont très concentrées pour les critère maximin et de Tsallis (pour lequel elles sont toutes situées sur une ligne dans le cas f_4/P_{10}), et plus éparses pour le critère de variance. Les fonctions ζ_N (colonne du milieu) et Δ_N (colonne de droite) s'écartent de 0 dans les zones où des réponses n'ont pas été observées. Dans le premier cas, on constate que le critère de variance est plus efficace car (par chance) en répartissant les points dans le domaine d'entrée, les observations correspondantes sont bien réparties dans le domaine de sortie, alors que par les critères maximin et de Tsallis les grandes valeurs de sortie n'ont pas été échantillonnées. Dans le deuxième cas, on constate que le critère de variance est bien plus mauvais du fait que le domaine de sortie n'est pas bien échantillonné.

Nous avons essayé de combiner les critère maximin et de Tsallis avec le critère de variance : le premier point est ajouté au maximum de la variance de krigeage, puis, alternativement, un point est ajouté à l'optimum du critère de puis au maximum du critère de variance. Les valeurs de I_ζ et I_Δ obtenues sont tracées sur la figure 5.18, où l'on constate, par comparaison avec la figure 5.15, que cette façon de procéder améliore les performances des deux critères pour les fonctions f_1 et f_3 où ils étaient moins performants que le critère de variance, mais les font globalement empirer dans les autres cas : avec le peu de mesures dont on dispose, il semble que les critères de diversité soient les plus à même à répondre au problème posé, mais il faut se poser la question de la recherche des bornes du domaine de sortie.

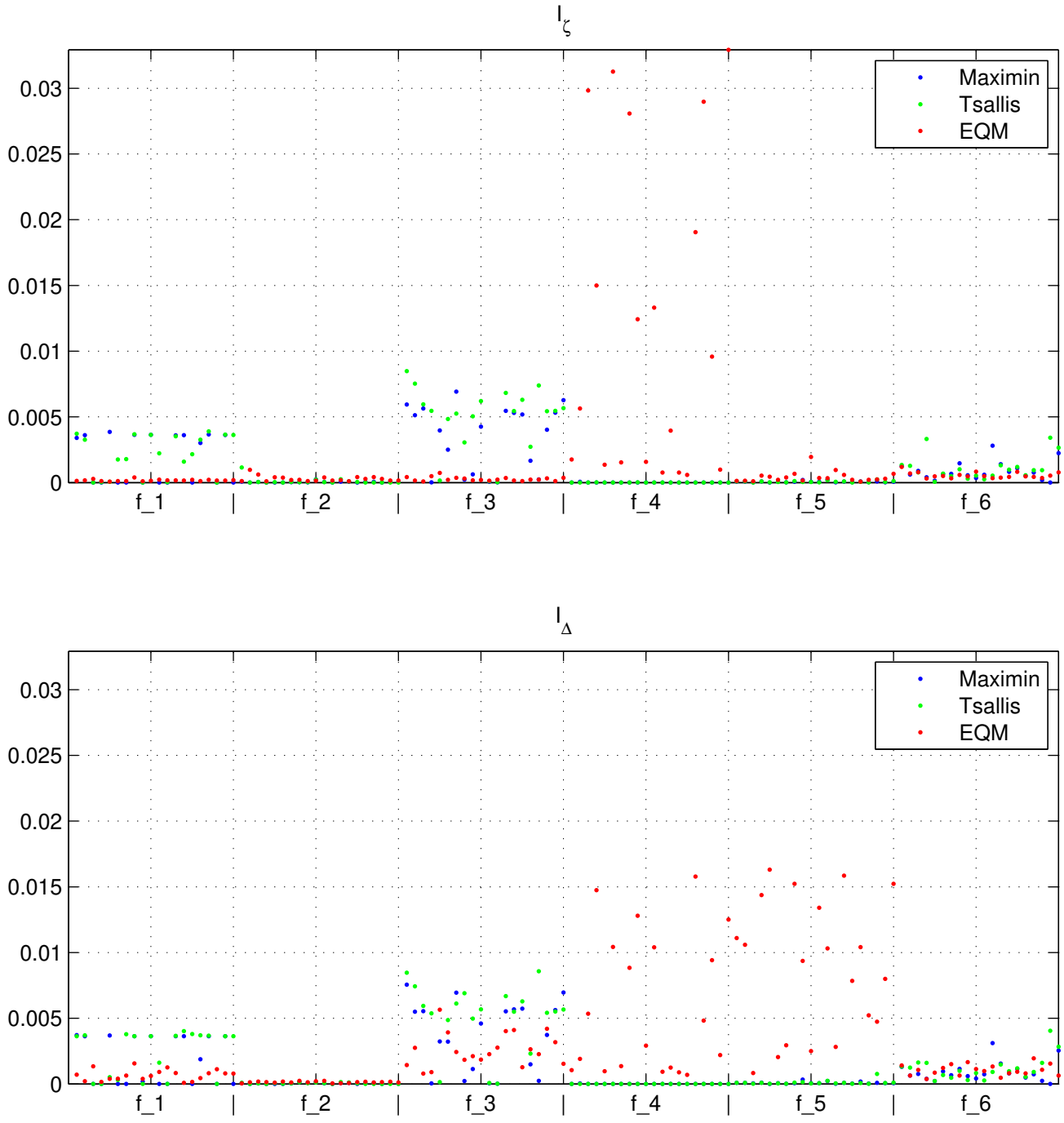

FIGURE 5.15 – Valeurs de I_ζ et I_Δ correspondant aux 120 plans finaux obtenus par le critère maximin (en bleu), de Tsallis (en vert) et de variance de krigeage (rouge). Pour chaque fonction-test, les valeurs sont ordonnées dans l'ordre croissant des plans (de P_1 à P_{20}).

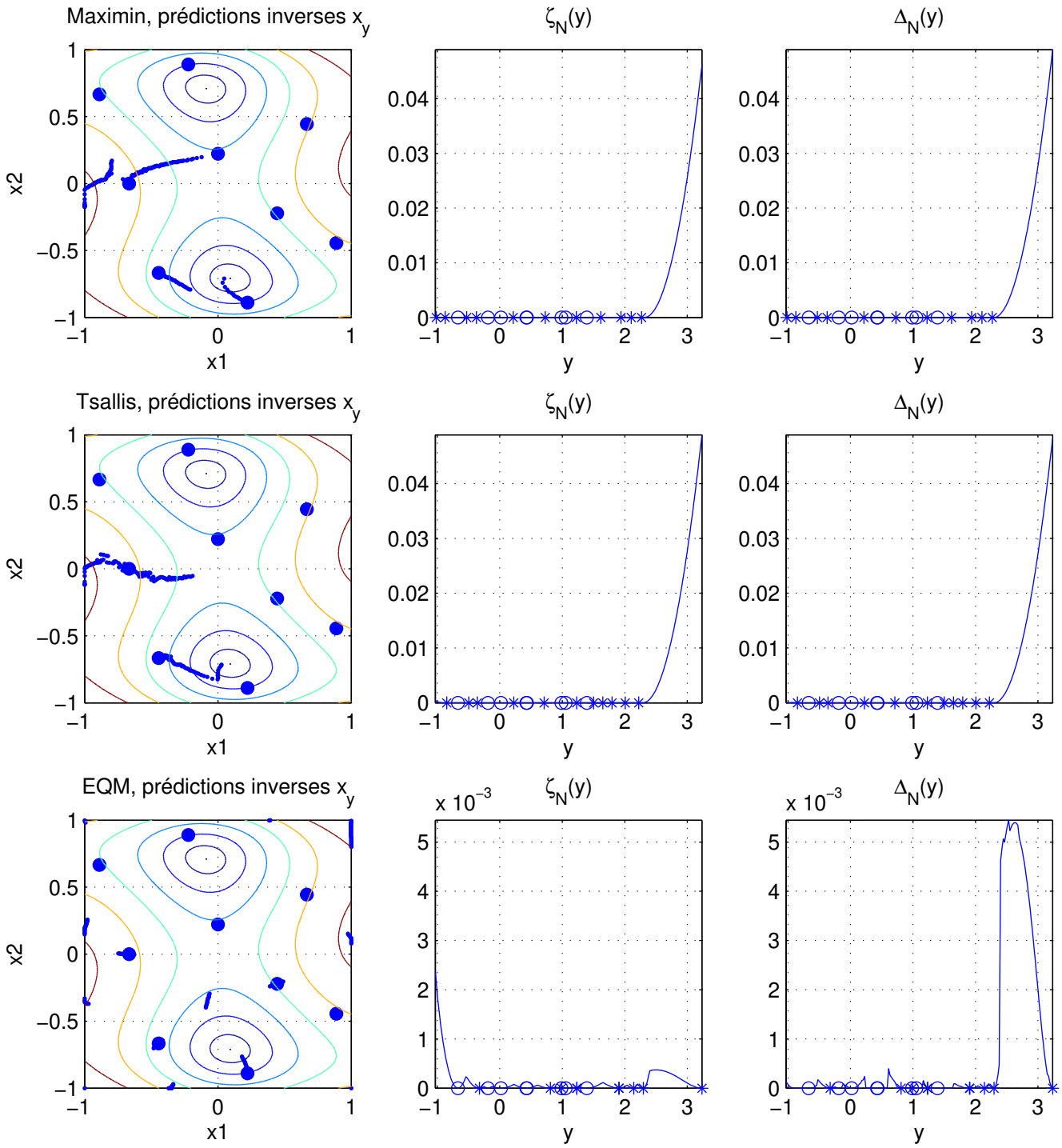

FIGURE 5.16 – Inversion du système à partir du plan final obtenu pour le critère maximin avec bornes connues (haut), Tsallis avec $h = 10^{-2}$ (milieu), et variance de krigeage (bas), avec la fonction-test f_1 et le plan initial P_1.

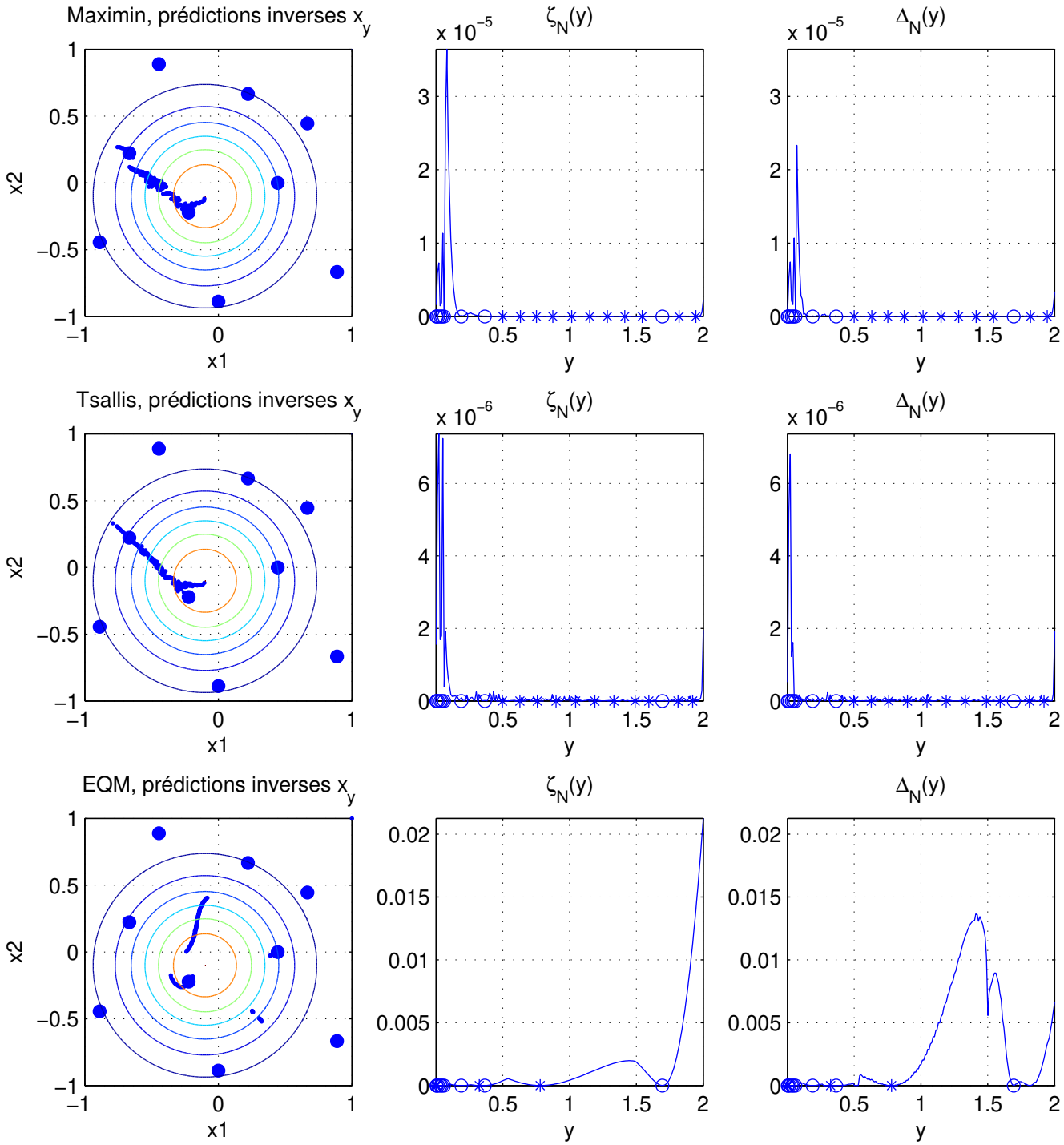

FIGURE 5.17 – Inversion du système à partir du plan final obtenu pour le critère maximin avec bornes connues (haut), Tsallis avec $h = 10^{-2}$ (milieu), et variance de krigeage (bas), avec la fonction-test f_4 et le plan initial P_{10}.

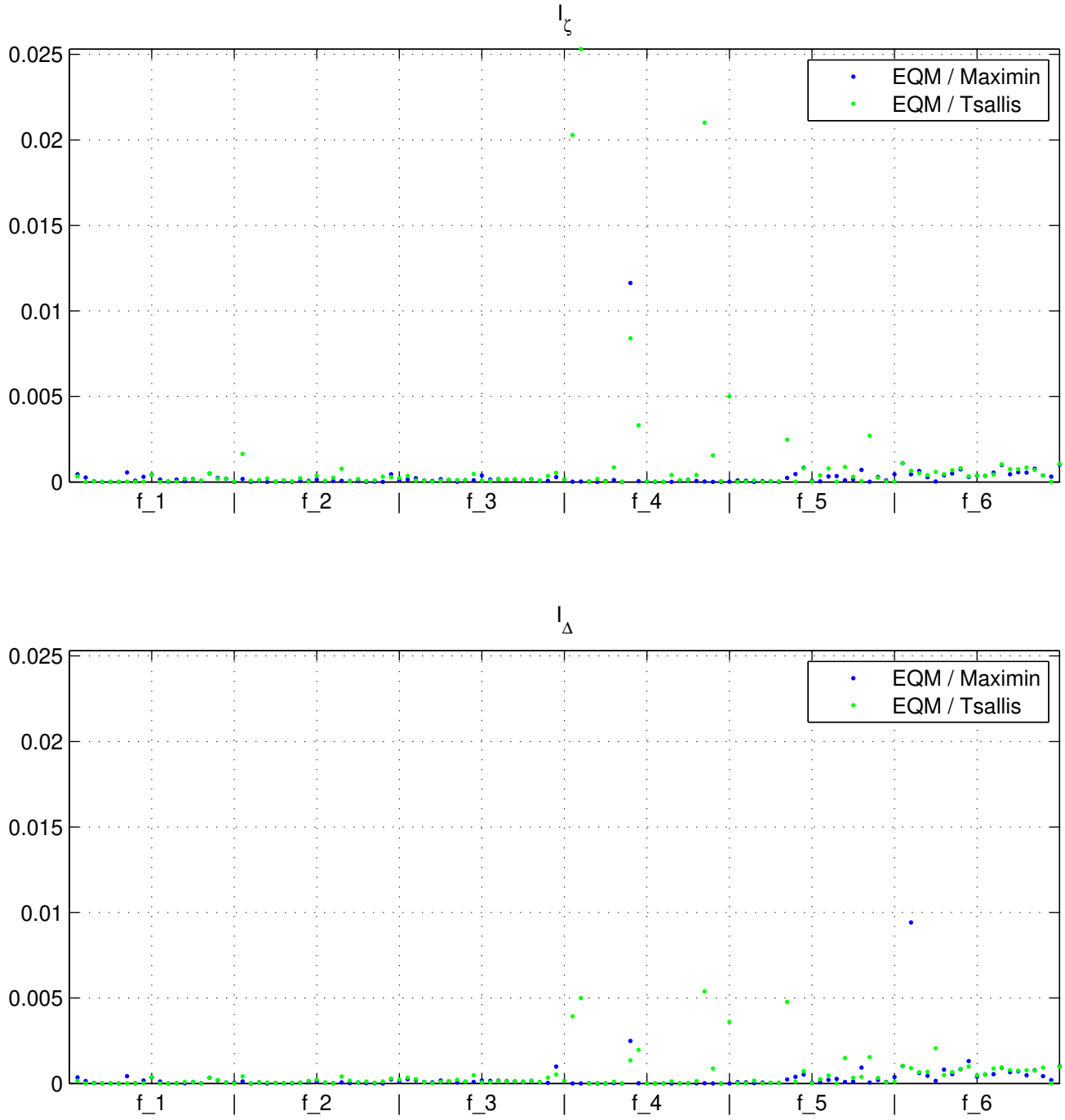

FIGURE 5.18 – Valeurs de I_ζ et I_Δ correspondant aux 120 plans finaux obtenus par le critère maximin (en bleu), de Tsallis (en vert), avec pour chaque critère un point sur deux ajouté au maximum de la variance de krigeage. Pour chaque fonction-test, les valeurs sont ordonnées dans l'ordre croissant des plans (de P_1 à P_{20}).

5.3 Cas où la fonction inconnue a 5 entrées et 2 sorties

Nous présentons maintenant l'algorithme tel qu'il est mis au point dans le cas de 5 entrées et 2 sorties. Afin de tester ses performances, une fonction, que l'on espère représentative du système physique, a été construite à partir d'un nombre limité de données obtenues sur un système très proche du système étudié. Nous allons lancer la procédure séquentielle sur cette fonction.

Le choix ainsi que la construction du plan initial retenu dans le cas de 5 facteurs d'entrée sont expliqués dans un premier temps. Puis, nous traitons de l'implémentation des critères de diversité dans l'algorithme ainsi que des particularités liées au fait qu'il y a maintenant 2 sorties. Finalement, nous donnons les résultats des tests effectués et constatons que l'approche simplifiée présentée en 4.2.3.4 donne en pratique des résultats satisfaisants pour les deux critères de diversité.

Notons que, pour des raisons de temps de calcul dues à l'augmentation des dimensions du problème, nous ne présentons pas de résultats liés à l'inversion du système tels que ceux du paragraphe 5.2.3 : les comparaisons des performances des critères se font en terme de répartition des sorties observées.

5.3.1 Fonction-test et plan initial

La fonction-test, construite à partir de données réelles, ne s'écrit pas sous une forme analytique simple. Nous illustrons ci-dessous ses variations afin de montrer sa complexité. Sur les figures suivantes sont représentées les courbes des sorties en fonction d'une entrée donnée, pour différentes valeurs du ratio Si/Al, les 3 entrées restantes étant fixées.

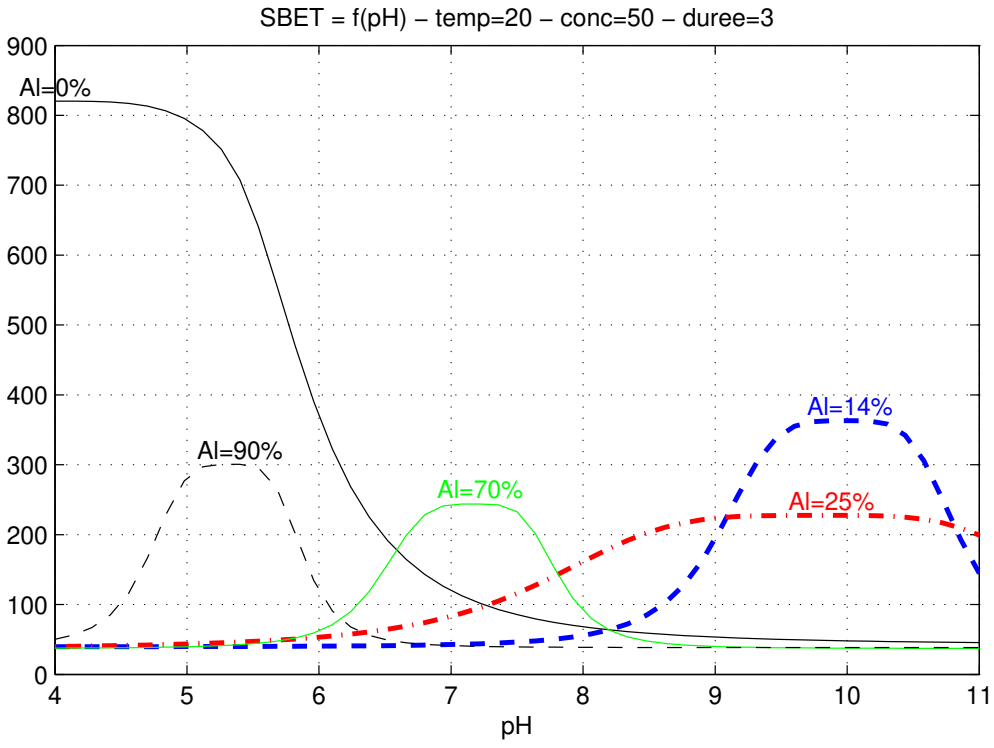

FIGURE 5.19 – Surface spécifique en fonction du pH, pour différentes valeurs du ratio Si/Al.

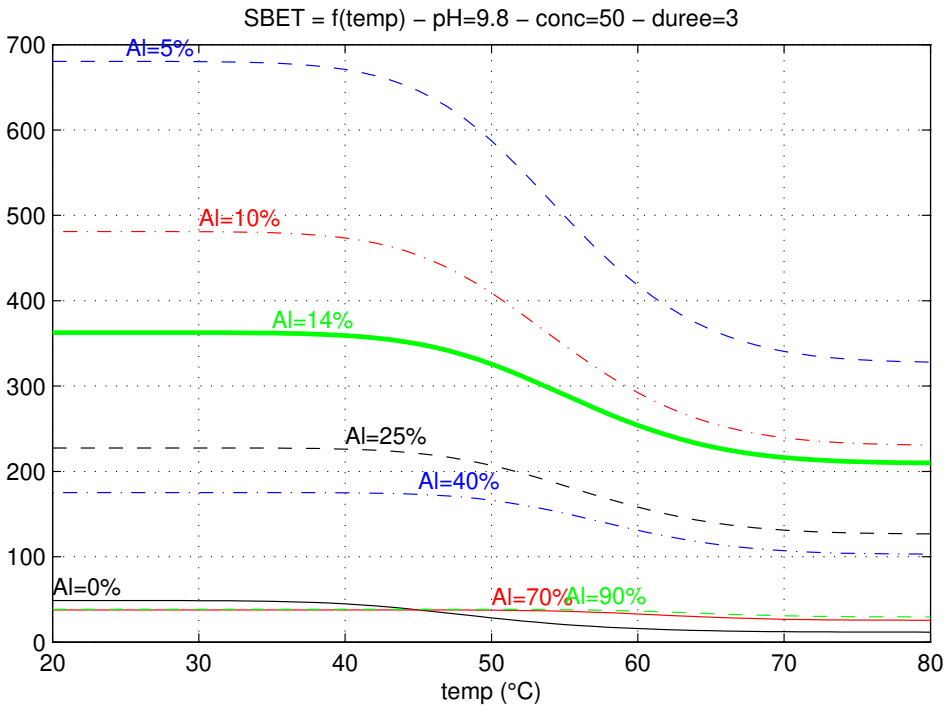

FIGURE 5.20 – Surface spécifique en fonction de la température, pour différentes valeurs du ratio Si/Al.

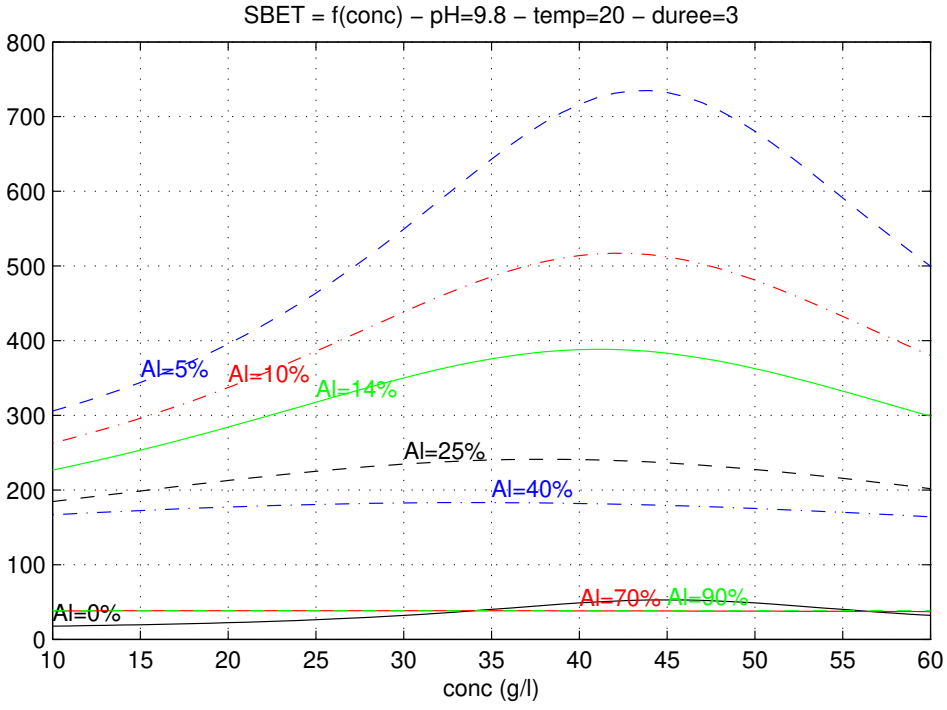

FIGURE 5.21 – Surface spécifique en fonction de la concentration, pour différentes valeurs du ratio Si/Al.

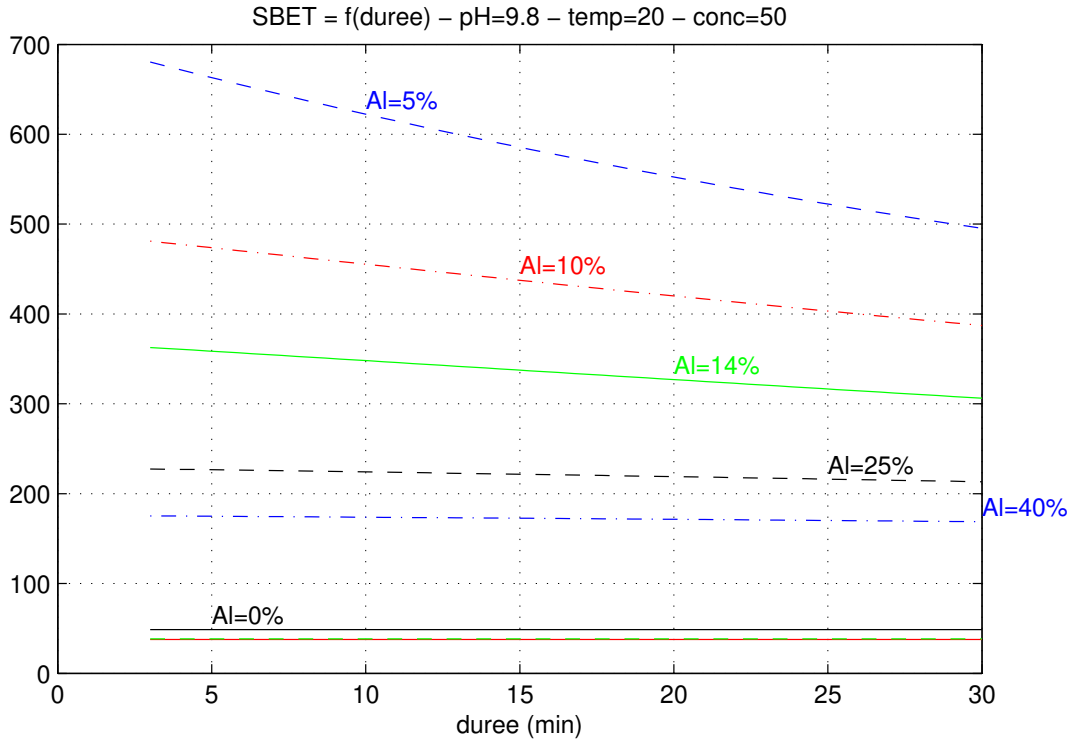

FIGURE 5.22 – Surface spécifique en fonction de la durée d'ajout, pour différentes valeurs du ratio Si/Al.

