Modélisation comportementale de systèmes non-linéaires multivariables par méthodes à noyaux et applications

Emmanuel Vazquez

N° D'ORDRE : 7872

UNIVERSITÉ PARIS XI
UFR SCIENTIFIQUE D'ORSAY

THÈSE

Présentée

Pour obtenir

Le GRADE de DOCTEUR EN SCIENCES

DE L'UNIVERSITÉ PARIS XI ORSAY

PAR

Emmanuel Vazquez

Sujet :

Modélisation comportementale de systèmes non-linéaires multivariables par méthodes à noyaux et applications

Soutenue le 12 mai 2005 devant la Commission d'examen

M. Bastin	Georges	Rapporteur
M. Schaback	Robert	Rapporteur
M. Bondon	Pascal	
M. Pronzato	Luc	Président
M. Wackernagel	Hans	
M. Walter	Éric	Directeur de thèse

Remerciements

Je souhaiterais en premier lieu exprimer toute ma gratitude à mon directeur de thèse M. Éric Walter *pour ses nombreux conseils, son soutien et ses qualités humaines. Je remercie les membres de mon jury pour leur lecture attentive de ce manuscrit et leurs appréciations très favorables. Je remercie également le Laboratoire des Signaux et Systèmes dans son ensemble, chercheurs permanents, thésards, informaticiens et personnes de l'administration qui m'ont offert un cadre de travail agréable et stimulant. Je remercie mes collègues de Supélec pour la confiance qu'ils m'ont accordée en me proposant de travailler avec eux. Enfin, je dédie ce travail à mes amis, à ma famille, et tout particulièrement à Virginie.*

Modélisation comportementale de systèmes non-linéaires multivariables par méthodes à noyaux et applications

Emmanuel Vazquez

Manuscrit compilé le 12 septembre 2005

Table des matières

Table des figures

Introduction

Régression et modélisation comportementale

La prédiction linéaire de processus aléatoires est une méthode de régression utilisable avec des données échantillonnées irrégulièrement. Elle est couramment utilisée en géostatistique sous l'appellation *krigeage* pour modéliser des phénomènes physiques définis spatialement (Matheron, 1963, 1973 ; Cressie, 1993 ; Stein, 1999). En dehors de la géostatistique, l'utilisation de processus aléatoires comme outil de régression non-linéaire a été abordée dans la théorie des splines (Kimeldorf et Wahba, 1970a ; Wahba, 1990), en automatique (Bastin et Gevers, 1985) et dans le domaine de la modélisation de simulations informatiques (Sacks et al., 1989), mais est toutefois restée confinée. L'une des raisons possibles de cette situation est que sa mise en œuvre nécessite des opérations matricielles dont le coût algorithmique croît proportionnellement au cube du nombre des données manipulées. Si cela a pu constituer un frein à son utilisation dans le passé, la puissance de calcul disponible depuis quelques années permet de considérer des problèmes avec plusieurs milliers de données. Si on les compare à d'autres méthodes comme les réseaux de neurones artificiels, les modèles par processus gaussiens offrent des avantages de simplicité conceptuelle et de flexibilité, et sont susceptibles de produire des approximations de fonction d'une qualité supérieure pour un coût algorithmique plus bas. Dans la seconde moitié des années 1990, on constate deux évolutions. La première est l'apparition des méthodes de régression à vecteurs de support (Vapnik et al., 1997 ; Smola, 1998) dont le principe repose sur la théorie des espaces hilbertiens à noyau reproduisant. Les méthodes de régression à noyaux reproduisants, dont les splines mentionnées ci-dessus, possèdent des liens étroits avec la théorie des processus aléatoires gaussiens, dont certains sont établis depuis les années 1960 (Parzen, 1962, 1963 ; Kimeldorf et Wahba, 1970a,b). La seconde évolution est la réapparition du krigeage sous une formulation bayésienne (Williams, 1997 ; Neal, 1997), sans référence d'ailleurs à la géostatistique. Ces méthodes commencent à être diffusées et appliquées dans des domaines tels que l'identification de systèmes dynamiques à temps discret (Girard et al., 2003).

La thèse présentée dans ce mémoire porte sur l'utilisation du krigeage et des méthodes à noyaux pour la modélisation de processus ou de systèmes artificiels, par exemple pour des tâches de conception, de commande ou de diagnostic[1]. L'étape de modélisation d'un processus ou d'un système est fondamentale mais souvent délicate. Les connaissances a priori ne permettent pas toujours d'établir un modèle physique. Même quand ces connaissances sont disponibles, on renonce parfois à établir des liens directs avec la physique car les problèmes posés sont de trop grande complexité.

[1] (Sacks et al., 1989) a largement contribué à notre réflexion initiale.

Cette complexité peut tenir à la dimension de l'espace d'état, à la nature du couplage entre les variables, à la nature des non-linéarités, et au coût algorithmique qui peut s'avérer incompatible avec l'usage prévu. On peut alors essayer des techniques analysant la dépendance entre les entrées (que nous appellerons *facteurs*) et des variables de sortie pertinentes, sans s'appuyer sur les lois de la physique gouvernant le système à modéliser. On parle alors de modèles *comportementaux* ou *boîte noire*. Ces modèles permettent d'effectuer des *prédictions* sur des caractéristiques non mesurées du système, comme la valeur de la sortie ou de ses dérivées en un point de l'espace des facteurs où aucune expérience n'a été conduite.

La modélisation comportementale des systèmes statiques est un problème d'interpolation ou d'approximation de fonctions à partir de données non nécessairement uniformément échantillonnées. Les systèmes dynamiques, qui ne sont pas abordés dans ce mémoire, peuvent être traités par des modèles comportementaux à temps discret autorégressifs dans lesquels intervient le même problème d'interpolation ou d'approximation de fonction. Dans ce travail, nous nous intéressons aux méthodes d'interpolation et d'approximation par prédiction linéaire de processus aléatoires et par régression régularisée dans des espaces hilbertiens à noyau reproduisant. Ces deux méthodes, dont nous verrons qu'elles sont pour l'essentiel identiques, permettent de construire des modèles comportementaux possédant plusieurs entrées et plusieurs sorties, éventuellement corrélées.

Objectifs et organisation du mémoire

Nous proposons ici une synthèse sur la modélisation boîte noire par processus gaussiens et par régression régularisée par des normes d'espaces à noyau reproduisant. Cette synthèse fait appel à des résultats obtenus dans des domaines très variés. Notre présentation s'inspire de la géostatistique (Matheron, 1971b, 1973 ; Chilès et Delfiner, 1999), de la théorie des processus aléatoires (Doob, 1953 ; Rozanov, 1967 ; Ibragimov et Rozanov, 1978 ; Brockwell et Davis, 1987), des statistiques (Huber, 1981 ; Stein, 1999), de la théorie des splines (Wahba, 1990), de l'analyse fonctionnelle (Mallat, 1999 ; Aubin, 2000), de la théorie de l'apprentissage (Vapnik, 1995 ; Schölkopf et Smola, 2002), etc. Un de nos objectifs est de faciliter l'accès à ces connaissances en les présentant dans un cadre unifié. Le sujet étant bien sûr très vaste, de nombreux points intéressants ne pourront pas être mentionnés. Par exemple, les développements récents dans la théorie statistique de l'apprentissage sur les bornes d'erreur de généralisation ne seront pas du tout mentionnés (Vapnik, 1995 ; Smola, 1998).

Nous plaçons le krigeage au centre de notre présentation pour deux raisons. D'une part parce que c'est l'une des premières utilisations de la théorie de la prédiction linéaire en réponse à des problèmes réels et en cela, l'apport de la géostatistique au problème de modélisation comportementale est essentiel. D'autre part, le point de vue probabiliste ainsi adopté est très utile pour comprendre en quoi le choix du noyau est important. Trop souvent, en effet, les noyaux sont choisis a priori, de type radial gaussien par exemple, indépendamment de la nature des données observées. En géostatistique au contraire, la phase de prédiction est toujours précédée d'une phase d'analyse des données observées dont l'objectif est de choisir une covariance aussi adaptée que possible à la physique du phénomène étudié. Cependant, nous nous éloignons de la présentation traditionnelle de la géostatistique afin de présenter les méthodes à noyaux dans un cadre unifié. Notre étude s'attache aux principaux aspects mathématiques de ces méthodes et aux questions de

mise en œuvre, avec notamment les problèmes liés aux choix du noyau reproduisant. Une place non négligeable est également consacrée à l'illustration de ces méthodes par des applications à des problèmes réels.

Le **chapitre 1** constitue une introduction à la modélisation de type boîte noire, destinée à faciliter la lecture du mémoire. Le principe de régression régularisée dans les espaces à noyau reproduisant y est rappelé, ainsi que des notions classiques concernant la théorie du krigeage.

La littérature consacrée aux processus aléatoires et à leur prédiction est très riche. Au cours du **chapitre 2**, nous rappelons quelques propriétés mathématiques de ces processus, puis le principe de prédiction linéaire que nous présentons d'abord sous sa forme la plus élémentaire, c'est-à-dire lorsque la structure du second ordre des processus aléatoires est supposée connue. L'objectif de ce chapitre est de donner les éléments essentiels à une bonne compréhension du krigeage, ce qui demande parfois de rentrer dans des aspects mathématiques relativement avancés. Nous rappelons également la notion de variables aléatoires à valeurs dans un espace hilbertien. Ceci permettra de définir au chapitre 3 des processus aléatoires à trajectoires dans des espaces hilbertiens à noyau reproduisant.

Le **chapitre 3** couvre les notions importantes sur les espaces hilbertiens à noyau reproduisant, dont le rôle est central dans la théorie de l'approximation de fonctions. Nous abordons ensuite la régression régularisée dans ces espaces et rappelons le théorème fondamental du représentant. La régression à vecteurs de support (à plusieurs facteurs et une seule sortie) est une application désormais bien connue de ce théorème. Nous rappelons rapidement ce type de régression avant de proposer une extension au cas de plusieurs sorties corrélées.

Dans les applications, il n'est généralement pas réaliste de supposer connue la moyenne du processus aléatoire qui modélise un système. Pour cette raison, nous consacrons le **chapitre 4** aux processus aléatoires intrinsèques, qui s'écrivent de manière informelle comme la somme d'un processus aléatoire de moyenne nulle et d'une fonction linéairement paramétrée inconnue. Nous rappelons la formulation du krigeage intrinsèque pour prédire de tels processus aléatoires intrinsèques. Une caractéristique importante de cette prédiction est que la partie paramétrique du processus aléatoire intrinsèque n'est pas régularisée. Cette propriété fournit un moyen efficace d'incorporer de l'information a priori dans un modèle boîte noire. Nous étendons le krigeage intrinsèque à la prédiction de dérivées, dont il nous semble que la formalisation n'avait jamais été publiée.

Le **chapitre 5** aborde le problème du choix de noyau. Notre objectif est d'abord de mieux comprendre quels sont les paramètres d'un noyau qui influent sur la qualité d'une prédiction. L'enjeu est important puisqu'un choix de noyau inadapté peut conduire à des prédicteurs inefficaces, comme cela sera constaté dans le chapitre 6. D'un point de vue théorique, cette partie utilise principalement des résultats de statistique de la littérature, que nous illustrerons par des expériences numériques. Le choix d'une structure de covariance adaptée aux données est généralement suivi d'une phase d'estimation de paramètres. Nous passons en revue différentes méthodes d'estimation utilisables en pratique, en accordant une place privilégiée à la méthode du maximum de vraisemblance.

Le **chapitre 6** présente des applications. Ce chapitre constitue une partie importante de notre travail dans laquelle nous tentons d'évaluer la pertinence pratique des méthodes étudiées. Nous avons choisi d'insister sur les aspects méthodologiques, ce qui nous conduit à ne pas rentrer dans

les détails des problèmes réels traités. Nous essayons enfin de dégager les points communs de ces applications et soulignons les difficultés fréquemment rencontrées tout en suggérant des approches pour y faire face.

Chapitre 1

Modèles statiques boîte noire

Résumé — Ce chapitre présente une introduction élémentaire à la modélisation de type boîte noire des systèmes à plusieurs entrées et plusieurs sorties. Il rappelle et utilise la notion de régression régularisée dans les espaces à noyau reproduisant. Une place particulière est accordée à la théorie du krigeage, qui nous semble occuper un rôle fondamental parmi les méthodes de régression régularisée et faciliter le choix d'un noyau pour un système donné. Dans la théorie du krigeage, en effet, le noyau reproduisant est vu comme une fonction de covariance qui décrit le comportement de ce système.

1.1 Modélisation boîte noire d'un système

La présentation proposée dans ce chapitre introductif se veut simple et didactique, et adopte le point de vue de l'identification de *systèmes statiques*. Les systèmes dynamiques, qui ne sont pas abordés dans ce mémoire, peuvent être traités en considérant par exemple des modèles autorégressifs à temps discret du type $y_{t+1} = f(y_t, y_{t-1}, \ldots, y_{t-m+1}, u_t, u_{t-1}, \ldots, u_{t-p+1})$ où y_t est la sortie du système au temps t et u_t est la commande du système au temps t. Ce type de modèle est formellement identique à un système statique comportant une sortie et $m + p$ entrées. Nous pensons toutefois que l'utilisation de tels modèles dans des applications réelles nécessite une réflexion théorique supplémentaire (par exemple, (Girard et al., 2003) s'intéresse à la propagation des incertitudes dans les modèles dynamiques à temps discret) qui pourra être envisagée comme une des perspectives de continuation de ce travail. Nous commençons par des notions de régression régularisée et d'espace de Hilbert à noyau reproduisant, point de départ essentiel. Nous accordons une place particulière à la théorie du krigeage qui offre des avantages sur lesquels nous insisterons et dont l'utilisation s'avère particulièrement simple. Les géostatisticiens utilisent couramment cette méthode pour modéliser des phénomènes de nature spatiale (Cressie, 1993 ; Chilès et Delfiner, 1999). Rappelons que le nom de la méthode a été choisi par le mathématicien français Georges Matheron (Matheron, 1963) qui a repris et formalisé dans les années soixante les travaux de Krige (Krige, 1951). (La traduction anglaise de krigeage est *Kriging*.)

Nous considérerons un *système* sous forme d'une fonction vectorielle multivariable $\boldsymbol{f}^* : \mathbb{R}^d \to \mathbb{R}^q$. Dans un premier temps, nous supposerons que ce système n'a qu'une seule sortie ($q = 1$). Le terme *facteur* sera employé pour désigner une entrée quelconque du système, un facteur pouvant être n'importe quelle grandeur caractérisant ses conditions d'opération. Nous noterons $\boldsymbol{x} \in \mathbb{R}^d$ le vecteur des facteurs. L'objectif de la modélisation est d'approximer de manière pertinente la ou les sorties $\boldsymbol{f}^*(\boldsymbol{x})$ du système. Lorsque celui-ci est complexe, et que son analyse à partir des équations de la physique se révèle trop compliquée pour aboutir à un modèle de connaissance, une solution est d'en établir un modèle boîte noire. Celui-ci est construit à partir d'un ensemble de données correspondant à des observations des sorties $\boldsymbol{f}^{\text{obs}}_{\boldsymbol{x}_i}$ associées à un ensemble de vecteurs de facteurs connus $\boldsymbol{x}_i$, $i = 1, \ldots, n$. Notons qu'en raison du bruit d'observation généralement présent, $\boldsymbol{f}^{\text{obs}}_{\boldsymbol{x}_i} \neq \boldsymbol{f}^*(\boldsymbol{x}_i)$.

Une approche des plus simples de la modélisation boîte noire est l'approximation linéaire du système dans une région de l'espace des facteurs, par exemple autour d'un point de fonctionnement. Dans le cas $q = 1$, on cherche alors une approximation du type $f(\boldsymbol{x}) = \boldsymbol{b}^\mathsf{T}\boldsymbol{x} + b_0$. Cette démarche est justifiée par la possibilité d'approximer localement une fonction non-linéaire par un développement de Taylor à l'ordre 1. Le vecteur $\boldsymbol{b}$ et le scalaire b_0 constituent les paramètres du modèle et doivent être estimés à partir des données observées. Plusieurs méthodes peuvent être envisagées, la méthode des moindres carrés étant sans aucun doute la plus utilisée. Elle minimise le coût

$$J(\boldsymbol{b}, b_0) = \sum_{i=1}^{n} (f^{\text{obs}}_{\boldsymbol{x}_i} - \boldsymbol{b}^\mathsf{T}\boldsymbol{x}_i - b_0)^2 \,.$$

Il est possible d'étendre cette formulation en incorporant des fonctions non-linéaires en $\boldsymbol{x}$ dans le modèle, habituellement des termes polynomiaux. L'approximation aura alors la forme

$$f(\boldsymbol{x}) = \boldsymbol{b}^\mathsf{T}\boldsymbol{r}(\boldsymbol{x}) \,. \tag{1.1}$$

On peut ainsi rendre compte de comportements non-linéaires autour d'un point de fonctionnement ou modéliser les tendances globales du système.

1.2 Prédiction par régression régularisée

Le nombre de paramètres du modèle (1.1) croît rapidement avec la dimension d de l'espace des facteurs (un polynôme complet de degré l à d indéterminées comporte C^d_{d+l} termes). Or un grand nombre de termes paramétriques conduit la plupart du temps à un problème d'estimation mal posé et à des instabilités numériques. Dans la méthode des moindres carrés, calculer le conditionnement de la matrice $\boldsymbol{R} = (\boldsymbol{r}(\boldsymbol{x}_1) \cdots \boldsymbol{r}(\boldsymbol{x}_n))$ permet de le vérifier. Si le problème est mal posé, la solution classique consiste à adapter la méthode des moindres carrés en incorporant un terme de régularisation à la fonction coût. Le problème régularisé peut par exemple être formulé comme

Minimiser le critère régularisé

$$J(\boldsymbol{b}) = C\|\boldsymbol{b}\|_2^2 + \sum_{i=1}^{n} (f^{\text{obs}}_{\boldsymbol{x}_i} - \boldsymbol{b}^\mathsf{T}\boldsymbol{r}(\boldsymbol{x}_i))^2 \,, \tag{1.2}$$

qui pénalise les grandes valeurs des composantes de $\boldsymbol{b}$ et où la constante positive C permet de contrôler le compromis entre régularisation et attache aux données.

Dans ce type de modèle paramétrique, la différence entre les données observées et l'approximation n'est pas nécessairement liée au bruit d'observation, car imposer une structure paramétrique de faible dimension pour le modèle conduit nécessairement à une capacité d'approximation limitée. Ainsi, si la dimension de $\boldsymbol{b}$ est faible, même si le bruit d'observation est négligeable, le modèle ne redonne pas les valeurs observées pour les sorties du système mais les approxime. Il est de même souvent difficile d'obtenir un modèle paramétrique capable de suivre les principales tendances du système tout en rendant localement compte des variations autour d'un point de fonctionnement. Une solution théoriquement envisageable serait d'augmenter la dimension de $\boldsymbol{b}$ en ajoutant autant de termes paramétriques que nécessaire. Cette solution conduit cependant à des approximations très irrégulières, avec des variations non désirées entre les points observés. Pour contrôler la régularité de l'approximation, il est possible de pénaliser les comportements irréguliers. Ceci complique la mise en œuvre, et on se restreint donc à des modèles de taille limitée. Il est cependant possible, comme nous le verrons dans la section qui suit, de construire des modèles mieux adaptés en partant de (1.1) et (1.2) grâce au concept de *noyaux reproduisants*. La régularité de l'approximation sera alors contrôlée par le choix du noyau.

1.3 Régression régularisée dans les espaces de Hilbert à noyau reproduisant

1.3.1 Mise en place du formalisme

Remarquons que la solution du problème d'optimisation sans contrainte (1.2) doit satisfaire la condition $\frac{\partial J}{\partial \boldsymbol{b}} = \boldsymbol{0}$, ce qui implique

$$\widehat{\boldsymbol{b}} = \frac{1}{C} \sum_{i=1}^{n} (f^{\text{obs}}_{\boldsymbol{x}_i} - \widehat{\boldsymbol{b}}^{\mathsf{T}} \boldsymbol{r}(\boldsymbol{x}_i)) \boldsymbol{r}(\boldsymbol{x}_i) \,.$$

Il existe par conséquent des scalaires $\widehat{a}_i$ tels que

$$\widehat{\boldsymbol{b}} = \sum_{i=1}^{n} \widehat{a}_i \boldsymbol{r}(\boldsymbol{x}_i) \,.$$

La formulation (1.1), (1.2) est équivalente à chercher $\hat{f}$ sous la forme $f(\boldsymbol{x}) = \sum_i a_i(\boldsymbol{r}(\boldsymbol{x}_i), \boldsymbol{r}(\boldsymbol{x}))$, où $(.,.)$ est le produit scalaire dans $\mathbb{R}^l$, en minimisant

$$J(a_i, i = 1, \ldots, n) = C \left\| \sum_{i=1}^{n} a_i \boldsymbol{r}(\boldsymbol{x}_i) \right\|^2 + \sum_{i=1}^{n} (f^{\text{obs}}_{\boldsymbol{x}_i} - f(\boldsymbol{x}_i))^2 \,.$$

Comme le problème est quadratique, les $\widehat{a}_i$ optimaux s'obtiennent simplement. De plus,

$$\left\| \sum_{i=1}^{n} a_i \boldsymbol{r}(\boldsymbol{x}_i) \right\|^2 = \sum_{i,j} a_i (\boldsymbol{r}(\boldsymbol{x}_i), \boldsymbol{r}(\boldsymbol{x}_j)) a_j \,,$$

de sorte que le nouveau problème régularisé s'exprime uniquement en fonction des produits scalaires $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{r}(\boldsymbol{x}_i), \boldsymbol{r}(\boldsymbol{x}_j))$, où $k(.,.)$ est une fonction appelée par la suite *noyau reproduisant*. Cette remarque est fondamentale parce qu'il devient possible de se donner a priori un produit scalaire sous la forme $k(\boldsymbol{x}, \boldsymbol{y})$ sans avoir à évaluer les objets $\boldsymbol{r}(\boldsymbol{x}_i)$.

La famille de vecteurs $\{\boldsymbol{r}(\boldsymbol{x}) \in \mathbb{R}^l, \boldsymbol{x} \in \mathbb{R}^d\}$ génère un espace $\mathcal{F} \subseteq \mathbb{R}^l$ auquel $\widehat{\boldsymbol{b}}$ appartient. Cet espace $\mathcal{F}$ est appelé l'*espace des caractéristiques* (*feature space* en anglais). Si $\mathcal{F}$ est de petite dimension, la capacité d'approximation du modèle est limitée. Toute la suite va être consacrée à voir comment on peut aisément augmenter la dimension de $\mathcal{F}$. Remarquons que l'on peut identifier $\mathcal{F}$ à un espace de fonctions puisqu'à tout élément $\boldsymbol{b} = \sum_{i=1}^{n} a_i \boldsymbol{r}(\boldsymbol{x}_i) \in \mathcal{F}$ correspond une fonction définie par

$$\forall \boldsymbol{x} \in \mathbb{R}^d, \quad f(\boldsymbol{x}) = (\boldsymbol{r}(\boldsymbol{x}), \sum_{i=1}^{n} a_i \boldsymbol{r}(\boldsymbol{x}_i)) \,.$$

(Nous utilisons par la suite la notation $f \equiv \boldsymbol{b}$ pour rappeler la propriété d'identification.) Dans $\mathcal{F}$, on définit une norme par

$$\|f\|_{\mathcal{F}}^2 = \|\boldsymbol{b}\|_2^2 = \sum_{i,j=1}^{n} a_i (\boldsymbol{r}(\boldsymbol{x}_i), \boldsymbol{r}(\boldsymbol{x}_j)) a_j \,.$$

Remarquons aussi le rôle particulier joué par $\boldsymbol{r}(\boldsymbol{x})$ qui permet d'évaluer toute fonction de $\mathcal{F}$ en un point $\boldsymbol{x}$ en formant le produit scalaire $f(\boldsymbol{x}) = (\boldsymbol{r}(\boldsymbol{x}), \boldsymbol{b})$. Il est donc possible d'identifier $\boldsymbol{r}(\boldsymbol{x})$ à l'opérateur d'évaluation $f \mapsto f(\boldsymbol{x})$, que nous noterons $\delta_{\boldsymbol{x}}$ par la suite. Cet opérateur d'évaluation ponctuelle est linéaire et continu en raison des propriétés du produit scalaire. Le fait que $(\boldsymbol{r}(\boldsymbol{y}), \boldsymbol{r}(\boldsymbol{x})) = (\boldsymbol{r}(\boldsymbol{x}), \boldsymbol{r}(\boldsymbol{y})) = k(\boldsymbol{x}, \boldsymbol{y})$, $\forall \boldsymbol{y}$, entraîne que l'on peut identifier $\boldsymbol{r}(\boldsymbol{x})$ et la fonction $k(\boldsymbol{x}, .)$ et que $k(\boldsymbol{x}, .) \in \mathcal{F}$.

L'objectif fixé peut maintenant être atteint, puisqu'il est possible de se donner a priori un produit scalaire $k(\boldsymbol{x}, \boldsymbol{y})$ et que le vecteur des caractéristiques $\boldsymbol{r}(\boldsymbol{x}) \equiv k(\boldsymbol{x}, .)$ n'a pas besoin d'être évalué lors de la minimisation de (1.2), ce qui est particulièrement intéressant s'il s'agit d'un vecteur de très grande dimension. Par la suite, $\boldsymbol{r}(\boldsymbol{x})$ sera d'ailleurs vu non plus comme un vecteur de dimension fini, mais comme une fonction (en tant que vecteur de dimension infini). L'espace des caractéristiques $\mathcal{F}$, engendré par les combinaisons linéaires $f = \sum_i a_i k(\boldsymbol{x}_i, .)$, sera alors lui aussi de dimension infinie. Les fonctions de $\mathcal{F}$ seront évaluées en un point $\boldsymbol{x}$ en formant leur produit scalaire avec $k(\boldsymbol{x}, .)$.

1.3.2 Principales propriétés des espaces à noyau reproduisant

Dans cette section, les principaux résultats concernant les espaces de Hilbert à noyau reproduisant sont rappelés de manière informelle. Ces résultats sont très classiques et le lecteur les trouvera dans de nombreux ouvrages consacrés à l'analyse fonctionnelle (par exemple (Yosida, 1980)). Par définition, une fonction $k : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ est dite *noyau reproduisant* d'un espace hilbertien $\mathcal{F}$ de fonctions sur $\mathbb{X} \subset \mathbb{R}^d$, muni du produit scalaire $(\cdot, \cdot)_{\mathcal{F}}$, si toute fonction f de $\mathcal{F}$ vérifie la propriété

$$f(\boldsymbol{x}) = (k(\boldsymbol{x}, .), f)_{\mathcal{F}} \,. \tag{1.3}$$

Cette définition suppose implicitement que la fonction $k(\boldsymbol{x}, .)$ appartient à $\mathcal{F}$. Elle explique l'emploi à la section précédente du terme noyau reproduisant de $\mathcal{F}$ pour $k(\boldsymbol{x}, .) \equiv \boldsymbol{r}(\boldsymbol{x})$.

Si k est un noyau reproduisant, en prenant $f = k(\boldsymbol{y}, .)$, et par symétrie du produit scalaire on obtient une première propriété importante :

$$k(\boldsymbol{y}, \boldsymbol{x}) = (k(\boldsymbol{x}, .), k(\boldsymbol{y}, .)_{\mathcal{F}} = k(\boldsymbol{x}, \boldsymbol{y}) \, .$$

Puisque $k(\boldsymbol{x}, \boldsymbol{y})$ est un produit scalaire, c'est une fonction de type positif, c'est-à-dire que si $f = \sum_{i=1}^{n} a_i k(\boldsymbol{x}_i, .)$ alors $\sum_{i,j} a_i a_j k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \|f\|_{\mathcal{F}}^2 \geq 0$. La deuxième propriété importante est que l'espace vectoriel $\tilde{\mathcal{F}}$ engendré par les fonctions $k(\boldsymbol{x}, .)$ lorsque $\boldsymbol{x}$ parcourt $\mathbb{X}$ est dense dans $\mathcal{F}$.

Il est naturel de se demander à quelles conditions un espace hilbertien de fonctions sur $\mathbb{X}$ admet un noyau reproduisant. Le résultat classique est qu'un espace $\mathcal{F}$ de fonctions $f : \mathbb{X} \to \mathbb{R}$ admet un noyau reproduisant si et seulement si

$$\forall \boldsymbol{x} \in \mathbb{X}, \exists M_{\boldsymbol{x}} \geq 0 \text{ tel que } |f(\boldsymbol{x})| \leq M_{\boldsymbol{x}} \|f\|_{\mathcal{F}} \, , \; \forall f \in \mathcal{F} \, . \tag{1.4}$$

Rappelons que $\delta_{\boldsymbol{x}}$ désigne la forme linéaire $f \mapsto f(\boldsymbol{x})$. La condition (1.4) exprime que cette forme linéaire est continue sur $\mathcal{F}$. Cette condition est assez forte, mais souvent le problème de savoir si elle est vérifiée ne se pose pas. On choisit en effet fréquemment d'abord un noyau, ce qui détermine implicitement l'espace de fonctions engendré. (Dans le cas des splines, cependant, l'espace de fonctions est construit en premier.)

Le dernier résultat important que nous mentionnerons est le *théorème de représentation* (Kimeldorf et Wahba, 1971). Ce théorème fondamental explicite la forme du minimiseur pour une version généralisée du problème (1.2). Soit $\mathcal{F}$ un espace hilbertien à noyau reproduisant, on cherche $\hat{f} \in \mathcal{F}$ qui minimise

$$C\|f\|_{\mathcal{F}}^2 + \sum_{i=1}^{n} l(f_{\boldsymbol{x}_i}^{\text{obs}}, f(\boldsymbol{x}_i)) \, , \tag{1.5}$$

où l est convexe en f. Alors $\hat{f}$ admet la représentation

$$\hat{f}(.) = \sum_{i=1}^{n} \widehat{a}_i k(\boldsymbol{x}_i, .) \, , \tag{1.6}$$

et (1.6) substitué dans (1.5) permet de calculer la fonction optimale par optimisation numérique dans un espace de dimension finie. Ce théorème montre que même si $\mathcal{F}$ est un espace de dimension infinie, la solution est toujours dans l'espace vectoriel de dimension finie engendré par les $k(\boldsymbol{x}_i, .)$, $i = 1, \ldots, n$ et peut donc être explicitée.

Les applications de ce théorème concernent non seulement le problème de régression en dimension finie vu plus haut, mais aussi les splines (Schoenberg, 1964 ; Wahba, 1990 ; Kybic et al., 2002), qui sont les solutions d'un problème régularisé, l'approximation par fonctions de base radiales (Micchelli, 1986 ; Powell, 1987), ainsi que les méthodes à vecteurs de support plus récemment développées (Vapnik, 1995 ; Schölkopf et Smola, 2002). Dans tout problème de régression régularisée, le choix du noyau reproduisant caractérise l'espace $\mathcal{F}$ et donc les propriétés de l'approximation ou du modèle boîte noire. Il est donc important de choisir ce noyau correctement. Nous pensons que ce choix est facilité en adoptant le point de vue de la modélisation par processus aléatoires, où le noyau est vu comme une fonction de covariance.

1.4 Modélisation par processus aléatoires

1.4.1 Processus aléatoires et krigeage

Les principes de la prédiction linéaire de processus aléatoires (ou krigeage) sont très sommairement présentés dans cette section, et nous montrerons dans la suivante que cette prédiction peut être vue comme une régression régularisée. Le système (supposé déterministe) est maintenant modélisé par un processus aléatoire noté $F(\boldsymbol{x})$. Ainsi, les données observées seront considérées comme des réalisations des variables aléatoires $F(\boldsymbol{x}_i)$. Si en outre la sortie est corrompue par un bruit de mesure additif, ce bruit est modélisé par des variables aléatoires indépendantes N_i. Pour simplifier la présentation, nous supposerons ici le bruit négligeable. La variable aléatoire $F(\boldsymbol{x})$ modélise donc l'incertitude sur la sortie qui n'a pas été observée. Cependant la connaissance d'observations au voisinage de $\boldsymbol{x}$ est susceptible de faire diminuer cette incertitude. En effet, si on a effectué une observation au point $\boldsymbol{x}_i$, on s'attend à ce que la valeur de la sortie ne change pas brusquement si le vecteur des facteurs reste au voisinage de $\boldsymbol{x}_i$. Cette hypothèse se traduit par l'existence d'une corrélation entre les variables aléatoires $F(\boldsymbol{x})$ et $F(\boldsymbol{x}_i)$, $i = 1, \ldots, n$. Si cette corrélation peut être déterminée, soit par analyse des données, soit d'après une connaissance a priori, il est possible de prédire $F(\boldsymbol{x})$ en fonction des $F(\boldsymbol{x}_i)$. Pour effectuer cette prédiction, les hypothèses sur $F(\boldsymbol{x})$ doivent être précisées.

Nous supposerons que $F(\boldsymbol{x})$ est un processus du second ordre, c'est-à-dire que la variance de $F(\boldsymbol{x})$ est finie pour tout $\boldsymbol{x}$. Nous supposerons aussi connues la moyenne de $F(\boldsymbol{x})$, ainsi que sa fonction de covariance. (L'hypothèse de connaissance de la moyenne sera assouplie plus loin.) Puisque la moyenne est supposée connue, elle peut être soustraite, et il est donc possible de considérer uniquement des processus $F(\boldsymbol{x})$ à moyenne nulle. $F(\boldsymbol{x})$ est alors essentiellement caractérisé par sa fonction de covariance $k(\boldsymbol{x}, \boldsymbol{y}) = \mathrm{Cov}[F(\boldsymbol{x}), F(\boldsymbol{y})] = \mathrm{E}[F(\boldsymbol{x})F(\boldsymbol{y})]$. Rappelons que $\mathrm{E}[XY]$ définit un produit scalaire dans l'espace des variables aléatoires de variance finie $L^2(\Omega, \mathcal{A}, P)$; $k(\boldsymbol{x}, \boldsymbol{y})$ sera donc vu plus loin comme un produit scalaire.

La méthode la plus simple pour prédire $F(\boldsymbol{x})$ est de calculer la meilleure projection linéaire $\widehat{F}(\boldsymbol{x})$ de $F(\boldsymbol{x})$ sur l'espace $\mathcal{H}_S$ généré par les variables aléatoires observées $F(\boldsymbol{x}_i)$, $i = 1, \ldots, n$. Ceci correspond à la formulation du krigeage, et signifie que l'on cherche un prédicteur linéaire $\widehat{F}(\boldsymbol{x}) = \sum_i \widehat{\lambda}_{i,\boldsymbol{x}} F(\boldsymbol{x}_i)$ tel que $\mathrm{Var}(\widehat{F}(\boldsymbol{x}) - F(x)) = \|\widehat{F}(\boldsymbol{x}) - F(x)\|^2$ soit minimum ou, de manière équivalente puisque la meilleure prédiction linéaire est la projection orthogonale sur $\mathcal{H}_S$, tel que

$$\begin{aligned}(\widehat{F}(\boldsymbol{x}) - F(\boldsymbol{x}), F(\boldsymbol{x}_i)) &= \mathrm{Cov}[\widehat{F}(\boldsymbol{x}) - F(x), F(\boldsymbol{x}_i)] \\ &= 0\,, \quad \forall i \in \{1, \ldots, n\}\,.\end{aligned} \tag{1.7}$$

En remplaçant $\widehat{F}(\boldsymbol{x})$ par son expression en fonction des $F(\boldsymbol{x}_i)$, on vérifie que le vecteur $\widehat{\boldsymbol{\lambda}}_{\boldsymbol{x}} = (\widehat{\lambda}_{1,\boldsymbol{x}}, \ldots, \widehat{\lambda}_{n,\boldsymbol{x}})^\mathsf{T}$ est solution du système linéaire

$$\boldsymbol{K}\widehat{\boldsymbol{\lambda}}_{\boldsymbol{x}} = \boldsymbol{k}_{\boldsymbol{x}}\,, \tag{1.8}$$

où $\boldsymbol{K}$ est la matrice des covariances $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ et $\boldsymbol{k}_{\boldsymbol{x}}$ est le vecteur des covariances $k(\boldsymbol{x}, \boldsymbol{x}_i)$. La matrice $\boldsymbol{K}$ est en principe de rang plein. Cependant, notamment lorsque certaines observations sont très rapprochées, cette matrice devient en général mal conditionnée et le calcul de $\widehat{\boldsymbol{\lambda}}_{\boldsymbol{x}}$ doit être fait avec précaution, par exemple à partir d'une décomposition en valeurs singulières tronquée de $\boldsymbol{K}$.

Notons que le prédicteur obtenu est sans biais, puisque la moyenne de $F(\boldsymbol{x})$ est connue. Lorsqu'il existe i tel que $\boldsymbol{x} = \boldsymbol{x}_i$, la meilleure prédiction linéaire en l'absence de bruit de mesure est naturellement $\widehat{F}(\boldsymbol{x}_i) = F(\boldsymbol{x}_i)$. Par conséquent, si on observe une sortie en un point de l'espace des facteurs, alors la prédiction est égale à cette observation. Le krigeage réalise donc une interpolation des sorties mesurées sur le système, utilisable comme modèle boîte noire construit, comme dans la section 1.1, à partir d'observations antérieures. Nous verrons plus loin que la prédiction obtenue par krigeage équivaut en fait à une régression régularisée. On peut également calculer la variance de l'erreur de prédiction, obtenue simplement par la relation de Pythagore

$$\begin{aligned} \operatorname{Var}(\hat{F}(\boldsymbol{x}) - F(\boldsymbol{x})) &= \operatorname{Var} F(\boldsymbol{x}) - \operatorname{Var} \hat{F}(\boldsymbol{x}) \\ &= k(\boldsymbol{x},\boldsymbol{x}) - \widehat{\boldsymbol{\lambda}}_{\boldsymbol{x}}^{\mathsf{T}} \boldsymbol{K} \widehat{\boldsymbol{\lambda}}_{\boldsymbol{x}} \\ &= k(\boldsymbol{x},\boldsymbol{x}) - \boldsymbol{k}_{\boldsymbol{x}}{}^{\mathsf{T}} \boldsymbol{K}^{-1} \boldsymbol{k}_{\boldsymbol{x}} \,. \end{aligned} \tag{1.9}$$

Il est ainsi possible d'évaluer l'incertitude associée à la prédiction obtenue en $\boldsymbol{x}$. Une illustration est présentée sur la figure 1.1, qui représente l'interpolation de données en dimension 1 et pour laquelle on donne des intervalles de confiances calculés à partir de la variance de l'erreur de prédiction. Dans ce cadre probabiliste, une autre approche pourrait en principe être envisagée pour prédire la sortie du système. Il serait en effet possible de simuler $F(\boldsymbol{x})$ conditionnellement aux données observées, ce qui est également montré sur la figure 3.1. Il existe évidemment une infinité de trajectoires possibles, ce qui justifie le calcul de la meilleure prédiction linéaire. Cependant, les simulations conditionnelles restent parfois utiles, notamment lorsque l'on s'intéresse à des traitements non-linéaires, comme la prédiction d'un seuillage.

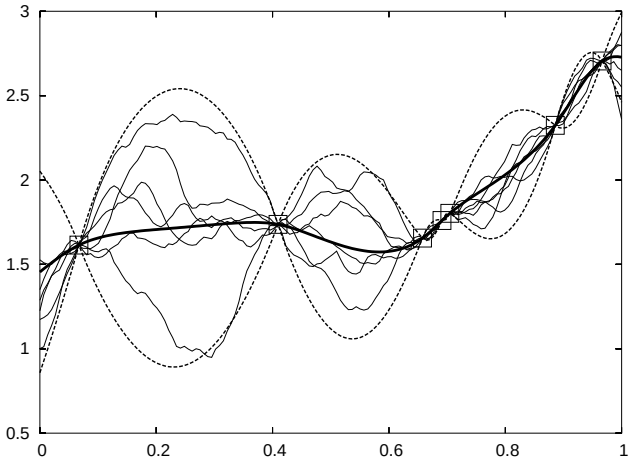


FIG. 1.1 – Exemple en dimension 1. Les observations matérialisées par les carrés sont interpolées par la courbe en trait gras continu obtenue par krigeage. Les courbes en pointillés représentent les intervalles de confiance à 95% pour la prédiction. L'incertitude augmente loin des observations. Les courbes en traits fins continus sont des simulations conditionnelles, qui sont également des interpolations. Le krigeage réalise une moyenne de ces trajectoires.

Avant de conclure cette section, remarquons que la prédiction linéaire est construite uniquement à partir des propriétés du second ordre de $F(\boldsymbol{x})$. Rappelons que si $F(\boldsymbol{x})$ est un processus

aléatoire gaussien, c'est-à-dire lorsque $\forall n \in \mathbb{N}$, $\forall\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ et $\forall\{\lambda_1, \ldots, \lambda_n\}$, $\sum_{i=1}^n \lambda_i F(\boldsymbol{x}_i)$ est une variable aléatoire de distribution gaussienne, alors la prédiction linéaire est optimale (il s'agit de l'espérance de $F(\boldsymbol{x})$ conditionnellement aux variables aléatoires observées). Si le processus est loin d'être gaussien, la prédiction linéaire peut ne pas être adaptée, et des méthodes spécifiques peuvent alors être utilisées (voir le krigeage non-linéaire dans (Chilès et Delfiner, 1999)).

1.4.2 Liens avec la régression régularisée

Des analogies entre régression régularisée et krigeage sont peut-être déjà apparues au lecteur dans la section précédente. Nous précisons maintenant ces analogies et rappelons un résultat classique sur l'équivalence entre krigeage et régression régularisée. Il est possible d'aller directement à la conclusion de cette section en première lecture.

Les analogies entre ces méthodes proviennent essentiellement du fait qu'elles utilisent des espaces de structure identique. Trois espaces isométriques peuvent ainsi être considérés.

Soit d'abord le processus aléatoire à moyenne nulle $F(\boldsymbol{x}) : \mathbb{R}^d \to L^2(\Omega, \mathcal{A}, P)$. Soit $\mathcal{H}$, l'espace complété du sous-espace de $L^2(\Omega, \mathcal{A}, P)$ engendré par $F(\boldsymbol{x})$, $\boldsymbol{x} \in \mathbb{R}^d$, muni du produit scalaire $(X, Y)_{\mathcal{H}} = \mathrm{E}[XY]$. La fonction $k(\boldsymbol{x}, \boldsymbol{y})$, définie comme la covariance de $F(\boldsymbol{x})$ et $F(\boldsymbol{y})$, permet de calculer les produits scalaires des éléments de $\mathcal{H}$ puisque

$$(\sum_i \lambda_i F(\boldsymbol{x}_i), \sum_j \mu_j F(\boldsymbol{x}_j))_{\mathcal{H}} = \sum_{i,j} \lambda_i k(\boldsymbol{x}_i, \boldsymbol{x}_j) \mu_j \,.$$

Considérons ensuite $\tilde{\Lambda}$ l'espace vectoriel des mesures à support fini sur $\mathbb{R}^d$. Les éléments de $\tilde{\Lambda}$ seront notés $\lambda = \sum_i \lambda_i \delta_{\boldsymbol{x}_i}$. On prolonge $F(\boldsymbol{x})$ sur $\tilde{\Lambda}$ en définissant l'application linéaire

$$\begin{array}{rccc} F: & \tilde{\Lambda} & \to & \mathcal{H} \\ & \lambda = \sum_i \lambda_i \delta_{\boldsymbol{x}_i} & \mapsto & F(\lambda) = \sum_i \lambda_i F(\boldsymbol{x}_i) \,. \end{array} \tag{1.10}$$

Cette application est injective si et seulement si la fonction $k(.,.)$ est définie positive, ce que nous supposerons. Dans $\tilde{\Lambda}$, on considère le produit scalaire $(\lambda, \mu)_{\tilde{\Lambda}} = (F(\lambda), F(\mu))_{\mathcal{H}}$. En particulier, $(\delta_{\boldsymbol{x}}, \delta_{\boldsymbol{y}})_{\tilde{\Lambda}} = k(\boldsymbol{x}, \boldsymbol{y})$. Notons Λ le complété de $\tilde{\Lambda}$ par rapport à ce produit scalaire. F est alors une isométrie bijective entre Λ et $\mathcal{H}$.

Considérons enfin Λ comme le dual topologique d'un espace hilbertien de fonctions $\mathcal{F}$, c'est-à-dire comme l'ensemble des formes linéaires continues sur $\mathcal{F}$. Puisque $\delta_{\boldsymbol{x}} \in \Lambda$, $\mathcal{F}$ est par construction un espace de Hilbert à noyau reproduisant car (1.4) est vérifiée. On sait alors, d'après ce qui précède, que le noyau de $\mathcal{F}$ est $k(\boldsymbol{x}, \boldsymbol{y})$, la fonction de covariance de $F(\boldsymbol{x})$, et que $\mathcal{F}$ est engendré par ce noyau. En conclusion, Λ, $\mathcal{F}$ et $\mathcal{H}$ possèdent le même produit scalaire et on peut s'attendre à ce que, pour une fonction de covariance donnée $k(\boldsymbol{x}, \boldsymbol{y})$, la prédiction par krigeage du processus aléatoire $F(\boldsymbol{x})$ et la régression régularisée à noyau $k(\boldsymbol{x}, \boldsymbol{y})$ soient équivalentes.

Ces résultats d'équivalence ont été démontrés dans (Kimeldorf et Wahba, 1970a,b) et (Matheron, 1981). En l'absence de bruit d'observation, le critère de régression régularisée correspondant à (1.5) peut être modifié pour obtenir une interpolation équivalente au krigeage, en disant que l'on minimise $\|f\|^2_{\mathcal{F}}$ sous les contraintes

$$f(\boldsymbol{x}_i) = f^{\mathrm{obs}}_{\boldsymbol{x}_i} \,, \quad \forall i \in \{1, \ldots, n\} \,. \tag{1.11}$$

Sans donner ici la preuve de cette équivalence, remarquons simplement comme point de départ que les contraintes (1.11) peuvent être réécrites

$$(k(\boldsymbol{x}_i,.), f - f^{\text{obs}})_{\mathcal{F}} = 0\,, \quad \forall i\,,$$

où $f^{\text{obs}} \in \mathcal{F}$ est la fonction générant les données observées. C'est donc l'orthogonalité de $f - f^{\text{obs}}$ avec l'espace $\mathcal{F}_S = \text{vect}\{k(\boldsymbol{x}_1,.), \ldots, k(\boldsymbol{x}_n,.)\}$ qui ressemble à la condition $\widehat{F}(\boldsymbol{x}) - F(\boldsymbol{x}) \perp \mathcal{H}_S$ vue plus haut. Nous présenterons à la section suivante le cas du krigeage avec bruit et verrons que la prédiction obtenue est alors équivalente à la solution du problème (1.5).

Quel que soit le point de vue adopté, le concept de noyau reproduisant occupe un rôle essentiel. Il est possible de partir d'un espace hilbertien de fonctions, par exemple en spécifiant une norme, ou en explicitant un vecteur de caractéristiques $\boldsymbol{r}(\boldsymbol{x})$, et d'en déduire le noyau reproduisant correspondant. Il est aussi possible de se donner arbitrairement un noyau. Cependant, choisir un noyau adapté à des données observées peut s'avérer délicat. Nous pensons que l'avantage du point de vue probabiliste est qu'il donne des outils statistiques pour choisir un noyau à partir des observations (voir la section 1.5).

1.4.3 Autres formes de krigeage

Dans cette section nous rappelons comment on peut adapter le krigeage pour obtenir des modèles de type boîte noire lorsque les observations sont bruitées. Nous montrons ensuite comment modéliser des systèmes à plusieurs sorties corrélées. Nous indiquons enfin ce qu'il faut faire lorsque la moyenne du processus est inconnue.

Prise en compte du bruit d'observation

Nous avons vu que le krigeage est une prédiction linéaire optimale, c'est à dire que l'on projette $F(\boldsymbol{x})$ (toujours supposé à moyenne nulle pour l'instant) sur l'espace $\mathcal{H}_S$ des variables aléatoires $F(\boldsymbol{x}_i)$. Il est naturel d'étendre cette formulation au cas où l'on projette sur d'autres variables aléatoires que les $F(\boldsymbol{x}_i)$. Par exemple, si la sortie est corrompue par un bruit blanc additif centré N de variance σ_N^2, les variables aléatoires observées sont $F^{\text{obs}}(\boldsymbol{x}_i) = F(\boldsymbol{x}_i) + N$. On cherche alors le prédicteur $\widehat{F}(\boldsymbol{x}) = \sum_{i=1}^n \lambda_{i,\boldsymbol{x}} F^{\text{obs}}(\boldsymbol{x}_i)$ tel que $(\widehat{F}(\boldsymbol{x}) - F(\boldsymbol{x}), F^{\text{obs}}(\boldsymbol{x}_i)) = 0$, $\forall i \in \{1, \ldots, n\}$. La covariance de $F^{\text{obs}}(\boldsymbol{x}_i)$ et $F^{\text{obs}}(\boldsymbol{x}_j)$ est $k(\boldsymbol{x}_i, \boldsymbol{x}_j) + \sigma_N^2$ si $\boldsymbol{x}_i = \boldsymbol{x}_j$, et $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ sinon. Le système linéaire à résoudre devient

$$(\boldsymbol{K} + \sigma_N^2 \boldsymbol{I}_n)\widehat{\boldsymbol{\lambda}}_{\boldsymbol{x}} = \boldsymbol{k}_{\boldsymbol{x}}\,,$$

où $\boldsymbol{I}_n$ est la matrice identité.

Système à plusieurs sorties corrélées

Considérerons maintenant un système à plusieurs sorties. Soit $F_\alpha(\boldsymbol{x})$, un processus aléatoire modélisant la sortie indicée α du système. Connaissant les covariances et inter-covariances des processus aléatoires

$$k_{\alpha,\beta}(\boldsymbol{x}, \boldsymbol{y}) = \text{Cov}[F_\alpha(\boldsymbol{x}), F_\beta(\boldsymbol{y})]\,,$$

on peut prédire par projection orthogonale n'importe quelle sortie ou combinaison linéaire de ces sorties en fonction d'observations $F_{\alpha_i}(\boldsymbol{x}_i)$. Les méthodes permettant de choisir les covariances et

les inter-covariances sont les mêmes que pour des covariances simples (voir la section 1.5). Notons que l'on peut écrire formellement $k_{\alpha,\beta}(\boldsymbol{x},\boldsymbol{y}) = k([\alpha\ \boldsymbol{x}],[\beta\ \boldsymbol{y}])$, ce qui revient à considérer que l'on ajoute un facteur supplémentaire dont la valeur indice la sortie du système. En conséquence, l'ensemble des covariances et inter-covariances peut être considéré comme un noyau reproduisant unique (en ajoutant un facteur supplémentaire) et ce noyau peut être utilisé sans changement dans le cadre de la régression régularisée (voir par exemple (Vazquez et Walter, 2003b)). Le point de vue probabiliste du krigeage permet donc très simplement de construire des régressions à plusieurs entrées et plusieurs sorties, en tenant compte des éventuelles corrélations entre sorties.

Moyenne du processus inconnue

Supposons maintenant que $\mathrm{E}[F(\boldsymbol{x})] = b$, avec $b \in \mathbb{R}$ inconnu. On remarque que les combinaisons linéaires telles que $\sum_i \lambda_i F(\boldsymbol{x}_i)$ avec $\sum_i \lambda_i = 0$ filtrent la moyenne inconnue de $F(\boldsymbol{x})$ dans le sens où $\mathrm{E}[\sum_i \lambda_i F(\boldsymbol{x}_i)] = 0$. L'idée est donc de se ramener au cas à moyenne nulle en utilisant de tels accroissements. On considère plus généralement des accroissements d'ordre r de $F(\boldsymbol{x})$ définis par les variables aléatoires

$$F(\lambda) = \sum_{i=1}^{n} \lambda_i F(\boldsymbol{x}_i) \tag{1.12}$$

telles que pour tout $\boldsymbol{l} = (l_1 \dots l_d)^\mathsf{T}$, $0 \leq l_1 + \dots + l_d \leq r$,

$$\sum_{i=1}^{n} \lambda_i \boldsymbol{x}_i^{\boldsymbol{l}} = 0\,, \tag{1.13}$$

où, pour $\boldsymbol{x} = (x_{[1]} \cdots x_{[d]})^\mathsf{T}$, $\boldsymbol{x}^{\boldsymbol{l}}$ est le monôme $x_{[1]}^{l_1} \cdots x_{[d]}^{l_d}$. Cette dernière propriété peut être vue comme l'orthogonalité des mesures $\lambda \in \Lambda$ avec les monômes $\boldsymbol{x}^{\boldsymbol{l}}$. Par conséquent, ces mesures filtrent toute moyenne polynomiale de $F(\boldsymbol{x})$ (d'ordre inférieur ou égal à r). Sur l'ensemble de ces accroissements d'ordre r, on considère des covariances généralisées $k(\boldsymbol{x},\boldsymbol{y})$ qui permettent de calculer les covariances :

$$\mathrm{Cov}[F(\lambda), F(\mu)] = \sum_{i,j} \lambda_i k(\boldsymbol{x}_i, \boldsymbol{x}_j) \mu_j\,. \tag{1.14}$$

Les covariances généralisées sont de type *conditionnellement positif*, c'est-à-dire qu'elles doivent garantir $\mathrm{Var}[\tilde{F}(\lambda)] \geq 0$ pour tout accroissement $F(\lambda)$ d'ordre r. La classe des covariances est donc incluse dans celle des covariances généralisées.

Le krigeage dit *intrinsèque* (Matheron, 1973) construit un prédicteur $\hat{F}(\boldsymbol{x}) = \sum_{i=1}^{n} \widehat{\lambda}_{i,\boldsymbol{x}} F(\boldsymbol{x}_i)$ minimisant $\mathrm{Var}[\widehat{F}(\boldsymbol{x}) - F(\boldsymbol{x})]$ sous la contrainte que $\widehat{F}(\boldsymbol{x}) - F(\boldsymbol{x})$ soit un accroissement d'ordre r. On est ainsi amené a résoudre le système linéaire :

$$\begin{pmatrix} \boldsymbol{K} & \boldsymbol{P}^\mathsf{T} \\ \boldsymbol{P} & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{\lambda}_{\boldsymbol{x}} \\ \boldsymbol{\mu}_{\boldsymbol{x}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{k}_{\boldsymbol{x}} \\ \boldsymbol{p}_{\boldsymbol{x}} \end{pmatrix}, \tag{1.15}$$

où $\boldsymbol{P}$ est une matrice $m \times n$ dont les éléments sont les $\boldsymbol{x}_i^{\boldsymbol{l}}$ (m étant le nombre de monômes de d variables de degré inférieur ou égal à r), $\boldsymbol{\mu}_{\boldsymbol{x}}$ est un vecteur de coefficients de Lagrange et $\boldsymbol{p}_{\boldsymbol{x}}$ est le vecteur des monômes $\boldsymbol{x}^{\boldsymbol{l}}$.

La théorie du krigeage intrinsèque et des fonctions conditionnellement positives est mathématiquement plus difficile que celle du krigeage des processus aléatoires à moyenne connue. Sa mise

en œuvre reste toutefois d'un niveau de simplicité élementaire et, outre le fait qu'il est possible de traiter les processus aléatoires à moyenne inconnue, le krigeage intrinsèque permet de considérer une classe de covariances plus vaste (notons, par exemple, que les splines de type « plaques minces » (Wahba, 1990) sont fondées sur des noyaux conditionnellement positifs). Le krigeage intrinsèque permet également d'inclure de l'information a priori dont on ne souhaite pas qu'elle soit régularisée (l'idée étant alors d'inclure des fonctions particulières en plus des monômes considérés plus haut Vazquez et Walter (2005)). Retenons que la théorie du krigeage, avec le point de vue des processus aléatoires, offre de larges possibilités pour construire des modèles de type boîte noire dans des situations variées.

1.5 Choix d'un noyau

Le choix d'un noyau pour une application donnée est un problème délicat, plus ou moins difficile selon la dimension de l'espace des facteurs et les propriétés du système étudié. Il est bien sûr possible de choisir un noyau a priori et d'espérer des résultats satisfaisants. Ce procédé peut éventuellement être amélioré par une phase de validation croisée. Cependant, l'analyse des données observées ainsi que les connaissances a priori sur le système peuvent orienter le choix. Le point de vue du krigeage permet comme nous l'avons mentionné de considérer un noyau comme une covariance et les méthodes d'estimation statistique deviennent alors applicables. Les géostatisticiens considèrent que la phase d'analyse des données (qu'ils appellent *analyse structurelle*) est beaucoup plus importante et délicate que la phase de prédiction proprement dite. Cette analyse porte d'abord sur la nature des observations : type des facteurs (valeurs numériques, catégories), présence de bruit dans les observations, incertitudes sur les facteurs, informations disponibles a priori, etc. Le choix reflète ensuite la structure de la corrélation : nature des corrélations courte et longue portée, phénomènes non stationnaires, non gaussiens (méthodes de prédiction spécifiques à employer), etc.

Choisir une covariance comporte en fait deux aspects : le choix d'une famille paramétrée de fonctions de covariance d'une part, en favorisant si possible les covariances ayant un nombre faible de paramètres, et l'estimation des paramètres de cette famille d'autre part.

1.5.1 Classes de fonctions de covariances

Traditionnellement, une covariance est choisie sous la forme d'une fonction paramétrée, mais cette forme est souvent quelque peu arbitraire. Le choix peut toutefois être orienté par les connaissances a priori ou l'analyse des données. L'hypothèse la plus classique est de supposer la covariance invariante par translation ($\mathrm{Cov}[F(\boldsymbol{x}), F(\boldsymbol{y})] = k(\boldsymbol{x} - \boldsymbol{y})$) ce qui implique un modèle stationnaire. Notons cependant qu'il n'est généralement pas possible en pratique de tester la stationnarité à partir d'un ensemble fini de données, observées sur un horizon borné. On peut tout au plus tester la présence de certaines formes de non stationnarité, comme par exemple la présence de tendances polynomiales dans les données. Mais, même dans ce cas, des ambiguïtés demeurent car on peut interpréter une tendance comme moyenne du processus aléatoire ou bien comme des variations dues à des corrélations à longue portée par rapport à l'horizon d'observation. Les modèles stationnaires permettent surtout de simplifier la mise en œuvre. Si les données présentent des tendances

(polynomiales ou autres), on considérera des accroissements du processus pour filtrer ces tendances, et on utilisera des covariances généralisées stationnaires. Une autre hypothèse classique est l'isotropie des données, et dans ce cas la covariance dépend seulement de la distance euclidienne entre les facteurs. Ceci n'est pas du tout requis et peut s'avérer maladroit, notamment quand les facteurs sont de natures diverses. On choisit alors souvent une covariance du type $k(h)$ avec $h = \sqrt{(\boldsymbol{x}-\boldsymbol{y})^\mathsf{T}\boldsymbol{A}(\boldsymbol{x}-\boldsymbol{y})}$, où $\boldsymbol{A}$ est une matrice définie positive à choisir.

Une fonction de covariance étant de type positif, on la choisira parmi les familles classiques, données par exemple dans (Yaglom, 1986), (Stein, 1999) ou la littérature géostatistique. Les covariances classiques sont par exemple les covariances gaussiennes ($k(h) = \sigma^2 e^{-1/2(h/\rho)^2}$), exponentielles ($k(h) = \sigma^2 e^{-(|h|/\rho)^\nu}$), cubiques, etc. Il est possible de régler deux comportements principalement. Le premier est la *portée de la corrélation entre les données* (réglée ci-dessus par ρ). Le deuxième comportement est la *régularité de la covariance*. Comme (Stein, 1999) le montre dans un cadre asymptotique, si l'on considère deux covariances et que l'on utilise l'une pour prédire un processus ayant en réalité l'autre covariance, on obtient les mêmes variances d'erreur que si on avait utilisé cette dernière, à condition que les covariances aient la même régularité. La régularité d'une covariance correspond par exemple à sa dérivabilité à l'origine. Elle correspond aussi à la décroissance de la densité spectrale du processus aléatoire en hautes fréquences. (Stein, 1999) préconise l'utilisation de la fonction de Matérn qui est caractérisée par trois paramètres servant à ajuster indépendamment la variance, la portée et la régularité. La densité spectrale $g(u)$ correspondante est du type $g(u) = \gamma(1 + (u/u_0)^2)^{-\nu-1/2}$, où le réel positif ν permet de régler la décroissance à l'infini. La fonction de Matérn peut s'écrire

$$k(h) = \frac{\sigma}{2^{\nu-1}\Gamma(\nu)} \left(\frac{2\nu^{1/2}h}{\rho}\right)^\nu \mathcal{K}_\nu \left(\frac{2\nu^{1/2}h}{\rho}\right), \tag{1.16}$$

où $\mathcal{K}_\nu$ est la fonction de Bessel modifiée de seconde espèce. Numériquement, il faut éviter des trop grandes valeurs de ν (typiquement, on choisit $\nu < 5$).

1.5.2 Estimation des paramètres

En géostatistique, cette étape consiste traditionnellement à estimer de manière empirique la covariance du processus à partir des données et à ajuster ensuite le vecteur $\boldsymbol{\theta}$ des paramètres du modèle de covariance pour se rapprocher de cette covariance empirique. Estimer la covariance à partir des données est en fait une étape importante car elle permet aussi de vérifier dans une certaine mesure les hypothèses sur $F(\boldsymbol{x})$. Pour cette estimation, les géostatisticiens utilisent un *variogramme* qui correspond à la variance des différences $F(\boldsymbol{x}_i) - F(\boldsymbol{x}_j)$ en fonction de la distance entre les points $\boldsymbol{x}_i$ et $\boldsymbol{x}_j$ dans l'espace des facteurs. La variance des différences est généralement petite pour des distances faibles et augmente avec la distance.