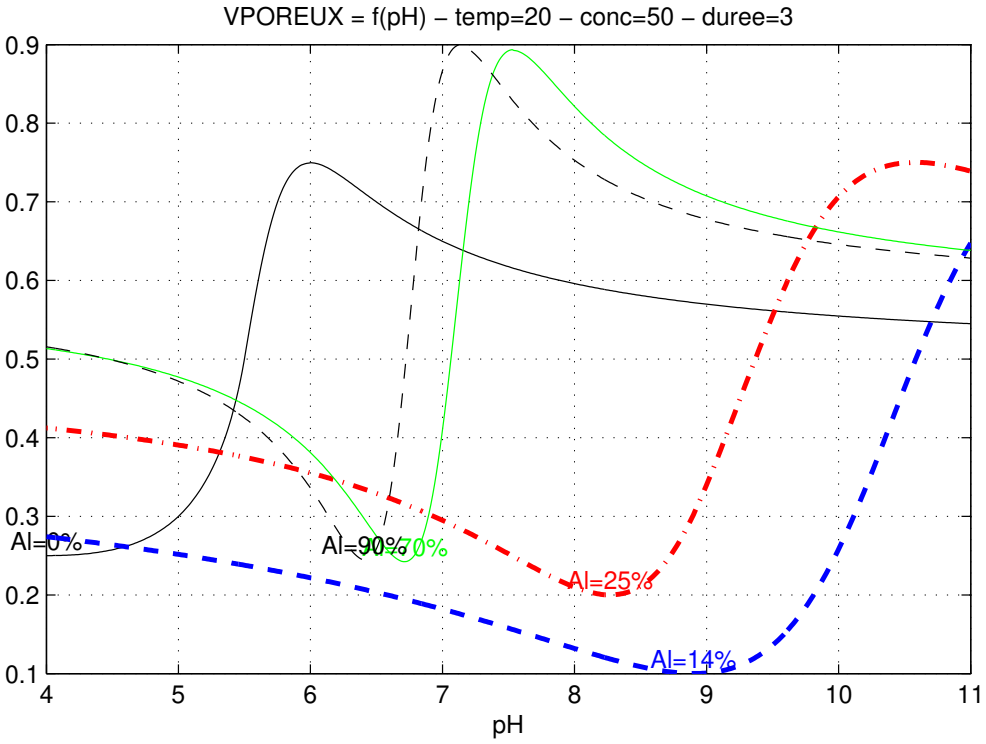

FIGURE 5.23 – Volume mésoporeux en fonction du pH, pour différentes valeurs du ratio Si/Al.

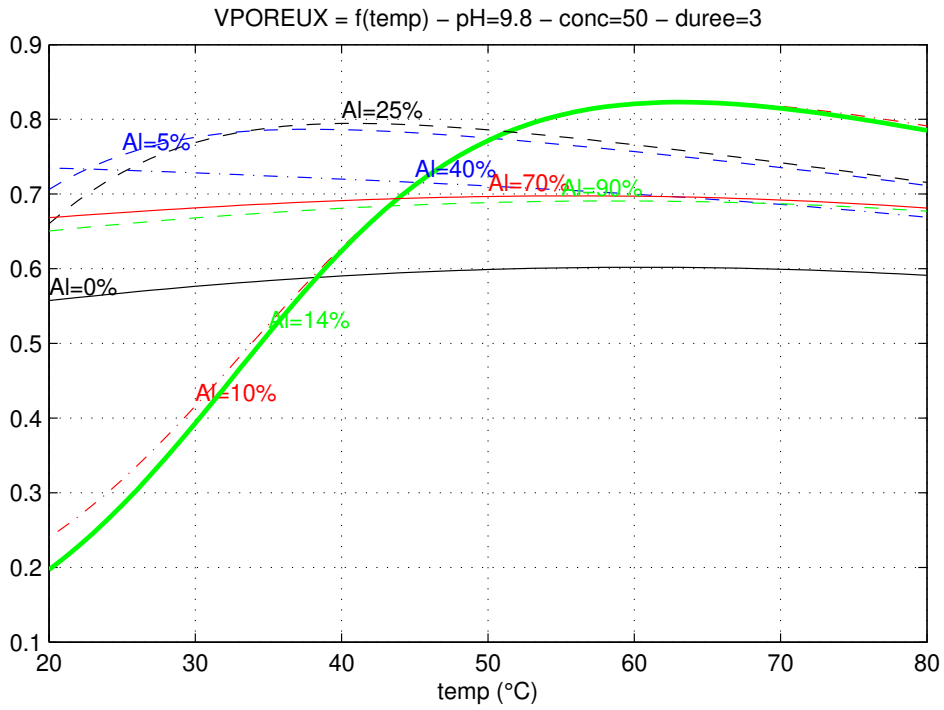

FIGURE 5.24 – Volume mésoporeux en fonction de la température, pour différentes valeurs du ratio Si/Al.

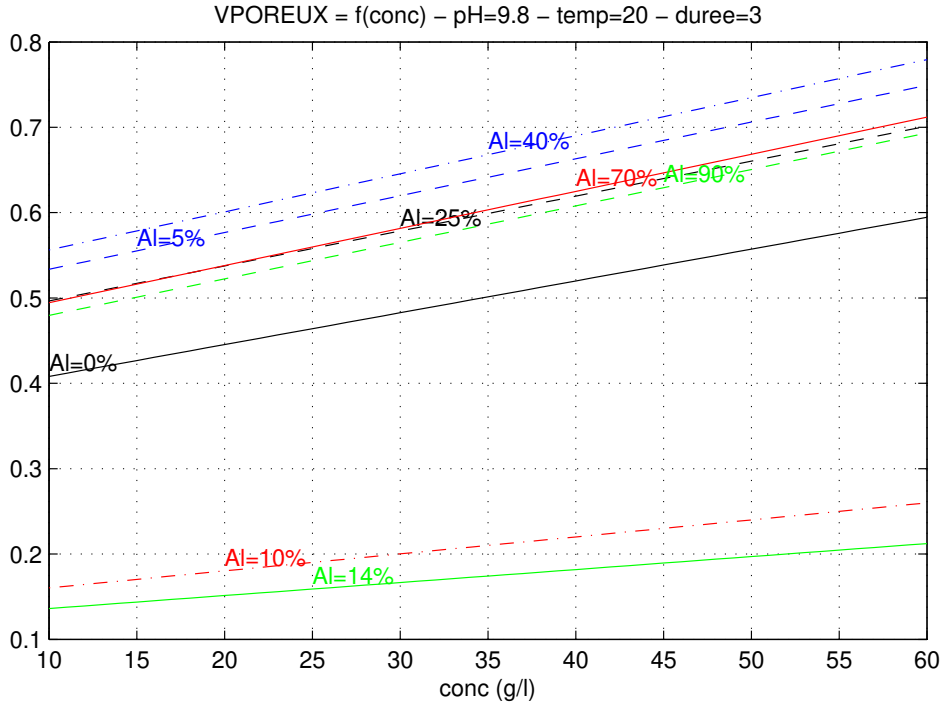

FIGURE 5.25 – Volume mésoporeux en fonction de la concentration, pour différentes valeurs du ratio Si/Al.

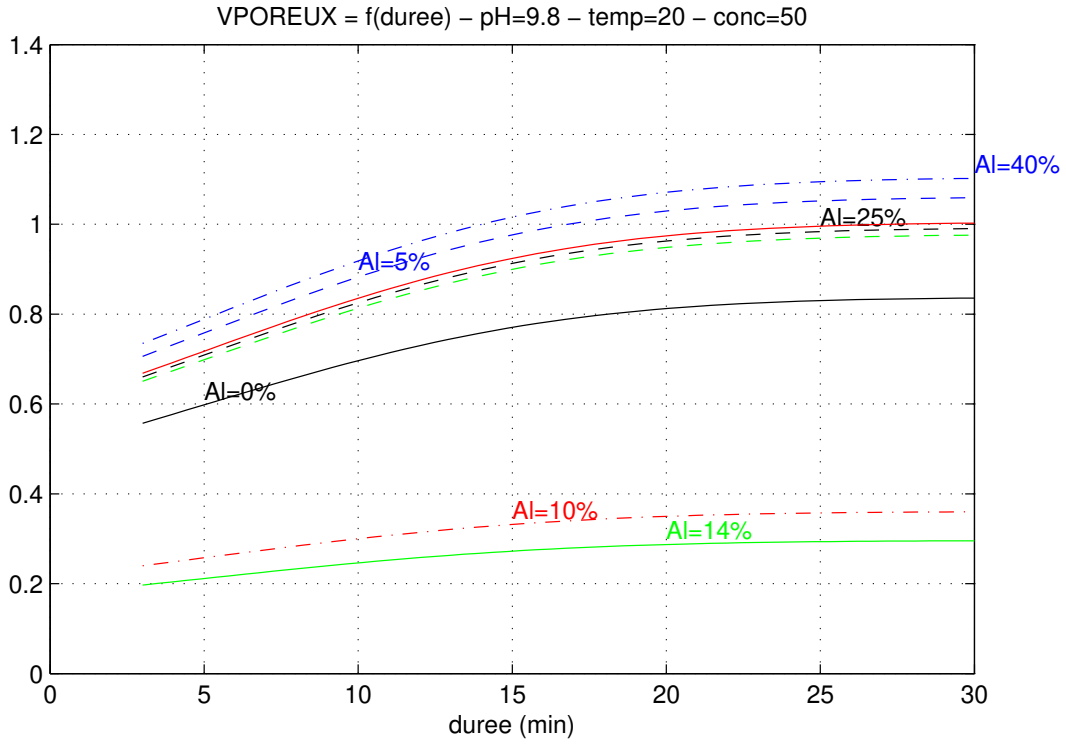

FIGURE 5.26 – Volume mésoporeux en fonction de la durée d'ajout, pour différentes valeurs du ratio Si/Al.

En ce qui concerne le choix du plan initial utilisé pour les tests (qui sera ensuite utilisé comme première série de mesures du système physique), nous nous intéressons tout d'abord au nombre de points qu'il doit contenir. Lors d'une procédure d'optimisation séquentielle, le choix de la taille du plan initial est importante : il s'agit de ne pas faire trop d'essais au départ, afin de ne pas « gâcher » des expériences que l'on pourrait faire de façon plus avisée ensuite en utilisant le modèle, mais il importe tout de même de disposer d'une quantité suffisante de points afin que le modèle initial ne soit pas trop mauvais. Sachant que l'on va disposer d'un budget d'environ 500 essais au total, le choix d'une cinquantaine de points a été choisie pour le plan initial.

Ayant remarqué que le modèle de krigeage est meilleur si l'on connaît les valeurs du système aux coins du domaine expérimental, il a été décidé d'inclure tous les coins du cube de dimension 5 au plan initial, ce qui fait déjà $2^5 = 32$ points. Les 16 points restants sont définis en s'inspirant des tableaux orthogonaux (§ 3.1.1.2) : nous souhaitons construire un plan dont les projections d-dimensionnelles soient toutes identiques, pour $d = 1, \ldots, 4$. Si l'on ramène le domaine expérimental à $[0, 1]^5$, les 16 points ajoutés aux 32 coins sont les suivants,

$$\frac{1}{3} \times \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 2 & 2 \\ 1 & 1 & 2 & 1 & 2 \\ 1 & 1 & 2 & 2 & 1 \\ 1 & 2 & 1 & 2 & 2 \\ 1 & 2 & 1 & 2 & 1 \\ 1 & 2 & 2 & 1 & 1 \\ 1 & 2 & 2 & 2 & 2 \\ 2 & 1 & 1 & 1 & 2 \\ 2 & 1 & 1 & 2 & 1 \\ 2 & 1 & 2 & 1 & 1 \\ 2 & 1 & 2 & 2 & 2 \\ 2 & 2 & 1 & 1 & 1 \\ 2 & 2 & 1 & 2 & 2 \\ 2 & 2 & 2 & 1 & 2 \\ 2 & 2 & 2 & 2 & 1 \end{pmatrix}.$$

Les projections selon 2 et 3 coordonnées du plan ci-dessus complété avec les 32 coins sont représentées sur la figure (5.27).

Remarque 5.3.1 *Le plan dual, obtenu en échangeant les valeurs* $1/3$ *et* $2/3$*, se projette de façon identique.*

Nous disposons donc d'un plan initial contenant 48 points.

5.3.2 Tests

Nous allons maintenant comparer les critères de diversité dans les dimensions du problème réel, en utilisant la fonction-test et le plan initial présentés précédemment. Rappelons qu'en pratique, les observations sont entâchées de bruit de mesure, se font par séries de 6 et les résultats des mesures arrivent avec un retard d'une série.

Nous présentons dans un premier temps les modifications de l'algorithme effectuées pour prendre en compte cette nouvelle configuration. Puis, nous comparons les deux critères de diversité, en ajoutant tout d'abord les points un par un sans bruit de mesure, puis avec bruit, et

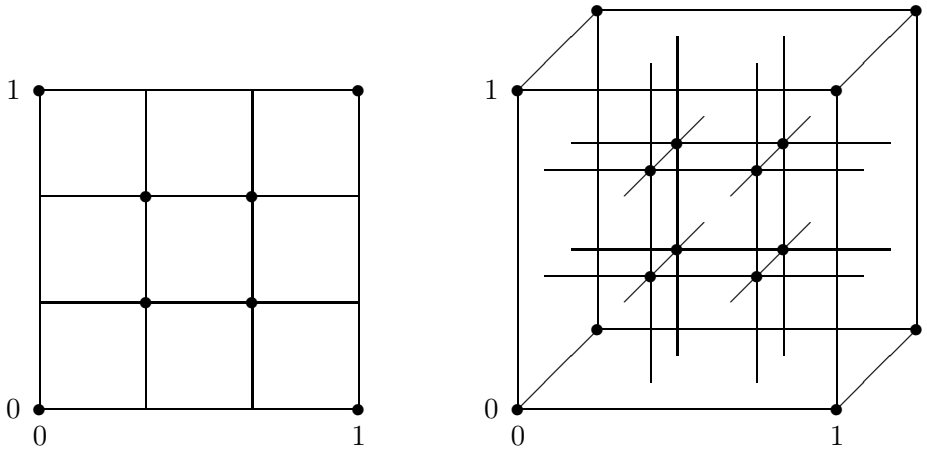

FIGURE 5.27 – Projections selon 2 et 3 facteurs des points du plan initial.

enfin par séries de 6 avec retard. En raison de l'augmentation du temps de calcul, nous aurons recours à l'approche simplifiée présentée au paragraphe 4.2.3.4

5.3.2.1 Modifications de l'algorithme

L'algorithme a dû être adapté au cas de 5 entrées et 2 sorties, car l'optimisation se fait maintenant sur un plus grand nombre de points, et les calculs de diversité se font différemment dans le cas de deux sorties pour le critère maximin (pas de formule analytique).

Discrétisation du domaine d'entrée. Nous avons dans un premier temps essayé de faire de l'optimisation globale du critère de diversité en utilisant les polyèdres de Delaunay (§ J.1) : des optimisations locales sont effectuées avec pour valeurs initiales les centres des polyèdres de Delaunay construits à partir de l'ensemble des observations, puis le meilleur optimum local obtenu est considéré comme réalisant l'optimum global du critère. Cependant, cette façon de procéder est très coûteuse en temps de calcul dans le cas de 5 entrées, car le nombre de polyèdres de Delaunay augmente très rapidement avec le nombre de points du plan (voir la figure J.2, où est tracée l'évolution du nombre de polyèdres en fonction du nombre de points), et donc avec lui le nombre d'optimisations locales à effectuer à chaque ajout de point. Afin de remédier à cette difficulté, que l'on peut rattacher au *fléau de la dimension* [92], nous nous sommes résolus à discrétiser le domaine d'entrée en une grille $\mathcal{G}_{\mathcal{X}}$. Ainsi, à chaque itération de l'algorithme, c'est simplement le point de la grille optimisant le critère de diversité qui est ajouté au plan courant. La grille a été choisie après discussion avec l'expérimentateur, qui s'est basé sur sa connaissance du système ainsi que sur la précision avec laquelle il peut obtenir les valeurs de facteurs d'entrée : la grille contient $13 \times 17 \times 13 \times 6 \times 5 = 86190$ points.

Calcul des prédictions par krigeage. Le calcul de la moyenne et de la variance de krigeage aux points de la grille se fait vectoriellement par appel de la routine predictor de DACE, qui a pour argument d'entrée l'ensemble des points de la grille. L'augmentation du nombre de points par rapport au cas de 2 entrées fait que le temps de calcul augmente de façon notable. Afin de contourner ce problème, nous avons choisi d'appeler séquentiellement la routine predictor avec pour argument d'entrée des sous-ensembles disjoints de points de la grille, dont la taille a dû

être choisie. Faire les calculs point par point étant coûteux en raison du grand nombre d'appels de la routine, il a fallu trouver un compromis. Nous avons remarqué que prendre des séries de 1000 points est un compromis raisonnable (aller au-delà de 1000 points à la fois rallonge le temps de calcul). Les 2 sorties sont modélisées de façon indépendante, ce qui signifie qu'un modèle de krigeage est construit séparément pour chacune des deux sorties.

Inclusion de bruit de mesure. Il est possible en théorie d'inclure et d'estimer un bruit de mesure dans le modèle de krigeage (voir le § 2.4). En pratique, le nombre de données d'observation relativement faible ainsi que la complexité du système étudié font qu'il est peu raisonnable d'espérer obtenir une estimation précise du bruit. En conséquence, la variance du bruit de mesure a été estimée dans un premier temps : un modèle quadratique de la variance des réponses a été construit en se servant de l'expertise de l'expérimentateur, puis ses paramètres ont été estimés en utilisant un plan d'expériences. Le bruit de mesure est donc supposé connu.

La routine DACE a été modifiée pour pouvoir inclure un bruit de mesure connu dans le modèle de krigeage, et estimer les paramètres en remplaçant les formules du maximum de vraisemblance par celles données au paragraphe 2.4. Le reste de la routine n'a pas été modifié.

Calcul de la diversité pour le critère maximin. Pour l'évaluation du critère de diversité (4.16), la méthode de calcul proposée basée sur une évaluation numérique à l'intérieur des cellules de Voronoi s'est révélée trop coûteuse en temps de calcul. En pratique, le domaine des sorties (supposé connu) est discrétisé en une grille de taille 100×100, et l'intégrale (4.16) est évaluée en prenant la moyenne de l'intégrande sur les points de la grille. La précision de la méthode n'a pas été évaluée, mais nous avons remarqué qu'en pratique l'optimum obtenu par cette méthode est le même que celui obtenu à partir de l'intégration numérique, et le gain en temps de calcul est conséquent. L'algorithme Matlab appelle une routine Fortran qui effectue le calcul ci-dessus, ce qui accélére encore le temps de calcul.

5.3.2.2 Comparaison des résultats selon le critère utilisé

Nous testons maintenant les critères de diversité dans les conditions réelles. Afin de comparer l'efficacité des critères dans le cas de 2 sorties, les points sont tout d'abord ajoutés 1 par 1 au plan initial contenant 48 points jusqu'à atteindre 500 points, puis nous comparons la répartition finale des sorties et des entrées. Un bruit de mesure est ensuite incorporé au modèle de krigeage, et l'on constate qu'une grande proportion d'observations sont répétées, ce qui va nous conduire à utiliser la réinterpolation (§ 2.4.2) afin de garantir que les observations ne seront pas répétées. Nous comparons finalement les deux critères dans le cas bruité où les observations se font par paquets de 6 et les résultats des mesures arrivent avec retard.

Points ajoutés 1 par 1 sans bruit de mesure. Les 500 valeurs de sortie du plan final sont représentées sur la figure 5.28, pour le critère maximin (haut) et le critère de Tsallis (bas), où les cercles pleins représentent les 48 points du plan initial, les croix les 452 points ajoutés, et le fond gris est le domaine de sortie $f(\mathcal{G}_{\mathcal{X}})$ atteignable par les points de la grille d'entrée (on ne peut pas espérer obtenir des valeurs en-dehors de cette zone). Comme on peut le constater, le critère maximin répartit plutôt bien les points, à part dans la zone située à droite, mais la fonction étant très compliquée, le modèle de krigeage est imprécis. L'ensemble du domaine est exploré, mais rappelons que les bornes du domaine atteignable sont supposées connues. Le critère de Tsallis a été utilisé avec une valeur du paramètre $h = 10^{-2}$ afin de s'assurer qu'il soit bien discriminant, en utilisant la formule de la remarque 4.2.3. Nous pouvons observer dans la partie basse de la figure que les points ajoutés avec le critère de Tsallis semblent moins bien répartis,

et surtout sont très concentrés dans la partie basse du domaine où il y a déjà de nombreuses observations du plan initial (cercles pleins). Bien que les bornes du domaine atteignable en sortie soient supposées inconnues ici, le critère a bien exploré les bords du domaine ce qui est très satisfaisant. Le critère de Tsallis a l'avantage d'être plus économe en temps de calcul : en effet, le temps de calcul total est d'environ 34h10min pour le critère maximin (le dernier point a été ajouté en 8min8sec), alors que le temps de calcul total est de 18h12min dans le cas du critère de Tsallis (dernier point ajouté en 5min21s).

Comparons alors le placement des facteurs d'entrée correspondants, représentés sur la figure 5.29 et la figure 5.30. Le graphique de coordonnées (i, j) correspond dans les deux cas au facteur x_i en abcisse contre le facteur x_j en ordonnée (ce qui explique la répartition diagonale des points sur les graphiques de coordonnées (i, i)). Les cercles sont les points du plan initial, les étoiles les points ajoutés. On remarque que le nombre des facteurs explorés est limité dans les deux cas (il n'y a par exemple aucune observation correspondant à $x_4 = 20$). Pour les 2 derniers facteurs x_4 et x_5, le petit nombre de niveaux observés s'explique en partie par la discrétisation grossière (6 et 5 niveaux respectivement).

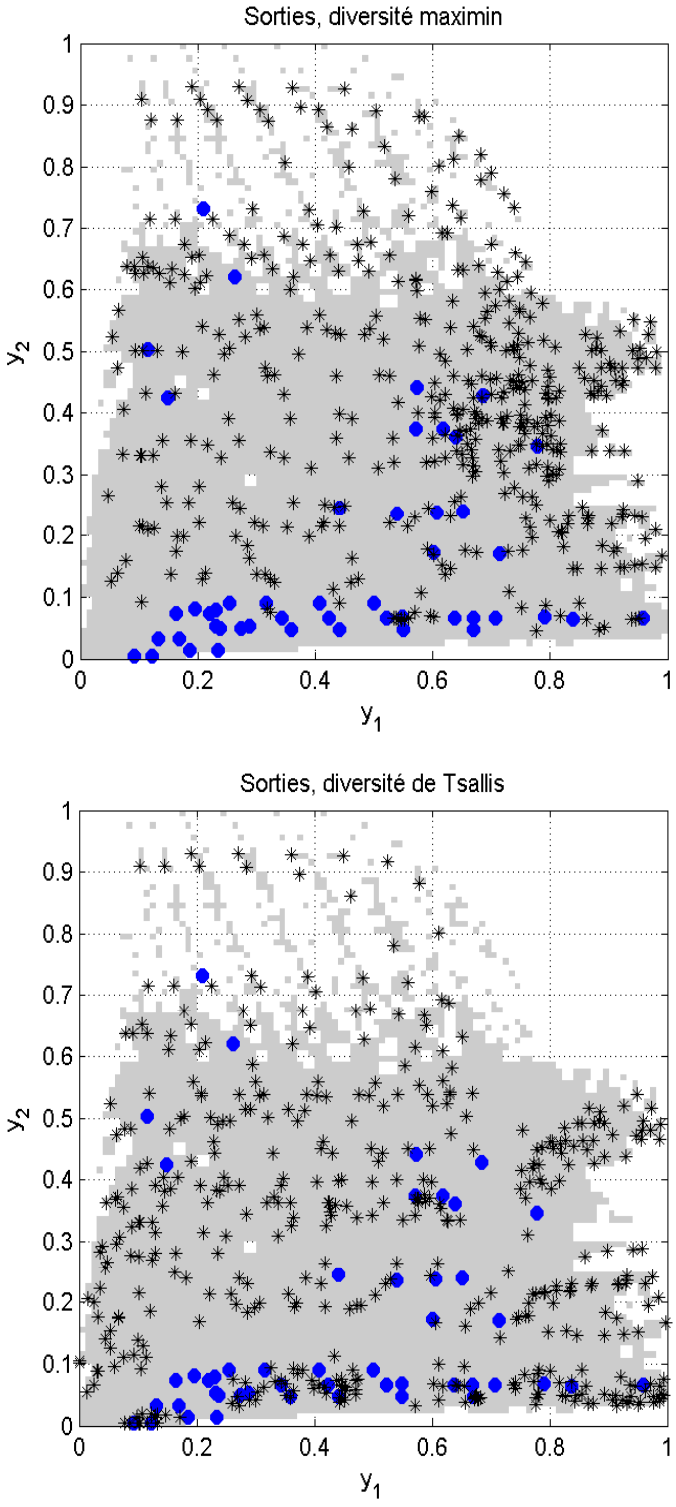

FIGURE 5.28 – Sorties obtenues par les critères de diversité maximin et de Tsallis.

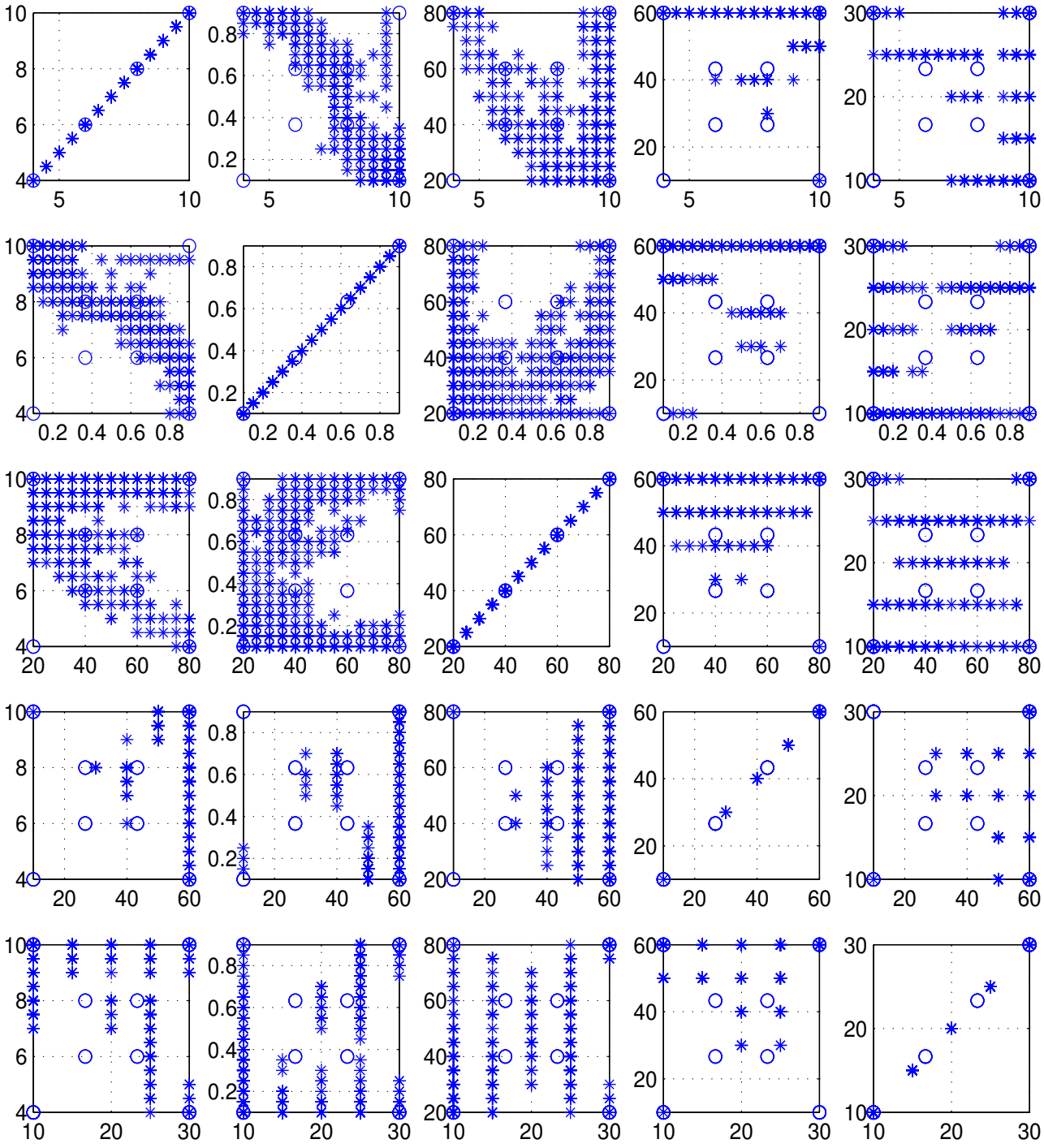

FIGURE 5.29 – Placement des entrées, critère de diversité maximin.

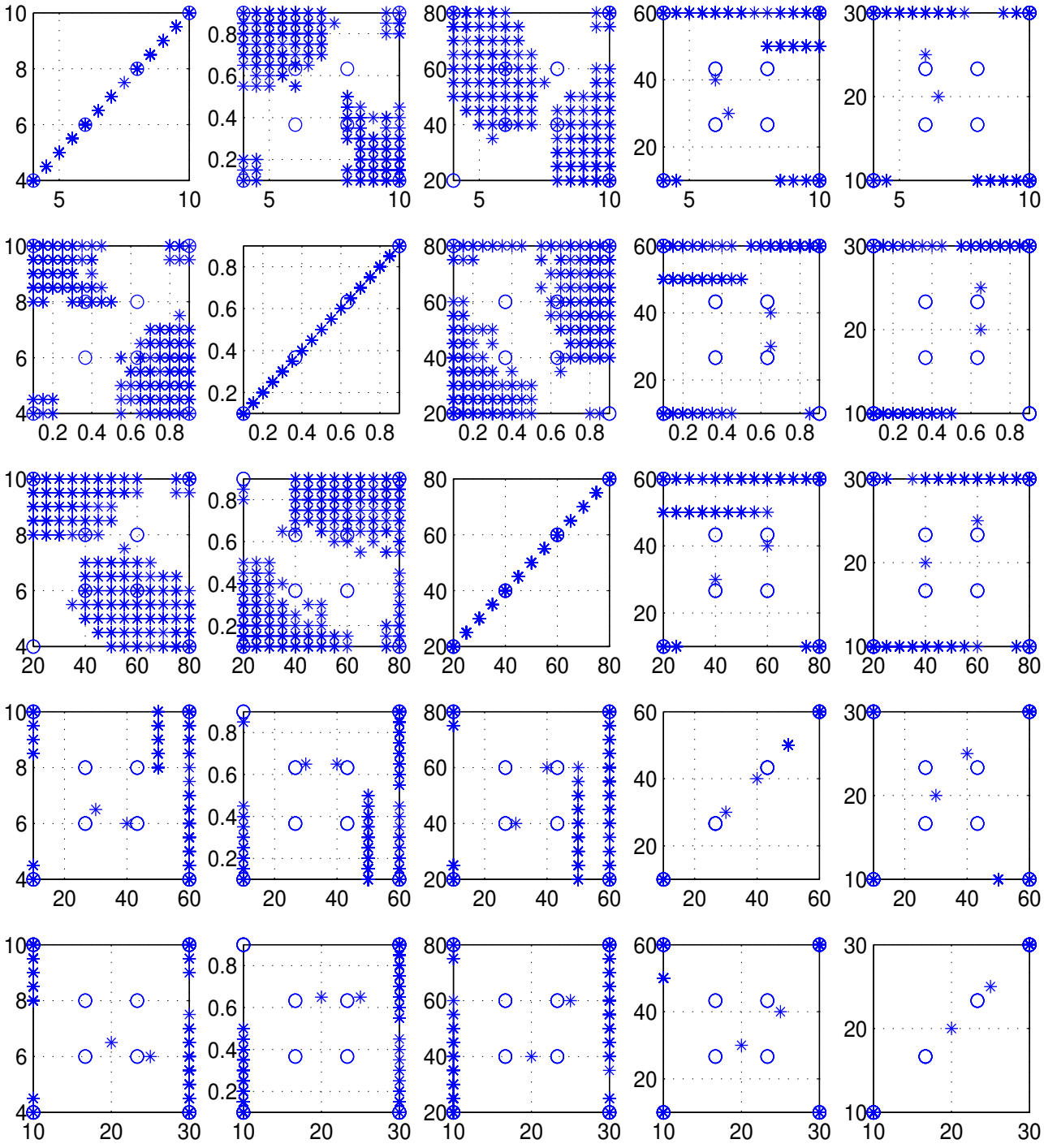

FIGURE 5.30 – Placement des entrées, critère de diversité de Tsallis.

Nous avons ensuite testé une détermination empirique du pas h pour le critère de Tsallis dans le cas de 2 sorties. S'inspirant du pas empirique dans le cas d'une seule sortie, les termes de la matrice Σ du corollaire 4.2.5 sont estimés en utilisant la formule empirique (4.7) pour chacune des deux sorties,

$$\sigma_j = \max_{i=1,\dots,n-1} \frac{y^j_{(i+1)} - y^j_{(i)}}{6}, \qquad j = 1, 2,$$

avec $y^j_{(i)}$ la i^e plus grande observation de la sortie numéro j. Les sorties du plan final correspondant sont représentées sur la figure 5.31, et l'on constate que les résultats sont moins bons que sur la figure 5.28 car les grandes valeurs de y_2 ne sont pas explorées (la formule empirique n'est pas satisfaisante car les pas calculés ne tiennent pas compte des distances entre les points mais seulement des distances marginales entre leurs coordonnées).

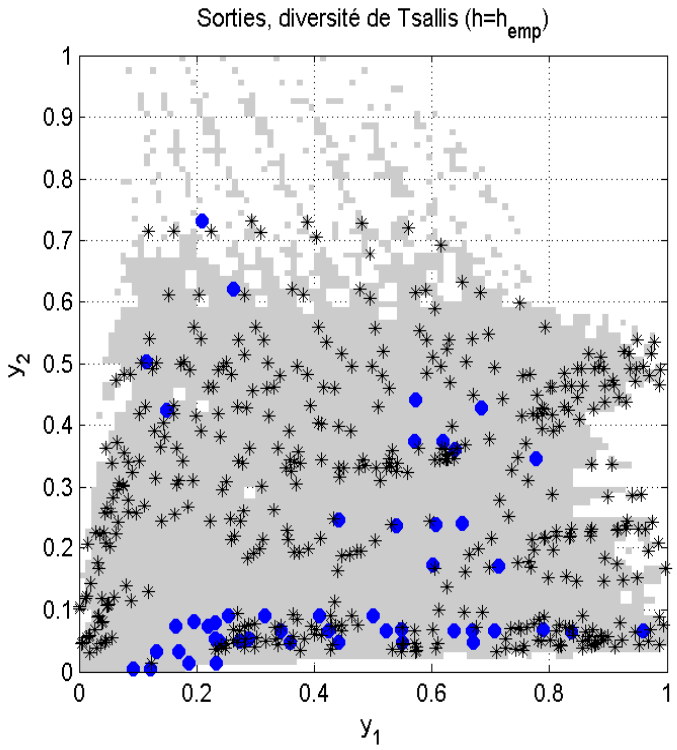

FIGURE 5.31 – Sorties obtenues par diversité de Tsallis et h_{emp}.

Points ajoutés 1 par 1 avec bruit de mesure. En ajoutant systématiquement aux observations de la fonction-test un bruit gaussien d'écart-type égal à 5% de la plage de variation de chaque sortie, et utilisant un modèle de krigeage avec inclusion de ce bruit supposé connu, nous avons obtenu avec les critères maximin et de Tsallis une liste de 500 points représentés sur la figure 5.32, où les points du plan initial sont représentés par des cercles pleins et les points ajoutés ensuite par des étoiles (le domaine atteignable étant le fond gris).

Nous constatons qu'il y a un grand nombre de répétitions : 336 observations ont été faîtes en double sur les 500 dont nous disposons pour le critère maximin (ce qui ne laisse que 164 points différents dans le plan final), 219 sont en double pour le critère de Tsallis (ce qui ne laisse que 281 points différents dans le plan final). Nous pensons que cela n'est pas acceptable pour espérer obtenir un modèle relativement fiable du système, et appliquons donc la technique de réinterpolation présentée en 2.4.2 : à chaque itération, le modèle de krigeage obtenu est modifié de sorte que la variance de prédiction soit nulle aux observations, garantissant ainsi la non-répétition des observations. Les plans finaux obtenus sont représentés sur la figure 5.33. Les sorties ne sont alors pas aussi bien réparties que dans le cas sans bruit de la figure 5.28 (ce qui est raisonnable), mais on observe, pour le critère maximin, une nette amélioration dans le remplissage de l'espace par rapport au cas bruité sans réinterpolation représenté sur la figure 5.32. Le critère de Tsallis produit un plan final globalement meilleur.

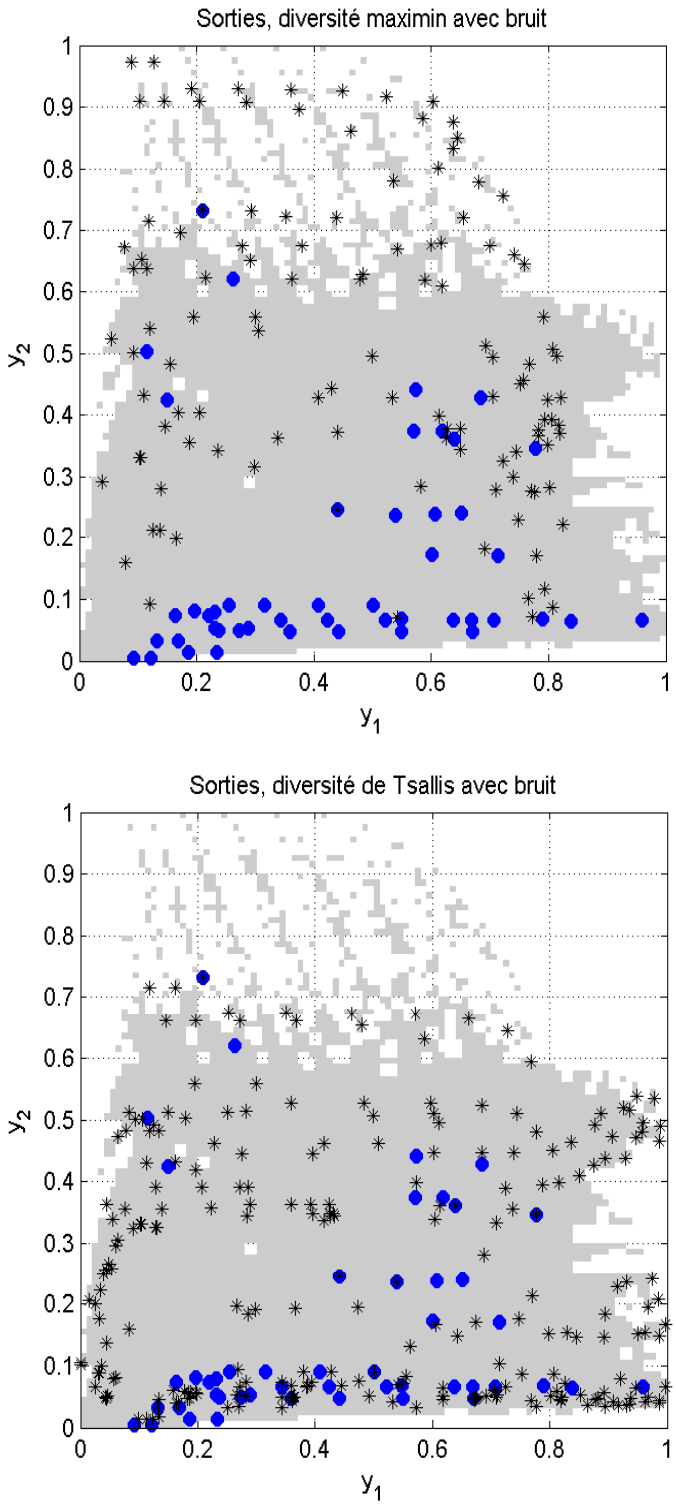

FIGURE 5.32 – Sorties obtenues par diversité maximin et de Tsallis avec un bruit de 5%.

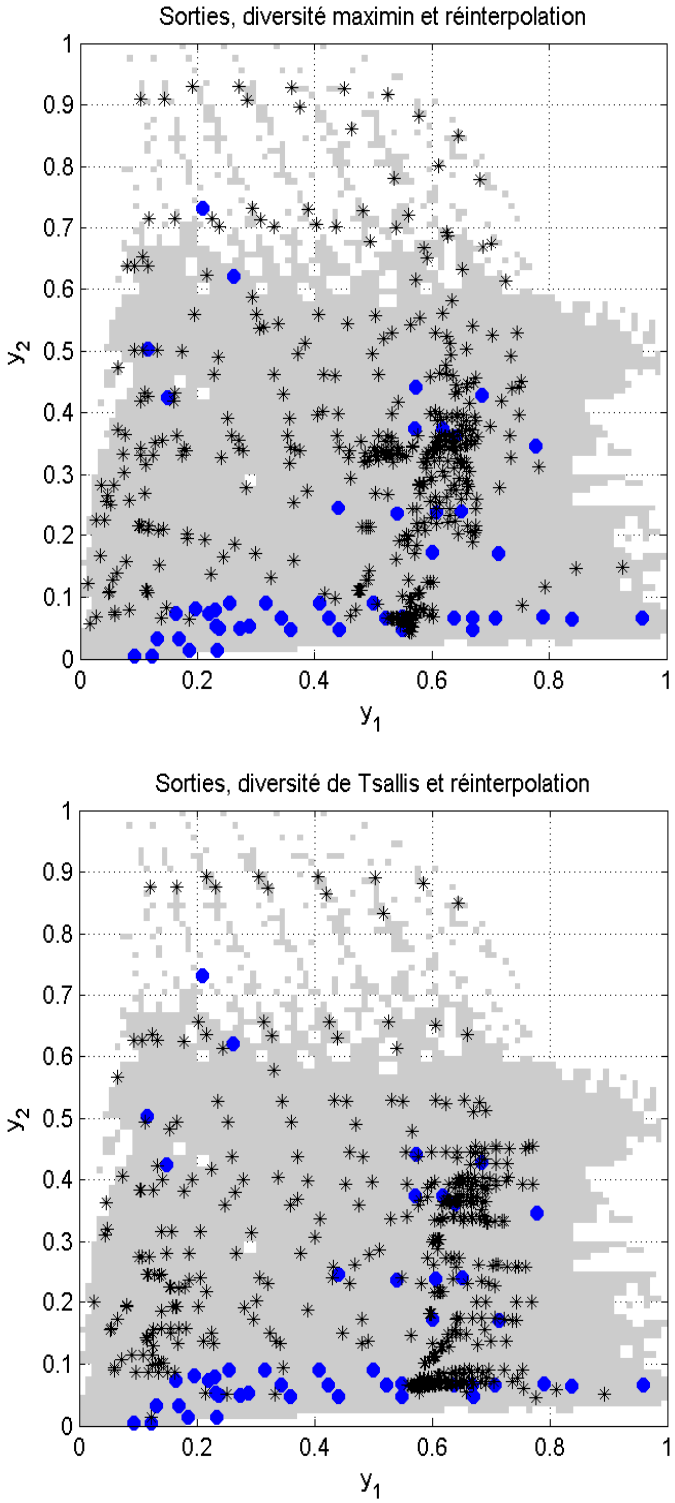

FIGURE 5.33 – Sorties obtenues par diversité maximin et de Tsallis, avec réinterpolation avec un bruit de 5%.

Points ajoutés 6 par 6, avec bruit de mesure et réinterpolation, et retard d'arrivée des mesures. Nous comparons finalement le critère maximin et le critère de Tsallis dans les conditions du problème réel, en utilisant l'approche simplifiée présentée au § 4.2.3. Afin de trouver les 6 meilleurs optima locaux du critère de diversité, un algorithme d'optimisations locales simultanées sur une grille a été utilisé : à partir d'un vecteur de points initiaux, l'algorithme effectue des optimisations locales dans les directions de plus grande pente et retourne le vecteur correspondant d'optima locaux. Ne pouvant utiliser comme points initiaux les centres des polyèdres de Delaunay en raison de leur nombre prohibitif dans le cas de 5 entrées (voir le § J.1, figure J.2), nous choisissons de prendre comme points initiaux de la recherche les points du plan courant. Nous avons considéré 75 étapes de planification et ajouté aux 48 points du plan initial 75 séries de 6 points, ce qui fait un total de $48 + 6 \times 75 = 498$ points. Les résultats obtenus par cette méthode sont présentés sur la figure 5.34, et sont très satisfaisants quel que soit le critère utilisé. Le plan final obtenu avec le critère maximin est représenté en haut, celui obtenu par le critère de Tsallis en bas. Dans les deux cas, les 48 points du plan initial sont les cercles pleins et les 450 points ajoutés les étoiles. Le domaine atteignable, en fond gris, est bien échantillonné dans les deux cas, avec les bords relativement bien atteints (les bords sont plus nettement atteints avec le critère maximin, mais celui-ci intégre la connaissance des bornes atteignables par les sorties). La répartition des points est plutôt satisfaisante si l'on considère le petit nombre d'essais à disposition, même si tout le domaine n'est pas atteint, et il y a des zones vides de points alors que d'autres contiennent une grande concentration de points. Le temps de calcul total est bien moins important (d'un facteur au moins 4) que dans le cas où les points étaient ajoutés 1 par 1, ce qui s'explique par le fait que l'on fait 6 fois moins d'optimisations. Pour le critère maximin, le temps de calcul total est de 6h46min (la dernière série de points ajoutée en 10min5s), alors que pour le critère de Tsallis le temps de calcul total est de 4h22s (la dernière série de points ajoutée en 8min53s). Remarquons cependant que la dernière série de 6 points est ajoutée en plus de temps que le dernier point dans le cas 1 par 1 (page 161), car le nombre d'optimisations locales à chaque itération augmente avec le nombre de points contenu dans le plan (une initialisation est effectuée en chaque point du plan courant) et devient à terme plus coûteux qu'un tri des valeurs du critère sur la grille des entrées.

En pratique, il est possible que l'on dispose d'un budget de moins de 500 mesures. C'est pourquoi le remplissage du domaine de sorties pour 100, 200, 300 et 400 points est représenté respectivement sur la figure 5.35 pour le critère maximin et sur la figure 5.36 pour le critère de Tsallis. On peut constater que le remplissage du domaine devient rapidement satisfaisant dans les deux cas.

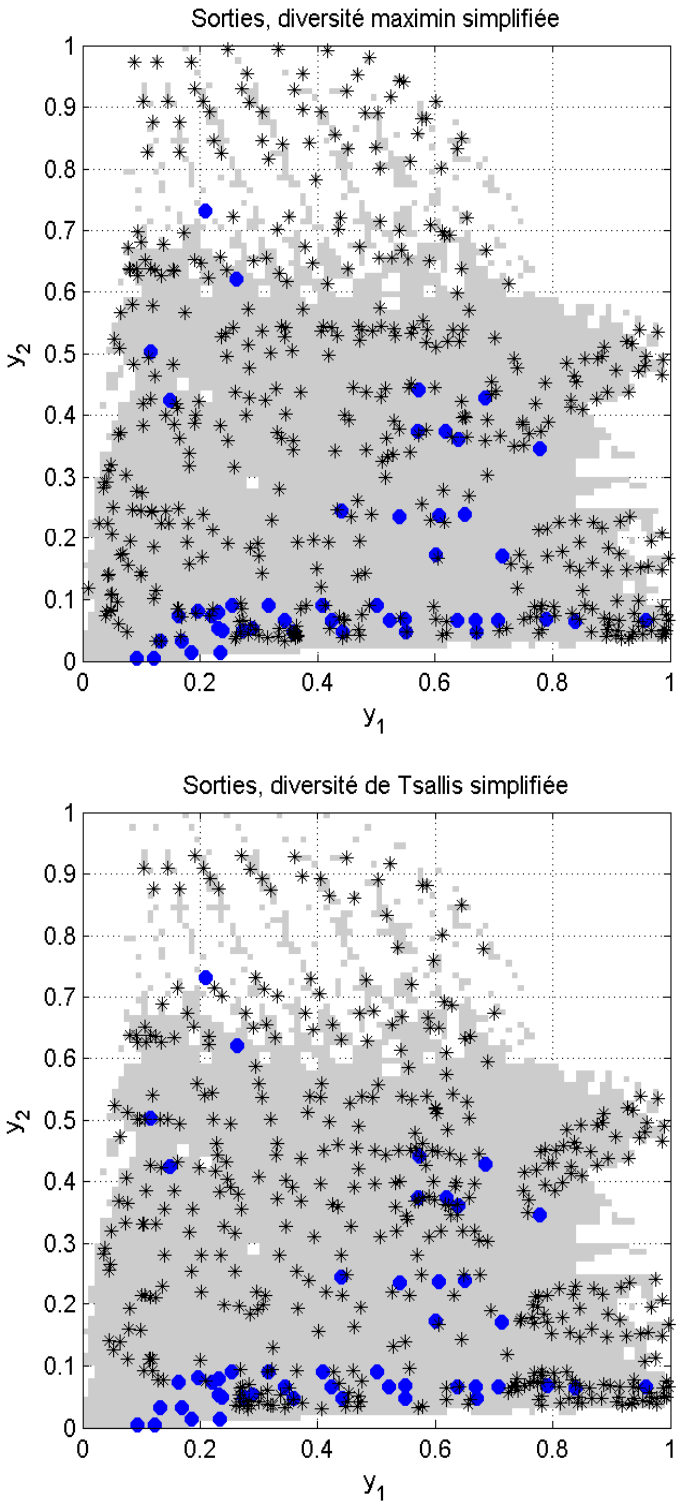

FIGURE 5.34 – Sorties obtenues par diversité maximin et de Tsallis, avec bruit de mesure de 5%, essais par séries de 6 et retard d'arrivée des résultats.

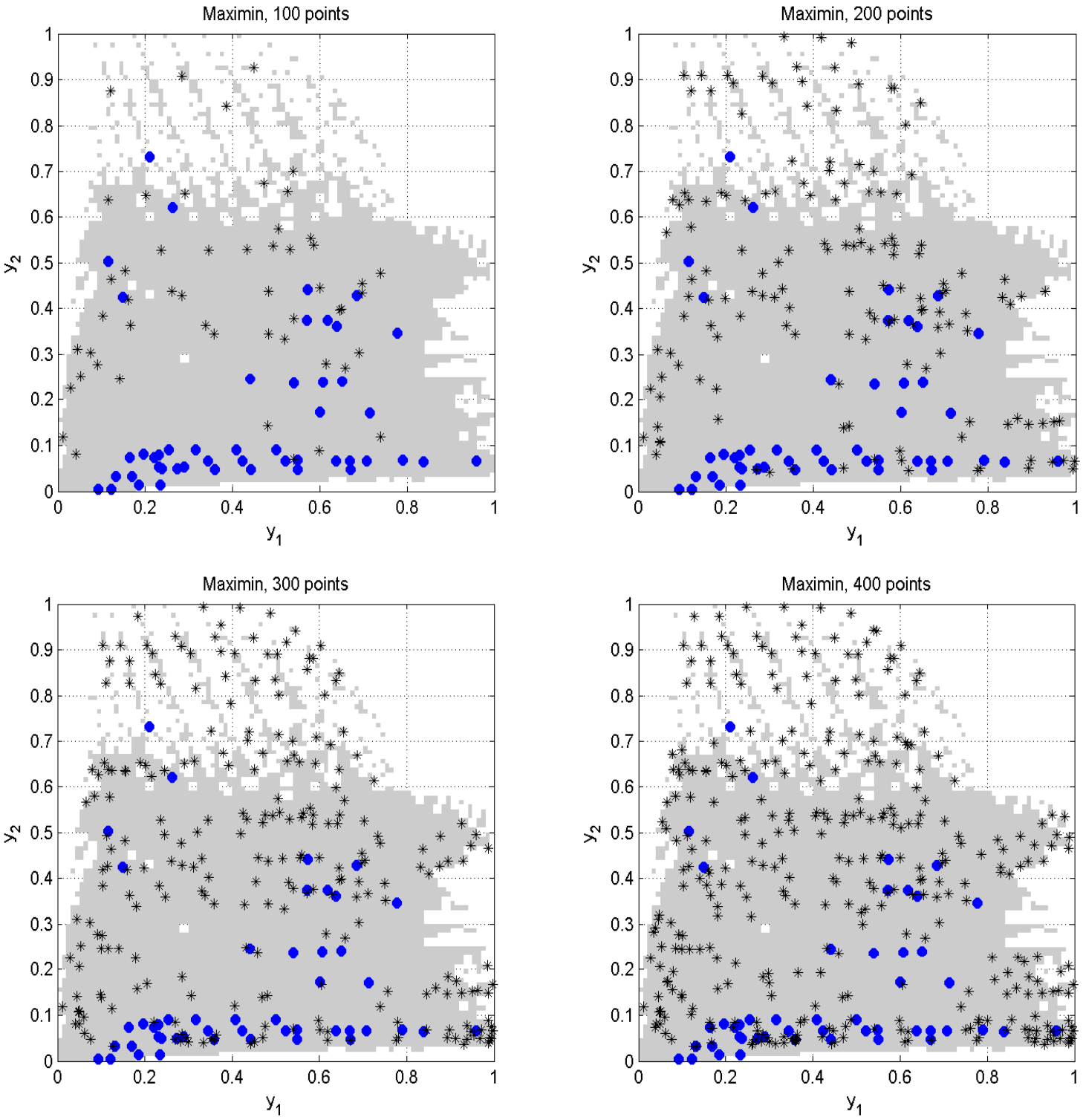

FIGURE 5.35 – Remplissage du domaine de sortie par diversité maximin, avec bruit de mesure de 5%, essais par séries de 6 et retard d'arrivée des résultats.

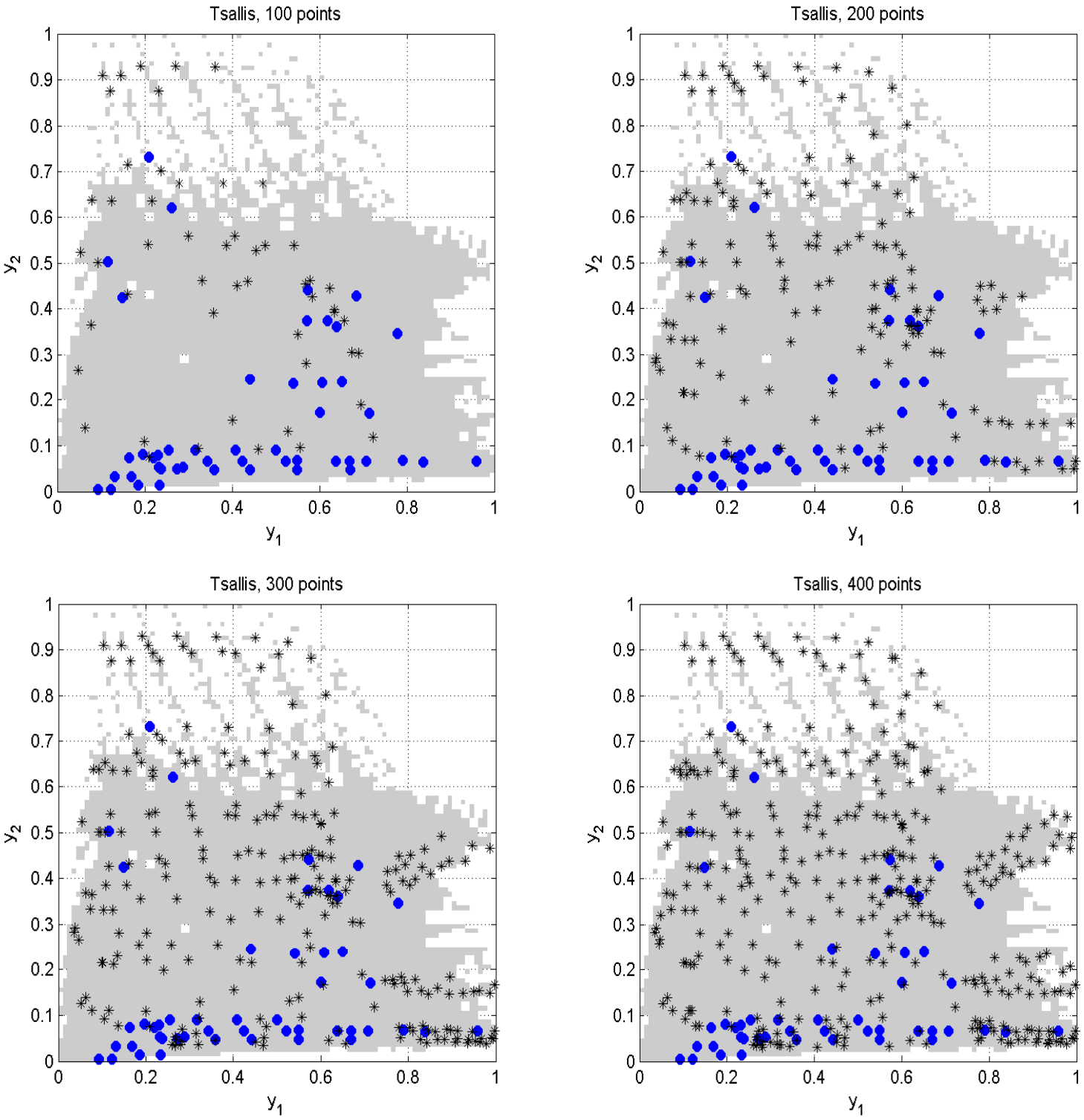

FIGURE 5.36 – Remplissage du domaine de sortie par diversité de Tsallis, avec bruit de mesure de 5%, essais par séries de 6 et retard d'arrivée des résultats.