Si le processus aléatoire est gaussien et le nombre des données modeste, la méthode du maximum de vraisemblance nous semble préférable. Rappelons que la log–vraisemblance d'une réalisation $\boldsymbol{f}^{\text{obs}} \in \mathbb{R}^n$ d'un vecteur aléatoire gaussien de moyenne nulle $\boldsymbol{F}^{\text{obs}} = (F(\boldsymbol{x}_1) \cdots F(\boldsymbol{x}_n))^\mathsf{T}$ est donnée par

$$L(\boldsymbol{f}^{\text{obs}}, \boldsymbol{\theta}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log\det \boldsymbol{K}(\boldsymbol{\theta}) - \frac{1}{2}\boldsymbol{f}^{\text{obs}\mathsf{T}}\boldsymbol{K}(\boldsymbol{\theta})^{-1}\boldsymbol{f}^{\text{obs}}, \tag{1.17}$$

où la matrice de covariance dépend de $\boldsymbol{\theta}$. La maximisation de la log–vraisemblance (1.17) peut être conduite avec des méthodes de type gradient conjugué, par exemple. Pour des vecteurs aléatoires gaussiens de moyenne inconnue, il est possible d'utiliser la méthode du maximum de vraisemblance restreint (*REML*).

1.6 Conclusions

Ce chapitre a rapidement présenté les fondements de la régression régularisée, en accordant une place particulière au krigeage, qui nous semble offrir un cadre adapté pour le choix du noyau. Les méthodes à noyau, parfois appelées méthodes non paramétriques, peuvent être obtenues à partir des méthodes dites paramétriques plus répandues. Les chapitres 2 et 3 reviennent plus en détail sur ces notions. Soulignons que les modèles boîte noire obtenus sont particulièrement simples à obtenir et à utiliser. Les applications potentielles de ces méthodes de régression sont innombrables. Le chapitre 6 illustrera leur mise en œuvre dans des domaines très variés pour des dimensions de l'espace des facteurs allant jusqu'à plusieurs dizaines.

Chapitre 2

Prédiction linéaire des processus aléatoires stationnaires à moyenne nulle

Résumé — Ce chapitre rappelle des notions sur les processus aléatoires qui sont très classiques mais nécessaires pour la mise en place de modèles boîte noire par krigeage. Le cadre formel adopté permet d'évoquer les propriétés des trajectoires d'un processus aléatoire et de soulever les problèmes liés à la continuité des facteurs. Il présente le principe de prédiction linéaire et les équations du krigeage pour les processus aléatoires à moyenne nulle ou connue.

Les modèles par processus aléatoires constituent un outil fondamental pour traiter des *systèmes incertains* de type entrées–sorties statiques. Dans le cadre de la modélisation de type boîte noire, un système peut être incertain, par exemple parce qu'il n'est pas possible d'expliciter analytiquement les relations entre entrées et sorties du système, ou parce que le processus a été observé partiellement ou encore parce que les observations sont entachées d'erreurs de mesure. Un processus aléatoire est un *modèle* du système étudié. La sortie du système réel correspond à *une* trajectoire particulière (inconnue) du processus aléatoire. L'utilisation de modèles par processus aléatoires permet de quantifier l'incertitude sur les trajectoires du système et d'effectuer des prédictions de nature probabiliste sur celles-ci.

L'objectif de la section 2.1 est de définir les termes utilisés ci-dessus et de rappeler de manière formelle la construction d'un processus aléatoire, telle qu'on la trouve dans de nombreux ouvrages mais dans des contextes différents de celui de la modélisation boîte noire de systèmes. La section 2.2 est consacrée aux processus gaussiens qui seront les modèles utilisés par la suite. Ces brefs rappels nous semblent justifiés, car parler de processus à paramètres continus nécessite quelques précautions. La section 2.3 présente des notions plus avancées qui seront utilisées au chapitre 3. Dans ces trois premières sections, il s'agit surtout de donner les concepts fondamentaux qui se-

ront nécessaires par exemple lorsque nous nous intéresserons à la prédiction linéaire et au choix de modèle. Les sections 2.4 et 2.5 forment la partie essentielle de ce chapitre, où les propriétés importantes des modèles boîte noire par krigeage sont présentées.

2.1 Notions élémentaires sur les processus aléatoires

Cette section rappelle quelques notions très classiques sur les processus aléatoires. L'un des points de vue classiques est celui des processus aléatoires indexés par le temps. Dans ce cas, les questions portent naturellement sur les propriétés spectrales, sur la prédiction des phénomènes, etc. Lorsque l'on s'intéresse à des phénomènes à temps discret, on parle de *séries chronologiques*. Une présentation très complète de la théorie des séries chronologiques est donnée par exemple dans (Brockwell et Davis, 1987). Les problèmes de prédiction nécessitant un traitement à temps continu sont moins fréquents en pratique mais l'utilisation de processus aléatoires à temps continu reste indispensable pour la modélisation de nombreux phénomènes. Il existe bien sûr une quantité très importante de publications traitant des processus aléatoires à temps continu. Pour donner quelques repères, citons (Doob, 1953 ; Cramér et Leadbetter, 1967 ; Gikhman et Skorohod, 1974) qui peuvent être lus comme premières introductions. Les cas particuliers des processus aléatoires régis par des équations différentielles stochastiques, ou des processus *ARMA* à temps continu, appelés *CARMA* (voir par exemple (Durbin, 1961)), constituent des domaines très étudiés. Des travaux sur les processus *CARMA* comme (Jones, 1981 ; Jones et Vecchia, 1993) sont utiles pour le problème du choix de modèle, qui sera abordé au chapitre 5. Notons que les propriétés théoriques de tels processus à temps continu sont aussi plus difficiles à étudier et plus subtiles que dans le cas discret. Dans le cas de la modélisation de systèmes, nous considérerons des processus aléatoires ayant comme paramètres les facteurs du modèle. On aura donc essentiellement besoin de traiter des processus à *plusieurs paramètres continus*. La plupart des concepts et des résultats valables dans le cadre temporel restent vrais dans des cas plus généraux. Les aspects probabilistes des processus du second ordre sont traités par de nombreux auteurs. Par exemple, (Doob, 1953 ; Rozanov, 1967 ; Cramér et Leadbetter, 1967 ; Yaglom, 1973 ; Billingsley, 1995) constituent des références importantes. (Ibragimov et Rozanov, 1978) aborde en détails la théorie des processus gaussiens. Il peut également être intéressant de consulter (Adler, 1981) pour les questions relatives à la nature des trajectoires. Les ouvrages de géostatistique comme (Cressie, 1993 ; Wackernagel, 1995 ; Chilès et Delfiner, 1999) sont aussi des références très importantes pour cette étude. En géostatistique, les processus aléatoires sont indexés sur un espace physique de type géographique, typiquement un espace continu de dimension deux ou trois. Les problèmes que nous traiterons dans le cadre de la modélisation de systèmes sont de même nature que ceux rencontrés en géostatistique, mais avec des espaces de facteurs de dimension parfois bien supérieure. Nous souhaitons insister sur l'intérêt des ouvrages de G. Matheron, publiés par l'École des Mines de Paris, qui gardent toute leur pertinence dans ce contexte élargi.

2.1.1 Définitions élémentaires, construction

Soit un espace de probabilité $(\Omega, \mathcal{A}, \mathsf{P})$. Il s'agit d'un ensemble d'éléments $\omega \in \Omega$, sur lequel on définit une σ-algèbre (ou tribu) $\mathcal{A}$, que l'on interprète comme un ensemble d'événements. Cette

σ-algèbre est constituée d'ensembles mesurables $A \subseteq \Omega$, et l'on munit $(\Omega, \mathcal{A})$ d'une mesure de probabilité P. Les ensembles mesurables de Ω sont les ensembles auxquels on peut assigner une probabilité (voir par exemple (Rudin, 1987) pour une présentation de la théorie de la mesure).

Les fonctions mesurables $X(\omega) : \Omega \to \mathbb{R}$ sont appelées variables aléatoires. Rappelons qu'une fonction X de Ω vers $\mathbb{R}$ est mesurable si pour tout ouvert B de $\mathbb{R}$, $X^{-1}(B) \in \mathcal{A}$. La plus petite σ-algèbre générée par les ouverts de $\mathbb{R}$ s'appelle σ-algèbre des boréliens, et se notera $\mathcal{B}(\mathbb{R})$. Nous noterons également $\mathcal{B}(\mathbb{R}^n)$ la σ-algèbre générée par les ouverts de $\mathbb{R}^n$. Soit X une variable aléatoire. On note $\sigma(X)$, et on appelle σ-algèbre générée par X, la plus petite σ-algèbre sur Ω rendant X mesurable. On a donc

$$\sigma(X) = \{X^{-1}(B);\ B \in \mathcal{B}(\mathbb{R})\}\,.$$

L'application

$$\begin{array}{rccl} \mathsf{P}_X : & \mathcal{B}(\mathbb{R}) & \to & [0,1] \\ & B & \mapsto & \mathsf{P}_X(B) = \mathsf{P}\{X \in B\} \end{array} \tag{2.1}$$

est une probabilité sur $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Cette application est la loi de probabilité de X.

Définition 1. Un ensemble de variables aléatoires $F(\omega, \boldsymbol{x})$, lorsque $\boldsymbol{x}$ parcourt l'ensemble $\mathbb{X}$, constitue une *fonction aléatoire* de paramètre $\boldsymbol{x}$.

En pratique, l'espace des paramètres $\mathbb{X}$, que nous appellerons *espace des facteurs* dans le cadre de la modélisation boîte noire[1], est un espace $\mathbb{R}^d$ tout entier. L'espace d'étude est en général plus petit, puisqu'il s'agit quasiment toujours d'un domaine borné de $\mathbb{R}^d$, souvent un pavé borné fermé. Une telle fonction aléatoire sur un domaine de $\mathbb{R}^d$ s'appelle encore *processus aléatoire*, ou *champ aléatoire*, pour insister sur le fait que la dimension de l'espace des paramètres est supérieure à un. Les valeurs de cette fonction aléatoire sont donc les variables aléatoires notées $F(\boldsymbol{x})$. Nous écrirons aussi $F(\boldsymbol{x})$ pour désigner la fonction aléatoire $(F(\omega, \boldsymbol{x}), \boldsymbol{x} \in \mathbb{X})$. Si $\omega \in \Omega$ est fixé, la fonction réelle $F(\omega, .)$ de $\boldsymbol{x} \in \mathbb{X}$ est appelée *réalisation* de la fonction aléatoire, ou encore *trajectoire*.

Il est donc naturel de considérer un espace $\mathcal{F}$ de fonctions réelles sur $\mathbb{X}$ qui contient toutes les trajectoires possibles de la fonction aléatoire $F(\boldsymbol{x})$. Un tel espace peut correspondre simplement à l'ensemble de toutes les fonctions réelles $\mathbb{R}^{\mathbb{X}}$. On se fixe l'objectif de construire une loi de probabilité pour $F(\boldsymbol{x})$. Ceci passe par le choix des ensembles mesurables sur $\mathbb{R}^{\mathbb{X}}$.

Définition 2. Soient les ensembles de fonctions $U_B \subset \mathbb{R}^{\mathbb{X}}$ qui sont de la forme

$$\{f \in \mathbb{R}^{\mathbb{X}};\ (f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_n)) \in B\}\,,$$

où $B \in \mathcal{B}(\mathbb{R}^n)$, $n \geq 1$. Ces ensembles élémentaires U_B sont appelés *ensembles cylindriques* et génèrent une σ-algèbre sur $\mathbb{R}^{\mathbb{X}}$, noté $\mathcal{C}(\mathbb{R}^{\mathbb{X}})$.

Proposition 1. *Avec ce choix d'ensembles mesurables de $\mathbb{R}^{\mathbb{X}}$, l'application de $(\Omega, \mathcal{A}, \mathsf{P})$ dans $(\mathbb{R}^{\mathbb{X}}, \mathcal{C}(\mathbb{R}^{\mathbb{X}}))$, qui à tout $\omega \in \Omega$ associe une trajectoire $F(\omega, .) \in \mathbb{R}^{\mathbb{X}}$, est mesurable.*

[1] La sortie d'un modèle dépend typiquement de deux types de paramètres. Les premiers permettent un ajustement du modèle car nous considérons en fait une famille de modèles. En ce sens, nous parlons de *modèle paramétré*. Les paramètres du second type sont les *entrées* du modèle, qui caractérisent l'état d'opération du système modélisé. Nous appelons ces paramètres *facteurs* pour les distinguer des premiers.

Démonstration. Soit $U_B \in \mathcal{C}(\mathbb{R}^{\mathbb{X}})$. L'image réciproque de U_B par $F(\omega, \boldsymbol{x})$,

$$\{\omega \in \Omega;\ F(\omega, .) \in U_B\} = \{\omega \in \Omega;\ (F(\omega, \boldsymbol{x}_1), \ldots, F(\omega, \boldsymbol{x}_n)) \in B\}\,, \quad B \in \mathcal{B}(\mathbb{R}^n)\,,$$

est bien un ensemble de $\mathcal{A}$ (d'après, par exemple, le théorème 1.8 de (Rudin, 1987)). □

Définition 3. La probabilité sur l'espace mesurable $(\mathbb{R}^{\mathbb{X}}, \mathcal{C}(\mathbb{R}^{\mathbb{X}}))$ définie par

$$\mathsf{P}_F(U) = \mathsf{P}\{F(\omega, .) \in U\}, \quad U \in \mathcal{C}(\mathbb{R}^{\mathbb{X}}) \tag{2.2}$$

est appelée *loi de probabilité* de la fonction aléatoire $F(\boldsymbol{x})$. (Les ensembles $\{F(\omega, .) \in U\}$ contiennent les éléments $\omega \in \Omega$ tels que $(F(\omega, \boldsymbol{x}_1), \ldots, F(\omega, \boldsymbol{x}_n)) \in B$, où $B \in \mathcal{B}(\mathbb{R}^n)$.)

En résumé, à une collection de fonctions mesurables sur $(\Omega, \mathcal{A}, \mathsf{P})$ paramétrées par $\boldsymbol{x} \in \mathbb{X}$, on peut associer une loi de probabilité P_F sur l'espace mesurable de fonctions $(\mathbb{R}^{\mathbb{X}}, \mathcal{C}(\mathbb{R}^{\mathbb{X}}))$.

Remarque. Les problèmes que l'on se pose en pratique pour la modélisation boîte noire de systèmes font généralement intervenir un nombre au plus dénombrable et le plus souvent fini de coordonnées de fonctions de $\mathbb{R}^{\mathbb{X}}$, et donc une suite finie ou dénombrable de variables aléatoires $F(\boldsymbol{x}_i)$. Ainsi, la plupart des questions sur le modèle (notamment les questions de prédiction) peuvent recevoir une réponse probabiliste. En effet, avec le choix des ensembles mesurables effectué ci-dessus, les fonctions mesurables d'une variable aléatoire $F(\boldsymbol{x})$, voire d'un ensemble dénombrable de variables aléatoires $F(\boldsymbol{x}_i)$ sont des fonctions mesurables. Cependant, l'examen de propriétés faisant intervenir un ensemble infini non dénombrable de coordonnées de $\mathbb{R}^{\mathbb{X}}$ pose problème car une réunion ou une intersection non dénombrable d'ensembles mesurables n'est pas mesurable. De telles propriétés ne peuvent donc pas recevoir de réponse probabiliste (ou de manière équivalente, être mesurées). On pourrait penser que ces problèmes n'ont pas d'intérêt pratique. Pourtant, les propriétés de continuité des trajectoires, de dérivabilité, etc., font intervenir un ensemble non dénombrable de variables aléatoires. Il semble donc gênant de vouloir modéliser un système en utilisant un concept qui ne permet pas de répondre à des questions aussi simples que celle de la continuité. Ces points seront revus plus bas.

Si l'on se donne une fonction $F(\omega, \boldsymbol{x})$ à valeurs réelles et une mesure de probabilité P_F sur $(\mathbb{R}^{\mathbb{X}}, \mathcal{C}(\mathbb{R}^{\mathbb{X}}))$, on peut définir une σ-algèbre sur Ω ainsi qu'une mesure de probabilité. La notation $\sigma(F(\boldsymbol{x}))$ introduite plus haut pour une variable aléatoire $F(\boldsymbol{x})$, peut être étendue à un ensemble fini de variables aléatoires de telle sorte que $\sigma(F(\boldsymbol{x}_1), \ldots, F(\boldsymbol{x}_n))$ désigne la σ-algèbre générée par $F(\boldsymbol{x}_1), \ldots, F(\boldsymbol{x}_n)$. Plus généralement, la σ-algèbre sur Ω définie par $\sigma(F(\mathbb{X})) = \{F^{-1}(U);\, U \in \mathcal{C}(\mathbb{R}^{\mathbb{X}})\}$ est dite générée par $F(\omega, .)$. Il s'agit de la plus petite σ-algèbre contenant les ensembles de Ω de la forme $\{(F(\boldsymbol{x}_1), \ldots, F(\boldsymbol{x}_n)) \in B\}$, où $B \in \mathcal{B}(\mathbb{R}^n)$. La relation $\mathsf{P}\{F(\omega, .) \in U\} = \mathsf{P}_F(U)$, $U \in \mathcal{C}(\mathbb{R}^{\mathbb{X}})$, définit bien une probabilité sur $\sigma(F(\mathbb{X}))$, et est telle que $F(\boldsymbol{x})$ soit une fonction aléatoire sur l'espace $(\Omega, \sigma(F(\mathbb{X})), \mathsf{P})$ admettant la loi de probabilité P_F.

En pratique, il est fondamental de pouvoir définir la loi d'un processus aléatoire en spécifiant une famille de probabilités sur des ensembles mesurables définis en considérant seulement un nombre fini de variables aléatoires $F(\boldsymbol{x}_i)$, $i = 1, \ldots, n$.

Définition 4. Les *répartitions finies* $\mathsf{P}_{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n}$ d'un processus aléatoire $F(\boldsymbol{x})$ sont les probabilités sur les espaces vectoriels $\mathbb{R}^n$ telles que

$$\mathsf{P}_{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n}(B) = \mathsf{P}\{(F(\boldsymbol{x}_1), \ldots, F(\boldsymbol{x}_n)) \in B\} \quad B \in \mathcal{B}(\mathbb{R}^n)$$

Proposition 2. *Si deux processus aléatoires ont les mêmes répartitions finies, ils ont les mêmes lois de probabilité (on dit qu'ils sont* équivalents*).*

Démonstration. Les répartitions finies déterminent les probabilités sur les ensembles cylindriques de manière unique. Ce résultat provient du fait que les ensembles cylindriques forment une algèbre de $\mathbb{R}^{\mathbb{X}}$ (ensembles stables par passage au complémentaire et par union et intersection finie) et de l'utilisation du théorème de Carathéodory : soit μ une mesure sur une algèbre $\mathcal{A}$, alors μ se prolonge en une mesure sur $\sigma(\mathcal{A})$. Si de plus μ est σ-finie, ce prolongement est unique (voir par exemple (Billingsley, 1995), section 3). □

Les répartitions finies satisfont une notion de cohérence puisque pour $m < n$, $B \in \mathcal{B}(\mathbb{R}^m)$,

$$\mathsf{P}_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n}(B \times \mathbb{R}^{n-m}) = \mathsf{P}_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_m}(B)$$

Théorème 1 (d'extension de Kolmogorov). *Étant donnée une famille de répartitions finies cohérentes, $\mathsf{P}_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n}$, sur les espaces de dimension finie $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$, il existe une mesure de probabilité P sur $(\Omega, \sigma(F(\mathbb{X})))$ unique satisfaisant*

$$\mathsf{P}_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n}(B) = \mathsf{P}\{(F(\boldsymbol{x}_1), \ldots, F(\boldsymbol{x}_n)) \in B\}$$

Démonstration. La preuve (délicate) de ce théorème se trouve dans certains ouvrages de probabilité (Tucker, 1967 ; Billingsley, 1995). Le résultat est parfois attribué à (Daniell, 1919). □

Ce théorème fondamental indique que la loi d'un processus aléatoire est entièrement caractérisée par l'ensemble des lois des vecteurs aléatoires extraits de taille finie, ce qui permet de construire des processus aléatoires à partir d'un choix de répartitions de dimension finie cohérentes.

Cependant, comme indiqué plus haut, la donnée de répartitions finies ne caractérise pas entièrement les propriétés du processus aléatoire. En effet, l'intersection d'un nombre fini ou infini dénombrable d'ensembles mesurables est un ensemble mesurable, de sorte qu'une grandeur qui dépend d'un nombre fini ou même infini dénombrable de variables aléatoires est en général une fonction mesurable. Dans le cas d'un processus aléatoire à paramètres continus, qui est donc une famille non dénombrable de variables aléatoires, les répartitions finies ne contiennent donc pas toutes les caractéristiques sur le processus. Par exemple les questions portant sur la continuité des trajectoires, la dérivabilité du processus, ou des bornes, ne peuvent pas être caractérisées par des fonctions mesurables.

Exemple (Doob, 1953). Soit $\Omega = [0, 1]$, muni de la σ-algèbre des boréliens de $[0, 1]$ et de la mesure de Lebesgue. Soient les processus aléatoires $F_1(\omega, x)$ et $F_2(\omega, x)$, $x \in \mathbb{R}$ tels que

$$F_1(\omega, x) = 0\,, \quad \forall \omega\,, \forall x$$

$$F_2(\omega, x) = \begin{cases} 1 & \text{si } x = \omega, \\ 0 & \text{sinon.} \end{cases}$$

Remarquons d'abord que les répartitions finies des deux processus sont égales puisque $\mathsf{P}\{F_{1/2}(\omega, x) = 0\} = 1, \forall x \in \mathbb{R}$. Il s'agit donc de deux processus équivalents. On a même $\mathsf{P}\{F_1(\omega, x) = F_2(\omega, x)\} = 1, \forall x \in \mathbb{R}$ et $F_2(x)$ est donc une modification de $F_1(x)$. Mais

$$\mathsf{P}\{x \mapsto F_1(\omega, x) \text{ est continue sur } \mathbb{R}\} = 1$$

$$\mathsf{P}\{x \mapsto F_2(\omega, x) \text{ est continue sur } \mathbb{R}\} = 0$$

De même

$$\mathsf{P}\{\sup_{x\in[0,1]} F_1(\omega, x) < \frac{1}{2}\} = 1$$

$$\mathsf{P}\{\sup_{x\in[0,1]} F_2(\omega, x) < \frac{1}{2}\} = 0$$

Le comportement que l'on voudrait privilégier est celui auquel on s'attend plus naturellement, c'est-à-dire celui du processus F_1. Cet exemple illustre le fait qu'un modèle aléatoire défini à partir de répartitions finies est insuffisant pour répondre à des questions du type : est-ce que le système dépasse un certain seuil ? Considérer un processus aléatoire à paramètres continus est donc délicat.

Il nous semble donc pertinent de rappeler rapidement dans cette présentation la notion de séparabilité introduite par (Doob, 1953), qui permet d'exclure certains cas pathologiques en requérant un comportement régulier des processus aléatoires considérés. On souhaite notamment empêcher qu'un processus ait des propriétés différentes selon que l'on considère des ensembles dénombrable ou non dénombrable de coordonnées. En effet, l'hypothèse de séparabilité permet de déterminer les propriétés des trajectoires du processus aléatoires à partir d'un ensemble éventuellement dense mais au plus dénombrable de l'espace des facteurs. La définition de la séparabilité permet notamment d'exclure le processus aléatoire F_2 défini plus haut.

Définition 5 (séparabilité). Soit $\mathcal{I}_F$ la classe des intervalles fermés de $\mathbb{R}$. Un processus aléatoire $F(\omega, \boldsymbol{x})$ est dit *séparable* (sous-entendu, relativement à la classe des intervalles fermés), s'il existe une suite $(\boldsymbol{x}_j)$ de paramètres de $\mathbb{X}$, telle que pour tout intervalle $I \in \mathcal{I}_F$ et tout ouvert O de $\mathbb{X}$, les ensembles de Ω

$$\{F(\omega, \boldsymbol{x}) \in I, \forall \boldsymbol{x} \in O\}, \quad \{F(\omega, \boldsymbol{x}_j) \in I, \forall \boldsymbol{x}_j \in O\}$$

différent au plus d'un ensemble négligeable $N \subset \Omega$ (N est de mesure nulle car on suppose la mesure de probabilité complète).

Une manière équivalente de caractériser un processus aléatoire séparable est de dire qu'il existe une suite au plus dénombrable $(\boldsymbol{x}_j)$ d'éléments de $\mathbb{X}$ et un ensemble $N \subset \Omega$ négligeable tels que si $\omega \notin N$

$$\inf_{\boldsymbol{x}\in O} F(\omega, \boldsymbol{x}) = \inf_{\boldsymbol{x}_j \in O} F(\omega, \boldsymbol{x}_j)$$

$$\sup_{\boldsymbol{x}\in O} F(\omega, \boldsymbol{x}) = \sup_{\boldsymbol{x}_j \in O} F(\omega, \boldsymbol{x}_j)$$

pour tout ouvert $O \subset \mathbb{X}$. Si $F(\boldsymbol{x})$ est séparable à l'aide d'une suite $(\boldsymbol{x}_j)$, il est séparable à l'aide de toute séquence plus grande.

Ainsi, dans l'exemple introduit ci-dessus, $F_2(x)$ n'est pas séparable puisque

$$\mathsf{P}\{0 \le F_2(x) \le \frac{1}{2}, \quad \forall x \in]0, 1[\,\} = 0$$

alors que

$$\mathsf{P}\{0 \le F_2(x) \le \frac{1}{2}, \quad \forall x \in]0, 1[\cap \mathcal{J}\,\} = 1 - \mathsf{P}(\mathcal{J}) = 1\,,$$

où $\mathcal{J}$ est tout ensemble au plus dénombrable d'éléments de $\Omega = [0,1]$. Plus généralement, si F est un processus aléatoire tel que $\mathsf{P}\{F(\boldsymbol{x}) = 0\} = 1$, $\forall \boldsymbol{x} \in \mathbb{X}$, alors, pour toute suite $(\boldsymbol{x}_j)$, $\mathsf{P}\{F(\boldsymbol{x}_j) = 0,\ j \geq 1\} = 1$. Si le processus est de plus séparable, alors on a $\mathsf{P}\{F(\boldsymbol{x}) = 0,\ \boldsymbol{x} \in\ \mathbb{X}\} = 1$.

Si $F(\boldsymbol{x})$ est séparable et si B est un ouvert de $\mathbb{X}$, $\inf_{\boldsymbol{x}\in B} F(\boldsymbol{x})$, $\sup_{\boldsymbol{x}\in B} F(\boldsymbol{x})$, $\liminf_{\boldsymbol{x}\to\boldsymbol{y}} F(\boldsymbol{x})$ et $\limsup_{\boldsymbol{x}\to\boldsymbol{y}} F(\boldsymbol{x})$ sont des variables aléatoires. Une autre propriété importante est que si $F(\boldsymbol{x})$ est séparable à l'aide d'une suite $(\boldsymbol{x}_j)$ dense, on peut trouver deux sous-suites $\boldsymbol{y}_1 < \boldsymbol{y}_2 < \cdots < \boldsymbol{y}_n \uparrow \boldsymbol{x}$ et $\boldsymbol{y}'_1 > \boldsymbol{y}'_2 > \cdots > \boldsymbol{y}'_n \downarrow \boldsymbol{x}$ telles que, avec probabilité 1,

$$\lim_{n\to\infty} \inf F(\boldsymbol{y}_n) = \lim_{\boldsymbol{y}\uparrow\boldsymbol{x}} \inf F(\boldsymbol{y}) \leq \lim_{\boldsymbol{y}\downarrow\boldsymbol{x}} \inf F(\boldsymbol{y}) = \lim_{n\to\infty} \inf F(\boldsymbol{y}'_n)\,.$$

Par suite, si $F(\boldsymbol{x})$ est presque sûrement continu (voir section 2.4.4) sur la suite $(\boldsymbol{x}_j)$, il est presque sûrement continu sur $\mathbb{X}$.

Le théorème d'existence suivant est très important, car il permet de s'assurer qu'à partir de répartitions finies cohérentes quelconques, il est toujours possible de construire un processus aléatoire séparable.

Théorème 2 (Existence de processus séparables). *Soit $F(\omega, \boldsymbol{x})$, avec $\omega \in \Omega$, $\boldsymbol{x} \in \mathbb{X}$, où $\mathbb{X}$ est un ensemble de paramètres continu.* Il existe un processus aléatoire séparable $\tilde{F}(\boldsymbol{x})$, *défini sur le même espace Ω, et tel que*

$$\mathsf{P}\{\tilde{F}(\boldsymbol{x}) = F(\boldsymbol{x})\} = 1\,, \quad \forall \boldsymbol{x} \in \mathbb{X}\,.$$

(On dit que $\tilde{F}(\boldsymbol{x})$ est une modification *de $F(\boldsymbol{x})$.)*

Démonstration. Voir (Doob, 1953). □

Si $\tilde{F}(\boldsymbol{x})$ est une modification de $F(\boldsymbol{x})$, les répartitions finies de $F(\boldsymbol{x})$ et $\tilde{F}(\boldsymbol{x})$ sont identiques et $F(\boldsymbol{x})$ et $\tilde{F}(\boldsymbol{x})$ sont donc aussi des processus équivalents d'après la proposition 2. En d'autres termes, pour tout processus aléatoire $F(\boldsymbol{x})$, on peut trouver un processus équivalent $\tilde{F}(\boldsymbol{x})$ séparable. Dans la suite, les processus aléatoires seront toujours implicitement supposés séparables.

En résumé, le théorème d'extension de Kolmogorov garantit l'existence d'un processus aléatoire à partir de la donnée de répartitions finies cohérentes et le théorème d'existence de processus séparables nous dit que l'on peut trouver des processus aléatoires avec un comportement raisonnable ; plus précisément, que l'on peut trouver un processus aléatoire pour lequel il existe un ensemble dénombrable de coordonnées du processus qui caractérise le processus sur tout son ensemble de définition. Ces deux théorèmes réunis sont fondamentaux pour construire rigoureusement un processus aléatoire servant à la modélisation d'un système.

Nous avons rappelé ci-dessus les éléments permettant de définir un processus aléatoire paramétré par un ensemble de facteurs pour modéliser une sortie particulière d'un système. Le cas d'un système à plusieurs sorties se traite en considérant plusieurs processus aléatoires. Il n'y a pas de difficultés à étendre la définition d'un processus aléatoire à valeurs dans $\mathbb{R}$ à un processus aléatoire à valeurs dans un espace vectoriel $\mathbb{R}^q$. Nous utilisons la notation $F_\alpha(\boldsymbol{x})$, $\alpha \in \{1, \ldots, q\}$ pour désigner un élément du vecteur aléatoire $\boldsymbol{F}(\omega, \boldsymbol{x})$, paramétré par $\boldsymbol{x}$.

Définition 6. Un processus aléatoire *à valeurs vectorielles* $\boldsymbol{F}(\omega, \boldsymbol{x}) = (F_1(\omega, \boldsymbol{x}), \ldots, F_q(\omega, \boldsymbol{x}))^\mathsf{T}$ est une collection de variables aléatoires $F_\alpha(\omega, \boldsymbol{x})$, $\forall \alpha \in \{1, \ldots, q\}$, $\forall \boldsymbol{x} \in \mathbb{X}$

Au vu de cette définition, on se ramène immédiatement au cas d'un processus aléatoire à valeur dans $\mathbb{R}$ en considérant l'indice α comme un paramètre supplémentaire et en posant $F(\alpha, \boldsymbol{x}) = F_\alpha(\boldsymbol{x})$. Nous utiliserons fréquemment ce changement de notation pour modéliser les systèmes à plusieurs sorties.

2.1.2 Moments, fonction de covariance

La *fonction de covariance* d'un processus aléatoire occupera une place importante par la suite. En effet, la résolution des problèmes de prédiction linéaire que nous traiterons est essentiellement fondée sur la connaissance de celle-ci. Pour parler d'un processus aléatoire, nous ne nous intéresserons ainsi généralement qu'à sa moyenne et à sa fonction de covariance, sans préciser les répartitions finies du processus, susceptibles de faire intervenir des moments d'ordres supérieurs. Cette section rappelle les notions élémentaires sur les fonctions de covariances.

Une variable aléatoire X est du *second ordre* si $\mathrm{E}[X^2] = \int_\Omega X(\omega)^2 d\mathsf{P}(\omega)$ est définie. L'espace des variables aléatoires du second ordre est noté usuellement $L^2(\Omega, \mathcal{A}, \mathsf{P})$, puisqu'il s'agit des variables aléatoires de carré intégrable sur Ω pour la mesure P (à une équivalence près). Muni du produit scalaire $(X, Y) = \mathrm{E}[XY]$, $L^2(\Omega, \mathcal{A}, \mathsf{P})$ est un espace de Hilbert. Comme $L^2(\Omega, \mathcal{A}, \mathsf{P}) \subset L^1(\Omega, \mathcal{A}, \mathsf{P})$, $\mathrm{E}[X^2] < \infty$ implique $\mathrm{E}[|X|] < \infty$ et $\mathrm{E}[X]$ est donc bien défini.

Par la suite, les processus considérés seront *toujours* du second ordre, c'est-à-dire $\forall \boldsymbol{x} \in \mathbb{X}$, $F(\boldsymbol{x}) \in L^2(\Omega, \mathcal{A}, \mathsf{P})$. De plus, nous dirons souvent qu'un processus est du second ordre sans préciser l'espace $L^2(\Omega, \mathcal{A}, \mathsf{P})$. Notons $m(\boldsymbol{x}) = \mathrm{E}[F(\boldsymbol{x})]$, la moyenne du processus aléatoire, et $k(\boldsymbol{x}, \boldsymbol{y}) = \mathrm{Cov}(F(\boldsymbol{x}), F(\boldsymbol{y})) = \mathrm{E}[(F(\boldsymbol{x}) - m(\boldsymbol{x}))(F(\boldsymbol{y}) - m(y))] = \mathrm{E}[F(\boldsymbol{x})F(\boldsymbol{y})] - m(\boldsymbol{x})m(\boldsymbol{y})$, la fonction appelée *autocovariance*, ou plus simplement *covariance* du processus aléatoire $F(\boldsymbol{x})$. Pour un processus du second ordre, la moyenne existe, de même que la covariance (d'après l'inégalité de Schwarz). La covariance est symétrique : $k(\boldsymbol{x}, \boldsymbol{y}) = k(\boldsymbol{y}, \boldsymbol{x})$. On définit également la fonction de *corrélation* par

$$\rho(\boldsymbol{x}, \boldsymbol{y}) = \frac{k(\boldsymbol{x}, \boldsymbol{y})}{\sqrt{k(\boldsymbol{x}, \boldsymbol{x}) k(\boldsymbol{y}, \boldsymbol{y})}} .$$

Elle est telle que $|\rho(\boldsymbol{x}, \boldsymbol{y})| \leq 1$.

Lorsque la fonction de covariance est invariante par translation (ce qui est le cas lorsque le processus est stationnaire, voir la section 2.4.1) c'est-à-dire lorsque $k(\boldsymbol{x} + \boldsymbol{\tau}, \boldsymbol{y} + \boldsymbol{\tau}) = k(\boldsymbol{x}, \boldsymbol{y})$, $\forall \boldsymbol{\tau}$, nous utiliserons souvent la notation $k(\boldsymbol{h})$, avec $\boldsymbol{h} = \boldsymbol{x} - \boldsymbol{y}$. Quand la covariance est de plus invariante par rotation (ce qui est le cas lorsque le processus est isotrope), nous noterons aussi $k(\|\boldsymbol{h}\|)$ ou $k(h)$ avec $h = \|\boldsymbol{x} - \boldsymbol{y}\|$.

Quelques propriétés fondamentales des covariances sont rappelées ci-dessous (le chapitre 5 reviendra sur les fonctions de covariances du point de vue du choix de modèle).

Définition 7. Une fonction $r(\boldsymbol{x}, \boldsymbol{y})$ est de *type positif* (*non-négatif*), si

$$\forall n \in \mathbb{N}, \forall \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{X}, \forall \lambda_1, \ldots, \lambda_n \in \mathbb{R}, \quad \sum_{i,j=1}^{n} \lambda_i \lambda_j r(\boldsymbol{x}_i, \boldsymbol{x}_j) \geq 0 .$$

Une fonction de covariance est donc de type positif puisque

$$\mathrm{Var}[\sum_{i=1}^{n} \lambda_i F(\boldsymbol{x}_i)] = \mathrm{E}[(\sum_{i=1}^{n} \lambda_i F(\boldsymbol{x}_i) - \sum_{i=1}^{n} \lambda_i m(\boldsymbol{x}_i))^2] = \sum_{i,j=1}^{n} \lambda_i \lambda_j k(\boldsymbol{x}_i, \boldsymbol{x}_j) \geq 0 .$$

Toute matrice de covariance $\boldsymbol{K}$ d'éléments $k_{i,j} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ est donc symétrique non-négative. Les fonctions de type positif peuvent être caractérisées par le théorème de Bochner (rappelé dans la section 2.4.2). En pratique, les covariances sont choisies dans des familles de fonctions paramétrées admissibles (voir le chapitre 5).

Proposition 3. *Soient m une fonction de $\mathbb{R}^d$ dans $\mathbb{R}$ et k une fonction de $\mathbb{R}^d \times \mathbb{R}^d$ dans $\mathbb{R}$, symétrique et positive. Alors il existe un processus $F(\boldsymbol{x})$ admettant $m(\boldsymbol{x})$ comme moyenne et $k(\boldsymbol{x}, \boldsymbol{y})$ comme covariance.*

Démonstration. Par construction d'un processus gaussien $F(\boldsymbol{x})$ (voir la section 2.2.2) grâce à une application immédiate du théorème de Kolmogorov (Chonavel, 2002). (La loi de probabilité P_F de $F(\boldsymbol{x})$ définie sur l'espace $(\mathbb{R}^{\mathbb{X}}, \mathcal{C}(\mathbb{R}^{\mathbb{X}}))$ est unique.) □

Répétons que la donnée d'une moyenne et d'une fonction de covariance ne définit pas de manière unique un processus aléatoire. Voir un exemple dans la section 2.1.3.

Lorsque l'on considère plusieurs processus aléatoires $F_1(\boldsymbol{x}), \ldots, F_q(\boldsymbol{x})$ définis sur un même espace de probabilité et un même espace de paramètres, on considère les moyennes $m_\alpha(\boldsymbol{x}) = \mathrm{E}[F_\alpha(\boldsymbol{x})]$, ainsi que les covariances et inter-covariances définies par $k_{\alpha,\beta}(\boldsymbol{x}, \boldsymbol{y}) = \mathrm{E}[(F_\alpha(\boldsymbol{x}) - m_\alpha(\boldsymbol{x}))(F_\beta(\boldsymbol{y}) - m_\beta(y))]$, $\alpha, \beta \in \{1, \ldots, q\}$.

Pour les fonctions d'inter-covariance, remarquons que $k_{\alpha,\beta}(\boldsymbol{x}, \boldsymbol{y}) = k_{\beta,\alpha}(\boldsymbol{y}, \boldsymbol{x})$. En revanche, on a généralement $k_{\alpha,\beta}(\boldsymbol{x}, \boldsymbol{y}) \neq k_{\alpha,\beta}(\boldsymbol{y}, \boldsymbol{x})$.

Exemple 1. Soit $k(\boldsymbol{h})$, $\boldsymbol{h} \in \mathbb{R}^d$, une covariance admissible (voir la section 2.4.2 et le chapitre 5) ; le modèle dit *proportionnel* est donné par

$$\begin{aligned} k_{1,1}(\boldsymbol{h}) &= k_{2,2}(\boldsymbol{h}) = k(\boldsymbol{h}) = k(-\boldsymbol{h}), \quad \boldsymbol{h} \in \mathbb{R}^d, \\ k_{1,2}(\boldsymbol{h}) &= k_{2,1}(\boldsymbol{h}) = \gamma k(\boldsymbol{h}), \end{aligned} \tag{2.3}$$

où le scalaire γ doit être choisi dans $[-1, 1]$ pour satisfaire l'inégalité de Schwarz

$$|k_{\alpha,\beta}(\boldsymbol{h})| \leq \sqrt{k_{\alpha,\alpha}(0) k_{\beta,\beta}(0)}. \tag{2.4}$$

Exemple 2. Soient les processus $F_1(x)$, $F_2(x)$ et $F_3(x)$ définis sur $\mathbb{X} = \mathbb{R}$ par

$$F_2(x) = F_1(x - \tau) + F_3(x),$$

où $F_1(x)$ et $F_3(x)$ ne sont pas corrélés et où $\mathrm{Cov}[F_1(x), F_1(y)] = k_{1,1}(|x - y|)$. Ces processus sont caractérisés par des inter-covariances non symétriques en raison du retard τ. Dans ce cas, les inter-covariances entre $F_1(\boldsymbol{x})$ et $F_2(\boldsymbol{x})$ sont telles que $k_{1,2}(x, y) = k_{1,1}(|x - y + \tau|) \neq k_{1,2}(y, x) = k_{1,1}(|x - y - \tau|)$.

Il nous arrivera par la suite de considérer l'indice α comme un paramètre supplémentaire du processus aléatoire (d'après une remarque faite dans la section précédente), et nous utiliserons dans ce cas la notation $k([\alpha, \boldsymbol{x}], [\beta, \boldsymbol{y}])$. C'est un moyen commode pour se ramener formellement à une seule fonction de covariance.

2.1.3 Exemples de processus aléatoires

Processus aléatoire défini à partir de la donnée d'une séquence aléatoire. Une manière explicite de définir une probabilité sur l'espace des trajectoires est de se donner la collection des variables aléatoires définissant le processus. En transmissions numériques, on considère par exemple des processus indicés par le temps $t \in \mathbb{R}$, sous la forme

$$F(t) = \sum_{k \in \mathbb{Z}} A_k \mathbb{1}_{[0,T[}(t - kT)$$

où les A_k sont des variables aléatoires indépendantes qui constituent des symboles transmis et $\mathbb{1}_{[0,T[}(t)$ désigne la fonction indicatrice de l'intervalle $[0, T[$, valant un si $t \in [0, T[$, et zéro ailleurs. Les variables A_k sont à valeurs dans un alphabet fini $\{a_1, \ldots, a_m\}$, avec des probabilités $p_1, \ldots, p_m$ pour l'apparition des symboles. La loi du processus est entièrement caractérisée par les répartitions finies, en particulier par les probabilités des événements $B = \{F(t_1) \leq b_1, \ldots, F(t_n) \leq b_n\}$, tels que $\mathsf{P}(B) = \mathrm{E}[\mathbb{1}_B]$, où $\mathbb{1}_B$ est la variable aléatoire indicatrice de l'événement B. La loi du n-uplet $(F(t_1), \ldots, F(t_n))$ est en particulier la même que celle de $(F(t_1 + kT), \ldots, F(t_n + kT))$ (il s'agit d'un processus cyclostationnaire). Dans le cadre de ce mémoire, une telle construction de processus aléatoire ne sera jamais utilisée.

Définition à partir de la donnée de moments. Nous définirons plutôt un processus aléatoire à partir de sa loi et en général, nous nous contenterons de spécifier les moments d'ordre un et deux. Le type de processus qui sera sans doute le plus intéressant pour la modélisation boîte noire est le processus gaussien, dont la loi est uniquement déterminée par la donnée des moments d'ordre un et deux (la définition précise est rappelée dans la section 2.2). De manière informelle, un processus gaussien est caractérisé par une moyenne et une fonction de corrélation qui indique comment la valeur du processus en un point influence les valeurs en des points voisins. Typiquement, cette fonction de corrélation dépend de la distance entre les points de l'espace des facteurs et décroît lorsque cette distance augmente. Un processus du second ordre de moyenne nulle, spécifié uniquement par la corrélation entre les points, permet de construire de façon simple et relativement efficace des modèles de type boîte noire. La prédiction d'un processus aléatoire gaussien est abordée dans la section 2.5.1 et s'avère très simple à mettre en œuvre. La figure 2.1 représente des trajectoires possibles d'un processus gaussien pour deux fonctions de covariance. La diversité des trajectoires ainsi obtenues justifie l'intérêt de la modélisation de systèmes par des processus aléatoires gaussiens.

Nous présentons maintenant plus spécifiquement l'*exemple des bandes tournantes* (Matheron, 1973 ; Stein, 1999), introduit ici pour illustrer les difficultés liées au fait de ne spécifier que les moments d'ordre un et deux d'un processus aléatoire. G. Matheron propose la procédure suivante pour simuler un processus aléatoire sur $\mathbb{R}^d$. Soit $F_s(\omega, x)$ un processus aléatoire du second ordre défini sur $(\Omega, \mathcal{A}, \mathsf{P})$, paramétré par $x \in \mathbb{R}$, de moyenne nulle et admettant une covariance stationnaire $k_s(x-y)$. Soit de plus un vecteur aléatoire $\boldsymbol{U}(\omega) \in \mathbb{R}^d$, défini également sur $(\Omega, \mathcal{A}, \mathsf{P})$, de norme 1 pour tout ω, avec une loi admettant une densité uniforme sur la sphère unité ; $\sigma(\boldsymbol{U})$ et $\sigma(F_s(\mathbb{R}))$ sont supposées indépendantes. Considérons alors la fonction aléatoire $F_d(\boldsymbol{x}) = F_s(\,(\boldsymbol{x}, \boldsymbol{U})_{\mathbb{R}^d}\,)$, $\boldsymbol{x} \in \mathbb{R}^d$, où $(.,.)_{\mathbb{R}^d}$ désigne le produit scalaire dans $\mathbb{R}^d$. La moyenne de $F_d(\boldsymbol{x})$ est nulle puisque

$$\mathrm{E}[F_d(\boldsymbol{x})] = \mathrm{E}[\mathrm{E}[F_s(\,(\boldsymbol{x}, \boldsymbol{U})_{\mathbb{R}^d}\,) \mid \boldsymbol{U}]] = 0\,,$$

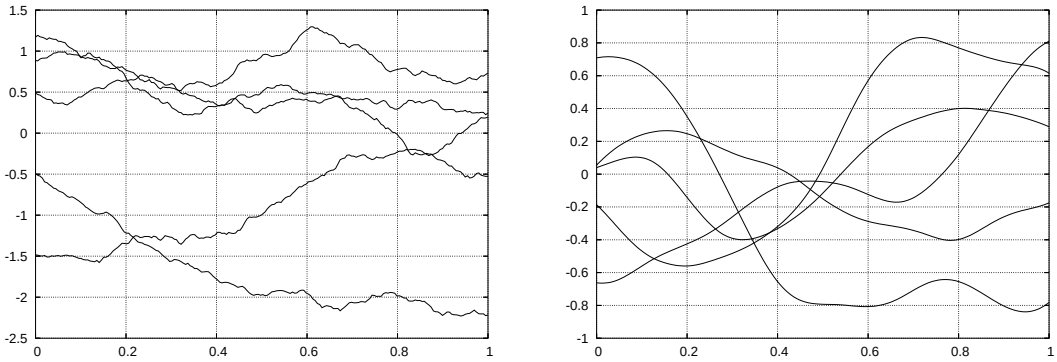


FIG. 2.1 – Trajectoires de processus gaussiens stationnaires de moyenne nulle simulées sur $[0,1]$. Le comportement de la covariance à l'origine influence la régularité des trajectoires (voir la section 2.4.4). À gauche, les trajectoires sont simulées en prenant une covariance continue mais non dérivable à l'origine. À droite, la covariance utilisée est deux fois continûment dérivable à l'origine.

où $\mathrm{E}[X \mid Y]$ désigne l'espérance conditionnelle de X sachant Y (quelques rappels sont présentés dans la section 2.5.1). La covariance de $F_d(\boldsymbol{x})$ s'écrit

$$\begin{aligned} k_d(\boldsymbol{x},\boldsymbol{y}) = \mathrm{Cov}[F_d(\boldsymbol{x}), F_d(\boldsymbol{y})] &= \mathrm{E}\left[\mathrm{E}\left[F_s\big((\boldsymbol{x},\boldsymbol{U})_{\mathbb{R}^d}\big)F_s\big((\boldsymbol{y},\boldsymbol{U})_{\mathbb{R}^d}\big) \mid \boldsymbol{U}\right]\right] \\ &= \mathrm{E}\left[k\big((\boldsymbol{x}-\boldsymbol{y},\boldsymbol{U})_{\mathbb{R}^d}\big)\right] \\ &= \int_{\partial B(0,1)} k\big((\boldsymbol{x}-\boldsymbol{y},\boldsymbol{u})_{\mathbb{R}^d}\big) d\mathsf{P}_{\partial B(0,1)}(\boldsymbol{u}) \end{aligned}$$

où $\mathsf{P}_{\partial B(0,1)}$ est la probabilité concentrée sur la sphère unité, invariante par rotation.

Pour des raisons de symétries, k_d ne dépend que de $h = \|\boldsymbol{x}-\boldsymbol{y}\|$. (Matheron, 1973) donne la forme analytique de la covariance k_d en fonction de la covariance ${k_s}^2$:

$$k_d(h) = \frac{2\Gamma(\frac{1}{2}d)}{\pi^{\frac{1}{2}}\Gamma(\frac{1}{2}(d-1))} \int_0^1 k_s(vh)(1-v^2)^{\frac{1}{2}(d-3)} dv\,.$$

La covariance est donc isotrope mais les réalisations du processus ne reflètent pas cette caractéristique puisqu'à ω fixé, la trajectoire $F_d(\omega,\boldsymbol{x})$ est constante le long des hyperplans tels que $(\boldsymbol{x},\boldsymbol{U})_{\mathbb{R}^d}$ soit constant. Cet exemple illustre donc le fait que les moments d'ordre deux d'un processus aléatoire sont loin de déterminer toutes ses caractéristiques. En pratique, il est nécessaire de faire des hypothèses supplémentaires sur le processus aléatoire. Dans de nombreux cas, il est pertinent de modéliser la sortie d'un système par un processus aléatoire gaussien, dont la loi est caractérisée uniquement par les moments d'ordre deux. Dans ce cas, les trajectoires possèdent des propriétés spécifiques (voir la section 2.4.4) et les procédures de prédiction linéaire que nous présenterons dans la section 2.5.1 sont optimales. Si l'hypothèse gaussienne n'est pas adaptée, il est possible dans certains cas de trouver des transformations permettant de s'y ramener (voir par exemple Chilès et Delfiner (1999)). Nous ne présenterons pas de telles méthodes dans ce mémoire.

[2]Cette formule est importante d'un point de vue théorique car toute covariance isotrope en dimension d s'exprime en fonction d'une covariance admissible en dimension un sous cette forme.

Un processus fondamental : le mouvement brownien. Le mouvement brownien $W(t)$, $t \in \mathbb{R}^+$ (aussi appelé processus de Wiener) est un processus extrêmement important en probabilités. Ses trajectoires sont continues, ses accroissement sont indépendants et c'est un processus markovien. Il est par exemple d'utilisation constante dans la théorie des équations différentielles stochastiques. Rappelons sa définition.

Définition 8. Un processus aléatoire $W(t)$, $t \in \mathbb{R}^+$ est un mouvement brownien si :

1. $W(t)$ est un processus gaussien
2. $\mathrm{E}[W(t)] = 0$, $\mathrm{E}[W(t)W(s)] = \min(t, s)$
3. Pour presque tout ω, les trajectoires $t \mapsto W(\omega, t)$ sont continues

Notons que les points 1 et 2 dans la définition précédente caractérisent complètement les répartitions finies du mouvement brownien. L'ajout de la condition 3 permet de spécifier la nature des trajectoires.

$W(t)$ est à accroissement orthogonaux puisque, pour $t \geq s \geq 0$, $h > 0$,

$$\mathrm{Cov}[W(s), W(t+h) - W(t)] = \\ \mathrm{E}[W(s)(W(t+h) - W(t))] = \mathrm{E}[W(s)W(t+h)] - \mathrm{E}[W(s)W(t)] = s - s = 0$$

Une telle propriété sera par exemple utilisée au chapitre 4.

Bruit blanc. Nous terminons cette section sur la notion de bruit blanc à temps ou paramètres continus. Il est bien connu que la définition d'un bruit blanc à temps continu est délicate. Le cas du bruit blanc à temps discret ne pose pas de problème puisqu'il s'agit d'une suite de variables aléatoires indépendantes, de même moyenne et de même variance. Le bruit blanc, stationnaire, admet alors une mesure spectrale absolument continue par rapport à la mesure de Lebesgue sur $]-\pi, \pi]$, avec une densité spectrale constante. Dans le cas du temps continu, si l'on considère une densité spectrale constante sur $]-\infty, \infty[$, il n'existe pas de processus aléatoire du second ordre admettant cette densité. La définition rigoureuse d'un bruit blanc nécessite des outils théoriques supplémentaires pour être correctement établie. Dans les paragraphes suivants, nous présentons brièvement une définition possible de bruit blanc défini sur $\mathbb{X} = \mathbb{R}^d$.

Considérons une mesure μ σ-finie sur $\mathcal{B}(\mathbb{R}^d)$. Soit une fonction aléatoire $W(\omega, B)$ paramétrée par les ensembles B de mesure finie de $\mathcal{B}(\mathbb{R}^d)$, à valeurs dans $\mathbb{R}$, admettant la propriété d'additivité suivante : quels que soient B_1 et B_2 disjoints, $W(B_1 \cup B_2) = W(B_1) + W(B_2)$ avec probabilité 1. Supposons de plus $\mathrm{E}[W(B)] = 0$, $\forall B \in \mathcal{B}(\mathbb{R}^d)$, et

$$\mathrm{Cov}[W(B_1), W(B_2)] = \mu(B_1 \cap B_2)\,, \quad \forall B_1, B_2 \in \mathcal{B}(\mathbb{R}^d)\,.$$

Une telle fonction d'ensembles W s'appelle une *mesure aléatoire*. Nous en rappellerons la définition précise dans la section 2.4.2. Les expressions intégrales du type

$$U = \int_{\mathbb{X}} \varphi(\boldsymbol{x}) dW(\boldsymbol{x})\,, \tag{2.5}$$

où $\varphi(\boldsymbol{x})$ est une fonction de $L^2(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \mu)$, sont des *intégrales stochastiques* (Doob, 1953 ; Rozanov, 1967) dont le sens précis sera revu dans la section 2.4.2. Les variables aléatoires U ainsi

définies forment un espace complet, lorsque φ parcourt $L^2(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \mu)$. Formellement, nous pouvons considérer la dérivée $N(\boldsymbol{x}) = (dW/d\mu)_{\boldsymbol{x}}$ et écrire (2.5) sous la forme $\int \varphi(\boldsymbol{x})N(\boldsymbol{x})d\mu(\boldsymbol{x})$. De même nous pouvons écrire l'expression $\mathrm{Cov}[N(\boldsymbol{x}), U]$ à la place de $d\,\mathrm{Cov}[W(B), U]/d\mu$. L'objet $N(\boldsymbol{x})$ ainsi construit s'appelle *bruit blanc*.

À moins que le point $\boldsymbol{x}$ n'ait une mesure positive, $N(\boldsymbol{x})$ n'a pas de signification propre. Seules les expressions intégrales $\int \varphi N d\mu$, $\varphi \in L^2(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \mu)$, et les covariances $\mathrm{Cov}[N(\boldsymbol{x}), U]$, $\boldsymbol{x} \in \mathbb{R}^d$ ont un sens correctement défini comme ci-dessus. Nous n'entrerons pas davantage dans les détails de la théorie des distributions aléatoires. Pour résumer, le bruit blanc à paramètres continus est un processus aléatoire qui ne peut pas être évalué en un point $\boldsymbol{x}$. Pour cette raison, nous éviterons l'emploi de tels bruits blancs dans les modèles aléatoires que nous considérerons. Toutefois, la notion de bruit blanc joue un rôle important dans la modélisation de perturbations, en particulier lorsque l'on commet des erreurs sur les observations de la sortie du système. Nous distinguerons principalement deux cas.

Le cas d'erreurs d'observation sera modélisé par des variables aléatoires N_j, $j = 1, \ldots, n$, indépendantes. Il n'y a pas de raisons en effet de modéliser des erreurs d'observations, nécessairement en nombre fini, par un processus aléatoire à paramètres continus.

S'il existe des raisons spécifiques de modéliser le système à l'aide d'un processus aléatoire comportant une composante de type bruit blanc, nous adopterons le point de vue des géostatisticiens qui utilisent dans ce cas la notion d'« effet de pépites ». Un processus aléatoire comportant une composante dite de « pépites » est un processus présentant des variations à très petite échelle. Ces variations à très petite échelle se caractérisent donc par une fonction de corrélation $\rho(\boldsymbol{x}, \boldsymbol{y})$ qui décroît rapidement dès que $\|\boldsymbol{x} - \boldsymbol{y}\|$ augmente au voisinage de zéro. Cela signifie qu'à l'échelle du domaine d'étude, la structure de corrélation est caractérisée par un saut à l'origine mais il ne s'agit pas d'une discontinuité théorique. D'un point de vue numérique *uniquement*, un « effet de pépite » se traduit par une discontinuité de la fonction de covariance à l'origine.

2.2 Processus gaussiens

2.2.1 Rôle des processus gaussiens

De multiples raisons expliquent le rôle central des processus gaussiens dans les modèles que nous considérons. Tout d'abord, l'utilisation de variables aléatoires gaussiennes peut souvent être justifiée par le principe du *maximum d'entropie*. En effet, lorsque nous modélisons la sortie d'un système par un processus aléatoire, les connaissances a priori dont nous disposons ne permettent pas en général de déterminer la loi du processus. Tout au plus, pouvons-nous estimer ses moments d'ordre un et deux. Parmi tous les processus aléatoires de moyenne et de covariance données, le principe du maximum d'entropie suggère alors de choisir ceux comportant le moins d'*information* supplémentaire (au sens de la *théorie de l'information*). L'information est mesurée par l'*entropie* et il s'agit donc de choisir un processus aléatoire tel que son entropie soit la plus grande possible. Or, parmi toutes les lois de moyenne et de covariance données, la loi normale est celle qui a l'entropie maximale. Le *principe du maximum d'entropie* peut être vu comme un principe de simplicité.

D'autre part, les processus gaussiens possèdent des propriétés théoriques spécifiques aujourd'hui très bien comprises (il existe des études détaillées comme par exemple Ibragimov et Rozanov

(1978) ; Adler (1990)). Enfin, la forme très simple des répartitions finies d'un processus gaussien, entièrement caractérisées par des moments d'ordre un et deux seulement, permet d'obtenir facilement des solutions analytiques exactes pour nombre de problèmes théoriques et appliqués, comme par exemple celui de la prédiction. Ainsi, les opérations de conditionnement et de marginalisation sur des variables aléatoires gaussiennes redonnent des variables aléatoires gaussiennes et il suffit souvent d'effectuer des opérations simples d'algèbre matricielle pour trouver les paramètres correspondants.

2.2.2 Définitions

Rappelons, sans démonstrations, les notions bien connues sur les processus gaussiens qui seront nécessaires dans cette présentation. Une variable aléatoire X réelle est gaussienne si sa fonction caractéristique

$$\phi_X(u) = \int_{\mathbb{R}} e^{\imath u x} d\mathsf{P}_X(x)\,,$$

est de la forme

$$\phi_X(u) = \exp\left(\imath u m - \frac{1}{2}\sigma^2 u^2\right)\,.$$

On montre que $m = \mathrm{E}[X]$ et $\sigma^2 = \mathrm{Var}[X]$. La loi de X admet alors la densité

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(x-m)^2}{\sigma^2}\right)\,.$$

Définition 9. Un *vecteur aléatoire* $(X_1, \ldots, X_n)^\mathsf{T}$ est *gaussien* si toute combinaison linéaire de ses composantes $X_1, \ldots, X_n$ est une variable aléatoire gaussienne (avec la convention $X = 0$ est une variable aléatoire gaussienne).

Si $(X_1, \ldots, X_n)^\mathsf{T}$ est un vecteur aléatoire du second ordre, rappelons que $\boldsymbol{K} = (k_{i,j})_{i,j=1,\ldots,n}$, où $k_{i,j} = \mathrm{Cov}(X_i, X_j)$, $i, j = 1 \ldots n$, est la *matrice de covariance* de ce vecteur aléatoire.

Proposition 4. *Un vecteur aléatoire du second ordre $\boldsymbol{X}$, de moyenne $\boldsymbol{m}$ et de matrice de covariance $\boldsymbol{K}$ est gaussien si et seulement si sa fonction caractéristique*

$$\phi_{\boldsymbol{X}}(\boldsymbol{u}) = \int_{\mathbb{R}^n} \exp(\imath(\boldsymbol{u}, \boldsymbol{x})) d\mathsf{P}_{\boldsymbol{X}}(\boldsymbol{x})\,,$$

où $(\boldsymbol{u}, \boldsymbol{x})$ désigne le produit scalaire canonique de $\mathbb{R}^n$, s'écrit

$$\phi_{\boldsymbol{X}}(\boldsymbol{u}) = \exp\left(\imath(\boldsymbol{u}, \boldsymbol{m}) - \frac{1}{2}(\boldsymbol{K}\boldsymbol{u}, \boldsymbol{u})\right)\,. \tag{2.6}$$

Démonstration. Voir par exemple (Ibragimov et Rozanov, 1978). □

La moyenne est telle que $\forall \boldsymbol{u} \in \mathbb{R}^n$,

$$(\boldsymbol{u}, \boldsymbol{m}) = \int_{\mathbb{R}^n} (\boldsymbol{u}, \boldsymbol{x}) d\mathsf{P}_{\boldsymbol{X}}(\boldsymbol{x})\,.$$

La matrice de covariance est telle que, $\forall \boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n$,

$$(\boldsymbol{K}\boldsymbol{u}, \boldsymbol{v}) = \int_{\mathbb{R}^n} [(\boldsymbol{u}, \boldsymbol{x}) - (\boldsymbol{u}, \boldsymbol{m})][(\boldsymbol{v}, \boldsymbol{x}) - (\boldsymbol{v}, \boldsymbol{m})] d\mathsf{P}_{\boldsymbol{X}}(\boldsymbol{x})\,.$$

Remarque. On dit qu'une probabilité sur $\mathbb{R}^n$ est gaussienne si elle est la loi d'un vecteur aléatoire gaussien.