Nous avons dans ce chapitre comparé les critères de diversité maximin et de Tsallis. Dans le cas du critère maximin, les bornes du domaine atteignable en sortie étaient supposées connues, ce qui rend la méthode applicable à condition d'avoir au moins une idée *a priori* des valeurs de ces bornes. Le critère de Tsallis ne suppose pas ces bornes connues. Pour le choix de la valeur du pas, nous l'avons fixé à $h = 10^{-2}$, les sorties observées étant ramenées à $[0, 1]$ à chaque itération. Le critère maximin avec bornes connues donne les résultats les plus satisfaisants en ce qui concerne la répartition finale des sorties, mais le critère de Tsallis donne aussi une répartition satisfaisante et est plus économe en temps de calcul du fait qu'il se calcule analytiquement. Nous avons testé une approche simplifiée donnant de bons résultats pratiques, mais peu satisfaisante sur le plan théorique (réinterpolation, meilleurs optima locaux). L'approche théoriquement satisfaisante présentée en 4.2.3.3 a été abordée en pratique dans le cas de 2 entrées et 1 sortie : le peu de résultats obtenus semblent encourageants, mais il reste à mettre au point un algorithme d'optimisation efficace et règler les problèmes de conditionnement des matrices de corrélation dans le cas du retard d'arrivée des mesures, en utilisant par exemple une fonction de covariance de Matérn du type (2.14) ou (2.15). Finalement, l'inversion du système à partir des plans finaux obtenus à l'aide de ces critères d'ajout a été étudiée dans le cas de deux entrées et une sortie, et il a été observé que les deux critères de diversité retenus sont pertinents pour la recherche des facteurs d'entrée où la réponse est la plus proche d'une valeur cible donnée. Le problème de la recherche des bornes du domaine atteignable par les sorties reste cependant à étudier.

Conclusions et perspectives

Conclusions La première partie de ce travail a consisté en un exposé théorique des notions utilisées dans ce mémoire.

Nous avons tout d'abord voulu écrire de façon claire et détaillée les liens bien connus existant entre des méthodes telles que le krigeage, les splines, les ondelettes et les machines à vecteurs de support. Après une brève introduction aux espaces de Hilbert à noyau reproduisant, l'énoncé du théorème du représentant sous diverses formes a montré que ces méthodes consistent à minimiser la somme d'un risque empirique et d'un terme de pénalisation. Il a ainsi été constaté que le krigeage intrinsèque est équivalent aux splines « plaque mince » (lorsque le noyau est conditionnellement semi-défini positif), et le fait que les machines à vecteurs de support forment un cadre général regroupant les autres méthodes citées précédemment (notamment, la fonction de risque empirique n'est pas nécessairement quadratique dans le cas de la SVR, mais le noyau est supposé continu). Notre choix d'utiliser le krigeage pour modéliser le système a été motivé par le fait qu'il permet d'évaluer naturellement l'incertitude (à travers le cadre probabiliste) pour un coût de calcul moindre, ce qui nous a permis de prendre en compte l'erreur (quadratique moyenne) de prédiction dans nos critères d'ajout.

Nous avons ensuite détaillé davantage l'exposé théorique concernant le krigeage, en commençant par présenter les processus gaussiens qui forment la base probabiliste de la méthode. Il a ainsi été constaté que la régularité des trajectoires dépend de la régularité de la fonction de covariance. Puis, après quelques rappels de statistique, la définition du prédicteur de krigeage en un point a été donnée ainsi que son expression analytique et celle de l'erreur (quadratique moyenne) associée. Nous nous sommes ensuite intéressés aux qualités des estimateurs du maximum de vraisemblance, dont les « bonnes » propriétés asymptotiques dépendent généralement d'observations placées de plus en plus loin (*i.e.* dans un domaine de taille croissante), ce qui ne correspond pas à la plupart des situations pratiques où le domaine d'observation est borné. Il est rappelé que choisir une mauvaise fonction de covariance n'est pas nécessairement un obstacle à la prédiction optimale, du moment que la fonction de covariance choisie est compatible, en un certain sens, avec la vraie fonction de covariance. Finalement, les modifications à intégrer au modèle de krigeage afin de prendre en compte le bruit de mesure ont été présentées.

Le chapitre sur les plans d'expériences a conclu cette présentation théorique. Il s'agissait plutôt de donner les idées générales intervenant dans la mise au point d'un plan d'expériences que de rentrer dans les détails. Les plans remplissant l'espace (ou *space-filling*) dont la construction est indépendante d'un modèle ont été présentés dans un premier temps. Ce type de plan peut être utilisé lorsqu'on ne dispose d'aucune information *a priori* sur le système et que l'on souhaite observer « un peu partout ». Nous avons retenu les hypercubes latins maximin pour leur facilité de construction. Ensuite, une liste de plans d'expériences construits à partir d'un modèle a été présentée. Il a ainsi été rappelé que les critères traditionnels (D-optimalité par exemple), définis pour un modèle de régression classique, peuvent être généralisés pour un modèle de krigeage.

Finalement, des critères d'ajout séquentiel de points utilisés pour l'optimisation globale d'une fonction, issus de la littérature, ont été présentés. Nous nous sommes inspirés de ces critères pour mettre au point des critères d'ajout visant à optimiser la diversité des réponses, critères présentés dans la deuxième partie de ce travail.

Cette seconde partie a été consacrée à la mise au point d'un critère d'ajout dans le but de maximiser la diversité des réponses (afin de pouvoir, par la suite, obtenir un ensemble de points en entrée, les « prédictions inverses », permettant de bien prédire l'ensemble des valeurs atteignables par les sorties). Dans un premier temps, des critères à base de discrépance, de distance et d'entropie ont été testés dans le cas simple où l'on souhaite ajouter un point à un ensemble existant de sorte que le nouvel ensemble obtenu soit aussi dispersé que possible. Des essais préliminaires dans $\mathbb{R}$ ont permis de retenir deux critères, l'un à base de distance, l'autre à base d'entropie (de Tsallis), qui ont ensuite été testés plus en détail. Chacun des deux critères permet la prise en compte de l'incertitude donnée par le modèle de krigeage pour un coût de calcul faible dans le cas d'une seule sortie. Dans le cas de deux sorties, le critère à base de distance ne s'écrit plus sous forme analytique et doit être calculé en effectuant une intégration numérique, ce qui allonge le temps de calcul. Le critère à base d'entropie de Tsallis s'écrit sous forme analytique quel que soit le nombre de sorties. La mise en forme des critères afin de prendre en compte les contraintes de l'étude (bruit, mesures effectuées plusieurs à la fois et retard d'arrivée des résultats) a ensuite été discutée. Si le critère à base de distance ne s'adapte pas naturellement à ces contraintes pratiques, il a été montré que c'est le cas pour le critère à base d'entropie de Tsallis. Cependant, la mise en œuvre pratique peut s'avérer délicate, en raison notamment du mauvais conditionnement des matrices de covariance et du coût de calcul inhérent à l'optimisation d'un critère multi-variable (plusieurs points étant ajoutés à chaque étape). C'est pourquoi nous avons finalement proposé une approche empirique qui permet en pratique d'utiliser les deux critères dans le cadre du problème posé.

Le fonctionnement de l'algorithme d'optimisation a ensuite été détaillé. Un plan d'expériences initial est enrichi séquentiellement, en utilisant un critère de diversité, jusqu'à avoir épuisé le budget de mesures. La procédure a tout d'abord été mise au point dans le cas où le système a 2 entrées et 1 sortie. Des tests effectués sur une bibliothèque de plans initiaux et de fonctions-test ont montré que le critère à base de distance donne les meilleurs résultats quand les bornes du domaine de sortie sont connues (ce qui n'est pas le cas en pratique). Le critère à base d'entropie de Tsallis donne des résultats satisfaisants quand ces bornes sont inconnues. Nous avons montré la pertinence des critères d'ajout proposés en les comparant a la technique classique consistant à ajouter séquentiellement des points au maximum de la variance de krigeage, technique qui ne permet pas de construire un ensemble de réponses convenablement dispersées. Nous avons effectué quelques tests avec le critère d'entropie de Tsallis en prenant en compte les contraintes pratiques de l'étude (retards et expérimentations groupées). Les résultats obtenus sont satisfisants au niveau de la répartition finale des réponses, mais l'optimisation du critère est coûteuse en temps de calcul. En ce qui concerne les performances des prédictions inverses, nous avons observé que les deux critères d'ajout proposés donnent des résultats satisfaisants, mais que la recherche des bornes du domaine atteignable par les sorties devrait être incorporée à l'algorithme d'ajout. Nous avons finalement mis au point la procédure d'ajout dans le cas où le système comporte 5 entrées et 2 sorties (comme le système réel), et avons fait des tests sur un modèle physique du système. Les deux critères de diversité ont donné des résultats satisfaisants lorsqu'on utilise l'approche empirique en présence de retard et d'expérimentations groupées.

Perspectives Il a été montré que le critère de diversité utilisant l'entropie de Tsallis s'adapte naturellement en théorie aux contraintes de l'étude (retard et expérimentations groupées). D'un point de vue pratique, celui-ci n'est cependant pas encore utilisable en raison du temps de calcul élevé pour déterminer la prochaine série de points de mesure à chaque itération. Il serait donc intéressant de mettre au point un algorithme de recherche plus efficace permettant de trouver la solution dans un temps raisonnable. L'approche empirique proposée donne des résultats satisfaisants en pratique, mais sa justification théorique reste à établir.

Nous avons observé que dans certains cas, les deux critères d'ajout proposés n'étaient pas entièrement satisfaisant pour le calcul des prédictions inverses : en effet, lorsque les bornes du domaine de sortie ne sont pas explorées, leurs prédictions inverses ne donnent pas des prédictions satisfaisantes. Il serait donc profitable d'imaginer une procédure d'ajout prenant en compte la recherche des bornes du domaine de sortie.

Nous avons supposé ici que le phénomène étudié pouvait être modélisé par un processus stationnaire. L'extension de la méthode au cas non stationnaire (sans doute plus proche de la réalité) mériterait d'être étudiée. Il faudrait alors faire intervenir des fonctions de corrélation non stationnaires dans la méthode de krigeage telles que celles proposées par Paciorek [123] et Stein [165], ou utiliser le krigeage intrinsèque.

Le calcul de l'incertitude sur la prédiction par krigeage utilisée dans la méthode d'ajout de points ne prend pas en compte la part due à l'estimation des paramètres du modèle (paramètres de la fonction de corrélation). Une correction du type de celles introduites par Abt [4, 5] ou Zimmerman et Cressie [201] pourrait être utilisée pour estimer la variance de krigeage, en gardant à l'esprit qu'il y a des cas où ces corrections font empirer la qualité de l'estimation. En ce qui concerne les changements apportés au critère de diversité, seule la variance de la loi normale de la nouvelle sortie serait alors modifiée.

Il aurait également été possible de prendre en compte le coût de fabrication. En effet, plusieurs choix différents des variables d'entrée peuvent conduire aux mêmes valeurs de sortie, et tous les choix des variables d'entrée n'ont pas le même coût de fabrication. Pour ce faire, il serait envisageable d'introduire une fonction de coût dans l'intégrale qui définit le critère de diversité.

Il aurait finalement été intéressant de pouvoir tester l'approche proposée sur le système réel. Celui-ci n'était cependant pas encore prêt à recevoir des mesures au moment de l'écriture de ces lignes. Pour le moment, des études de répétabilité (visant à vérifier que des valeurs d'entrée identiques produisent à peu près des mêmes valeurs en sortie) ont conduit à restreindre l'espace des facteurs, ce qui va diminuer le nombre de points de la grille des entrées et ainsi accélérer la procédure d'optimisation. À ce jour, les mesures ont été effectuées aux coins du domaine, le reste des points du plan initial devrait avoir été observé dans quelques mois puis l'algorithme d'optimisation sera appliqué. Souhaitons que les « critères de diversité » puissent à l'avenir trouver d'autres applications pratiques !

Annexe A

Ergodicité

Soit Y un processus aléatoire à valeurs réelles défini sur $\mathcal{X} \times \Omega$, avec un espace d'indexation $\mathcal{X} \subset \mathbb{R}^d$ et un espace de probabilité $(\Omega, \mathcal{T}, \mathcal{P})$. Nous rappelons ici quelques éléments relatifs au concept d'ergodicité, et montrons que l'on peut, sous certaines hypothèses, estimer des espérances mathématiques à partir d'observations d'une seule trajetoire Y_ω dans l'espace des paramètres $\mathcal{X}$.

Définition A.0.2 *[101] Notons $\mathbb{R}^{\mathcal{X}}$ l'ensemble des fonctions à valeurs réelles définies sur $\mathcal{X}$. On appelle* cylindre *de $\mathbb{R}^{\mathcal{X}}$ tout ensemble de fonctions $y(x) \in \mathbb{R}^{\mathcal{X}}$ défini par un nombre fini d'inégalités du type*

$$a_1 < y(x_1) \leq b_1, \ldots, a_n < y(x_n) \leq b_n,$$

avec $a_i, b_i \in \mathbb{R}, x_i \in \mathcal{X}$. Les cylindres forment une algèbre notée $\mathcal{I}$.
La tribu borélienne *$\mathcal{B}_{\mathcal{X}}$ de $\mathbb{R}^{\mathcal{X}}$ est la tribu engendrée par les cylindres,*

$$\mathcal{B}_{\mathcal{X}} = \sigma(\mathcal{I}).$$

On rappelle que la mesure image $\mathcal{P}_Y$ est définie sur l'ensemble mesurable $(\mathbb{R}^{\mathcal{X}}, \mathcal{B}_{\mathcal{X}})$ par

$$\mathcal{P}_Y(B) = \mathcal{P}(Y^{-1}(B)) \quad \forall B \in \mathcal{B}_{\mathcal{X}}.$$

Définition A.0.3 *Soit $(\Omega, \mathcal{T}, \mathcal{P})$ un espace de probabilité. L'application mesurable $g : (\Omega, \mathcal{T}, \mathcal{P}) \longrightarrow (\Omega, \mathcal{T}, \mathcal{P})$* préserve la mesure $\mathcal{P}$ *si*

$$\mathcal{P}(g^{-1}(A)) = \mathcal{P}(A) \quad \forall A \in \mathcal{T}.$$

L'application g est appelée transformation préservant la mesure $\mathcal{P}$.
L'ensemble $A \in \mathcal{T}$ est dit invariant par *g si $g^{-1}(A) = A$.*

Une transformation préserve donc la mesure si la probabilité d'un évènement est la même avant et après la transformation.

Voyons maintenant un type de transformation préservant la mesure $\mathcal{P}_Y$ lorsque le processus aléatoire Y est stationnaire.

Définition A.0.4 *On appelle* décalage (shift) *la transformation τ_t définie sur $(\mathbb{R}^{\mathcal{X}}, \mathcal{B}_{\mathcal{X}}, \mathcal{P}_Y)$ par*

$$\tau_t \circ y(x) = y(x + t).$$

La transformation τ_t préserve la mesure $\mathcal{P}_Y$ si le processus aléatoire Y est stationnaire.

Il est maintenant possible de définir la notion d'ergodicité.

Définition A.0.5 *Le processus aléatoire Y est dit* ergodique *si la probabilité de chacun de ses ensembles invariants par τ_t vaut 0 ou 1, pour tout $t \in \mathcal{X}$.*

Dans le cas d'un processus gaussien stationnaire, on a une condition suffisante simple d'ergodicité sur sa fonction de covariance $k(\cdot)$.

Théorème A.0.6 *[31, 101] Un processus gaussien stationnaire est ergodique si* $\lim_{\|\tau\|\to\infty} k(\tau) = 0$.

Le résultat important suivant permet d'estimer l'espérance mathématique de la variable aléatoire $Y(0)$ (ou de $Y(x)$, pour une valeur quelconque de x) à partir de la moyenne spatiale d'une trajectoire Y_ω.

Théorème A.0.7 *[101] Soit Y un processus stationnaire (définition 2.1.13) sur $\mathbb{R} \times \Omega$, avec $\mathbb{E}|Y(x)| < \infty \quad \forall x \in \mathbb{R}$, et à trajectoires intégrables (presque sûrement). Alors, si Y est ergodique,*

$$\frac{1}{t}\int_0^t Y(x)\mathrm{d}x \underset{t\to+\infty}{\longrightarrow} \mathbb{E}\{Y(0)\} \qquad (= \mathbb{E}\{Y(x)\} \ \forall x) \quad \mathcal{P}\text{-presque sûrement.}$$

Remarque A.0.8 *Le résultat d'ergodicité ci-dessus concerne concerne la moyenne du processus. Sous certaines hypothèses, il est possible d'obtenir des résultats similaires de convergence presque sûre pour la variance et la covariance,*

$$\sigma^2 = \mathbb{E}\left\{(Y(x) - m)^2\right\} = \lim_{t\to\infty} \frac{1}{t}\int_0^t (Y(x) - m)^2 \mathrm{d}x$$

et

$$k(\tau) = \mathbb{E}\left\{(Y(x) - m)\,(Y(x+\tau) - m)\right\} = \lim_{t\to\infty} \frac{1}{t-\tau}\int_0^{t-\tau} (Y(x) - m)(Y(x+\tau) - m)\mathrm{d}x.$$

Annexe B

Régularité en moyenne quadratique

Nous rappelons ici la notion de régularité en moyenne quadratique (M.Q.) d'un processus aléatoire. Contrairement à la régularité des trajectoires, la relation entre la régularité en M.Q. et la fonction de covariance est très simple et ne nécessite pas d'hypothèse de séparabilité du processus.

Définition B.0.9 *Soit Y un processus aléatoire défini sur $\mathcal{X} \times \Omega$, avec $\mathcal{X} \subset \mathbb{R}^d$. Le processus aléatoire Y est dit* continu en moyenne quadratique *en $x \in \mathcal{X}$ si*

$$\lim_{t \to x} \mathbb{E}\{Y(t) - Y(x)\}^2 = 0.$$

On dira que Y est continu en M.Q. sur $\mathcal{X}$ si il est continu en M.Q. en x, pour tout $x \in \mathcal{X}$.

Si Y est stationnaire, de fonction de covariance $k(\cdot)$, on peut écrire $\mathbb{E}\{Y(x) - Y(x')\}^2 = 2\{k(0) - k(x - x')\}$, d'où le résultat suivant.

Théorème B.0.10 *[2] Le processus aléatoire stationnaire Y est continu en M.Q. sur $\mathbb{R}^d$ si, et seulement si, sa fonction de covariance $k(\cdot)$ est continue en* 0.

La fonction de covariance d'un processus stationnaire continu en M.Q. est donc continue partout (c.f. remarque 2.1.15).

La continuité en M.Q. sur $\mathcal{X}$ n'implique pas en général la continuité des trajectoires, comme illustré par le processus Z de l'exemple 2.1.6. Pour un processus gaussien séparable, la continuité en M.Q. implique la continuité des trajectoires si le théorème 2.1.19 est verifié, ce qui est toujours le cas en pratique [2].
La réciproque est fausse aussi en général : la continuité des trajectoires n'entraîne pas la continuité en M.Q. (sauf si le processus est borné [17]).

Exemple B.0.11 *[17] Soit $Y : \mathcal{X} \times \Omega \to \mathbb{R}$ un processus aléatoire, avec $\mathcal{X} = [0, 1]$ et $\Omega = [0, 1]$ muni de la loi uniforme, défini par*

$$Y(x, \omega) = \begin{cases} \frac{1}{x - \frac{1}{2}} \mathbb{1}_{]\frac{1}{2}, x[}(\omega) \ \textit{si} \ x \in]\frac{1}{2}, 1] \, ; \\ 0 \ \textit{si} \ x \in [0, \frac{1}{2}]. \end{cases}$$

Les trajectoires $Y_\omega(\cdot)$ de Y sont continues sur $[0, \frac{1}{2}]$ et $]\frac{1}{2}, 1]$. De plus, $\mathcal{P}\{Y_\omega(x) \to 0$ quand $x \to \frac{1}{2}\} = 1$, donc les trajectoires de Y sont continues. Cependant, $\mathbb{E}[Y^2(x, \omega)] \to \infty$ quand $x \to \frac{1}{2}^+$, et $\mathbb{E}[Y^2(x, \omega)] = 0$ si $x \in [0, \frac{1}{2}]$, ce qui montre que Y n'est pas continu en M.Q.

Nous rappelons maintenant la notion de différentiabilité en moyenne quadratique. On pourra remarquer les nombreux points communs avec la différentiabilité classique. Pour alléger la présentation, on notera dans la suite $Y(x) = Y(x, \omega)$, tout en gardant à l'esprit qu'il s'agit d'un processus aléatoire.

Définition B.0.12 *[17] Un processus aléatoire $Y(x), x \in \mathbb{R}^d$ est dit* différentiable en moyenne quadratique *en x si $\forall h \in \mathbb{R}^d$, il existe un processus aléatoire $L_x(h)$, linéaire en h, tel que*

$$Y(x+h) = Y(x) + L_x(h) + R(x,h), \quad \textit{avec } \frac{R(x,h)}{\| h \|} \underset{\|h\| \to 0}{\longrightarrow} 0 \textit{ dans } \mathrm{L}^2,$$

où L^2 désigne le quotient de l'espace de Hilbert des variables aléatoires de carré intégrable par la relation d'équivalence « égalité presque sûre ». La différentielle $L_x(h)$ est donc définie à un ensemble de mesure $\mathcal{P}$-nulle près.
Nous pouvons remarquer la différence entre les notions de différentiabilité en M.Q. et p.s., dûe au fait que la convergence en M.Q. et la convergence p.s. (du processus $R(x,h)/\| h \|$) ne sont pas liées généralement [60].

Remarque B.0.13 *[17] On peut montrer que la différentiabilité en M.Q. entraîne la continuité en M.Q., comme dans le cas déterministe.*

Définition B.0.14 *Soit $Y(x), x \in \mathbb{R}^d$ un processus aléatoire. On appelle* dérivée directionnelle M.Q. *de direction v, où $v \in \mathbb{R}^d$, le processus aléatoire*

$$D_v Y(x) = \lim_{\varepsilon \to 0} \frac{Y(x+\varepsilon v) - Y(x)}{\varepsilon} \quad \textit{dans } \mathrm{L}^2.$$

On appelle i^e dérivée partielle M.Q. *le processus $D_{e_i} Y$, avec e_i le i^e vecteur de la base canonique de $\mathbb{R}^d$.*

Si Y est différentiable en M.Q., on peut montrer que $D_v Y(x) = L_x(v)$ presque partout. Tout comme dans la théorie différentielle classique, l'existence des dérivées directionnelles M.Q. ne garantit cependant même pas la continuité en M.Q.

Exemple B.0.15 *[17] Soit $\{Y(x), x = (x_1, x_2) \in \mathbb{R}^2\}$ le processus aléatoire défini par*

$$Y(x) = \begin{cases} \frac{x_1 x_2^2}{x_1^2 + x_2^4} V \textit{ si } x \neq 0, \textit{ avec } V \sim \mathcal{N}(0,1)\,; \\ 0 \textit{ si } x = 0. \end{cases}$$

On vérifie que $D_v Y(0) = \frac{v_2^2}{v_1} V \ \ \forall v = (v_1, v_2), v_1 \neq 0$, et $D_v Y(0) = 0 \ \ \forall v = (0, v_2)$, d'où l'existence de toutes les dérivées directionnelles au point 0. Cependant, $\mathbb{E}\{Y(x) - Y(0)\}^2 = \frac{1}{4}$ si $x_1 = x_2^2$, Y n'est donc pas continu en M.Q. en 0.

Remarque B.0.16 *La différentiabilité en moyenne quadratique est parfois définie comme l'existence de toutes les dérivées directionnelles M.Q. [2]. Cependant, cette définition conduit à l'existence de processus aléatoires dérivables en M.Q. mais pas continus en M.Q. (exemple B.0.15).*

Voyons maintenant les liens entre la régularité M.Q. d'un processus aléatoire stationnaire et sa fonction de covariance.

Proposition B.0.17 *[17] Soit $Y(x), x \in \mathbb{R}^d$ un processus aléatoire stationnaire de fonction de covariance $k(\cdot)$. On suppose que les dérivées partielles d'ordre 2 de $k(\cdot)$ existent et sont continues. Alors, si on note $k_v(\cdot)$ la fonction de covariance de la dérivée directionnelle M.Q. $D_vY(\cdot)$, on a*

$$k_v(\tau) = -{}^t v H(\tau) v, \text{ avec } H(\tau)_{(i,j)} = \frac{\partial^2 k}{\partial \tau_i \partial \tau_j}(\tau).$$

Le théorème suivant, conséquence de résultats donnés dans [2, 161], donne une condition nécessaire et suffisante de dérivabilité d'ordre q en M.Q. pour un processus aléatoire stationnaire dont les moments d'ordre 2 sont finis.

Théorème B.0.18 *Soit Y un processus aléatoire stationnaire de fonction de covariance $k(\cdot)$. Alors*

$$Y \text{ est } q \text{ fois dérivable en M.Q.} \iff k^{(2\kappa)}(0) \text{ existe } \forall \kappa, |\kappa| = q,$$

avec l'expression de la dérivée de la fonction de covariance donnée dans la proposition 2.1.22.

Si on admet qu'en pratique le théorème 2.1.19 est vrai pour toute fonction de corrélation stationnaire continue, comme suggeré dans [2], on peut appliquer le théorème B.0.18 pour déterminer la régularité des trajectoires d'un processus gaussien stationnaire (en enlevant « en M.Q. » dans l'énoncé).

Pour conclure, et afin d'illustrer la différence entre les deux notions de dérivabilité, nous donnons un exemple de processus aléatoire pouvant avoir des trajectoires analytiques et n'être même pas dérivable en moyenne quadratique.

Exemple B.0.19 *[161] Soit $Y(x) = \cos(U + xV)$, avec U et V des variables aléatoires indépendantes de lois respectives uniforme sur $[0, 2\pi]$ et Cauchy standard. On peut montrer que $\mathbb{E}\{Y(x)\} = 0$ et $\mathrm{cov}(Y(x), Y(x')) = \frac{1}{2}e^{-|x-x'|}$, donc le théorème B.0.18 permet d'affirmer que le processus stationnaire Y n'est pas dérivable en moyenne quadratique, bien que ses trajectoires soient analytiques.*

Annexe C

Krigeage bayésien

Nous rappelons ici comment la théorie bayésienne peut être utilisée pour estimer les paramètres d'un processus gaussien [56, 136]. Nous commençons par quelques rappels sur la méthodologie bayésienne, puis présentons des méthodes d'estimation des paramètres.

C.1 Formalisme bayésien

On commence par se donner une famille de modèles (ou *hypothèses*) $\mathcal{H}_i$, pouvant servir à modéliser un ensemble de données $\mathcal{D}$, avec pour chacun un ensemble de paramètres w_i. On utilise ensuite nos connaissances du système étudié pour définir :

- une probabilité $\mathcal{P}(\mathcal{H}_i)$, qui permet de distinguer les modèles *a priori* plus ou moins adaptés ;
- une probabilité $\mathcal{P}(w_i|\mathcal{H}_i)$, qui permet de quantifier, pour un modèle $\mathcal{H}_i$ donné, les valeurs des paramètres *a priori* les plus pertinentes.

Une fois l'ensemble des données d'apprentissage $\mathcal{D}$ collectées, on effectue deux inférences successives :

1. pour tout i, inférence des paramètres w_i du modèle $\mathcal{H}_i$ à partir des données. On utilise pour cela la formule de Bayes,

$$\underbrace{\mathcal{P}(w_i|\mathcal{D},\mathcal{H}_i)}_{\text{a posteriori}} = \frac{\overbrace{\mathcal{P}(\mathcal{D}|w_i,\mathcal{H}_i)}^{\text{vraisemblance}}\ \overbrace{\mathcal{P}(w_i|\mathcal{H}_i)}^{\text{a priori}}}{\underbrace{\mathcal{P}(\mathcal{D}|\mathcal{H}_i)}_{\text{évidence}}}. \tag{C.1}$$

   La *vraisemblance*, que l'on peut calculer, est la probabilité d'obtenir les données à partir du modèle $\mathcal{H}_i$ avec les valeurs des paramètres w_i (voir page 55 dans le cas d'un processus gaussien). L'*évidence* est inconnue, mais constante ($\mathcal{D}$ et $\mathcal{H}_i$ sont fixés à ce niveau), c'est donc une constante de normalisation qui peut être omise dans les calculs. On obtient donc la densité de probabilité *a posteriori* comme le produit de la vraisemblance par l'*a priori* (à une constante multiplicative près) ;

2. comparaison des modèles $\mathcal{H}_i$. Supposons que l'on puisse maintenant, pour tout i, calculer l'évidence (comme constante de normalisation) en utilisant la formule (C.1). En appliquant une nouvelle fois le théorème de Bayes, on obtient

$$\mathcal{P}(\mathcal{H}_i|\mathcal{D}) = \frac{\mathcal{P}(\mathcal{D}|\mathcal{H}_i)\mathcal{P}(\mathcal{H}_i)}{\mathcal{P}(\mathcal{D})}, \tag{C.2}$$

où $\mathcal{P}(\mathcal{D})$, qui ne dépend que des données, est une constante de normalisation. La densité de probabilité *a posteriori* $\mathcal{P}(\mathcal{H}_i|\mathcal{D})$ nous permet finalement de ranger les modèles $\mathcal{H}_i$ par ordre de pertinence.

En pratique, il est souvent très difficile d'évaluer l'évidence dans l'équation (C.1), qui peut être une intégrale très compliquée, et il faut se résoudre à des approximations de type Monte Carlo. Le choix de l'*a priori* n'est pas non plus toujours évident.