Proposition 5. *Une probabilité gaussienne sur $\mathbb{R}^n$ admet une densité sur $\mathbb{R}^n$*

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{n/2}\det(\boldsymbol{K})^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{K}^{-1}(\boldsymbol{x}-\boldsymbol{m}),(\boldsymbol{x}-\boldsymbol{m}))\right),$$

si et seulement si $\boldsymbol{K}$ est non-dégénérée.

Démonstration. Voir (Ibragimov et Rozanov, 1978). □

Si $\boldsymbol{X}$ est un vecteur aléatoire gaussien et $\boldsymbol{K}$ n'est pas inversible, il existe $\boldsymbol{u} \neq 0 \in \mathbb{R}^n$, tel que $\boldsymbol{K}\boldsymbol{u} = 0$. La variable aléatoire $Y = (\boldsymbol{u}, \boldsymbol{X})$, de moyenne $(\boldsymbol{u}, \boldsymbol{m})$ et de variance $(\boldsymbol{K}\boldsymbol{u}, \boldsymbol{u}) = 0$, est telle que $\mathsf{P}\{Y = (\boldsymbol{u}, \boldsymbol{m})\} = 1$. Donc $\boldsymbol{X}$ est presque sûrement dans l'hyperplan affine $(\boldsymbol{x}, \boldsymbol{u}) = (\boldsymbol{u}, \boldsymbol{m})$ orthogonal à $\boldsymbol{u}$. Autrement dit, la probabilité gaussienne $\mathsf{P}_{\boldsymbol{X}}$ est portée par le plan $P = \boldsymbol{m} + \boldsymbol{K}\mathbb{R}^n$, de dimension $p \leq n$, p étant le rang de $\boldsymbol{K}$. On a donc $\mathsf{P}_{\boldsymbol{X}}(\mathbb{R}^n \backslash P) = 0$.

Si X_1 et X_2 sont deux variables aléatoires indépendantes de fonctions caractéristiques ϕ_{X_1} et ϕ_{X_2}, la fonction caractéristique de la somme est donnée par $\phi_{X_1+X_2} = \phi_{X_1}\phi_{X_2}$. Toute combinaison linéaire de variables aléatoires gaussiennes indépendantes est donc gaussienne. De même, la limite d'une suite de variables aléatoires gaussiennes $(X_n)_{n\in\mathbb{N}}$ convergente en loi est une variable aléatoire gaussienne. Pour s'en assurer, on peut utiliser le théorème général de Lévy suivant.

Théorème 3. *Soit une suite $(X_n)_{n\in\mathbb{N}}$ de variables aléatoires à valeurs dans $\mathbb{R}^d$.*

1. *Si la suite (X_n) converge en loi vers X, alors ϕ_{X_n} converge simplement vers ϕ_X.*
2. *Si les ϕ_{X_n} convergent simplement vers une fonction complexe ϕ sur $\mathbb{R}^d$, et si cette fonction est continue en 0, alors c'est la fonction caractéristique d'une variable aléatoire X et (X_n) converge en loi vers X.*

Démonstration. Voir (Billingsley, 1995), section 26. □

Définition 10. Un processus aléatoire est gaussien si toutes ses répartitions finies sont gaussiennes. Chacune des répartitions finies $\mathsf{P}_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n}$ du processus aléatoire est caractérisée par sa moyenne $(m(\boldsymbol{x}_1),\ldots,m(\boldsymbol{x}_n))$, où $m(\boldsymbol{x})$, $\boldsymbol{x} \in \mathbb{X}$, est la moyenne de $F(\boldsymbol{x})$, et par sa matrice de covariance $\boldsymbol{K} = (k(\boldsymbol{x}_i, \boldsymbol{x}_j))_{i,j=1\ldots n}$, où $k(\boldsymbol{x}, \boldsymbol{y})$, $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{X}$, est la covariance de $F(\boldsymbol{x})$.

Par conséquent, la loi d'un processus aléatoire gaussien est définie de manière unique par la donnée de sa moyenne $m(\boldsymbol{x})$ et de sa covariance $k(\boldsymbol{x}, \boldsymbol{y})$.

2.3 Variables aléatoires à valeurs dans un espace hilbertien

Dans les sections précédentes, nous avons rappelé les définitions des variables, vecteurs et processus aléatoires. Nous souhaitons maintenant généraliser ces notions et présenter celle de *variables aléatoires à valeurs dans des espaces hilbertiens*, telle que définie par exemple dans (Ibragimov et Rozanov, 1978). La notion de variable aléatoire à valeurs dans un espace de fonctions permet de construire un processus aléatoire d'une autre manière que celle considérée ci-dessus. Dans les paragraphes précédents, un processus aléatoire avait été construit comme une collection de variables

aléatoires à valeurs réelles, en mettant l'accent sur les répartitions finies. Ici, nous considérons directement une application de Ω dans un espace de fonctions $\mathcal{F}$. Le cas où $\mathcal{F}$ est un espace de fonctions hilbertien présente un aspect théorique intéressant. L'intérêt de ce formalisme apparaîtra plus précisément au chapitre 3, lorsque nous verrons les espaces hilbertiens de fonctions à noyau reproduisant.

La notion de variables aléatoires à valeurs dans un espace hilbertien s'apparente à celle de vecteur aléatoire gaussien ; rappelons qu'un vecteur aléatoire $\boldsymbol{X}$ est gaussien si et seulement si $\boldsymbol{X}(\boldsymbol{u}) = (\boldsymbol{X}, \boldsymbol{u})$ est une variable aléatoire gaussienne, pour tout $\boldsymbol{u} \in \mathbb{R}^n$. Soit $\mathcal{F}$ un espace de Hilbert et $F(\omega)$ une fonction définie sur un espace probabilisé $(\Omega, \mathcal{A}, \mathsf{P})$ à valeurs dans $\mathcal{F}$.

Définition 11. $F : \Omega \to \mathcal{F}$ est une *variable aléatoire à valeurs dans* $\mathcal{F}$, si $\forall u \in \mathcal{F}$, $F(u) = (F, u)_{\mathcal{F}}$ est une fonction mesurable sur $(\Omega, \mathcal{A}, \mathsf{P})$.

F définit ainsi un processus aléatoire $F(u)$, $u \in \mathcal{F}$, à valeurs réelles. Par définition, F est une variable aléatoire gaussienne si la variable aléatoire $F(u)$ est gaussienne pour tout u. Si $\mathcal{F}$ est séparable, on dit que F est à valeurs séparables.

Proposition 6. *Les deux affirmations qui suivent sont équivalentes :*
- *la variable aléatoire F est gaussienne,*
- *le processus aléatoire $F(u)$ est gaussien.*

Démonstration. Supposons F gaussienne. Alors pour tous $\lambda_1, \ldots, \lambda_n \in \mathbb{R}^n$ et $u_1, \ldots, u_n \in \mathcal{F}$, la combinaison linéaire

$$\sum_{i=1}^{n} \lambda_i F(u_i) = (\sum_{i=1}^{n} \lambda_i u_i, F)_{\mathcal{F}}$$

est gaussienne. Donc le vecteur aléatoire $(F(u_1), \ldots, F(u_n))$ est gaussien ce qui montre que $F(u)$ est un processus gaussien. La réciproque est évidente. □

Définition 12. Soit $\mathcal{F}$ un espace hilbertien. Un ensemble $U \subset \mathcal{F}$ est dit *cylindrique* s'il est de la forme

$$U = \{f \in \mathcal{F} \,;\, ((u_1, f)_{\mathcal{F}}, \ldots, (u_n, f)_{\mathcal{F}}) \in B\}$$

où l'on choisit $n \geq 1$, $u_1, \ldots, u_n \in \mathcal{F}$, et un ensemble borélien $B \subset \mathbb{R}^n$. Les ensembles cylindriques génèrent sur $\mathcal{F}$ une σ-algèbre, que l'on note $\mathcal{C}(\mathcal{F})$.

Remarquons que si $\mathcal{C}(\mathcal{F})$ désigne la σ-algèbre cylindrique sur $\mathcal{F}$, alors $\mathcal{C}(\mathcal{F}) \subseteq \mathcal{B}(\mathcal{F})$, où $\mathcal{B}(\mathcal{F})$ désigne la σ-algèbre borélienne de $\mathcal{F}$ pour la topologie de la norme de l'espace hilbertien $\mathcal{F}$ (en effet, on peut montrer que tout ensemble cylindrique est ouvert).

Proposition 7. *Soit F une variable aléatoire dans l'espace hilbertien $\mathcal{F}$. Avec le choix d'ensembles mesurables $\mathcal{C}(\mathcal{F})$ de $\mathcal{F}$, l'application F de $(\Omega, \mathcal{A}, \mathsf{P})$ dans $(\mathcal{F}, \mathcal{C}(\mathcal{F}))$ est mesurable.*

Démonstration. Ce résultat a déjà été rencontré dans la section 2.1.1. □

Définition 13. Si F est une variable aléatoire sur $(\Omega, \mathcal{A}, \mathsf{P})$ à valeurs dans l'espace hilbertien $\mathcal{F}$, la probabilité sur $(\mathcal{F}, \mathcal{C}(\mathcal{F}))$ définie par

$$\mathsf{P}_F(U) = \mathsf{P}\{F^{-1}(U)\}\,, \quad U \in \mathcal{C}(\mathcal{F})\,,$$

est la *loi* de F.

Les notions de moyenne et de covariance s'étendent également aux variables aléatoires dans un espace hilbertien. Une mesure μ sur $(\mathcal{F}, \mathcal{C}(\mathcal{F}))$ est dite d'*ordre p au sens faible* si

$$\int_{\mathcal{F}} |(u, f)_{\mathcal{F}}|^p d\mu(f) < \infty$$

pour tout $u \in \mathcal{F}$. Une variable aléatoire F est d'ordre p (au sens faible) si sa loi possède la propriété correspondante.

Proposition 8. *Soit $(\Omega, \mathcal{A}, \mathsf{P})$ un espace de probabilités, $\mathcal{F}$ un espace hilbertien, et $F : \Omega \to \mathcal{F}$ une variable aléatoire d'ordre un au sens faible. Alors il existe un élément $\mathrm{E}[F] \in \mathcal{F}$ tel que*

$$\int_{\Omega} (F(\omega), u)_{\mathcal{F}} d\mathsf{P}(\omega) = \int_{\mathcal{F}} (f, u)_{\mathcal{F}} d\mathsf{P}_F(f) = (\mathrm{E}[F], u)_{\mathcal{F}} \tag{2.7}$$

pour tout $u \in \mathcal{F}$.

Démonstration. Voir (Ibragimov et Rozanov, 1978). □

L'élément $\mathrm{E}[F]$ défini dans la proposition 8 est appelé moyenne de F (il s'agit d'une intégrale dite de Pettis). Remarquons que (2.7) peut être réécrite

$$\mathrm{E}[(F, u)_{\mathcal{F}}] = (\mathrm{E}[F], u)_{\mathcal{F}} .$$

La *moyenne* de $F(u)$ est notée

$$m(u) = \mathrm{E}[F(u)] = (\mathrm{E}[F], u)_{\mathcal{F}} , \quad u \in \mathcal{F} .$$

L'application $m(u)$ est linéaire et continue sur $\mathcal{F}$ (Ibragimov et Rozanov, 1978).

Proposition 9. *Soit $\mathcal{F}$ un espace hilbertien de fonctions et μ une mesure du second ordre au sens faible définie sur $\mathcal{C}(\mathcal{F})$. La forme bilinéaire définie sur $\mathcal{F} \times \mathcal{F}$ par*

$$A_\mu(u, v) = \int_{\mathcal{F}} (f, u)(f, v) d\mu(f) - \int_{\mathcal{F}} (f, u) d\mu(f) \int_{\mathcal{F}} (f, v) d\mu(f)$$

pour tous $u, v \in \mathcal{F}$, est continue pour la topologie de la norme de $\mathcal{F}$. L'opérateur linéaire K défini par

$$(Ku, v)_{\mathcal{F}} = A_\mu(u, v) \tag{2.8}$$

est auto-adjoint, positif et continu sur $\mathcal{F}$.

Démonstration. Voir (Ibragimov et Rozanov, 1978). □

Définition 14. K défini par (2.8) est appelé *opérateur de covariance*. L'opérateur de covariance d'une variable aléatoire F du second ordre, caractérisée par sa loi P_F, est donc défini par

$$\begin{aligned} (Ku, v) &= \mathrm{E}[(u, F)(v, F)] - \mathrm{E}[(u, F)]\,\mathrm{E}[(v, F)] \\ &= \mathrm{E}[((u, F) - (\mathrm{E}[F], u))((v, F) - (\mathrm{E}[F], v))] . \end{aligned}$$

Soit F du second ordre, K son opérateur de covariance et $k(u, v)$ la fonction de covariance de $F(u)$. Alors

$$(Ku, v) = k(u, v) .$$

Le résultat suivant récapitule les relations vues ci-dessus entre l'ordre d'une variable aléatoire F et l'ordre du processus aléatoire $F(u)$ qu'elle définit.

Proposition 10. *Soit F une variable aléatoire à valeurs dans $\mathcal{F}$ et $F(u)$ le processus aléatoire défini par F.*

1. *Si F est du premier ordre, alors $F(u)$ est du premier ordre.*
2. *Si F est du second ordre, alors $F(u)$ est du second ordre.*

Démonstration. Si F est du premier ordre, chaque $F(u)$ est intégrable car

$$\mathrm{E}[|F(u)|] = \mathrm{E}[|(F,u)|] < \infty\,.$$

De même, si F est du second ordre, $F(u)$ est du second ordre pour tout u car

$$\mathrm{E}[F(u)^2] = \mathrm{E}[(F,u)^2] < \infty$$

□

Les propriétés réciproques de la proposition précédente sont *fausses* : l'ordre d'un processus aléatoire $F(u)$ ne détermine pas l'ordre de la variable aléatoire F correspondante. (Lukic et Beder, 2001) en donne un exemple très instructif dans un contexte d'espaces hilbertiens à noyau reproduisant. Nous reproduisons ci-dessous une partie de cet exemple.

Exemple. Considérons l'espace $\mathcal{F} = \ell^2(\mathrm{a})$ des suites de carré sommable avec poids $a_n = n^2$, muni du produit scalaire $(\mathrm{u},\mathrm{v}) = \sum_{n=1}^{\infty} n^2 u_n v_n$. Soit la fonction $F : \mathbb{N} \to \mathcal{F}$, définie par

$$F(n) = \mathrm{e}_n\,, \quad \forall n \in \mathbb{N}\,,$$

où $\mathrm{e}_n = (0,\ldots,\frac{1}{n},0,\ldots)$, $n = 1,2,\ldots$, forme une base orthonormée de $\mathcal{F}$. En choisissant l'espace mesuré $(\mathbb{N},\mathcal{P}(\mathbb{N}),\mathsf{P})$ avec

$$\mathsf{P}(n) = p_n\,, \quad p_n > 0 \quad \sum_n p_n = 1\,,$$

F est une variable aléatoire à valeurs dans $\mathcal{F}$. Clairement, $F(\mathrm{e}_n) = (F,\mathrm{e}_n)_{\mathcal{F}}$, $n \in \mathbb{N}$, est un processus aléatoire à trajectoires dans $\mathcal{F}$, du second ordre, avec $\mathrm{E}[F(\mathrm{e}_n)] = \mathrm{E}[F(\mathrm{e}_n)^2] = p_n$ et $\mathrm{E}[F(\mathrm{e}_n)F(\mathrm{e}_m)] = 0$, si $n \neq m$ (remarquer que $F(i,\mathrm{e}_n)$ vaut 1 si et seulement si $i = n$). Pour tout $\mathrm{u} \in \mathcal{F}$, $F(\mathrm{u}) = (F,\mathrm{u})_{\mathcal{F}} = \sum nF(\mathrm{e}_n)u_n$ et

$$\mathrm{E}[F(\mathrm{u})^2] = \mathrm{E}[(F,\mathrm{u})_{\mathcal{F}}^2] = \sum_n \mathrm{E}[F(\mathrm{e}_n)^2]n^2u_n^2 \leq \|\mathrm{u}\|^2 < \infty\,,$$

ce qui montre que $F(\mathrm{u})$ est du second ordre pour tout $\mathrm{u} \in \mathcal{F}$.

La loi de probabilité P_F de F est une probabilité sur $(\ell^2(\mathrm{a}),\mathcal{C})$, où $\mathcal{C}$ est la σ-algèbre cylindrique de $\ell^2(\mathrm{a})$. P_F est portée par l'ensemble $\{\mathrm{e}_i\}$ et telle que $\mathsf{P}_F(\mathrm{e}_n) = \mathsf{P}(n) = p_n$. En choisissant par exemple $p_n = C/n^{\frac{3}{2}}$, où C est un coefficient de normalisation, $\mathrm{m} = \mathrm{E}[F] = (C/n^{\frac{3}{2}})_n$ n'est pas un élément de l'espace hilbertien $\mathcal{F} = \ell^2(\mathrm{a})$. Par conséquent F n'est pas une variable aléatoire du premier ordre, contrairement à $F(\mathrm{u})$.

2.4 Propriétés d'un processus aléatoire

Rappelons quelques notions fondamentales utilisées dans le cadre de la prédiction linéaire.

2.4.1 Stationnarité

Définition 15. Un processus aléatoire $F(\boldsymbol{x})$, $\boldsymbol{x} \in \mathbb{R}^d$, est *stationnaire* (au sens strict) si ses répartitions finies sont invariantes par translation.

La stationnarité d'un processus aléatoire implique que ses moments, s'ils existent, sont invariants par translation.

Définition 16. Un processus aléatoire du second ordre est *stationnaire au second ordre* si ses moments d'ordre deux sont invariants par translation. Dans ce cas $m(\boldsymbol{x}) = m$ et $k(\boldsymbol{x}, \boldsymbol{y}) = k(\boldsymbol{x}-\boldsymbol{y})$, avec l'abus de notation usuel. On parle alors de *covariance stationnaire*.

Dans le cas des processus gaussiens, la stationnarité au second ordre implique la stationnarité au sens strict. Dans la littérature, on trouve aussi le terme de *champ aléatoire homogène* pour désigner un processus aléatoire stationnaire sur $\mathbb{R}^d$.

Remarques sur la forme des covariances stationnaires. Une covariance stationnaire est telle que $|k(\boldsymbol{h})| \leq k(\boldsymbol{0})$ $\forall \boldsymbol{h} \in \mathbb{R}^d$, d'après l'inégalité de Schwarz. La covariance à l'origine $k(0)$ est la variance du processus. Les propriétés d'un processus aléatoire gaussien stationnaire sont déterminées par la forme de la corrélation. Ceci explique que dans le cas de processus stationnaires, certains ouvrages ne font pas la distinction entre fonction de corrélation et de covariance.

Nous supposons dans la suite de ce chapitre que les processus sont stationnaires, et ceci pour deux raisons. La première est que l'invariance des propriétés du modèle aléatoire par translation dans l'espace des facteurs est importante dans l'étape de l'estimation du modèle (choix de la covariance, voir le chapitre 5). En effet, les observations de la sortie du système correspondent à une seule réalisation du processus aléatoire. Or l'inférence statistique des paramètres du modèle nécessite en principe une répétition des réalisations. On espère donc pouvoir remplacer la *répétition des réalisations* par une *répétition dans l'espace des facteurs*. La seconde raison fondamentale qui justifie l'hypothèse de stationnarité est que les processus stationnaires admettent une *représentation spectrale* (voir la section 2.4.2). Au chapitre 4, le cas de processus aléatoires à moyenne non stationnaires sera traité en se ramenant à des processus stationnaires.

Quand plusieurs processus aléatoires sont considérés simultanément, on peut donner la définition suivante.

Définition 17. Les processus aléatoires $F_\alpha(\boldsymbol{x})$, $\alpha = 1, \ldots, q$, $\boldsymbol{x} \in \mathbb{R}^d$ sont *stationnairement corrélés* si leur moyenne est constante, c'est-à-dire, $\mathrm{E}[F_\alpha(\boldsymbol{x})] = m_\alpha$, $\alpha = 1, \ldots, q$, et si les fonctions de covariance et d'inter-covariances sont invariantes par translation :

$$k_{\alpha,\beta}(\boldsymbol{x} + \boldsymbol{h}, \boldsymbol{y} + \boldsymbol{h}) = k_{\alpha,\beta}(\boldsymbol{x}, \boldsymbol{y})\,,\ \forall \boldsymbol{h}\,.$$

2.4.2 Éléments de représentation spectrale des processus stationnaires

Cette section rappelle les quelques notions de représentation spectrale qui seront utilisées dans la section 2.4.3 et au chapitre 5.

Théorème 4 (Bochner). *Une fonction réelle $k(\boldsymbol{h})$, $\boldsymbol{h} \in \mathbb{R}^d$ est de type positif si et seulement si elle peut être représentée sous la forme*

$$k(\boldsymbol{h}) = \int_{\mathbb{R}^d} e^{\imath(\boldsymbol{u},\boldsymbol{h})} d\xi(\boldsymbol{u}) \,,$$

où $\xi(\boldsymbol{u})$ est une mesure positive bornée sur $\mathbb{R}^d$.

Démonstration. Voir par exemple (Gikhman et Skorohod, 1974). □

Soit $F(\boldsymbol{x})$, $\boldsymbol{x} \in \mathbb{R}^d$, un processus aléatoire stationnaire de moyenne nulle et de covariance $k(\boldsymbol{h})$. La mesure ξ correspondant à une covariance est appelée *mesure spectrale* de $F(\boldsymbol{x})$. Lorsque la mesure spectrale admet une densité, on parle de *densité spectrale*. Les propriétés de la densité spectrale sont utiles pour caractériser un processus aléatoire du second ordre. Par exemple, on peut relier la densité spectrale à la régularité du processus (voir la section 2.4.4) ou établir la forme des covariances isotropes en dimension d (voir le chapitre 5).

En vue d'obtenir une représentation spectrale d'un processus aléatoire du second ordre stationnaire, il est classique de définir la notion de processus à accroissement orthogonaux, puis la notion d'intégrale stochastique. Nous nous limiterons à quelques rappels de définitions. Pour plus de détails, voir (Rozanov, 1967 ; Brockwell et Davis, 1987) ou des traités d'analyse fonctionnelle comme (Riesz et Nagy-Sz., 1965), qui traite des représentations spectrales de manière générale.

Définition 18 (Mesure aléatoire). Une fonction Φ définie sur la tribu borélienne $\mathcal{B}(\mathbb{R}^d)$, à valeurs dans $L^2(\Omega, \mathcal{A}, \mathsf{P})$, est une *mesure aléatoire* si elle possède la propriété d'additivité dénombrable :

$$\forall (B_j)_{j\in\mathbb{N}} \in \mathcal{B}(\mathbb{R}^d) \text{ tel que } B_j \cap B_k = \emptyset \text{ pour } j \neq k \,, \Phi(\cup_{j\in\mathbb{N}} B_j) = \sum_{j\in\mathbb{N}} \Phi(B_j) \,,$$

et la propriété d'orthogonalité :

$$\forall B_1, B_2 \in \mathcal{B}(\mathbb{R}^d) \text{ tel que } B_1 \cap B_2 = \emptyset \quad \mathrm{E}[\Phi(B_1)\Phi(B_2)] = 0 \,.$$

À une mesure aléatoire, on associe une mesure positive définie par

$$\xi_\Phi(B) = \|\Phi(B)\|^2 \,.$$

Pour construire des intégrales stochastiques, on considère d'abord le cas des fonctions étagées à support compact de la forme $\sum_{j\in\mathcal{J}} a_j \mathbb{1}_{B_j}$, avec $(B_j)_{j\in\mathcal{J}}$ une partition finie d'un compact, auxquelles on associe les grandeurs $\sum_j a_j \Phi(B_j)$. Par un critère de densité, on définit ensuite l'intégrale stochastique d'une fonction $f \in L^2(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \xi_\Phi)$ notée

$$\int_{\mathbb{R}^d} f(\boldsymbol{u}) d\Phi(\boldsymbol{u}) \,,$$

et telle que

$$\left\| \int_{\mathbb{R}^d} f(\boldsymbol{u}) d\Phi(\boldsymbol{u}) \right\|^2 = \int_{\mathbb{R}^d} |f(\boldsymbol{u})|^2 d\xi_\Phi(\boldsymbol{u}) \,.$$

(Une fonction f sur $\mathbb{R}^d$ est donc intégrable par rapport à une mesure aléatoire Φ si et seulement si $f \in L^2(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \xi_\Phi)$.)

Théorème 5 (Représentation spectrale). *Soit $F(\boldsymbol{x})$, $\boldsymbol{x} \in \mathbb{R}^d$, un processus aléatoire du second ordre de moyenne nulle, stationnaire et à covariance continue. ($F(\boldsymbol{x})$ est donc continu en moyenne quadratique, voir la section 2.4.4.) Il existe une mesure aléatoire Φ_F unique (à une équivalence près) telle que*

$$F(\boldsymbol{x}) = \int_{\mathbb{R}^d} e^{\imath(\boldsymbol{u},\boldsymbol{x})} d\Phi_F(\boldsymbol{u}) \,.$$

Démonstration. Voir (Rozanov, 1967). □

Remarque. Le théorème de Stone (Riesz et Nagy-Sz., 1965 ; Yosida, 1980) est un outil fondamental pour établir le théorème de représentation spectrale dans le cas de processus à paramètres continus. Son utilisation suppose la continuité de la covariance.

2.4.3 L'espace $\mathcal{H}$

Nous nous intéressons maintenant à l'espace généré par un processus $F(\boldsymbol{x})$, $\boldsymbol{x} \in \mathbb{R}^d$, du second ordre et de moyenne nulle. Notons $\tilde{\Lambda}$ l'espace vectoriel des mesures à support fini sur $\mathbb{R}^d$. Les éléments de $\tilde{\Lambda}$ sont de la forme $\lambda = \sum_{i=1}^n \lambda_i \delta_{\boldsymbol{x}_i}$, avec $\lambda_i \in \mathbb{R}$, $\forall i$, et $\delta_{\boldsymbol{x}_i}$ la mesure telle que $\delta_{\boldsymbol{x}_i}(B)$ vaut un si $\boldsymbol{x}_i \in B$, et zéro sinon. Considérons aussi $\tilde{\mathcal{H}}$ le sous-espace vectoriel de $L^2(\Omega, \mathcal{A}, \mathsf{P})$ généré par les combinaisons linéaires de variables aléatoires $F(\boldsymbol{x})$, $\boldsymbol{x} \in \mathbb{R}^d$.

Le processus aléatoire $F(\boldsymbol{x})$, vu plus haut comme une fonction $\mathbb{R}^d \to \tilde{\mathcal{H}}$, est étendu en considérant l'application *linéaire*

$$\begin{array}{rccc} F: & \tilde{\Lambda} & \to & \tilde{\mathcal{H}} \\ & \lambda = \sum_{i=1}^n \lambda_i \delta_{\boldsymbol{x}_i} & \mapsto & F(\lambda) = \sum_{i=1}^n \lambda_i F(\boldsymbol{x}_i) \,. \end{array} \tag{2.9}$$

Les éléments de $\tilde{\mathcal{H}}$ sont donc des éléments $F(\lambda)$, $\lambda \in \tilde{\Lambda}$. Remarquons que $F(\lambda)$ peut être vu comme une intégrale stochastique.

La covariance de $F(\boldsymbol{x})$ est étendue au processus $F(\lambda)$, $\lambda \in \tilde{\Lambda}$:

$$\begin{array}{rccc} k: & \tilde{\Lambda} \times \tilde{\Lambda} & \to & \mathbb{R} \\ & (\lambda, \mu) & \mapsto & k(\lambda, \mu) = \sum_{i,j=1}^{n,m} \lambda_i \mu_j k(\boldsymbol{x}_i, \boldsymbol{y}_j) \,. \end{array}$$

L'opérateur $k(\lambda, \mu)$ est bilinéaire, symétrique et positif sur $\tilde{\Lambda} \times \tilde{\Lambda}$. On peut encore considérer la covariance comme un opérateur sur $\tilde{\mathcal{H}} \times \tilde{\mathcal{H}}$ en posant $k(F(\lambda), F(\mu)) = k(\lambda, \mu)$. Si l'on requiert de plus que la covariance $k(\boldsymbol{x}, \boldsymbol{y})$ soit définie positive, alors $\|F(\lambda)\| = 0$ implique $\lambda = 0$ et $F(\lambda) = 0$. Dans ce cas, l'opérateur de covariance définit un produit scalaire sur $\tilde{\mathcal{H}}$.

Moyennant cette condition, $\tilde{\mathcal{H}}$ est préhilbertien. Soit $\mathcal{H}$ un complété de $\tilde{\mathcal{H}}$ pour le produit scalaire défini par k. $\mathcal{H}$ est un espace hilbertien appelé espace généré par $F(\boldsymbol{x})$, $\boldsymbol{x} \in \mathbb{R}^d$. De même, l'espace $\tilde{\Lambda}$, muni du produit scalaire $k(\lambda, \mu)$, est préhilbertien. Soit Λ un complété de $\tilde{\Lambda}$.

L'application linéaire F se prolonge sur Λ par continuité. Les espaces $\mathcal{H}$ et Λ ainsi construits sont isométriques.

Dans les paragraphes suivants, nous regardons les propriétés de $\mathcal{H}$ lorsque l'on suppose $F(\boldsymbol{x})$ stationnaire. Au processus aléatoire $F(\boldsymbol{x})$ correspond alors une mesure spectrale ξ_F, ainsi qu'une mesure aléatoire Φ_F telle que $F(\boldsymbol{x})$ admet la représentation spectrale

$$F(\boldsymbol{x}) = \int_{\mathbb{R}^d} e^{\imath(\boldsymbol{u},\boldsymbol{x})} d\Phi_F(\boldsymbol{u})\,.$$

La proposition suivante permet d'établir une correspondance entre les éléments de $\mathcal{H}$ et la représentation spectrale de $F(\boldsymbol{x})$.

Théorème 6. *L'opérateur linéaire*

$$\varrho: \quad \textstyle\sum_{i=1}^n \lambda_i e^{\imath(\boldsymbol{u},\boldsymbol{x}_i)} \quad \mapsto \quad F(\lambda)\,, \quad \lambda = \sum_{i=1}^n \lambda_i \delta_{\boldsymbol{x}_i}$$

se prolonge en un isomorphisme de $L^2(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \xi_F)$ dans $\mathcal{H}$

Démonstration. Voir (Rozanov, 1967). La preuve utilise le fait que l'ensemble des combinaisons linéaires $\sum_{i=1}^n \lambda_i e^{\imath(\boldsymbol{u},\boldsymbol{x}_i)}$ est un sous-ensemble dense de $L^2(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \xi_F)$. □

Grâce à ce résultat, on sait qu'à tout élément $X \in \mathcal{H}$ correspond une fonction $\varphi_X(\boldsymbol{u}) \in L^2(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \xi_F)$. Notons que $L^2(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \xi_F)$ est muni du produit scalaire

$$(\varphi_X, \varphi_Y)_{L^2(\xi_F)} = (X, Y)_{\mathcal{H}} = \int_{\mathbb{R}^d} \varphi_X(\boldsymbol{u}) \varphi_Y(\boldsymbol{u})^* d\xi_F(u)\,.$$

De plus (Rozanov, 1967), X et φ_X sont liés par

$$X = \int_{\mathbb{R}^d} \varphi_X(\boldsymbol{u}) d\Phi_F(\boldsymbol{u})\,. \tag{2.10}$$

La formule (2.10) nous donne par conséquent une *représentation spectrale* de tout élément appartenant à l'espace $\mathcal{H}$. La fonction φ_X dans (2.10) est appelée *caractéristique spectrale* de X.

Soit T un opérateur linéaire sur $\mathcal{H}$, alors l'équation

$$Y = TX = \int_{\mathbb{R}^d} \varphi_Y(\boldsymbol{u}) d\Phi_F(\boldsymbol{u}) = T\Big[\int_{\mathbb{R}^d} \varphi_X(\boldsymbol{u}) d\Phi_F(\boldsymbol{u})\Big] = \int_{\mathbb{R}^d} [T'\varphi_X](\boldsymbol{u}) d\Phi_F(\boldsymbol{u})$$

définit un opérateur linéaire T' sur $L^2(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \xi_F)$. Inversement, tout opérateur linéaire T' sur $L^2(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \xi_F)$ correspond à un opérateur linéaire T sur $\mathcal{H}$. Par exemple, l'opérateur de translation $\tau_{\boldsymbol{h}}$ dans $\mathcal{H}$ correspond à l'opérateur de multiplication par $e^{\imath(\boldsymbol{h},\boldsymbol{u})}$.

Nous nous intéressons plus spécifiquement maintenant aux opérations de filtrage linéaire.

Définition 19 (Filtrage linéaire). Un processus $F_1(\boldsymbol{x})$ $\boldsymbol{x} \in \mathbb{R}^d$ est un *filtrage linéaire* de $F(\boldsymbol{x})$ s'il peut être représenté sous la forme

$$F_1(\boldsymbol{x}) = \int_{\mathbb{R}^d} e^{\imath(\boldsymbol{u},\boldsymbol{x})} \varphi_1(\boldsymbol{u}) d\Phi_F(\boldsymbol{u})\,,$$

où $\varphi_1(\boldsymbol{u})$ est une fonction de $L^2(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \xi_F)$. Par conséquent, $F_1(\boldsymbol{x}) \in \mathcal{H}$, $\forall \boldsymbol{x} \in \mathbb{R}^d$.

Remarque. Il est important de se limiter aux fonctions de $L^2(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \xi_F)$ afin que le processus $F_1(\boldsymbol{x})$ soit du second ordre, car

$$\|F_1(\boldsymbol{x})\|^2 = \int_{\mathbb{R}^d} |\varphi_1(\boldsymbol{u})|^2 d\xi_F(\boldsymbol{u})\,.$$

Proposition 11. *$F(\boldsymbol{x})$ et $F_1(\boldsymbol{x})$ sont stationnairement corrélés.*

Démonstration.

$$(F_1(\boldsymbol{x}), F(\boldsymbol{y}))_{\mathcal{H}} = \int_{\mathbb{R}^d} e^{\imath(\boldsymbol{u},\boldsymbol{x})} \varphi_1(\boldsymbol{u}) e^{-\imath(\boldsymbol{u},\boldsymbol{y})} d\xi_F(\boldsymbol{u}) = \int_{\mathbb{R}^d} e^{\imath(\boldsymbol{u},\boldsymbol{x}-\boldsymbol{y})} \varphi_1(\boldsymbol{u}) d\xi_F(\boldsymbol{u})$$

ne dépend que de $\boldsymbol{x} - \boldsymbol{y}$. □

Remarquons qu'un processus aléatoire $F_1(\boldsymbol{x})$ stationnairement corrélé avec $F(\boldsymbol{x})$ est issu d'un filtrage linéaire de $F(\boldsymbol{x})$ si et seulement si il existe un $\boldsymbol{x}_0 \in \mathbb{R}^d$ tel que $F_1(\boldsymbol{x}_0) \in \mathcal{H}$; on a $F_1(\boldsymbol{x}) = \tau_{(\boldsymbol{x}-\boldsymbol{x}_0)} F_1(\boldsymbol{x}_0)$, $\forall \boldsymbol{x}$.

Par la suite, nous devrons déterminer la covariance du processus filtré ainsi que les intercovariances entre $F(\boldsymbol{x})$ et $F_1(\boldsymbol{x})$. On obtient classiquement les relations appelées *formules des interférences* (rappelées par exemple dans (Chonavel, 2002)).

Exemple : dérivée de $F(x)$. Considérons un processus aléatoire du second ordre $F(x)$, $x \in \mathbb{R}$. $F(x)$ est dérivable en moyenne quadratique si la limite de $\|h^{-1}(F(x+h) - F(x))\|_{\mathcal{H}}$ existe quand h tend vers 0 (voir la section 2.4.4). Comme

$$h^{-1}(F(x+h) - F(x)) = \int_{\mathbb{R}} h^{-1}(e^{\imath u(x+h)} - e^{\imath u x}) d\Phi_F(u)\,,$$

$F(x)$ est dérivable en moyenne quadratique si la fonction

$$h \mapsto h^{-1}(e^{\imath u(x+h)} - e^{\imath u x})$$

converge dans $L^2(\mathbb{R}, \mathcal{B}(\mathbb{R}), \xi_F)$. La dérivée de $F(x)$ admet alors la représentation

$$F'(x) = \int_{\mathbb{R}} \imath u e^{\imath u x} d\Phi_F(u)\,.$$

2.4.4 Propriétés sur des trajectoires

Différentes notions de continuité

Si la sortie effective du système à modéliser est continue en fonction des facteurs, il est bien sûr pertinent de choisir un modèle aléatoire ayant aussi des trajectoires continues. En pratique, on voudrait donc savoir si un modèle aléatoire donné par sa loi ou par ses moments a des propriétés de continuité. La notion de continuité pour un processus aléatoire peut toutefois être définie en plusieurs sens. Rappelons les différentes notions de continuité utilisées ainsi que leurs relations.

Définition 20 (Continuité des trajectoires). $F(\boldsymbol{x})$ possède des *trajectoires continues avec probabilité* 1 si

$$\mathsf{P}\{\lim_{\boldsymbol{x}\to\boldsymbol{x}_0} F(\boldsymbol{x}) - F(\boldsymbol{x}_0) = 0\,, \quad \forall \boldsymbol{x}_0 \in \mathbb{X}\} = 1$$

La continuité des trajectoires est souvent la propriété désirée pour un modèle de type boîte noire. Cependant, il s'agit d'une propriété forte. Des notions de continuité plus faibles existent.

Définition 21 (Continuité en probabilité). Un processus aléatoire $F(\boldsymbol{x})$ est dit *continu en probabilité* en $\boldsymbol{x}_0$ si $F(\boldsymbol{x})$ converge en probabilité vers $F(\boldsymbol{x}_0)$ quand $\boldsymbol{x}$ tend vers $\boldsymbol{x}_0$, c'est à dire si

$$\lim_{\boldsymbol{x}\to\boldsymbol{x}_0} \mathsf{P}\{|F(\boldsymbol{x}) - F(\boldsymbol{x}_0)| > \varepsilon\} = 0 \quad \forall \varepsilon > 0\,.$$

Définition 22 (Continuité presque sûre). Un processus aléatoire $F(\boldsymbol{x})$ est dit *continu presque sûrement* en $\boldsymbol{x}_0$ si $F(\boldsymbol{x})$ converge presque sûrement vers $F(\boldsymbol{x}_0)$ quand $\boldsymbol{x}$ tend vers $\boldsymbol{x}_0$, c'est à dire si

$$\mathsf{P}\{\lim_{\boldsymbol{x}\to\boldsymbol{x}_0} F(\boldsymbol{x}) = F(\boldsymbol{x}_0)\} = 1$$

Définition 23 (Continuité en moyenne quadratique). Un processus aléatoire $F(\boldsymbol{x})$ est dit *continu en moyenne quadratique* en $\boldsymbol{x}_0$ si $F(\boldsymbol{x})$ converge en moyenne quadratique vers $F(\boldsymbol{x}_0)$ quand $\boldsymbol{x}$ tend vers $\boldsymbol{x}_0$, c'est à dire si

$$\lim_{\boldsymbol{x}\to\boldsymbol{x}_0} \mathrm{E}[(F(\boldsymbol{x}) - F(\boldsymbol{x}_0))^2] = 0$$

Pour chacune des définitions 21 à 23, un processus est continu sur $\mathbb{X}$ s'il est continu pour tout $\boldsymbol{x} \in \mathbb{X}$. Les définitions 21 à 23 correspondent bien à des notions plus faibles que la définition 20, et qui ne garantissent pas la continuité des trajectoires. Ainsi dans l'exemple de la section 2.1.1, le processus $F_2(\boldsymbol{x})$ est continu en probabilité sur $[0, 1]$, puisque

$$\mathsf{P}\{|F(x) - F(x_0)| > \varepsilon\} = \mathsf{P}\{w = x\} + \mathsf{P}\{w = x_0\} = 0\,,$$

alors que les trajectoires de $F_2(x)$ ne sont pas continues. On peut vérifier que $F_2(x)$ est encore continu presque sûrement et en moyenne quadratique.

Les relations bien connues entre convergences impliquent les relations correspondantes entre les continuités. Ainsi, la continuité en moyenne quadratique n'implique pas en général la continuité presque sûre et réciproquement. Par contre, les continuités presque sûre et en moyenne quadratique impliquent la continuité en probabilité.

Lorsque l'on dispose d'hypothèses supplémentaires sur le processus aléatoire considéré, il est possible de trouver des relations supplémentaires entre les continuités. Nous souhaitons avant tout présenter l'influence des moments d'ordre deux sur les continuités. Ce point est important, car dans le cadre de la prédiction linéaire, ce sont les moments d'ordre deux que nous aurons à choisir.

Proposition 12. *Soit $F(\boldsymbol{x})$ un processus aléatoire du second ordre à moyenne continue. Alors $F(\boldsymbol{x})$ est continu en moyenne quadratique en $\boldsymbol{x}_0$ si et seulement si sa covariance $k(\boldsymbol{x}, \boldsymbol{y})$ est continue en $(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x}_0, \boldsymbol{x}_0)$. Si $k(\boldsymbol{x}, \boldsymbol{y})$ est continue sur sa diagonale $\boldsymbol{x} = \boldsymbol{y}$, alors elle est continue partout.*

Démonstration. Après avoir soustrait la moyenne :

$$\begin{aligned} |k(\boldsymbol{x}, \boldsymbol{y}) - k(\boldsymbol{x}_0, \boldsymbol{x}_0)| &= \mathrm{E}[F(\boldsymbol{x})F(\boldsymbol{y}) - F(\boldsymbol{x}_0)^2] \\ &= \mathrm{E}[(F(\boldsymbol{x}) - F(\boldsymbol{x}_0))(F(\boldsymbol{y}) - F(\boldsymbol{x}_0)] \\ &\quad - \mathrm{E}[F(\boldsymbol{x}_0)(F(\boldsymbol{x}_0) - F(\boldsymbol{y}))] + \mathrm{E}[F(\boldsymbol{x}_0)(F(\boldsymbol{x}) - F(\boldsymbol{x}_0))]\,. \end{aligned}$$

Si $F(\boldsymbol{x})$ est continu en moyenne quadratique en $\boldsymbol{x}_0$, les trois termes de droite tendent vers 0 lorsque $(\boldsymbol{x}, \boldsymbol{y}) \to (\boldsymbol{x}_0, \boldsymbol{x}_0)$, d'après l'inégalité de Schwarz, et donc $k(\boldsymbol{x}, \boldsymbol{y})$ est continue en $(\boldsymbol{x}_0, \boldsymbol{x}_0)$. La réciproque est immédiate.

Nous ne donnons pas la preuve de la deuxième partie de la proposition. Voir (Stein, 1999) dans le cas stationnaire et (Adler, 1981) pour le cas général. □

Dans le cas de processus stationnaires, une relation très importante existe entre la continuité en moyenne quadratique et la mesure spectrale. On peut en effet utiliser le fait qu'il existe une dépendance entre la régularité d'une fonction $f : \mathbb{R} \to \mathbb{R}$ et la décroissance de sa transformée de Fourier $\tilde{f}(u)$ quand u tend vers l'infini pour obtenir la proposition suivante.

Proposition 13 (Régularité et décroissance (en dimension 1)). *Une fonction intégrable $f : \mathbb{R} \to \mathbb{R}$ est bornée avec ses dérivées jusqu'à l'ordre p continues et bornées si*

$$\int_{\mathbb{R}} |\tilde{f}(u)|(1 + |u|^p)du < +\infty\,. \tag{2.11}$$

Démonstration. Si $\tilde{f}(u) \in L^1(\mathbb{R})$, alors on vérifie à l'aide de la formule de Fourier inverse que f est continue et bornée :

$$|f(x)| \leq \frac{1}{2\pi} \int_{\mathbb{R}} |e^{\imath ux} \tilde{f}(u)\ |du = \frac{1}{2\pi} \int_{\mathbb{R}} |\tilde{f}(u)\ |du \tag{2.12}$$

et

$$|f(x) - f(y)| \leq \frac{1}{2\pi} \int_{\mathbb{R}} |(e^{\imath ux} - e^{\imath uy})\tilde{f}(u)\ |du \overset{x \to y}{\longrightarrow} 0$$

d'après le théorème de convergence dominée.
La transformée de Fourier de la dérivée de f à l'ordre k est $(\imath u)^k \tilde{f}(u)$. D'après (2.11) et (2.12), $\int_{\mathbb{R}} |\tilde{f}(u)||u|^k du\ < +\infty$ pour tout $k \leq p$, ce qui montre que $f^{(k)}(x)$ est continue et bornée. □

Ainsi, si la densité spectrale $d\xi_F(u)/du$ de $F(x)$, $x \in \mathbb{R}$, est telle que l'on ait

$$|d\xi_F(u)/du| \leq \frac{K}{1 + |u|^{1+\varepsilon}}$$

avec K et $\varepsilon > 0$, alors $F(x)$ est continue en moyenne quadratique. Le résultat se généralise aux ordres supérieurs (voir le paragraphe *Différentiabilité*, page 45).

Théorèmes généraux impliquant la continuité des trajectoires

Nous avons mentionné plus haut que la continuité en moyenne quadratique n'implique pas en général la continuité des trajectoires. Il existe de nombreux résultats permettant de garantir la continuité des trajectoires à partir de conditions sur les lois des processus ou même sur les moments (nous avons vu précédemment que de telles propriétés, qui impliquent des ensembles non dénombrables de variables aléatoires, ne peuvent être établies qu'à une modification près des processus aléatoires, ce que nous sous-entendrons par la suite). Les approches modernes de ces questions fournissent des résultats fins mais dont les preuves sont souvent très techniques (Dudley, 1973 ; Talagrand, 1987 ; Adler, 1990). Pour établir la continuité des trajectoires, on cherche souvent à majorer des queues de distribution de probabilité. Nous rappelons seulement

ci-dessous quelques conditions suffisantes de continuité des trajectoires qui sont les plus utiles en pratique. Des démonstrations peuvent par exemple être trouvées dans (Belyaev, 1961 ; Cramér et Leadbetter, 1967 ; Adler, 1981).

Proposition 14 (processus sur $\mathbb{R}$). *S'il existe des constantes $C > 0$ et $\eta > \alpha > 0$, telles que pour $h > 0$ suffisamment petit,*

$$\mathrm{E}\left[|F(x+h) - F(x)|^{\alpha}\right] \leq \frac{C|h|}{|\log|h||^{1+\eta}},$$

alors il existe une modification (définition page 25) de $F(x)$ qui possède des trajectoires continues avec probabilité 1. Notons que de nombreux processus satisfont une inégalité plus forte du type

$$\mathrm{E}\left[|F(x+h) - F(x)|^{\alpha}\right] \leq C|h|^{1+\eta}$$

avec $\alpha > 0$ et $\eta > 0$.

Démonstration. Sans donner la démonstration (Cramér et Leadbetter, 1967), il peut être intéressant d'indiquer de manière informelle que l'on commence par obtenir une condition suffisante de continuité des trajectoires en majorant des queues de probabilité : il existe deux fonctions $\varepsilon(h)$ et $g(h)$ satisfaisant certaines conditions telle que, pour tout $\boldsymbol{h}$ tel que $\|\boldsymbol{h}\| \leq h$,

$$\mathsf{P}\{|F(\boldsymbol{x}+\boldsymbol{h}) - F(\boldsymbol{x})| > \varepsilon(h)\} \leq g(h).$$

On applique ensuite l'inégalité de Markov

$$\mathsf{P}\{|F(\boldsymbol{x}+\boldsymbol{h}) - F(\boldsymbol{x})| > \varepsilon(\|\boldsymbol{h}\|)\} \leq \frac{\mathrm{E}[|F(\boldsymbol{x}+\boldsymbol{h}) - F(\boldsymbol{x})|^{\alpha}]}{|\varepsilon(\|\boldsymbol{h}\|)|^{\alpha}}$$

avec une fonction $\varepsilon(h)$ correctement choisie. □

Proposition 15 (processus sur $\mathbb{R}^d$). *Soit $F(\boldsymbol{x})$, $\boldsymbol{x} \in \mathbb{R}^d$ un processus aléatoire. S'il existe des constantes $C > 0$ et $\eta > \alpha > 0$ telles que dans un voisinage de $\|\boldsymbol{h}\| = 0$,*

$$\mathrm{E}\left[|F(\boldsymbol{x}+\boldsymbol{h}) - F(\boldsymbol{x})|^{\alpha}\right] \leq \frac{C\|\boldsymbol{h}\|^{2d}}{|\log\|\boldsymbol{h}\||^{1+\eta}},$$

alors $F(\boldsymbol{x})$ possède des trajectoires continues sur tout intervalle compact de $\mathbb{R}^d$ avec probabilité 1.

Démonstration. Ce résultat (plus faible pour $d = 1$ que la proposition 14) est donné sans preuve dans Adler (1981) d'après Belyaev (1972). □

Exemple. On sait que le mouvement brownien possède des trajectoires continues avec probabilité 1. Les incréments $W(t+h) - W(t)$ sont des variables aléatoires gaussiennes centrées de variance h. Par conséquent, $\mathrm{E}[|W(t+h) - W(t)|^p] = C_p|h|^{p/2}$, où C_p est le moment d'ordre p d'une variable aléatoire de loi gaussienne centrée de variance unité. On peut vérifier que, pour h suffisamment petit,

$$\mathrm{E}\left[|W(t+h) - W(t)|^4\right] = C_4|h|^2 < C_4|h|/|\log(|h|)|^6$$

et donc la condition suffisante de la proposition 14 est vérifiée pour $\alpha = 4$.

Cette condition suffisante est toutefois restrictive si on ne connaît que des moments d'ordre deux d'un processus aléatoire. Pour un processus stationnaire avec $\alpha = 2$, par exemple, le comportement de la borne impose une fonction de covariance avec une pente tendant relativement rapidement vers 0 au voisinage de l'origine (en $|h|^2/|\log(|h|)|^\beta$). Pour les processus aléatoires gaussiens, il est possible d'obtenir une condition suffisante beaucoup moins restrictive donnée par la proposition suivante.

Proposition 16. *Si $F(\boldsymbol{x})$ est un processus aléatoire gaussien du second ordre, de moyenne nulle et admettant une covariance continue, et s'il existe des constantes $C > 0$ et $\varepsilon > 0$, telles que pour $\|\boldsymbol{x} - \boldsymbol{y}\| < 1$*

$$\mathrm{E}[|F(\boldsymbol{x}) - F(\boldsymbol{y})|^2] \leq \frac{C}{|\log\|\boldsymbol{x} - \boldsymbol{y}\|\,|^{1+\varepsilon}}\,, \tag{2.13}$$

alors $F(\boldsymbol{x})$ admet des trajectoires continues avec probabilité 1.

Démonstration. Voir (Adler, 1981) pour une preuve assez technique. □

(D'autres résultats concernant le cas gaussien pourront être trouvés dans (Cramér et Leadbetter, 1967 ; Belyaev, 1972 ; Dudley, 1973).) Si l'on suppose de plus $F(\boldsymbol{x})$ stationnaire, la condition (2.13) peut se réécrire simplement à l'aide de la fonction de corrélation $\rho(\boldsymbol{h})$ de $F(\boldsymbol{x})$: s'il existe $C > 0$ et $\varepsilon > 0$, telle que pour $\|\boldsymbol{h}\| < 1$

$$1 - \rho(\boldsymbol{h}) \leq \frac{C}{|\log\|\boldsymbol{h}\||^{1+\varepsilon}}\,, \tag{2.14}$$

alors $F(\boldsymbol{x})$ admet des trajectoires continues avec probabilité un. La borne de (2.14) est une fonction croissante en $\|\boldsymbol{h}\|$ tendant vers zéro à l'origine et vers l'infini lorsque $\|\boldsymbol{h}\|$ tend vers un. Remarquons qu'un changement d'échelle n'a pas d'influence sur les propriétés de continuité et que la condition $\|\boldsymbol{h}\| < 1$ pourrait éventuellement être remplacée par $\|\boldsymbol{h}\| < \eta$, avec η arbitrairement petit ($0 < \eta < 1$). Par conséquent, la condition (2.14) porte sur la vitesse de convergence de la corrélation à l'origine. La figure 2.2 représente la borne $|\log h|^{-(1+\varepsilon)}$ pour différentes valeurs de ε. Cette figure suggère que la pente de la borne croît très rapidement quand h se rapproche de zéro et on peut vérifier par le calcul que la dérivée tend vers l'infini quand h tend vers zéro.

En se fondant sur cette observation, (Abrahamsen, 1997) affirme que toutes les fonctions de covariances continues utilisées en pratique respectent cette borne. *En pratique*, il est par conséquent possible de considérer qu'un processus aléatoire gaussien à moyenne et à covariance continue possède des trajectoires continues avec probabilité un.

Dérivabilité

Nous nous intéressons dans cette section aux propriétés des dérivées des trajectoires d'un processus aléatoire. En automatique, par exemple, il est fréquent de devoir utiliser les dérivées des sorties d'un système. Un modèle boîte noire de ce système doit donc permettre une estimation de ces dérivées. Dans la théorie des processus aléatoires, les notions de dérivabilité se traitent de la même façon que les notions de continuité. Les notions les plus intéressantes à définir sont celles de la dérivabilité en moyenne quadratique et dérivabilité des trajectoires avec probabilité un.

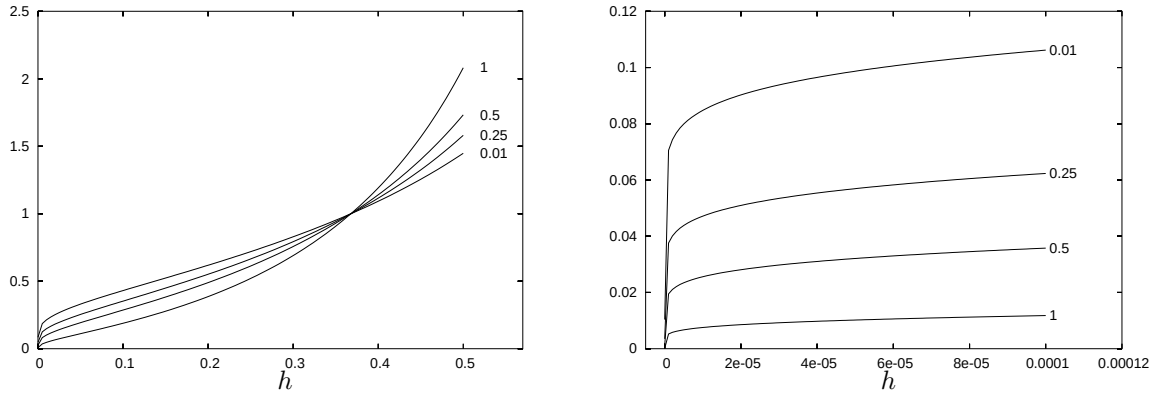


FIG. 2.2 – Fonction $|\log h|^{-(1+\varepsilon)}$ tracée pour plusieurs valeurs de ε (0.01, 0.25, 0.5 et 1). La figure de droite est un agrandissement autour de l'origine de la figure de gauche.

Soit $F(\boldsymbol{x})$, $\boldsymbol{x} \in \mathbb{R}^d$, un processus aléatoire du second ordre. Considérons le processus $F_{\boldsymbol{h},t}(\boldsymbol{x})$, $\boldsymbol{h} \in \mathbb{R}^d$, $t \in \mathbb{R}$, défini par

$$F_{\boldsymbol{h},t}(\boldsymbol{x}) = \frac{F(\boldsymbol{x}+t\boldsymbol{h}) - F(\boldsymbol{x})}{t}$$

Le processus $F(\boldsymbol{x})$ est dérivable en moyenne quadratique en $\boldsymbol{x}_0$ dans la direction $\boldsymbol{h}$ si $F_{\boldsymbol{h},t}(\boldsymbol{x}_0)$ converge en moyenne quadratique lorsque t tend vers zéro. Si la limite existe quelle que soit la direction $\boldsymbol{h}$, on dira que le processus est dérivable en moyenne quadratique. De même, on dira que $F(\boldsymbol{x})$ possède des trajectoires dérivables en $\boldsymbol{x}_0$ avec probabilité un, s'il existe un ensemble N de mesure négligeable telle pour tout $\omega \in \Omega\backslash N$ et $\forall \boldsymbol{h} \in \mathbb{R}^d$, la limite de $F_{\boldsymbol{h},t}(\omega, \boldsymbol{x}_0)$ existe lorsque t tend vers zéro. On montre aisément que si les limites $\lim_{t\to 0} F_{\boldsymbol{h},t}(\boldsymbol{x}_0)$ existent à la fois en moyenne quadratique et avec probabilité un, alors les deux limites sont égales.

Si $F(\boldsymbol{x})$ est dérivable en $\boldsymbol{x}_0$, nous notons $\nabla F(\boldsymbol{x})$ le processus aléatoire *gradient* à valeurs vectorielles dans $\mathbb{R}^d$ tel que, $\forall \boldsymbol{h}$,

$$\lim_{t\to 0} F_{\boldsymbol{h},t}(\boldsymbol{x}_0) = (\nabla F(\boldsymbol{x}_0), \boldsymbol{h})\,.$$

Dans le cas d'un processus défini sur $\mathbb{R}$, le processus dérivé sera noté $F'(x)$ ou $F^{(1)}(x)$.

Les moments du second ordre du processus gradient $\nabla F(\boldsymbol{x})$ (sous condition d'existence de celui-ci) se déduisent simplement des moments de $F(\boldsymbol{x})$. Si l'on pose $m(\boldsymbol{x}) = \mathrm{E}[F(\boldsymbol{x})]$, alors

$$\mathrm{E}[\nabla F(\boldsymbol{x})] = \nabla m(\boldsymbol{x})\,.$$

Le vecteur d'inter-covariances entre $F(\boldsymbol{x})$ et $\nabla F(\boldsymbol{x})$ est tel que

$$\mathrm{Cov}[F(\boldsymbol{x}), \nabla F(\boldsymbol{y})] = \nabla_{\boldsymbol{y}} k(\boldsymbol{x}, \boldsymbol{y})\,,$$

et les éléments de la matrice d'inter-covariances $\boldsymbol{K}^{(2)}$ entre les composantes de $\nabla F(\boldsymbol{x})$ valent

$$k^{(2)}_{i,j} = \frac{\partial^2}{\partial x_{[i]} \partial y_{[j]}} k(\boldsymbol{x}, \boldsymbol{y})\,.$$

Nous ne démontrons pas ces expressions mais dans le cas où $x \in \mathbb{R}$ et $F(x)$ est stationnaire, on peut noter que le processus $F_t(x) = (F(x+t) - F(x))/t$ a pour fonction de covariance

$$k_t(h) = \frac{1}{h^2}(2k(h) - k(h+t) - k(h-t))\,,$$

et que l'on a bien $\lim_{t\to 0} k_t(h) = -k^{(2)}(h)$.

Les conditions de dérivabilité des trajectoires de $F(\boldsymbol{x})$ avec probabilité 1 s'obtiennent de la même manière que les conditions de continuité des trajectoires, en transposant les résultats précédents aux composantes du vecteur gradient $\nabla F(\boldsymbol{x})$. Il n'est pas nécessaire d'expliciter cette généralisation.

Une condition suffisante de dérivabilité en moyenne quadratique est rappelée dans la proposition suivante.

Proposition 17. *Soit $F(\boldsymbol{x})$, $\boldsymbol{x} \in \mathbb{R}^d$, un processus aléatoire du second ordre admettant la fonction de covariance $k(\boldsymbol{x}, \boldsymbol{y})$. Si les dérivées partielles $\partial^2 k(\boldsymbol{x}, \boldsymbol{y})/\partial x_{[i]}\partial y_{[i]}$, $i = 1, \ldots, d$, existent et sont finies en $(\boldsymbol{x}_0, \boldsymbol{x}_0)$, alors $F(\boldsymbol{x})$ est dérivable en moyenne quadratique en $\boldsymbol{x}_0$.*

Démonstration. Voir (Cramér et Leadbetter, 1967) par exemple. □

Dans le cas $x \in \mathbb{R}$, on peut remarquer le lien entre dérivabilité (à l'ordre p) en moyenne quadratique et la décroissance de la densité spectrale. D'après la proposition 13, si

$$\Big|\frac{d\xi_F(u)}{du}\Big| \leq \frac{K}{1 + |u|^{2p+\varepsilon}}$$

avec K et $\varepsilon > 0$, alors $F(x)$ est dérivable à l'ordre p en moyenne quadratique.

2.5 Prédiction linéaire des processus du second ordre de moyenne nulle, krigeage

Dans les sections précédentes, nous avons présenté une synthèse des éléments classiques de la théorie des processus aléatoires. Par la suite, nous allons modéliser la sortie d'un système par un processus aléatoire paramétré par un ensemble de facteurs caractérisant l'état du système. Pour le moment, nous traitons le cas d'un processus à moyenne nulle ou connue (le cas général sera vu au chapitre 4). Ceci nous permet de spécifier les propriétés des trajectoires du processus aléatoire en choisissant essentiellement une fonction de covariance. Dans ce type de modèle, nous supposons que la sortie du système correspond à une trajectoire particulière du processus aléatoire. Cette trajectoire, qui n'est généralement observée que partiellement, doit donc être prédite afin d'utiliser le modèle quelles que soient les valeurs des facteurs.