C.2 Application au krigeage

Nous pouvons maintenant présenter le krigeage bayésien [65, 144]. On commence par se donner une loi *a priori* pour le vecteur de paramètres ${}^t({}^t\beta, \sigma^2, {}^t\psi)$. Les modèles $\mathcal{H}_i$ précédents correspondent ici à des processus gaussiens dont la fonction de corrélation est paramétrée par un vecteur ψ ; on peut donc écrire $\mathcal{H}_i = \mathcal{H}_\psi$. Suivant le formalisme présenté en C.1, la densité de probabilité $f_\psi(\cdot)$, qui décrit la loi du paramètre ψ, correspond à la probabilité $\mathcal{P}(\mathcal{H}_i)$. La densité de probabilité $f_{[w|\psi]}(\cdot)$ correspond à la probabilité $\mathcal{P}(w_i|\mathcal{H}_i)$, qui décrit la loi du vecteur des paramètres $w = {}^t({}^t\beta, \sigma^2)$ à ψ fixé. Partant de la loi (2.19),

$$\left[\begin{pmatrix} Y_0 \\ Y^n \end{pmatrix} \middle| w, \psi \right] \sim \mathcal{N}_{1+n} \left[\begin{pmatrix} {}^t m_0 \\ M \end{pmatrix} \beta, \sigma^2 \begin{pmatrix} 1 & {}^t r_0 \\ r_0 & R \end{pmatrix} \right], \tag{C.3}$$

on va chercher à évaluer la loi du meilleur prédicteur $\mathbb{E}\{Y_0|Y^n\}$ (théorème 2.2.4), en tenant compte de la distribution *a priori* des paramètres. Cette loi *a posteriori* du prédicteur est appelée *loi prédictive*, et sa densité de probabilité s'écrit

$$\begin{aligned} f_{[Y_0|Y^n]} &= f_{[Y_0|Y^n,\psi]} f_{[\psi|Y^n]} \\ &= f_{[Y_0|Y^n,w,\psi]} f_{[w|Y^n,\psi]} f_{[\psi|Y^n]}. \end{aligned} \tag{C.4}$$

La densité $f_{[Y_0|Y^n,w,\psi]}(\cdot)$ est donnée par la proposition 2.2.12. Pour obtenir l'expression des deux autres densités $f_{[w|Y^n,\psi]}(\cdot)$ et $f_{[\psi|Y^n]}(\cdot)$, on applique la théorie bayésienne présentée en C.1.
On effectue tout d'abord le premier niveau d'inférence (C.1),

$$f_{[w|Y^n,\psi]} = \frac{f_{[Y^n|w,\psi]} f_{[w|\psi]}}{f_{[Y^n|\psi]}},$$

qui permet d'obtenir la densité $f_{[w|Y^n,\psi]}(\cdot)$ et l'évidence $f_{[Y^n|\psi]}(\cdot)$ par intégration, en utilisant le fait que l'*a posteriori* $f_{[w|Y^n,\psi]}(\cdot)$ est d'intégrale 1. Par le deuxième niveau d'inférence (C.2), on obtient ensuite la densité *a posteriori* de ψ,

$$f_{[\psi|Y^n]} = \frac{f_{[Y^n|\psi]} f_\psi}{f_{Y^n}}.$$

On peut alors finalement calculer (C.4). Le problème est qu'en pratique les intégrales à évaluer sont souvent très compliquées, et il faut se résoudre à faire des hypothèses simplificatrices ou des approximations [142].

– Approximation par méthodes de Monte Carlo : l'idée est d'approcher la distribution prédictive en écrivant

$$\begin{aligned} f_{[Y_0|Y^n]}(\cdot) &= \int f_{[Y_0|Y^n,(\beta,\sigma^2,\psi)]}(\cdot) f_{[(\beta,\sigma^2,\psi)|Y^n]}(\zeta) \mathrm{d}\zeta \\ &\approx \frac{1}{l} \sum_{i=1}^{l} f_{[Y_0|Y^n,(\beta,\sigma^2,\psi)_i]}, \end{aligned} \tag{C.5}$$

où $\{(\beta, \sigma^2, \psi)_i, i = 1, \ldots, l\}$ est un échantillon obtenu à partir de la loi *a posteriori* des paramètres, de densité

$$f_{[(\beta,\sigma^2,\psi)|Y^n]} = \frac{f_{[Y^n|(\beta,\sigma^2,\psi)]} f_{(\beta,\sigma^2,\psi)}}{f_{Y^n}}.$$

Des méthodes de Monte Carlo, dont les *MCMC (Markov Chain Monte Carlo)* qui utilisent une chaîne de Markov dont la loi stationnaire est $[(\beta, \sigma^2, \psi)|Y^n]$, sont discutées dans [117, 188, 189].

– *Maximisation de l'évidence (evidence maximization)* [110] : l'idée est d'approcher la densité conditionnelle $f_{[Y_0|Y^n]}(\cdot)$ en utilisant les valeurs des paramètres les plus probables, déterminées à partir de la relation

$$\begin{aligned} f_{[\beta,\sigma^2,\psi|Y^n]} &= \frac{f_{[Y^n|\beta,\sigma^2,\psi]} f_{(\beta,\sigma^2,\psi)}}{f_{Y^n}} \\ &\propto f_{[Y^n|\beta,\sigma^2,\psi]} f_{(\beta,\sigma^2,\psi)}. \end{aligned}$$

La méthode de maximisation de l'évidence sélectionne les valeurs des paramètres qui maximisent la fonction $f_{[\beta,\sigma^2,\psi|Y^n]} \propto f_{[Y^n|\beta,\sigma^2,\psi]} f_{(\beta,\sigma^2,\psi)}$. En passant au logarithme, on obtient

$$(\widehat{\beta}, \widehat{\sigma^2}, \widehat{\psi}) = \underset{(\beta,\sigma^2,\psi)}{\operatorname{argmax}} \left[l(\beta, \sigma^2, \psi|Y^n) + \log f(\beta, \sigma^2, \psi) \right], \tag{C.6}$$

avec $l(\beta, \sigma^2, \psi|Y^n)$ la log-vraisemblance (2.30) et $f = f_{(\beta,\sigma^2,\psi)}$ la densité *a priori* des paramètres. On remarque le terme correctif $f(\beta, \sigma^2, \psi)$ ajouté à la log-vraisemblance, qui permet la prise en compte de l'information *a priori* sur la loi des paramètres (notons que cela ne garantit normalement pas contre d'éventuels problèmes de multi-modalité de la log-vraisemblance, mais il est possible de choisir une loi *a priori* qui garantit contre la multi-modalité tout en contenant l'information *a priori* dont on dispose [56]). Le vecteur $(\widehat{\beta}, \widehat{\sigma^2}, \widehat{\psi})$ est appelé *mode a posteriori* de la densité $f_{[Y^n|\beta,\sigma^2,\psi]}$.
La méthode de maximisation de l'évidence est basée sur l'hypothèse simplificatrice que la densité *a posteriori* des paramètres $f_{[\beta,\sigma^2,\psi|Y^n]}$ est très concentrée autour de son mode, par rapport aux variations de la loi prédictive $f_{[Y_0|Y^n]}$ [56] (voir (C.5)). On utilise alors l'approximation suivante de la distribution prédictive de Y_0,

$$f_{[Y_0|Y^n]} \approx f_{[Y_0|Y^n,\widehat{\beta},\widehat{\sigma^2},\widehat{\psi}]}.$$

La complexité algorithmique de la méthode de maximisation de l'évidence est du même ordre que pour le maximum de vraisemblance, en $\mathcal{O}(n^3)$. L'obtention des valeurs optimales des paramètres est facilitée si on peut calculer le gradient de (C.6).

– Hypothèses simplificatrices sur la loi *a priori* des paramètres $f_{(\beta,\sigma^2,\psi)}(\cdot)$ [144]. Si la loi *a priori* des paramètres a une forme simple, on peut parfois obtenir explicitement la loi prédictive :
– si σ^2 et ψ sont supposés connus, la loi de départ (C.3) devient

$$\left[\begin{pmatrix} Y_0 \\ Y^n \end{pmatrix} \middle| \beta \right] \sim \mathcal{N}_{1+n} \left[\begin{pmatrix} {}^t m_0 \\ M \end{pmatrix} \beta, \sigma^2 \begin{pmatrix} 1 & {}^t r_0 \\ r_0 & R \end{pmatrix} \right].$$

Si l'on suppose une loi *a priori* normale pour β, la loi prédictive obtenue est aussi normale (voir [144]). Un fait intéressant est que mettre un *a priori* non-informatif uniforme sur β,

$$\beta \sim 1,$$

qualifié d'*impropre* car la distribution uniforme sur $\mathbb{R}^p$ n'est pas une loi de probabilité, donne la loi prédictive

$$[Y_0|Y^n] \sim \mathcal{N}(\widehat{Y}_0, \mathrm{EQM}(x_0)),$$

où $\widehat{Y}_0$ et $\mathrm{EQM}(x_0)$ sont donnés respectivement par les formules (2.23) et (2.24).

– si seul ψ est connu, partant de la loi

$$\left[\begin{pmatrix} Y_0 \\ Y^n \end{pmatrix} \middle| \beta, \sigma^2\right] \sim \mathcal{N}_{1+n}\left[\begin{pmatrix} {}^t m_0 \\ M \end{pmatrix} \beta, \sigma^2 \begin{pmatrix} 1 & {}^t r_0 \\ r_0 & R \end{pmatrix}\right],$$

et utilisant la relation

$$f_{[\beta,\sigma^2]} = f_{[\beta|\sigma^2]} f_{\sigma^2},$$

on peut, suivant les lois *a priori* mises sur chacun des membres $[\beta|\sigma^2]$ et $[\sigma^2]$, obtenir pour loi prédictive une loi de Student. Notons qu'en mettant un *a priori* uniforme sur $[\beta, \sigma^2]$, on peut retrouver $\widehat{Y}_0$ comme moyenne de la loi prédictive (voir [144] pour les détails).

– si aucun des paramètres β, σ^2, ψ n'est connu, on peut faire l'hypothèse simplificatrice que les paramètres d'échelle β et σ^2 sont indépendants du paramètre de corrélation ψ, *i.e.*

$$f_{[\beta,\sigma^2,\psi]} = f_{[\beta,\sigma^2]} f_{\psi}.$$

En mettant un *a priori* sur $[\beta, \sigma^2]$ comme à l'item précédent, on peut obtenir la loi prédictive vue comme une loi conditionnelle au paramètre inconnu ψ, notée $[Y_0|Y^n, \psi]$. La densité de la loi prédictive s'écrit donc

$$\begin{aligned} f_{[Y_0|Y^n]} &= \int f_{[Y_0,\psi|Y^n]} \mathrm{d}\psi \\ &= \int f_{[Y_0|Y^n,\psi]} f_{[\psi|Y^n]} \mathrm{d}\psi, \end{aligned}$$

avec la densité *a posteriori* de ψ $f_{[\psi|Y^n]}(\cdot)$ qui peut s'obtenir en écrivant

$$\begin{aligned} f_{[\psi|Y^n]} &= \int f_{[\beta,\sigma^2,\psi|Y^n]} \mathrm{d}\beta \, \mathrm{d}\sigma^2 \\ &= \int f_{[Y^n|\beta,\sigma^2,\psi]} f_{[\beta,\sigma^2,\psi]} \mathrm{d}\beta \, \mathrm{d}\sigma^2, \end{aligned}$$

une intégrale $(p+1)$-dimensionnelle qui peut se traiter en utilisant des lois *a priori* simples [144].

Remarque C.2.1 *Si l'ensemble des paramètres $\{\beta, \sigma^2, \psi\}$ est supposé connu, l'interprétation bayésienne du krigeage est la suivante : on se donne un* a priori *sur le modèle $Y(x)$, sous la forme d'un processus gaussien de paramètres connus β, σ^2, ψ. Après avoir observé le vecteur des réponses Y^n, le processus* a posteriori $[Y(x)|Y^n]$ *est aussi un processus gaussien, dont la moyenne et la covariance* a posteriori *sont données par la proposition 2.2.12.*

En résumé, l'approche bayésienne permet l'évaluation de l'incertitude liée aux paramètres du modèle, au prix de calculs numériques souvent coûteux.

Annexe D

Cokrigeage

Nous nous intéressons dans ce paragraphe à la modélisation par krigeage dans le cas où l'on observe simultanément plusieurs réponses d'un même système, et l'on souhaite prendre en compte le fait que les réponses sont corrélées. Ce type de modélisation, appelée *cokrigeage* [26, 144], utilise des processus vectoriels. Nous présentons dans un premier temps le modèle de cokrigeage, puis le calcul du prédicteur est effectué dans une seconde partie.

D.1 Modèle de cokrigeage

Dans le modèle de cokrigeage, les q réponses du système sont chacune modélisées sous la forme

$$Y_i(x) = {}^t m_i(x)\beta_i + Z_i(x), \quad i = 1, \ldots, q, \tag{D.1}$$

avec $m_i(\cdot)$ une fonction connue à valeurs dans $\mathbb{R}^{p_i}$, $\beta_i \in \mathbb{R}^{p_i}$ un vecteur de paramètres inconnu, et $Z_i(\cdot)$ un processus gaussien stationnaire de moyenne nulle et fonction de covariance $k_i(\cdot)$ connue. Chaque réponse est donc modelisée comme en (2.18) dans le cas d'un modèle de krigeage avec une seule réponse, mais les processus aléatoires $Z_i(\cdot)$ sont liés entre eux à travers une structure de covariance : le processus

$$Z(x) = {}^t(Z_1(x), \ldots, Z_q(x))$$

est supposé *multidimensionnel stationnaire*, de moyenne nulle

$$\mathbb{E}\left\{{}^t(Z_1(x), \ldots, Z_q(x))\right\} = 0_{q\times 1} \quad \forall x \in \mathcal{X},$$

et *fonctions de covariance croisées stationnaires* définies par

$$k_{ij}(\tau) = \mathbb{E}\left\{Z_i(x)Z_j(x+\tau)\right\}.$$

Notons que $k_{ii}(\tau) = k_i(\tau)$.

Remarque D.1.1 *[144] Il faut prendre garde au fait que, contrairement aux fonctions de covariance stationnaires classiques, on a en général $k_{ij}(\tau) \neq k_{ij}(-\tau)$ si $i \neq j$. Cependant, $k_{ij}(\tau) = k_{ji}(-\tau)$.*

Les fonctions de covariance croisées continues vérifient une propriété semblable au théorème 2.1.27 (de Bochner).

Proposition D.1.2 *[26] Une fonction de covariance croisée stationnaire continue $k_{ij}(\cdot)$ admet la représentation spectrale*

$$k_{ij}(\tau) = \int \mathrm{e}^{2\pi i\langle u,\tau\rangle} F_{ij}(\mathrm{d}u), \tag{D.2}$$

où $F_{ij}(\mathrm{d}u)$ est appelée mesure spectrale croisée.

Les fonctions $k_{ij}(\cdot)$ ne peuvent pas être choisies arbitrairement dans la famille des fonctions de covariance, mais doivent être compatibles entre elles pour assurer que la variance de toute combinaison linéaire des processus Z_i est positive.

Théorème D.1.3 *[26] Soit $\{k_{ij}(\cdot), 1 \leq i,j \leq q\}$ une famille de fonctions continues s'écrivant sous la forme (D.2). Il existe un processus multidimensionnel stationnaire d'ordre 2 ayant les $k_{ij}(\cdot)$ comme fonctions de covariance croisée si, et seulement si, la matrice $(F_{ij}(B))_{1\leq i,j\leq q}$ est semi-définie positive pour tout ensemble de Borel $B \subset \mathbb{R}^q$.*

Pour des éléments sur la régularité des processus aléatoires multivariés, on pourra consulter [17].

D.2 Calcul du prédicteur de cokrigeage

Supposons, pour tout $i = 1, \ldots, q$, que la réponse $y_i(\cdot)$ a été observée en $x_1^i, \ldots, x_{n_i}^i$ (l'ensemble d'observation peut donc être différent selon la réponse). Notons $Y_i^{n_i} = {}^t(Y_i(x_1^i), \ldots, Y_i(x_{n_i}^i))$ le vecteur aléatoire des observations correspondant à la sortie i, et $y_i^{n_i}$ le vecteur des valeurs observées. On souhaite prédire la valeur des réponses en un nouveau point x_0. Supposons pour simplifier, quitte à réarranger la numérotation des réponses, que l'on souhaite prédire $Y_1(x_0)$. La démarche est la même que dans le cas d'une seule réponse : on recherche un prédicteur linéaire de la forme

$$\widehat{Y}_1(x_0) = \sum_{i=1}^{n_1} \lambda_{i1} Y_1(x_i^1) + \sum_{i=1}^{n_2} \lambda_{i2} Y_2(x_i^2) + \cdots + \sum_{i=1}^{n_q} \lambda_{iq} Y_q(x_i^q) = {}^t c_0 Y^n,$$

avec

$$c_0 = {}^t(\lambda_{11}, \ldots, \lambda_{n_1 1}, \ldots, \lambda_{1q}, \ldots, \lambda_{n_q q})$$

et

$${}^t Y^n = ({}^t Y_1^{n_1}, \ldots, {}^t Y_q^{n_q}).$$

On souhaite que le prédicteur soit sans biais par rapport à $Y_1(x_1^1), \ldots, Y_1(x_{n_1}^1)$. Une manière de remplir cette condition est donnée par

$$\sum_{i=1}^{n_1} \lambda_{i1} = 1 \text{ et } \sum_{i=1}^{n_j} \lambda_{ij} = 0 \text{ pour } j = 2, \ldots, q.$$

On cherche sous ces hypothèses à minimiser l'erreur quadratique moyenne

$$\begin{aligned} \mathbb{E}\left[\left(\widehat{Y}_1(x_0) - Y_1(x_0)\right)^2\right] &= \sum_{l,m=1}^{q} \sum_{i,j=1}^{n_l} \lambda_{il}\lambda_{jm} k_{lm}(x_i^l - x_j^m) - 2\sum_{l=1}^{q}\sum_{i=1}^{n_l} \lambda_{il} k_{l1}(x_i^l - x_0) + k_{11}(0) \\ &= {}^t c_0 K c_0 - 2\,{}^t c_0 k_0 + k_{11}(0), \end{aligned}$$

avec

$$k_0 = {}^t\Big(k_{11}(x_1^1 - x_0), \ldots, k_{11}(x_{n_1}^1 - x_0), \ldots, k_{q1}(x_1^q - x_0), \ldots, k_{q1}(x_{n_q}^q - x_0)\Big)$$

et

$$K = \begin{pmatrix} K_1 & K_{12} & \ldots & K_{1q} \\ {}^tK_{12} & K_2 & \ldots & K_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ {}^tK_{1q} & {}^tK_{2q} & \ldots & K_q \end{pmatrix},$$

où $K_i = \mathrm{cov}(Y_i^{n_i})$ pour $i = 1, \ldots, q$ et $K_{ij} = \mathrm{cov}(Y_i^{n_i}, Y_j^{n_j})$ pour $1 \leq i < j \leq q$. Le système de cokrigeage à résoudre est donc

$$\begin{cases} \underset{c_0 \in \mathbb{R}^n}{\mathrm{argmin}}\, {}^tc_0 K c_0 - 2{}^tc_0 k_0 \\ {}^tMc_0 = m_0, \end{cases} \tag{D.3}$$

avec

$$\begin{aligned} M &= \begin{pmatrix} M_1 & \ldots & 0_{n_1 \times p_q} \\ \vdots & \ddots & \vdots \\ 0_{n_q \times p_1} & \ldots & M_q \end{pmatrix} \\ {}^tm_0 &= \big({}^tm_1(x_0), 0_{1\times(p-p_1)}\big), \end{aligned}$$

où $M_i = {}^t\big(m_i(x_1^i), \ldots, m_i(x_{n_i}^i)\big)$ pour $i = 1, \ldots, q$, et $p = \sum_{i=1}^q p_i$. On reconnaît la forme du système de krigeage universel (2.22), où les matrices de corrélation stationnaires R et r_0 sont remplacées par les matrices « de covariance » K et k_0. On a vu que la solution du système (D.3) est donnée par

$$\widehat{Y}_1(x_0) = {}^tm_0\widehat{\beta} + {}^tk_0K^{-1}(Y^n - M\widehat{\beta}),$$

avec $\widehat{\beta} = ({}^tMK^{-1}M)^{-1}{}^tMK^{-1}Y^n$ l'estimateur des moindres carrés pondérés de $\widehat{\beta}$, et

$$\mathrm{EQM}(x_0) = k_{11}(0) - {}^tk_0K^{-1}k_0 + {}^t\gamma({}^tMK^{-1}M)^{-1}\gamma,$$

où $\gamma = m_0 - {}^tMK^{-1}k_0$.

Remarque D.2.1 *[144] Le même raisonnement que dans le cas d'une seule réponse permet de retrouver l'expression du prédicteur quand l'ensemble des paramètres du modèle sont supposés connus. D'après le théorème 2.2.4, le meilleur prédicteur de $Y_0 = Y_1(x_0)$ est donné par*

$$\widehat{Y}_1(x_0) = \mathbb{E}\left\{Y_0 | Y_1^{n_1} = y_1^{n_1}, \ldots, Y_q^{n_q} = y_q^{n_q}\right\}. \tag{D.4}$$

Cette espérance conditionnelle se calcule aisément dans le cas gaussien. Or, d'après le modèle (D.1),

$${}^t\big(Y_0, Y_1^{n_1}, \ldots, Y_q^{n_q}\big) \sim \mathcal{N}_{1+\sum_{i=1}^q n_i}\left(M'\beta, K'\right),$$

avec

$$M' = \begin{pmatrix} {}^tm_1(x_0) & \ldots & 0_{1\times p_q} \\ M_1 & \ldots & 0_{n_1 \times p_q} \\ \vdots & \ddots & \vdots \\ 0_{n_q \times p_1} & \ldots & M_q \end{pmatrix},$$

$\beta = ({}^t\beta_1, \ldots, {}^t\beta_q)$ *et*

$$K' = \begin{pmatrix} \sigma_{11}^2 & {}^tk_1 & {}^tk_{12} & \ldots & {}^tk_{1q} \\ k_1 & K_1 & K_{12} & \ldots & K_{1q} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ k_{1q} & {}^tK_{1q} & {}^tK_{2q} & \ldots & K_q \end{pmatrix},$$

où $\sigma_{11}^2 = k_{11}(0)$, $k_1 = \mathrm{cov}(Y_1^{n_1}, Y_0)$ *et* $k_{1i} = \mathrm{cov}(Y_i^{n_i}, Y_0)$ *pour* $i = 2, \ldots, q$. *D'après le corollaire 2.2.14, l'espérance conditionnelle est donnée par*

$$\widehat{Y}_1(x_0) = {}^tm_0\beta + {}^tk_0K^{-1}(Y^n - M\beta).$$

Exemple D.2.2 *(Utilisation des dérivées)[144, 178] Supposons que l'on souhaite faire de la prédiction d'un processus gaussien stationnaire et que l'on dispose d'information* a priori *sur les valeurs des dérivées (partielles) du processus en certains points du domaine* $\mathcal{X}$, *données par la physique du système. On peut alors utiliser le cokrigeage pour tenir compte de cette information : le système est modélisé par le processus*

$$Y(x) = {}^tm(x)\beta + Z(x),$$

avec $m(\cdot)$ *une fonction connue et dérivable à valeurs dans* $\mathbb{R}^p$, $\beta \in \mathbb{R}^p$ *un vecteur de paramètres inconnu, et* $Z(\cdot)$ *un processus gaussien stationnaire de moyenne nulle et fonction de covariance* $k(\cdot)$ *connue. Sous certaines conditions de régularité de la fonction de covariance (§ 2.1.2.2), le processus* Y *est dérivable et le processus « i^e dérivée partielle » est donné par*

$$Y_i(x) = {}^tm_i(x)\beta + Z_i(x), \quad i = 1, \ldots, d,$$

avec $m_i(x) = \frac{\partial m(x)}{\partial x_i}$ *et* Z_i *un processus gaussien centré dont la fonction de covariance est calculée ci-dessous. La prédiction en un nouveau point* $Y(x_0)$ *s'obtient par cokrigeage en utilisant l'ensemble des observations disponibles pour chacun des processus* $Y, Y_1, \ldots, Y_d$. *Les covariances croisées s'obtiennent à partir des formules*

$$\begin{aligned} \mathrm{cov}(Y(x), Y_i(x')) &= \frac{\partial k(x,x')}{\partial x_i'}; \\ \mathrm{cov}(Y_i(x), Y_j(x')) &= \frac{\partial^2 k(x,x')}{\partial x_i \partial x_j'}. \end{aligned}$$

Notons que la méthode se généralise aux dérivées d'ordre quelconque si la moyenne et la fonction de covariance du processus sont choisies suffisamment régulières.

Annexe E

Krigeage intrinsèque

L'hypothèse d'un processus aléatoire stationnaire peut s'avérer mal adaptée pour la modélisation de certains systèmes, alors qu'une simple fonction polynômiale par morceaux serait adéquate. Nous avons vu au chapitre 1 que les splines plaque mince (polynômiales par morceaux) sont équivalentes au krigeage intrinsèque : il est donc intéressant de se placer du point de vue stochastique afin de disposer de tout l'appareillage probabiliste. Le krigeage intrinsèque modélise le système par une classe de processus plus générale que les processus stationnaires, les *processus intrinsèques (intrinsic random fields, IRFs)*, dont les incréments sont stationnaires.

Dans un premier temps, nous présentons les processus aléatoires intrinsèques habituellement utilisés, qui correspondent en toute rigueur aux processus intrinsèques d'ordre 0. Puis nous passons à la généralisation aux processus intrinsèques d'ordre q, et finalement donnons l'expression du prédicteur. Dans toute la suite, les processus seront supposés à valeurs réelles.

E.1 Processus aléatoires intrinsèques (d'ordre 0)

Définition E.1.1 *[26] Un processus aléatoire* $\{Y(x), x \in \mathbb{R}^d\}$ *est dit* intrinsèque d'ordre 0 (Intrinsic Random Field, IRF–0) *si le processus incrémentiel*

$$Y_h(x) = Y(x+h) - Y(x)$$

est stationnaire pour tout h*, ou de manière équivalente, si*

- $\mathbb{E}\left[Y(x+h) - Y(x)\right] = m(h)$;
- $\mathrm{var}\left[Y(x+h) - Y(x)\right] = 2\gamma(h)$.

Le terme $m(h)$ *est appelé* dérive (drift) *du processus intrinsèque. La fonction* $2\gamma(\cdot)$ *est appelée* variogramme*, et* $\gamma(\cdot)$ *est appelée* semi-variogramme.

Remarque E.1.2

- *On peut montrer que la dérive* $m(h)$ *est une fonction linéaire de* h *:* $m(h) = \langle a, h\rangle$*, avec* $a \in \mathbb{R}^d$ *fixé. Sans perte de généralité, on pourrait supposer* $m(h) = 0 \quad \forall h$ *: ce point sera expliqué dans la remarque E.2.3 ;*
- *dans la littérature, c'est la fonction* $\gamma(\cdot)$ *qui est parfois appelée variogramme ;*
- *si* Y *est un processus intrinsèque, la variance de la v.a.* $\sum_{i=1}^n \lambda_i Y(x_i)$ *est finie si, et*

seulement si, $\sum_{i=1}^{n} \lambda_i = 0$. *À l'aide du variogramme, on obtient*

$$\begin{aligned} \operatorname{var}\left(\sum_{i=1}^{n} \lambda_i Y(x_i)\right) &= -\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j \gamma(x_i - x_j)\,; \\ \operatorname{cov}\left(\sum_{i=1}^{n} \lambda_i Y(x_i), \sum_{j=1}^{N} \mu_j Y(x_j)\right) &= -\sum_{i=1}^{n}\sum_{j=1}^{N} \lambda_i \mu_j \gamma(x_i - x_j), \end{aligned}$$

si $\sum_{i=1}^{n} \lambda_i = 0$ *et* $\sum_{j=1}^{N} \mu_j = 0$.

Proposition E.1.3 *Le variogramme vérifie les propriétés suivantes :*
- $2\gamma(h) = 2\gamma(-h)$;
- $2\gamma(h) \geq 0$;
- $2\gamma(0) = 0$.

Ces conditions ne sont cependant pas suffisantes pour définir un variogramme admissible. Une condition nécessaire et suffisante sera donnée par le théorème E.1.5.

Définition E.1.4 *Une fonction symétrique* $g : \mathbb{R}^d \to \mathbb{R}$ *est dite* conditionnellement semi-définie négative *si* $\forall l, \forall\, x_1, \ldots, x_l \in \mathbb{R}^d, \forall\, \lambda_1, \ldots \lambda_l \in \mathbb{R}$ *vérifiant* $\sum_{i=1}^{l} \lambda_i = 0$,

$$\sum_{i,j=1}^{l} \lambda_i \lambda_j g(x_i - x_j) \leq 0.$$

Le théorème suivant est l'analogue du théorème 2.1.11 pour les processus stationnaires.

Théorème E.1.5 *[31] Toute fonction conditionnellement semi-définie négative est le variogramme d'un processus aléatoire intrinsèque.*

Faisons maintenant le lien avec les processus stationnaires.