La prédiction linéaire est une méthode des plus simples mais se révèle satisfaisante dans la plupart des cas rencontrés. Le krigeage, qui est en soit une prédiction linéaire, a été développé en géostatistique avec un point de vue très proche du problème de modélisation de système que nous nous posons. Nous présentons dans un premier temps les aspects élémentaires du krigeage, puis les notions permettant de modéliser des systèmes à plusieurs sorties observées en présence de bruit. Nous présentons enfin différentes mises en œuvre de la méthode, comme le krigeage dual ou une méthode récursive inspirée d'un algorithme de prédiction de série chronologique.

2.5.1 Krigeage

Prédiction

Dans cette section, nous introduisons le principe de la prédiction linéaire du point de vue de la *projection de meilleure approximation* sur le sous-espace vectoriel généré par les variables aléatoires observées. Il s'agit donc de s'intéresser aux estimateurs linéaires optimaux en moyenne quadratique. Il existe d'autres points du vue permettant de retrouver les équations de la prédiction linéaire optimale. En particulier, si l'on considère des processus gaussiens, le point de vue bayésien et le point de vue du maximum de vraisemblance conduisent aux mêmes équations (Williams et Rasmussen, 1996 ; Williams, 1997). Plusieurs approches déterministes sont aussi envisageables. Mentionnons l'approche d'optimisation directe des poids (Roll et al., 2003) et l'approche de la régression linéaire régularisée (Nashed et Wahba, 1974 ; Nashed, 1976 ; Tikhonov et Arsenin, 1977). Le point de vue de la régression linéaire régularisée fournit une interprétation fondamentale étudiée au chapitre 3.

Soit $(\Omega, \mathcal{A}, \mathsf{P})$ un espace de probabilité. On observe un processus aléatoire du second ordre $F(\boldsymbol{x})$. Supposons de plus que le processus soit à moyenne $m(\boldsymbol{x})$ *connue*. Dans ce cas, il revient au même de s'intéresser uniquement aux processus à moyenne nulle, puisque l'on peut toujours soustraire la moyenne à $F(\boldsymbol{x})$ et raisonner sur le processus centré. Soient $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ des points du domaine $\mathbb{X}$ correspondant à des observations du processus aléatoire $F(\omega, \boldsymbol{x}_1) = f_1^{\text{obs}}, \ldots, F(\omega, \boldsymbol{x}_n) = f_n^{\text{obs}}$, pour ω fixé. On souhaite prédire la valeur du processus en un point $\boldsymbol{x}$ du domaine $\mathbb{X}$. *Prédire* revient à chercher une fonction des variables aléatoires $F(\boldsymbol{x}_1), \ldots, F(\boldsymbol{x}_n)$ pour approximer $F(\omega, \boldsymbol{x})$, ω étant inconnu. Nous utiliserons le terme de *prédicteur* pour désigner la variable aléatoire définie sur $(\Omega, \mathcal{A}, \mathsf{P})$, fonction de $F(\boldsymbol{x}_1), \ldots, F(\boldsymbol{x}_n)$ et notée $\hat{F}(\boldsymbol{x})$, qui approxime $F(\boldsymbol{x})$. La valeur prédite est la réalisation de $\hat{F}(\boldsymbol{x})$ pour $\omega \in \Omega$ correspondant aux observations $f_1^{\text{obs}}, \ldots, f_n^{\text{obs}}$ (la valeur prédite est déterministe mais ω reste inconnu).

Espérance conditionnelle

Il est naturel de chercher un prédicteur qui minimise l'erreur $F(\boldsymbol{x}) - \hat{F}(\boldsymbol{x})$ en un certain sens. Un critère classique d'optimalité est la minimisation de la moyenne quadratique de l'erreur de prédiction. On cherche alors une fonction mesurable $\hat{F}(\boldsymbol{x}) = \hat{h}(F(\boldsymbol{x}_1), \ldots, F(\boldsymbol{x}_n))$, du second ordre et telle que

$$\mathrm{E}[(F(\boldsymbol{x}) - \hat{F}(\boldsymbol{x}))^2] = \int_\Omega (F(\boldsymbol{x}) - \hat{F}(\boldsymbol{x}))^2 d\mathsf{P}(\omega)$$

soit la plus petite possible. Comme le processus est de moyenne nulle, $\mathrm{E}[(F(\boldsymbol{x}) - \hat{F}(\boldsymbol{x}))^2] = \mathrm{Var}[(F(\boldsymbol{x}) - \hat{F}(\boldsymbol{x}))]$, et le prédicteur cherché est aussi à variance minimale.

Notons $\mathcal{S} = \sigma(F(\boldsymbol{x}_1), \ldots, F(\boldsymbol{x}_n))$, la σ-algèbre générée par les variables aléatoires $F(\boldsymbol{x}_1), \ldots, F(\boldsymbol{x}_n)$ [3]. Chercher $h(F(\boldsymbol{x}_1), \ldots, F(\boldsymbol{x}_n))$ dans $L^2(\Omega, \mathcal{A}, \mathsf{P})$, mesurable, revient à chercher une variable aléatoire de $L^2(\Omega, \mathcal{A}, \mathsf{P})$, $\mathcal{S}$-mesurable et la plus proche possible de $F(\boldsymbol{x})$ en norme L^2. Il est bien connu que la solution est donnée par l'espérance de $F(\boldsymbol{x})$ conditionnellement à $\mathcal{S}$ dans

[3] $\mathcal{S}$ porte l'information des variables aléatoires $F(\boldsymbol{x}_1), \ldots, F(\boldsymbol{x}_n)$, puisqu'il s'agit de l'ensemble des événements dont l'observation de ces variables permet de savoir s'ils ont eu lieu.

$L^2(\Omega, \mathcal{A}, \mathsf{P})$, notée $\mathrm{E}[F(\boldsymbol{x}) \mid \mathcal{S}]$ [4]. Autrement dit,

$$\hat{h}(F(\boldsymbol{x}_1), \ldots, F(\boldsymbol{x}_n)) = \mathrm{E}[F(\boldsymbol{x}) \mid \mathcal{S}]$$

est la meilleure approximation dans $L^2(\Omega, \mathcal{A}, \mathsf{P})$ de $F(\boldsymbol{x})$ par une fonction de $F(\boldsymbol{x}_1), \ldots, F(\boldsymbol{x}_n)$.

L'opérateur $F(\boldsymbol{x}) \mapsto \mathrm{E}[F(\boldsymbol{x}) \mid \mathcal{S}]$ est donc le projecteur orthogonal dans $L^2(\Omega, \mathcal{A}, \mathsf{P})$ sur le sous-espace vectoriel $L^2(\Omega, \mathcal{S}, \mathsf{P}_{\mathcal{S}})$, où $\mathsf{P}_{\mathcal{S}}$ est la restriction de P à $\mathcal{S}$.

Malheureusement, il n'existe pas d'algorithme efficace permettant de calculer cette espérance conditionnelle pour un processus aléatoire dans le cas général. On peut en revanche restreindre la classe des fonctions h pour obtenir des expressions analytiques simples. Par la suite, nous considérons le cas du krigeage, où l'on se restreint à la classe des combinaisons linéaires des $F(\boldsymbol{x}_i)$ [5].

Prédicteur linéaire

Considérons la classe des prédicteurs linéaires, en notant bien que cette restriction n'implique pas que le modèle boîte noire obtenu soit linéaire en les facteurs. La classe des prédicteurs linéaires est d'ailleurs suffisamment riche pour la plupart des applications usuelles et possède l'avantage d'être extrêmement simple à mettre en œuvre (voir le chapitre 6 pour des exemples d'application du krigeage). Parfois, la classe des prédicteurs linéaires est encore trop vaste, notamment lorsque l'on dispose de beaucoup de données et que l'on souhaite obtenir des solutions dites *creuses* (voir la section 6.8 pour quelques développements sur ce sujet).

Puisque le prédicteur cherché est linéaire, il est de la forme

$$\hat{F}(\boldsymbol{x}) = \sum_{i=1}^{n} \widehat{\lambda}_{i,\boldsymbol{x}} F(\boldsymbol{x}_i)\,.$$

Le meilleure prédicteur $\hat{F}(\boldsymbol{x})$ est le projeté orthogonal sur le sous-espace vectoriel

$$\mathcal{H}_S = \mathrm{vect}\{F(x_i), i = 1, \ldots, n\}\,,$$

et on constate que les conditions d'orthogonalité s'expriment en fonction des moments d'ordre deux seulement. En effet,

$$(F(\boldsymbol{x}) - \hat{F}(\boldsymbol{x}), F(\boldsymbol{x}_i)) = 0 \Rightarrow k(\boldsymbol{x}, \boldsymbol{x}_i) - \sum_{j=1}^{n} \widehat{\lambda}_{j,\boldsymbol{x}} k(\boldsymbol{x}_j, \boldsymbol{x}_i) = 0\,,$$

[4] Pour mémoire (Gikhman et Skorohod, 1974), l'espérance conditionnelle $\mathrm{E}[F(\boldsymbol{x}) \mid \mathcal{S}]$ dans $L^1(\Omega, \mathcal{A}, \mathsf{P})$, où $\mathcal{S}$ est une sous-σ-algèbre de $\mathcal{A}$, est par définition l'unique variable aléatoire de $L^1(\Omega, \mathcal{A}, \mathsf{P})$, $\mathcal{S}$-mesurable, telle que $\forall B \in \mathcal{S}$,

$$\int_B F(\boldsymbol{x}) d\mathsf{P} = \int_B \mathrm{E}[F(\boldsymbol{x}) \mid \mathcal{S}] d\mathsf{P}\,.$$

Si de plus $F(\boldsymbol{x}) \in L^2(\Omega, \mathcal{A}, \mathsf{P})$, pour toute variable aléatoire $Y \in L^2(\Omega, \mathcal{A}, \mathsf{P})$ $\mathcal{S}$-mesurable, on a

$$\mathrm{E}[Y F(x)] = \mathrm{E}\left[Y\, \mathrm{E}[F(\boldsymbol{x}) \mid \mathcal{S}]\right].$$

[5] Nous excluons ici la présentation d'autres classes de prédicteurs. Deux raisons expliquent ce choix. D'une part, la classe des prédicteurs linéaires se révèle satisfaisante dans la plupart des applications à traiter. D'autre part, l'étape essentielle qui conditionne la qualité d'un modèle est le choix correct la structure du second ordre du processus $F(\boldsymbol{x})$. Le recours à des prédicteurs non-linéaires est toutefois envisageable pour des besoins spécifiques (par exemple, des problèmes de seuillages). Une approche proposée en géostatistique consiste à transformer les données afin d'utiliser des prédicteurs linéaires (technique d'anamorphose, voir Chilès et Delfiner (1999)).

pour $i = 1, \ldots, n$. Par conséquent, on obtient un système d'équations linéaires en $\widehat{\lambda}_{j,\boldsymbol{x}}$, $j = 1, \ldots, n$.

Équations du krigeage

En réécrivant ce système linéaire sous une forme matricielle, on reconnaît les équations du krigeage (dit *simple* dans la littérature géostatistique).

Proposition 18 (krigeage). *Soit $F(\boldsymbol{x})$ un processus aléatoire du second-ordre, de moyenne nulle, admettant la fonction de covariance $k(\boldsymbol{x}, \boldsymbol{y})$. La meilleure prédiction linéaire de $F(\boldsymbol{x})$ à partir des variables aléatoires $F(\boldsymbol{x}_i)$, $i = 1, \ldots, n$, est la projection orthogonale $\hat{F}(\boldsymbol{x})$ sur $\mathcal{H}_S = \mathrm{vect}\{F(x_i); i = 1, \ldots, n\}$, qui s'écrit*

$$\hat{F}(\boldsymbol{x}) = \sum_{i=1}^{n} \widehat{\lambda}_{i,\boldsymbol{x}} F(\boldsymbol{x}_i)\,. \tag{2.15}$$

Les $\widehat{\lambda}_{i,\boldsymbol{x}}$ s'obtiennent en résolvant le système linéaire

$$\boldsymbol{K}_S \widehat{\boldsymbol{\lambda}}_{S,\boldsymbol{x}} = \boldsymbol{k}_{S,\boldsymbol{x}}\,, \tag{2.16}$$

où $\boldsymbol{K}_S$ est la matrice de covariance de taille $n \times n$ du vecteur aléatoire des observations $\boldsymbol{F}_S = (F(\boldsymbol{x}_1), \ldots, F(\boldsymbol{x}_n))^\mathsf{T}$, $\widehat{\boldsymbol{\lambda}}_{S,\boldsymbol{x}} = (\widehat{\lambda}_{1,\boldsymbol{x}}, \ldots, \widehat{\lambda}_{n,\boldsymbol{x}})^\mathsf{T}$ est le vecteur de $\mathbb{R}^n$ des coefficients du krigeage, et $\boldsymbol{k}_{S,\boldsymbol{x}} = (k(\boldsymbol{x}, \boldsymbol{x}_1), \ldots, k(\boldsymbol{x}, \boldsymbol{x}_n))^\mathsf{T}$ est le vecteur de $\mathbb{R}^n$ des covariances entre les variables observées et la variable aléatoire à prédire.

Nous omettrons généralement l'indice S par la suite et noterons $\boldsymbol{K}\widehat{\boldsymbol{\lambda}}_{\boldsymbol{x}} = \boldsymbol{k}_{\boldsymbol{x}}$ quand il n'y a pas d'ambiguïté.

La solution en $\widehat{\boldsymbol{\lambda}}_{\boldsymbol{x}}$ de (2.16) est unique tant que la matrice de covariance $\boldsymbol{K}$ est de rang plein (ce qui est normalement le cas), et elle s'écrit analytiquement $\widehat{\boldsymbol{\lambda}}_{\boldsymbol{x}} = \boldsymbol{K}^{-1}\boldsymbol{k}_{\boldsymbol{x}}$. Insistons sur le fait que les méthodes numériques pour calculer $\widehat{\boldsymbol{\lambda}}_{\boldsymbol{x}}$ n'utilisent pas ces formes analytiques. En effet, la matrice de covariance $\boldsymbol{K}$ est symétrique définie positive, ce qui appelle des techniques particulières pour la résolution du système linéaire (2.16) (voir la section 6.8). Lorsque la matrice de covariance n'est pas de rang plein alors que la covariance est définie positive, cela signifie qu'une donnée apparaît plusieurs fois et il est normalement possible d'éviter une telle situation. Toutefois, il se peut aussi que la matrice de covariance soit mal conditionnée, notamment lorsque deux données sont très proches dans l'espace des facteurs, avec une corrélation proche de un. Dans ce cas, on cherche à réduire la dimension (le rang) de la matrice en supprimant des données ou à partir de factorisations de la matrice $\boldsymbol{K}$. La section 6.8 revient sur ces problèmes plus en détail.

Propriétés d'interpolation

Puisque le prédicteur est une projection sur $\mathcal{H}_S$, on a $\hat{F}(\boldsymbol{x}_i) = F(\boldsymbol{x}_i)$, $\forall i = 1, \ldots, n$, ce qui implique que pour tout ω, la valeur prédite en $\boldsymbol{x}_i$ est la valeur observée f_i^{obs} en ce point. Par conséquent, le krigeage réalise toujours une interpolation (dans le cas où aucun bruit d'observation n'est pris en compte). Cependant, rien n'indique pour le moment que l'interpolation obtenue soit nécessairement une fonction continue. Une manière directe d'obtenir la continuité de la prédiction $\hat{F}(\omega, \boldsymbol{x})$ à ω fixé, est de s'intéresser à la forme duale du krigeage (voir la section 2.5.3).

Exemple graphique

La figure 2.3 illustre de la propriété d'interpolation. Des exemples similaires seront présentés au chapitre 6. Contrairement à d'autres méthodes plus classiques de régression (régression linéaire sur une base de polynômes ou réseau de neurones artificiels, voir par exemple (Bishop, 1995)), le krigeage fournit un modèle passant précisément par les données observées. L'interpolation obtenue par krigeage constitue donc un modèle comportemental pertinent lorsque la sortie du système est observée sans bruit. Le choix de la fonction de covariance permet de garder le contrôle du type d'interpolation, comme cela sera vu plus précisément au chapitre 5. Remarquons que la prédiction obtenue n'a pas d'« oscillations » parasites comme on en rencontre avec une interpolation polynomiale. Nous verrons en effet au chapitre 3 que le krigeage constitue une méthode de régression régularisée. D'après la figure 2.3, la prédiction est meilleure lorsqu'elle est réalisée à proximité de données observées. En général, le modèle est meilleur si le nombre de données est important sur le domaine d'étude[6]. Cette affirmation sera justifiée au chapitre 5.

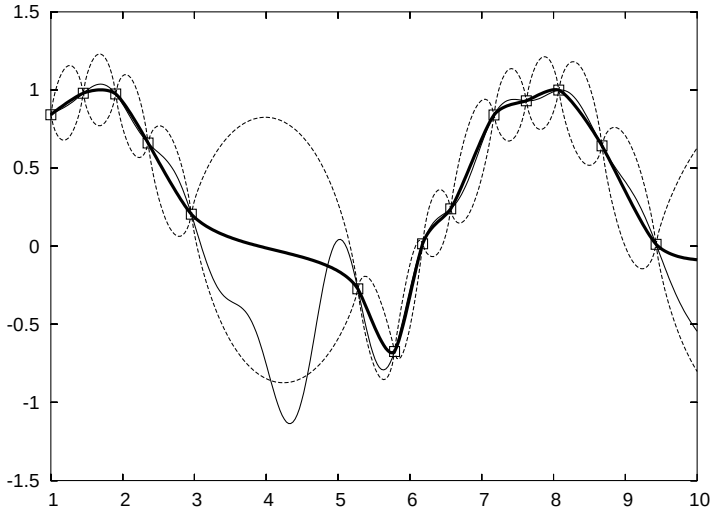


FIG. 2.3 – Exemple de prédiction par krigeage. (Les conventions graphiques sont celles de la figure 1.1.)

Cas des processus gaussiens

Supposons maintenant $F(\boldsymbol{x})$ gaussien. Dans ce cas il est bien connu que l'espérance conditionnelle peut être calculée analytiquement.

Proposition 19. *Soit* $(\boldsymbol{X}, \boldsymbol{Y})$ *un vecteur gaussien à valeurs dans* $\mathbb{R}^{n+q}$. *On suppose* $\det(\boldsymbol{K}_{\boldsymbol{X}}) \neq 0$. *Alors, pour toute* $f \in L^1$,

$$\mathrm{E}[f(\boldsymbol{Y}) \mid \boldsymbol{X}] = \phi_f(\boldsymbol{X}) \text{ p.s., avec } \phi_f(\boldsymbol{x}) = \mathrm{E}[f(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{Z})]\,,$$

[6]Si l'on possédait un nombre suffisant de données observées sans bruit, il serait équivalent d'un point de vue opérationnel de décrire le système par un modèle boîte noire ou par un modèle physique. Le modèle boîte noire pourrait même s'avérer meilleur que le modèle physique car ce dernier est souvent obtenu à partir d'approximations.

où $\boldsymbol{A} = \boldsymbol{K}_{\boldsymbol{Y},\boldsymbol{X}}\boldsymbol{K}_{\boldsymbol{X}}^{-1}$, $\boldsymbol{Z}$ *suit une loi normale* $\mathcal{N}(\mathrm{E}[\boldsymbol{Y}] - \boldsymbol{A}\,\mathrm{E}[\boldsymbol{X}], \boldsymbol{K}_{\boldsymbol{Y}} - \boldsymbol{A}\boldsymbol{K}_{\boldsymbol{X},\boldsymbol{Y}})$ *et* $\boldsymbol{K}_{\boldsymbol{Y},\boldsymbol{X}} = \mathrm{E}[(\boldsymbol{Y} - \mathrm{E}[\boldsymbol{Y}])(\boldsymbol{X} - \mathrm{E}[\boldsymbol{X}])^{\mathsf{T}}]$. *En particulier,*

$$\mathrm{E}[\boldsymbol{Y} \mid \boldsymbol{X}] - \mathrm{E}[\boldsymbol{Y}] = \boldsymbol{K}_{\boldsymbol{Y},\boldsymbol{X}}\boldsymbol{K}_{\boldsymbol{X}}^{-1}(\boldsymbol{X} - \mathrm{E}[\boldsymbol{X}]) \text{ p.s.}$$

Démonstration. Voir par exemple (Chonavel, 2002). □

Par conséquent, dans le cas des processus gaussiens, le prédicteur linéaire est optimal parmi tous les prédicteurs minimisant l'erreur en moyenne quadratique. Pour cette raison, la prédiction linéaire est bien adaptée dans le cas des processus gaussiens et on peut supposer que cela reste vrai si le processus n'est pas trop loin d'être gaussien. Dans le cadre de modèles boîte noire, il est en général pertinent de se restreindre aux processus gaussiens. Cependant, comme le note Stein (1999), il est relativement facile de construire des processus aléatoires non gaussiens pour lesquels la prédiction linéaire n'est plus du tout adaptée. Lorsque le modèle gaussien n'est pas tenable, on pourra recourir à des prédictions non-linéaires parmi celles proposées par les géostatisticiens (Chilès et Delfiner, 1999).

Erreur de prédiction

Dans le cas d'un processus aléatoire de moyenne connue, le prédicteur du krigeage est sans biais. D'autre part, l'un des avantages du cadre probabiliste retenu est que l'on caractérise facilement l'incertitude de la prédiction. La variance de l'erreur de prédiction est en effet donnée par

$$\begin{aligned}
\mathrm{Var}[F(\boldsymbol{x}) - \hat{F}(\boldsymbol{x})] &= \mathrm{E}[(F(\boldsymbol{x}) - \hat{F}(\boldsymbol{x}))^2] \\
&= k(\boldsymbol{x},\boldsymbol{x}) - 2\widehat{\boldsymbol{\lambda}}_{\boldsymbol{x}}^{\mathsf{T}}\boldsymbol{k}_{\boldsymbol{x}} + \widehat{\boldsymbol{\lambda}}_{\boldsymbol{x}}^{\mathsf{T}}\boldsymbol{K}\widehat{\boldsymbol{\lambda}}_{\boldsymbol{x}} \\
&= \mathrm{E}[F(\boldsymbol{x})^2] - \mathrm{E}[\hat{F}(\boldsymbol{x})^2] \quad \text{(relation de Pythagore)} \\
&= k(\boldsymbol{x},\boldsymbol{x}) - \widehat{\boldsymbol{\lambda}}_{\boldsymbol{x}}^{\mathsf{T}}\boldsymbol{K}\widehat{\boldsymbol{\lambda}}_{\boldsymbol{x}}
\end{aligned} \tag{2.17}$$

La variance de l'erreur de prédiction est typiquement utilisée pour déterminer la qualité de la prédiction en donnant des intervalles de confiance. Ainsi dans le cas gaussien, un intervalle de largeur $2a$ fois l'écart type de l'erreur de prédiction centré sur la valeur prédite contient une fraction égale à $\mathrm{erf}(a/\sqrt{2})$ des trajectoires possibles du processus (pour $a = 1.96$, $\mathrm{erf}(1.96/\sqrt{2}) \approx 0.95$). Dans la section 6.6, nous donnons un exemple d'utilisation de la variance de l'erreur de prédiction pour déterminer un plan d'expériences (il s'agit de choisir les observations pour minimiser l'erreur de prédiction).

Remarquons d'après (2.17), que la variance du prédicteur est nécessairement plus petite que la variance du processus aléatoire. On dit ainsi parfois que la prédiction « lisse ». On peut encore interpréter cette propriété comme le fait que la prédiction opère un moyennage des trajectoires possibles du processus aléatoire, conditionnellement aux observations. Nous reviendrons sur ce point dans la section 2.6 pour parler du rôle des simulations conditionnelles.

Une autre possibilité pour caractériser l'erreur de prédiction est d'utiliser la proposition suivante, classique en algèbre linéaire (Hirsch et Lacombes, 1999).

Proposition 20. *Soit* $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ *une famille libre d'un espace préhilbertien réel* E *et* F *le sous-espace généré par cette famille. Pour tout* $\boldsymbol{x} \in E$

$$d^2(\boldsymbol{x}, F) = \frac{\det[G(\boldsymbol{x}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)]}{\det[G(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)]},$$

où on note $G(\ldots)$ la matrice de Gram d'une famille de vecteurs, et $d(\boldsymbol{x}, F)$, la distance euclidienne de $\boldsymbol{x}$ à F. (Par définition, la matrice de Gram d'une famille de vecteurs $\boldsymbol{x}_i$, $i = 1, \ldots, n$, est la matrice $n \times n$ des produits scalaires $(\boldsymbol{x}_i, \boldsymbol{x}_j)$.)

Démonstration. (Indications) Considérer $\boldsymbol{y}$ le projeté orthogonal de $\boldsymbol{x}$ sur F. Dans le calcul de $\det[G(\boldsymbol{x}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)]$, remplacer $\boldsymbol{x}$ par $(\boldsymbol{x} - \boldsymbol{y}) + \boldsymbol{y}$ et utiliser la linéarité du déterminant par rapport à la première colonne. □

Cette formule est directement exploitable car la matrice de covariance est une matrice de produits scalaires, puisque l'on suppose $F(\boldsymbol{x})$ de moyenne nulle. Elle fournit deux interprétations intéressantes. Nous pouvons tout d'abord constater de manière heuristique que l'erreur de prédiction est de l'ordre de la plus petite valeur propre de la matrice de covariance. Cette remarque est importante dans le cadre de la résolution numérique du système (2.16). La théorie de l'information fournit une seconde interprétation. En effet, un vecteur aléatoire gaussien de dimension d et de matrice de covariance $\boldsymbol{K}$ possède une entropie s'écrivant sous la forme $\frac{1}{2}\log_2[\det \boldsymbol{K}] + \log_2[(2\pi e)^{d/2}] + C$. Ceci permet d'interpréter l'erreur de prédiction comme approximativement la différence d'information entre le vecteur aléatoire $\boldsymbol{F}_S$ et le vecteur $(F(\boldsymbol{x}), \boldsymbol{F}_S)^\mathsf{T}$.

Avant de conclure cette section, nous souhaitons discuter le comportement de l'erreur lorsque le nombre d'observations devient dense sur un domaine d'étude. Cette question sera reprise plus en détails dans la section 5.3. Pour effectuer la prédiction en un point $\boldsymbol{x}$ du domaine d'étude, considérons une suite d'observations aux points $\boldsymbol{x}_i \neq \boldsymbol{x}$, $i \in \mathbb{N}$, telle que $\boldsymbol{x}$ soit un point d'adhérence de la suite $(\boldsymbol{x}_n)$. Appelons $P_{\mathcal{H}_{S_n}}$, l'opérateur de projection orthogonale sur $\mathcal{H}_{S_n} = \mathrm{vect}\{F(\boldsymbol{x}_1), \ldots, F(\boldsymbol{x}_n)\}$, et $P_{F(\boldsymbol{x}_{t_n})}$ l'opérateur de projection sur le sous-espace généré par $F(\boldsymbol{x}_{t_n})$, avec $t_n = \mathrm{argmin}_{i \leq n} \|\boldsymbol{x} - \boldsymbol{x}_i\|_2$. Alors, pour tout $n \in \mathbb{N}$, il est clair que

$$\begin{aligned} \mathrm{Var}[F(\boldsymbol{x}) - P_{\mathcal{H}_{S_n}} F(\boldsymbol{x})] &\leq \mathrm{Var}[F(\boldsymbol{x}) - P_{F(\boldsymbol{x}_{t_n})} F(\boldsymbol{x})]\,, & (2.18) \\ &\leq k(\boldsymbol{x}, \boldsymbol{x}) - \frac{k(\boldsymbol{x}, \boldsymbol{x}_{t_n})^2}{k(\boldsymbol{x}, \boldsymbol{x})}\,. & (2.19) \end{aligned}$$

Supposons $k(\boldsymbol{x}, \boldsymbol{y})$ continue, stationnaire. Si la densité de points augmente dans un voisinage de $\boldsymbol{x}$, $\boldsymbol{x} - \boldsymbol{x}_{t_n}$ tend vers zéro et la variance de l'erreur tend également vers zéro. La vitesse de convergence est liée à la régularité de la covariance à l'origine. Nous verrons dans la section 5.3 que l'on peut trouver $l > 0$ et $C \geq 0$ tel que,

$$\mathrm{Var}[F(\boldsymbol{x}) - P_{\mathcal{H}_{S_n}} F(\boldsymbol{x})] \leq C \|\boldsymbol{x} - \boldsymbol{x}_{t_n}\|^l\,, \quad \forall n\,.$$

Notons que les bornes d'erreur optimales lorsque l'échantillonnage est non uniforme sont difficiles à obtenir en général. Très peu de résultats sont établis si l'on suppose de plus que la covariance utilisée pour la prédiction linéaire diffère de la covariance du processus (ce qui est en pratique toujours le cas, voir la section 5.3).

2.5.2 Prédiction dans le cas de plusieurs processus aléatoires

La méthode appelée *cokrigeage* (voir par exemple (Chilès et Delfiner, 1999)) permet d'effectuer une prédiction lorsque l'on observe plusieurs processus aléatoires $F_\alpha(\boldsymbol{x})$, $\alpha = 1, \ldots, q$. On peut, par exemple, vouloir prédire $F_1(\boldsymbol{x})$ en fonction d'observations $F_{\alpha_i}(\boldsymbol{x}_i)$, $i = 1, \ldots, n$. Le principe

reste le même que dans le cas d'un seul processus aléatoire. On cherche la projection orthogonale $\hat{F}_1(\boldsymbol{x})$ de $F_1(\boldsymbol{x})$ sur $\mathcal{H}_S = \text{vect}\{F_{\alpha_i}(\boldsymbol{x}_i); i = 1, \ldots, n\}$ qui s'écrit

$$\hat{F}_1(\boldsymbol{x}) = \sum_{i=1}^{n} \widehat{\lambda}_{i,\boldsymbol{x}} F_{\alpha_i}(\boldsymbol{x}_i)\,.$$

Les $\widehat{\lambda}_{i,\boldsymbol{x}}$ s'obtiennent, comme dans le cas du krigeage, par la résolution d'un système d'équations linéaires. Ce système s'obtient en fonction des covariances et des inter-covariances entre les processus aléatoires, et peut s'écrire sous forme matricielle :

$$\begin{pmatrix} k_{\alpha_1,\alpha_1}(\boldsymbol{x}_1,\boldsymbol{x}_1) & k_{\alpha_1,\alpha_2}(\boldsymbol{x}_1,\boldsymbol{x}_2) & \cdots & k_{\alpha_1,\alpha_n}(\boldsymbol{x}_1,\boldsymbol{x}_n) \\ k_{\alpha_2,\alpha_1}(\boldsymbol{x}_2,\boldsymbol{x}_1) & k_{\alpha_2,\alpha_2}(\boldsymbol{x}_2,\boldsymbol{x}_2) & & \vdots \\ \vdots & & \ddots & \vdots \\ \vdots & & & k_{\alpha_n,\alpha_n}(\boldsymbol{x}_n,\boldsymbol{x}_n) \end{pmatrix} \widehat{\boldsymbol{\lambda}}_{S,\boldsymbol{x}} = \begin{pmatrix} k_{1,\alpha_1}(\boldsymbol{x},\boldsymbol{x}_1) \\ \vdots \\ k_{1,\alpha_n}(\boldsymbol{x},\boldsymbol{x}_n) \end{pmatrix}$$

Remarquons que l'on se ramène très simplement au cas du krigeage. Pour un processus $F_\alpha(\boldsymbol{x})$, l'indice α peut en effet être interprété comme un paramètre supplémentaire (voir la même remarque dans la section 2.1.1). Formellement, le cas de plusieurs processus aléatoires devient identique au cas de la prédiction d'un seul processus et peut se traiter au moyen d'une seule fonction de covariance $k([\alpha,\boldsymbol{x}],[\beta,\boldsymbol{y}])$.

Dans le contexte de la modélisation de type boîte noire, l'utilisation de ce type de prédiction est naturellement intéressante quand le système comporte plusieurs sorties et que l'on souhaite exploiter les corrélations éventuelles entre elles.

Observations bruitées. Un exemple d'application important de cette méthode se présente lorsque l'on a un système dont la sortie observée est perturbée par un bruit additif gaussien. Chaque observation est alors modélisée par une variable aléatoire

$$F^{\text{obs}}_{\boldsymbol{x}_i} = F(\boldsymbol{x}_i) + N_i\,,$$

où $F(\boldsymbol{x})$ modélise la sortie du système sans bruit. Pour calculer le prédicteur linéaire optimal au sens du krigeage de $F(\boldsymbol{x})$,

$$\hat{F}(\boldsymbol{x}) = \sum_{i=1}^{n} \widehat{\lambda}_{i,\boldsymbol{x}} F^{\text{obs}}_{\boldsymbol{x}_i},$$

on calcule la projection orthogonale de $F(\boldsymbol{x})$ sur $\text{vect}\{F^{\text{obs}}_{\boldsymbol{x}_i};\ i = 1, \ldots, n\}$, ce qui conduit à résoudre le système linéaire

$$(\boldsymbol{K} + \sigma_N^2 \boldsymbol{I}_n)\widehat{\boldsymbol{\lambda}}_{\boldsymbol{x}} = \boldsymbol{k}_{\boldsymbol{x}}\,,$$

où σ_N^2 est la variance du bruit d'observation, supposée identique pour chaque observation. La prédiction obtenue est une approximation de la sortie du système. Nous n'avons plus dans ce cas la propriété d'interpolation et la variance de l'erreur de prédiction en un point $\boldsymbol{x}$ est minorée par la variance du bruit d'observation. Notons que la méthode se généralise aisément au cas de bruits d'observation corrélés.

Prédiction de la dérivée d'une sortie. Formellement, nous pouvons considérer un système à deux sorties, l'une étant la sortie effective du système, l'autre étant la dérivée de celle-ci. Pour simplifier la présentation, supposons que le modèle ne dépende que d'un seul facteur $x \in \mathbb{R}$. La sortie est alors modélisée par le processus aléatoire $F(x)$, supposé stationnaire (pour simplifier encore davantage), de moyenne nulle (le cas d'une moyenne non nulle est traité dans la section 4.6.2) et de covariance $k(h)$. La prédiction de la dérivée de $F(x)$ nécessite que $F(x)$ soit *dérivable en moyenne quadratique*, ce qui impose l'existence de la dérivée seconde de la covariance avec notamment $k^{(2)}(0) < \infty$. Le prédicteur linéaire $\widehat{F'}(x)$ de $F'(x)$ à partir de $F_{x_1}^{\text{obs}}, \ldots, F_{x_n}^{\text{obs}}$ s'écrit sous la forme

$$\widehat{F'}(x) = \sum_{i=1}^{n} \widehat{\lambda}_{i,x} F_{x_i}^{\text{obs}},$$

où les $\widehat{\lambda}_{i,x}$ s'obtiennent comme solution du système

$$(\boldsymbol{K} + \sigma_N^2 \boldsymbol{I}_n)\widehat{\boldsymbol{\lambda}}_x = \boldsymbol{k'}_x\,,$$

avec σ_N^2 la variance du bruit et

$$\boldsymbol{k'}_x = (\text{sgn}(x - x_1)k'(|x - x_1|), \ldots, \text{sgn}(x - x_n)k'(|x - x_n|))^{\mathsf{T}}.$$

Terminons par quelques remarques à propos de la prédiction des dérivées. Tout d'abord, notons que la méthode permet de dériver des signaux bruités, ce qui est généralement considéré comme un problème délicat de traitement du signal. La section 6.1.2 montre sur des exemples que la méthode est très satisfaisante. La généralisation à la prédiction des dérivées d'ordre supérieur ne pose pas de problème particulier, à condition de choisir des covariances suffisamment dérivables. Dans le cas où $\boldsymbol{x} \in \mathbb{R}^d$, on doit naturellement s'intéresser à la prédiction des composantes du gradient de $F(\boldsymbol{x})$. Enfin, il peut être intéressant dans certaines situations de supposer des dérivées partielles (ou le gradient) connues en des points de l'espace des facteurs. Par exemple, on peut vouloir prendre en compte des connaissances a priori sur la physique du système, comme des conditions aux limites. Dans ce cas, nous pouvons considérer ces connaissances comme les observations des vecteurs aléatoires $\nabla F(\boldsymbol{y}_j)$, $j = 1, \ldots, m$, et calculer alors le prédicteur optimal en projetant orthogonalement les composantes de $\nabla F(\boldsymbol{x})$ sur

$$\text{vect}\{F(\boldsymbol{x}_i); i = 1, \ldots, n\} \oplus \text{vect}\{\nabla F(\boldsymbol{y}_j), j = 1, \ldots, m\}.$$

La section 6.1.2 propose des exemples de prise en compte d'information a priori sur les dérivées et d'intégration.

2.5.3 Krigeage dual

Le *krigeage dual* (Chilès et Delfiner, 1999) permet d'éviter de calculer les coefficients $\widehat{\lambda}_{i,\boldsymbol{x}}$ de la combinaison linéaire pour tous les $\boldsymbol{x}$ où la prédiction doit être faite. Ceci présente un intérêt non négligeable en pratique. Le terme « dual » vient de l'analyse de l'équivalence entre krigeage et régression régularisée (historiquement, il s'agissait en fait de splines (Matheron, 1981)).

Si la matrice de covariance du vecteur aléatoire des observations est de rang plein, pour tout $\boldsymbol{x} \in \mathbb{X}$, le prédicteur s'écrit

$$\begin{aligned} \hat{F}(\boldsymbol{x}) &= (\widehat{\boldsymbol{\lambda}}_{\boldsymbol{x}}, \boldsymbol{F}_S) \\ &= (\boldsymbol{K}^{-1}\boldsymbol{k}_{\boldsymbol{x}}, \boldsymbol{F}_S) \\ &= (\boldsymbol{k}_{\boldsymbol{x}}, \boldsymbol{K}^{-1}\boldsymbol{F}_S)\,. \end{aligned}$$

Le terme « dual » vient de l'opération effectuée dans la dernière équation, qui utilise le fait que la matrice $\boldsymbol{K}^{-1}$ est symétrique (auto-adjointe).

Notons $\boldsymbol{A}_S = (A_1, \ldots, A_n)^\mathsf{T}$ le vecteur aléatoire $\boldsymbol{K}^{-1}\boldsymbol{F}_S$. Pour ω fixé, le prédicteur s'exprime donc comme une combinaison linéaire des covariances :

$$\hat{F}(\boldsymbol{x}) = \sum_{i=1}^{n} A_i k(\boldsymbol{x}_i, \boldsymbol{x})\,,$$

avec les variables aléatoires A_i données par la solution du système d'équations linéaires à coefficients aléatoires

$$\boldsymbol{K}\boldsymbol{A}_S = \boldsymbol{F}_S\,. \tag{2.20}$$

Autrement dit, il n'est pas nécessaire de résoudre un nouveau système linéaire pour chaque $F(\boldsymbol{x})$ à prédire et il suffit d'avoir résolu une fois pour toute le système linéaire (2.20). La proposition qui suit permet d'affirmer que le krigeage réalise une interpolation continue des données observées.

Proposition 21. *Soit $F(\boldsymbol{x})$ un processus aléatoire du second ordre, de moyenne nulle et de covariance continue. Si la matrice de covariance du vecteur aléatoire des observations $F(\boldsymbol{x}_i)$, $i = 1, \ldots, n$, est de rang plein, alors les valeurs prédites $\hat{F}(\omega, \boldsymbol{x})$ à partir des observations constituent, à ω fixé, une fonction continue qui interpole les observations $F(\omega, \boldsymbol{x}_i)$.*

Démonstration. Nous avons déjà vu la propriété $\hat{F}(\omega, \boldsymbol{x}_i) = F(\omega, \boldsymbol{x}_i)$, $\forall i \in \{1, \ldots, n\}$ et $\forall \omega \in \Omega$. Comme $\hat{F}(\boldsymbol{x})$ est une combinaison linéaire (aléatoire) de fonctions continues, $\hat{F}(\omega, \boldsymbol{x})$ est une fonction continue à ω fixé. □

2.5.4 Une méthode récursive de prédiction linéaire

La méthode proposée dans cette section est une formulation de l'*algorithme des innovations* (voir par exemple Brockwell et Davis, 1987) pour la prédiction d'un processus aléatoire paramétré par un espace de facteurs quelconque. Elle exploite de manière intéressante la propriété de projection orthogonale pour calculer de manière récursive la solution du système (2.16), ce qui présente un intérêt dans les applications où des données sont susceptibles d'être rajoutées au fur et à mesure. Nous présentons cet algorithme et nous évaluons sa complexité.

Proposition 22 (Algorithme des innovations). *Soit $F(\boldsymbol{x})$ un processus aléatoire du second ordre, de moyenne nulle et de covariance $k(\boldsymbol{x}, \boldsymbol{y})$. Soit $\hat{F}_{S_n}(\boldsymbol{x}_{n+1})$, la projection orthogonale de $F(\boldsymbol{x}_{n+1})$ sur*

$$\mathcal{H}_{S_n} = \mathrm{vect}\{F(\boldsymbol{x}_1), \ldots, F(\boldsymbol{x}_n)\}\,.$$

On a

$$\hat{F}_{S_n}(\boldsymbol{x}_{n+1}) = \begin{cases} 0 & \text{si } n = 1, \\ \sum_{j=1}^{n} \theta_{n+1,j}(F(\boldsymbol{x}_j) - \hat{F}_{S_{j-1}}(\boldsymbol{x}_j)) & \text{si } n \geq 2\,, \end{cases} \tag{2.21}$$

avec

$$\begin{aligned} \theta_{n+1,m} &= v_m^{-1}\left[k(\boldsymbol{x}_{n+1},\boldsymbol{x}_m) - \sum_{j=1}^{m-1}\theta_{m,j}\theta_{n+1,j}v_j\right],\ m = 1,\ldots,n\,, \\ v_1 &= k(\boldsymbol{x}_1,\boldsymbol{x}_1)\,, \\ v_n &= k(\boldsymbol{x}_n,\boldsymbol{x}_n) - \sum_{j=1}^{n-1}\theta_{n,j}^2 v_j\,. \end{aligned}$$

On peut calculer récursivement v_1, $\theta_{2,1}$, v_2, $\theta_{3,1}$, $\theta_{3,2}$, v_3, *etc.*

Démonstration. La famille $\{F(\boldsymbol{x}_1) - \hat{F}_{S_1}(\boldsymbol{x}_1), \ldots, F(\boldsymbol{x}_n) - \hat{F}_{S_{n-1}}(\boldsymbol{x}_n)\}$ est orthogonale puisque pour $i < j$,

$$F(\boldsymbol{x}_i) - \hat{F}_{S_{i-1}}(\boldsymbol{x}_i) \in \mathcal{H}_{S_{j-1}}\,,$$

et

$$F(\boldsymbol{x}_j) - \hat{F}_{S_{j-1}}(\boldsymbol{x}_j) \perp \mathcal{H}_{S_{j-1}}\,.$$

Posons

$$v_n = \|F(\boldsymbol{x}_n) - \hat{F}_{S_{n-1}}(\boldsymbol{x}_n)\|^2\,.$$

Formons le produit scalaire des deux côtés de l'égalité (2.21) avec $F(\boldsymbol{x}_m) - \hat{F}_{S_{m-1}}(\boldsymbol{x}_m)$. Il vient

$$(\hat{F}_{S_n}(\boldsymbol{x}_{n+1}), F(\boldsymbol{x}_m) - \hat{F}_{S_{m-1}}(\boldsymbol{x}_m)) = \theta_{n+1,m}v_m\,, \quad \forall m \in \{1,\ldots,n\}$$

La propriété d'orthogonalité implique alors :

$$\begin{aligned} \theta_{n+1,m} &= v_m^{-1}(\hat{F}_{S_n}(\boldsymbol{x}_{n+1}), F(\boldsymbol{x}_m) - \hat{F}_{S_{m-1}}(\boldsymbol{x}_m)) \\ &= v_m^{-1}(F(\boldsymbol{x}_{n+1}), F(\boldsymbol{x}_m) - \hat{F}_{S_{m-1}}(\boldsymbol{x}_m)) \\ &= v_m^{-1}\left[k(\boldsymbol{x}_{n+1},\boldsymbol{x}_m) - \sum_{j=1}^{m-1}\theta_{m,j}(F(\boldsymbol{x}_{n+1}), F(\boldsymbol{x}_j) - \hat{F}_{S_{j-1}}(\boldsymbol{x}_j))\right] \\ &= v_m^{-1}\left[k(\boldsymbol{x}_{n+1},\boldsymbol{x}_m) - \sum_{j=1}^{m-1}\theta_{m,j}\theta_{n+1,j}v_j\right] \end{aligned}$$

□

Remarque. Cet algorithme s'applique traditionnellement dans le cas des séries chronologiques, où les observations, modélisées par des variables aléatoires X_n, sont espacées régulièrement dans le temps. Il permet de calculer les prédictions en fonction des innovations $X_n - \hat{X}_n$, alors que l'algorithme récursif de Durbin–Levinson calcule les prédictions directement en fonction des X_n. Ce dernier algorithme exploite la structure Toeplitz de la matrice de covariance et n'est pas utilisable pour le krigeage si l'échantillonnage est non uniforme.

Évaluons la complexité numérique de cet algorithme. Le nombre des opérations impliquées à l'étape n est de $n^2 + 3n - 2$ multiplications et $0.5n^2 + 3.5n$ additions. Notons que $\sum_{k=1}^{n} k^2 = \frac{1}{6}n(n+1)(2n+1)$. L'ordre de grandeur du nombre total des opérations pour le calcul récursif de $\hat{F}_{S_n}(\boldsymbol{x}_{n+1})$ figure dans la table 2.1, où l'on voit que la complexité de la méthode des innovations est comparable à celle des méthodes classiques de résolution de systèmes linéaires.

Algorithme	additions	multiplications
Gauss	$\frac{n^3}{3}$	$\frac{n^3}{3}$
Cholesky	$\frac{n^3}{6}$	$\frac{n^3}{6}$
Innovations	$\frac{n^3}{6}$	$\frac{n^3}{3}$

TAB. 2.1 – Ordre de grandeur du nombre total d'opérations pour différentes méthodes de calcul de $\hat{F}(\boldsymbol{x}_{n+1})$

2.5.5 Limites de la prédiction linéaire

Un intérêt fondamental du krigeage est sa simplicité algorithmique. Cette simplicité est d'ailleurs liée à celle des hypothèses sur la loi du processus $F(\boldsymbol{x})$, considéré comme approximativement gaussien. L'inférence des paramètres de $F(\boldsymbol{x})$ à partir des données observées est ainsi facilitée. D'autres hypothèses simplificatrices sont utilisées en pratique, comme celle de l'existence de propriétés d'invariance dans l'espace des facteurs, notamment celle de stationnarité. Le krigeage est une prédiction optimale seulement dans le cas gaussien (section 2.5.1) mais nous pouvons encore utiliser une prédiction linéaire même lorsque le modèle gaussien n'est pas adapté (si les moments d'ordre un et deux du processus aléatoire sont connus). Comme le note Stein (1999), section 1.4, la prédiction linéaire peut devenir largement sous-optimale pour certains processus aléatoires, mais de tels processus sont assez peu réalistes comme modèles de système.

L'une des limitations de la prédiction linéaire est qu'il n'est pas possible de garantir la positivité de la prédiction. Des solutions ont toutefois été proposées en utilisant des algorithmes de programmation quadratique et des contraintes de positivité sur les coefficients $\widehat{\lambda}_{i,\boldsymbol{x}}$ du krigeage (Barnes et Johnson, 1984 ; Chilès et Delfiner, 1999). La classe des prédicteurs est restreinte aux combinaisons linéaires à coefficients positifs (la solution d'un tel problème est unique car il s'agit de trouver la meilleure projection sur un sous-ensemble convexe fermé de $\mathcal{H}_S \subset \mathcal{H}$). Il est également possible d'imposer des contraintes d'inégalité quelconques sur la prédiction (Langlais, 1990 ; Chilès et Delfiner, 1999).

Les méthodes de krigeage non-linéaire permettent de s'affranchir de la limitation du caractère gaussien du processus modélisant le système. Ces techniques permettent par exemple de prédire des dépassements de seuils. Si des raisons spécifiques motivent le choix d'un processus non-gaussien comme modèle du système, une analyse plus détaillée des données est généralement nécessaire.

Enfin, mentionnons une limitation de la prédiction linéaire liée au nombre de données. Si le nombre n de données est important, le coût algorithmique de la méthode en $\mathrm{O}(n^3)$ peut s'avérer incompatible avec l'usage prévu du modèle. La solution la plus simple pour remédier à ce problème est alors d'effectuer une prédiction en un point donné de l'espace des facteurs en ne gardant que les données situées dans un voisinage de ce point. Toutefois, cette solution ne permet par d'éliminer

les données redondantes. Afin d'obtenir des modèles *creux*, les méthodes de régression à vecteurs de support sont envisageables (voir la section 3.6.4), au prix d'une complexité numérique accrue.

2.6 Éléments sur la simulation de processus aléatoires

Dans de nombreux cas, il est utile de générer des trajectoires d'un processus aléatoire. On parle alors de *simulations* de trajectoires. Nous présentons très brièvement les principes généraux que nous avons utilisés. On pourra consulter (Chilès et Delfiner, 1999) pour une présentation très complète des méthodes de simulation dans un cadre géostatistique. (Robert et Casella, 1999 ; Lantuejoul, 2002) constituent des références classiques sur les méthodes de simulation en statistiques.

2.6.1 Décomposition de la matrice covariance

Soit $F(\boldsymbol{x})$ un processus aléatoire de moyenne nulle et de covariance $k(\boldsymbol{x}, \boldsymbol{y})$. Nous souhaitons obtenir des réalisations du vecteur $\boldsymbol{F}_S = (F(\boldsymbol{x}_1), \ldots, F(\boldsymbol{x}_n))^\mathsf{T}$. Nous nous restreignons ici au cas où $F(\boldsymbol{x})$ est gaussien, pour lequel les méthodes de génération de réalisations sont simples à mettre en œuvre. (Lorsque $F(\boldsymbol{x})$ n'est pas gaussien, il est possible de se contenter d'approximations au second ordre et de générer des réalisations au sens faible, en ne tenant compte que des moments d'ordre un et deux.) Les méthodes de simulation les plus simples reposent sur des décompositions de la matrice de covariance. On calcule ainsi la factorisation de Cholesky $\boldsymbol{C}\boldsymbol{C}^\mathsf{T}$ de la matrice de covariance $\boldsymbol{K}_S$ de $\boldsymbol{F}_S$, où $\boldsymbol{C}$ est une matrice triangulaire inférieure (une telle décomposition existe si la matrice de covariance est positive (Ciarlet, 1998)). Si $\boldsymbol{X}$ est un vecteur de n variables aléatoires gaussiennes indépendantes de variance unité, alors le vecteur aléatoire $\boldsymbol{C}\boldsymbol{X}$ admet la même loi que $\boldsymbol{F}_S$.

Il suffit donc de générer des réalisations de $\boldsymbol{X}$ pour obtenir des réalisations de $\boldsymbol{F}_S$. En revanche, la factorisation de Cholesky comporte $\mathrm{O}(n^3)$ opérations, ce qui restreint en pratique cette méthode à la simulation de vecteurs à quelques centaines d'éléments. Il est parfois possible de tirer avantage de la structure de la matrice de covariance lorsque les points de l'espace des facteurs sont régulièrement disposés sur une grille (Zimmerman, 1989). D'autres possibilités existent, comme la séparation du domaine de simulation en plusieurs sous-domaines (Vecchia, 1988).

2.6.2 Simulations conditionnelles

On peut vouloir générer des réalisations d'un processus aléatoire qui prennent la valeur de la sortie aux points observés. On parle alors de *simulation conditionnelle*.

Soit $\boldsymbol{F}_{S_0} = (F(\boldsymbol{y}_1), \ldots, F(\boldsymbol{y}_m))$ et $\boldsymbol{F}_{S_1} = (F(\boldsymbol{x}_1), \ldots, F(\boldsymbol{x}_n))$ deux vecteurs gaussiens extraits du processus aléatoire $F(\boldsymbol{x})$. La loi conditionnelle de $\boldsymbol{F}_{S_1}$ sachant $\boldsymbol{F}_{S_0} = \boldsymbol{f}_{S_0}$ s'écrit d'après la règle

de Bayes

$$\begin{aligned}
&\mathsf{P}\{F(\boldsymbol{x}_1) \in [f_{\boldsymbol{x}_1}, f_{\boldsymbol{x}_1} + df_{\boldsymbol{x}_1}], \ldots, F(\boldsymbol{x}_n) \in [f_{\boldsymbol{x}_n}, f_{\boldsymbol{x}_n} + df_{\boldsymbol{x}_n}] \mid \boldsymbol{f}_{S_0}\} \\
&\quad = \mathsf{P}\{F(\boldsymbol{x}_n) \in [f_{\boldsymbol{x}_n}, f_{\boldsymbol{x}_n} + df_{\boldsymbol{x}_n}] \mid \boldsymbol{f}_{S_0}, f_{\boldsymbol{x}_1}, \ldots, f_{\boldsymbol{x}_{n-1}}\} \\
&\qquad \times \mathsf{P}\{F(\boldsymbol{x}_{n-1}) \in [f_{\boldsymbol{x}_{n-1}}, f_{\boldsymbol{x}_{n-1}} + df_{\boldsymbol{x}_{n-1}}] \mid \boldsymbol{f}_{S_0}, f_{\boldsymbol{x}_1}, \ldots, f_{\boldsymbol{x}_{n-2}}\} \\
&\qquad \vdots \\
&\qquad \times \mathsf{P}\{F(\boldsymbol{x}_1) \in [f_{\boldsymbol{x}_1}, f_{\boldsymbol{x}_1} + df_{\boldsymbol{x}_1}] \mid \boldsymbol{f}_{S_0}\}
\end{aligned}$$

Par conséquent, la simulation de $\boldsymbol{F}_{S_1}$ conditionnellement à $\boldsymbol{F}_{S_0} = \boldsymbol{f}_{S_0}$ peut être abordée par une *approche séquentielle* en générant successivement des réalisations selon les lois conditionnelles $\mathsf{P}\{F(\boldsymbol{x}_i) \in [f_{\boldsymbol{x}_i}, f_{\boldsymbol{x}_i} + df_{\boldsymbol{x}_i}] \mid \boldsymbol{f}_{S_0}, f_{\boldsymbol{x}_1}, \ldots, f_{\boldsymbol{x}_{i-1}}\}$, $i = 1, \ldots, n$. Cette approche très générale nécessite toutefois d'être capable de déterminer les lois conditionnelles, ce qui ne pose pas de problème dans le cas d'un processus gaussien (on peut par exemple déterminer les paramètres de la loi conditionnelle en utilisant la moyenne et la variance du prédicteur par krigeage).

Une seconde méthode aisément mise en œuvre est le *conditionnement par krigeage*. Soit $F(\boldsymbol{x})$ un processus aléatoire gaussien et $\hat{F}(\boldsymbol{x})$ le meilleur prédicteur linéaire de $F(\boldsymbol{x})$ à partir des observations en $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$. Pour tout $\omega \in \Omega$, $F(\omega, \boldsymbol{x}_i) - \hat{F}(\omega, \boldsymbol{x}_i) = 0$, $i = 1, \ldots, n$. Si l'on suppose que les sorties observées du système correspondent à un certain ω_0, le processus aléatoire défini par

$$\hat{F}(\omega_0, \boldsymbol{x}) + (F(\omega, \boldsymbol{x}) - \hat{F}(\omega, \boldsymbol{x})) \tag{2.22}$$

est tel que ses réalisations passent toutes par les observations. On peut vérifier que le processus (2.22) admet la loi de $F(\boldsymbol{x})$ conditionnellement aux observations (Chilès et Delfiner, 1999). Par conséquent, la méthode de simulation conditionnelle par krigeage consiste à

- calculer par krigeage l'interpolation $\hat{F}(\omega_0, \boldsymbol{x})$ des observations $F(\omega_0, \boldsymbol{x}_i)$,
- générer des simulations non conditionnelles de $F(\omega, \boldsymbol{x})$ en utilisant par exemple une factorisation de Cholesky de la matrice de covariance,
- calculer par krigeage les interpolations de ces simulations échantillonnées aux points $\boldsymbol{x}_i$,
- appliquer la formule (2.22).

Notons que les coefficients $\widehat{\lambda}_i$ du krigeage obtenus pour interpoler les observations sont réutilisées lorsque l'on effectue les interpolations des simulations non conditionnelles.

2.6.3 Autres techniques

La simulation de processus aléatoires est un sujet très classique dans la littérature des probabilités et des statistiques (Lantuejoul, 2002). Les difficultés principales sont de deux types. Tout d'abord, simuler des processus non gaussiens est souvent difficile parce qu'il n'est pas facile d'obtenir les lois conditionnelles. Par ailleurs si le nombre de points d'échantillonnage est important, peu de méthodes efficaces existent. La technique des bandes tournantes, évoquée dans la section 2.1.3, ne permet de simuler que des processus aléatoires gaussiens isotropes définis sur des espaces de facteurs de dimension d. Une difficulté supplémentaire est la détermination de la covariance en dimension un à partir de la donnée de celle en dimension d. Les techniques de simulation par chaînes de Markov sont sans doute les plus connues et les plus conseillées (algorithmes de Gibbs et de Metropolis-Hastings).

2.7 Conclusions

Ce chapitre a fait une présentation élémentaire de la théorie des processus aléatoires, en commençant par leur construction. Ces éléments sont extrêmement classiques mais nous avons jugé pertinent de les rappeler. En effet, la notion de processus aléatoires à paramètres continus fait appel à des notions parfois subtiles. Les deux points qui nous paraissent importants à considérer sont les propriétés de continuité des trajectoires induites par la structure du second ordre et la notion de variables aléatoires à valeurs dans des espaces hilbertiens. Cependant, notre objectif n'était pas de rentrer dans des considérations avancées que l'on peut trouver dans des ouvrages comme (Ibragimov et Rozanov, 1978) ou (Adler, 1981).

Nous avons également rappelé le principe de prédiction linéaire qui correspond à la formulation élémentaire du krigeage. Il s'agit là encore de notions très classiques, notamment en géostatistique, mais notre présentation a été orientée ici vers la modélisation boîte noire des systèmes. Pour cette raison, nous avons traité plus spécifiquement le cas des observations bruitées, celui des systèmes à plusieurs sorties, la prédiction de dérivées, etc. Nous avons montré que l'algorithme des innovations (qui appartient traditionnellement à la littérature des séries chronologiques) pouvait être utilisé dans notre contexte pour effectuer des prédictions en ligne.

Au chapitre suivant, dans lequel nous aborderons la modélisation boîte noire par méthodes à noyaux, nous rappellerons l'équivalence entre méthodes à noyaux et la prédiction linéaire de processus aléatoires.

Chapitre 3

Régression régularisée dans les espaces hilbertiens à noyau reproduisant

Résumé — La notion d'espace hilbertien à noyau reproduisant est fondamentale pour la modélisation boîte noire. Dans un premier temps, le concept d'espace hilbertien à noyau reproduisant est rappelé et illustré. Nous décrivons ensuite l'utilisation de tels espaces dans le cadre de la régression régularisée et examinons les liens entre régression régularisée et prédiction linéaire dans un cadre probabiliste.

3.1 Espaces hilbertiens à noyau reproduisant

La théorie des espaces hilbertiens à noyau reproduisant est très classique en analyse fonctionnelle. Les premières études ont été effectuées avant 1950 et l'article (Aronszajn, 1950) constitue une synthèse de référence. Nous commençons par donner les définitions et les principales propriétés. Dans la section 3.2, nous présenterons des exemples tirés de la littérature qui nous paraissent pertinents pour la construction de modèles boîtes noires.

Nous noterons $\mathcal{F}$ un espace vectoriel de fonctions $\boldsymbol{x} \mapsto f(\boldsymbol{x})$ définies sur $\mathbb{X}$, muni d'une structure d'espace hilbertien avec le produit scalaire $(.,.)_{\mathcal{F}}$.

Définition 24. Une fonction $k(\boldsymbol{x}, \boldsymbol{y}) : \mathbb{X}\times\mathbb{X} \to \mathbb{R}$ est un *noyau reproduisant* d'un espace hilbertien $\mathcal{F}$ de fonctions sur $\mathbb{X}$, si toute fonction f de $\mathcal{F}$ vérifie la *propriété de reproduction*

$$f(\boldsymbol{x}) = (k(\boldsymbol{x}, .), f)_{\mathcal{F}}\,. \tag{3.1}$$

Cette définition suppose implicitement que les fonctions $k(\boldsymbol{x}, .)$ appartiennent à $\mathcal{F}$, pour tout $\boldsymbol{x}$ appartenant à $\mathbb{X}$.

Les propriétés fondamentales d'un noyau reproduisant sont rappelées dans la proposition suivante.

Proposition 23 (Propriétés d'un noyau reproduisant). *Si k est un noyau reproduisant d'un espace hilbertien $\mathcal{F}$, alors*

1. *$k(\boldsymbol{x}, \boldsymbol{y})$ est unique,*
2. *$\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{X}$, $k(\boldsymbol{x}, \boldsymbol{y}) = k(\boldsymbol{y}, \boldsymbol{x})$,*
3. *$\sum_{i,j=1}^{n} \lambda_i \lambda_j k(\boldsymbol{x}_i, \boldsymbol{x}_j) \geq 0$, pour toute suite finie de points $\boldsymbol{x}_i \in \mathbb{X}$ et de scalaires λ_i,*
4. *$(k(\boldsymbol{x}, .), k(\boldsymbol{y}, .))_{\mathcal{F}} = k(\boldsymbol{x}, \boldsymbol{y})$.*

Démonstration. Ces propriétés se démontrent sans difficulté. Si k et k' sont deux noyaux reproduisants, on a $\forall \boldsymbol{x} \in \mathbb{X}$

$$\|k(\boldsymbol{x}, .) - k'(\boldsymbol{x}, .)\|_{\mathcal{F}}^2 = (k(\boldsymbol{x}, .) - k'(\boldsymbol{x}, .), k(\boldsymbol{x}, .))_{\mathcal{F}} - (k(\boldsymbol{x}, .) - k'(\boldsymbol{x}, .), k'(\boldsymbol{x}, .))_{\mathcal{F}} = 0\,,$$

ce qui montre l'unicité. En prenant $f = k(\boldsymbol{y}, .)$ dans (3.1), on montre que $k(\boldsymbol{x}, \boldsymbol{y}) = k(\boldsymbol{y}, \boldsymbol{x})$. En prenant $f = \sum_{i=1}^{n} \lambda_i k(\boldsymbol{x}_i, .)$, et en exprimant $\|f\|_{\mathcal{F}}$, on prouve la propriété de positivité. La quatrième propriété résulte de la propriété de reproduction. □

Notons que les noyaux reproduisants sont des fonctions symétriques et de type positif. Le fait qu'un noyau reproduisant possède ainsi les mêmes propriétés qu'une fonction de covariance constitue un point essentiel (voir la section 3.5).