Proposition E.1.6 *[31] Si le processus* Y *est stationnaire de fonction de covariance* $k(\cdot)$, *alors il est intrinsèquement stationnaire et*

$$2\gamma(h) = 2\left(k(0) - k(h)\right).$$

La réciproque est fausse.

Exemple E.1.7 *[26] Le mouvement brownien* $\{W(x), x \in \mathbb{R}^d\}$ *vérifie* $\operatorname{var}\left[W(x+h) - W(x)\right] = \|h\|$, *mais* $\operatorname{cov}(W(x), W(x')) = (\|x\| + \|x'\| - \|x - x'\|)/2$, *qui n'est pas une fonction de* $x - x'$.

L'outil de base du cas stationnaire, la fonction de covariance $k(\cdot)$, est remplacée par le variogramme $\gamma(\cdot)$. On y gagne en généralité, la famille des variogrammes étant plus large que la famille des fonctions de covariance. Cependant, les calculs de covariances sont maintenant limités aux combinaisons linéaires admissibles. Dans la pratique, c'est parfois le variogramme qui est utilisé comme outil de base d'analyse : on consultera [26, 31] pour une présentation beaucoup plus détaillée du variogramme.

E.2 Processus intrinsèques d'ordre q $(q \in \mathbb{N})$

Nous présentons ici la généralisation du paragraphe précédent. Avant de lire la suite, on peut retourner au paragraphe 1.3.1 pour revoir les mesures admissibles d'ordre q et les fonctions conditionnellement semi-définies positives d'ordre q.

E.2.1 Définition et propriétés

Introduisons tout d'abord quelques notations. Pour $x = (x_{[1]}, \ldots, x_{[d]}) \in \mathbb{R}^d$, $l = (l_1, \ldots, l_d) \in \mathbb{N}^d$, on notera x^l le monôme $x_{[1]}^{l_1} \ldots x_{[n]}^{l_n}$ et $|l| = l_1 + \cdots + l_d$ le degré de x^l. On vérifie de façon immédiate que la mesure $\lambda = \sum_{i=1}^n \lambda_i \delta_{x_i}$ est admissible d'ordre q ($\lambda \in \Lambda_q^d$) si, et seulement si

$$\sum_{i=1}^n \lambda_i x_i^l = 0 \quad \forall l,\ |l| = 0, 1, \ldots, q,$$

autrement dit si tous les moments de la mesure λ d'ordre q ou moins sont nuls.

Définition E.2.1 *On appelle* mesure translatée *d'une mesure discrète λ par le vecteur h, notée $\tau_h\lambda$, la mesure possédant les mêmes poids que λ, mais appliqués aux points du support de λ translatés du vecteur h : si $\lambda = \sum_{i=1}^n \lambda_i \delta_{x_i}$, alors*

$$\tau_h\lambda = \sum_{i=1}^n \lambda_i \delta_{x_i+h}.$$

L'action de la mesure translatée sur une fonction f de $\mathbb{R}^d$ sera donc

$$\tau_h\lambda \cdot f = \sum_{i=1}^n \lambda_i f(x_i + h).$$

Nous pouvons maintenant donner une définition simple d'un processus aléatoire intrinsèque d'ordre q. Une définition plus abstraite est possible, à base de classes d'équivalence [26, 178] (les processus définis ici sont en fait des représentants de ces classes).

Définition E.2.2 *Un processus aléatoire $Y(x)$ est dit* intrinsèque d'ordre q (IRF–q) *si pour toute mesure admissible $\lambda = \sum_{i=1}^n \lambda_i \delta_{x_i} \in \Lambda_q^d$, le processus aléatoire*

$$Y_\lambda(x) = \tau_x\lambda \cdot Y = \sum_{i=1}^n \lambda_i Y(x_i + x)$$

est stationnaire de moyenne nulle.

Cette définition équivaut à
- $\mathbb{E}[Y_\lambda(x)] = 0$ (voir la remarque E.2.3) ;
- $\mathbb{E}[Y_\lambda(x)Y_\lambda(x')] = k_\lambda(x - x') \quad \forall x, x' \in \mathbb{R}^d, \lambda \in \Lambda_q^d$.

Remarque E.2.3 *[26, 31]*
- *Le modèle intrinsèque usuel du paragraphe E.1 correspond au cas $q = 0$ (la condition de la Définition E.1.1 est satisfaite en prenant $\lambda = \delta_h - \delta_0 \in \Lambda_0^d$) ;*

- *puisque* $\Lambda_{q+1}^d \subset \Lambda_q^d$, *un IRF–q est aussi un IRF–*$(q+1)$. *Un processus stationnnaire est donc intrinsèque à tout ordre (formellement, un processus stationnaire correspond au cas* $q = -1$*) ;*
- *l'hypothèse de moyenne nulle du processus* Y_λ, *faite pour simplifier la présentation, ne nuit pas à la généralité : la moyenne d'un IRF–q est en effet un polynôme de degré au plus q, qui est filtré par l'action de la mesure* $\tau_x\lambda$. *On voit ici apparaître les classes d'équivalence : un IRF–q n'est en fait défini qu'à un polynôme de degré q près ;*
- *par définition d'un IRF–q* Y, $\mathbb{E}[Y_\lambda(x)]^2 < \infty \quad \forall \lambda \in \Lambda_q^d$ *; cependant,* $\mathbb{E}[Y(x)]^2$ *peut être infinie, ou dépendre de* x. *L'hypothèse de stationnarité des incréments autorise donc des non-stationnarités de la moyenne et de la variance.*

Exemple E.2.4 *[26, 31]*

- *Si* Y *est un processus stationnaire défini sur* $\mathbb{R}$, *de moyenne nulle, son intégrale*

$$T_0(x) = \int_0^x Y(t)\,\mathrm{d}t$$

*est un IRF–*0. *Plus généralement, l'intégrale* $(q+1)^e$ *d'un processus stationaire* Y *sur* $\mathbb{R}$ *de moyenne nulle, définie par*

$$T_q(x) = \int_0^x \frac{(x-t)^q}{q!}\,Y(t)\,\mathrm{d}t,$$

est un IRF–q ;

- *inversement, si un IRF–q défini sur* $\mathbb{R}$ *est différentiable* $(q+1)$ *fois, sa dérivée d'ordre* $(q+1)$ *est stationnaire.*

E.2.2 Covariance généralisée

La structure de corrélation d'un processus stationnaire est définie par sa fonction de covariance, celle d'un processus intrinsèque par son variogramme. Plus généralement, la structure de corrélation d'un IRF–q est caractérisée par une *fonction de covariance généralisée*. La famille des covariances généralisées est plus grande que celle des covariances, mais la covariance n'est définie que pour les combinaisons linéaires admissibles d'ordre q du processus.

Définition E.2.5 *Soit* Y *un IRF–q. Une fonction* $k(\cdot)$, *définie sur* $\mathbb{R}^d$, *est appelée* fonction de covariance généralisée (CG, *ou* Generalized Covariance, GC) *de* Y *si*

$$\mathbb{E}\left[(\lambda \cdot Y)(\mu \cdot Y)\right] = \sum_{i=1}^{n}\sum_{j=1}^{N} \lambda_i \mu_j k(x_i - t_j) \tag{E.1}$$

pour tous n, N, *et toute paire de mesures admissibles* $\lambda = \sum_{i=1}^n \lambda_i \delta_{x_i}$ *et* $\mu = \sum_{j=1}^N \mu_i \delta_{t_j} \in \Lambda_q^d$.

Remarque E.2.6 *[26]*

- *Il suffit de vérifier la condition ci-dessus pour* $\lambda = \mu$, *comme on peut le constater en développant* $[(\lambda+\mu)\cdot Y]^2$. *La condition que doit vérifier une CG est donc*

$$\mathbb{E}[\lambda \cdot Y]^2 = \sum_{i,j=1}^{n} \lambda_i \lambda_j k(x_i - x_j) \qquad \forall \lambda = \sum_{i=1}^{n} \lambda_i \delta_{x_i} \in \Lambda_q^d\,; \tag{E.2}$$

- *en pratique, la fonction $k(\cdot)$ s'utilise comme une fonction de covariance ordinaire, mais les formules (E.1) et (E.2) ne sont valables que pour des combinaisons linéaires admissibles d'ordre q ;*
- *la famille des fonctions de covariance généralisée d'ordre q coïncide avec la famille des fonctions $k(\cdot)$ symétriques sur $\mathbb{R}^d$ et satisfaisant la condition*

$$\sum_{i,j=1}^{n} \lambda_i \lambda_j k(x_i - x_j) \geq 0,$$

pour toute mesure admissible $\lambda = \sum_{i=1}^{n} \lambda_i \delta_{x_i} \in \Lambda_q^d$, autrement dit avec la famille des fonctions conditionnellement semi-définies positives d'ordre q (Définition 1.3.6).

Théorème E.2.7 *[26] Tout IRF–q possède une covariance généralisée $k(\cdot)$, unique à une classe d'équivalence près : toute autre CG s'écrit $k(h) + p(h)$, où $p(\cdot)$ est un polynôme pair de degré au plus $2q$.*

Exemple E.2.8 *[26]*

- *Si Y est stationnaire de covariance $k(\cdot)$, alors $k(\cdot)$ est aussi une covariance généralisée ;*
- *si Y est un IRF–0, alors $k(h) = -\gamma(h) + a$, avec $a \in \mathbb{R}$: la covariance généralisée vaut $-\gamma(\cdot)$ à un polynôme pair de degré 0 près,* i.e. *une constante ;*
- *si Y est stationnaire de moyenne nulle sur $\mathbb{R}$, nous avons vu que sa $(q+1)^e$ intégrale*

$$T_q(x) = \int_0^x \frac{(x-t)^q}{q!} Y(t)\,\mathrm{d}t$$

est un IRF–q. On peut montrer que sa covariance généralisée s'écrit

$$k_q(h) = (-1)^{q+1} \int_0^h \frac{(h-t)^{2q+1}}{(2q+1)!} k(t)\,\mathrm{d}t,$$

où $k(\cdot)$ est la fonction de covariance stationnaire de Y.

E.2.3 Représentation spectrale

Il existe un analogue au théorème 2.1.27 (de Bochner) pour les covariances généralisées.

Théorème E.2.9 *[26, 84, 161] Une fonction continue $k(\cdot)$ est la covariance généralisée d'un IRF–q défini sur $\mathbb{R}^d$ si, et seulement si,*

$$k(h) = \int \left\{ \cos({}^t vh) - P_q({}^t vh) \mathbb{1}_{B_d}(v) \right\} \mu(\mathrm{d}v) + Q(h), \tag{E.3}$$

avec $P_q(x) = \sum_{i=0}^{q} \frac{(-x^2)^i}{(2i)!}$, $\mathbb{1}_{B_d}(\cdot)$ la fonction indicatrice de la boule unité de $\mathbb{R}^d$, $Q(\cdot)$ un polynôme de degré $\leq 2q$ quelconque, et μ une mesure positive et symétrique vérifiant

$$\int_{\mathbb{R}^d} \frac{\|v\|^{2q+2}}{\left(1 + \|v\|^2\right)^{q+1}} \mu(\mathrm{d}v) < \infty, \tag{E.4}$$

appellée mesure spectrale *de l'IRF–q.*

Remarque E.2.10 *[26]*

- *Puisque $|\cos({}^t vx) - P_q({}^t vx)| = \mathcal{O}(\|v\|^{2q+2})$, l'intégrale (E.3) est bien définie pour une mesure μ satisfaisant (E.4) ;*
- *le choix de B_d est arbitraire : tout voisinage borné de l'origine conviendrait.*

E.3 Prédicteur de krigeage intrinsèque

Il s'agit maintenant d'écrire les équations du prédicteur et de l'EQM de krigeage intrinsèque. Le processus $Y(x)$ est supposé être un IRF–q (q connu), de covariance généralisée $k(\cdot)$ connue. On souhaite prédire $Y_0 = Y(x_0)$ en utilisant un prédicteur linéaire de la forme

$$\widehat{Y}(x_0) = \sum_{i=1}^{n} \lambda_i Y(x_i).$$

Un choix convenable des coefficients λ_i garantit que le terme d'erreur $\widehat{Y}(x_0) - Y(x_0)$ est une combinaison linéaire admissible d'ordre q, dont on peut donc calculer la norme quadratique.

E.3.1 Calcul de l'erreur quadratique moyenne

Soit $\delta_{(x_0)} = \sum_{i=1}^{n} \lambda_i \delta_{x_i} - \delta_{x_0}$ une mesure admissible d'ordre q ($\delta_{(x_0)} \in \Lambda_q^d$). Alors

$$\delta_{(x_0)} \cdot Y = \sum_{i=1}^{n} \lambda_i Y(x_i) - Y(x_0) = \widehat{Y}(x_0) - Y(x_0)$$

est une combinaison linéaire admissible d'ordre q, et la formule (E.2) donne l'expression de l'erreur quadratique moyenne,

$$\begin{aligned} \mathbb{E}\left[\left(\widehat{Y}(x_0) - Y(x_0)\right)^2\right] &= \sum_{i,j=1}^{n} \lambda_i \lambda_j k(x_i - x_j) - 2\sum_{i=1}^{n} \lambda_i k(x_i - x_0) + k(0) \\ &= {}^t c_0 K c_0 - 2\,{}^t c_0 k_0 + k(0), \end{aligned}$$

avec $c_0 = {}^t(\lambda_1, \ldots, \lambda_n), K = (k(x_i - x_j))_{1 \le i,j \le n}$ et $k_0 = (k(x_i - x_0))_{1 \le i \le n}$. Le prédicteur de krigeage intrinsèque sera obtenu en minimisant (E.5), sous la contrainte $\delta_{(x_0)} \in \Lambda_q^d$.

E.3.2 Équations du krigeage intrinsèque

Minimiser l'équation (E.5) sachant que $\delta_{(x_0)}$ est une mesure admissible d'ordre q conduit au système

$$\begin{cases} \underset{c_0 \in \mathbb{R}^n}{\operatorname{argmin}}\, {}^t c_0 K c_0 - 2\,{}^t c_0 k_0 \\ {}^t M c_0 = m_0, \end{cases} \tag{E.5}$$

avec

$$\begin{aligned} M &= (x_i^l)_{i=1,\ldots,n}^{|l|=0,\ldots,q}; \\ m_0 &= (x_0^l)_{|l|=0,\ldots,q}, \end{aligned}$$

où les exposants l sont rangés d'une certaine façon. On reconnaît le système de krigeage universel (2.22), où les matrices de corrélation stationnaires R et r_0 sont remplacées par les matrices de covariance généralisées K et k_0 (le remplacement des matrices de covariance par les matrices de corrélation est possible dans le cas stationnaire mais pas dans le cas intrinsèque). On a vu que la solution du système (E.5) est

$$\widehat{Y}_0 = {}^t m_0 \widehat{\beta} + {}^t k_0 K^{-1}(Y^n - M\widehat{\beta}),$$

avec $\widehat{\beta} = ({}^tMK^{-1}M)^{-1}{}^tMK^{-1}Y^n$ l'estimateur des moindres carrés pondérés de $\widehat{\beta}$, et

$$\mathrm{EQM}(x_0) = k(0) - {}^tk_0K^{-1}k_0 + {}^t\gamma({}^tMK^{-1}M)^{-1}\gamma,$$

avec $\gamma = m_0 - {}^tMK^{-1}k_0$.

Remarque E.3.1 *[26, 31]*

- *Le système obtenu est le même que pour le krigeage universel, la covariance généralisée remplaçant la covariance stationnaire ;*
- *le prédicteur et la variance de krigeage ne dépendent que de la classe d'équivalence de la fonction de covariance généralisée considérée ;*
- *il a été supposé implicitement que le système du krigeage intrinsèque (E.5) a toujours une solution. Ceci n'est vrai que si m_0 appartient à l'espace vectoriel généré par les colonnes de M : cette condition devant être vérifiée pour tout m_0, les colonnes de M doivent être linéairement indépendantes (revoir la remarque 1.3.20). Cette condition, ajoutée à une covariance généralisée conditionnellement semi-définie positive d'ordre q et des données distinctes, assure que le système de krigeage intrinsèque admet une solution unique ;*
- *la propriété d'absence de biais est vérifiée par définition d'un IRF–q, car la combinaison linéaire admissible d'ordre q $\widehat{Y}(x_0) - Y(x_0)$ est de moyenne nulle ;*
- *l'utilisation de combinaisons linéaires admissibles implique la propriété d'invariance suivante : l'ajout d'un polynôme quelconque $\sum_{|l|=0}^{q} a_l x^l$ à $Y(x)$ ne change pas la valeur de $\widehat{Y}(x_0) = Y(x_0)$;*
- *afin de calculer l'estimateur « plug-in » de krigeage intrinsèque, il faut estimer les paramètres de covariance. Pour cela, il est impossible d'utiliser le maximum de vraisemblance, car la vraisemblance n'est pas définie pour certaines combinaisons linéaires du processus. On peut cependant utiliser le maximum de vraisemblance restreint.*

Annexe F

Premiers moments (tronqués) de la loi normale mono-dimensionnelle

Nous donnons ici les formules permettant de calculer les moments (tronqués) d'ordre 1 et 2 d'une loi normale univariée. Rappelons tout d'abord que

$$\frac{1}{\sqrt{2\pi}} \int_a^b \exp\left(-\frac{x^2}{2}\right) \mathrm{d}x = \Phi(b) - \Phi(a),$$

où $\Phi(\cdot)$ est la fonction de répartition de la loi normale standard $\mathcal{N}(0,1)$. Posant $u = (a-m)/\sigma$ et $v = (b-m)/\sigma$, on obtient facilement

$$\begin{aligned}
\frac{1}{\sigma\sqrt{2\pi}} \int_a^b x \exp\left\{-\frac{(x-m)^2}{2\sigma^2}\right\} \mathrm{d}x &= \frac{\sigma}{\sqrt{2\pi}} \left[-\exp\left(-\frac{x^2}{2}\right)\right]_u^v + m\left(\Phi(v) - \Phi(u)\right) ; \\
\frac{1}{\sigma\sqrt{2\pi}} \int_a^b x^2 \exp\left\{-\frac{(x-m)^2}{2\sigma^2}\right\} \mathrm{d}x &= (m^2+\sigma^2)\left(\Phi(v) - \Phi(u)\right) + \frac{\sigma^2}{\sqrt{2\pi}} \left[-x\exp\left(-\frac{x^2}{2}\right)\right]_u^v \\
&\quad + \frac{2\sigma m}{\sqrt{2\pi}} \left[-\exp\left(-\frac{x^2}{2}\right)\right]_u^v .
\end{aligned}$$

Annexe G

Formules de discrépance

Nous démontrons ici les formules de discrépance pour l'ajout d'un nouveau point énoncées au § 4.1.1.

G.1 Discrépance L^∞

Pour fixer les idées, considérons tout d'abord le cas fictif $n = 0$ (aucun point échantillonné pour le moment). On cherche à évaluer la discrépance L^∞ au point $y \in [0, 1]$,

$$D_\infty(y) = \sup_{z \in [0,1]} |F_1(z) - F_{\mathcal{U}}(z)|,$$

avec $F_1(\cdot)$ la fonction de répartition empirique de l'échantillon $\{y\}$ et $F_{\mathcal{U}}(\cdot)$ la fonction de répartition de la loi uniforme sur $[0, 1]$.

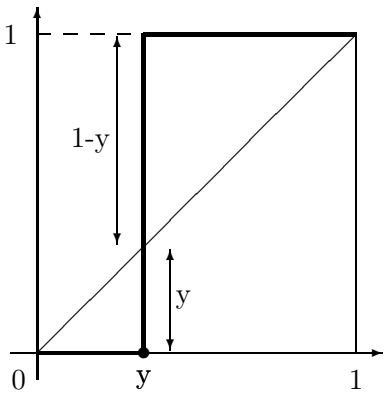

FIGURE G.1 – Calcul de la discrépance au nouveau point y, avec aucun point encore échantillonné. La fonction de répartition empirique est représentée en trait fort, la fonction de répartition uniforme en trait fin.

On remarque sur la figure G.1, où les courbes représentatives des fonctions $F_1(\cdot)$ et $F_{\mathcal{U}}(\cdot)$ sont représentées respectivement en trait fort et en trait fin, pour une valeur arbitraire de y repérée

par un point, que le maximum est atteint au point y, et on obtient

$$D_\infty(y) = \sup(y, 1-y).$$

Revenons alors au cas général (rappelons que l'on a supposé $0 \leq y_1 < y_2 \ \cdots < y_n \leq 1$).

G.1.1 Démonstration de la proposition 4.1.1

Par analogie avec le cas trivial présenté ci-dessus, on remarque tout d'abord que le supremum de la quantité

$$|F_{n+1}(z) - F_{\mathcal{U}}(z)|,$$

avec $F_{n+1}(\cdot)$ la fonction de répartition empirique de l'échantillon $\{y_1, \ldots, y_n, y\}$ et $F_{\mathcal{U}}(\cdot)$ la fonction de répartition de la loi uniforme sur $[0,1]$, est atteint en l'un (ou plusieurs) des points $y_1, \ldots, y_n, y$. Il reste donc à évaluer les quantités $d^-(z) = |F_{n+1}(z^-) - F_{\mathcal{U}}(z^-)|$ et $d^+(z) = |F_{n+1}(z^+) - F_{\mathcal{U}}(z^+)|$ en ces points. Ces quantités dépendent de la position du point y.

– Si $0 \leq y_1 < \ldots y_i < y < y_{i+1} < \cdots < y_n \leq 1$, on obtient le tableau de valeurs suivant.

	y_1	$\ldots$	y_i	y	y_{i+1}	$\ldots$	y_n
d^-	y_1	$\ldots$	$\lvert y_i - \frac{i-1}{n+1}\rvert$	$\lvert y - \frac{i}{n+1}\rvert$	$\lvert y_{i+1} - \frac{i+1}{n+1}\rvert$	$\ldots$	$\lvert y_n - \frac{n}{n+1}\rvert$
d^+	$\lvert y_1 - \frac{1}{n+1}\rvert$	$\ldots$	$\lvert y_i - \frac{i}{n+1}\rvert$	$\lvert y - \frac{i+1}{n+1}\rvert$	$\lvert y_{i+1} - \frac{i+2}{n+1}\rvert$	$\ldots$	$1 - y_n$

– Si $y = y_i$ pour un certain i, les valeurs prises par d^+ et d^- sont données ci-dessous.

	y_1	$\ldots$	y_{i-1}	$y = y_i$	y_{i+1}	$\ldots$	y_n
d^-	y_1	$\ldots$	$\lvert y_{i-1} - \frac{i-2}{n+1}\rvert$	$\lvert y_i - \frac{i-1}{n+1}\rvert$	$\lvert y_{i+1} - \frac{i+1}{n+1}\rvert$	$\ldots$	$\lvert y_n - \frac{n}{n+1}\rvert$
d^+	$\lvert y_1 - \frac{1}{n+1}\rvert$	$\ldots$	$\lvert y_{i-1} - \frac{i-1}{n+1}\rvert$	$\lvert y_i - \frac{i+1}{n+1}\rvert$	$\lvert y_{i+1} - \frac{i+2}{n+1}\rvert$	$\ldots$	$1 - y_n$

Donc,

$$D_\infty^n(y) = \begin{cases} \sup(y, |y - \frac{1}{n+1}|, R_{0,1}), & \text{si } y \in [0, y_1[\,; \\ \sup(|y - \frac{i}{n+1}|, |y - \frac{i+1}{n+1}|, R_{i,i+1}), & \text{si } y \in]y_i, y_{i+1}[, 1 \leq i < n\,; \\ \sup(|y - \frac{n}{n+1}|, 1-y, R_{n,n+1}), & \text{si } y \in]y_n, 1], \end{cases}$$

avec

$$\begin{aligned}
R_{0,1} &= \sup\left(\left|y_1 - \frac{1}{n+1}\right|, \left|y_1 - \frac{2}{n+1}\right|, \ldots, \left|y_i - \frac{i}{n+1}\right|, \left|y_i - \frac{i+1}{n+1}\right|, \ldots, \left|y_n - \frac{n}{n+1}\right|, 1-y_n\right); \\
R_{i,i+1} &= \sup\left(y_1, \left|y_1 - \frac{1}{n+1}\right|, \ldots, \left|y_i - \frac{i-1}{n+1}\right|, \left|y_i - \frac{i}{n+1}\right|, \left|y_{i+1} - \frac{i+1}{n+1}\right|, \left|y_{i+1} - \frac{i+2}{n+1}\right|, \ldots,\right. \\
&\quad \left.\left|y_n - \frac{n}{n+1}\right|, 1-y_n\right) \quad \text{pour } i = 1, \ldots, n-1\,; \\
R_{n,n+1} &= \sup\left(y_1, \left|y_1 - \frac{1}{n+1}\right|, \ldots, \left|y_i - \frac{i-1}{n+1}\right|, \left|y_i - \frac{i}{n+1}\right|, \ldots, \left|y_n - \frac{n-1}{n+1}\right|, \left|y_n - \frac{n}{n+1}\right|\right).
\end{aligned}$$

Remarquons au passage que, pour $i = 1, \ldots, n$,

$$\begin{aligned}
D_\infty^n(y_i) &= \sup\left(y_1, \left|y_1 - \frac{1}{n+1}\right|, \ldots, \left|y_{i-1} - \frac{i-2}{n+1}\right|, \left|y_{i-1} - \frac{i-1}{n+1}\right|, \left|y_i - \frac{i-1}{n+1}\right|,\right. \\
&\quad \left.\left|y_i - \frac{i+1}{n+1}\right|, \left|y_{i+1} - \frac{i+1}{n+1}\right|, \left|y_{i+1} - \frac{i+2}{n+1}\right|, \ldots, \left|y_n - \frac{n}{n+1}\right|, 1-y_n\right).
\end{aligned}$$

On écrit ensuite l'expression de la discrépance sans les valeurs absolues,

$$D_\infty^n(y) = \begin{cases} \sup(y, \frac{1}{n+1} - y, R_{0,1}), & \text{si } y \in [0, y_1[\cap[0, \frac{1}{n+1}[\,; \\ \sup(y, R_{0,1}), & \text{si } y \in [0, y_1[\cap[\frac{1}{n+1}, 1]\,; \\ \sup(\frac{i+1}{n+1} - y, R_{i,i+1}), & \text{si } y \in]y_i, y_{i+1}[\cap[0, \frac{i}{n+1}[\,; \\ \sup(y - \frac{i}{n+1}, \frac{i+1}{n+1} - y, R_{i,i+1}), & \text{si } y \in]y_i, y_{i+1}[\cap[\frac{i}{n+1}, \frac{i+1}{n+1}[\,; \\ \sup(y - \frac{i}{n+1}, R_{i,i+1}), & \text{si } y \in]y_i, y_{i+1}[\cap[\frac{i+1}{n+1}, 1]\,; \\ \sup(1 - y, R_{n,n+1}), & \text{si } y \in]y_n, 1] \cap [0, \frac{n}{n+1}]\,; \\ \sup(y - \frac{n}{n+1}, 1 - y, R_{n,n+1}), & \text{si } y \in]y_n, 1] \cap [\frac{n}{n+1}, 1], \end{cases}$$

ce qui permet déjà de constater que la fonction $D_\infty^n(\cdot)$ est affine par morceaux, de pentes $-1, 0, 1$. En prenant des intervalles encore plus petits, de façon à ne plus avoir de suprema, et en supposant (provisoirement) que D_∞^n est continue (ce qui permet de fermer tous les intervalles), on retrouve la formule de la proposition 4.1.1.