Proposition 24. *Si $\mathcal{F}$ admet $k(\boldsymbol{x}, \boldsymbol{y})$ comme noyau reproduisant, l'espace vectoriel $\tilde{\mathcal{F}}$ engendré par les fonctions $k(\boldsymbol{x}, .)$, lorsque $\boldsymbol{x}$ parcourt $\mathbb{X}$, est dense dans $\mathcal{F}$.*

Démonstration. Notons $\mathcal{F}^*$ le dual topologique de $\mathcal{F}$, c'est-à-dire l'ensemble des formes linéaires continues sur $\mathcal{F}$. Pour montrer que $\tilde{\mathcal{F}} = \text{vect}\{k(\boldsymbol{x}, .), x \in \mathbb{X}\}$ est dense dans $\mathcal{F}$, il suffit de montrer que toute forme linéaire $\lambda \in \mathcal{F}^*$ s'annulant sur $\tilde{F}$ est identiquement nulle. Notons ϱ l'opérateur de dualité de $\mathcal{F}$ (rappelons qu'il s'agit d'une isométrie surjective de $\mathcal{F}$ sur $\mathcal{F}^*$). Soit $\lambda \in \mathcal{F}^*$ s'annulant sur $\tilde{\mathcal{F}}$ et $f \in \mathcal{F}$ telle que $\lambda = \varrho f$. Pour tout $\boldsymbol{x} \in \mathbb{X}$,

$$f(\boldsymbol{x}) = (f, k(\boldsymbol{x}, .))_{\mathcal{F}} = \langle \varrho f, k(x, .) \rangle_{\mathcal{F}^*, \mathcal{F}} = \langle \lambda, k(x, .) \rangle_{\mathcal{F}^*, \mathcal{F}} = 0\,,$$

où $\langle \cdot, \cdot \rangle$ désigne le produit de dualité. Donc $f = 0$ et par suite, $\lambda = \varrho f = 0$. □

Proposition 25. *Si $\mathcal{F}$ admet $k(\boldsymbol{x}, .)$ comme noyau reproduisant, $\forall f \in \mathcal{F}$ et $\forall \boldsymbol{x} \in \mathbb{X}$,*

$$|f(\boldsymbol{x})| \leq \sqrt{k(\boldsymbol{x}, \boldsymbol{x})} \|f\|_{\mathcal{F}}\,.$$

La convergence forte (en norme) d'une suite de fonction (f_n) vers f, ainsi que la convergence faible (convergence des produits scalaires (f_n, g) vers (f, g) pour tout $g \in \mathcal{F}$), implique la convergence simple de f_n vers f.

Démonstration. Ces propriétés découlent de l'inégalité de Schwarz. □

Un espace hilbertien à noyau reproduisant $\mathcal{F}$ est un espace de fonctions définies en tout point de $\mathbb{X}$. Ceci exclut en particulier le cas de $L^2(\mathbb{X})$ qui n'est pas à proprement parlé un espace de fonctions, mais plus exactement un espace de classes d'équivalence de fonctions. La proposition 25 peut être reformulée en disant que si deux fonctions de $\mathcal{F}$ sont proches pour la topologie induite par $\|\cdot\|_{\mathcal{F}}$, alors les valeurs de ces fonctions en tout point de $\mathbb{X}$ sont également proches. Il s'agit d'une propriété importante.

Théorème 7. *Soit $k : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$, une fonction symétrique de type positif. Alors il existe un espace hilbertien $\mathcal{F}$ de fonctions sur $\mathbb{X}$ admettant k comme noyau reproduisant.*

Démonstration par construction explicite de $\mathcal{F}$. (D'après une preuve présentée dans (Aubin, 2000).) Soit $\tilde{\Lambda}$ le sous-espace vectoriel de $\mathbb{R}^{\mathbb{X}*}$ généré par les formes linéaires

$$\delta_{\boldsymbol{x}} : f \in \mathbb{R}^{\mathbb{X}} \mapsto f(\boldsymbol{x}) \in \mathbb{R}\,.$$

La première étape consiste à remarquer que $\mathbb{R}^{\mathbb{X}}$, l'espace de toutes les fonctions sur $\mathbb{X}$, s'identifie au dual algébrique de $\tilde{\Lambda}$. En effet, si Φ est une forme linéaire $\tilde{\Lambda} \to \mathbb{R}$, définissons la fonction $\phi \in \mathbb{R}^{\mathbb{X}}$ par $\phi(\boldsymbol{x}) = \Phi(\delta_{\boldsymbol{x}})$ pour tout $\boldsymbol{x} \in \mathbb{X}$. Alors ϕ peut être identifiée à Φ car $\forall \lambda = \sum_{i=1}^{n} \lambda_i \delta_{x_i} \in \tilde{\Lambda}$,

$$\Phi(\lambda) = \sum_{i=1}^{n} \lambda_i \Phi(\delta_{\boldsymbol{x}_i}) = \sum_{i=1}^{n} \lambda_i \phi(\boldsymbol{x}_i)\,.$$

La fonction $k : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ permet de définir sur $\tilde{\Lambda} \times \tilde{\Lambda}$ la forme bilinéaire

$$(\lambda, \mu)_{\tilde{\Lambda}} = \sum_{i,j} \lambda_i \mu_j k(\boldsymbol{x}_i, \boldsymbol{y}_j)\,,$$

avec $\lambda = \sum_{i=1}^{n} \lambda_i \delta_{\boldsymbol{x}_i}$ et $\mu = \sum_{j=1}^{m} \mu_j \delta_{\boldsymbol{y}_j}$. La forme $(.,.)_{\tilde{\Lambda}}$ est symétrique et positive car k est symétrique et de type positif. Il s'agit donc d'un semi-produit scalaire.

Muni du semi-produit scalaire $(.,.)_{\tilde{\Lambda}}$, $\tilde{\Lambda}$ est un espace préhilbertien non séparé. Rappelons que le séparé d'un espace préhilbertien non séparé $\tilde{\mathcal{E}}$ est par définition l'espace préhilbertien quotient $\tilde{\mathcal{E}}/\mathcal{K}$, où $\mathcal{K}$ est le sous-espace vectoriel $\mathcal{K} = \{f \in \tilde{\mathcal{E}},$ tels que $(f, g)_{\tilde{\mathcal{E}}} = 0, \forall g \in \tilde{\mathcal{E}}\}$. Notons Λ le séparé complété de $\tilde{\Lambda}$ et $\mathcal{F}$ son dual. $\mathcal{F}$ est un espace de Hilbert (en tant que dual de Λ). Puisque toute forme linéaire sur $\tilde{\Lambda}$ s'identifie à une fonction de $\mathbb{R}^{\mathbb{X}}$, $\mathcal{F}$ s'identifie à un espace hilbertien de fonctions.

Montrons finalement que k est le noyau reproduisant de $\mathcal{F}$. Soit ϑ l'injection canonique de $\tilde{\Lambda}$ dans son séparé complété Λ. Soit ϱ^{-1} l'isométrie bijective de Λ sur $\mathcal{F}$. Alors, $\forall f \in \mathcal{F}$,

$$f(\boldsymbol{x}) = \langle \vartheta\delta_{\boldsymbol{x}}, f\rangle_{\Lambda,\mathcal{F}} = (\varrho^{-1}\vartheta\delta_{\boldsymbol{x}}, f)_{\mathcal{F}}\,, \tag{3.2}$$

où $\langle\cdot,\cdot\rangle$ désigne le produit de dualité. Pour montrer que k est un noyau reproduisant, il suffit d'établir, d'après (3.2), que $k(\boldsymbol{x}, .) = \varrho^{-1}\vartheta\delta_{\boldsymbol{x}}$. Or d'après la définition du semi-produit scalaire sur $\tilde{\Lambda}$,

$$k(\boldsymbol{x}, \boldsymbol{y}) = (\delta_{\boldsymbol{x}}, \delta_{\boldsymbol{y}})_{\tilde{\Lambda}} = (\vartheta\delta_{\boldsymbol{x}}, \vartheta\delta_{\boldsymbol{y}})_{\Lambda} = \langle \vartheta\delta_{\boldsymbol{y}}, \varrho^{-1}\vartheta\delta_{\boldsymbol{x}}\rangle_{\Lambda,\mathcal{F}}\,.$$

□

Ce théorème est d'utilisation constante pour établir des modèles de type boîte noire par régression régularisée dans des espaces hilbertiens à noyau reproduisant. En pratique, on choisit une fonction symétrique de type positif et le théorème garantit l'existence d'un espace hilbertien admettant cette fonction comme noyau reproduisant. L'intérêt principal de l'espace hilbertien à noyau reproduisant ainsi construit est la simplicité des calculs de produits scalaires. En effet, si $f = \sum_{i=1}^{n} \lambda_i k(\boldsymbol{x}_i, .)$ et $g = \sum_{i=1}^{m} \mu_i k(\boldsymbol{y}_i, .)$ alors

$$(f, g)_{\mathcal{F}} = \sum_{i,j=1}^{n,m} \lambda_i \mu_j k(\boldsymbol{x}_i, \boldsymbol{y}_j) \,.$$

Un cas important d'application de ce théorème apparaît lorsque l'on choisit un noyau reproduisant sous la forme d'une fonction de covariance. Nous verrons dans la section 3.5.5 l'utilisation pratique de cette possibilité.

Réciproquement, est-il possible d'associer à tout espace hilbertien $\mathcal{F}$ de fonctions sur $\mathbb{X}$ un noyau reproduisant k tel que $\mathcal{F}$ soit engendré par les fonctions $k(\boldsymbol{x}, .)$ lorsque $\boldsymbol{x}$ parcourt $\mathbb{X}$?

Théorème 8. *Soit $\mathcal{F}$ un espace hilbertien de fonctions $f : \mathbb{X} \to \mathbb{R}$ vérifiant*

$$\forall \boldsymbol{x} \in \mathbb{X}, \exists M_{\boldsymbol{x}} > 0, \ \textit{tel que } |f(x)| \leq M_{\boldsymbol{x}} \|f\|_{\mathcal{F}} \,, \quad \forall f \in \mathcal{F} \,. \tag{3.3}$$

Alors $\mathcal{F}$ possède un noyau reproduisant k.

Démonstration. La condition (3.3) exprime que la forme linéaire $\delta_{\boldsymbol{x}} : f \mapsto f(\boldsymbol{x})$ est continue sur $\mathcal{F}$, donc qu'elle est élément de $\mathcal{F}^*$. Si ϱ est l'opérateur de dualité de $\mathcal{F}$ sur $\mathcal{F}^*$, posons $\tilde{k}(\boldsymbol{x}) = \varrho^{-1}\delta_{\boldsymbol{x}}$ et $k(\boldsymbol{x}, \boldsymbol{y}) = \tilde{k}(\boldsymbol{x})(\boldsymbol{y})$. Alors k est bien le noyau reproduisant de $\mathcal{F}$, puisque $\forall f \in \mathcal{F}$, $f(\boldsymbol{x}) = \langle \delta_{\boldsymbol{x}}, f \rangle_{\mathcal{F}^*,\mathcal{F}} = (\varrho^{-1}\delta_{\boldsymbol{x}}, f)_{\mathcal{F}} = (k(\boldsymbol{x}, .), f)_{\mathcal{F}}$. □

Remarquons que le théorème 8 établit la réciproque de la proposition 25. Il est d'ailleurs fréquent de donner la définition d'un espace à noyau reproduisant à partir de l'hypothèse (3.3) qui requiert la continuité de la forme linéaire d'évaluation ponctuelle.

3.2 Constructions d'espaces hilbertiens à noyau reproduisant

3.2.1 Exemples en dimension finie

Considérons un espace vectoriel $\mathcal{F}$ de dimension finie l. Tout élément $f \in \mathcal{F}$ peut s'écrire à l'aide d'une base $B = \{r_1, \ldots, r_l\}$ de $\mathcal{F}$

$$f = \sum_{i=1}^{l} b_i r_i \,.$$

Notons $\boldsymbol{b} = (b_1, \ldots, b_l)^\mathsf{T}$ le vecteur représentant f dans la base B. Soit $\boldsymbol{K} = (k_{i,j})_{i,j=1}^{l}$ une matrice symétrique définie positive (une telle matrice peut correspondre à une matrice de covariance). $\boldsymbol{K}$ est inversible et son inverse est noté $\boldsymbol{Q} = (q_{i,j})_{i,j=1}^{l}$. Muni du produit scalaire

$$(f_1, f_2)_{\mathcal{F}} = \boldsymbol{b}_1^\mathsf{T} \boldsymbol{K}^{-1} \boldsymbol{b}_2 = \boldsymbol{b}_1^\mathsf{T} \boldsymbol{Q} \boldsymbol{b}_2 \,,$$

$\mathcal{F}$ est un espace hilbertien. $\boldsymbol{Q}$ est la matrice de produit scalaire dans B telle que $q_{i,j} = (r_i, r_j)_{\mathcal{F}}$.

Exemple 1

Soit $\mathcal{F}$ un espace vectoriel de fonctions admettant une base de fonctions $r_1(\boldsymbol{x}), \ldots, r_l(\boldsymbol{x})$, $\boldsymbol{x} \in \mathbb{X}$. Toute fonction de $\mathcal{F}$ peut s'écrire $f(\boldsymbol{x}) = \sum_i b_i r_i(\boldsymbol{x})$. Alors $\mathcal{F}$ admet le noyau reproduisant

$$k(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{r}(\boldsymbol{x})^\mathsf{T} \boldsymbol{K} \boldsymbol{r}(\boldsymbol{y}) \,, \tag{3.4}$$

où $\forall \boldsymbol{x} \in \mathbb{X}$, $\boldsymbol{r}(\boldsymbol{x}) = (r_1(\boldsymbol{x}), \ldots, r_l(\boldsymbol{x}))^\mathsf{T}$. En effet, comme fonction de $\boldsymbol{y}$, $k(\boldsymbol{x}, \boldsymbol{y})$ défini par (3.4) est représenté par le vecteur $\boldsymbol{K}\boldsymbol{r}(\boldsymbol{x})$ dans la base $r_1(\boldsymbol{x}), \ldots, r_l(\boldsymbol{x})$ et par conséquent

$$(f, k(\boldsymbol{x}, .))_\mathcal{F} = \boldsymbol{b}^\mathsf{T} \boldsymbol{K}^{-1} \boldsymbol{K} \boldsymbol{r}(\boldsymbol{x}) = f(\boldsymbol{x}) \,.$$

Le cas particulier $\boldsymbol{K} = \boldsymbol{I}_l$ correspond au choix de la norme $\|f\|_\mathcal{F}^2 = \sum_{i=1}^l b_i^2$ (voir aussi la section 3.4.1).

Exemple 2

Soit $\mathcal{F}$ un espace vectoriel de fonctions réelles sur un ensemble fini $\mathbb{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_l\}$ ($\mathcal{F} = \mathbb{R}^l$). Tout élément de $\mathcal{F}$ est représenté par ses coordonnées $f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_l)$ dans la base canonique $\{r_1, \ldots, r_l\}$. Le produit scalaire dans $\mathcal{F}$ s'écrit $(f_1, f_2)_\mathcal{F} = \sum_{i,j=1}^l f(\boldsymbol{x}_i) q_{i,j} f(\boldsymbol{x}_j)$. Alors $\mathcal{F}$ admet le noyau reproduisant $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = k_{i,j}$. On vérifie en effet que

$$(f, k(\boldsymbol{x}_l, .))_\mathcal{F} = \sum_{i,j=1}^l f(\boldsymbol{x}_i) q_{i,j} k(\boldsymbol{x}_l, \boldsymbol{x}_j) = \sum_{i,j=1}^l f(\boldsymbol{x}_i) q_{i,j} k_{l,j} = f(\boldsymbol{x}_l) \,.$$

Ces deux exemples sont construits de la même manière. Dans le premier, on utilise une base finie de fonctions. Dans l'autre cas, on utilise le fait que la fonction est déterminée par un nombre fini de ses coordonnées de $\mathbb{X}$. Un cas intéressant apparaît lorsqu'un nombre fini de coordonnées de $\mathbb{X}$ détermine de manière unique les coordonnées dans une base de fonctions (par exemple, quand cette base de fonctions est une base de polynômes de degré fini).

3.2.2 Exemples en dimension infinie

Exemple 1

Pour cet exemple (d'après Hirsch et Lacombes, 1999), on se fixe une mesure borélienne μ sur $\mathbb{R}$. Si $h \in L^2(\mu)$, on note f_h la fonction sur $\mathbb{R}$ définie par

$$f_h(x) = \int e^{\imath t x} h(t) d\mu(t).$$

L'application $h \mapsto f_h$ est injective. En effet, si $\forall x \int e^{\imath t x} h(t) d\mu(t) = 0$, alors h est nulle μ-presque partout (voir par exemple (Hirsch et Lacombes, 1999)). Soit alors $\mathcal{F} = \{f_h, \text{ avec } h \in L^2(\mu)\}$. Pour $h_1, h_2 \in L^2(\mu)$, on pose

$$(f_{h_1}, f_{h_2})_\mathcal{F} = \int h_1 \overline{h_2} d\mu.$$

On vérifie alors aisément que $\mathcal{F}$ est un espace de Hilbert. De plus, $\mathcal{F}$ admet le noyau reproduisant

$$k(x, y) = \int e^{\imath t (x-y)} d\mu(t)$$

car $t \mapsto e^{-\imath t y} \in L^2(\mu)$ et $(f_h, k(., y))_\mathcal{F} = \int h e^{\imath t y} d\mu = f_h(y)$. (Cet exemple sera revu dans la section 3.3.3 consacrée à la représentation spectrale des noyaux.)

Exemple 2

Nous présentons maintenant un exemple inspiré de (Adler, 1990) où une fonction de covariance est utilisée comme noyau reproduisant. Soit $W(t)$ un mouvement brownien défini $\forall t \in T = [0,1]$. La covariance de $W(t)$ est donc $k(t,s) = \min(s,t)$. On s'intéresse à la nature de l'espace de Hilbert $\mathcal{F}$ admettant la fonction $k(t,s)$ comme noyau reproduisant. Considérons deux fonctions de $\mathcal{F}$, $f_1(.) = \sum_{i=1}^{n} \lambda_i k(t_i,.)$ et $f_2(.) = \sum_{j=1}^{m} \mu_j k(s_j,.)$. Leur produit scalaire dans $\mathcal{F}$ s'écrit

$$(f_1, f_2)_{\mathcal{F}} = \sum_{i,j=1}^{n,m} \lambda_i \mu_j \min(t_i, s_j)\,.$$

En remarquant que la dérivée de $k(t,s)$ par rapport à t peut s'écrire $\mathbb{1}_{[0,s]}(t)$, on obtient une nouvelle expression du produit scalaire sous la forme :

$$\begin{aligned}(f_1, f_2)_{\mathcal{F}} &= \sum_{i,j=1}^{n,m} \lambda_i \mu_j \int_0^1 \mathbb{1}_{[0,t_i]}(t)\mathbb{1}_{[0,s_j]}(t)dt \\ &= \int_0^1 \left(\sum_i \lambda_i \mathbb{1}_{[0,t_i]}(t) \sum_j \mu_j \mathbb{1}_{[0,s_j]}(t)\right) dt \\ &= \int_0^1 f_1^{(1)}(t) f_2^{(1)}(t)dt\,.\end{aligned}$$

Par conséquent, $\mathcal{F}$ est constitué de l'ensemble de fonctions

$$\left\{ f \in \mathbb{R}^T,\ \text{telle que } f(t) = \int_0^t f^{(1)}(s)ds \text{ et } \int_0^1 (f^{(1)}(t))^2 dt < \infty \right\}, \tag{3.5}$$

muni du produit scalaire

$$(f_1, f_2)_{\mathcal{F}} = \int_0^1 f_1^{(1)}(t) f_2^{(1)}(t)dt\,.$$

En effet, il est immédiat de vérifier que $k(s,t)$ est dans l'ensemble défini par (3.5) et que de plus, $\forall t \in T$,

$$(f, k(t,.))_{\mathcal{F}} = \int_0^1 f^{(1)}(s)\mathbb{1}_{[0,t]}(s)ds = f(t)\,.$$

$\mathcal{F}$ est constitué de fonctions presque partout dérivables. Cet exemple illustre une caractéristique importante : l'espace hilbertien engendré par une fonction de covariance k est un espace de fonctions possédant des propriétés *comparables* à celles des trajectoires du processus aléatoire gaussien admettant k comme covariance.

Cet exemple est également important parce qu'il est à la base de la théorie de nombreux types de *splines*. Les splines sont en effet des solutions dans un espace hilbertien à noyau reproduisant d'un problème d'approximation régularisé (par la norme de l'espace hilbertien). Ainsi, l'espace $\mathcal{F}$ construit ci-dessus correspond à des splines linéaires en dimension 1. Les solutions régularisées d'un problème d'approximation seront dans ce cas des fonctions linéaires par morceaux. Nous n'abordons pas explicitement la théorie des splines dans cette présentation (un point de vue plus général est celui de la régression régularisée dans des espaces hilbertiens à noyau reproduisant, présenté dans la section 3.4). Notons que la théorie des splines nécessite en principe d'introduire la notion de noyau conditionnellement positif, qui sera présentée au chapitre 4. ((Wahba, 1990) constitue une référence classique sur le sujet.)

3.2.3 Notion d'espace des caractéristiques

Cette section explique de manière informelle le rôle fondamental des noyaux reproduisants dans la théorie de l'apprentissage (Vapnik, 1995). La fonction principale d'une procédure d'apprentissage est de classer des données de manière automatique. Ces données correspondent à des catégories dépendant de certains facteurs, représentés comme dans le cas de la modélisation boîte noire par un vecteur $\boldsymbol{x} \in \mathbb{X}$. On met en correspondance l'ensemble des catégories avec un ensemble d'étiquettes numériques $\mathcal{L}$ (par exemple, s'il y a deux catégories, on associe un ensemble d'étiquettes $\mathcal{L} = \{+1, -1\}$). Les données sont des couples $(\boldsymbol{x}_i, f^{\text{obs}}_{\boldsymbol{x}_i}) \in \mathbb{X} \times \mathcal{L}$, $i = 1, \ldots, n$.

Un problème à deux catégories est dit linéairement séparable s'il existe un hyperplan dans l'espace des facteurs $\mathbb{X}$, tel que les données avec des étiquettes différentes se trouvent de part et d'autre de cet hyperplan. On cherche donc une fonction linéaire $\boldsymbol{x} \mapsto f(\boldsymbol{x})$ telle que, $\forall i \in \{1, \ldots, n\}$, $f(\boldsymbol{x}_i) f^{\text{obs}}_{\boldsymbol{x}_i} > 0$. Il existe pour cela plusieurs méthodes (réseaux de neurones formels, détection linéaire à partir des moments d'ordre deux, machines à vecteurs de support) et dans la plupart des cas, il est facile de montrer (voir (Vapnik, 1995) et la section 3.4.2) que l'on obtient des problèmes d'approximation linéaire pouvant s'exprimer uniquement en fonction des produits scalaires $(\boldsymbol{x}, \boldsymbol{y})_{\mathbb{R}^d}$, $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{X}$. Lorsque le problème n'est plus linéairement séparable dans $\mathbb{X}$, l'approche utilisée dans la plupart des ouvrages de la littérature de la théorie de l'apprentissage est connue sous le nom d'*astuce du noyau*, traduction du terme anglais *kernel trick* (Aizerman et al., 1964), popularisée dans le cadre de la classification de données par (Boser et al., 1992).

Cette approche passe par l'introduction d'une fonction $\phi : \boldsymbol{x} \in \mathbb{X} \mapsto \phi(\boldsymbol{x}) \in \mathcal{F}$ qui transforme l'espace des facteurs initial $\mathbb{X}$ en un espace dans lequel la séparation linéaire redevient possible. $\mathcal{F}$ est appelé *espace des caractéristiques* (*feature space* en anglais) et peut être un espace de dimension infinie, typiquement un espace de fonctions elles-mêmes définies sur $\mathbb{X}$. Le problème de séparation linéaire dans $\mathcal{F}$ fait alors intervenir des produits scalaires $(\phi(\boldsymbol{x}), \phi(\boldsymbol{y}))_{\mathcal{F}}$. L'*astuce du noyau* consiste à ne jamais expliciter les transformations $\phi(\boldsymbol{x})$ puisque seule la fonction $k(\boldsymbol{x}, \boldsymbol{y}) = (\phi(\boldsymbol{x}), \phi(\boldsymbol{y}))_{\mathcal{F}}$ a besoin d'être évaluée. Or dans un espace de Hilbert à noyau reproduisant, on a la propriété $k(\boldsymbol{x}, \boldsymbol{y}) = (k(\boldsymbol{x}, .), k(\boldsymbol{y}, .))_{\mathcal{F}}$. L'idée est donc de poser $\phi(\boldsymbol{x}) = k(\boldsymbol{x}, .)$ et l'espace des caractéristiques est donc un espace hilbertien à noyau reproduisant. Récemment, de nombreux problèmes ont bénéficié de ce point de vue. Parmi ces méthodes, les *machines à vecteurs de support* (Vapnik, 1995 ; Schölkopf et Smola, 2002) connaissent un succès indéniable de même que les méthodes d'analyse en composantes principales à noyau, *kernel principal component analysis* (*KPCA*) en anglais (Schölkopf et al., 1998 ; Mika et al., 1999 ; Müller et al., 2001). La théorie de l'apprentissage a donc joué un rôle important dans le développement des méthodes à noyau[1].

3.3 Quelques représentations d'un noyau reproduisant

Nous présentons dans les paragraphes suivants des notions sur les représentations possibles d'un noyau reproduisant. Ceci facilitera la compréhension du rôle de la norme d'un espace hilbertien à noyau reproduisant donné (dans la section 3.4, nous utiliserons des normes d'espaces à noyau reproduisant pour régulariser des problèmes de régression).

[1] On peut regretter cependant que certaines méthodes classiques, notamment les méthodes de régression régularisée avec attache aux données quadratique aient été réinventées. Par exemple, les méthodes dites « proximal vector machines », ou « least square SVM » sont clairement de simples reformulations de méthodes bien connues.

3.3.1 Représentation de Mercer sur des domaines compacts

Lorsque le domaine de définition $\mathbb{X}$ des fonctions de $\mathcal{F}$ est compact, il existe une représentation des noyaux reproduisants utilisant le théorème de Mercer que nous souhaitons rappeler. Dans la littérature cette représentation est souvent mentionnée malgré le caractère restrictif de l'hypothèse de compacité de $\mathbb{X}$ [2]. Nous adoptons le point de vue formel de la théorie des opérateurs de Hilbert–Schmidt, telle qu'elle est exposée dans la plupart des ouvrages d'analyse fonctionnelle (voir par exemple (Riesz et Nagy-Sz., 1965 ; Hirsch et Lacombes, 1999 ; Aubin, 2000)).

Définition 25 (Opérateur de Hilbert–Schmidt). Soit $\mathcal{E}$ un espace de Hilbert, séparable, de dimension infinie. $\mathcal{L}(\mathcal{E})$ désigne l'ensemble des opérateurs linéaires sur $\mathcal{E}$. Si $(e_n)_{n\in\mathbb{N}}$ est une base hilbertienne de $\mathcal{E}$, on dit qu'un opérateur $T \in \mathcal{L}(\mathcal{E})$ est un *opérateur de Hilbert–Schmidt* lorsque la série numérique $\sum_{n=0}^{\infty}\|Te_n\|^2$ est convergente. On démontre (voir par exemple (Aubin, 2000)) que cette définition est indépendante de la base hilbertienne considérée.

Proposition 26. *Les opérateurs de Hilbert–Schmidt sont compacts.*

Démonstration. Il s'agit d'un résultat classique (Hirsch et Lacombes, 1999 ; Aubin, 2000). □

Définition 26 (Opérateur intégral). Soit $\mathbb{X}$ un domaine compact de $\mathbb{R}^d$ et $k : \mathbb{X}\times\mathbb{X} \to \mathbb{R}$ une fonction continue de carré sommable sur $\mathbb{X}\times\mathbb{X}$. On définit l'opérateur linéaire intégral $T_k : L^2(\mathbb{X}) \to L^2(\mathbb{X})$ par la relation,

$$T_k f(\boldsymbol{x}) = \int_{\mathbb{X}} k(\boldsymbol{x},\boldsymbol{y}) f(\boldsymbol{y}) dy\,, \quad \forall \boldsymbol{x} \in \mathbb{X}.$$

La fonction k est appelée *noyau* de l'opérateur intégral T_k.

Théorème 9 (du noyau). *Soit $\mathbb{X}$ un compact de $\mathbb{R}^d$. ($L^2(\mathbb{X})$ est donc séparable.) Un opérateur $T \in \mathcal{L}(L^2(\mathbb{X}))$ est de Hilbert–Schmidt si et seulement si il est associé à un opérateur intégral à noyau $k \in L^2(\mathbb{X}\times\mathbb{X})$.*

Démonstration. Voir Aubin (2000). □

Proposition 27. *Si $k \in L^2(\mathbb{X}\times\mathbb{X})$ est symétrique, T_k est un opérateur auto-adjoint. Si de plus k est une fonction de type positif, alors T_k est un opérateur positif (c'est-à-dire $(T_k f, f)_{L^2(\mathbb{X})} \geq 0$, $\forall f \in L^2(\mathbb{X})$).*

Démonstration. Il s'agit encore de résultats classiques sur les opérateurs de Hilbert–Schmidt (Hirsch et Lacombes, 1999). □

Rappelons ensuite les propriétés des valeurs propres d'un opérateur auto-adjoint compact.

Théorème 10 (Valeurs propres des opérateurs auto-adjoints compacts). *Soit T un opérateur auto-adjoint compact quelconque sur un espace hilbertien $\mathcal{E}$ et $\mathrm{vp}(T)$ l'ensemble de ses valeurs propres. Pour chaque valeur propre $\lambda \in \mathrm{vp}(T)$, on note $\mathcal{E}_\lambda$ l'espace propre associé.*

- *L'ensemble $\mathrm{vp}(T)$ est une partie infinie, dénombrable et bornée de $\mathbb{R}$, dont le seul point d'accumulation est 0.*

[2] (Courant et Hilbert, 1965) traite à un niveau élémentaire des opérateurs intégraux et du théorème de Mercer.

- *Si T est positif, alors toutes ses valeurs propres sont positives ou nulles.*
- *Tous les sous-espaces propres de T correspondant à des valeurs propres non nulles sont de dimension finie.*
- *Deux sous-espaces propres de T correspondant à des valeurs propres différentes sont orthogonaux*
- *Soit, pour chaque valeur propre non nulle λ de T, P_λ le projecteur orthogonal sur $\mathcal{E}_\lambda$. Alors $T = \sum_{\lambda \in \mathrm{vp}(T)0} \lambda P_\lambda$, au sens des familles sommables dans $\mathcal{L}(\mathcal{E})$.*

Démonstration. Voir (Hirsch et Lacombes, 1999). □

Corollaire 1. *L'espace $\overline{\mathrm{Im}\, T}$ admet une base hilbertienne dénombrable $(\varphi_n)_{n\in\mathbb{N}}$ formée de vecteurs propres de T correspondant à des valeurs propres non nulles. Pour tout $f \in \overline{\mathrm{Im}\, T}$,*

$$f = \sum_{n\in\mathbb{N}} (f, \varphi_n)\varphi_n \,.$$

Nous supposerons par la suite que k est symétrique et que T_k n'est pas de rang fini. Soit $(\varphi_n)_{n\in\mathbb{N}}$ une base hilbertienne de $\overline{\mathrm{Im}\, T_k}$ formé de vecteurs propres de T_k et $(\lambda_n)_{n\in\mathbb{N}}$ la suite des valeurs propres non nulles correspondantes. Si k est continu, symétrique et de type positif, k définit un noyau reproduisant. Le théorème de Mercer est fondé sur la représentation spectrale de l'opérateur T_k et permet donc d'établir une représentation d'un noyau reproduisant $k \in L^2(\mathbb{X}\times\mathbb{X})$ lorsque $\mathbb{X}$ est compact.

Théorème 11 (Mercer). *Soit T_k un opérateur de Hilbert–Schmidt sur $L^2(\mathbb{X})$ avec $\mathbb{X}$ compact. Avec les notations et hypothèses ci-dessus,*

$$\iint |k(\boldsymbol{x},\boldsymbol{y})|^2 d\boldsymbol{x} d\boldsymbol{y} = \sum_{n=0}^{+\infty} \lambda_n^2$$

et

$$k(\boldsymbol{x},\boldsymbol{y}) = \sum_{n=0}^{+\infty} \lambda_n \varphi_n(\boldsymbol{x})\overline{\varphi_n(\boldsymbol{y})}\,. \tag{3.6}$$

Ces séries convergent absolument et uniformément sur $L^2(\mathbb{X},\mathbb{X})$.

Démonstration. Voir (Riesz et Nagy-Sz., 1965 ; Aubin, 2000). □

La représentation (3.6) peut être vue comme une généralisation de (3.4). On peut montrer de plus que $\mathcal{F}$ peut être caractérisé à l'aide des valeurs et des vecteurs propres de T_k (voir Cucker et Smale (2001) par exemple). En effet, $\mathcal{F}$ s'identifie à l'espace

$$\left\{ f \in L^2(\mathbb{X}) \text{ telles que } f = \sum_{n=0}^{\infty} c_n \varphi_n \,,\ \left(\frac{c_n}{\sqrt{\lambda_n}}\right) \in \ell^2(\mathbb{N}) \right\}$$

avec la norme

$$\|f\|_{\mathcal{F}}^2 = (\sum_{n=0}^{\infty} c_n\varphi_n, \sum_{m=0}^{\infty} c_m\varphi_m)_{\mathcal{F}} = \sum_{n=0}^{\infty} \frac{c_n^2}{\lambda_n} < +\infty\,.$$

Par conséquent, les fonctions de l'espace hilbertien $\mathcal{F}$ à noyau reproduisant k peuvent être représentées dans une base $(\varphi_n)_{n\in\mathbb{N}}$. De plus, les coefficient c_n de cette représentation décroissent

d'autant plus vite que les valeurs propres λ_n tendent rapidement vers zéro. La représentation de Mercer permet donc de mettre en relation un noyau et la régularité des fonctions générées par ce noyau.

3.3.2 Utilisation de frames

La théorie des frames (Duffin et Schaeffer, 1952 ; Daubechies, 1992 ; Mallat, 1999) a été introduite dans le cadre des problèmes de reconstruction de fonctions à partir d'échantillonnages réguliers ou irréguliers. Il n'est donc pas surprenant que l'on puisse établir des liens entre la théorie des frames et celle des noyaux reproduisants qui est également utilisée pour l'approximation de fonctions (ce que nous verrons dans la section 3.4). Ces liens ont par exemple été explicités dans (Gao et al., 2001 ; Rakotomamonjy et Canu, 2002). Dans les paragraphes suivants, nous rappelons quelques éléments de la théorie des frames (Mallat, 1999) et la démarche permettant de construire un noyau reproduisant à partir de la donnée d'une frame.

Une frame d'un espace hilbertien $\mathcal{F}$ de fonctions définies sur $\mathbb{X}$ est une suite $(\phi_j)_{j\in\mathcal{J}}$ d'éléments de $\mathcal{F}$, où $\mathcal{J}$ est un ensemble fini ou dénombrable, telle que toute fonction $f \in \mathcal{F}$ est caractérisée entièrement par les produits scalaires $((f,\phi_j)_{\mathcal{F}})_{j\in\mathcal{J}}$. Les éléments de la frame ne sont pas nécessairement linéairement indépendants. La définition précise d'une frame est rappelée ci-dessous.

Définition 27. Une suite de fonctions $(\phi_j)_{j\in\mathcal{J}}$ est une *frame* de $\mathcal{F}$ s'il existe deux constantes $0 < A \leq B < \infty$ telles que

$$\forall f \in \mathcal{F} \quad A\|f\|_{\mathcal{F}}^2 \leq \sum_{j\in\mathcal{J}} |(f,\phi_j)_{\mathcal{F}}|^2 \leq B\|f\|_{\mathcal{F}}^2 . \tag{3.7}$$

Une *frame ajustée* est telle que $A = B$. A et B sont les *bornes de frame*.

La relation (3.7) est une condition de stabilité. Comme $((f,\phi_j)_{\mathcal{F}})_{j\in\mathcal{J}} \in \ell^2(\mathcal{J})$, il est licite de définir l'opérateur de frame U par

$$\begin{array}{rccc} U: & \mathcal{F} & \to & \ell^2(\mathcal{J}) \\ & f & \mapsto & ((f,\phi_j)_{\mathcal{F}})_{j\in\mathcal{J}} . \end{array}$$

On définit également U^*, l'adjoint de U tel que $(Uf,x)_{\ell^2(\mathcal{J})} = (f,U^*x)_{\mathcal{F}}$.

Exemple. Une base orthonormale d'un espace hilbertien $\mathcal{F}$ est un cas particulier de frame avec $A = B = 1$. Par exemple, la famille $\{T^{-1}h_T(t_nT), n \in \mathbb{N}\}$ où $h_T(t) = sin(\pi t/T)/(\pi t/T)$ est une frame de l'espace des fonctions de $L^2(\mathbb{R})$ dont la transformée de Fourier est à support dans $[-\pi/T, \pi/T]$. Dans ce cas $Uf = (f(nT))_{n\in\mathbb{N}}$.

Théorème 12. *Soit $(\phi_j)_{j\in\mathcal{J}}$ une frame de $\mathcal{F}$, avec des bornes A et B. On définit la frame duale $(\tilde{\phi}_j)_{j\in\mathcal{J}}$ de $\mathcal{F}$ par*

$$\tilde{\phi}_j = (U^*U)^{-1}\phi_j .$$

Alors $\forall f \in \mathcal{F}$,

$$\frac{1}{B}\|f\|_{\mathcal{F}}^2 \leq \sum_{j\in\mathcal{J}} |(f,\tilde{\phi}_j)_{\mathcal{F}}|^2 \leq \frac{1}{A}\|f\|_{\mathcal{F}}^2$$

et

$$f = \sum_{j\in\mathcal{J}} (f,\tilde{\phi}_j)_{\mathcal{F}}\phi_j = \sum_{j\in\mathcal{J}} (f,\phi_j)_{\mathcal{F}}\tilde{\phi}_j .$$

Démonstration. Voir par exemple (Mallat, 1999). □

Théorème 13. *Soit $\mathcal{F}$ un espace de Hilbert et $(\phi_j)_{j\in\mathcal{J}}$ une frame de $\mathcal{F}$. Si $\forall \boldsymbol{x} \in \mathbb{X}$,*

$$\Big\| \sum_{j\in\mathcal{J}} \phi_j(\boldsymbol{x})\tilde{\phi}_j(.) \Big\|_{\mathcal{F}} < \infty \,,$$

alors $\mathcal{F}$ est un espace hilbertien à noyau reproduisant.

Démonstration. Soit $f \in \mathcal{F}$. Alors $\forall \boldsymbol{x} \in \mathbb{X}$,

$$|f(\boldsymbol{x})| = \Big| \sum_{j\in\mathcal{J}} (f, \tilde{\phi}_j)_{\mathcal{F}} \phi_j(\boldsymbol{x}) \Big| \,,$$

et par suite,

$$\begin{aligned} |f(\boldsymbol{x})| &= \Big| (f, \sum_{j\in\mathcal{J}} \phi_j(\boldsymbol{x})\tilde{\phi}_j)_{\mathcal{F}} \Big| \\ &\leq \|f\|_{\mathcal{F}} \Big\| \sum_{j\in\mathcal{J}} \phi_j(\boldsymbol{x})\tilde{\phi}_j \Big\|_{\mathcal{F}} \end{aligned}$$

Donc $\mathcal{F}$ est un espace hilbertien à noyau reproduisant. □

Le théorème suivant (voir par exemple Rakotomamonjy et Canu (2002)) donne la forme du noyau reproduisant en fonction des éléments de la frame.

Théorème 14. *Soit $\mathcal{F}$ un espace hilbertien séparable à noyau reproduisant. S'il existe une frame $(\phi_j)_{j\in\mathcal{J}}$ de $\mathcal{F}$, alors le noyau reproduisant admet la représentation*

$$k(\boldsymbol{x}, \boldsymbol{y}) = \sum_{j\in\mathcal{J}} \tilde{\phi}_j(\boldsymbol{x})\phi_j(\boldsymbol{y}) \,. \tag{3.8}$$

Démonstration. Pour toute fonction $f \in \mathcal{F}$, et pour tout $\boldsymbol{x} \in \mathbb{X}$

$$\begin{aligned} f(\boldsymbol{x}) &= \sum_{j\in\mathcal{J}} (f, \tilde{\phi}_j)_{\mathcal{F}} \phi_j(\boldsymbol{x}) \\ &= (f, \sum_{i\in\mathcal{J}} \phi_j(\boldsymbol{x})\tilde{\phi}_j)_{\mathcal{F}} \,. \end{aligned}$$

Ceci montre que le noyau reproduisant de $\mathcal{F}$ peut bien se mettre sous la forme (3.8). □

Les théorèmes précédents permettent par conséquent des représentations plus générales que celle obtenue avec le théorème de Mercer, puisque la famille de fonctions ϕ_j ne constitue pas nécessairement une base de l'espace $\mathcal{F}$. Remarquons également que lorsque la frame est de dimension finie (par suite, $\mathcal{F}$ est de dimension finie), on retrouve les résultats établis dans la section 3.2.1, et en particulier la relation (3.4).

3.3.3 Représentation spectrale pour les noyaux invariants par translation

La représentation vue dans la section 3.3.1, valable pour des espaces $\mathbb{X}$ compacts, n'est pas adaptée lorsque l'on utilise les noyaux classiques définis sur $\mathbb{R}^d \times \mathbb{R}^d$, ce qui est habituellement le cas. Si le noyau est invariant par translation sur $\mathbb{R}^d \times \mathbb{R}^d$, le théorème de Bochner, rappelé dans la section 2.4.2, peut être utilisé pour établir une représentation du noyau dans le domaine de Fourier. L'hypothèse d'invariance par translation permet d'écrire $k(\boldsymbol{x}, \boldsymbol{y}) = k(\|\boldsymbol{x} - \boldsymbol{y}\|)$. Si $k(\boldsymbol{h}) \in L^2(\mathbb{R}^d)$, ce que nous supposerons par la suite, alors

$$k(\boldsymbol{x} - \boldsymbol{y}) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} e^{\imath(\boldsymbol{u}, \boldsymbol{x}-\boldsymbol{y})} \tilde{k}(u) d\boldsymbol{u}\,,$$

où $\tilde{k}(\boldsymbol{u})$ est la transformée de Fourier de $k(\boldsymbol{h})$.

Proposition 28. *Soit $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ un noyau reproduisant invariant par translation et $\mathcal{F}$ l'espace hilbertien engendré par ce noyau. Si $k(\boldsymbol{h}) \in L^2(\mathbb{R}^d)$, alors $\mathcal{F} \subset L^2(\mathbb{R}^d)$ et de plus, $\forall f \in \mathcal{F}$,*

$$\|f\|_{\mathcal{F}}^2 = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} |\tilde{f}(\boldsymbol{u})|^2 \tilde{k}(\boldsymbol{u})^{-1}\ d\boldsymbol{u}\,. \tag{3.9}$$

Démonstration. (Inspirée de (Aubin, 2000).) Rappelons que $\tilde{\mathcal{F}}$ est l'espace vectoriel des combinaisons linéaires de k. Si $k(\boldsymbol{h}) \in L^2(\mathbb{R}^d)$, alors $\tilde{\mathcal{F}} \subset L^2(\mathbb{R}^d)$ et $\forall f \in \tilde{\mathcal{F}}$, la transformée de Fourier de f s'écrit

$$\tilde{f}(\boldsymbol{u}) = \sum_{i=1}^{n} \lambda_i e^{-\imath(\boldsymbol{u}, \boldsymbol{x}_i)} \tilde{k}(\boldsymbol{u})\,.$$

Par suite, on vérifie très simplement que $\forall f \in \tilde{\mathcal{F}}$

$$\|f\|_{\mathcal{F}}^2 = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} |\tilde{f}(\boldsymbol{u})|^2 \tilde{k}(\boldsymbol{u})^{-1} d\boldsymbol{u}\,.$$

Comme $\tilde{k}(\boldsymbol{u})$ est borné et positif, il existe $c > 0$ tel que $\tilde{k}(\boldsymbol{u})^{-1} \geq c$, $\forall \boldsymbol{u} \in \mathbb{R}^d$. Si $f \in \tilde{\mathcal{F}}$,

$$||f||_2^2 = \frac{1}{(2\pi)^d} \int |\tilde{f}(\boldsymbol{u})|^2 d\boldsymbol{u} \leq \frac{1}{c(2\pi)^d} \int |\tilde{f}(\boldsymbol{u})|^2 \tilde{k}(\boldsymbol{u})^{-1} = \frac{1}{c} \|f\|_{\mathcal{F}}^2$$

montre que l'injection ϑ de $\tilde{\mathcal{F}}$ dans $L^2(\mathbb{R}^d)$ est continue. Notons alors ϱ l'isométrie de $\tilde{\mathcal{F}}$ dans le complété $\mathcal{F}$ et $\bar{\vartheta}$ l'unique application de $\mathcal{F}$ dans $L^2(\mathbb{R}^d)$ prolongeant (par continuité) ϑ au sens où $\vartheta = \bar{\vartheta}\varrho$. Montrer que $\mathcal{F} \subset L^2(\mathbb{R}^d)$ consiste à montrer que $\bar{\vartheta}$ est injective. Soit $f \in \mathcal{F}$ telle que $\bar{\vartheta}(f) = 0$. Par définition du complété $\mathcal{F}$, il existe une suite $(f_n)_{n\in\mathbb{N}}$ de $\tilde{\mathcal{F}}$ telle que $f = \lim_{n\to\infty} \varrho(f_n)$. La suite (f_n) est de Cauchy dans $\tilde{\mathcal{F}}$ et donc $(\vartheta(f_n))$ est une suite de Cauchy dans $L^2(\mathbb{R}^d)$ puisque ϑ est continue. Notons u la limite de $\vartheta(f_n)$ dans $L^2(\mathbb{R}^d)$. Comme $\forall n$, $\bar{\vartheta}\varrho(f_n) = \vartheta(f_n)$, $\vartheta(f_n)$ converge vers $u = 0$. Donc (f_n) converge vers 0, $f = 0$ et $\bar{\vartheta}$ est injective.

On a donc montré que $\bar{\vartheta}(\mathcal{F})$ est un complété de $\tilde{\mathcal{F}}$ qui est contenu dans $L^2(\mathbb{R}^d)$ ($\bar{\vartheta}(\mathcal{F})$ est un espace hilbertien pour le produit scalaire $(\bar{\vartheta}(f), \bar{\vartheta}(g))_{\bar{\vartheta}(\mathcal{F})} = (f, g)_{\mathcal{F}}$, isométrique à $\mathcal{F}$). En identifiant $\bar{\vartheta}(\mathcal{F})$ et $\mathcal{F}$, on peut écrire $\mathcal{F} \subset L^2(\mathbb{R}^d)$. L'espace $\mathcal{F}$ est constitué de l'ensemble de fonctions

$$\left\{ f \in L^2(\mathbb{R}^d),\ \text{t.q.}\ \int_{\mathbb{R}^d} |\tilde{f}(\boldsymbol{u})|^2 \tilde{k}(\boldsymbol{u})^{-1} d\boldsymbol{u} < +\infty \right\},$$

muni du produit scalaire

$$(f,g) = \frac{1}{(2\pi)^d} \int \tilde{f}(\boldsymbol{u})\overline{\tilde{g}(\boldsymbol{u})}\tilde{k}(\boldsymbol{u})^{-1} d\boldsymbol{u}\,.$$

On vérifie que $\forall f \in \mathcal{F}$, $\forall \boldsymbol{x} \in \mathbb{R}^d$,

$$(k(\cdot - \boldsymbol{x}), f) = \frac{1}{(2\pi)^d} \int (\tilde{k}(\boldsymbol{u}) e^{-\imath(\boldsymbol{u},\boldsymbol{x})})\overline{\tilde{f}(\boldsymbol{u})}\tilde{k}(\boldsymbol{u})^{-1} d\boldsymbol{u} = f(\boldsymbol{x})\,.$$

□

La norme de $\mathcal{F}$ écrite sous la forme (3.9) permet une interprétation intéressante en terme de propriété de régularité. Remarquons en effet que la transformée de Fourier $\tilde{k}$ intervient avec une puissance négative dans la norme. Or $\tilde{k}$ décroît lorsque les fréquences augmentent. Ceci montre que la norme de $\mathcal{F}$ pénalise les composantes hautes fréquences des fonctions de $\mathcal{F}$. Cette propriété fondamentale est à l'origine de la régression régularisée qui sera présentée dans la section 3.4.

Remarquons enfin que l'expression du noyau

$$k(\boldsymbol{x} - \boldsymbol{y}) = \frac{1}{(2\pi)^d} \int e^{\imath(\boldsymbol{x}-\boldsymbol{y},\boldsymbol{u})} \tilde{k}(\boldsymbol{u}) d\boldsymbol{u} = \frac{1}{(2\pi)^d} \int e^{\imath(\boldsymbol{x},\boldsymbol{u})} e^{-\imath(\boldsymbol{y},\boldsymbol{u})} \tilde{k}(\boldsymbol{u}) d\boldsymbol{u} \tag{3.10}$$

permet une analogie avec la représentation de Mercer (3.6).

3.4 Régression régularisée

3.4.1 Généralités sur la régression régularisée en dimension finie

Dans cette section, nous reprenons le point de vue adopté au chapitre 1. Nous souhaitons montrer comment la notion de noyau reproduisant peut être mise en place naturellement dans le cadre de la régression régularisée pour la construction de modèles boîte noire.

Considérons une fonction de plusieurs variables $f^* : \mathbb{X} \subset \mathbb{R}^d \to \mathbb{R}$, qui peut par exemple représenter la sortie d'un système dépendant d'un ensemble de facteurs caractérisant son point de fonctionnement. L'objectif est d'approximer $f^*(\boldsymbol{x})$ à partir d'un ensemble fini d'observations $\mathcal{S} = \{(\boldsymbol{x}_i, f^{\text{obs}}_{\boldsymbol{x}_i}),\ i = 1, \cdots, n\}$. Les observations peuvent être bruitées, et dans ce cas $f^{\text{obs}}_{\boldsymbol{x}_i} \neq f^*(\boldsymbol{x}_i)$.

Une approximation de la fonction $f^* : \mathbb{X} \subset \mathbb{R}^d \to \mathbb{R}$ peut correspondre à un simple modèle linéaire défini sur un domaine de l'espace des facteurs, par exemple autour d'un point de fonctionnement du système. Cette approximation linéaire peut s'écrire $f(\boldsymbol{x}) = \boldsymbol{b}^\mathsf{T}\boldsymbol{x} + b_0$, où le vecteur $\boldsymbol{b} \in \mathbb{R}^d$ et le scalaire b_0 sont des paramètres à estimer à partir des données et de l'information a priori sur le système éventuellement disponible. La méthode la plus classique pour estimer ces paramètres est sans doute la méthode des moindres carrés. Elle correspond à minimiser le critère d'erreur d'approximation quadratique

$$J(\boldsymbol{b}, b_0) = \sum_{i=1}^{n} (f^{\text{obs}}_{\boldsymbol{x}_i} - \boldsymbol{b}^\mathsf{T}\boldsymbol{x}_i - b_0)^2\,.$$

On peut généraliser très facilement cette régression linéaire en incorporant dans le modèle des termes non-linéaires en $\boldsymbol{x}$, tels que des monômes ou des fonctions exponentielles. L'approximation s'écrit alors sous la forme

$$f(\boldsymbol{x}) = \boldsymbol{b}^\mathsf{T}\boldsymbol{r}(\boldsymbol{x})\,, \tag{3.11}$$

où $\boldsymbol{b} = (b_{[1]} \cdots b_{[l]})^\mathsf{T}$ et $\boldsymbol{r}(\boldsymbol{x}) = (r_{[1]}(\boldsymbol{x}) \cdots r_{[l]}(\boldsymbol{x}))^\mathsf{T}$. Ce type d'approximation permet par exemple la prise en compte de non-linéarités locales de $f^*(\boldsymbol{x})$, ou bien la modélisation de tendances polynomiales sur le domaine d'étude $\mathbb{X}$.

Le nombre de termes paramétriques dans (3.11) croît potentiellement très rapidement avec le nombre de facteurs considérés. Par exemple, dans le cas d'approximations de type polynomial, on peut noter qu'un polynôme complet de degré m à d variables comporte $\binom{d+m}{d}$ monômes. Or l'estimation d'un grand nombre de paramètres peut conduire à un problème mal conditionné (ou mal posé) et à des instabilités numériques. (Pour détecter ces situations dans le cadre de la méthode des moindres carrés, il est préférable de calculer le conditionnement de la matrice de régression $\boldsymbol{R} = (\boldsymbol{r}(\boldsymbol{x}_1) \cdots \boldsymbol{r}(\boldsymbol{x}_n))$. Une solution classique pour un problème mal conditionné consiste à ajouter un terme de régularisation dans le critère d'erreur d'approximation. Par exemple, on peut chercher à minimiser

$$J(\boldsymbol{b}) = C\|\boldsymbol{b}\|_2^2 + \sum_{i=1}^{n} (f_{\boldsymbol{x}_i}^{\text{obs}} - \boldsymbol{b}^\mathsf{T}\boldsymbol{r}(\boldsymbol{x}_i))^2 \,. \tag{3.12}$$

Dans ce cas, on pénalise les grandes valeurs des éléments de $\boldsymbol{b}$. Le paramètre $C \in \mathbb{R}^+$ permet de régler le compromis entre régularisation et attache aux données. Ce type de régularisation est connu sous le nom de régularisation de Tikhonov et possède une théorie mathématique bien établie (Tikhonov et Arsenin, 1977). En particulier, il est possible d'établir des conditions suffisantes pour obtenir un problème bien posé, c'est-à-dire telles que les solutions existent de manière unique et soient stables. De manière informelle, on peut dire que le terme de régularisation dans (3.12) permet de restreindre l'espace de recherche des fonctions optimales pour retrouver un problème bien posé.

Toutefois, le principe de régularisation introduit dans (3.12) ne garantit pas pour autant une approximation de qualité satisfaisante. Tout d'abord, la *capacité d'approximation* du modèle (3.11) ne permet pas forcément une description pertinente des données. Cette capacité d'approximation est liée, entre autres, à la dimension de l'espace des fonctions dans lequel on recherche le minimiseur de (3.12), c'est-à-dire au nombre de termes paramétriques dans le modèle. Si la capacité d'approximation est suffisante, le modèle est capable d'interpoler les données, ce qui est utile par exemple si l'on sait que le bruit d'observation du système est négligeable. On peut évidemment chercher à augmenter cette capacité d'approximation en ajoutant autant de termes paramétriques que nécessaire. Mais ce faisant, on risque d'obtenir une approximation très irrégulière (en anglais, ce phénomène correspond à l'*overfitting*) si quelques précautions ne sont pas prises. Ce phénomène est illustré par la figure 3.1, où l'on utilise un modèle comprenant 201 termes paramétriques, obtenus en minimisant (3.12) (avec une petite valeur de C pour obtenir une quasi interpolation des données, représentées par les carrés). Cet exemple montre qu'il est nécessaire d'utiliser un schéma de régularisation adéquat. Nous nous intéressons à cette question dans la section suivante.

3.4.2 Choix d'un schéma de régularisation par utilisation d'un noyau

Soit $\widehat{\boldsymbol{b}}$ le minimiseur du critère (3.12). Nous souhaitons montrer que le modèle $\widehat{\boldsymbol{b}}^\mathsf{T}\boldsymbol{r}(\boldsymbol{x})$ peut être vu comme le résultat d'une régression à noyau reproduisant. Puisque $\widehat{\boldsymbol{b}}$ satisfait $\frac{\partial J}{\partial \boldsymbol{b}}(\widehat{\boldsymbol{b}}) = \boldsymbol{0}$,

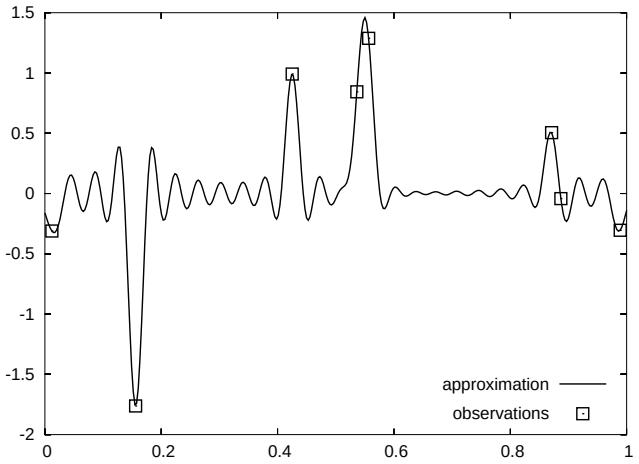


FIG. 3.1 – Exemple d'interpolation de données avec une régression mal régularisée. Le vecteur de régression est $\boldsymbol{r}(x) = [1 \ \cos(\omega_0 x) \ \sin(\omega_0 x) \quad \cdots \quad \cos(m\omega_0 x) \ \sin(m\omega_0 x)]^\mathsf{T}$, où $m = 100$

on peut écrire

$$\widehat{\boldsymbol{b}} = \frac{1}{C}\sum_{i=1}^{n}(f^{\text{obs}}_{\boldsymbol{x}_i} - \widehat{\boldsymbol{b}}^\mathsf{T}\boldsymbol{r}(\boldsymbol{x}_i))\boldsymbol{r}(\boldsymbol{x}_i)\,.$$

Par conséquent, il existe des scalaires $\widehat{a}_i$ tels que

$$\widehat{\boldsymbol{b}} = \sum_{i=1}^{n}\widehat{a}_i\boldsymbol{r}(\boldsymbol{x}_i)\,.$$

Le problème de régression initial peut alors être reformulé comme celui de la recherche d'une fonction $\widehat{f}(\boldsymbol{x}) = \sum_i \widehat{a}_i(\boldsymbol{r}(\boldsymbol{x}_i), \boldsymbol{r}(\boldsymbol{x}))$, où $(.,.)$ est le produit scalaire canonique de $\mathbb{R}^l$, et où les $\widehat{a}_i$ minimisent le critère

$$J(a_i, i = 1, \cdots, n) = \left\| \sum_{i=1}^{n} a_i\boldsymbol{r}(\boldsymbol{x}_i)\right\|^2 + \frac{1}{C}\sum_{i=1}^{n}(f^{\text{obs}}_{\boldsymbol{x}_i} - \widehat{f}(\boldsymbol{x}_i))^2\,.$$

Ce problème, exprimé sous une forme dite duale, est évidemment de type moindres carrés et admet une solution numérique qui s'obtient en résolvant un système d'équations linéaires en $\widehat{a}_i$. Puisque

$$\left\| \sum_{i=1}^{n} a_i\boldsymbol{r}(\boldsymbol{x}_i)\right\|^2 = \sum_{i,j} a_i(\boldsymbol{r}(\boldsymbol{x}_i), \boldsymbol{r}(\boldsymbol{x}_j))a_j,$$

le problème régularisé fait intervenir uniquement des produits scalaires de la forme $k(\boldsymbol{x}_i, \boldsymbol{x}_j) \triangleq (\boldsymbol{r}(\boldsymbol{x}_i), \boldsymbol{r}(\boldsymbol{x}_j))$. Cela signifie que les objets $\boldsymbol{r}(\boldsymbol{x}_i)$ n'ont pas besoin d'être évalués directement, si l'expression analytique de k est connue.

Appelons $\mathcal{F}$ l'*espace des caractéristiques* défini par $\text{vect}\{\boldsymbol{r}(\boldsymbol{x}) \in \mathbb{R}^l, \boldsymbol{x} \in \mathbb{X}\} \subseteq \mathbb{R}^l$. Tout $\boldsymbol{b} = \sum_{i=1}^{n} a_i\boldsymbol{r}(\boldsymbol{x}_i) \in \mathcal{F}$ peut être identifié à une fonction f de la forme

$$\forall \boldsymbol{x} \in \mathbb{X}, \quad f(\boldsymbol{x}) = (\boldsymbol{r}(\boldsymbol{x}), \sum_{i=1}^{n} a_i\boldsymbol{r}(\boldsymbol{x}_i))\,.$$

Ainsi, $\mathcal{F}$ peut être identifié à un espace de fonctions, ces fonctions étant principalement caractérisées par les objets $\boldsymbol{r}(\boldsymbol{x})$. (Nous utilisons par la suite la notation $f \equiv \boldsymbol{b}$ pour rappeler la propriété d'identification.) On peut ensuite définir la norme de $f \equiv \boldsymbol{b} \in \mathcal{F}$ par

$$\|f\|_{\mathcal{F}}^2 = \sum_{i,j=1}^{n} a_i(\boldsymbol{r}(\boldsymbol{x}_i), \boldsymbol{r}(\boldsymbol{x}_j))a_j \,.$$

Dans $\mathcal{F}$, $\boldsymbol{r}(\boldsymbol{x})$ est identifiable à l'opérateur d'évaluation ponctuelle. En effet, toute fonction $f \equiv \boldsymbol{b}$ de $\mathcal{F}$ peut être évaluée en un point $\boldsymbol{x}$ en formant le produit scalaire $f(\boldsymbol{x}) = (\boldsymbol{r}(\boldsymbol{x}), \boldsymbol{b})$ (voir (3.11)). L'opérateur $\boldsymbol{r(x)}$ d'évaluation en $\boldsymbol{x}$ est continu puisque $|f(\boldsymbol{x})| \leq \|\boldsymbol{r}(\boldsymbol{x})\| \, \|f\|_{\mathcal{F}}$. Par conséquent, $\mathcal{F}$ est un espace hilbertien à noyau reproduisant ; son noyau reproduisant est $k(\boldsymbol{x}, \boldsymbol{y})$, puisque l'on peut identifier la fonction $k(\boldsymbol{x}, .) \in \mathcal{F}$ à l'opérateur d'évaluation $\boldsymbol{r}(\boldsymbol{x})$ ($k(\boldsymbol{x}, .) \equiv \boldsymbol{r}(\boldsymbol{x})$ en utilisant le fait que $k(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{r}(\boldsymbol{y}), \boldsymbol{r}(\boldsymbol{x}))$, $\forall\, \boldsymbol{y}$).

Si l'on considère une fonction $f \equiv \boldsymbol{b} \in \mathcal{F}$, on a $\|f\|_{\mathcal{F}}^2 = \sum_{i=1}^{l} b_{[i]}^2$, et donc l'espace des caractéristiques $\mathcal{F}$ est muni d'une norme qui pénalise de la même manière tous les termes paramétriques $r_{[i]}(\boldsymbol{x})$ de $\boldsymbol{r}(\boldsymbol{x})$. Comme on l'a vu plus haut ceci n'est pas forcément désirable. Pour éviter les comportements irréguliers de f, il faut pouvoir pénaliser davantage les termes à l'origine des variations importantes. Typiquement, ces termes correspondent aux hautes fréquences. Une expérience numérique simple, présentée dans la figure 3.2, confirme que la notion de choix du schéma de régularisation est essentielle.

La régularisation peut être choisie a priori, par exemple, comme cela a été fait dans le cas de la figure 3.2, en multipliant $\boldsymbol{r(x)}$ élément par élément par des coefficients choisis arbitrairement mais de telle façon que ces coefficients décroissent lorsque la fréquence augmente. Cette approche n'est toutefois pas très élégante et difficile à généraliser dans le cas d'espace de facteurs de dimension supérieure à un[3].

L'approche probablement la plus simple consiste à choisir le schéma de régularisation en utilisant la norme d'un espace hilbertien à noyau reproduisant offrant de bonnes propriétés de régularité (sans nécessairement se restreindre à des dimensions finies comme ci-dessus). Cette idée permet d'effectuer des régressions régularisées dans des espaces de fonctions de dimension infinie (grandes capacités d'approximation possibles) en veillant à ce que les problèmes d'approximation restent bien posés. En effet, nous avons vu dans la section 3.3 que les noyaux reproduisants, ainsi que les éléments d'un espace à noyau, admettent des représentations spectrales sous certaines conditions. Typiquement (par exemple comme dans (3.9)), la norme d'un espace à noyau pénalise les hautes fréquences. Par suite, il est intéressant d'utiliser cette norme comme terme de régularisation. La régression régularisée dans des espaces à noyau est détaillée dans la section suivante.

Enfin, notons que le schéma de régularisation (même à noyau reproduisant) doit résulter en pratique d'une procédure d'adaptation aux données. En effet, le fait d'utiliser un noyau reproduisant arbitraire parmi les familles connues de fonctions symétriques positives ne garantit pas la qualité de l'approximation. Ce point sera abordé au chapitre 5.

[3]Nous présentons cependant dans les sections 5.6 et 6.5 des approches conceptuellement très proches, où le schéma de régularisation est adapté automatiquement aux données observées en optimisant un critère de qualité d'approximation.

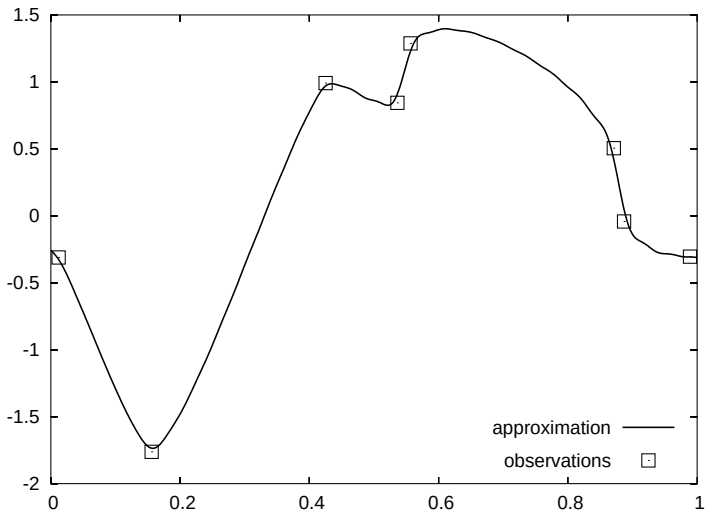


FIG. 3.2 – Les données et la structure du modèle sont les mêmes que dans la figure 3.1, mais le schéma de régularisation a été adapté, en utilisant le vecteur de régression $\boldsymbol{r}(x) = [1 \ \frac{\cos(\omega_0 x)}{1+\omega_0^{1.3}} \ \frac{\sin(\omega_0 x)}{1+\omega_0^{1.3}} \ \cdots \ \frac{\cos(m\omega_0 x)}{1+(m\omega_0)^{1.3}} \ \frac{\sin(m\omega_0 x)}{1+(m\omega_0)^{1.3}}]^\mathsf{T}$

3.4.3 Problème régularisé et forme des solutions

La forme générale d'un problème de régression dans un espace à noyau, régularisée par une approche de Tikhonov (Tikhonov et Arsenin, 1977) est maintenant rappelée.

Considérons un système statique à plusieurs entrées et une seule sortie, c'est-à-dire une fonction $f^* : \mathbb{X} \to \mathbb{R}$, ainsi qu'un ensemble fini d'observations $S = \{(\boldsymbol{x}_i, f^{\text{obs}}_{\boldsymbol{x}_i}) \in \mathbb{X} \times \mathbb{R},\ i = 1, \ldots, n\}$. Si le système est observé sans bruit, on a $f^*(\boldsymbol{x}_i) = f^{\text{obs}}_{\boldsymbol{x}_i}$, $\forall i$. L'objectif est de chercher une fonction $\hat{f}$ dans un ensemble de fonctions $\mathcal{F}$ minimisant un critère quantifiant l'erreur d'approximation de f^*, noté $J_{f^*}(f)$.

Soit une fonction $l(\boldsymbol{x}, s, t) \in \mathbb{R}^+ \cup +\infty$, $\boldsymbol{x} \in \mathbb{R}^d$, $s, t \in \mathbb{R}$ satisfaisant $l(\boldsymbol{x}, s, s) = 0$, $\forall x \in \mathbb{X}$, $\forall s \in \mathbb{R}$. La fonction $\boldsymbol{x} \mapsto l(\boldsymbol{x}, f(\boldsymbol{x}), f^*(\boldsymbol{x}))$ sert à assigner un coût lorsque $f(\boldsymbol{x})$ s'écarte de la vraie valeur $f^*(\boldsymbol{x})$. l est donc appelée une *fonction de coût* (en anglais, *loss function*). En général, l ne dépend pas de son premier paramètre $\boldsymbol{x}$ et la notation simplifiée $l(s, t)$ sera utilisée par la suite. Par exemple, $l(s, t) = (s - t)^2$ est la fonction de coût dite quadratique (ou L^2) et $l(t, s) = |s - t|$ est la fonction de coût dite L^1.

Le critère d'erreur d'approximation que l'on souhaite minimiser, formé à partir des données et de la fonction de coût, peut s'écrire

$$J_{f^*}(f) = \sum_{i=1}^{n} l(f(\boldsymbol{x}_i), f^{\text{obs}}_{\boldsymbol{x}_i}) \,. \tag{3.13}$$

Selon l'espace $\mathcal{F}$ choisi, les solutions optimales $\hat{f} = \arg\min_{f \in \mathcal{F}} J_{f^*}(f)$ possèdent des comportements très différents. Si $\mathcal{F}$ est suffisamment large, il est même possible de trouver une infinité de fonctions qui annulent exactement le critère. En général, l'unicité de la solution n'est donc que rarement vérifiée et la minimisation de J_{f^*} est un problème mal posé. Pour garantir l'unicité de

la solution ainsi que sa stabilité, il est nécessaire de s'assurer que $\mathcal{F}$ est suffisamment petit ou de modifier le critère d'erreur d'approximation. L'utilisation de la norme de $\mathcal{F}$ comme terme de régularisation obéit en quelque sorte à ces deux principes (on modifie le critère pour s'assurer que $f \in \mathcal{F}$ ne « s'éloigne pas trop » de la fonction nulle).

Le théorème du représentant (prouvé dans (Kimeldorf et Wahba, 1970a) dans le cas quadratique, et dans (Schölkopf et al., 2001) dans le cas général) permet d'établir la forme des solutions.

Théorème 15 (dit du représentant). *Soit $\mathcal{F}$ un espace hilbertien à noyau reproduisant et soit le critère d'erreur d'approximation régularisée de Tikhonov, défini sur $\mathcal{F}$ par*

$$J_{f^*,\mathrm{reg}}(f) = \sum_{i=1}^{n} l(f(\boldsymbol{x}_i), f^{\mathrm{obs}}_{\boldsymbol{x}_i}) + C\|f\|^2_{\mathcal{F}}\,, \tag{3.14}$$

où $C > 0$. Alors, si $l(s,t)$ est convexe en s, le minimiseur de $J_{f^*,reg}$ dans $\mathcal{F}$ est unique et peut s'écrire sous la forme

$$\hat{f}(\boldsymbol{x}) = \sum_{i=1}^{n} a_i k(\boldsymbol{x}, \boldsymbol{x}_i)\,.$$

où $\boldsymbol{a} = (a_1, \ldots, a_n)^{\mathsf{T}}$ est la solution d'un problème bien posé.

Démonstration. Voir (Schölkopf et al., 2001). □

Proposition 29 (Cas d'une fonction de coût quadratique). *Si $l(s,t) = (s-t)^2$, le minimiseur de $J_{f^*,\mathrm{reg}}(f)$ s'obtient en résolvant le système d'équations linéaires*

$$(C\boldsymbol{I}_n + \boldsymbol{K})\boldsymbol{a} = \boldsymbol{f}^{\mathrm{obs}}\,, \tag{3.15}$$

où $\boldsymbol{K}$ est la matrice $n \times n$ d'éléments $k_{i,j} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ et où $\boldsymbol{f}^{\mathrm{obs}} = (f^{\mathrm{obs}}_{\boldsymbol{x}_1}, \ldots, f^{\mathrm{obs}}_{\boldsymbol{x}_n})^{\mathsf{T}}$.

On constate que (3.15) est le système obtenu dans la méthode du krigeage dual avec bruit d'observation. On peut donc interpréter C comme la variance du bruit d'observation. Il est maintenant important de s'interroger sur les liens précis qui existent entre la régression régularisée par une norme d'espace à noyau reproduisant et le krigeage (ou la prédiction linéaire). Les sections suivantes montrent que les méthodes sont équivalentes, ce qui est en fait peu surprenant.

3.5 Espaces hilbertiens à noyau reproduisant et processus aléatoires

Les constructions des espaces $\mathcal{H}$ et $\mathcal{F}$ effectuées dans les sections 2.4.3 et 3.1 sont très similaires. Nous souhaitons présenter plus précisément les liens entre ces deux espaces, ainsi que le rôle joué par l'espace Λ obtenu comme le complété de l'espace vectoriel des mesures à support fini pour la topologie induite par le noyau. Les premiers travaux qui étudient précisément le rôle des espaces à noyau reproduisant dans le domaine des processus aléatoires sont à notre connaissance (Parzen, 1962, 1963, 1970), ainsi que (Hajek, 1962) qui ne recourt toutefois pas explicitement à la notion de noyau reproduisant. En se fondant sur ces premières études[4], Kimeldorf et Wahba (1970b) ont établi un résultat important concernant l'équivalence entre la régression régularisée par une norme d'espace à noyau reproduisant et la prédiction linéaire, que nous présenterons dans la section 3.5.5.

[4] Anecdote : E. Parzen était directeur de la thèse de G. Wahba.

3.5.1 Trois espaces ayant la même structure

Soit $F(\boldsymbol{x})$, $\boldsymbol{x} \in \mathbb{X}$, un processus aléatoire du second ordre, de moyenne nulle et de covariance $k(\boldsymbol{x}, \boldsymbol{y})$. Considérons $\tilde{\mathcal{H}}$ l'espace engendré par les combinaisons linéaires de $F(\boldsymbol{x})$ et soit $\mathcal{H}$ un complété de $\tilde{\mathcal{H}}$ pour la norme de $L^2(\Omega, \mathcal{A}, \mathsf{P})$. Rappelons que $\tilde{\Lambda}$ désigne l'espace des mesures à support fini sur $\mathbb{X}$. Ainsi, si $\lambda \in \tilde{\Lambda}$, cette mesure peut s'écrire $\lambda = \sum_{i=1}^{n} \lambda_i \delta_{\boldsymbol{x}_i}$. On munit $\tilde{\Lambda}$ du produit scalaire

$$(\lambda, \mu)_{\tilde{\Lambda}} = \sum_{i,j=1}^{n,m} \lambda_i \mu_j k(\boldsymbol{x}_i, \boldsymbol{y}_j)\,,$$

et on désigne par Λ un complété de $\tilde{\Lambda}$ pour ce produit scalaire. L'application linéaire $F : \Lambda \to \mathcal{H}$, telle que $F(\lambda) = \sum_{i=1}^{n} \lambda_i F(\boldsymbol{x}_i)$ pour $\lambda \in \tilde{\Lambda}$, et qui se prolonge par continuité sur Λ, définit une isométrie bijective entre Λ et $\mathcal{H}$. Nous avons également vu plus haut que Λ est le dual d'un espace hilbertien de fonctions $\mathcal{F}$ admettant le noyau reproduisant $k(\boldsymbol{x}, \boldsymbol{y})$. Il existe donc une isométrie bijective entre Λ et $\mathcal{F}$.

Par conséquent, Λ, $\mathcal{H}$ et $\mathcal{F}$ ont la même structure et sont isomorphes. Les topologies sur ces espaces sont toutes liées au noyau $k(\boldsymbol{x}, \boldsymbol{y})$. Dans $\mathcal{H}$, $k(\boldsymbol{x}, \boldsymbol{y})$ caractérise la corrélation entre deux variables aléatoires $F(\boldsymbol{x})$ et $F(\boldsymbol{y})$ qui augmente lorsque les facteurs $\boldsymbol{x}$ et $\boldsymbol{y}$ sont proches. Les propriétés des trajectoires de $F(\boldsymbol{x})$ (notamment l'échelle caractéristique des variations de ces trajectoires, la régularité, etc.) dépendent de la forme de $k(\boldsymbol{x}, \boldsymbol{y})$. Dans $\mathcal{F}$, toute fonction peut être reconstruite ou approximée par des combinaisons linéaires de $k(\boldsymbol{x}, \boldsymbol{y})$. Donc les fonctions de $\mathcal{F}$ sont également caractérisées par la forme de $k(\boldsymbol{x}, \boldsymbol{y})$.

3.5.2 Espace hilbertien engendré par une fonction de covariance

Nous avons vu précédemment que toute fonction de covariance $k(\boldsymbol{x}, \boldsymbol{y})$ d'un processus aléatoire $F(\boldsymbol{x})$ est admissible comme noyau reproduisant. Les paragraphes suivants explicitent l'isomorphisme entre l'espace $\mathcal{H}$ et l'espace hilbertien $\mathcal{F}$ à noyau reproduisant k.

Soit $\mathcal{H}$ l'espace hilbertien engendré par un processus aléatoire $F(\boldsymbol{x})$ du second ordre, de moyenne nulle et de covariance $k(\boldsymbol{x}, \boldsymbol{y})$. Considérons l'application linéaire $\varrho : \mathcal{H} \to \mathbb{R}^{\mathbb{X}}$ définie par

$$(\varrho X)(\boldsymbol{x}) = (F(\boldsymbol{x}), X)_{\mathcal{H}}\,, \quad \forall \boldsymbol{x} \in \mathbb{X}\,.$$

Soit $\mathcal{F} = \operatorname{Im} \rho$, l'image de $\mathcal{H}$ par cette application, c'est-à-dire l'espace vectoriel des fonctions f telles que $f = \varrho X$, lorsque X parcourt $\mathcal{H}$. Si $f = \varrho X \in \mathcal{F}$, pour un certain $X \in \mathcal{H}$, on a donc la relation fondamentale (Parzen, 1962, 1963, 1970)

$$f(\boldsymbol{x}) = (F(\boldsymbol{x}), X)_{\mathcal{H}}\,, \quad \forall \boldsymbol{x} \in \mathbb{X}.$$

Supposons de plus $k(\boldsymbol{x}, \boldsymbol{y})$ définie positive. Alors $\varrho X_1 = \varrho X_2$, $X_1, X_2 \in \mathcal{H}$, implique $X_1 = X_2$, si bien que $\varrho : \mathcal{H} \to \mathcal{F}$ est inversible. Le produit scalaire sur $\mathcal{F}$ défini par

$$(f, g)_{\mathcal{F}} = (\varrho^{-1} f, \varrho^{-1} g)_{\mathcal{H}}\,,$$

dote $\mathcal{F}$ d'une structure d'espace hilbertien et nous constatons que ϱ est une isométrie bijective entre $\mathcal{H}$ et $\mathcal{F}$ (parfois appelée isométrie de Loève (Lukic et Beder, 2001)).

L'espace $\mathcal{F}$ ainsi construit est-il un espace à noyau reproduisant ? On montre facilement que $k(\boldsymbol{x},.) \in \mathcal{F}$, que $\varrho^{-1}k(\boldsymbol{x},.) = F(\boldsymbol{x})$ et

$$(k(\boldsymbol{x},.),k(\boldsymbol{y},.))_{\mathcal{F}} = (F(\boldsymbol{x}),F(\boldsymbol{y}))_{\mathcal{H}} = k(\boldsymbol{x},\boldsymbol{y})\,, \quad \forall \boldsymbol{x},\boldsymbol{y} \in \mathbb{X}\,.$$

De même, si $f = \varrho X \in \mathcal{F}$, pour un $X \in \mathcal{H}$, on a

$$f(\boldsymbol{x}) = (F(\boldsymbol{x}),X)_{\mathcal{H}} = (k(\boldsymbol{x},.),f)_{\mathcal{F}}\,, \quad \forall \boldsymbol{x} \in \mathbb{X}.$$

Par conséquent, $\mathcal{F}$ admet le noyau reproduisant $k(\boldsymbol{x},\boldsymbol{y})$. Notons que $\mathcal{F}$ est engendré par les fonctions $k(\boldsymbol{x},.)$, lorsque $\boldsymbol{x}$ parcourt $\mathbb{X}$; les fonctions $k(\boldsymbol{x},.)$ sont identifiables aux éléments $F(\boldsymbol{x})$, qui engendrent $\mathcal{H}$. Par la suite, nous dirons simplement que $\mathcal{F}$ est engendré par la covariance $k(\boldsymbol{x},\boldsymbol{y})$ et il sera sous-entendu que $\mathcal{F}$ est muni d'une structure d'espace hilbertien avec le produit scalaire dérivant de $k(\boldsymbol{x},\boldsymbol{y})$.

Dans la section 2.4.3, nous avons rappelé qu'un filtrage linéaire de $F(\boldsymbol{x})$ génère un processus aléatoire appartenant à $\mathcal{H}$. Nous verrons dans les paragraphes suivants que l'espace des fonctions engendrées par la covariance du processus filtré est un sous-espace de $\mathcal{F}$. Considérons un opérateur linéaire continu $T : \mathcal{H} \to \mathcal{H}$ et notons $\mathcal{F}_1$ l'espace vectoriel engendré par la covariance $k_1(\boldsymbol{x},\boldsymbol{y})$ du processus $F_1(\boldsymbol{x}) = TF(\boldsymbol{x})$. L'espace $\mathcal{F}_1$ est donc constitué de l'ensemble des fonctions s'exprimant sous la forme

$$f_1(\boldsymbol{x}) = (F_1(\boldsymbol{x}),X)_{\mathcal{H}_1} = (F_1(\boldsymbol{x}),X)_{\mathcal{H}} = (\varrho_1 X)(\boldsymbol{x})\,, \quad X \in \mathcal{H}_1\,,$$

où $\mathcal{H}_1$ est l'espace hilbertien de variables aléatoires généré par $F_1(\boldsymbol{x})$, $\boldsymbol{x} \in \mathbb{X}$, et $\varrho_1 : \mathcal{H}_1 \to \mathcal{F}_1$ est une application bijective. Muni du produit scalaire $(f_1,g_1)_{\mathcal{F}_1} = (\varrho_1^{-1}f_1, \varrho_1^{-1}g_1)_{\mathcal{H}_1}$, $\mathcal{F}_1$ est un espace hilbertien à noyau reproduisant $k_1(\boldsymbol{x},\boldsymbol{y})$.

À tout opérateur linéaire continu $T : \mathcal{H} \to \mathcal{H}$ nous faisons correspondre un opérateur linéaire $\mathcal{F} \to \mathcal{F}$, noté (abusivement) T, tel que pour $f \in \mathcal{F}$,

$$(Tf)(\boldsymbol{x}) = (TF(\boldsymbol{x}), \varrho^{-1}f)_{\mathcal{H}}\,, \quad \forall \boldsymbol{x} \in \mathbb{X}\,.$$

On a donc, en notant T^* l'opérateur adjoint de T, $Tf = \varrho T^* \varrho^{-1} f$, ce qui montre que $T : \mathcal{F} \to \mathcal{F}$ est continu.

Théorème 16. *On a $\mathcal{F}_1 \subseteq \mathcal{F}$ et plus précisément, les relations suivantes :*

$$\mathcal{F}_1 = \{TT^*f\,,\ f \in \mathcal{F}\}\,, \tag{3.16}$$

$$k_1(\boldsymbol{x},.) = TT^*k(\boldsymbol{x},.)\,, \quad \forall \boldsymbol{x} \in \mathbb{X}\,, \tag{3.17}$$

$$(TT^*f, TT^*g)_{\mathcal{F}_1} = (TT^*f, g)_{\mathcal{F}}\,, \quad \forall f,g \in \mathcal{F}\,. \tag{3.18}$$

Démonstration. Soit $f_1 \in \mathcal{F}_1$, alors $\exists X \in \mathcal{H}$ tel que $\forall \boldsymbol{x} \in \mathbb{X}$,

$$f_1(\boldsymbol{x}) = (TF(\boldsymbol{x}),TX)_{\mathcal{H}} = (T^*TF(\boldsymbol{x}),X)_{\mathcal{H}} = (TT^*f)(\boldsymbol{x})\,,$$

avec $f \in \mathcal{F}$ (en utilisant la relation $ABf = \varrho(BA)^*\varrho^{-1}f$). Réciproquement, si $f = \varrho X \in \mathcal{F}$, alors

$$(TT^*f)(\boldsymbol{x}) = (T^*TF(\boldsymbol{x}),X)_{\mathcal{H}} = (TF(\boldsymbol{x}),TX) \in \mathcal{F}_1\,.$$

D'autre part $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{X}$,

$$k_1(\boldsymbol{x}, \boldsymbol{y}) = (TF(\boldsymbol{y}), TF(\boldsymbol{x}))_{\mathcal{H}} = (T^*TF(\boldsymbol{y}), F(\boldsymbol{x})) = (TT^*k(\boldsymbol{x}, .))(\boldsymbol{y}) \,,$$

ce qui établit la relation (3.17) entre les noyaux reproduisants k et k_1.

Enfin, $\forall f, g \in \mathcal{F}$, $(TT^*f, TT^*g)_{\mathcal{F}_1}$ est une forme bilinéaire continue dans $\mathcal{F}$. En particulier, à f fixé, il existe $\tilde{f} \in \mathcal{F}$ telle que $\forall g \in \mathcal{F}$,

$$(TT^*f, TT^*g)_{\mathcal{F}_1} = (\tilde{f}, g)_{\mathcal{F}} \,.$$

En prenant $g = k(\boldsymbol{x}, .)$, on a $\forall \boldsymbol{x} \in \mathbb{X}$

$$\tilde{f}(\boldsymbol{x}) = (\tilde{f}, k(\boldsymbol{x}, .))_{\mathcal{F}} = (TT^*f, TT^*k(\boldsymbol{x}, .))_{\mathcal{F}_1} = (TT^*f, k_1(\boldsymbol{x}, .))_{\mathcal{F}_1} = (TT^*f)(\boldsymbol{x}) \,,$$

ce qui démontre (3.18). □

3.5.3 Opérateur de domination

Dans les paragraphes suivants, nous rappelons la notion de domination entre normes et les relations connues dans le cas des espaces hilbertiens à noyau reproduisant (Aronszajn, 1950).

Définition 28. Soient $\|.\|_0$ et $\|.\|_1$ deux normes définies sur un espace vectoriel $\tilde{\mathcal{F}}$. On dit que la norme $\|.\|_0$ *domine* la norme $\|.\|_1$ lorsque

$$\|f\|_0 \leq \|f\|_1 \quad \forall f \in \tilde{\mathcal{F}}.$$

Par suite, le complété $\mathcal{F}_1$ de $\tilde{\mathcal{F}}$ pour la norme $\|.\|_1$ est un sous-espace vectoriel du complété $\mathcal{F}_0$ de $\mathcal{F}$ pour la norme $\|.\|_0$ ($\mathcal{F}_1 \subseteq \mathcal{F}_0$).

Soient k_0 et k_1 deux noyaux reproduisants, et $\mathcal{F}_0$ et $\mathcal{F}_1$ les espaces hilbertiens à noyau reproduisant correspondant.

Définition 29. On dit que le noyau k_0 *domine* le noyau k_1, ce que l'on écrit $k_0 \geq k_1$, si $\mathcal{F}_1 \subseteq \mathcal{F}_0$. ($\mathcal{F}_1$ est un sous-espace vectoriel de $\mathcal{F}_0$ mais possède un produit scalaire différent de celui de $\mathcal{F}_0$. $\mathcal{F}_1$ n'est donc pas un sous-espace hilbertien de $\mathcal{F}_0$.)

Théorème 17. *Soient k_0 et k_1 tels que $k_0 \geq k_1$. Alors*

$$\|f\|_{\mathcal{F}_0} \leq \|f\|_{\mathcal{F}_1}, \quad \forall f \in \mathcal{F}_1 \,.$$

De plus, il existe un opérateur unique $L : \mathcal{F}_0 \to \mathcal{F}_0$ dont l'image est contenue dans $\mathcal{F}_1$, et tel que

$$(f, g)_{\mathcal{F}_0} = (Lf, g)_{\mathcal{F}_1} \,, \quad \forall f \in \mathcal{F}_0 \,, \forall g \in \mathcal{F}_1 \,.$$

En particulier,

$$Lk_0(\boldsymbol{x}, .) = k_1(\boldsymbol{x}, .) \,, \quad \forall \boldsymbol{x} \in \mathbb{X} \,.$$

L'opérateur L est continu, positif et auto-adjoint.

Réciproquement, soit $L : \mathcal{F}_0 \to \mathcal{F}_0$ un opérateur continu, positif et auto-adjoint. Alors

$$k_1(\boldsymbol{x}, \boldsymbol{y}) = (Lk_0(\boldsymbol{x}, .), k_0(\boldsymbol{y}, .))_{\mathcal{F}_0}, \quad \boldsymbol{x}, \boldsymbol{y} \in \mathbb{X} \,,$$

définit un noyau reproduisant tel que $k_1 \leq k_0$.

Démonstration. Ce théorème rassemble un certain nombre de résultats prouvés dans (Aronszajn, 1950). □

L'opérateur L du théorème 17 est appelé *opérateur de domination* de k_1 par k_0. On dit que la domination est *nucléaire* si L est un *opérateur nucléaire* (voir par exemple Ibragimov et Rozanov, 1978 ; Yosida, 1980 ; Lukic et Beder, 2001). Nous notons alors $k_0 \gg k_1$ la relation de domination nucléaire.

3.5.4 Processus aléatoires à trajectoires dans un espace à noyau reproduisant

Cette section énonce les conditions pour que les trajectoires d'un processus aléatoire de covariance k_1 appartiennent à un espace hilbertien à noyau k_0 (Hajek, 1962 ; Driscoll, 1973 ; Lukic et Beder, 2001). Rappelons que nous avons vu dans la section 2.3 la notion de variable aléatoire dans un espace de Hilbert $\mathcal{F}$. Nous présentons maintenant plus précisément les propriétés de ces objets lorsque $\mathcal{F}$ est un espace hilbertien à noyau reproduisant.

Proposition 30. *Soit $\mathcal{F}$ un espace hilbertien à noyau reproduisant k, $(\Omega, \mathcal{A}, \mathsf{P})$ un espace de probabilités et $F(\omega, \boldsymbol{x})$ un processus aléatoire tel que ses trajectoires sont dans $\mathcal{F}$ avec probabilité un. Alors*

$$F(\omega) = F(\omega, .)\,, \quad \omega \in \Omega\,, \tag{3.19}$$

définit une variable aléatoire à valeurs dans $\mathcal{F}$ et l'on a

$$F(\boldsymbol{x}) = (F, k(\boldsymbol{x}, .))_{\mathcal{F}} \quad \forall \boldsymbol{x} \in \mathbb{X}. \tag{3.20}$$

Réciproquement, soit F une variable aléatoire définie sur $(\Omega, \mathcal{A}, \mathsf{P})$ à valeurs dans $\mathcal{F}$, alors (3.20) définit un processus aléatoire sur $(\Omega, \mathcal{A}, \mathsf{P})$ et (3.19) est vérifiée.

Démonstration. Voir (Lukic et Beder, 2001). □

Par la suite, $\mathcal{F}_0$ désigne un espace hilbertien à noyau reproduisant k_0 et $F(\boldsymbol{x})$ un processus aléatoire, du second ordre, de covariance k_1. L'espace hilbertien engendré par le noyau k_1 est noté $\mathcal{F}_1$. Lorsque les trajectoires de $F(\boldsymbol{x})$ sont dans $\mathcal{F}_0$ avec probabilité un, il existe d'après la proposition 30 une variable aléatoire F à valeurs dans $\mathcal{F}_0$, telle que $F(\boldsymbol{x}) = (F, k(\boldsymbol{x}, .))_{\mathcal{F}_0}$, $\forall x \in \mathbb{X}$. Nous avons vu dans la section 2.3 que si $F(\boldsymbol{x})$ n'est pas du second ordre alors F ne peut pas être du second ordre. Supposons que F soit du second ordre et notons K l'opérateur de covariance associé (voir la section 2.3), tel que

$$k_1(\boldsymbol{x}, \boldsymbol{y}) = (K k_0(\boldsymbol{x}, .), k_0(\boldsymbol{y}, .))_{\mathcal{F}_0}\,.$$

On a alors $k_0 \geq k_1$ et $\mathcal{F}_1 \subseteq \mathcal{F}_0$. Notons que K est un opérateur de domination.

Le résultat suivant permet d'établir la réciproque.

Théorème 18.

- *$k_0 \geq k_1$ si et seulement si F est du second ordre.*
- *$k_0 \gg k_1$ si et seulement si F est du second ordre et il existe une modification de F à valeurs séparables.*

Démonstration. Voir (Lukic et Beder, 2001). □

Nous examinons dans les paragraphes suivants les conditions sous lesquelles un processus aléatoire de covariance k_1 possède des trajectoires dans un espace à noyau reproduisant k_0. Ces conditions ont été explicitées dans (Lukic et Beder, 2001). On trouve toutefois dans (Hajek, 1962 ; Driscoll, 1973) les prémices de la plupart de ces résultats. Nous pouvons résumer le résultat essentiel sous la forme

$$\mathsf{P}\{F(\omega,.) \in \mathcal{F}_0\} = 0 \text{ ou } 1\,,$$

selon la relation de domination entre k_0 et k_1. Nous rappelons plus précisément ci-dessous les principales conclusions de Lukic et Beder (2001) (sans démonstration).

Le premier résultat montre que la relation de domination nucléaire garantit l'appartenance des trajectoires d'un processus aléatoire à un espace à noyau reproduisant.

Théorème 19. *Soit $F(\boldsymbol{x})$, $\boldsymbol{x} \in \mathbb{X}$, un processus aléatoire du second ordre, de covariance k_1, et soit k_0 le noyau reproduisant d'un espace hilbertien $\mathcal{F}_0$ tel que $k_0 \gg k_1$. Si de plus la moyenne $m(\boldsymbol{x}) = \mathrm{E}[F(\boldsymbol{x})]$ appartient à $\mathcal{F}_0$, alors il existe une modification de F dont les trajectoires appartiennent à $\mathcal{F}_0$ avec probabilité un.*

Dans le cas général, la condition de domination nucléaire n'est pas nécessaire.

Théorème 20. *Soient k_0 et k_1 les noyaux reproduisants des espaces hilbertiens $\mathcal{F}_0$ et $\mathcal{F}_1$ tels que $\mathcal{F}_1$ soit un sous-ensemble séparable de $\mathcal{F}_0$. Alors il existe un processus aléatoire $F(\boldsymbol{x})$, $\boldsymbol{x} \in \mathbb{X}$, du second ordre et de covariance k_1 avec ses trajectoires dans $\mathcal{F}_0$.*

Dans le cas des processus gaussiens, la condition de domination nucléaire est cependant nécessaire (ce résultat est connu depuis assez longtemps, voir par exemple (Hajek, 1962 ; Ibragimov et Rozanov, 1978)). On a en effet le théorème suivant.

Théorème 21. *Soit $F(\boldsymbol{x})$, $\boldsymbol{x} \in \mathbb{X}$, un processus aléatoire gaussien de moyenne $m(\boldsymbol{x})$ et de covariance $k_1(\boldsymbol{x}, \boldsymbol{y})$. Soit $\mathcal{F}_0$ un espace hilbertien à noyau reproduisant k_0 tel que $m(.) \in \mathcal{F}_0$. Si les trajectoires de $F(\boldsymbol{x})$ appartiennent à $\mathcal{F}_0$ avec probabilité un, alors la variable aléatoire F définie par le processus $F(\boldsymbol{x})$ est gaussienne et $k_0 \gg k_1$.*

3.5.5 Équivalence entre régression régularisée et prédiction linéaire

Nous commencerons par envisager le cas des observations non bruitées, avant de voir les conséquences de la prise en compte d'un bruit de mesure.

Observations non bruitées

Si la capacité d'approximation du modèle est suffisante, on peut toujours imposer les conditions

$$\hat{f}(\boldsymbol{x}_i) = f^{\mathrm{obs}}_{\boldsymbol{x}_i}\,, \quad \forall i = 1, \ldots, n\,.$$

(Ceci est pertinent si les observations ne sont pas entachées par des erreurs de mesure.) Considérons alors le problème de régression sous contraintes, régularisé par une norme d'espace à noyau

reproduisant k :

$$\text{minimiser} \qquad \|f\|_{\mathcal{F}}^2\,, \quad f \in \mathcal{F} \tag{3.21}$$

$$\text{sous les contraintes} \qquad f(\boldsymbol{x}_i) = f_{\boldsymbol{x}_i}^{\text{obs}}\,, \quad \forall i = 1, \ldots, n\,. \tag{3.22}$$

Proposition 31. *Soient $\hat{f}$ la solution du problème (3.21)–(3.22) et $F(\boldsymbol{x})$ un processus aléatoire du second ordre, de moyenne nulle et de covariance k. Alors $\forall \boldsymbol{x} \in \mathbb{X}$, $\hat{f}(\boldsymbol{x}) = \hat{F}(\omega_0, \boldsymbol{x})$, où $\hat{F}(\boldsymbol{x})$ est le prédicteur par krigeage de $F(\boldsymbol{x})$ à partir des variables $F(\boldsymbol{x}_i)$, $i = 1, \ldots n$, et ω_0 tel que $F(\omega_0, \boldsymbol{x}_i) = f_{\boldsymbol{x}_i}^{\text{obs}}$, pour tout $i = 1, \ldots, n$.*

La preuve présentée ci-dessous nous semble sensiblement plus simple que les démonstrations très formelles de Kimeldorf et Wahba (1970a) ou Matheron (1981)[5].

Démonstration. Soit $f^* \in \mathcal{F}$ ($\mathcal{F}$ est l'espace de recherche des solutions) telle que $f^*(\boldsymbol{x}_i) = f_{\boldsymbol{x}_i}^{\text{obs}}$, $i = 1, \ldots, n$. Si $\mathcal{F}$ est un espace hilbertien à noyau reproduisant, (3.22) peut être exprimée par la condition d'orthogonalité

$$(k(\boldsymbol{x}_i, .), f - f^*)_{\mathcal{F}} = 0\,, \tag{3.23}$$

Si $\mathcal{F}_S$ est le sous-espace vectoriel de $\mathcal{F}$ défini par $\text{vect}\{k(\boldsymbol{x}_i, .), i = 1, \ldots, n\}$, (3.23) exprime la condition

$$f - f^* \perp \mathcal{F}_S\,. \tag{3.24}$$

Comme on souhaite minimiser $\|f\|_{\mathcal{F}}$, on a nécessairement $f \in \mathcal{F}_S$, et donc

$$f = \sum_{i=1}^{n} a_i k(\boldsymbol{x}_i, .)\,. \tag{3.25}$$

On reconnaît le résultat du théorème de représentation, qui vient donc d'être démontré dans le cas de l'interpolation. D'un point de vue géométrique, il est équivalent de chercher à minimiser l'erreur $\|f - f^*\|$ sous la contrainte $f \in \mathcal{F}_S$.

Puisque la fonction $k(\boldsymbol{x}_i, .)$ s'identifie à la variable aléatoire $F(\omega, \boldsymbol{x}_i)$ (d'après la section 3.5.1), l'espace $\mathcal{F}_S$ s'identifie au sous-espace vectoriel

$$\mathcal{H}_S = \text{vect}\{F(\omega, \boldsymbol{x}_1), \ldots, F(\omega, \boldsymbol{x}_n)\}.$$

Notons que $\hat{F}(\boldsymbol{x}) = \sum_{i=1}^{n} \lambda_{i,\boldsymbol{x}} F(\boldsymbol{x}_i)$, la meilleure prédiction linéaire de $F(\boldsymbol{x})$, s'identifie à la projection orthogonale $\hat{k}(\boldsymbol{x}, .)$ de $k(\boldsymbol{x}, .)$ sur $\mathcal{F}_S$ (et non pas au minimiseur $\hat{f}$ du problème régularisé sous contraintes). On établit que la meilleure prédiction linéaire et la régression régularisée sont équivalentes en remarquant que l'évaluation de $\hat{f}$ au point $\boldsymbol{x}$ et la valeur estimée $\hat{F}(\omega_0, \boldsymbol{x})$ sachant $F(\omega_0, \boldsymbol{x}_i) = f_{\boldsymbol{x}_i}^{\text{obs}}$, $\forall i = 1, \ldots, n$ coïncident. En effet, l'évaluation de $\hat{f} \in \mathcal{F}_S$ au point $\boldsymbol{x}$ peut s'écrire

$$\begin{aligned} \hat{f}(\boldsymbol{x}) &= (k(\boldsymbol{x}, .), \hat{f})_{\mathcal{F}} \\ &= (k(\boldsymbol{x}, .) - \hat{k}(\boldsymbol{x}, .), \hat{f})_{\mathcal{F}} + (\hat{k}(\boldsymbol{x}, .), \hat{f})_{\mathcal{F}} \\ &= (\hat{k}(\boldsymbol{x}, .), \hat{f})_{\mathcal{F}} \\ (&= (\hat{k}(\boldsymbol{x}, .), f^*)_{\mathcal{F}}) \end{aligned} \tag{3.26}$$

[5] Il n'existe pas d'autre démonstration à notre connaissance dans la littérature.

(Remarquons que d'après (3.26), $\hat{k}$ est le noyau reproduisant de $\mathcal{F}_S$.)

Concluons en notant que des variables aléatoires observées $F(\omega_0, \boldsymbol{x}_i) \in \mathbb{R}$, $i = 1, \ldots, n$, s'expriment comme

$$F(\omega_0, \boldsymbol{x}_i) = f^{\text{obs}}_{\boldsymbol{x}_i} = (k(\boldsymbol{x}_i, .), f^*)_{\mathcal{F}} = (k(\boldsymbol{x}_i, .), \hat{f})_{\mathcal{F}} \,.$$

Donc,

$$\hat{F}(\omega_0, \boldsymbol{x}) = \left(\sum_{i=1}^{n} \hat{\lambda}_{i,\boldsymbol{x}} k(\boldsymbol{x}_i, .), \hat{f} \right)_{\mathcal{F}} = (\hat{k}(\boldsymbol{x}, .), \hat{f})_{\mathcal{F}} = \hat{f}(\boldsymbol{x}) \,.$$

□

En résumé, des considérations géométriques simples permettent de montrer l'équivalence entre l'interpolation régularisée dans un espace à noyau reproduisant et la prédiction linéaire sans bruit d'observation.

Observations bruitées

Si les observations sont bruitées, ou si la capacité d'approximation du modèle est limitée, nous pouvons considérer le problème de régression régularisée suivant avec une fonction de coût quadratique :

$$\text{minimiser } \|f\|^2_{\mathcal{F}} + \frac{1}{C} \sum_{i=1}^{n} (f(\boldsymbol{x}_i) - f^{\text{obs}}_{\boldsymbol{x}_i})^2 \,, \quad f \in \mathcal{F} \tag{3.27}$$

Proposition 32. *Soit $\hat{f}$ la solution du problème (3.27) et soit $F(\boldsymbol{x})$ un processus aléatoire du second ordre, de moyenne nulle et de covariance k. Considérons les variables aléatoires*

$$F^{\text{obs}}_{\boldsymbol{x}_i} = F(\boldsymbol{x}_i) + N_i \,, \quad i = 1, \ldots, n,$$

où les N_i sont des variables aléatoires indépendantes, gaussiennes, centrées, de variance $\sigma^2_N = C$. Alors $\forall \boldsymbol{x} \in \mathbb{X}$, $\hat{f}(\boldsymbol{x}) = \hat{F}(\omega_0, \boldsymbol{x})$, où $\hat{F}(\omega, \boldsymbol{x})$ est le prédicteur par krigeage de $F(\boldsymbol{x})$ à partir des variables $F^{\text{obs}}_{\boldsymbol{x}_i}$, $i = 1, \ldots, n$, et ω_0 est tel que $F^{\text{obs}}_{\boldsymbol{x}_i}(\omega_0) = f^{\text{obs}}_{\boldsymbol{x}_i}$, pour tout $i = 1, \ldots, n$.

La démonstration que nous proposons s'avère très similaire à celle du cas non bruité (nous adaptons la démonstration de Kimeldorf et Wahba (1970a) pour la rendre plus lisible). Introduisons quelques notations avant de traduire le problème (3.27) sous forme géométrique.

Soit $\mathcal{F}_N = \text{vect}\{e_i(\boldsymbol{x}), i = 1, \ldots, n\}$, l'espace vectoriel de dimension n engendré par les fonctions $e_i : \mathbb{X} \to \mathbb{R}$ définies par $e_i(\boldsymbol{x}) = C$ si $\boldsymbol{x} = \boldsymbol{x}_i$ et 0 partout ailleurs. Remarquons que tout $v \in \mathcal{F}_N$ est représenté dans la base (e_i) par le vecteur

$$\boldsymbol{v} = \left(\frac{v(\boldsymbol{x}_1)}{C}, \ldots, \frac{v(\boldsymbol{x}_n)}{C} \right)^{\mathsf{T}} .$$

Munissons $\mathcal{F}_N$ du produit scalaire

$$(v, w)_{\mathcal{F}_N} = C^{-1} \sum_{i=1}^{n} v(\boldsymbol{x}_i) w(\boldsymbol{x}_j) = C \boldsymbol{v}^{\mathsf{T}} \boldsymbol{w}.$$

D'après la section 3.2.1, $k_N(\boldsymbol{x},\boldsymbol{y}) = C^{-1}\boldsymbol{e}(\boldsymbol{x})^\mathsf{T}\boldsymbol{e}(\boldsymbol{y})$, avec $\boldsymbol{e}(\boldsymbol{x}) = (e_1(\boldsymbol{x}),\ldots,e_n(\boldsymbol{x}))^\mathsf{T}$, est le noyau reproduisant de $\mathcal{F}_N$. De plus, $\mathcal{F}_N$ est isomorphe à l'espace de variables aléatoires

$$\mathcal{H}_N = \mathrm{vect}\{N_i, i = 1,\ldots,n\}\,,$$

avec la correspondance $e_i \equiv N_i$.

Soit $\mathcal{F}' = \mathcal{F} \oplus \mathcal{F}_N$ (cette décomposition est unique). Considérons le produit scalaire sur $\mathcal{F}'$ défini pour $f', g' \in \mathcal{F}'$ par

$$(f', g')_{\mathcal{F}'} = (f, g)_{\mathcal{F}} + (v, w)_{\mathcal{F}_N}\,,$$

avec $f' = f + v$, $g' = g + w$, où f et g appartiennent à $\mathcal{F}$, et v et w appartiennent à $\mathcal{F}_N$. Sous ce produit scalaire, $\mathcal{F}$ et $\mathcal{F}_N$ sont deux espaces orthogonaux. Définissons également l'espace vectoriel

$$\mathcal{F}'_S = \mathrm{vect}\{k(\boldsymbol{x}_i,.) + k_N(\boldsymbol{x}_i,.),\ i = 1,\ldots,n\} \subset \mathcal{F}_S + \mathcal{F}_N\,,$$

qui est de dimension n. Supposons enfin les observations générées par une fonction $f^{*'} = f^* + v^* \in \mathcal{F}'$, avec $f^* \in \mathcal{F}$ et $v^* \in \mathcal{F}_N$.

Lemme 1. *Sous les hypothèses et avec les notations précédentes, le problème (3.27) est équivalent au problème suivant*

$$\textit{minimiser} \qquad \|f'\|^2_{\mathcal{F}'}\,, \quad f' \in \mathcal{F}' \tag{3.28}$$

$$\textit{sous la contrainte} \qquad f^{*'} - f' \perp \mathcal{F}'_S\,. \tag{3.29}$$

Démonstration. Si $f' = f + v$, $f \in \mathcal{F}$, $v \in \mathcal{F}_N$, satisfait (3.29), on a

$$(f^* + v^* - f - v, k(\boldsymbol{x}_i,.) + k_N(\boldsymbol{x}_i,.))_{\mathcal{F}'} = 0\,, \quad \forall i \in \{1,\ldots,n\}\,.$$

Donc $\forall i \in \{1,\ldots,n\}$,

$$\begin{aligned}(v, k_N(\boldsymbol{x}_i,.))_{\mathcal{F}'} &= v(\boldsymbol{x}_i)\\ &= (f^* + v^* - f, k(\boldsymbol{x}_i,.) + k_N(\boldsymbol{x}_i,.))_{\mathcal{F}'}\\ &= f^{*'}(\boldsymbol{x}_i) - f(\boldsymbol{x}_i)\,.\end{aligned}$$

Ceci permet d'écrire

$$\|v\|^2_{\mathcal{F}'} = \|v\|^2_{\mathcal{F}_N} = \frac{1}{C}\sum_{i=1}^{n}(f^{*'}(\boldsymbol{x}_i) - f(\boldsymbol{x}_i))^2\,.$$

Puisque $\|f'\|^2_{\mathcal{F}'} = \|f\|^2_{\mathcal{F}} + \|v\|^2_{\mathcal{F}_N}$, la fonction $\hat{f} = \hat{f}' - \hat{v}$, où $\hat{f}'$ est solution du problème (3.28)–(3.29), est bien le minimiseur de (3.27). □

Démonstration de la proposition 32. Comme précédemment, on obtient une démonstration immédiate du théorème de représentation dans le cas d'une fonction de coût quadratique. En effet, la solution $\hat{f}'$ du problème (3.28)–(3.29) satisfait nécessairement $\hat{f}' = \hat{f} + \hat{v} \in \mathcal{F}'_S$ et par suite, $\hat{f} \in \mathcal{F}_S$.

Soit $\hat{k}(\boldsymbol{x},.)$ le projeté orthogonal de $k(\boldsymbol{x},.)$ sur $\mathcal{F}'_S$, isomorphe à $\hat{F}(\boldsymbol{x})$.

$$\begin{aligned}\hat{f}(\boldsymbol{x}) &= (k(\boldsymbol{x},.) - \hat{k}(\boldsymbol{x},.), \hat{f} + \hat{v})_{\mathcal{F}'} - (k(\boldsymbol{x},.) - \hat{k}(\boldsymbol{x},.), \hat{v})_{\mathcal{F}'} + (\hat{k}(\boldsymbol{x},.), \hat{f})_{\mathcal{F}'}\\ &= (\hat{k}(\boldsymbol{x},.), \hat{f} + \hat{v})_{\mathcal{F}'} - (k(\boldsymbol{x},.), \hat{v})_{\mathcal{F}'}\\ &= (\hat{k}(\boldsymbol{x},.), \hat{f}')_{\mathcal{F}'}\,.\end{aligned}$$

De plus,

$$F(\omega_0, \boldsymbol{x}_j) + N_j(\omega_0) = f^{*'}(\boldsymbol{x}_j) = (k(\boldsymbol{x}_j, .) + k_N(\boldsymbol{x}_j, .), \hat{f}')_{\mathcal{F}'} ,$$

implique

$$\hat{F}(\omega_0, \boldsymbol{x}) = \left(\sum_{i=1}^{n} \lambda_{i,\boldsymbol{x}} (k(\boldsymbol{x}_i, .) + k_N(\boldsymbol{x}_i, .)), \hat{f}' \right)_{\mathcal{F}'} = (\hat{k}(\boldsymbol{x}, .), \hat{f}')_{\mathcal{F}'} = \hat{f}(\boldsymbol{x}) .$$

□

Des considérations géométriques simples permettent donc de traduire le problème d'approximation avec coût quadratique, régularisée par une norme d'espace à noyau reproduisant, en termes de prédiction linéaire d'un processus aléatoire. Dans le cadre de la modélisation boîte noire de système, il n'y a donc pas de différence entre ces deux approches. Seul le cadre interprétatif change. Toutefois, cette interprétation double conduit à différentes méthodes pour le choix du noyau, comme nous le verrons au chapitre 5. Dans les sections suivantes, nous généralisons l'interprétation probabiliste lorsque l'on utilise d'autres fonctions coût que la fonction quadratique.

3.6 Fonctions d'attache aux données

3.6.1 Généralités

Le choix d'une fonction de coût est nécessairement lié à la possibilité d'optimiser ensuite le critère correspondant. Le cas de la fonction coût quadratique est particulièrement simple de ce point de vue, ce qui la rend intéressante malgré ses défauts de robustesse vis-à-vis des données aberrantes. La fonction de coût ε–insensible utilisée dans les méthodes à vecteurs de support (Vapnik, 1995 ; Vapnik et al., 1997 ; Schölkopf et Smola, 2002) est aussi un choix pertinent, comme nous le verrons dans les sections 3.6.3 et 3.6.4. Dans cette section, le problème du choix du terme d'attache aux données est discuté en introduisant la notion de *risque*.

Fonction d'attache aux données et fonction de coût

Une fonction d'attache aux données, notée

$$c\left((\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n), (f^{\text{obs}}_{\boldsymbol{x}_1}, \ldots, f^{\text{obs}}_{\boldsymbol{x}_n}), (f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_n))\right)$$

est une fonction à valeurs dans $\mathbb{R}^+ \cup +\infty$ des vecteurs de facteurs $\boldsymbol{x}_i$, des observations $f^{\text{obs}}_{\boldsymbol{x}_i}$ et des valeurs prédites $f(\boldsymbol{x}_i)$ correspondantes. En général, les fonctions retenues ne prennent pas en compte la dépendance éventuelle entre les observations, si bien que l'on considère en fait des fonctions d'attache aux données sous la forme

$$c\left((\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n), (f^{\text{obs}}_{\boldsymbol{x}_1}, \ldots, f^{\text{obs}}_{\boldsymbol{x}_n}), (f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_n))\right) = \sum_{j=1}^{n} l(\boldsymbol{x}_j, f^{\text{obs}}_{\boldsymbol{x}_j}, f(\boldsymbol{x}_j))$$

La fonction $l(\boldsymbol{x}, s, t) : \mathbb{X} \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}^+$ est une fonction de coût, qui doit satisfaire $l(\boldsymbol{x}, s, s) = 0$, $\forall \boldsymbol{x} \in \mathbb{X}$ et $\forall s \in \mathbb{R}$. On souhaite généralement que le coût d'une prédiction augmente avec l'écart $|f^{\text{obs}}_{\boldsymbol{x}} - f(\boldsymbol{x})|$. Par ailleurs, il est souhaitable que l soit convexe en t pour garantir l'unicité des solutions du problème régularisé. Une simplification supplémentaire consiste à omettre la dépendance de l en $\boldsymbol{x}$, ce qui conduit à considérer la fonction d'attache aux données (3.13).

Risque en espérance

Dans la théorie de l'apprentissage, voir (Vapnik, 1995 ; Schölkopf et Smola, 2002), il est classique de relier la fonction d'attache aux données à la notion de *risque empirique*. Dans ce contexte, on suppose que les données observées sont des réalisations d'un vecteur aléatoire $(\boldsymbol{X}, F^{\text{obs}}) \in \mathbb{X}\times\mathbb{R}$ admettant une loi de probabilité $\mathsf{P}_{\boldsymbol{X},F^{\text{obs}}}$. Étant donné une fonction de coût l, le *risque en espérance* d'une fonction f, défini par

$$R[f] = \mathrm{E}[l(\boldsymbol{X}, F^{\text{obs}}, f(\boldsymbol{X}))] = \int_{\mathbb{X},\mathbb{R}} l(\boldsymbol{x}, f^{\text{obs}}_{\boldsymbol{x}}, f(\boldsymbol{x}))d\mathsf{P}_{\boldsymbol{X},F^{\text{obs}}}(\boldsymbol{x}, f^{\text{obs}}_{\boldsymbol{x}}) \,,$$

quantifie la qualité de l'approximation.

Risque empirique

Si la probabilité $\mathsf{P}_{\boldsymbol{X},F^{\text{obs}}}$ est inconnue, $R[f]$ n'est pas calculable. $\mathsf{P}_{\boldsymbol{X},F^{\text{obs}}}$ peut dans ce cas être approximée par la mesure de probabilité empirique obtenue à partir des données

$$\mathsf{P}_{\text{emp}} = \frac{1}{n}\sum_{j=1}^{n} \delta_{\boldsymbol{x}_i}(\boldsymbol{x})\delta_{f^{\text{obs}}_{\boldsymbol{x}_i}}(f^{\text{obs}}_{\boldsymbol{x}}) \,.$$

Le *risque empirique* $R_{\text{emp}}[f]$ s'obtient en remplaçant la vraie probabilité des observations par la probabilité empirique dans le risque en espérance :

$$R_{\text{emp}}[f] = \frac{1}{n}\sum_{j=1}^{n} l(\boldsymbol{x}_i, f^{\text{obs}}_{\boldsymbol{x}_i}, f(\boldsymbol{x}_i)) \,. \tag{3.30}$$

Le risque empirique correspond donc à une fonction classique d'attache aux données[6]. Cependant, introduire la notion de risque empirique ne fournit aucun critère permettant de choisir la fonction de coût. Dans les sections suivantes, cette question est abordée en utilisant deux points de vue.

3.6.2 Maximum a posteriori

La loi des observations est ici supposée connue. Plus précisément, on considère un processus aléatoire $F(\boldsymbol{x})$, supposé gaussien, de moyenne nulle et de covariance $k(\boldsymbol{x}, \boldsymbol{y})$, et $N_1, \ldots, N_n$, des variables aléatoires à valeurs dans $\mathbb{R}$, indépendantes, admettant une densité de probabilité notée p_N. Les variables aléatoires $F^{\text{obs}}_{\boldsymbol{x}_i} = F(\boldsymbol{x}_i) + N_i$, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{X}$, modélisent les observations. On définit les vecteurs aléatoires $\boldsymbol{F}_S = (F(\boldsymbol{x}_1), \ldots, F(\boldsymbol{x}_n))^{\mathsf{T}}$, $\boldsymbol{N} = (N_1, \ldots, N_n)^{\mathsf{T}}$, $\boldsymbol{F}^{\text{obs}} = (F^{\text{obs}}_{\boldsymbol{x}_1}, \ldots, F^{\text{obs}}_{\boldsymbol{x}_n})^{\mathsf{T}}$.

On souhaite décomposer le problème de prédiction de $F(\boldsymbol{x})$ en deux sous problèmes distincts. Le premier problème consiste à prédire le vecteur $\boldsymbol{F}_S$ à partir de $\boldsymbol{F}^{\text{obs}}$. De manière équivalente, il s'agit d'estimer les variables aléatoires N_i. Le second problème consiste à prédire $F(\boldsymbol{x})$ à partir de $\boldsymbol{F}_S$. Il s'agit d'un problème d'interpolation que l'on sait bien traiter. Nous nous intéressons donc au premier problème et choisissons l'approche du *maximum a posteriori*, qui conduit à calculer

$$\arg\max_{\boldsymbol{f}} p(\boldsymbol{f} \mid \boldsymbol{f}^{\text{obs}}) = \arg\max_{\boldsymbol{f}} \frac{p(\boldsymbol{f}^{\text{obs}} \mid \boldsymbol{f})p(\boldsymbol{f})}{p(\boldsymbol{f}^{\text{obs}})} = \arg\max_{\boldsymbol{f}} p(\boldsymbol{f}^{\text{obs}} \mid \boldsymbol{f})p(\boldsymbol{f}) \,.$$

[6] Dans la littérature de la théorie de l'apprentissage, le discours usuel consiste à remarquer que, contrairement au risque en espérance, le risque empirique n'est pas un critère d'approximation satisfaisant (il peut toujours être annulé si on considère une classe de fonctions suffisamment large), ce qui justifie l'utilisation d'un terme de régularisation.

La loi de probabilité du couple de vecteurs aléatoires $(\boldsymbol{F}^{\text{obs}}, \boldsymbol{F}_S)$ admet la densité

$$p(\boldsymbol{f}^{\text{obs}}, \boldsymbol{f}) = \frac{1}{(\det\ 2\pi\boldsymbol{K})^{1/2}} \exp\Big(-\frac{1}{2}\boldsymbol{f}^{\mathsf{T}}\boldsymbol{K}^{-1}\boldsymbol{f}\Big) \prod_{i=1}^{n} p_N\big(f^{\text{obs}}_{\boldsymbol{x}_i} - f_{[i]}\big)\,.$$

Le logarithme de cette densité s'écrit

$$\log p(\boldsymbol{f}, \boldsymbol{f}^{\text{obs}}) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\det\boldsymbol{K} - \frac{1}{2}\boldsymbol{f}^{\mathsf{T}}\boldsymbol{K}^{-1}\boldsymbol{f} + \sum_{i=1}^{n}\log p_N\big(f^{\text{obs}}_{\boldsymbol{x}_i} - f_{[i]}\big)\,.$$

Estimer $\boldsymbol{F}_S$ par la méthode du maximum a posteriori conduit par conséquent à minimiser la quantité

$$L(\boldsymbol{f}) = \frac{1}{2}\boldsymbol{f}^{\mathsf{T}}\boldsymbol{K}^{-1}\boldsymbol{f} - \sum_{i=1}^{n}\log p_N\big(f^{\text{obs}}_{\boldsymbol{x}_i} - f_{[i]}\big)\,. \tag{3.31}$$

Proposition 33. *Soit la fonction de coût*

$$l(s,t) = -2\log p_N(t-s)\,.$$

Avec ce choix, le problème de régression régularisée qui consiste à minimiser (3.14) est équivalent au problème suivant, lui-même constitué de deux sous problèmes distincts :

i. Estimer $\boldsymbol{F}$ par la méthode du maximum a posteriori, c'est-à-dire minimiser (3.31) pour obtenir $\hat{\boldsymbol{f}}$.

ii. Chercher $\hat{f} \in \mathcal{F}$ telle que $\|\hat{f}\|_{\mathcal{F}}$ soit minimale sous contrainte $\hat{f}(\boldsymbol{x}_i) = \hat{f}_{[i]}$ (problème d'interpolation).

Démonstration. Soit $\mathcal{F}$ l'espace hilbertien admettant le noyau reproduisant k. Supposons que $\hat{f} \in \mathcal{F}$ minimise (3.14). D'après le théorème du représentant, $\exists a_1, \ldots, a_n$ tels que

$$\hat{f}(\boldsymbol{x}) = \sum_{i=1}^{n} a_i k(\boldsymbol{x}_i, \boldsymbol{x})\,.$$

Donc $\hat{\boldsymbol{f}} = (\hat{f}(\boldsymbol{x}_1), \ldots, \hat{f}(\boldsymbol{x}_n))^{\mathsf{T}}$ est tel que $\hat{\boldsymbol{f}} = \boldsymbol{K}\boldsymbol{a}$, avec $\boldsymbol{a} = (a_1, \ldots, a_n)^{\mathsf{T}}$ et par suite,

$$\|\hat{f}\|^2_{\mathcal{F}} = \boldsymbol{a}^{\mathsf{T}}\boldsymbol{K}\boldsymbol{a} = \hat{\boldsymbol{f}}^{\mathsf{T}}\boldsymbol{K}^{-1}\hat{\boldsymbol{f}}\,.$$

Par conséquent, $\hat{\boldsymbol{f}}$ est solution du problème (*i*). D'autre part, $\hat{f}$ est solution du problème (*ii*) puisque elle est déterminée de manière unique par $\hat{\boldsymbol{f}}$.