G.1.2 Preuve de la continuité de $D_\infty^n(\cdot)$

Afin de démontrer que $D_\infty^n(\cdot)$ est continue, il suffit de démontrer que D_∞^n est continue aux points $y_1, y_2, \ldots, y_n$, car on vient de voir qu'entre les points $0, y_1, y_2, \ldots, y_n, 1$, D_∞^n est affine de pente $-1, 0$, ou 1. Montrons par exemple que D_∞^n est continue à gauche en $y_i, 1 < i < n$ (les autres cas se démontrent de manière analogue). On a vu que

$$\begin{aligned} D_\infty^n(y_i) \;=\; & \sup\Big(y_1, \Big|y_1 - \frac{1}{n+1}\Big|, \ldots, \Big|y_{i-1} - \frac{i-2}{n+1}\Big|, \Big|y_{i-1} - \frac{i-1}{n+1}\Big|, \Big|y_i - \frac{i-1}{n+1}\Big|, \\ & \Big|y_i - \frac{i+1}{n+1}\Big|, \Big|y_{i+1} - \frac{i+1}{n+1}\Big|, \Big|y_{i+1} - \frac{i+2}{n+1}\Big|, \ldots, \Big|y_n - \frac{n}{n+1}\Big|, 1 - y_n\Big). \end{aligned}$$

Lorsque $y_{i-1} < y < y_i$,

$$\begin{aligned} D_\infty^n(y) \;=\; & \sup\Big(y_1, \Big|y_1 - \frac{1}{n+1}\Big|, \ldots, \Big|y_{i-1} - \frac{i-2}{n+1}\Big|, \Big|y_{i-1} - \frac{i-1}{n+1}\Big|, \Big|y - \frac{i-1}{n+1}\Big|, \\ & \Big|y - \frac{i}{n+1}\Big|, \Big|y_i - \frac{i}{n+1}\Big|, \Big|y_i - \frac{i+1}{n+1}\Big|, \ldots, \Big|y_n - \frac{n}{n+1}\Big|, 1 - y_n\Big). \end{aligned}$$

En faisant tendre y vers y_i dans $D_\infty^n(y)$, on constate que seul le terme $\left|y_i - \frac{i}{n+1}\right|$ n'apparaît pas dans $D_\infty^n(y_i)$, ce qui implique que

$$D_\infty^n(\cdot) \text{ n'est pas continue à gauche en } y_i \Leftrightarrow \Big|y_i - \frac{i}{n+1}\Big| > D_\infty^n(y_i).$$

En particulier, si $D_\infty^n(\cdot)$ n'est pas continue à gauche en y_i, alors

$$\Big|y_i - \frac{i}{n+1}\Big| > \sup\left(\Big|y_i - \frac{i-1}{n+1}\Big|, \Big|y_i - \frac{i+1}{n+1}\Big|\right),$$

ce qui est impossible. Donc $D_\infty^n(\cdot)$ est continue à gauche en y_i. □

G.2 Discrépance L^1

Ordonnons l'échantillon $\{y_1, \ldots, y_n, y\}$ dans l'ordre croissant $y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(n+1)}$. Supposons pour commencer que les $y_i, i = 1, \ldots, n+1$ sont tous distincts. La fonction de répartition empirique s'écrit

$$F_{n+1}(z) = \begin{cases} 0, & \text{si } z \in [0, y_{(1)}[\,; \\ \frac{1}{n+1}, & \text{si } z \in [y_{(1)}, y_{(2)}[\,; \\ \quad\vdots & \\ \frac{n}{n+1}, & \text{si } z \in [y_{(n)}, y_{(n+1)}[\,; \\ 1, & \text{si } z \in [y_{(n+1)}, 1]. \end{cases}$$

Nous cherchons à évaluer

$$D_1^n(y) = \int_{[0,1]} |z - F_{n+1}(z)| \, \mathrm{d}z.$$

$$\begin{aligned} D_1^n(y) &= \int_0^{y_{(1)}} |z| \, \mathrm{d}z + \sum_{i=1}^{n} \int_{y_{(i)}}^{y_{(i+1)}} \left| z - \frac{i}{n+1} \right| \mathrm{d}z + \int_{y_{(n+1)}}^{1} |z-1| \, \mathrm{d}z \\ &= \frac{y_{(1)}^2}{2} + \sum_{i=1}^{n} \left\{ \int_{y_{(i)}}^{\frac{i}{n+1}} \left| z - \frac{i}{n+1} \right| \mathrm{d}z + \int_{\frac{i}{n+1}}^{y_{(i+1)}} \left| z - \frac{i}{n+1} \right| \mathrm{d}z \right\} + \frac{(1 - y_{(n+1)})^2}{2} \\ &= \frac{y_{(1)}^2}{2} + \sum_{i=1}^{n} \left\{ (-1)^{\mathbb{1}_{]\frac{i}{n+1}, \infty[}(y_{(i)})} \frac{(y_{(i)} - \frac{i}{n+1})^2}{2} - (-1)^{\mathbb{1}_{]\frac{i}{n+1}, \infty[}(y_{(i+1)})} \frac{(y_{(i+1)} - \frac{i}{n+1})^2}{2} \right\} + \\ &\quad \frac{(1 - y_{(n+1)})^2}{2}. \end{aligned}$$

Regroupant les termes en $y_{(i)}$, on obtient la formule voulue,

$$D_1^n(y) = \frac{1}{2} \sum_{i=1}^{n+1} \left\{ (-1)^{\mathbb{1}_{]\frac{i}{n+1}, \infty[}(y_{(i)})} \left(y_{(i)} - \frac{i}{n+1} \right)^2 - (-1)^{\mathbb{1}_{]\frac{i-1}{n+1}, \infty[}(y_{(i)})} \left(y_{(i)} - \frac{i-1}{n+1} \right)^2 \right\},$$

qui est bien un polynôme de degré 2 par morceaux continu en y. La proposition 4.1.3 est ainsi démontrée.

G.3 Discrépance L^2

Ordonnons l'échantillon $\{y_1, \ldots, y_n, y\}$ dans l'ordre croissant comme au paragraphe précédent, $y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(n+1)}$. On peut alors calculer

$$\begin{aligned}
D_2^n(y) &= \sqrt{\int_0^1 (z - F_{n+1}(z))^2 \mathrm{d}z} \\
&= \sqrt{\int_0^{y_{(1)}} z^2 \mathrm{d}z + \sum_{i=1}^{n} \int_{y_{(i)}}^{y_{(i+1)}} \left(z - \frac{i}{n+1}\right)^2 \mathrm{d}z + \int_{y_{(n+1)}}^1 (1-z)^2 \mathrm{d}z} \\
&= \sqrt{\frac{y_{(1)}^3}{3} + \frac{1}{3} \sum_{i=1}^{n} \left[\left(y_{(i+1)} - \frac{i}{n+1}\right)^3 - \left(y_{(i)} - \frac{i}{n+1}\right)^3 \right] + \frac{\left(1 - y_{(n+1)}\right)^3}{3}} \\
&= \sqrt{\frac{1}{3} \sum_{i=1}^{n+1} \left[\left(y_{(i)} - \frac{i-1}{n+1}\right)^3 - \left(y_{(i)} - \frac{i}{n+1}\right)^3 \right]},
\end{aligned}$$

qui est une fonction continue en y, et la proposition 4.1.4 est démontrée.

Annexe H

Distance de Wasserstein

Nous définissons ici dans un premier temps la distance de Wasserstein, puis démontrons la formule du critère d'ajout de point proposé à la remarque 4.1.2.

H.1 Définition et propriétés

Nous considérons dans la suite la version L^2 de la distance de Wasserstein [37, 138] par simplicité des calculs associés.

Définition H.1.1 *[67] Soient μ et ν deux lois de probabilité définies sur un même espace probabilisé. On appelle* distance de Wasserstein *entre μ et ν, et on note $W(\mu,\nu)$, la quantité*

$$W(\mu,\nu) = \inf_{X\sim\mu, Y\sim\nu} \sqrt{\mathbb{E}(X-Y)^2} = \inf_{X\sim\mu, Y\sim\nu} \sqrt{\mathbb{E}X^2 + \mathbb{E}Y^2 - 2\mathbb{E}XY}.$$

Les lois μ et ν étant connues, la quantité à minimiser dépend de la loi jointe du couple (X,Y) : la recherche du minimum ci-dessus est équivalente à la recherche de $\sup_{X\sim\mu, Y\sim\nu} \mathbb{E}XY$.

Théorème H.1.2 *(Théorème de la réciproque)[67] La distance de Wasserstein est atteinte pour les v.a. suivantes, définies sur $([0,1], \mathcal{B}([0,1]), \mathrm{d}x)$,*

$$\begin{aligned} X_0(\omega) &= \inf\{x\in\mathbb{R} | \mu(]-\infty, x]\geq\omega)\} = \inf\{x\in\mathbb{R} | F_\mu(x)\geq\omega)\} \\ Y_0(\omega) &= \inf\{x\in\mathbb{R} | \nu(]-\infty, x]\geq\omega)\} = \inf\{x\in\mathbb{R} | F_\nu(x)\geq\omega)\}, \end{aligned}$$

appelées réarrangées croissantes *de μ et ν.*

On a donc

$$W(\mu,\nu) = \sqrt{\mathbb{E}X_0^2 + \mathbb{E}Y_0^2 - 2\mathbb{E}X_0Y_0}.$$

H.2 Critère d'ajout de point utilisant la distance de Wasserstein

Nous allons construire un critère d'ajout de point utilisant les résultats du paragraphe précédent. Considérons un ensemble d'observations distinctes $\{y_1,\ldots,y_n\}$ dans $[0,1]$. On souhaite observer un nouveau point $y = y_{n+1}$ dans $[0,1]$. Afin de guider notre choix, nous allons minimiser la distance de Wasserstein entre la mesure empirique associée à l'échantillon $\{y_1,\ldots,y_n,y_{n+1}\}$ et la mesure uniforme sur $[0,1]$. On considère la statistique d'ordre $y_{(1)} < y_{(2)} < \cdots < y_{(n)} < y_{(n+1)}$.

Proposition H.2.1 *Si $\mu \sim \mathcal{U}([0,1])$ et $\nu_n(y) \sim 1/(n+1)(\sum_{i=1}^n \delta_{y_i} + \delta_y)$, alors*

$$W_n(y) = W(\mu, \nu_n(y)) = \sqrt{\frac{1}{3} + \frac{1}{n+1}\sum_{i=1}^{n+1} y_i^2 - \frac{1}{(n+1)^2}\sum_{i=1}^{n+1}(2i-1)y_{(i)}}\,,$$

où $y_{n+1} = y$.

Preuve *Construisons les réarrangées croissantes du théorème H.1.2. Clairement, pour $\omega \in [0,1]$,*

$$\begin{array}{rcccl} X_0(\omega) & = & \inf\{x \in \mathbb{R} | F_{\mathcal{U}}(x) \geq \omega)\} & = & \omega \\ Y_0(\omega) & = & \inf\{x \in \mathbb{R} | F_{\nu_n(y)}(x) \geq \omega)\} & = & y_{(1)}\mathbb{1}_{[0,\frac{1}{n+1}]}(\omega) + \sum_{i=2}^{n+1} y_{(i)}\mathbb{1}_{]\frac{i-1}{n+1},\frac{i}{n+1}]}(\omega). \end{array}$$

Appelons μ_0 et ν_0 les lois suivies respectivement par X_0 et Y_0. On obtient immédiatement que la densité de la loi jointe du couple (X_0, Y_0) est donnée par

$$\mathrm{d}(\mu_0, \nu_0)(x,y) = \sum_{i=1}^{n+1} \mathbb{1}_{[\frac{i-1}{n+1},\frac{i}{n+1}]}(x)\,\mathrm{d}x\,\delta_{y_{(i)}}(y).$$

On en déduit que

$$\begin{aligned} W_n(y)^2 &= W(\mu_0,\nu_0)^2 \\ &= \mathbb{E}X_0^2 + \mathbb{E}Y_0^2 - 2\mathbb{E}X_0Y_0 \\ &= \int_{[0,1]} x^2 \mathrm{d}x + \frac{1}{n+1}\sum_{i=1}^{n+1} y_{(i)}^2 - 2\int_{\mathbb{R}^2} xy\,\mathrm{d}(\mu_0,\nu_0)(x,y) \\ &= \frac{1}{3} + \frac{1}{n+1}\sum_{i=1}^{n+1} y_i^2 - 2\sum_{i=1}^{n+1} y_{(i)} \int_{[\frac{i-1}{n+1},\frac{i}{n+1}]} x\,\mathrm{d}x \\ &= \frac{1}{3} + \frac{1}{n+1}\sum_{i=1}^{n+1} y_i^2 - \sum_{i=1}^{n+1} y_{(i)}\left[\left(\frac{i}{n+1}\right)^2 - \left(\frac{i-1}{n+1}\right)^2\right] \\ &= \frac{1}{3} + \frac{1}{n+1}\sum_{i=1}^{n+1} y_i^2 - \frac{1}{(n+1)^2}\sum_{i=1}^{n+1}(2i-1)y_{(i)}. \end{aligned}$$

□

Annexe I

Une distance pour mesurer l'éloignement de deux plans

Nous montrons ici que la fonction d_I introduite au chapitre 5 est bien une distance. Rappelons tout d'abord les notations utilisées. Si $A = \{a_1, \ldots, a_n\}$ et $B = \{b_1, \ldots, b_n\}$ sont deux ensembles à n points de $\mathbb{R}^d$, on note

$$\mathrm{d}_I(A, B) = \frac{1}{n} \min_{\tau \in \mathcal{S}_n} \sum_{i=1}^{n} \left\| a_i - b_{\tau(i)} \right\|_2,$$

où $\mathcal{S}_n$ désigne l'ensemble des permutations de $\{1, \ldots, n\}$.

Proposition I.0.2 *La fonction* $\mathrm{d}_I(\cdot, \cdot)$ *est une distance sur* $(\mathbb{R}^d)^n$.

Preuve

– *(symétrie)*

$$\begin{aligned}
\mathrm{d}_I(A, B) &= \frac{1}{n} \min_{\tau \in \mathcal{S}_n} \sum_{i=1}^{n} \left\| a_{\tau^{-1}(i)} - b_i \right\|_2 \\
&= \frac{1}{n} \min_{\tau \in \mathcal{S}_n} \sum_{i=1}^{n} \left\| b_i - a_{\tau^{-1}(i)} \right\|_2 \\
&= \frac{1}{n} \min_{\tau' \in \mathcal{S}_n} \sum_{i=1}^{n} \left\| b_i - a_{\tau'(i)} \right\|_2 \\
&= \mathrm{d}_I(B, A)\,;
\end{aligned}$$

– *(positivité)* $\mathrm{d}_I(A, B) \geq 0$, *et*

$$\begin{aligned}
\mathrm{d}_I(A, B) = 0 &\Leftrightarrow \exists \tau \in \mathcal{S}_n, \left\| a_i - b_{\tau(i)} \right\|_2 = 0 \quad \forall i = 1, \ldots, n \\
&\Leftrightarrow \exists \tau \in \mathcal{S}_n, a_i = b_{\tau(i)} \quad \forall i = 1, \ldots, n \\
&\Leftrightarrow A = B\,;
\end{aligned}$$

– *(inégalité triangulaire) Soit* $C = \{c_1, \ldots, c_n\} \in (\mathbb{R}^d)^n$*, alors* $\forall \tau, \tau' \in \mathcal{S}_n$*,*

$$\begin{aligned}
\mathrm{d}_I(A,C) &\leq \frac{1}{n}\sum_{i=1}^{n} \left\|a_i - c_{\tau(i)}\right\|_2 \\
&= \frac{1}{n}\sum_{i=1}^{n} \left\|a_i - b_{\tau'(i)} + b_{\tau'(i)} - c_{\tau(i)}\right\|_2 \\
&\leq \frac{1}{n}\sum_{i=1}^{n} \left(\left\|a_i - b_{\tau'(i)}\right\|_2 + \left\|b_{\tau'(i)} - c_{\tau(i)}\right\|_2\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} \left\|a_i - b_{\tau'(i)}\right\|_2 + \frac{1}{n}\sum_{i=1}^{n} \left\|b_{\tau'(i)} - c_{\tau(i)}\right\|_2 \\
&= \frac{1}{n}\sum_{i=1}^{n} \left\|a_i - b_{\tau'(i)}\right\|_2 + \frac{1}{n}\sum_{i=1}^{n} \left\|b_i - c_{\tau\tau'^{-1}(i)}\right\|_2 \\
&= \frac{1}{n}\sum_{i=1}^{n} \left\|a_i - b_{\tau'(i)}\right\|_2 + \frac{1}{n}\sum_{i=1}^{n} \left\|b_i - c_{\tau''(i)}\right\|_2,
\end{aligned}$$

avec $\tau'' = \tau\tau'^{-1}$*. L'inégalité étant vraie pour tous* $\tau', \tau'' \in \mathcal{S}_n$*, on en déduit*

$$\begin{aligned}
\mathrm{d}_I(A,C) &\leq \frac{1}{n}\min_{\tau\in\mathcal{S}_n}\sum_{i=1}^{n} \left\|a_i - b_{\tau(i)}\right\|_2 + \frac{1}{n}\min_{\tau\in\mathcal{S}_n}\sum_{i=1}^{n} \left\|b_i - c_{\tau(i)}\right\|_2 \\
&= \mathrm{d}_I(A,B) + \mathrm{d}_I(B,C).
\end{aligned}$$

□

Annexe J

Triangles de Delaunay et cellules de Voronoi

Soit S un ensemble de points de $\mathbb{R}^d$, appelés *sites*. Les *polyèdres de Delaunay* et les *diagrammes de Voronoi* [14, 16, 52] sont des outils duaux permettant de partager l'espace en régions à partir de l'ensemble S. Nous illustrons ici le principe des deux méthodes en dimension $d = 2$ (on appellera alors les constructions *triangles de Delaunay* et *cellules de Voronoi*), et donnons des applications possibles de ces constructions. Dans la suite, on supposera donné un ensemble de sites $x_1, \ldots, x_n \in \mathbb{R}^2$.

J.1 Triangles de Delaunay

Les triangles de Delaunay sont construits en prenant pour sommets les éléments de S.

Définition J.1.1 *[52] La* triangulation de Delaunay *de S est l'unique famille de triangles ayant pour sommets l'ensemble des éléments de S, telle qu'aucun élément de S ne soit à l'intérieur du cercle circonscrit d'aucun des triangles de la famille.*

Remarque J.1.2
- *Il est sous-entendu que cette décomposition existe toujours et est unique ;*
- *il peut arriver que deux cercles circonscrits soient confondus. Dans ce cas, le « triangle » de Delaunay correspondant n'est plus un triangle, mais un quadrilatère (en toute rigueur, les « triangles » de Delaunay sont des polygones).*

On a tracé sur la figure J.1 la triangulation de Delaunay de 20 points dans $[0, 1]^2$, représentés par des cercles, obtenue avec la commande Matlab delaunay (notons que cette commande retourne des triangles même si des cercles circonscrits sont confondus).

La triangulation de Delaunay peut être utilisée quand on fait de l'optimisation globale d'une fonction (voir [18]). Supposons que l'on cherche le maximum global d'une fonction f sur un domaine $\mathcal{X}$, et que $S \subset \mathcal{X}$. Si l'on sait que f est unimodale à l'intérieur de chaque triangle de Delaunay, un moyen assez efficace de trouver l'optimum de f sur le domaine tout entier est le suivant :
- construire l'ensemble des triangles de Delaunay $T_1, \ldots, T_N$ correspondant à l'ensemble des sites S ;
- pour chaque triangle de Delaunay T_i, calculer les coordonnées de son centre de gravité c_i, pour $i = 1, \ldots, N$;

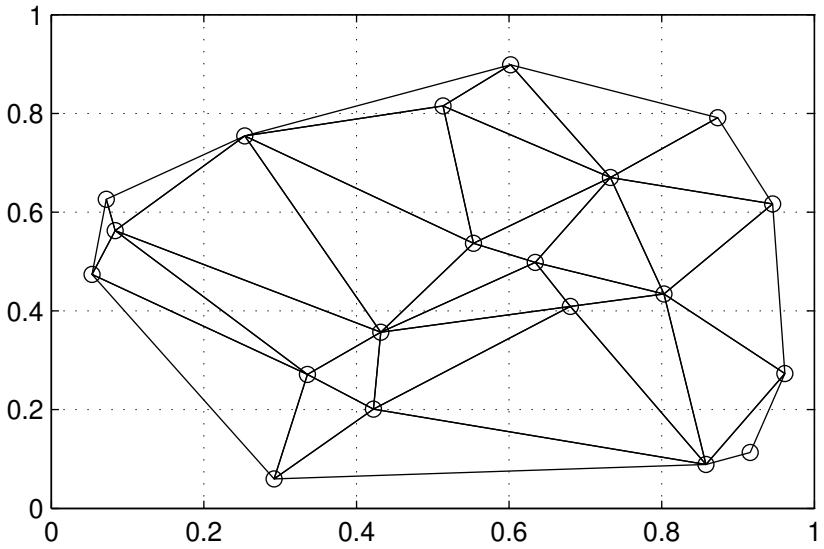

FIGURE J.1 – Triangulation de Delaunay à partir de 20 points dans $[0,1]^2$.

- lancer N fois un algorithme d'optimisation locale en prenant les c_i comme points initiaux. L'algorithme fournit alors N positions de maxima locaux $o_1, \ldots, o_N$;
- choisir comme position du maximum global

$$x^* = \underset{i=1,\ldots,N}{\operatorname{argmax}} f(o_i).$$

Remarque J.1.3 *Cette méthode suppose que les optima locaux o_i trouvés par l'algorithme appartiennent au triangle T_i. Certaines configurations de la triangulation, avec des triangles « plats », peuvent donner lieu à un point o_i situé dans un triangle adjacent au triangle T_i. On risque alors de rater l'optimum global.*

Cette technique peut être utilisée par exemple pour trouver le maximum de la variance de krigeage, qui vaut 0 aux observations $x_1, \ldots, x_n$ et grandit quand on s'éloigne des sites (en gardant à l'esprit le fait que la méthode ne garantit pas la découverte du maximum global). La construction des polyèdres de Delaunay devient cependant rapidement coûteuse en temps de calcul quand la dimension du domaine grandit (voir la figure J.2, où est tracée l'évolution du nombre de polyèdres de Delaunay en fonction du nombre de sites quand le domaine d'étude est de dimension 5).

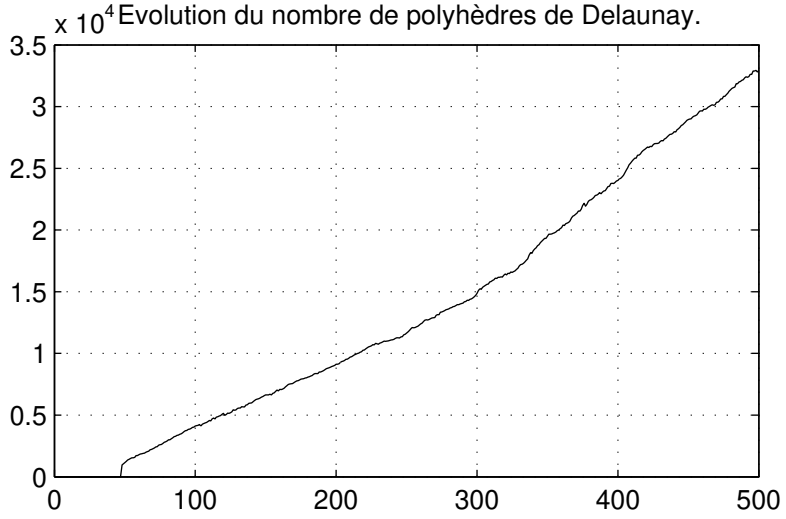

FIGURE J.2 – Nombre de polyhèdres de Delaunay en fonction de la taille du plan d'expériences à 5 facteurs (à partir de 48 points).

J.2 Cellules de Voronoi

Le diagramme de Voronoi d'un ensemble de sites de $\mathbb{R}^2$ est une partition du plan en zones déterminées par le site le plus proche.

Définition J.2.1 *Le* diagramme de Voronoi *de S est une partition de $\mathbb{R}^2$ en régions $C_1, \ldots, C_n$, appelées* cellules*, chaque cellule C_i correspondant au site x_i, de telle façon que les points situés dans la cellule C_i soient plus proches du site x_i que de tous les autres sites de S.*

Les cellules de Voronoi s'utilisent dans la recherche du plus proche voisin (§ 4.1.4.2). Si l'on doit, de façon répétée, trouver le site le plus proche d'un point du domaine, il suffit de calculer une fois pour toute les cellules de Voronoi correspondantes.

Sur la figure J.3 sont tracées les cellules de Voronoi associées aux mêmes 20 sites que ceux de la figure J.1, représentés par des cercles. Les cellules sont limitées par le fait que le domaine est borné, mais se prolongent à l'infini.

Il existe de nombreuses propriétés ainsi que d'extensions possibles des ensembles de Delaunay et de Voronoi présentés ici, pour un aperçu desquels on pourra consulter [16].

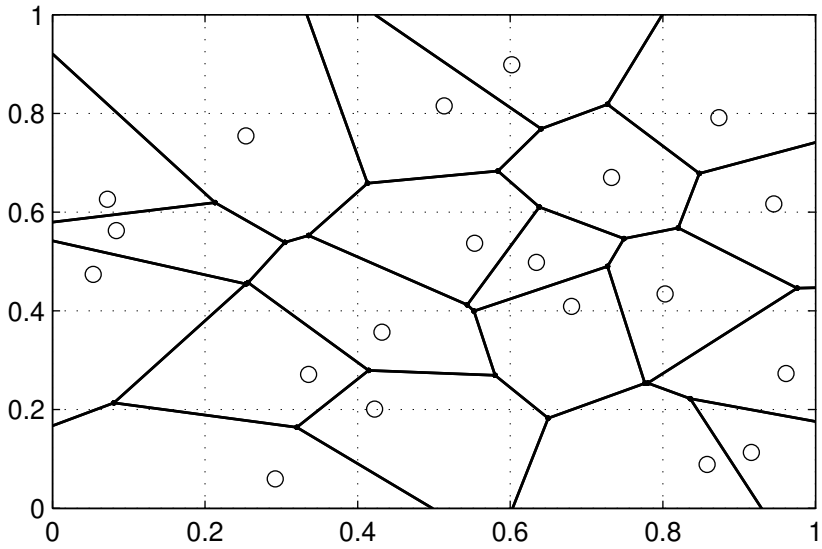

FIGURE J.3 – Cellules de Voronoi construites à partir de 20 points dans $[0,1]^2$.

Bibliographie

[1] R. Ababou, A.C. Bagtzoglou, and E.F. Wood. On the condition number of covariance matrices in kriging, estimation, and simulation of random fields. *Mathematical Geology*, 26(1) :99–133, 1994.

[2] P. Abrahamsen. A review of gaussian random fields and correlation functions. Technical Report 917, Norwegian Computing Center, Box 114, Blindern, N-0314, Oslo, Norway, 1997.

[3] M. Abramowitz and I.A. Stegun, editors. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover, New York, ninth edition, 1972.

[4] M. Abt. Approximating the mean squared prediction error in linear models under the family of exponential correlations. *Statistica Sinica*, 8 :511–526, 1998.

[5] M. Abt. Estimating the prediction mean squared error in Gaussian stochastic processes with exponential correlation structure. *The Scandinavian Journal of Statistics*, 26 :563–578, 1999.

[6] M. Abt and W.J. Welch. Fisher information and maximum-likelihood estimation of covariance parameters in Gaussian stochastic processes. *The Canadian Journal of Statistics*, 26(1) :127–137, 1998.

[7] R.A. Adams. *Sobolev Spaces*. Academic Press, 1975.

[8] J. Aldrich. R.A. Fisher and the making of maximum likelihood 1912 – 1922. *Statistical Science*, 12(3) :162–176, 1997.

[9] U. Amato, A. Antoniadis, and M. Pensky. Wavelet kernel penalized estimation for non-equispaced design regression. *Statistics and Computing*, 16(1) :37–55, 2006.

[10] W. An and Y. Sun. An equivalence between SILF–SVR and ordinary kriging. *Neural Processing Letters*, 23 :133–141, 2006.

[11] A. Antoniadis. Wavelets in statistics : a review (with discussion). *Journal of the Italian Statistical Society*, 6, 1997.

[12] A. Antoniadis. Wavelet methods in statistics : Some recent developments and their applications. *Statistics Surveys*, 1 :16–55, 2007.

[13] G. Arfken. *Mathematical Methods for Physicists*, chapter Lagrange Multipliers, pages 945–950. Orlando, FL : Academic Press, 1985.

[14] J.-M. Arnaudies and J. Bertin. *Groupes, Algèbre et Géométrie*, volume II. Ellipses, 1995.

[15] A.C. Atkinson and A.N. Donev. *Optimum Experimental Designs*, volume 8. Oxford : Clarendon Press, 1992.

[16] F. Aurenhammer and R. Klein. Voronoi diagrams. In J.-R. Sack and G. Urrutia, editors, *Handbook of Computational Geometry*, pages 201–290. Elsevier Science Publishing, 2000.

[17] S. Banerjee and A.E. Gelfand. On smoothness properties of spatial processes. *Journal of Multivariate Analysis*, 84 :85–100, 2003.

[18] R.A. Bates and L. Pronzato. Emulator-based global optimisation using lattices and Delaunay tesselation. In P. Prado and R. Bolado, editors, *Proceedings of the Third International Symposium on Sensitivity Analysis of Model Output*, pages 189–192, Madrid, June 18-20, 2001.

[19] J. Beirlant, E.J. Dudewicz, L. Györfi, and E.C. van der Meulen. Nonparametric entropy estimation : an overview. *International Journal of Mathematical and Statistical Sciences*, 6(1) :17–39, 1997.

[20] K. Benhenni and S. Cambanis. Sampling designs for estimating integrals of stochastic processes. *The Annals of Statistics*, 20(1) :161–194, 1992.

[21] A.J. Booker, J.E. Dennis Jr, P.D. Frank, D.B. Serafini, V. Torczon, and M.W. Trosset. A rigorous framework for optimization of expensive functions by surrogates. *Structural optimization*, 17 :1–13, 1999.

[22] J. Bretagnolle. Formule de Chernoff pour les lois empiriques de variables à valeurs dans des espaces généraux. *Astérisque*, 68 :33–52, 1979.

[23] M. Broniatowski. Estimation of the Kullback-Leibler divergence. *Mathematical Methods of Statistics*, 12(4) :391–409, 2003.

[24] J.N. Cawse. *Experimental Design for Combinatorial and High Throughput Materials Development*. Wiley Interscience, 3rd edition, 2003.

[25] H.-S. Chen, D.G. Simpson, and Z. Ying. Infill asymptotics for a stochastic process model with measurement error. *Statistica Sinica*, 10 :141–156, 2007.

[26] J.-P. Chilès and P. Delfiner. *Geostatistics : Modeling Spatial Uncertainty*. Wiley Series in Probability and Statistics. Wiley, 1999.