Réciproquement, supposons $\hat{\boldsymbol{f}} = (\hat{f}_{[1]}, \ldots, \hat{f}_{[n]})^{\mathsf{T}}$ solution du problème (*i*) et soit $\hat{f} \in \mathcal{F}$ la fonction de norme minimale, vérifiant $\hat{f}(\boldsymbol{x}_i) = \hat{f}_{[i]}$, $i = 1, \ldots, n$. Comme $\|\hat{f}\|^2_{\mathcal{F}} = \hat{\boldsymbol{f}}^{\mathsf{T}}\boldsymbol{K}^{-1}\hat{\boldsymbol{f}}$, on identifie le premier terme de (3.31) au terme de régularisation de (3.14). Le second terme de (3.31) correspond au terme d'attache aux données de (3.14), $-2\sum_{i=1}^{n}\log p_N(f^{\text{obs}}_{\boldsymbol{x}_i} - \hat{f}(\boldsymbol{x}_i))$. Donc $\hat{f}$ est bien la solution du problème (3.14). □

Exemple — Si le bruit d'observation est gaussien, centré et de variance σ^2_N, (3.31) s'exprime sous la forme déjà rencontrée

$$L(\boldsymbol{f}) = \frac{1}{2}\Big(\|f\|^2_{\mathcal{F}} + \frac{1}{\sigma^2_N}\sum_{i=1}^{n}(f^{\text{obs}}_{\boldsymbol{x}_i} - f(\boldsymbol{x}_i))^2\Big)\,.$$

Dans le cas d'une fonction de coût $l(s,t) = \rho(s-t)$, la proposition 33 montre donc que le problème (3.14) se ramène à un problème de prédiction d'un processus aléatoire gaussien observé avec un bruit additif, bruit dont la loi s'exprime en fonction de ρ sous la forme

$$p_N(s) = \frac{1}{Z} \exp\left(-\frac{1}{2}\rho(s)\right), \quad Z \in \mathbb{R}.$$

Nous constatons également que le terme d'attache aux données peut être choisi de façon à obtenir une reconstruction de f^* optimale au sens du maximum a posteriori, à condition de connaître la loi du bruit d'observation.

En pratique cependant, la loi du bruit d'observation est inconnue. Notons d'ailleurs que nous avons choisi d'attribuer le rôle de bruit d'observation aux variables N_i, mais que cela est arbitraire, dans la mesure où il existe une infinité de modèles possibles pour la sortie du système. Les variables aléatoires N_i pourraient aussi modéliser une erreur structurelle, commise en choisissant un processus aléatoire particulier plutôt qu'un autre.

Choisir les variables de bruit avec une loi normale conduit à une solution simple du problème de régression, comme nous l'avons vu précédemment (d'un point de vue analytique et numérique). Dans la plupart des cas, nous utiliserons ce type de modèle. Cependant, cette approche présente l'inconvénient d'être relativement sensible aux erreurs de modélisation du bruit d'observation, notamment s'il existe des données dites aberrantes. Dans la section suivante, nous rappelons la notion d'*estimation robuste*. Nous parlerons ensuite de la méthode de *régression à vecteurs de support*.

3.6.3 Estimation robuste

Le coût quadratique est optimal lorsque le modèle du bruit d'observation est gaussien et à variance connue, comme nous l'avons vu ci-dessus. Il est cependant bien connu que l'estimation par moindre carrés peut conduire à des résultats insatisfaisants lorsque la loi du bruit s'éloigne de la normalité (notamment si cette loi possède une densité décroissant lentement loin de la moyenne). L'une des solutions pour rendre un estimateur robuste à la loi du bruit consiste à enlever les observations extrêmes (celles qui correspondent par exemple à des valeurs aberrantes) avant d'effectuer la régression par moindre carrés. Une autre approche consiste à utiliser un coût moins vulnérable aux réalisations extrêmes du bruit que le coût quadratique. Les *M-estimateurs* (estimateurs au sens du maximum de vraisemblance), introduits par Huber (1964), constituent une famille très classique de méthodes de régression robuste.

Ces estimateurs sont traditionnellement utilisés dans le cas de modèles linéairement paramétrés $f_{\boldsymbol{b}}(\boldsymbol{x}) = \boldsymbol{b}^\mathsf{T}\boldsymbol{r}(\boldsymbol{x})$, où $\boldsymbol{b} \in \mathbb{R}^l$ est un vecteur de paramètres. Plus précisément, supposons le modèle générant les données observées de la forme

$$F^{\text{obs}}_{\boldsymbol{x}_i} = f_{\boldsymbol{b}^*}(\boldsymbol{x}_i) + N_i\,, \quad i = 1, \ldots, n\,,$$

où les variables aléatoires N_i, supposées indépendantes et identiquement distribuées, modélisent le bruit d'observation.

Considérons alors le problème de régression non régularisée suivant : minimiser le risque empirique (3.13) dans la classe de fonctions linéairement paramétrées $\{f_{\boldsymbol{b}} : \boldsymbol{x} \mapsto \boldsymbol{b}^\mathsf{T}\boldsymbol{r}(\boldsymbol{x}), \boldsymbol{b} \in \mathbb{R}^l\}$.

Notons $e_i = f_{\hat{\boldsymbol{b}}}(\boldsymbol{x}_i) - f^{\text{obs}}_{\boldsymbol{x}_i}$ les résidus du modèle (les différences entre les sorties du modèle et les données expérimentales). Par définition, un M-estimateur de $\boldsymbol{b}^*$ minimise le risque empirique (3.13), où l'on considère une fonction de coût $l(s,t) = \rho(s-t)$, avec la fonction $\rho : \mathbb{R} \to \mathbb{R}^+$ telle que

- $\rho(0) = 0$,
- $\rho(e) = \rho(-e)$,
- $\rho(e_i) \geq \rho(e_j)$ si $|e_i| \geq |e_j|$.

Le risque empirique s'écrit donc explicitement sous la forme

$$J_{f_{\boldsymbol{b}^*}}(f_{\boldsymbol{b}}) = \sum_{i=1}^{n} \rho(f_{\boldsymbol{b}}(\boldsymbol{x}_i) - f^{\text{obs}}_{\boldsymbol{x}_i}) \tag{3.32}$$

Notons ψ la dérivée première de ρ. En différenciant (3.32) par rapport à $\boldsymbol{b}$, et en annulant le gradient, nous obtenons un système d'équations que nous pouvons écrire sous la forme matricielle

$$\sum_{i=1}^{n} \psi(e_i)\boldsymbol{r}(\boldsymbol{x}_i) = \boldsymbol{0}\,. \tag{3.33}$$

Afin de comprendre de manière heuristique la notion de robustesse, il est pratique de définir la *fonction poids* $w(e) = \psi(e)/e$. (Cette fonction n'est pas nécessairement bien définie pour tout e, mais il s'agit d'une présentation informelle.) Le système (3.33) peut alors se réécrire sous la forme

$$\sum_{i=1}^{n} w(e_i)e_i\boldsymbol{r}(\boldsymbol{x}_i) = \boldsymbol{0}\,.$$

Résoudre ce dernier système est équivalent à un problème de moindre carrés pondérés (Walter et Pronzato, 1997), où le critère à minimiser est $\sum(w(e_i)e_i)^2$.

Dans le cas de l'estimateur des moindres carrés, $w(e) = 1$ pour tout e. Afin de minimiser l'influence des valeurs extrêmes, l'idée de la régression robuste est de prendre des poids décroissants lorsque $|e|$ augmente. Notons que l'on peut également voir le problème de régression robuste comme celui de choisir des fonctions ψ croissant moins vite que la fonction linéaire. La figure 3.3 montre des fonctions ψ et w correspondant à des fonctions coût classiques. Dans le cas du coût *ε-insensible* (voir la section 3.6.4), introduire la fonction poids n'a pas vraiment de sens. (Remarquons que prendre un coût constant au voisinage de 0 est un choix très singulier.) Cela permet toutefois de constater la ressemblance de ce coût avec celui de Huber. L'analyse rigoureuse de la robustesse nécessiterait l'introduction les fonctions d'influence (Huber, 1981) que nous ne présenterons pas ici[7].

La notion d'*efficacité* d'estimation permet de quantifier la qualité d'un M-estimateur en présence d'un bruit de loi donnée (pour un bruit gaussien, nous verrons que le coût quadratique est optimal au sens de l'efficacité d'estimation). Dans les paragraphes suivants, nous nous intéressons plus spécifiquement au choix du paramètre ε introduit dans les coûts du table 3.6.3. Remarquons que les M-estimateurs de $\boldsymbol{b}^*$ ne sont pas invariants par rapport à des changements d'échelle des résidus (quantifiée par exemple par leur écart type σ_e). On constate, en regardant par exemple

[7] Remarque : la solution numérique d'un problème de régression avec poids peut être obtenue par exemple à l'aide de l'algorithme itératif *IRLS* (*iteratively reweighted least-squares*). Dans le cas du coût ε-insensible, la solution s'obtient par programmation quadratique, et correspond aux méthodes à vecteurs de support, qui seront vues dans la section 3.6.4.

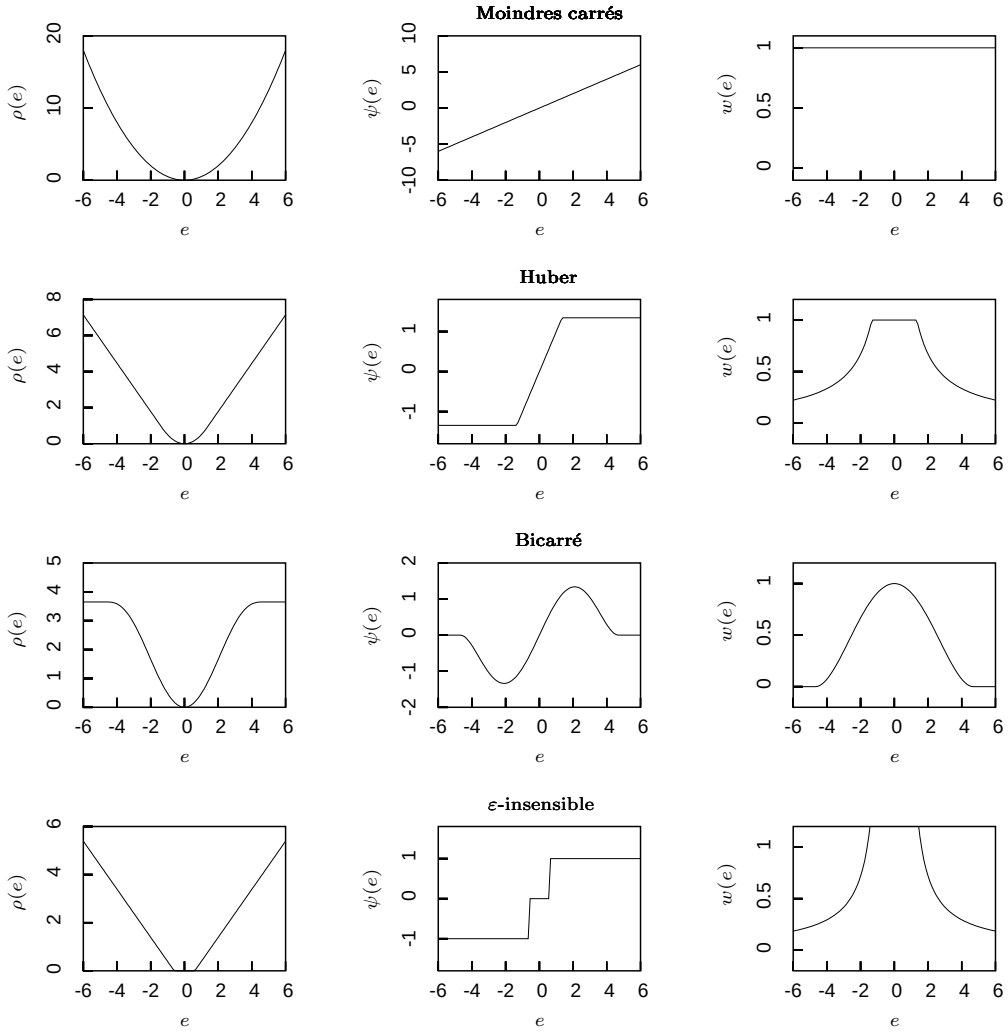


FIG. 3.3 – Fonctions coût ρ, dérivée ψ et poids w pour l'estimateur des moindres carrés (première ligne), de Huber (deuxième ligne), bicarré (troisième ligne) et ε-insensible (dernière ligne). Sur ces graphes, nous avons choisi $\varepsilon = 1.345$ pour l'estimateur de Huber, $\varepsilon = 4.685$ pour l'estimateur bicarré et $\varepsilon = 0.61$ pour l'estimateur ε-insensible.

Estimateur	fonction coût	fonction poids
Moindres carrés	$\rho(e) = e^2$	$w(e) = 1$
Huber	$\rho(e) = \begin{cases} \frac{1}{2}e^2 & \text{pour } \lvert e\rvert \le \varepsilon \\ \varepsilon\lvert e\rvert - \frac{1}{2}e^2 & \text{pour } \lvert e\rvert > \varepsilon \end{cases}$	$w(e) = \begin{cases} 1 & \text{pour } \lvert e\rvert \le \varepsilon \\ \varepsilon/\lvert e\rvert & \text{pour } \lvert e\rvert > \varepsilon \end{cases}$
Bicarré	$\rho(e) = \begin{cases} \frac{\varepsilon^2}{6}\left\{1 - \left[1 - \left(\frac{e}{\varepsilon}\right)^2\right]^3\right\}, & \lvert e\rvert \le \varepsilon \\ \frac{\varepsilon^2}{6}, & \lvert e\rvert > \varepsilon \end{cases}$	$w(e) = \begin{cases} \left[1 - \left(\frac{e}{\varepsilon}\right)^2\right]^2, & \lvert e\rvert \le \varepsilon \\ 0, & \lvert e\rvert > \varepsilon \end{cases}$
ε-insensible	$\rho(e) = \begin{cases} 0 & \text{pour } \lvert e\rvert \le \varepsilon \\ \lvert e\rvert - \varepsilon & \text{pour } \lvert e\rvert > \varepsilon \end{cases}$	$w(e) = \begin{cases} +\infty & \text{pour } \lvert e\rvert \le \varepsilon \\ \frac{1}{\lvert e\rvert - \varepsilon} & \text{pour } \lvert e\rvert > \varepsilon \end{cases}$

TAB. 3.1 – Fonctions coût et poids correspondant aux estimateurs des moindres carrés, de Huber, bicarrés et ε-insensible.

le coût de Huber, que prendre ε petit rend l'estimateur moins vulnérable aux valeurs extrêmes. Toutefois, l'estimateur devient alors moins performant au sens de l'efficacité, dont nous rappelons la définition ci-dessous.

Notons $\hat{\boldsymbol{b}}$ un estimateur non biaisé de $\boldsymbol{b}^*$. L'inégalité de Cramér-Rao fournit une borne inférieure de la matrice de covariance $\boldsymbol{B}_{\hat{\boldsymbol{b}}}$ du vecteur aléatoire $\hat{\boldsymbol{b}}$, égale à l'inverse de la matrice d'information de Fisher $\boldsymbol{I}_{\boldsymbol{b}^*}$ d'éléments

$$\boldsymbol{I}_{\boldsymbol{b}^*,[i,j]} = \mathrm{E}_{p(\boldsymbol{f}^{\mathrm{obs}}|\boldsymbol{b}^*)}\left[\partial_{\boldsymbol{b}^*{}_{[i]}} \log p(\boldsymbol{F}^{\mathrm{obs}} \mid \boldsymbol{b}^*) \cdot \partial_{\boldsymbol{b}^*{}_{[j]}} \log p(\boldsymbol{F}^{\mathrm{obs}} \mid \boldsymbol{b}^*)\right] .$$

(Le rôle de l'information de Fisher sera revu plus en détails dans la section 5.4.3.) On définit l'*efficacité* par $e = 1/(\det \boldsymbol{I}_{\boldsymbol{b}^*}\boldsymbol{B}_{\hat{\boldsymbol{b}}})$. On dit qu'un estimateur est efficace s'il atteint la borne de Cramér-Rao ($e = 1$).

Proposition 34. *Pour un estimateur du type*

$$\hat{\boldsymbol{b}} = \arg\min_{\boldsymbol{b}} c(\boldsymbol{F}^{\mathrm{obs}}, \boldsymbol{f}_{\boldsymbol{b}}) ,$$

où c est une fonction d'attache aux données deux fois différentiable par rapport $\boldsymbol{b}$, on a asymptotiquement $\boldsymbol{B}_{\hat{\boldsymbol{b}}} = \boldsymbol{Q}^{-1}\boldsymbol{G}\boldsymbol{Q}^{-1}$, avec $\boldsymbol{Q} = (q_{i,j})_{i,j=1}^{l}$, $\boldsymbol{G} = (g_{i,j})_{i,j=1}^{l}$, et

$$\begin{aligned} g_{i,j} &= \mathrm{Cov}_{p(\boldsymbol{f}^{\mathrm{obs}}|\boldsymbol{b}^*)}[\partial_{b_{[i]}} c(\boldsymbol{F}^{\mathrm{obs}}, \boldsymbol{f}_{\boldsymbol{b}^*}), \partial_{b_{[j]}} c(\boldsymbol{F}^{\mathrm{obs}}, \boldsymbol{f}_{\boldsymbol{b}^*})] \\ q_{i,j} &= \mathrm{E}_{p(\boldsymbol{f}^{\mathrm{obs}}|\boldsymbol{b}^*)}[\partial^2_{b_{[i]},b_{[j]}} c(\boldsymbol{F}^{\mathrm{obs}}, \boldsymbol{f}_{\boldsymbol{b}^*})]. \end{aligned}$$

Démonstration. Voir par exemple Murata et al. (1994), lemme 3. □

D'après la proposition précédente $e = (\det \boldsymbol{Q})^2/(\det \boldsymbol{I}_{\boldsymbol{b}^*}\boldsymbol{G})$. L'estimateur du maximum de vraisemblance $\hat{\boldsymbol{b}} = \arg\min_{\boldsymbol{b}} \log p(\boldsymbol{F}^{\mathrm{obs}} \mid \boldsymbol{b})$ est asymptotiquement efficace[8] (dans ce cas en effet, $\boldsymbol{I}_{\boldsymbol{b}^*} = \boldsymbol{G} = -\boldsymbol{Q}$).

La notion d'efficacité rappelée ci-dessus permet d'établir un critère de choix du paramètre d'échelle ε des M-estimateurs précédents. Généralement, on ajuste ε en fonction de l'écart type σ_N du bruit d'observation de manière à obtenir une efficacité (asymptotique) relativement élevée lorsque le bruit d'observation possède une loi gaussienne $\mathcal{N}(0, \sigma_N^2)$.

[8] Attention aux conditions supplémentaires à vérifier pour que l'estimateur du maximum de vraisemblance soit asymptotiquement efficace (Monfort, 1997).

Proposition 35. *La matrice de covariance $\boldsymbol{B}_{\hat{\boldsymbol{b}}}$ d'un M-estimateur est asymptotiquement égale à*

$$\frac{\mathrm{E}[\psi(N)^2]}{\mathrm{E}[\psi'(N)]^2}\boldsymbol{K}^{-1}\,,$$

où N est une variable aléatoire de bruit d'observation[9] *et $\boldsymbol{K} = \sum_{i=1}^n \boldsymbol{r}(\boldsymbol{x}_i)\boldsymbol{r}(\boldsymbol{x}_i)^\mathsf{T}$.*

Démonstration. Calculons la matrice $\boldsymbol{G}$ correspondant à la fonction d'attache aux données

$$c(\boldsymbol{F}^{\mathrm{obs}}, \boldsymbol{f}_{\boldsymbol{b}}) = \sum_{i=1}^n \rho(f_{\boldsymbol{b}}(\boldsymbol{x}_i) - F^{\mathrm{obs}}_{\boldsymbol{x}_i}).$$

On a

$$\begin{aligned}\partial_{\boldsymbol{b}} c(\boldsymbol{F}^{\mathrm{obs}}, \boldsymbol{f}_{\boldsymbol{b}^*}) &= \sum_{i=1}^n \psi(f_{\boldsymbol{b}^*}(\boldsymbol{x}_i) - F^{\mathrm{obs}}_{\boldsymbol{x}_i})\boldsymbol{r}(\boldsymbol{x}_i) \\ &= -\sum_{i=1}^n \psi(N_i)\boldsymbol{r}(\boldsymbol{x}_i)\,.\end{aligned}$$

De plus, $\mathrm{E}[\psi(N_i)] = 0$, $i = 1, 2\ldots$, car ψ est impaire et les variables N_i sont de moyenne nulle et ont des densités paires. Par conséquent,

$$\begin{aligned}\boldsymbol{G} &= \mathrm{E}\left[\left(\sum_i \psi(N_i)\boldsymbol{r}(\boldsymbol{x}_i)\right)\left(\sum_j \psi(N_j)\boldsymbol{r}(\boldsymbol{x}_j)^\mathsf{T}\right)\right] \\ &= \sum_i \mathrm{E}[\psi(N_i)^2]\boldsymbol{r}(\boldsymbol{x}_i)\boldsymbol{r}(\boldsymbol{x}_i)^\mathsf{T} \\ &= \mathrm{E}[\psi(N)^2]\boldsymbol{K}\,,\end{aligned}$$

avec $\boldsymbol{K} = \sum_{i=1}^n \boldsymbol{r}(\boldsymbol{x}_i)\boldsymbol{r}(\boldsymbol{x}_i)^\mathsf{T}$. En utilisant

$$\partial^2_{\boldsymbol{b},\boldsymbol{b}} c(\boldsymbol{F}^{\mathrm{obs}}, \boldsymbol{f}_{\boldsymbol{b}^*}) = \sum_{i,j} \psi'(N_i)\boldsymbol{r}(\boldsymbol{x}_i)\boldsymbol{r}(\boldsymbol{x}_i)^\mathsf{T}\,,$$

nous obtenons $\boldsymbol{Q} = \mathrm{E}[\psi'(N)]\boldsymbol{K}$. □

D'après les éléments de la preuve précédente, la matrice d'information de Fisher correspondant à un bruit d'observation gaussien de variance σ_N^2 est $\sigma_N^{-2}\boldsymbol{K}$. Par conséquent, l'efficacité asymptotique d'un M-estimateur pour un bruit d'observation gaussien est

$$\frac{\mathrm{E}[\psi'(N)]^2}{\mathrm{E}[\psi(N)^2]}\sigma_N^2.$$

Par exemple, pour l'estimateur de Huber, $\varepsilon = 1.345\sigma_N$ conduit à une efficacité de 0.95 lorsque le bruit d'observation est gaussien, en offrant de plus une protection contre les valeurs observées extrêmes. Pour l'estimateur bicarré, la même efficacité est obtenue lorsque $\varepsilon = 4.685\sigma_N$. Détaillons le calcul pour le coût ε-insensible (Smola et al., 1998). On a

$$\mathrm{E}[\psi(N)^2] = 2\int_\varepsilon^\infty p_N(s)ds = 1 - \mathrm{erf}\left(\frac{\varepsilon}{\sqrt{2\sigma_N^2}}\right)\,,$$

[9] Rappelons que le bruit d'observation correspond à des variables aléatoires N_i indépendantes, de loi symétrique et de variance finie.

et

$$\mathrm{E}[\psi'(N)] = 2p_N(\varepsilon) = \sqrt{\frac{2}{\pi\sigma_N^2}} \exp\left(-\frac{\varepsilon}{2\sigma_N^2}\right) .$$

Par conséquent,

$$e = \frac{2}{\pi} \exp\left(-\frac{\varepsilon^2}{\sigma_N^2}\right) \left(1 - \mathrm{erf}\left(\frac{\varepsilon}{\sqrt{2\sigma_N^2}}\right)\right)^{-1} .$$

Nous constatons que l'efficacité de l'estimateur ε-insensible pour un bruit gaussien est une fonction du rapport ε/σ_N. L'efficacité atteint un maximum d'environ 0.81 pour $\varepsilon \approx 0.61\sigma_N$ (voir la figure 3.4). On constate que l'estimateur ε-insensible possède une efficacité supérieure à celle d'un estimateur L^1 (d'efficacité 0.63). Il y a donc avantage à utiliser un estimateur ε-insensible au lieu de d'un estimateur L^1. Cependant, l'estimation correcte de la variance du bruit d'observation est essentielle car l'efficacité de l'estimateur ε-insensible chute significativement si ε est choisi trop grand.

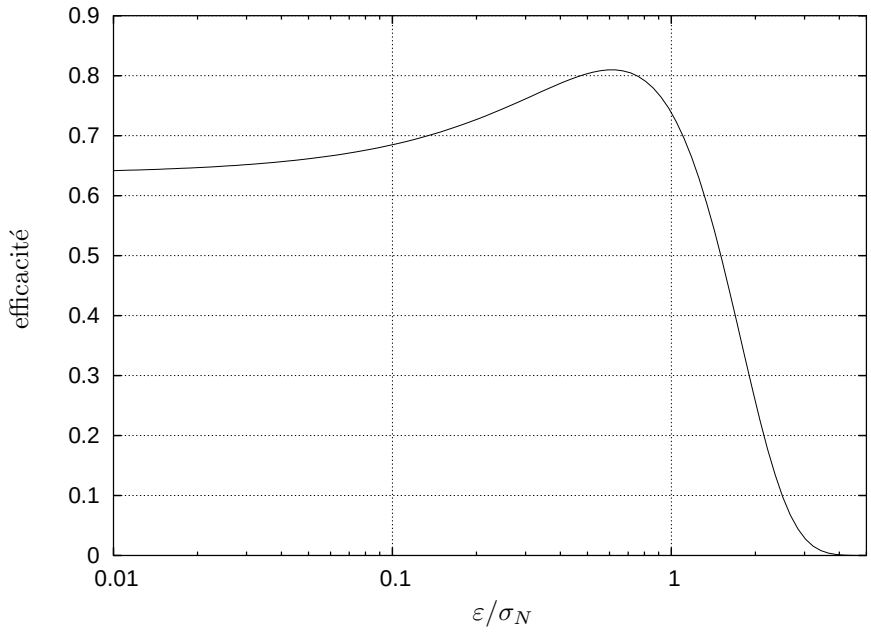


FIG. 3.4 – Efficacité de l'estimateur ε-insensible pour un bruit gaussien en fonction logarithmique du rapport entre le paramètre ε et l'écart-type du bruit

Notons cependant que la présentation effectuée ci-dessus s'applique seulement à l'estimation non régularisée des paramètres d'un modèle linéaire. Nous n'avons pas cherché à étendre les conclusions sur la robustesse des M-estimateurs au cas général de la régression régularisée.

3.6.4 Application à la régression à vecteurs de support

Méthodes à vecteurs de support

Les méthodes à vecteurs de supports, *support vector machines* (*SVM*) en anglais, ont été introduites par Vapnik dans le cadre de la classification, (Vapnik, 1995). Dans le cas de deux classes (ou catégories) linéairement séparables, une procédure de classification consiste à trouver un hyperplan tel que les données avec des étiquettes différentes soient de part et d'autre de cet hyperplan (voir la section 3.2.3). L'idée naturelle de Vapnik est de trouver l'hyperplan séparateur $\hat{H}$ de telle sorte que la distance d entre les données et l'hyperplan soit maximale. Chercher $\hat{H}$ maximisant d est un problème variationnel admettant une solution simple. Parmi tous les hyperplans séparateurs parallèles à $\hat{H}$, il existe nécessairement deux hyperplans, appelés marges, portant au moins une donnée. Les données portées par ces marges s'appellent *vecteurs de support*. L'un des aspects rendant les *SVM* populaires est que les vecteurs de support constituent en quelque sorte une *représentation creuse* des données (*sparse representation*) puisque le problème de classification linéaire ne nécessite que la connaissance de ces vecteurs de support.

La méthode *SVM* développée pour la classification a donné naissance à une méthode similaire dans le cas de la régression. La régression à vecteurs de support, *support vector regression* (*SVR*) en anglais, introduite par Vapnik et al. (1997) et Smola (1998), possède ainsi une propriété de représentation creuse des données. Toutefois, son intérêt principal est sans doute qu'il s'agit d'une méthode de régression robuste (au sens de la section 3.6.3) pour laquelle il existe des mises en œuvre efficaces. Les méthodes à vecteurs de support sont très bien comprises aujourd'hui et ont été traitées de manière extensive dans la littérature (voir par exemple (Schölkopf et Smola, 2002) et les nombreux articles consacrés au sujet). Les paragraphes suivants sont consacrés à une présentation plus détaillée de la méthode.

Formulation de la *SVR*

Soit $\mathcal{F}$ un espace à noyau reproduisant k. Une *SVR* construit une fonction $\hat{f}$ sous la forme

$$\begin{array}{rccc} f: & \mathbb{R}^d & \to & \mathbb{R} \\ & \boldsymbol{x} & \mapsto & (f_0, k(\boldsymbol{x}, .))_{\mathcal{F}} + b \end{array} \tag{3.34}$$

minimisant le critère régularisé

$$\frac{1}{2}\|f_0\|^2_{\mathcal{F}} + C\sum_{i=1}^{n}[f^{\text{obs}}_{\boldsymbol{x}_i} - f(\boldsymbol{x}_i)]_\varepsilon \,. \tag{3.35}$$

Dans ces équations, b est un terme de biais facultatif et $C \in \mathbb{R}^{+*}$ permet d'ajuster le compromis entre régularisation et attache aux données. Remarquons que $f_0 \in \mathcal{F}$ s'identifie à f en l'absence de b. L'originalité d'une *SVR* tient au choix du terme d'attache aux données, formé à partir de la fonction coût $[\cdot]_\varepsilon$ définie par

$$[s]_\varepsilon = \max(0, |s - \varepsilon|)\,,$$

appelée fonction de coût ε-*insensible* de Vapnik. La fonction de coût $[\cdot]_\varepsilon$, qu'on peut voir comme une variante d'une fonction de coût L^1 ou de la fonction de coût de Huber, offre deux avantages. D'une part, la minimisation de (3.35) s'effectue facilement d'un point de vue numérique et d'autre

part, l'approximation à vecteurs de support est robuste vis-à-vis des données aberrantes, comme nous l'avons vu plus haut.

Solution numérique

En introduisant des variables auxiliaires ξ_i, ξ_i^*, on peut reformuler (3.35) sous la forme

$$\begin{array}{ll} \text{minimiser} & \frac{1}{2}\|f_0\|_{\mathcal{F}}^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \\ \text{sous les contraintes} & \left\{ \begin{array}{lcl} f_{\boldsymbol{x}_i}^{\text{obs}} - (f_0, k(\boldsymbol{x}_i, .))_{\mathcal{F}} - b & \leq & \varepsilon + \xi_i \\ (f_0, k(\boldsymbol{x}_i, .))_{\mathcal{F}} + b - f_{\boldsymbol{x}_i}^{\text{obs}} & \leq & \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* & \geq & 0 \end{array} \right. \end{array}$$

Le lagrangien de ce problème peut s'écrire

$$\begin{aligned} L(f_0, b, \xi_i, \xi_i^*, \alpha_i, \alpha_i^*, \eta_i, \eta_i^*) \;=\; & \frac{1}{2}\|f_0\|_{\mathcal{F}}^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \\ & - \sum_{i=1}^{n}\alpha_i(\varepsilon + \xi_i - f_{\boldsymbol{x}_i}^{\text{obs}} + (f_0, k(\boldsymbol{x}_i, .))_{\mathcal{F}} + b) \\ & - \sum_{i=1}^{n}\alpha_i^*(\varepsilon + \xi_i^* + f_{\boldsymbol{x}_i}^{\text{obs}} - (f_0, k(\boldsymbol{x}_i, .))_{\mathcal{F}} - b) \\ & - \sum_{i=1}^{n}(\eta_i\xi_i + \eta_i^*\xi_i^*)\,, \end{aligned}$$

avec $\alpha_i, \alpha_i^*, \eta_i, \eta_i^* \geq 0$, $\forall i \in \{1, \ldots, n\}$. En utilisant le théorème de représentation, qui montre que $\hat{f}_0$ se met sous la forme

$$\hat{f}_0(\boldsymbol{x}) = \sum_{i=1}^{n}\hat{a}_i k(\boldsymbol{x}_i, \boldsymbol{x})\,,$$

et le fait que les dérivées partielles de L par rapport aux variables a_i, b, ξ et ξ^* s'annulent à l'optimum, on trouve que $\hat{f}$ se réécrit sous la forme

$$\hat{f}(\boldsymbol{x}) = \sum_{i=1}^{n}(\hat{\alpha}_i - \hat{\alpha}_i^*)k(\boldsymbol{x}_i, \boldsymbol{x}) + \hat{b}\,,$$

où les $\hat{\alpha}_i$ et $\hat{\alpha}_i^*$, sont solutions du problème d'optimisation quadratique

$$\begin{array}{ll} \text{maximiser} & -\varepsilon\sum_{i=1}^{n}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{n}f_{\boldsymbol{x}_i}^{\text{obs}}(\alpha_i - \alpha_i^*) - \frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)k(\boldsymbol{x}_i, \boldsymbol{x}_j) \\ \text{sous les contraintes} & \left\{ \begin{array}{l} \sum_{i=1}^{n}(\alpha_i - \alpha_i^*) = 0\,, \\ 0 \leq \alpha_i^*, \alpha_i \leq C, \quad i = 1, \ldots, n. \end{array} \right. \end{array}$$

Une caractéristique importante de la solution de ce problème d'optimisation est que les $(\hat{\alpha}_i - \hat{\alpha}_i^*)$ sont nuls pour certains i. Les données pour lesquelles les $(\hat{\alpha}_i - \hat{\alpha}_i^*)$ diffèrent de zéro sont

appelées les *vecteurs de support*, par référence au cas de la classification. Pour expliquer cette propriété plus précisément, on écrit les conditions de Karush–Kuhn–Tucker (KKT) qui indiquent que pour la solution optimale, les produits entre les variables duales et les contraintes s'annulent. Par conséquent,

$$\begin{aligned} \hat{\alpha}_i(\varepsilon + \hat{\xi}_i - f^{\text{obs}}_{\boldsymbol{x}_i} + \hat{f}(\boldsymbol{x}_i)) &= 0, \\ \hat{\alpha}^*_i(\varepsilon + \hat{\xi}_i + f^{\text{obs}}_{\boldsymbol{x}_i} - \hat{f}(\boldsymbol{x}_i)) &= 0\,. \end{aligned}$$

Seuls les indices i correspondant aux données pour lesquelles $|\hat{f}(\boldsymbol{x}_i) - f^{\text{obs}}_{\boldsymbol{x}_i}| \geq \varepsilon$ sont tels que les $\hat{\xi}_i$ ou $\hat{\xi}^*_i$ diffèrent de zéro. Les données qui sont à l'intérieur d'un tube de largeur ε autour de $\hat{f}(\boldsymbol{x})$, sont donc telles que $\hat{\alpha}_i = \hat{\alpha}^*_i = 0$.

3.6.5 *SVR* multi-sorties

Cette section montre une application de l'utilisation d'une fonction de covariance comme noyau reproduisant. Nous présentons une extension de la régression à vecteurs de support pour traiter le cas de systèmes à plusieurs sorties. L'approche traditionnelle qui consiste à prédire indépendamment chaque sortie du système est sous-optimale si les sorties sont corrélées entre elles, ou autrement dit, si les données observées sur une sortie d'un système sont susceptibles de contenir de l'information sur d'autres sorties de ce système. Dans la section 2.5.2, le passage de la prédiction linéaire dans le cas d'un seul processus aléatoire à celui de plusieurs processus a été effectué très simplement en introduisant les covariances et inter-covariances des processus. L'utilisation de ces covariances sous la forme

$$k[(\alpha, \boldsymbol{x}), (\beta, \boldsymbol{y})]\,,$$

où α et β, considérés comme des facteurs supplémentaires, servent à indicer la sortie du système, permet de se ramener formellement au cas d'un seul processus aléatoire. Si l'on choisit $k[(\alpha, \boldsymbol{x}), (\beta, \boldsymbol{y})]$ comme noyau reproduisant de $\mathcal{F}$, il devient trivial d'utiliser la régression à vecteur de support pour modéliser plusieurs sorties corrélées.

Biais identiques sur chaque sortie

S'il est possible de considérer que le terme de biais b est identique, pour chaque sortie du système, alors une *SVR* multi-sortie s'obtient en cherchant une fonction sous la forme

$$\begin{array}{rccc} f: & \{1,\ldots,q\} \times \mathbb{R}^d & \to & \mathbb{R} \\ & \alpha, \boldsymbol{x} & \mapsto & (f_0, k([\alpha, \boldsymbol{x}], [\cdot,\cdot]))_{\mathcal{F}} + b \end{array}$$

minimisant le critère

$$\frac{1}{2}\|f_0\|^2_{\mathcal{F}} + C \sum_{i,\alpha} [f^{\text{obs}}_{\alpha,\boldsymbol{x}_i} - f(\alpha, \boldsymbol{x}_i)]_\varepsilon \tag{3.36}$$

Notons que

$$f_0 = \sum_{\beta,i} a_{\beta,i} k([\beta, \boldsymbol{x}_i], [\cdot,\cdot])$$

et que le produit scalaire dans (3.36) s'écrit

$$(f_0, k([\alpha, \boldsymbol{x}], [\cdot, \cdot]))_{\mathcal{F}} = \sum_{\beta,i} a_{\beta,i} k([\beta, \boldsymbol{x}_i], [\alpha, \boldsymbol{x}]) \,.$$

En résumé, le seul changement à effectuer dans le cas d'une *SVR* multi-sortie et d'ajouter un paramètre α pour sélectionner la sortie du modèle $\hat{f}$. Cette formulation ne nécessite aucune modification des algorithmes d'optimisation quadratique utilisés habituellement pour les machines à vecteurs de support. La difficulté principale consiste à choisir un noyau $k([\alpha, \boldsymbol{x}], [\beta, \boldsymbol{y}])$, c'est-à-dire à déterminer les covariances et inter-covariances des processus aléatoires modélisant le système.

Remarquons que l'on peut prédire également très facilement le résultat de toute transformation linéaire sur les sorties. Par exemple,

$$f(1, \boldsymbol{x}) - f(2, \boldsymbol{x}) = (f_0, k([1, \boldsymbol{x}], [\cdot, \cdot]) - k([2, \boldsymbol{x}], [\cdot, \cdot]))_{\mathcal{F}} \,.$$

Biais différents sur chaque sortie

Pour traiter des moyennes inconnues et éventuellement différentes sur chaque sortie, il devient nécessaire de considérer des formulations semi-paramétriques des *SVR* (Smola et al., 1999). Nous reviendrons plus généralement sur les formulations semi-paramétriques dans le chapitre 4. Notons qu'une *SVR* qui comporte un biais b peut déjà être considéré comme une formulation semi-paramétrique. Nous n'expliquerons pas la façon dont on calcule les paramètres d'une *SVR* semi-paramétrique. L'approche de Smola et al. (1999) utilise les conditions KKT. Pour simplifier la présentation, commençons par le cas où les différentes moyennes sont supposées ne pas être corrélées entre elles. Alors, la *SVR* est étendue en considérant des termes paramétriques dans f sous la forme

$$\begin{array}{rccl} f: & \{1, \ldots, q\} \times \mathbb{R}^d & \rightarrow & \mathbb{R} \\ & \alpha, \boldsymbol{x} & \mapsto & (f_0, k([\alpha, \boldsymbol{x}], [\cdot, \cdot]))_{\mathcal{F}} + \sum_\beta b_\beta \delta_\beta(\alpha) \end{array}$$

où $\delta_\beta(\alpha)$ vaut un si $\beta = \alpha$ et zéro sinon.

Si les moyennes sont corrélées, la formulation précédente peut encore être utilisée. Dans la plupart des cas en effet, on dispose de suffisamment de données pour permettre une estimation indépendante des moyennes de chaque sortie du système. Il serait en principe possible de prendre en compte explicitement ces corrélations au prix d'une complication (peu justifiée) de la mise en œuvre.

Pour finir, il arrive fréquemment que dans un système à plusieurs sorties le bruit d'observation soit différent sur chaque sortie. Pour tenir compte de cette caractéristique, la solution consiste à envisager des termes d'attaches aux données différents pour chaque sortie dans le critère (3.36). Plus précisément, nous suggérons d'utiliser un terme d'attache aux données sous la forme

$$\sum_{i,\alpha} C_\alpha [f^{\text{obs}}_{\alpha,\boldsymbol{x}_i} - f(\alpha, \boldsymbol{x}_i)]_{\varepsilon_\alpha} \,.$$

Cette solution peut notamment être utilisée pour la prédiction robuste de dérivées par *SVR* (Vazquez et Walter, 2003b).

3.7 Conclusions

Ce chapitre a exposé la théorie des espaces hilbertiens à noyaux reproduisant en partant d'éléments classiques et en allant jusqu'à ses développements récents comme par exemple la représentation des noyaux dans des frames. Nous avons été conduit à réécrire les démonstrations de certains résultats afin de les présenter dans un cadre unifié et facilement accessible.

Ce chapitre a également effectué une synthèse sur l'équivalence entre prédiction linéaire de processus aléatoires et régression régularisée par noyaux reproduisants. En résumé, ces deux points de vue sont des interprétations différentes et complémentaires de la même méthode. Par exemple, le point de vue aléatoire permet de comprendre comment formuler le problème de régression régularisée pour l'approximation de fonctions à valeurs vectorielles pour modéliser des systèmes à plusieurs sorties corrélées (Vazquez et Walter, 2003b). Nous avons également rappelé l'intérêt de la régression à vecteurs de support dans le cadre de l'estimation robuste. Le point de vue fonctionnel nous permet de mieux comprendre le problème de la régularisation et son importance. Nous avons rappelé des résultats sur les processus aléatoires à trajectoires dans des espaces hilbertiens à noyau reproduisant. Nous espérons avoir ainsi écrit un document de travail utile pour de futurs travaux.

Au chapitre suivant, nous introduirons la notion d'espace hilbertien à noyau conditionnellement positif et sa relation avec les processus aléatoires intrinsèques.

Chapitre 4

Processus aléatoires intrinsèques

Résumé — Ce chapitre présente quelques aspects essentiels de la théorie des fonctions aléatoires intrinsèques et du krigeage intrinsèque (Matheron, 1971a, 1973). L'objectif est d'effectuer des prédictions linéaires lorsque la moyenne d'un processus aléatoire est inconnue. Les moyennes considérées sont des fonctions linéairement paramétrées. Nous verrons également comment utiliser le krigeage intrinsèque pour prendre en compte dans des modèles boîte noire certaines formes d'a priori sur les systèmes. Dans la méthode ainsi suggérée, l'erreur de prédiction est en un sens orthogonale à l'a priori que l'on souhaite introduire. Cette méthode permet donc de passer très facilement de la notion de modèle boîte noire à celle de modèle boîte grise.

4.1 Objectifs

Ce chapitre concerne la modélisation de systèmes dont la sortie tend à évoluer autour d'une valeur constante ou même, plus généralement, autour d'une fonction sur $\mathbb{X}$. Il s'agit donc de modéliser la sortie du système en distinguant deux effets. L'un constitue une tendance et modélise des variations lentes de la sortie ; l'autre modélise les variations locales, plus rapides. Cette question, qui correspond au problème du choix de modèle, est laissée à l'appréciation de l'utilisateur. Nous verrons par la suite qu'elle est aussi liée aux connaissances disponibles a priori sur le système. L'approche retenue consiste à modéliser le système par un processus aléatoire à moyenne inconnue. Cette moyenne peut par exemple être une constante ou une fonction polynomiale des facteurs.

En géostatistique, le krigeage dit *universel* est le point de vue souvent utilisé pour construire un prédicteur linéaire dans le cas d'un processus aléatoire de moyenne inconnue. Le krigeage universel obéit au principe de meilleure prédiction linéaire non biaisée, *best linear unbiased prediction* (*BLUP*) en anglais. Comme le montre la proposition suivante, le caractère non biaisé du meilleur prédicteur linéaire s'obtient en ajoutant une contrainte sur les coefficients de la combinaison linéaire qui le constitue.

Proposition 36. *Soit $F(\boldsymbol{x})$, $\boldsymbol{x} \in \mathbb{X}$, un processus aléatoire du second ordre de moyenne b et de*

covariance $k(\boldsymbol{x}, \boldsymbol{y})$. Le meilleur prédicteur linéaire non biaisé de $F(\boldsymbol{x})$ s'écrit sous la forme

$$\hat{F}(\boldsymbol{x}) = \sum_{i=1}^{n} \widehat{\lambda}_{i,\boldsymbol{x}} F(\boldsymbol{x}_i)\,,$$

où les $\lambda_{i,x}$ sont les solutions du système d'équations linéaires

$$\begin{pmatrix} \boldsymbol{K} & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} \widehat{\boldsymbol{\lambda}}_{\boldsymbol{x}} \\ \mu \end{pmatrix} = \begin{pmatrix} \boldsymbol{k}_{\boldsymbol{x}} \\ 1 \end{pmatrix}, \tag{4.1}$$

avec les notations usuelles (de la section 2.5.1) et $\mu \in \mathbb{R}$, un coefficient de Lagrange.

Démonstration. Considérons un prédicteur affine $\hat{F}(\boldsymbol{x})$ de $F(\boldsymbol{x})$ à partir des variables aléatoires $F(\boldsymbol{x}_1), \ldots, F(\boldsymbol{x}_n)$ sous la forme

$$\hat{F}(\boldsymbol{x}) = \sum_{i=1}^{n} \lambda_{i,\boldsymbol{x}} F(\boldsymbol{x}_i) + \lambda_{0,\boldsymbol{x}}\,.$$

La moyenne quadratique de l'erreur de prédiction s'écrit dans ce cas

$$\mathrm{E}[(F(\boldsymbol{x}) - \hat{F}(\boldsymbol{x}))^2] = \mathrm{Var}[F(\boldsymbol{x}) - \hat{F}(\boldsymbol{x})] + \left[b\left(1 - \sum_{i=1}^{n} \lambda_{i,\boldsymbol{x}}\right) - \lambda_{0,\boldsymbol{x}} \right]^2 \tag{4.2}$$

La moyenne b étant inconnue, la seule façon de minimiser cette erreur est d'annuler la contribution de b en imposant la condition

$$\sum_{i=1}^{n} \lambda_{i,\boldsymbol{x}} = 1\,, \tag{4.3}$$

et par suite, $\lambda_{0,\boldsymbol{x}} = 0$ puisque $\lambda_{0,\boldsymbol{x}}$ n'apparaît que dans le terme de biais. Dans ce cas, $\mathrm{E}[F(\boldsymbol{x}) - \hat{F}(\boldsymbol{x})] = 0$ et $\hat{F}(\boldsymbol{x})$ est un prédicteur non biaisé. La minimisation de (4.2) sous la contrainte (4.3) s'effectue par la méthode du lagrangien et conduit à résoudre le système (4.1). □

Remarque 1. On peut aisément généraliser le principe *BLUP* à des moyennes non constantes.

Remarque 2. La procédure qui consisterait à estimer la moyenne dans une première étape, puis à ajuster un modèle par krigeage aux résidus est à éviter. Pour illustrer cette affirmation, supposons que l'on observe sans bruit des données sur le domaine $\mathbb{X} = [0, 1]$, telles que $f_{x_i}^{\mathrm{obs}} = b_1 x_i + b_0$, $\forall i = 1, \ldots, n$, avec $b_1, b_0 \in \mathbb{R}$. Dans ce cas très simple, on constate qu'estimer une droite de régression $\hat{b}_1 x + \hat{b}_0$ par moindres carrés n'est pas du tout équivalent au fait d'estimer la moyenne des données $\check{b}_0$ puis de déterminer une droite de régression $\check{b}_1 x$ ajustée aux résidus $f_{x_i}^{\mathrm{obs}} - \check{b}_0$. Il est donc très important de ne pas traiter les termes paramétriques séparément. Plus précisément, il est possible de montrer que la méthode des résidus ne se justifie que dans le cas où les termes paramétriques sont orthogonaux.

Le krigeage *intrinsèque* étend et formalise le principe de prédiction par krigeage universel, et par la suite nous utiliserons essentiellement ce point de vue. Le krigeage intrinsèque permet de traiter le cas de processus aléatoires à moyenne inconnue pouvant être exprimée sous forme d'une fonction linéairement paramétrée (dans le cas présenté ci-dessus, le terme paramétrique est une

constante). Le point de départ de la théorie du krigeage intrinsèque est de traduire (4.3) en terme de relation d'orthogonalité. En effet, (4.3) implique la propriété

$$\mathrm{E}[(F(\boldsymbol{x}) - \hat{F}(\boldsymbol{x}))1] = (F(\boldsymbol{x}) - \hat{F}(\boldsymbol{x}), \mathbb{1}_\Omega)_{L^2(\Omega,\mathcal{A},\mathsf{P})} = 0\,.$$

Le principe fondamental est donc le suivant : on impose que l'erreur de prédiction soit minimale au sens quadratique, tout en requérant qu'elle soit orthogonale en un certain sens à l'espace engendré par les termes paramétriques constituant la moyenne du processus. La condition d'orthogonalité est justifiée par le principe de meilleure approximation, au sens où il n'est pas possible d'améliorer l'erreur de prédiction en ajoutant au prédicteur toute combinaison linéaire de termes paramétriques. Dans la section 4.6.1, nous utiliserons ce principe pour montrer que l'on peut introduire de l'*information a priori* dans un modèle boîte noire de manière élégante. Notons enfin qu'un autre intérêt de la théorie des fonctions aléatoires intrinsèques est d'introduire la notion de *covariances généralisées* et qu'il est ainsi possible d'enrichir la classe des modèles considérés.

4.2 Définitions et construction des fonctions aléatoires intrinsèques

4.2.1 Processus généralisés, covariances généralisées

On souhaite modéliser un système dont la sortie évolue autour d'une fonction inconnue $m(\boldsymbol{x})$ pouvant s'écrire sous une forme linéairement paramétrée $m(\boldsymbol{x}) = \boldsymbol{b}^\mathsf{T}\boldsymbol{r}(\boldsymbol{x})$, où $\boldsymbol{b}$ et $\boldsymbol{r}(\boldsymbol{x})$ appartiennent à $\mathbb{R}^l$. Pour ce faire, on considère un processus aléatoire du second ordre $F(\boldsymbol{x})$ de moyenne $m(\boldsymbol{x})$ inconnue. $F(\boldsymbol{x})$ admet donc la décomposition

$$F(\boldsymbol{x}) = F_0(\boldsymbol{x}) + \boldsymbol{b}^\mathsf{T}\boldsymbol{r}(\boldsymbol{x})$$

où $F_0(\boldsymbol{x})$ est un processus aléatoire du second ordre, de moyenne nulle. Une approche naturelle pour modéliser l'incertitude sur le vecteur des paramètres $\boldsymbol{b}$ serait de considérer $\boldsymbol{b}$ comme la réalisation d'un vecteur aléatoire $\boldsymbol{B}(\omega)$. Cette démarche nécessiterait toutefois d'introduire des hypothèses supplémentaires pour spécifier la loi de $\boldsymbol{B}$. Ce n'est pas la voie retenue par la suite. L'objectif poursuivi est d'éviter d'introduire des hypothèses sur $\boldsymbol{b}$ à l'aide d'un modèle du système ne faisant pas « apparaître » $m(\boldsymbol{x})$. Plus précisément, l'idée des processus aléatoires intrinsèques est de trouver des transformations simples de $F(\boldsymbol{x})$ qui *filtrent* $m(\boldsymbol{x})$. Nous verrons que pour des moyennes polynomiales ces transformations sont des accroissements, ou des accroissements généralisés, et qu'il est possible de calculer les moments du second ordre de ces accroissements à l'aide de fonctions de covariance généralisée. En quelque sorte, il s'agit d'obtenir un nouveau modèle du système. Commençons par introduire une notion de processus généralisé.

Soient l'espace vectoriel de fonctions

$$\mathcal{N} = \left\{\boldsymbol{b}^\mathsf{T}\boldsymbol{r}(\boldsymbol{x}), \boldsymbol{b} \in \mathbb{R}^l\right\}\,, \quad \dim\mathcal{N} = l$$

et le sous-espace vectoriel $\tilde{\Lambda}_{\mathcal{N}^\perp} \subset \tilde{\Lambda}$ des mesures à support fini qui s'annulent sur $\mathcal{N}$, c'est-à-dire l'espace des mesures $\lambda \in \tilde{\Lambda}$ telles que

$$\langle \lambda, f\rangle = \sum_{i=1}^{n} \lambda_i f(\boldsymbol{x}_i) = 0\,, \quad \forall f \in \mathcal{N}\,.$$

(Par la suite, nous utiliserons la notation $\langle \lambda, f \rangle = \int_{\mathbb{X}} f d\lambda$)

Définition 30. Une fonction $k : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$, symétrique, est *conditionnellement positive* par rapport à $\mathcal{N}$ si pour toute mesure $\lambda \in \tilde{\Lambda}_{\mathcal{N}^\perp}$, la valeur de la forme quadratique $k(\lambda, \lambda)$ est positive, où $k(\lambda, \mu)$, $\lambda, \mu \in \tilde{\Lambda}_{\mathcal{N}^\perp}$, est définie par

$$k(\lambda, \mu) = \sum_{i,j=1}^{n,m} \lambda_i \mu_j k(\boldsymbol{x}_i, \boldsymbol{y}_j) \,.$$

Si de plus $k(\lambda, \lambda) = 0$ implique $\lambda = 0$, pour tout $\lambda \in \tilde{\Lambda}_{\mathcal{N}^\perp}$, $k(\boldsymbol{x}, \boldsymbol{y})$ est dite *conditionnellement définie positive*.

Soit une application *linéaire* $F_{\mathrm{G}}(\lambda)$, définie sur $\tilde{\Lambda}_{\mathcal{N}^\perp}$, à valeurs dans un espace de variables aléatoires du second ordre $L^2(\Omega, \mathcal{A}, \mathsf{P})$. Pour tout λ, nous supposons également que $F_{\mathrm{G}}(\lambda)$ est de moyenne nulle et que la covariance $\mathrm{Cov}[F_{\mathrm{G}}(\lambda), F_{\mathrm{G}}(\mu)] = k(\lambda, \mu)$ est déterminée par une fonction conditionnellement définie positive $k(\boldsymbol{x}, \boldsymbol{y})$. Ainsi, l'application $F_{\mathrm{G}}(\lambda)$ est un *processus aléatoire généralisé*, non plus défini sur l'espace des facteurs $\mathbb{X}$, mais sur un espace de mesures. La fonction $k(\boldsymbol{x}, \boldsymbol{y})$ s'appelle dans ce cas une *covariance généralisée*. Soit $\tilde{\mathcal{H}}_{\mathcal{N}^\perp}$ le sous-espace vectoriel de $L^2(\Omega, \mathcal{A}, \mathsf{P})$ généré par les éléments $F_{\mathrm{G}}(\lambda)$, $\lambda \in \tilde{\Lambda}_{\mathcal{N}^\perp}$. Les éléments de $\tilde{\mathcal{H}}_{\mathcal{N}^\perp}$ sont de moyenne nulle. Par conséquent, le produit scalaire $L^2(\Omega, \mathcal{A}, \mathsf{P})$ sur $\tilde{\mathcal{H}}_{\mathcal{N}^\perp}$ s'exprime en fonction de $k(\boldsymbol{x}, \boldsymbol{y})$ sous la forme

$$(F_{\mathrm{G}}(\lambda), F_{\mathrm{G}}(\mu))_{\tilde{\mathcal{H}}_{\mathcal{N}^\perp}} = k(\lambda, \mu) = \sum_{i,j} \lambda_i \mu_j k(\boldsymbol{x}_i, \boldsymbol{y}_j) \,,$$

où $\lambda = \sum_i \lambda_i \delta_{\boldsymbol{x}_i}$ et $\mu = \sum_j \mu_j \delta_{\boldsymbol{y}_j}$ appartiennent à $\tilde{\Lambda}_{\mathcal{N}^\perp}$. Remarquons que la covariance généralisée $k(\boldsymbol{x}, \boldsymbol{y})$ sert à calculer le produit scalaire, comme la fonction de covariance dans le cas des processus à moyenne nulle. Munissons $\tilde{\Lambda}_{\mathcal{N}^\perp}$ du produit scalaire $k(\lambda, \mu)$ et considérons les complétés $\mathcal{H}_{\mathcal{N}^\perp}$ et $\Lambda_{\mathcal{N}^\perp}$ de $\tilde{\mathcal{H}}_{\mathcal{N}^\perp}$ et $\tilde{\Lambda}_{\mathcal{N}^\perp}$ respectivement. Les espaces ainsi construits sont isomorphes. $F_{\mathrm{G}}(\lambda)$ se prolonge par continuité sur $\Lambda_{\mathcal{N}^\perp}$. Par la suite, le processus aléatoire généralisé $F_{\mathrm{G}}(\lambda)$ est utilisé comme modèle du système. Des hypothèses simplificatrices sont introduites dans la section suivante.

Remarque. Si, pour tout $\lambda \in \Lambda_{\mathcal{N}^\perp}$, les variables aléatoires $F_{\mathrm{G}}(\lambda)$ sont gaussiennes, toute combinaison linéaire de telles variables aléatoires est gaussienne. Dans ce cas, $F_{\mathrm{G}}(\lambda)$ est un processus aléatoire généralisé gaussien.

4.2.2 Fonctions aléatoires intrinsèques

L'objectif suivant est d'introduire des propriétés de stationnarité sur les processus aléatoires généralisés pour obtenir la notion de *fonctions* (ou processus) *aléatoires intrinsèques*. Les fonctions aléatoires intrinsèques, caractérisées essentiellement par une covariance généralisée stationnaire, constituent des modèles de systèmes simples. L'hypothèse de stationnarité permet de plus l'inférence des paramètres de la covariance à partir des données (voir le chapitre 5). Soit $\tau_{\boldsymbol{h}} : \tilde{\Lambda}_{\mathcal{N}^\perp} \to \tilde{\Lambda}$, l'opérateur linéaire de translation tel que pour $\lambda = \sum_i \lambda_i \delta_{\boldsymbol{x}_i} \in \tilde{\Lambda}_{\mathcal{N}^\perp}$, $\tau_{\boldsymbol{h}} \lambda = \sum_i \lambda_i \delta_{\boldsymbol{x}_i + \boldsymbol{h}}$. $\tilde{\Lambda}_{\mathcal{N}^\perp}$ est supposé stable par translation. Cette hypothèse implique que $\mathcal{N}$ doit lui-même être un espace stable par translation. Par la suite, nous supposerons en outre que la covariance généralisée $k(\boldsymbol{x}, \boldsymbol{y})$ est invariante par translations. Alors $\tau_{\boldsymbol{h}}$ est continu et se prolonge de manière unique sur $\Lambda_{\mathcal{N}^\perp}$.

Définition 31. Soit $F_{\mathrm{G}}(\lambda)$, un processus aléatoire généralisé défini sur $\Lambda_{\mathcal{N}^\perp}$, de moyenne nulle et de covariance généralisée $k(\boldsymbol{x},\boldsymbol{y})$. Si $k(\boldsymbol{x},\boldsymbol{y})$ est invariante par translation, le processus aléatoire $\boldsymbol{h}\mapsto F(\tau_{\boldsymbol{h}}\lambda)$, $\lambda\in\Lambda_{\mathcal{N}^\perp}$, est stationnaire au second ordre. Dans ce cas, on dit que $F_{\mathrm{G}}(\lambda)$, $\lambda\in\Lambda_{\mathcal{N}^\perp}$, est un *processus aléatoire intrinsèque*, ou *intrinsic random function* (*IRF*) en anglais.

$\mathcal{N}$ est de dimension *finie*. La condition de stabilité par le groupe des translations implique alors que $\mathcal{N}$ est un espace vectoriel de fonctions de type *exponentielle–polynôme* (Matheron, 1971a). Un tel espace est généré par une famille de fonctions de la forme $\boldsymbol{x}^{\boldsymbol{l}}e^{(\boldsymbol{a},\boldsymbol{x})}$, où $\boldsymbol{a}$ est réel ou complexe, $\boldsymbol{l}$ représente le multi-indice $(l_1,\ldots,l_d)$ et $\boldsymbol{x}^{\boldsymbol{l}}=x_{[1]}^{l_1}\cdots x_{[d]}^{l_d}$. (Pour un multi-indice $\boldsymbol{l}$, on notera également $|\boldsymbol{l}|=l_1+\cdots+l_q$.) Afin de garantir la stabilité des fonctions de $\mathcal{N}$ par combinaison linaire et par translation, les monômes $\boldsymbol{x}^{\boldsymbol{l}}$ doivent former une base polynomiale complète. Dans la théorie classique des processus aléatoires intrinsèques, on se restreint au cas où $\mathcal{N}$ est un espace vectoriel de polynômes de degré inférieur ou égal à l. On prend donc $\mathcal{N}_l=\mathrm{vect}\{\boldsymbol{x}^{\boldsymbol{l}},\ \forall\boldsymbol{l}\text{ tels que }|\boldsymbol{l}|\le l\}$. Posons alors $\tilde{\Lambda}_l=\tilde{\Lambda}_{\mathcal{N}_l^\perp}$ et notons Λ_l le complété de $\tilde{\Lambda}_l$ pour le produit scalaire $k(\lambda,\mu)$. Nous dirons qu'une *IRF* $F_{\mathrm{G}}(\lambda)$ définie sur Λ_l est une *IRF* d'ordre l, ou simplement une *IRF*(l).