[27] J.H. Conway and N.J.A. Sloane. *Sphere Packings, Lattices and Groups*. Springer, 3rd edition, 1999.

[28] D.D. Cox and S. John. A statistical method for global optimization. In *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics*, volume 2, pages 1241–1246, Chicago, IL, Institute for Electrical and Electronic Engineers, Inc., 1992.

[29] D.D. Cox and S. John. SDO : a statistical method for global optimization. In N. Alexandrov and M.Y. Hussaini, editors, *Multidisciplinary Design Optimization : State of the Art*, pages 315–329, 1995.

[30] H. Cramer and M.R. Leadbetter. *Stationary and Related Stochastic Processes*. Wiley, New York, 1967.

[31] N.A.C. Cressie. *Statistics for Spatial Data, Revised Edition*. Wiley, New York, 1993.

[32] N.A.C. Cressie and J. Kornak. Spatial statistics in the presence of location error with an application to remote sensing of the environment. *Statistical Science*, 18(4) :436–456, 2003.

[33] N.A.C. Cressie and S.N. Lahiri. The asymptotic distribution of REML estimators. *Journal of Multivariate Analysis*, 45 :217–233, 1993.

[34] N.A.C. Cressie and D.L. Zimmerman. On the stability of the geostatistical method. *Mathematical Geology*, 24(1) :45–59, 1992.

[35] F. Cucker and S. Smale. Best choices for regularization parameters in learning theory : on the bias-variance problem. *Foundations of Computational Mathematics*, 2(4) :413–428, 2002.

[36] P. Cucker and S. Smale. On the mathematical foundations of learning. *Bulletin (New Series) of the American Mathematical Society*, 39(1) :1–49, 2001.

[37] J.A. Cuesta and C. Matran. Notes on the Wasserstein metric in Hilbert spaces. *The Annals of Probability*, 17(3) :1264–1276, 1989.

[38] I. Daubechies and B. Han. The canonical dual frame of a wavelet frame. *Applied and Computational Harmonic Analysis*, 12(3) :269–285, 2002.

[39] P. Deheuvels. Strong limiting bounds for maximal uniform spacings. *Annals of Probability*, 10 :1058–1065, 1982.

[40] D. den Hertog, J.P.C. Kleijnen, and A.Y.D. Siem. The correct kriging variance estimated by bootstrapping. Technical Report 2004-46, Tilburg University, Center for Economic Research, 2004.

[41] L. Devroye and G. Lugosi. *Combinatorial Methods in Density Estimation*. Springer–Verlag, 2001.

[42] A. Dimnaku, R. Kincaid, and M.W. Trosset. Approximate solutions of continuous dispersion problems. *Annals of Operations Research*, 136(1) :65–80, 2005.

[43] J.L. Doob. *Stochastic Processes*. Wiley, 1953.

[44] N.R. Draper and D.K.J. Lin. Response surface designs. In S. Ghosh and C.R. Rao, editors, *Handbook of Statistics*, volume 13, pages 343–375. Elsevier Science B.V., 1996.

[45] J.-J. Droesbeke, J. Fine, and G. Saporta, editors. *Plans d'expériences. Applications à l'entreprise.* Technip, 1997.

[46] W.M. Duckworth. Some binary maximin distance designs. *Journal of Statistical Planning and Inference*, 88 :149–170, 2000.

[47] B. Efron and R.J. Tibshirani. *An Introduction to the Bootstrap*. Collection Monographs on Statistics and Applied Probability, Boca Raton (Calif.) : Chapman & Hall, New York, 1998.

[48] J. Fan and I. Gijbels. *Local Polynomial Modelling and its Applications*. Chapman and Hall, London, 1996.

[49] K.-T. Fang, R. Li, and A. Sudjianto. *Design and Modeling for Computer Experiments*. Chapman & Hall, 2005.

[50] I. Fazekas and A.G. Kukush. Kriging and measurement error. *Probability and Statistics*, 25 :139–159, 2005.

[51] A.I.J. Forrester, A.J. Keane, and N.W. Bressloff. Design and analysis of 'noisy' computer experiments. *AIAA Journal*, 44 :2331–2339, 2006.

[52] S. Fortune. Voronoi diagrams and Delaunay triangulations. In D.Z. Du and F. Hwang, editors, *Computing in Euclidean Geometry*, volume 1, 1992.

[53] S. Galanti and A. Jung. Low-discrepancy sequences : Monte carlo simulation of option prices. *Journal of Derivatives*, pages 63–83, 1997.

[54] J.B. Gao, S.R. Gunn, C.J. Harris, and M. Brown. A probabilistic framework for SVM regression and error bar estimation. *Machine Learning*, 46(1-3) :71–89, 2002.

[55] S. Geman, E. Bienenstock, and R. Doursat. Neural networks and the bias/variance dilemma. *Neural Computation*, 4 :1–58, 1992.

[56] M.N. Gibbs. *Bayesian Gaussian Processes for Regression and Classification*. PhD thesis, University of Cambridge, 1997.

[57] I.I. Gikhman and A.V. Skorokhod. *Introduction to the theory of random processes*. Dover, 1996.

[58] G.H. Golub and C.F. Van Loan. *Matrix computations*. Johns Hopkins Studies in the Mathematical Sciences. The Johns University Press, 3rd edition, 1996.

[59] Y. Gratton. Le krigeage : la méthode optimale d'interpolation spatiale. *Les articles de l'Institut d'Analyse Géographique*, 2002.

[60] G.R. Grimmett and D.R. Stirzaker. *Probability and Random Processes*. Clarendon Press, Oxford, 2nd edition, 1992.

[61] V. Guigue, A. Rakotomamonjy, and S. Canu. Estimation de signaux par noyau d'ondelettes. In *20^e colloque sur le traitement du signal et des images*, volume 23(5-6), pages 449–460, Louvain-la-neuve (Belgique), 2006.

[62] X. Guyon. *Statistique et économétrie. Du modèle linéaire... aux modèles non-linéaires.* Ellipses, 2001.

[63] J. Hadamard. Sur les problèmes aux derivées partielles et leur signification physique. *Princeton University Bulletin*, pages 49–52, 1902.

[64] P. Hall and S. Morton. On the estimation of entropy. *Annals of the Institute of Statistical Mathematics*, 45(1) :69–88, 1993.

[65] M.S. Handcock and M.L. Stein. A bayesian analysis of kriging. *Technometrics*, 35(4) :403–410, 1993.

[66] R.H. Hardin and N.J.A. Sloane. A new approach to the construction of optimal designs. *Journal of Statistical Planning and Inference*, 37 :339–369, 1993.

[67] G.H. Hardy, J.E. Littlewood, and G. Polya. *Inequalities*. Cambridge University Press, second edition, 1988.

[68] D.A. Harville. Bayesian inference for variance components using only error contrasts. *Biometrika*, 61(2) :383–385, 1974.

[69] D.A. Harville and D.R. Jeske. Mean squared error of estimation of prediction under a general linear model. *Journal of the American Statistical Association*, 87(419) :724–731, 1992.

[70] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning : Data Mining, Inference, and Prediction*. Springer–Verlag, 2001.

[71] M.E. Havrda and F. Charvát. Quantification method of classification processes : concept of structural α-entropy. *Kybernetika*, 3 :30–35, 1967.

[72] A.S. Hedayat, N.J.A. Sloane, and J. Stufken. *Orthogonal Arrays : Theory and Applications*. Springer, New York, 1999.

[73] J.A. Hoeting, R.A. Davis, A.A. Merton, and S.E. Thompson. Model selection for geostatistical models. *Ecological Applications*, 16(1) :87–98, 2006.

[74] B. Hofmann. *Regularization for Applied Inverse and Ill-Posed Problems*. J. Teubner, Leipzig, 1986.

[75] K.M. Irvine, A.I. Gitelman, and J.A. Hoeting. Spatial designs and properties of spatial correlation : effects on covariance estimation. *Journal of Agricultural, Biological, and Environmental Statistics*, 12(4) :450–469, 2007.

[76] S. Jaffard, Y. Meyer, and R.D. Ryan. *Wavelets : Tools for Science & Technology*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2001.

[77] R. Jin, W. Chen, and A. Sudjianto. An efficient algorithm for constructing optimal design of computer experiments. *Journal of Statistical Planning and Inference*, 134(1) :268–287, 2005.

[78] P.W.M. John, M.E. Johnson, L.M. Moore, and D. Ylvisaker. Minimax distance designs in two-level factorial experiments. *Journal of Statistical Planning and Inference*, 44 :249–263, 1995.

[79] M.E. Johnson, L.M. Moore, and D. Ylvisaker. Minimax and maximin distance designs. *Journal of Statistical Planning and Inference*, 26 :131–148, 1990.

[80] D.R. Jones. A taxonomy of global optimization methods based on response surfaces. *Journal of Global Optimization*, 21 :345–383, 2001.

[81] D.R. Jones, M. Schonlau, and W.J. Welch. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13 :455–492, 1998.

[82] R.N. Kackar and D.A. Harville. Approximations for standard errors of estimators of fixed and random effects in mixed linear models. *Journal of the American Statistical Association*, 79(388) :853–862, 1984.

[83] M.G. Kendall and A. Stuart. *The advanced theory of statistics*, volume 2. Griffin, London, 1973.

[84] J.T. Kent and K.V. Mardia. The link between kriging and thin-plate splines. *Probability, Statistics and Optimisation : a Tribute to Peter Whittle*, pages 325–339, 1994.

[85] A. Keziou. *Utilisation des divergences entre mesures en statistique inférentielle*. PhD thesis, Université Paris VI, novembre 2003.

[86] A.I. Khuri and J.A. Cornell. *Response Surfaces : Designs and Analyses*. Marcel Dekker, second edition, 1996.

[87] G. Kimeldorf and G. Wahba. Some results on tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*, 33 :82–95, 1971.

[88] S. Kirkpatrick, C.D. Gelatt, and M.P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598) :671–680, 1983.

[89] T.G. Kolda, R.M. Lewis, and V.J. Torczon. Optimization by direct search : New perspectives on some classical and modern methods. *SIAM Review*, 45(3) :385–482, 2003.

[90] L. Kozachenko and N. Leonenko. On statistical estimation of entropy of a random vector. *Problems of Information Transmission*, 23(2) :95–101, 1987.

[91] M.F. Kratz. Level crossings and other level functionals of stationary Gaussian processes. *Probability Surveys*, 3 :230–288, 2006.

[92] F.Y. Kuo and I.H. Sloan. Lifting the curse of dimensionality. *Notices of the AMS*, 52(11) :1320–1329, 2005.

[93] S.N. Lahiri. On inconsistency of estimators based on spatial data under infill asymptotics. *Sankhya : The Indian Journal of Statistics*, 58 :403–417, 1996.

[94] R.L.H. Lam, W.J. Welch, and S.S. Young. Uniform coverage designs for molecule selection. *Technometrics*, 44(2) :99–109(11), 2002.

[95] B. Laurent. Adaptive estimation of a quadratic functional of a density by model selection. *ESAIM : Probability and Statistics*, 9 :1–18, 2005.

[96] E.L. Lehmann and G. Casella, editors. *Theory of Point Estimation*. Springer-Verlag, New York, second edition, 1998.

[97] N. Leonenko, L. Pronzato, and V. Savani. A class of Rényi information estimators for multidimensional densities. *The Annals of Statistics*, 36(5) :2153–2182, 2008.

[98] R. Li and A. Sudjianto. Analysis of computer experiments using penalized likelihood in gaussian kriging models. *Technometrics*, 47(2) :111–120, 2005.

[99] W. Light and H. Wayne. On power functions and error estimates for radial basis function interpolation. *Journal of Approximation Theory*, 92 :245–266, 1998.

[100] R. Linder. *Les plans d'expériences. Un outil indispensable à l'expérimentateur.* Presses des Ponts, 2005.

[101] G. Lindgren. *Lectures on Stationary Stochastic Processes ; a Course for PhD students in Mathematical Statistics and other fields*, May 1999.

[102] W.L. Loh. Fixed-domain asymptotics for a subclass of Matérn-type gaussian random fields. *The Annals of Statistics*, 33(5) :2344–2394, 2005.

[103] W.L. Loh and T.K. Lam. Estimating structured correlation matrices in smooth gaussian random field models. *The Annals of Statistics*, 28(3) :880–904, 2000.

[104] S.N. Lophaven, H.B. Nielsen, and J. Sondergaard. Aspects of the matlab toolbox DACE. Technical Report IMM-REP-2002-13, Informatics and Mathematical Modelling, Technical University of Denmark, 2002.

[105] S.N. Lophaven, H.B. Nielsen, and J. Sondergaard. DACE, a matlab kriging toolbox. Technical Report IMM-REP-2002-12, Informatics and Mathematical Modelling, Technical University of Denmark, 2002.

[106] L.-T. Luh. The embedding theory of native spaces. *Approximation Theory & Its Applications*, 17(4) :90–104, 2001.

[107] L.-T. Luh. The equivalence theory of native spaces. *Approximation Theory & Its Applications*, 17(1) :76–96, 2001.

[108] L.-T. Luh. On Wu and Schaback's error bound. arXiv :math/0602087v1, 2006.

[109] M.N. Lukić and J.H. Beder. Stochastic processes with sample paths in reproducing kernel Hilbert spaces. *Transactions of the American Mathematical Society*, 353(10) :3945–3969, 2001.

[110] D.J.C. MacKay. Introduction to Gaussian processes. In C.M. Bishop, editor, *Neural Networks and Machine Learning*, NATO ASI, pages 133–166. Kluwer Academic Press, 1998.

[111] W.R. Madych and S.A. Nelson. Multivariate interpolation and conditionally positive definite functions. *Approximation Theory and its Applications*, 4(4) :77–89, 1988.

[112] W.R. Madych and S.A. Nelson. Multivariate interpolation and conditionally positive definite functions II. *Mathematics of Computation*, 54(189) :211–230, 1990.

[113] K.V. Mardia. Maximum likelihood estimation for spatial models. In D.A. Griffith, editor, *Spatial Statistics : Past, Present and Future*, pages 203–253. Michigan Document Services, 1990.

[114] K.V. Mardia and R.J. Marshall. Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*, 71(1) :135–146, 1984.

[115] K.V. Mardia and A.J. Watkins. On multimodality of the likelihood in the spatial linear model. *Biometrika*, 76(2) :289–295, 1989.

[116] A. Marrel. *Mise en œuvre et utilisation du métamodèle processus gaussien pour l'analyse de sensibilité de modèles numériques : application à un code de transport hydrogéologique.* PhD thesis, INSA de Toulouse, 2008.

[117] J. Martin and T. Simpson. A Monte Carlo simulation of the kriging model. In *10th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference*, Albany, New York, August 30-1, 2004.

[118] T.J. Mitchell and M.D. Morris. Bayesian design and analysis of computer experiments : Two examples. *Statistica Sinica*, 2 :359–379, 1992.

[119] J. Mockus, V. Tiesis, and A. Zilinskas. The application of bayesian methods for seeking the extremum. *Towards Global Optimisation*, 2 :117–129, 1978.

[120] M.D. Morris and T.J. Mitchell. Exploratory designs for computational experiments. *Journal of Statistical Planning and Inference*, 43 :381–402, 1995.

[121] W.G. Müller. *Collecting Spatial Data. Optimum Design of Experiments for Random Fields.* Physica-Verlag, 2nd revised edition, 2003.

[122] I.J. Myung. Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology*, 47 :90–100, 2003.

[123] C.J. Paciorek and M. Schervish. Spatial modelling using a new class of nonstationary covariance functions. *Environmetrics*, 17 :483–506.

[124] G. Pagès and J. Printems. Optimal quadratic quantization for numerics : the Gaussian case. *Monte Carlo Methods and Applications*, 2(9) :135–165, 2003.

[125] E. Pardo-Igúzquiza. MLREML : A computer program for the inference of spatial covariance parameters by maximum likelihood and restricted maximum likelihood. *Computers & Geosciences*, 23(2) :153–162, 1997.

[126] R. Penrose. A generalized inverse for matrices. In *Proceedings of the Cambridge Philosophical Society*, volume 51, pages 406–413, 1955.

[127] W.R. Pestman. *Mathematical Statistics : an Introduction.* Walter de Gruyter, Berlin, 1998.

[128] T. Poggio and S. Smale. The mathematics of learning : Dealing with data. *Notices of the American Mathematical Society*, 50(5) :537–544, 2003.

[129] M. Pontil. Learning with reproducing kernel Hilbert spaces : a guide tour. *Bulletin of the Italian Artificial Intelligence Association, AI*IA Notizie*, 2003.

[130] M.J.D. Powell. On the use of quadratic models in unconstrained minimization without derivatives. Technical Report DAMTP 2003/NA03, Numerical Analysis Group, Cambridge university, 2003.

[131] L. Pronzato. One-step ahead adaptative D-optimal design on a finite design space is asymptotically optimal. *Metrika*. To appear.

[132] L. Pronzato and E. Thierry. Sequential experimental design and response optimisation. *Statistical Methods and Applications*, 11(3) :277–292, 2002.

[133] L. Pronzato and E. Thierry. Robust design with nonparametric models : Prediction of second-order characteristics of process variability by kriging. In *13th IFAC Symposium on System Identification*, pages 560–565, Rotterdam, 27–29 août 2005.

[134] F. Pukelsheim. *Optimal design of experiments.* SIAM, 2006.

[135] A. Rakotomamonjy and S. Canu. Frames, reproducing kernels, regularization and learning. *The Journal of Machine Learning Research*, 6 :1485–1515, 2005.

[136] C.E. Rasmussen and C.K.I. Williams. *Gaussian Processes for Machine Learning.* The MIT Press, 2006.

[137] G.K. Robinson. That BLUP is a good thing : the estimation of random effects, with discussion. *Statistical Science*, 6(1) :15–51, 1991.

[138] L. Rüschendorf. Wasserstein metric. In M. Hazewinkel, editor, *Encyclopœdia of Mathematics*. Kluwer Academic Publisher, 2001.

[139] H. Rue and L. Held. *Gaussian Markov Random Fields : Theory and Applications*. CRC Press, 2005.

[140] J. Sacks and S. Schiller. Spatial designs. In S.S. Gupta and J.O. Berger, editors, *Statistical Decision Theory and Related Topics IV*, volume 2, pages 385–395. Springer-Verlag, 1988.

[141] J. Sacks, S.B. Schiller, and W.J. Welch. Designs for computer experiments. *Technometrics*, 31(1) :41–47, 1989.

[142] J. Sacks, W.J. Welch, T.J. Mitchell, and H.P. Wynn. Design and analysis of computer experiments, with dicussion. *Statistical Science*, 4(4) :409–435, 1989.

[143] T.R. Sahama and N.T. Diamond. Sample size considerations and augmentation of computer experiments. *The Journal of Statistical Computation and Simulation*, 68 :307–319, 2001.

[144] T.J. Santner, B.I. Williams, and W.I. Notz. *The Design and Analysis of Computer Experiments*. Springer-Verlag, 2003.

[145] M.J. Sasena, P.Y. Papalambros, and P. Goovaerts. Metamodeling sampling criteria in a global optimization framework. In *Proceedings of the 8th Symposium on Multidisciplinary Analysis and Optimization*, pages 189–192, Long Beach, CA, Sept. 6-8, 2000.

[146] R. Schaback. Native Hilbert spaces for radial basis functions I. In *International Series of Numerical Mathematics*, volume 132, pages 255–282, Birkhäuser, Basel, 1999. New developments in approximation theory (Dortmund, 1998).

[147] R. Schaback. A unified theory of radial basis functions. Native Hilbert spaces for radial basis functions II. *Journal of Computational and Applied Mathematics*, 121(1–2) :165–177, 2000. Numerical analysis in the 20th century, Vol. I, Approximation theory.

[148] C. Scheidt. *Analyse statistique d'expériences simulées : modélisation adaptative de réponses non régulières par krigeage et plans d'expériences, Application à la quantification des incertitudes en ingénierie des réservoirs pétroliers*. PhD thesis, Université Louis Pasteur - Strasbourg I, 2006.

[149] B. Schölkopf, R. Herbrich, A.J. Smola, and R. Williamson. A generalized representer theorem. Technical Report NC2–TR–2000-81, ESPRIT Working Group in Neural and Computational Learning II, NeuroCOLT2 27150, 2000.

[150] M. Schonlau and W.J. Welch. Global optimization with nonparametric function fitting. In *Proceedings of the Section on Physical and Engineering Sciences, American Statistical Association*, pages 183–186, 1996.

[151] M. Schonlau, W.J. Welch, and D. Jones. A data-analytic approach to bayesian global optimization. In *American Statistical Association Proceedings, Section of Physical Engineering Sciences*, pages 186–191, 1997.

[152] M. Schonlau, W.J. Welch, and D.R. Jones. Global versus local search in constrained optimization of computer models. In *New Developments and Applications in Experimental Design*, volume 34 of *Lecture Notes - Monograph Series*, pages 11–25, IMS, Hayward, 1998.

[153] C.E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27 :379–423, 1948.

[154] M.C. Shewry and H.P. Wynn. Maximum entropy sampling. *Journal of Applied Statistics*, 14 :165–170, 1987.

[155] J.S. Simonoff. *Smoothing Methods in Statistics*. Springer Series in Statistics. Springer–Verlag, 1996.

[156] A.J. Smola, T.T. Friess, and B. Schölkopf. General cost functions for support vector regression. In M.S. Kearns, S.A. Solla, and D.A. Cohn, editors, *Advances in Neural Information Processing Systems*, volume 11, pages 585–591, Cambridge, MA, 1999. MIT Press.

[157] A.J. Smola and B. Schölkopf. A tutorial on support vector regression. *Statistics and Computing*, 14(3) :199–222, 2004.

[158] A.J. Smola, B. Schölkopf, and K.-R. Müller. The connection between regularization operators and support vector kernels. *Neural Networks*, 11(4) :637–649, 1998.

[159] A.J. Smola, B. Schölkopf, and K.-R. Müller. General cost functions for support vector regression. In T. Downs, M. Frean, and M. Gallagher, editors, *Proceedings of the Ninth Australian Conference on Neural Networks*, pages 79–83, Brisbane, Australia, 1998.

[160] M. Stehlík. Some properties of exchange design algorithms under correlation. Research Report Series / Department of Statistics and Mathematics 28, 2006.

[161] M.L. Stein. *Interpolation of Spatial Data : Some Theory for Kriging*. Springer, New York, 1999.

[162] M.L. Stein. Predicting random fields with increasing dense observations. *The Annals of Applied Probability*, 9(1) :242–273, 1999.

[163] M.L. Stein. The screening effect in kriging. *The Annals of Statistics*, 30(1) :298–323, 2002.

[164] M.L. Stein. Equivalence of Gaussian measures for some nonstationary random fields. *Journal of Statistical Planning and Inference*, 123 :1–11, 2004.

[165] M.L. Stein. Nonstationary spatial covariance functions. Technical report, University of Chicago, Chicago, Illinois USA, 2005.

[166] M.L. Stein, Z. Chi, and L.J. Welty. Approximating likelihoods for large spatial data sets. *Journal of the Royal Statistical Society*, 66(2) :275–296, 2004.

[167] A.N. Tikhonov and V.Y. Arsenin. *Solutions of Ill-Posed Problems*. Wiley, New York, 1977.

[168] E. Tillier. Développement d'un algorithme d'optimisation pour la synthèse de catalyseurs à haut débit. Technical Report 57901, IFP–Lyon, 2004.

[169] V.J. Torczon. On the convergence of pattern search algorithms. *SIAM Journal on Optimization*, 7(1) :1–25, 1997.

[170] M.W. Trosset. Approximate maximin distance designs. In *Proceedings of the Section on Physical and Engineering Sciences*, pages 223–227. American Statistical Association, 1999.

[171] C. Tsallis. Possible generalization of Boltzmann-Gibbs statistics. *Journal of Statistical Physics*, 52(1) :479–487, 1988.

[172] E.R. van Dam. Two-dimensional minimax latin hypercube designs. Technical Report 2005–105, Tilburg University, 2005.

[173] E.R. van Dam, B. Husslage, D. den Hertog, and H. Melissen. Maximin latin hypercubes in two dimensions. *Operations Research*, 55(1) :158–169, 2007.

[174] A.W. van der Vaart. Maximum likelihood estimation of parameters under a spatial sampling scheme. *The Annals of Statistics*, 24(5) :2049–2057, 1996.

[175] A.W. van der Vaart. *Asymptotic Statistics*. Cambridge series in statistical and probabilistic mathematics. Cambridge University Press, 1998.

[176] V. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.

[177] V. Vapnik. *The Nature of Statistical Learning Theory*. Statistics for Engineering and Information Science. Springer, second edition, 2000.

[178] E. Vazquez. *Modélisation comportementale de systèmes non-linéaires multivariables par méthodes à noyaux et applications*. PhD thesis, Université Paris XI Orsay, mai 2005.

[179] E. Vazquez and J. Bect. On the convergence of the expected improvement algorithm. arXiv :0712.3744v2, 2008.

[180] J. Villemonteix, E. Vasquez, and E. Walter. An informational approach to the global optimization of expensive-to-evaluate functions. *Journal of Global Optimization*. To appear.

[181] G. Wahba. *Spline Models for Observational Data*, volume 59 of *CBMS–NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, second edition, 1992.

[182] G. Wahba. Support Vector Machines, Reproducing Kernel Hilbert Spaces and the randomized GACV. Technical Report 984rr, Department of Statistics, University of Wisconsin, 1998.

[183] M. Waldman, H. Li, and M. Hassan. Novel algorithms for the optimization of molecular diversity of combinatorial libraries. *Journal of Molecular Graphics and Modelling*, 18 :412–426, 2000.

[184] M.P. Wand and M.C. Jones. *Kernel Smoothing*. Chapman and Hall, London, 1995.

[185] A.G. Watson and R.J. Barnes. Infill sampling criteria to locate extremes. *Mathematical Geology*, 27(5) :589–608, 1995.

[186] C. Wei. *Bayesian Approach to Support Vector Machines*. PhD thesis, National University of Singapore, 2003.

[187] H. Wendland. *Scattered Data Approximation*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, Cambridge, UK, 2005.

[188] C.K.I. Williams. Regression with Gaussian processes. In S.W. Ellacott, J.C. Mason, and I.J. Anderson, editors, *Proceedings of the first international conference on Mathematics of neural network : models, algorithms and applications*, pages 378–382. Kluwer Academic Publishers, 1997.

[189] C.K.I. Williams and C.E. Rasmussen. Gaussian processes for regression. In D.S. Touretzky, M.C. Mozer, and M.E. Hasselmo, editors, *Advances in Neural Information Processing Systems*, volume 8, pages 514–520. MIT Press, 1996.

[190] Z.-M. Wu and R. Schaback. Local error estimates for radial basis function interpolation of scattered data. *IIMA Journal of Numerical Analysis*, 13(1) :13–27, 1993.

[191] H.P. Wynn. The sequential generation of D-optimum experimental designs. *Annals of Mathematical Statistics*, 41 :1655–1664, 1970.

[192] H.P. Wynn. Maximum entropy sampling and general equivalence theory. In A. Di Bucchianico, H. Laüter, and H.P. Wynn, editors, *mODa'7 - advances in model-oriented design and analysis, Proceedings of the 7th international workshop*, pages 211–218, Heeze, Netherlands, June 2007. Physica-Verlag (Heidelberg).

[193] Y. Xiong, W. Chen, D. Apley, and X. Ding. A non-stationary covariance-based kriging method for metamodelling in engineering design. *International Journal for Numerical Methods in Engineering*, 71 :733–756, 2007.

[194] Z. Ying. Asymptotic properties of a maximum likelihood estimator with data from a gaussian process. *Journal of Multivariate Analysis*, 36 :280–296, 1991.

[195] Z. Ying. Maximum likelihood estimation of parameters under a spatial sampling scheme. *The Annals of Statistics*, 21(3) :1567–1590, 1993.

[196] H. Zhang. Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association*, 99(465) :250–261, 2004.

[197] H. Zhang and D.L. Zimmerman. Towards reconciling two asymptotic frameworks in spatial statistics. *Biometrika*, 92(4) :921–936, 2005.

[198] Z. Zhu and M.L. Stein. Spatial sampling design for prediction with estimated parameters. *Journal of Agricultural, Biological, and Environmental Statistics*, 11(1) :24–44, 2006.

[199] Z. Zhu and H. Zhang. Spatial sampling design under the infill asymptotic framework. *Environmetrics*, 17(4) :323–337, 2005.

[200] D.L. Zimmerman. Optimal network design for spatial prediction, covariance parameter estimation and empirical prediction. *Environmetrics*, 17 :635–652, 2006.

[201] D.L. Zimmerman and N. Cressie. Mean squared prediction error in the spatial linear model with estimated covariance parameters. *Annals of the Institute of Statistical Mathematics*, 44(1) :27–43, 1992.

[202] D.L. Zimmerman and M.B. Zimmerman. A comparison of spatial semivariograms estimators and corresponding ordinary kriging predictors. *Technometrics*, 33(1) :77–91, 1991.

[203] V.M. Zolotarev. Lévy metric. In M. Hazewinkel, editor, *Encyclopædia of Mathematics*. Kluwer Academic Publisher, 2001.

[204] A. Zygmund. *Trigonometrical series*. Dover, 1955.