Proposition 37. *Une fonction aléatoire intrinsèque d'ordre l est encore une fonction aléatoire intrinsèque d'ordre $l+1$.*

Démonstration. Les espaces Λ_l sont emboîtés :

$$\Lambda_{l+1}\subset\Lambda_l\,.$$

Toute fonction aléatoire intrinsèque $F_{\mathrm{G}}(\lambda)$ sur Λ_l sera également stationnaire sur Λ_{l+1}. □

Une remarque importante est que les mesures de Λ_l sont des opérateurs de différences finies (accroissements ou accroissements généralisés). Pour $l=0$, la condition pour qu'une mesure $\lambda=\sum_{i=1}^n\lambda_i\delta_{\boldsymbol{x}_i}$ soit dans Λ_0 est que $\sum_{i=1}^n\lambda_i=0$. Donc $\lambda=\sum_{i=1}^n\lambda_i(\delta_{\boldsymbol{x}_i}-\delta_{\boldsymbol{x}_1})$ et par conséquent, λ est une combinaison linéaire de mesures d'accroissements, ou d'opérateurs de différences finies du type $\delta_{\boldsymbol{x}_i}-\delta_{\boldsymbol{x}_1}$. Pour $l>0$, on obtient naturellement des accroissements généralisés, correspondant à des opérateurs de différences finies d'ordre supérieur. Comme en dimension d l'espace vectoriel des polynômes de degré inférieur à l est de dimension $\binom{d+l}{l}$, le nombre de points nécessaires pour spécifier une mesure de Λ_l grandit rapidement avec la dimension de l'espace des facteurs.

4.2.3 Représentants des *IRF*

Une fonction aléatoire généralisée peut être vue comme une classe d'équivalence de processus aléatoires à moyenne dans $\mathcal{N}$. Si $F(\boldsymbol{x})$, $\boldsymbol{x}\in\mathbb{X}$, est un processus aléatoire du second ordre, à moyenne dans $\mathcal{N}$ et de fonction de covariance $k(\boldsymbol{x},\boldsymbol{y})$, il peut être étendu sur $\tilde{\Lambda}_{\mathcal{N}^\perp}$ en considérant l'application linéaire

$$\begin{array}{rccc} F: & \tilde{\Lambda}_{\mathcal{N}^\perp} & \to & \mathcal{H} \\ & \lambda=\sum_{i=1}^n\lambda_i\delta_{\boldsymbol{x}_i} & \mapsto & F(\lambda)=\sum_{i=1}^n\lambda_iF(\boldsymbol{x}_i)\,, \end{array}$$

où $\mathcal{H}$ est l'espace hilbertien généré par $F(\boldsymbol{x})$, $\boldsymbol{x}\in\mathbb{X}$. Comme $F(\boldsymbol{x})$ est à moyenne dans $\mathcal{N}$, $F(\lambda)$, $\lambda\in\Lambda_{\mathcal{N}^\perp}$, est de moyenne nulle puisque λ filtre, ou annule, toute fonction de $\mathcal{N}$. Si $k(\boldsymbol{x},\boldsymbol{y})$ est définie positive, alors la forme bilinéaire $(\lambda,\mu)_{\tilde{\Lambda}_{\mathcal{N}^\perp}}=(F(\lambda),F(\mu))_{\mathcal{H}}$ est un produit scalaire sur $\tilde{\Lambda}_{\mathcal{N}^\perp}$ et

F se prolonge par continuité sur le complété $\Lambda_{\mathcal{N}^\perp}$ de $\tilde{\Lambda}_{\mathcal{N}^\perp}$. Par conséquent, $F : \Lambda_{\mathcal{N}^\perp} \to \mathcal{H}$ est un processus aléatoire généralisé (de covariance généralisée $k(\boldsymbol{x}, \boldsymbol{y})$). Réciproquement, étant donné un processus aléatoire généralisé $F_{\mathrm{G}}(\lambda)$, quels sont les processus aléatoires $F(\boldsymbol{x})$ qui coïncident avec $F_{\mathrm{G}}(\lambda)$ sur $\Lambda_{\mathcal{N}^\perp}$?

Définition 32. Soit $F_{\mathrm{G}}(\lambda)$ une fonction aléatoire généralisée définie sur $\Lambda_{\mathcal{N}^\perp}$. Un processus aléatoire du second ordre $F(\boldsymbol{x})$, $\boldsymbol{x} \in \mathbb{X}$, est un *représentant* de $F_{\mathrm{G}}(\lambda)$ si

$$F_{\mathrm{G}}(\lambda) = F(\lambda), \quad \forall \lambda \in \Lambda_{\mathcal{N}^\perp} .$$

Par la suite, nous nous intéressons à la nature des représentants des *IRF*(l). De chaque *IRF*(l) $F_{\mathrm{G}}(\lambda)$, il est possible de construire un représentant. En effet, il existe un ensemble $S_q = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_q\}$ de points de $\mathbb{X}$ tel que toute fonction f de $\mathcal{N}_l$ est déterminée de manière unique par ses valeurs aux points de cet ensemble. S_q est dit *unisolvant*. Soit $\{p_{\boldsymbol{z}_1}, \ldots, p_{\boldsymbol{z}_q}\}$ une base de polynômes de Lagrange de $\mathcal{N}_l$, tels que $\forall f \in \mathcal{N}_l$

$$f(\boldsymbol{x}) = \sum_{i=1}^{q} p_{\boldsymbol{z}_i}(\boldsymbol{x}) f(\boldsymbol{z}_i) .$$

Soit $\delta_{(\boldsymbol{x})}$ la mesure à support fini définie par

$$\delta_{(\boldsymbol{x})} = \delta_{\boldsymbol{x}} - \sum_{i=1}^{q} p_{\boldsymbol{z}_i}(\boldsymbol{x}) \delta_{\boldsymbol{z}_i} .$$

Notons que $\delta_{(\boldsymbol{z}_i)} = 0$, $i = 1, \ldots, q$. La mesure $\delta_{(\boldsymbol{x})}$, paramétrée par $\boldsymbol{x}$, est telle que $\delta_{(\boldsymbol{x})} \in \Lambda_{\mathcal{N}^\perp}$, $\forall \boldsymbol{x} \in \mathbb{X}$.

Proposition 38.

$$F_0(\boldsymbol{x}) = F_{\mathrm{G}}(\delta_{(\boldsymbol{x})}) \tag{4.4}$$

est un représentant de $F_{\mathrm{G}}(\lambda)$, de moyenne nulle, et $F_0(\boldsymbol{z}_i) = 0$ presque sûrement, $\forall i \in \{1, \ldots, q\}$.

Démonstration. Soit l'application linéaire

$$\begin{array}{rccc} \vartheta : & \tilde{\Lambda}_{\mathcal{N}^\perp} & \to & \Lambda_{\mathcal{N}^\perp} \\ & \lambda = \sum_{i=1}^{n} \lambda_i \delta_{\boldsymbol{x}_i} & \mapsto & \sum_{i=1}^{n} \lambda_i \delta_{(\boldsymbol{x}_i)} . \end{array}$$

$\forall \lambda \in \tilde{\Lambda}_{\mathcal{N}^\perp}$,

$$\begin{aligned} \vartheta(\lambda) &= \sum_{i=1}^{n} \lambda_i \delta_{\boldsymbol{x}_i} - \sum_{i=1}^{n} \lambda_i \sum_{j=1}^{q} p_{\boldsymbol{z}_j}(\boldsymbol{x}_i) \delta_{\boldsymbol{z}_j} \\ &= \lambda - \sum_{j=1}^{q} \langle \lambda, p_{\boldsymbol{z}_j} \rangle \delta_{\boldsymbol{z}_j} = \lambda . \end{aligned}$$

Par conséquent, $\forall \lambda \in \tilde{\Lambda}_{\mathcal{N}^\perp}$,

$$F_0(\lambda) = \sum_{i=1}^{n} \lambda_i F_0(\boldsymbol{x}_i) = \sum_{i=1}^{n} \lambda_i F_{\mathrm{G}}(\delta_{(\boldsymbol{x}_i)}) = F_{\mathrm{G}}(\vartheta(\lambda)) = F_{\mathrm{G}}(\lambda) . \tag{4.5}$$

Comme ϑ est continue sur $\tilde{\Lambda}_{\mathcal{N}^\perp}$, elle se prolonge de manière unique sur $\Lambda_{\mathcal{N}^\perp}$, et les relations (4.5) se prolongent par continuité sur $\Lambda_{\mathcal{N}^\perp}$. □

Proposition 39. *Si $F_0(\boldsymbol{x})$ est le représentant défini par (4.4), on obtient d'autres représentants de $F_{\mathrm{G}}(\lambda)$ sous la forme*

$$F(\boldsymbol{x}) = F_0(\boldsymbol{x}) + \sum_{i=1}^{q} B_i p_{\boldsymbol{z}_i}(\boldsymbol{x})\,, \tag{4.6}$$

où les B_i sont des variables aléatoires du second ordre quelconques (indépendantes ou non).

Démonstration. Sans difficulté. □

Dans le cadre de la modélisation boîte noire, la notion d'*IRF* est utilisée comme modèle du système au même titre qu'un processus aléatoire du second ordre. Les trajectoires ou réalisations possibles d'une *IRF* correspondent aux trajectoires de l'ensemble des représentants.

Remarque. Dans le cas particulier où les variables aléatoires B_i sont constantes presque sûrement, le processus obtenu est déterministe aux points de l'ensemble unisolvant S_q. On peut interpréter ce cas comme lorsque l'on fixe des conditions initiales dans une équation différentielle. Dans la section suivante, nous montrons par exemple que le mouvement brownien est un représentant d'une *IRF*. Dans ce cas, le processus est nul presque sûrement à l'origine.

Mentionnons enfin le lien entre les processus aléatoires intrinsèques et les modèles *auto-regressive integrated moving average* (*ARIMA*) de séries chronologiques (Brockwell et Davis, 1987).

Définition 33. Une suite de variables aléatoires $(X_t)_{t\in\mathbb{N}}$ est un processus *ARIMA*(p,d,q) (avec p, d et q entiers positifs) si le processus $(Y_t)_{t\in\mathbb{N}}$ défini par

$$Y_t = (1-B)^d X_t\,, \tag{4.7}$$

où $B : X_t \mapsto X_{t-1}$ est l'opérateur retard, est un processus *ARMA*(p,q) causal.

La transformation $(1-B)^d$ est un opérateur de différences finies d'ordre d. Une propriété importante est que si on ajoute à (X_t) un polynôme en t quelconque de degré inférieur ou égal à $d-1$ on ne change pas l'équation (4.7). Un processus *ARIMA*(p,d,q) peut donc être vu comme une *IRF*$(d-1)$ à temps discret. Le processus (X_t) est stationnaire si et seulement si $d=0$. Le modèle *ARIMA* est en fait bien adapté aux séries chronologiques non stationnaires qui comportent une tendance de type polynomiale.

4.3 Covariances généralisées

4.3.1 Caractéristiques générales

Cette section rappelle très brièvement les principaux résultats sur les fonctions symétriques, conditionnellement positives lorsque $\mathcal{N}$ est un espace de polynômes de degré $\leq l$. Il s'agit donc de s'intéresser aux covariances généralisées $k(\boldsymbol{h})$, $\boldsymbol{h}\in\mathbb{X}$, d'ordre l (Matheron, 1973), telles que si $F_{\mathrm{G}}(\lambda)$ est une *IRF*(l), on a pour tout $\lambda,\mu\in\Lambda_l$

$$\mathrm{E}[F_{\mathrm{G}}(\lambda)F_{\mathrm{G}}(\mu)] = \iint k(\boldsymbol{x}-\boldsymbol{y})d\lambda(\boldsymbol{x})d\mu(\boldsymbol{y})\,.$$

Classes d'équivalences de covariance généralisées

Toute fonction de covariance est une covariance généralisée. La proposition suivante montre que les covariances généralisées forment des classes d'équivalences.

Proposition 40. *Si $k(\boldsymbol{h})$ est une covariance généralisée d'une IRF(l) $F_{\mathrm{G}}(\lambda)$, toute fonction de la forme $k(\boldsymbol{h}) + p(\boldsymbol{h})$, où $p(\boldsymbol{h})$ est un polynôme pair de degré $\leq 2l$, est également une covariance généralisée de $F_{\mathrm{G}}(\lambda)$. Réciproquement, toute covariance généralisée de $F_{\mathrm{G}}(\lambda)$ est de la forme $k(\boldsymbol{h}) + p(\boldsymbol{h})$.*

Cette proposition est importante pour le choix des covariances généralisées en pratique. Nous donnons quelques éléments de démonstration que nous pensons utiles pour comprendre l'origine du résultat. La démonstration complète (Matheron, 1971a) est relativement longue et d'un intérêt limité dans le contexte de cette présentation.

Lemme 2. *Pour qu'une fonction $k(\boldsymbol{x}, \boldsymbol{y})$ continue et symétrique sur $\mathbb{R}^d \times \mathbb{R}^d$ vérifie la relation*

$$\iint k(\boldsymbol{x}, \boldsymbol{y}) d\lambda(\boldsymbol{x}) d\lambda(\boldsymbol{y}) = 0\,, \quad \forall \lambda \in \Lambda_l$$

il faut et il suffit que $k(\boldsymbol{x}, \boldsymbol{y})$ soit de la forme :

$$k(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{q} a_i(\boldsymbol{y}) \boldsymbol{x}^{\boldsymbol{l}_i} + \sum_{i=1}^{q} a_i(\boldsymbol{x}) \boldsymbol{y}^{\boldsymbol{l}_i} - \sum_{i,j} b_{i,j} \boldsymbol{x}^{\boldsymbol{l}_i} \boldsymbol{y}^{\boldsymbol{l}_j}\,, \tag{4.8}$$

où les $a_i(.)$, $i = 1, \ldots, q$, sont des fonctions continues sur $\mathbb{R}^d$, où les $b_{i,j}$, $i, j = 1, \ldots, q$, sont des constantes, et où $\forall i \in \{1, \ldots, q\}$, $|\boldsymbol{l}_i| \leq l$.

Démonstration. Voir (Matheron, 1971a). □

Démonstration partielle de la proposition. Toute fonction de la forme $k(\boldsymbol{h}) + k'(\boldsymbol{h})$ est une covariance généralisée de $F_{\mathrm{G}}(\lambda)$ si et seulement si, $\forall \lambda, \mu \in \Lambda_l$,

$$\iint k'(\boldsymbol{x} - \boldsymbol{y}) d\lambda(\boldsymbol{x}) d\mu(\boldsymbol{y}) = 0\,,$$

ou, ce qui revient au même, $\forall \lambda \in \Lambda_l$,

$$\iint k'(\boldsymbol{x} - \boldsymbol{y}) d\lambda(\boldsymbol{x}) d\lambda(\boldsymbol{y}) = 0\,. \tag{4.9}$$

Il faut identifier parmi les fonctions continues symétriques $k(\boldsymbol{x}, \boldsymbol{y})$ s'écrivant sous la forme (4.8), celles qui se mettent sous la forme $k'(\boldsymbol{x} - \boldsymbol{y})$. Remarquons que l'on doit avoir $k'(-\boldsymbol{h}) = k'(\boldsymbol{h})$. Soient $k'(\boldsymbol{h})$ vérifiant (4.9) et $k(\boldsymbol{x}, \boldsymbol{y})$ de la forme (4.8), telle que $k(\boldsymbol{x}, \boldsymbol{y}) = k' \circ \Delta\,(\boldsymbol{x}, \boldsymbol{y})$, où $\Delta : (\boldsymbol{x}, \boldsymbol{y}) \mapsto \boldsymbol{x} - \boldsymbol{y}$. On suppose k dérivable $l + 1$ fois par rapport à chacune de ses $2d$ variables. D'après (4.8), si D est un opérateur de dérivation d'ordre $l + 1$ par rapport à l'une des variables de $\boldsymbol{x}$ et de $\boldsymbol{y}$, $Dk = 0$. Ceci signifie que les dérivées d'ordre $2l + 2$ par rapport à chacune des d variables de la fonction $k'(\boldsymbol{h})$ sont nulles. Par conséquent, $k'(\boldsymbol{h})$ est un polynôme pair de d variables de degré au plus $2l$. Réciproquement, tout polynôme $p(\boldsymbol{h})$ pair, de degré $\leq 2l$ est tel que $p(\boldsymbol{x} - \boldsymbol{y})$ se met sous la forme

$$\sum_{i,j \text{ t.q } |\boldsymbol{l}_i| + |\boldsymbol{l}_j| \leq 2l} c_{i,j} \boldsymbol{x}^{\boldsymbol{l}_i} \boldsymbol{y}^{\boldsymbol{l}_j}\,.$$

$p(\boldsymbol{h})$ vérifie (4.9) puisque qu'il y a toujours un facteur dans la somme de degré $\leq l$. Nous ne démontrons pas le résultat dans le cas où k n'est pas dérivable (Matheron, 1971a, page 24). □

Lien entre covariance généralisées et covariance des représentants

On souhaite établir la forme analytique des covariances des représentants d'une *IRF*(l) $F_{\mathrm{G}}(\lambda)$. Soit le représentant de moyenne nulle $F_0(\boldsymbol{x}) = F(\delta_{(\boldsymbol{x})})$. Alors

$$\begin{aligned}\mathrm{Cov}[F_0(\boldsymbol{x}), F_0(\boldsymbol{y})] &= k(\boldsymbol{x},\boldsymbol{y}) - \sum_{i=1}^{q} p_{\boldsymbol{z}_i}(\boldsymbol{x})k(\boldsymbol{z}_i,\boldsymbol{x}) - \sum_{i=1}^{q} p_{\boldsymbol{z}_i}(\boldsymbol{y})k(\boldsymbol{z}_i,\boldsymbol{y}) \\ &\quad + \sum_{i,j=1}^{q} p_{\boldsymbol{z}_i}(\boldsymbol{x})p_{\boldsymbol{z}_j}(\boldsymbol{y})k(\boldsymbol{z}_i,\boldsymbol{z}_j)\end{aligned}$$

Pour les représentants sous la forme (4.6), la covariance s'écrit

$$\begin{aligned}\mathrm{Cov}[F(\boldsymbol{x}), F(\boldsymbol{y})] &= \mathrm{Cov}[F_0(\boldsymbol{x}), F_0(\boldsymbol{y})] + \sum_{i=1}^{q} p_{\boldsymbol{z}_i}(\boldsymbol{x})\,\mathrm{Cov}[B_i, F_0(\boldsymbol{y})] \\ &\quad + \sum_{i=1}^{q} p_{\boldsymbol{z}_i}(\boldsymbol{y})\,\mathrm{Cov}[B_i, F_0(\boldsymbol{x})] + \sum_{i,j=1}^{q} p_{\boldsymbol{z}_i}(\boldsymbol{x})p_{\boldsymbol{z}_j}(\boldsymbol{y})\,\mathrm{Cov}[B_i, B_j]\,.\end{aligned}$$

Plus généralement, on a la proposition suivante.

Proposition 41. *Si $F_{\mathrm{G}}(\lambda)$ est une IRF(l) de covariance généralisée $k(\boldsymbol{x},\boldsymbol{y})$, les covariances $k_r(\boldsymbol{x},\boldsymbol{y})$ des représentants de cette IRF s'expriment sous la forme*

$$k_r(\boldsymbol{x},\boldsymbol{y}) = k(\boldsymbol{x},\boldsymbol{y}) + \sum_{i=1}^{q} a_i(\boldsymbol{y})\boldsymbol{x}^{\boldsymbol{l}_i} + \sum_{i=1}^{q} a_i(\boldsymbol{x})\boldsymbol{y}^{\boldsymbol{l}_i} + \sum_{i,j=1}^{q} b_{i,j}\boldsymbol{x}^{\boldsymbol{l}_i}\boldsymbol{y}^{\boldsymbol{l}_j}\,, \tag{4.10}$$

où les $a_i(.)$ sont des fonctions continues, où les $b_{i,j}$ sont des coefficients réels tels que la matrice $\boldsymbol{B} = (b_{i,j})_{i,j}$ soit symétrique définie positive, et où $\forall i \in \{1,\ldots,q\}$, $|\boldsymbol{l}_i| \leq l$.

Démonstration. Soit $F(\boldsymbol{x})$ un représentant de $F_{\mathrm{G}}(\lambda)$ sous la forme (4.6), et soit k_r sa fonction de covariance. Remarquons que pour tout $\lambda \in \Lambda_l$, $\mathrm{Var}[F(\lambda)] = \mathrm{Var}[F_G(\lambda)]$, donc $k_r(\boldsymbol{x},\boldsymbol{y}) - k(\boldsymbol{x},\boldsymbol{y})$ vérifie nécessairement les conditions du lemme 2. La covariance de $F(\boldsymbol{x})$ s'écrit

$$\begin{aligned}k_r(\boldsymbol{x},\boldsymbol{y}) = \mathrm{Cov}[F(\boldsymbol{x}), F(\boldsymbol{y})] &= \mathrm{E}[F_{\mathrm{G}}(\delta_{(\boldsymbol{x})})F_{\mathrm{G}}(\delta_{(\boldsymbol{y})})] \\ &\quad + \sum_{i=1}^{q} \mathrm{Cov}[B_i, F_{\mathrm{G}}(\delta_{(\boldsymbol{y})})]\boldsymbol{x}^{\boldsymbol{l}_i} + \sum_{i=1}^{q} \mathrm{Cov}[B_i, F_{\mathrm{G}}(\delta_{(\boldsymbol{x})})]\boldsymbol{y}^{\boldsymbol{l}_i} \\ &\quad + \sum_{i,j=1}^{q} \mathrm{Cov}[B_i, B_j]\boldsymbol{x}^{\boldsymbol{l}_i}\boldsymbol{y}^{\boldsymbol{l}_j}\,.\end{aligned}$$

Donc k_r est bien de la forme (4.10). □

Comportements des covariances généralisées

Les covariances classiques continues et stationnaires $k(\boldsymbol{h})$ sont des fonctions bornées (d'après l'inégalité de Schwarz) et tendent généralement vers 0 quand $\|\boldsymbol{h}\|$ tend vers l'infini (lemme de Riemann-Lebesgue). Pour une covariance généralisée d'ordre l, on a un comportement du type

$$\lim_{\|\boldsymbol{h}\|\to+\infty} \frac{k(\boldsymbol{h})}{\|h\|^{2l+2}} = 0\,.$$

Par exemple, la fonction

$$k(x,y) = |x-y|^3\,.$$

est une covariance généralisée d'ordre 1 associée à $\mathcal{N}_1 = \text{vect}\{1, x\}$.

Le résultat précédent peut être établi à partir de la théorie spectrale des covariances généralisées (voir (Matheron, 1973)) mais aussi à partir de la formule (4.10) liant une covariance généralisée aux covariances des représentants.

4.3.2 *IRF*(0), variogrammes

Ce paragraphe présente le cas important des covariances généralisées définies à partir d'un variogramme. Plus spécifiquement, on s'intéresse au cas élémentaire où $\mathcal{N}$ est constitué des *fonctions constantes* sur $\mathbb{X}$. Ce cas correspond aux *IRF*(0). Toutes les représentations d'une *IRF*(0) $F_{\text{G}}(\lambda)$ sont par conséquent de la forme $F_0(\boldsymbol{x}) + b$, où $F_0(\boldsymbol{x})$ est un processus aléatoire à moyenne nulle, nul presque sûrement en un point $z_1 \in \mathbb{X}$, le choix de l'origine z_1 étant arbitraire. Les fonctions aléatoires intrinsèques d'ordre 0 permettent donc de traiter des processus aléatoires dont la moyenne est inconnue mais constante. Les fonctions aléatoires d'ordres supérieurs permettent de traiter des processus dont la moyenne n'est pas constante. De tels modèles sont pertinents lorsque l'on souhaite modéliser un système qui présente des tendances à croître ou à décroître par exemple.

Définition 34. Soit $F(\boldsymbol{x})$ un processus aléatoire du second ordre, stationnaire, de moyenne non nécessairement nulle, son *variogramme* $\gamma(\boldsymbol{h})$, défini à partir de la variance des accroissements de $F(\boldsymbol{x})$, est donné par

$$\gamma(\boldsymbol{h}) = \frac{1}{2}\,\text{Var}[F(\boldsymbol{x}) - F(\boldsymbol{x}+\boldsymbol{h})]\,.$$

Le variogramme est une fonction paire. Il vérifie la relation $\gamma(\boldsymbol{h}) = k(\boldsymbol{0}) - k(\boldsymbol{h})$, où $k(\boldsymbol{h})$ est la covariance de $F(\boldsymbol{x})$. Un variogramme est toujours nul à l'origine. Si $\lambda \in \Lambda_0$, la variance de $F(\lambda)$ est donnée par

$$\text{Var}\left[\sum_{i=1}^{n} \lambda_i F(\boldsymbol{x}_i)\right] = -\sum_{i,j} \lambda_i \lambda_j \gamma(\boldsymbol{x}_i - \boldsymbol{x}_j)\,. \tag{4.11}$$

Par conséquent, $-\gamma(\boldsymbol{h})$ est une covariance généralisée d'ordre 0, c'est-à-dire, une fonction conditionnellement positive par rapport à $\mathcal{N}_0$. On peut donc utiliser $-\gamma(\boldsymbol{h})$ pour caractériser les moments du second ordre d'une *IRF*(0) $F_{\text{G}}(\lambda)$. On voit une fois encore que les *IRF*(0) sont des modèles portant sur des accroissements de la sortie du système.

Lorsque le processus $F(\boldsymbol{x})$ est stationnaire, son variogramme est borné mais la définition d'une *IRF* impose la stationnarité des accroissements et non celle des représentants. Il est en fait possible d'utiliser un variogramme pour caractériser un représentant non stationnaire. Dans ce cas, la

croissance du variogramme est majorée par la relation $|\gamma(\boldsymbol{h})| = o(\|\boldsymbol{h}\|^2)$. (À partir de l'analyse du comportement de variogrammes empiriques calculés à partir de données expérimentales, cette propriété peut servir à tester le caractère stationnaire d'un système.)

Notons enfin que dans le cas des représentants stationnaires admettant une covariance $k(\boldsymbol{x}, \boldsymbol{y})$, la relation

$$k(\boldsymbol{x}, \boldsymbol{y}) = -\gamma(\boldsymbol{x} - \boldsymbol{y}) + \frac{1}{2}(k(\boldsymbol{x}, \boldsymbol{x}) + k(\boldsymbol{y}, \boldsymbol{y}))$$

est un cas particulier de la formule (4.10), avec $q = 1$, $a_1(\boldsymbol{x}) = k(\boldsymbol{x}, \boldsymbol{x})$, $\boldsymbol{l}_1 = 0$, $b_{1,1} = 0$.

Exemple. Le mouvement brownien est un exemple fondamental de représentant d'une *IRF*(0). Soit $F_{\mathrm{G}}(\lambda)$ une *IRF*(0) gaussienne de covariance généralisée

$$k(x, y) = -\gamma(x - y) = -\frac{1}{2}|x - y|\,, \quad x, y \in \mathbb{R}^+\,.$$

Considérons le représentant $W(x) = F_{\mathrm{G}}(\delta_x - \delta_0)$. C'est un processus aléatoire de moyenne nulle, nul presque sûrement en 0. Pour $x, y \in \mathbb{R}^+$, sa covariance est définie par

$$k_W(x, y) = \mathrm{E}[W(x)W(y)] = \mathrm{E}[F_{\mathrm{G}}(\delta_x - \delta_0)F_{\mathrm{G}}(\delta_y - \delta_0)] = \frac{1}{2}(-|x - y| + |x| + |y|)\,.$$

On vérifie que $k_W(x, y) = \min(x, y)$. Le représentant ainsi construit est un mouvement brownien tel que défini dans la section 2.1.3. Par ailleurs, des intégrations successives du mouvement brownien partant de 0 définissent des processus aléatoires

$$W_k(x) = \int_0^x \frac{(x - t)^{k-1}}{(k - 1)!} W(t)dt\,,$$

et l'on peut montrer que les W_k sont des représentants d'*IRF*(k) (Matheron, 1973).

4.4 Espaces hilbertiens à noyau conditionnellement positif

Dans cette section nous montrons, en utilisant la démarche utilisée par (Schaback, 1999), que l'on peut généraliser la notion d'espace hilbertien à noyau reproduisant au cas où l'on utilise des noyaux symétriques conditionnellement positifs. Cette généralisation nous permettra d'interpréter le krigeage intrinsèque qui sera vu dans la section 4.5 en terme de régression régularisée avec des semi-normes.

Soit $k(\boldsymbol{x}, \boldsymbol{y})$ une fonction symétrique conditionnellement définie positive par rapport à l'espace vectoriel $\mathcal{N}_l$ des polynômes de degré au plus l. Rappelons que le produit scalaire sur Λ_l est défini par

$$(\lambda, \mu)_{\Lambda_l} = k(\lambda, \mu) = \sum_{i,j=1}^{n,m} \lambda_i \mu_j k(\boldsymbol{x}_i, \boldsymbol{y}_j)\,.$$

Λ_l est un espace hilbertien de mesures sur $\mathbb{X}$ mais il n'est pas possible pour le moment de définir par dualité des fonctions sur $\mathbb{X}$, parce que la forme linéaire d'évaluation ponctuelle $\delta_{\boldsymbol{x}}$ n'appartient pas à Λ_l. Une solution consiste à construire, pour tout $\boldsymbol{x} \in \mathbb{X}$, un substitut de la forme linéaire $\delta_{\boldsymbol{x}}$. Comme précédemment, considérons un ensemble unisolvant $S_q = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_q\}$ de points de $\mathbb{X}$

tel que toute fonction f de $\mathcal{N}_l$ soit déterminée de manière unique par ses valeurs aux points de cet ensemble. Soit $\{p_{\boldsymbol{z}_1}, \ldots, p_{\boldsymbol{z}_q}\}$ une base de $\mathcal{N}_l$ de polynômes de Lagrange. Alors, $\forall f \in \mathcal{N}_l$

$$f(\boldsymbol{x}) = \sum_{i=1}^{q} p_{\boldsymbol{z}_i}(\boldsymbol{x}) f(\boldsymbol{z}_i)\,.$$

La mesure à support fini $\delta_{(\boldsymbol{x})} \in \Lambda_l$ définie par

$$\delta_{(\boldsymbol{x})} = \delta_{\boldsymbol{x}} - \sum_{i=1}^{q} p_{\boldsymbol{z}_i}(\boldsymbol{x}) \delta_{\boldsymbol{z}_i}\,, \tag{4.12}$$

est telle que $\delta_{(\boldsymbol{z}_i)} = 0$, $i = 1, \ldots, q$, et annule toute composante polynomiale de degré inférieur à l. La suite de ce paragraphe est consacrée à expliciter la nature de l'espace de fonctions généré par $(\delta_{(\boldsymbol{x})}, \lambda)_{\Lambda_l}$, lorsque λ parcourt Λ_l. Soit $\tilde{\mathcal{F}}_{S_q}$ l'espace vectoriel vect$\{f = (\delta_{(.)}, \lambda)_{\Lambda_l},\ \lambda \in \tilde{\Lambda}_l\}$. Notons que les éléments de $\tilde{\mathcal{F}}_{S_q}$ s'annulent sur S_q.

Si $\lambda = \sum_{i=1}^{n} \lambda_i \delta_{\boldsymbol{x}_i} \in \tilde{\Lambda}_l$, alors

$$\begin{aligned}(\delta_{(\boldsymbol{x})}, \lambda)_{\Lambda_l} &= \sum_{i=1} \lambda_i \left(k(\boldsymbol{x}, \boldsymbol{x}_i) - \sum_{j=1}^{q} p_{\boldsymbol{z}_j}(\boldsymbol{x}) k(\boldsymbol{z}_j, \boldsymbol{x}_i) \right) \\ &= \langle \lambda, k(\boldsymbol{x}, .)\rangle - \sum_{j=1}^{q} p_{\boldsymbol{z}_j}(\boldsymbol{x}) \langle \lambda, k(\boldsymbol{z}_j, .)\rangle\,. \end{aligned} \tag{4.13}$$

Par suite, si μ est un autre élément de $\tilde{\Lambda}_l$, on obtient la relation

$$\langle \mu, (\delta_{(.)}, \lambda)\rangle = (\mu, \lambda)_{\Lambda_l}\,. \tag{4.14}$$

Ceci montre que l'application $\lambda \mapsto (\delta_{(.)}, \lambda)_{\Lambda_l}$ est injective sur $\tilde{\Lambda}_l$. Par conséquent, l'application bilinéaire

$$(f, g)_{\tilde{\mathcal{F}}_{S_q}} = (\lambda, \mu)_{\Lambda_l}\,,$$

où $f(\boldsymbol{x}) = (\delta_{(\boldsymbol{x})}, \lambda)_{\Lambda_l}$ et $g(\boldsymbol{x}) = (\delta_{(\boldsymbol{x})}, \mu)_{\Lambda_l}$, pour tout $\boldsymbol{x} \in \mathbb{X}$, définit un produit scalaire sur $\tilde{\mathcal{F}}_{S_q}$. Le complété de $\tilde{\mathcal{F}}_{S_q}$, noté $\mathcal{F}_{S_q}$, s'identifie à Λ_l et nous noterons ϱ l'isométrie bijective $\mathcal{F}_{S_q} \to \Lambda_l$. La relation (4.14) se prolonge par continuité sur les complétés et montre par conséquent que $\langle ., . \rangle$ met en dualité Λ_l et $\mathcal{F}_{S_q}$.

Proposition 42. *Notons $k_{S_q}(\boldsymbol{x}, \boldsymbol{y})$, la fonction $\mathbb{X} \times \mathbb{X} \to \mathbb{R}$ telle que $k_{S_q}(\boldsymbol{x}, \boldsymbol{y}) = (\delta_{(\boldsymbol{x})}, \delta_{(\boldsymbol{y})})_{\Lambda_l}$. Alors k_{S_q} est symétrique et de type positif.*

Démonstration. $k_{S_q}(\boldsymbol{x}, \boldsymbol{y})$ est bien symétrique (par symétrie du produit scalaire) et de type positif car pour toutes suites finies $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \in \mathbb{X}^n$ et $(\lambda_1, \ldots, \lambda_n) \in \mathbb{R}^n$,

$$\sum_{i,j=1}^{n} \lambda_i \lambda_j k_{S_q}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sum_{i,j=1}^{n} \lambda_i \lambda_j (\delta_{(\boldsymbol{x}_i)}, \delta_{(\boldsymbol{x}_j)})_{\Lambda_l} = \left\| \sum_{i=1}^{n} \lambda_i \delta_{(\boldsymbol{x}_i)} \right\|_{\Lambda_l}^2 \geq 0\,.$$

□

Proposition 43. *$k_{\mathcal{F}_{S_q}}$ est le noyau reproduisant de $\mathcal{F}_{S_q}$.*

Démonstration. $\forall f \in \mathcal{F}_{S_q}$, $\exists \lambda \in \Lambda_l$ telle que $f = (\delta_{(\boldsymbol{x})}, \lambda)_{\Lambda_l}$ et donc,

$$(k_{\mathcal{F}_{S_q}}(\boldsymbol{x}, .), f)_{\mathcal{F}_{S_q}} = (\delta_{(\boldsymbol{x})}, \lambda)_{\Lambda_l} = f(\boldsymbol{x}) .$$

□

L'espace $\mathcal{F}_{S_q}$ possède la propriété assez faible que ses éléments s'annulent sur S_q. L'objectif final est d'établir, dans le cas des fonctions conditionnellement positives, la construction d'un espace de fonctions indépendant du choix de S_q, avec des propriétés similaires à celles des espaces hilbertiens à noyau reproduisant. Commençons par remarquer que $\mathcal{F}_{S_q}$ et $\mathcal{N}_l$ forment une somme directe d'espaces vectoriels de fonctions. En effet, si $f \in \mathcal{F}_{S_q} \cap \mathcal{N}_l$, $f(\boldsymbol{x}) = \sum_{i=1}^{q} f(\boldsymbol{z}_q) p_{\boldsymbol{z}_q}(\boldsymbol{x})$ et $\exists \lambda \in \Lambda_l$ tel que $f = (\delta_{(.)}, \lambda)_{\Lambda_l}$. Pour tout $\mu \in \Lambda_l$, $\langle \mu, f \rangle = (\mu, \lambda)_{\Lambda_l} = 0$ et donc, $f = 0$. L'opérateur linéaire $P_{\mathcal{N}_l, S_q} : \mathcal{F}_{S_q} \oplus \mathcal{N}_l \to \mathcal{N}_l$ défini par

$$P_{\mathcal{N}_l, S_q} f(\boldsymbol{x}) = \sum_{i=1}^{q} p_{\boldsymbol{z}_i}(\boldsymbol{x}) f(\boldsymbol{z}_i) , \quad \forall f \in \mathcal{F} , \boldsymbol{x} \in \mathbb{X}$$

est une projection sur $\mathcal{N}_l$. On a également le fait que $f - P_{\mathcal{N}_l, S_q} f$ s'annule aux points de S_q et que $Id - P_{\mathcal{N}_l, S_q}$ est une projection sur $\mathcal{F}_{S_q}$. Pour toute fonction $f \in \mathcal{F}_{S_q} \oplus \mathcal{N}_l$, il existe $\lambda \in \Lambda_l$ telle que $f - P_{\mathcal{N}_l, S_q} f = (\delta_{(.)}, \lambda)_{\Lambda_l}$.

Proposition 44. *Soit $f \in \mathcal{F}_{S_q} \oplus \mathcal{N}_l$. Alors $\forall \boldsymbol{x} \in \mathbb{X}$,*

$$f(\boldsymbol{x}) = (P_{\mathcal{N}_l, S_q} f)(\boldsymbol{x}) + (k_{S_q}(\boldsymbol{x}, .), f - P_{\mathcal{N}_l, S_q} f)_{\mathcal{F}_{S_q}} . \tag{4.15}$$

Démonstration. D'après les remarques précédentes. □

L'équation (4.15) peut être vue comme une généralisation de la propriété de reproduction (3.1). Notons que son utilisation requiert le choix préliminaire d'un ensemble unisolvant S_q et d'un noyau $k_{\mathcal{F}_{S_q}}$ correspondant.

Proposition 45. *Soient S_q^1 et S_q^2 deux ensembles tels que toute fonction de $\mathcal{N}_l$ soit déterminée de manière unique par ses valeurs aux points de S_q^1 ou de S_q^2. Alors $\mathcal{F}_{S_q^1} \oplus \mathcal{N}_l = \mathcal{F}_{S_q^2} \oplus \mathcal{N}_l$.*

Démonstration. Notons $\delta^1_{(\boldsymbol{x})}$ et $\delta^2_{(\boldsymbol{x})}$ les substituts de la forme d'évaluation ponctuelle (4.12) pour S_q^1 et S_q^2 respectivement. D'après (4.13), $\forall \lambda \in \Lambda_l$, il existe un polynôme $p_{1,2,\lambda}(\boldsymbol{x}) \in \mathcal{N}_l$, tel que,

$$(\delta^1_{(\boldsymbol{x})}, \lambda)_{\Lambda_l} = (\delta^2_{(\boldsymbol{x})}, \lambda)_{\Lambda_l} + p_{1,2,\lambda}(\boldsymbol{x}) .$$

Soit $f \in \mathcal{F}_{S_q^1} \oplus \mathcal{N}_l$. Il existe $\lambda \in \Lambda_l$ telle que $(\delta^1_{(\boldsymbol{x})}, \lambda)_{\Lambda_l} = (f - P_{\mathcal{N}_l, S_q^1} f)(\boldsymbol{x})$. Donc, $\forall \boldsymbol{x} \in \mathbb{X}$,

$$\begin{aligned} f(\boldsymbol{x}) &= (P_{\mathcal{N}_l, S_q^1} f)(\boldsymbol{x}) + (\delta^1_{(\boldsymbol{x})}, \lambda) \\ &= (P_{\mathcal{N}_l, S_q^1} f)(\boldsymbol{x}) + p_{1,2,\lambda}(\boldsymbol{x}) + (\delta^2_{(\boldsymbol{x})}, \lambda)_{\Lambda_l} , \end{aligned}$$

ce qui montre que $f \in \mathcal{F}_{S_q^2} \oplus \mathcal{N}_l$. □

Définition 35. Un *espace* $\mathcal{F}$ de fonctions sur $\mathbb{X}$ *à noyau* $k(\boldsymbol{x}, \boldsymbol{y})$ *symétrique conditionnellement défini positif par rapport à* $\mathcal{N}_l$ est la somme de $\mathcal{N}_l$ et d'un espace $\mathcal{F}_{S_q}$. $\mathcal{F}$ est indépendant du choix de S_q. Par la suite, une tel espace sera doté du semi-produit scalaire défini par

$$(f, g)_{\mathcal{F}} = (f - P_{\mathcal{N}_l, S_q} f, g - P_{\mathcal{N}_l, S_q} g)_{\mathcal{F}_{S_q}} .$$

Ce semi-produit scalaire est indépendant du choix de S_q. L'espace $\mathcal{F}$ à noyau k est semi-hilbertien.

Exemples. Dans la section 3.2.2, nous avions mentionné un exemple à la base de la théorie des splines construit en utilisant la fonction de covariance du mouvement brownien comme noyau reproduisant. Dans la plupart des cas, les splines sont construites à partir de noyaux conditionnellement positifs. Donnons quelques exemples. Les splines cubiques 1-D (polynômes de degré 3 par morceaux), sont très couramment utilisées. Elles correspondent au choix de l'espace des fonctions définies sur $[0,1]$, une fois continûment dérivable, et de dérivée seconde de carré sommable, muni de la semi-norme

$$\|f\|_{\mathcal{F}}^2 = \int_0^1 |f^{(2)}(t)|^2 dt\,.$$

Dans ce cas, $\mathcal{N}$ est constitué des fonctions polynomiales de degré au plus un. Le noyau conditionnellement positif associé est (Chilès et Delfiner, 1999, page 273)

$$k(t,s) = |t-s|^3\,.$$

On peut généraliser cet exemple pour obtenir des splines « plaques minces ». En deux dimensions, on choisit la semi-norme (Wahba, 1990)

$$\|f\|_{\mathcal{F}}^2 = \iint dx_{[1]} dx_{[2]} \left(\left(\frac{\partial^2 f}{\partial x_{[1]}^2} \right)^2 + 2 \left(\frac{\partial^2 f}{\partial x_{[1]} \partial x_{[2]}} \right)^2 + \left(\frac{\partial^2 f}{\partial x_{[2]}^2} \right)^2 \right)\,,$$

Le noyau associé est (Chilès et Delfiner, 1999, page 273)

$$k(\boldsymbol{x},\boldsymbol{y}) = \|\boldsymbol{x}-\boldsymbol{y}\|^2 \log\|\boldsymbol{x}-\boldsymbol{y}\|\,,$$

et l'espace $\mathcal{N}$ correspondant est l'espace des polynômes de deux variables de degré ≤ 1, c'est-à-dire $\mathcal{N} = \text{vect}\{1, x_{[1]}, x_{[2]}\}$.

4.5 Prédiction linéaire avec des processus aléatoires intrinsèques

Cette section présente le *krigeage intrinsèque*, c'est-à-dire la prédiction linéaire dans le cas de processus aléatoires intrinsèques. Considérons un système modélisé par une *IRF*(l) $F_{\mathrm{G}}(\lambda)$, dont on observe la sortie du système pour un nombre fini de valeurs $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ du vecteur des facteurs. Ces observations sont modélisées par les réalisations des variables aléatoires $F(\boldsymbol{x}_1), \ldots, F(\boldsymbol{x}_n)$, où $F(\boldsymbol{x})$ est un représentant inconnu de $F_{\mathrm{G}}(\lambda)$. Pour prédire la sortie du système correspondant au vecteur des facteurs $\boldsymbol{x}$, on approxime $F(\boldsymbol{x})$ par le prédicteur $\hat{F}(\boldsymbol{x})$ obtenu comme combinaison linéaire des $F(\boldsymbol{x}_i)$,

$$\hat{F}(\boldsymbol{x}) = \sum_{i=1}^{n} \widehat{\lambda}_{i,\boldsymbol{x}} F(\boldsymbol{x}_i)\,.$$

Notons que $\hat{F}(\boldsymbol{x})$ ne possède pas explicitement de terme de $\mathcal{N}_l$. Comme dans les cas précédents de prédiction linéaire, on cherche à minimiser la variance de l'erreur de prédiction $F(\boldsymbol{x}) - \sum_{i=1}^n \widehat{\lambda}_{i,\boldsymbol{x}} F(\boldsymbol{x}_i) = F(\delta_x - \sum_{i=1}^n \widehat{\lambda}_{i,\boldsymbol{x}} \delta_{\boldsymbol{x}_i})$. Cependant, le représentant $F(\boldsymbol{x})$ est inconnu et il n'est

pas possible de calculer directement la variance de l'erreur sauf si $\delta_x - \sum_{i=1}^n \widehat{\lambda}_{i,\boldsymbol{x}} \delta_{\boldsymbol{x}_i} \in \Lambda_l$, auquel cas

$$\begin{aligned}\mathrm{Var}[F(\delta_x - \sum_{i=1}^n \widehat{\lambda}_{i,\boldsymbol{x}} \delta_{\boldsymbol{x}_i})] &= \mathrm{Var}[F_{\mathrm{G}}(\delta_x - \sum_{i=1}^n \widehat{\lambda}_{i,\boldsymbol{x}} \delta_{\boldsymbol{x}_i})] \\ &= k(\boldsymbol{x},\boldsymbol{x}) - 2\sum_{i=1}^n \widehat{\lambda}_{i,\boldsymbol{x}} k(\boldsymbol{x}_i,\boldsymbol{x}) + \sum_{i,j=1}^n \widehat{\lambda}_{i,\boldsymbol{x}} \widehat{\lambda}_{j,\boldsymbol{x}} k(\boldsymbol{x}_i,\boldsymbol{x}_j)\,.\end{aligned}$$

En minimisant la variance de l'erreur, on obtient alors la méthode du krigeage intrinsèque, qui se résume à la résolution d'un système d'équations linéaires, comme le montre la proposition suivante.

Proposition 46 (krigeage intrinsèque). *Soient $F_{\mathrm{G}}(\lambda)$ une IRF(l) admettant une covariance généralisée $k(\boldsymbol{x},\boldsymbol{y})$ et $F(\boldsymbol{x})$ un représentant quelconque. La meilleure approximation linéaire de $F(\boldsymbol{x})$ à partir des variables aléatoires $F(\boldsymbol{x}_i)$, $i = 1,\ldots,n$, est la variable aléatoire*

$$\hat{F}(\boldsymbol{x}) = \sum_{i=1}^n \widehat{\lambda}_{i,\boldsymbol{x}} F(\boldsymbol{x}_i)\,,$$

où les $\widehat{\lambda}_{i,\boldsymbol{x}}$ sont tels que $\mathrm{Var}[F(\delta_{\boldsymbol{x}} - \sum_{i=1}^n \widehat{\lambda}_{i,\boldsymbol{x}} \delta_{\boldsymbol{x}_i})]$ soit minimale sous la contrainte $\delta_{\boldsymbol{x}} - \sum_{i=1}^n \widehat{\lambda}_{i,\boldsymbol{x}} \delta_{\boldsymbol{x}_i} \in \Lambda_l$. Les $\widehat{\lambda}_{i,\boldsymbol{x}}$ s'obtiennent en résolvant le système linéaire

$$\begin{pmatrix} \boldsymbol{K} & \boldsymbol{P}^{\mathsf{T}} \\ \boldsymbol{P} & \boldsymbol{0} \end{pmatrix} \begin{pmatrix} \widehat{\boldsymbol{\lambda}}_{\boldsymbol{x}} \\ \boldsymbol{\mu} \end{pmatrix} = \begin{pmatrix} \boldsymbol{k}_{\boldsymbol{x}} \\ \boldsymbol{p}_{\boldsymbol{x}} \end{pmatrix}, \tag{4.16}$$

où $\boldsymbol{K} = (k(\boldsymbol{x}_i,\boldsymbol{x}_j))_{i,j=1}^n$ est la matrice de covariance généralisée, $\boldsymbol{k}_{\boldsymbol{x}} = [k(\boldsymbol{x}_1,\boldsymbol{x}),\ldots,k(\boldsymbol{x}_n,\boldsymbol{x})]^{\mathsf{T}}$, $\widehat{\boldsymbol{\lambda}}_{\boldsymbol{x}} = [\widehat{\lambda}_{1,\boldsymbol{x}},\ldots,\widehat{\lambda}_{n,\boldsymbol{x}}]^{\mathsf{T}}$, $\boldsymbol{\mu} \in \mathbb{R}^q$ est un vecteur de coefficients de Lagrange, $\boldsymbol{P} = (\boldsymbol{x}_j^{\boldsymbol{l}_i})_{i,j=1}^{q,n}$ est la matrice $q \times n$ de fonctions de base de $\mathcal{N}_l$ évaluées sur S et $\boldsymbol{p}_{\boldsymbol{x}} = [\boldsymbol{x}^{\boldsymbol{l}_1},\ldots,\boldsymbol{x}^{\boldsymbol{l}_q}]^{\mathsf{T}}$.

Démonstration. Le système linéaire (4.16) est obtenu par la méthode du lagrangien. La figure 4.1 illustre géométriquement le principe du krigeage intrinsèque. □

La mise en œuvre pratique du krigeage intrinsèque est donc aussi simple que celle du krigeage classique. Le krigeage intrinsèque pourra être utilisé dans de nombreuses situations : moyenne inconnue, hypothèse de stationnarité mise en défaut (voir aussi la section 4.3.2), incorporation d'information a priori (voir la section 4.6.1). Rappelons que toute fonction de covariance peut être utilisée comme covariance généralisée. En particulier, lorsqu'on modélise le système par un processus aléatoire à moyenne inconnue, de covariance $k(\boldsymbol{x},\boldsymbol{y})$, le krigeage intrinsèque correspond exactement au cas de la meilleure prédiction non biaisée vue dans la section 4.1. Le krigeage intrinsèque est donc une généralisation des méthodes précédentes, avec notamment la possibilité d'utiliser des covariances généralisées, comme celles dérivant d'un variogramme qui spécifie la variance des accroissements du système.

Proposition 47. *La variance d'erreur du krigeage intrinsèque est donnée par*

$$\mathrm{Var}[F(\boldsymbol{x}) - \hat{F}(\boldsymbol{x})] = k(\boldsymbol{x},\boldsymbol{x}) - \sum_{i=1}^n \widehat{\lambda}_{i,\boldsymbol{x}} k(\boldsymbol{x}_i,\boldsymbol{x}) - \sum_{i=1}^q \mu_i \boldsymbol{x}^{\boldsymbol{l}_i}\,. \tag{4.17}$$

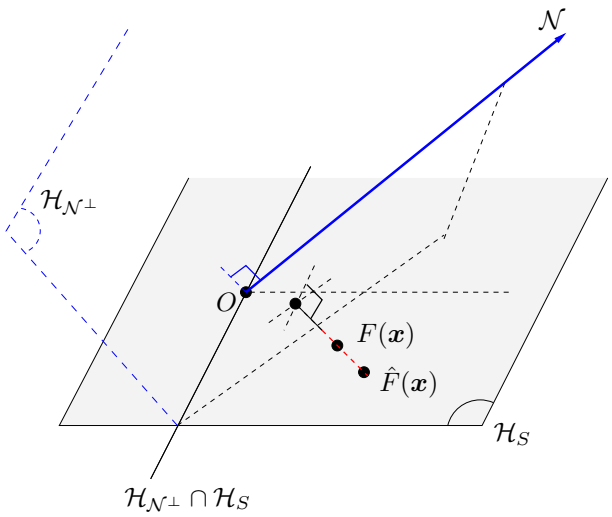


FIG. 4.1 – Meilleure approximation sous contrainte $\hat{F}(\boldsymbol{x}) - F(\boldsymbol{x}) \perp \mathcal{N}$

Démonstration.

$$\begin{aligned}
\mathrm{Var}[F(\boldsymbol{x}) - \hat{F}(\boldsymbol{x})] &= k(\boldsymbol{x},\boldsymbol{x}) - 2\widehat{\boldsymbol{\lambda}}_{\boldsymbol{x}}^\mathsf{T}\boldsymbol{k}_{\boldsymbol{x}} + \widehat{\boldsymbol{\lambda}}_{\boldsymbol{x}}^\mathsf{T}\boldsymbol{K}\widehat{\boldsymbol{\lambda}}_{\boldsymbol{x}} \\
&= k(\boldsymbol{x},\boldsymbol{x}) - 2\widehat{\boldsymbol{\lambda}}_{\boldsymbol{x}}^\mathsf{T}\boldsymbol{k}_{\boldsymbol{x}} + \widehat{\boldsymbol{\lambda}}_{\boldsymbol{x}}^\mathsf{T}(\boldsymbol{k}_{\boldsymbol{x}} - \boldsymbol{P}^\mathsf{T}\boldsymbol{\mu}) \\
&= k(\boldsymbol{x},\boldsymbol{x}) - \widehat{\boldsymbol{\lambda}}_{\boldsymbol{x}}^\mathsf{T}\boldsymbol{k}_{\boldsymbol{x}} - \boldsymbol{p}_{\boldsymbol{x}}^\mathsf{T}\boldsymbol{\mu}
\end{aligned}$$

□

4.6 Utilisation des processus aléatoires intrinsèques

4.6.1 Liens avec la régression régularisée et prise en compte d'information a priori

L'un des intérêts du krigeage intrinsèque est de nous donner la possibilité de prendre en compte de l'information a priori. Pour expliquer ce point, nous nous plaçons dans un premier temps du point de vue des méthodes de régression semi-régularisée. Dans le chapitre 3, il avait été rappelé qu'un problème de régression régularisée classique est du type

$$\text{minimiser } \|f\|_{\mathcal{F}}^2 + \frac{1}{C}\sum_{i=1}^{n}(f(\boldsymbol{x}_i) - f_{\boldsymbol{x}_i}^{\text{obs}})^2\,, \tag{4.18}$$

où la solution $\hat{f}$ est cherchée dans un espace hilbertien à noyau reproduisant $\mathcal{F}$. Pour modéliser une moyenne ou des tendances polynomiales, une idée naturelle est de rechercher la solution du problème régularisé dans un espace de fonctions $\mathcal{F}$ à noyau conditionnellement positif, qui est donc la somme d'un espace d'un espace à noyau reproduisant $\mathcal{F}_{S_q}$ et d'un espace de dimension finie $\mathcal{N}$. Toute fonction $f \in \mathcal{F}$ possède ainsi une composante dans $\mathcal{N}$. Cette composante dans $\mathcal{N}$ est une combinaison linéaire (d'éléments d'une base de fonctions) qui peut être vue comme une fonction

linéairement paramétrée. En ce sens, cette généralisation de la régression régularisée est parfois qualifiée de méthode semi-paramétrique. Ce principe est bien connu dans la théorie des splines (Wahba, 1990) et dans la théorie des fonctions de base radiales (Schaback, 1999). Il a aussi été suggéré pour la *SVR* (Smola et al., 1999) et s'apparente en géostatistique au krigeage avec *dérive externe* (Chilès et Delfiner, 1999).

Si l'on cherche la solution dans un espace à noyau conditionnellement positif, le terme de régularisation devient une semi-norme selon la définition 35 ($\mathcal{N}$ étant le noyau de cette semi-norme). Par conséquent, la composante paramétrique d'une fonction de $\mathcal{F}$ n'est pas régularisée dans le problème d'approximation (4.18). (Smola et al., 1999) remarque que cette propriété peut être mise à profit pour incorporer un a priori dans le modèle ; cet a priori doit correspondre au fait que la sortie du système est proche de certaines fonctions sur $\mathbb{X}$. Par conséquent, si les connaissances sur le système permettent de dire que son modèle est susceptible d'être amélioré en ajoutant certains termes paramétriques, l'utilisateur peut choisir de ne pas régulariser ces termes, et forcer ainsi le comportement du modèle. Toutefois, parce que les termes paramétriques ne sont pas régularisés dans cette méthode, on notera qu'il n'est pas possible de contrôler leur contribution. Le problème peut alors devenir mal posé : l'augmentation du nombre de degrés de liberté dans le modèle peut réduire considérablement la qualité de l'approximation.

En raison des liens entre la théorie des processus aléatoires intrinsèques et celle des espaces de fonctions à noyau conditionnellement positif, le krigeage intrinsèque peut aussi être qualifié de méthode semi-paramétrique. (Matheron, 1981) montre précisément le lien entre krigeage intrinsèque et les méthodes de régression semi-régularisée. En substance, ce lien est une équivalence du type de celle qui a été montrée dans la section 3.5.5 et s'établit de la même façon à partir de considérations géométriques. Les propriétés du krigeage intrinsèque, illustrées par la figure 4.1, montrent d'ailleurs que l'erreur de prédiction est orthogonale en un certain sens aux fonctions de $\mathcal{N}$. Ceci signifie que le prédicteur obtenu ne peut pas être amélioré en ajoutant une fonction de $\mathcal{N}$. En ce sens, si $\mathcal{N}$ correspond à des fonctions choisies d'après des connaissances a priori, on peut dire de manière informelle que l'erreur d'approximation est orthogonale à l'a priori.

Dans la théorie des processus aléatoires intrinsèques, les termes paramétriques forment une famille de fonctions linéairement indépendantes et l'espace engendré par ces fonctions de base doit être stable par translation. On a vu qu'un tel espace était constitué de fonctions polynômes–exponentielles (en pratique, les termes paramétriques sont des monômes $\boldsymbol{x}^l$). Ceci semble être une limite à l'intérêt pratique d'utiliser le krigeage intrinsèque pour prendre en compte des a priori parce que les fonctions que l'on souhaite incorporer dans le modèle peuvent être de natures très diverses[1].

Afin de contourner cette restriction, nous suggérons d'ajouter formellement des facteurs au modèle. Nous distinguons alors deux types de facteurs. Les facteurs du premier type sont les entrées naturelles du modèle du système. Les facteurs du second type sont introduits pour incorporer de l'information a priori. Ces facteurs sont des fonctions déterministes quelconques des facteurs du premier type. Il devient alors très facile d'utiliser des fonctions beaucoup plus générales que des

[1] Dans la littérature (voir par exemple (Sacks et al., 1989)), il est fréquent de lire que l'utilisation de termes paramétriques polynomiaux n'offre que peu d'intérêt en pratique. Pour soutenir ce point de vue, on peut se fonder sur (Matheron, 1973) qui établit que tout représentant d'une *IRF* est localement équivalent (sur un domaine borné de $\mathbb{X}$) à un processus stationnaire. Ceci suggère donc qu'il y a peu à gagner à utiliser le krigeage intrinsèque. Comme discuté plus haut, nous pensons au contraire que cette méthode est justifiée pour prendre en compte des a priori.

polynômes et d'étendre ainsi la portée de la théorie du krigeage intrinsèque. Enfin, ce point de vue permet aussi de mieux comprendre le rôle des facteurs dans un modèle : toute information susceptible de servir à prédire la sortie du système peut être utilisée comme facteur. Des exemples d'application seront traités au chapitre 6.

4.6.2 *IRF et dérivation*

Dans cette section, nous proposons de développer quelques propriétés sur la dérivation des *IRF*. L'un de nos objectifs est d'approximer les dérivées ou le gradient de la sortie d'un système, lorsque cette sortie est modélisée par une *IRF*. Pour simplifier la présentation, nous ne considérons qu'un seul facteur $x \in \mathbb{R}$. L'extension au cas multidimensionnel ne présente pas de difficulté particulière. Rappelons qu'un processus aléatoire $F(x)$ du second ordre, stationnaire, de moyenne nulle et de fonction de covariance $k(h)$ est dit dérivable en moyenne quadratique au point x si

$$F_h(x) = \frac{1}{h}(F(x+h) - F(x)) \tag{4.19}$$

converge en moyenne quadratique quand h tend vers zéro. La limite existe si et seulement si $k^{(2)}(0)$ existe et dans ce cas, $F(x)$ est dérivable pour tout x. Cette limite définit le processus dérivé noté $F^{(1)}(x)$. Les dérivées d'ordre supérieur s'obtiennent en itérant. De plus, il est immédiat de vérifier que

$$\mathrm{Cov}[F^{(q)}(x), F^{(r)}(y)] = (-1)^{(r)} k^{(q+r)}(x-y)\,. \tag{4.20}$$

Soit $\delta_x^{(1)}$ la mesure définie par la limite dans Λ de $(\delta_x - \delta_{x-h})/h$ quand h tend vers zéro, et $\delta_x^{(r)}$ les mesures obtenues en itérant la formule de dérivation. En utilisant l'extension linéaire continue de $F(x)$ sur Λ, on a $F^{(r)}(x) = F(\delta_x^{(r)})$, lorsqu'il est licite de dériver.

Notons que $\delta_x^{(r)} \in \Lambda_{r-1}$. En effet,

$$\langle \delta_{x_0}^{(r)}, x^l \rangle = \begin{cases} \frac{l!}{(l-r)!} x_0^{l-r} \text{ si } r \leq l\,, \\ 0 \text{ si } r > l\,. \end{cases}$$

Soit $F_{\mathrm{G}}(\lambda)$ une *IRF*(l), caractérisée par sa covariance généralisée $k(h)$. Nous souhaitons établir une notion de dérivabilité pour $F_{\mathrm{G}}(\lambda)$. Le point délicat est que $F_{\mathrm{G}}(\lambda)$ ou ses éventuelles dérivées ne peuvent pas être évaluées en un point x et qu'il n'est donc pas possible de définir la notion de dérivabilité en moyenne quadratique à partir de la variance d'une quantité comme (4.19).

Pour généraliser la notion de dérivée, il est donc nécessaire de recourir aux mesures de Λ_l. Considérons $\lambda = \sum_i \lambda_i \delta_{x_i} \in \Lambda_l$. Puisque $\tau_h \lambda \in \Lambda_l$, $\forall h \in \mathbb{R}$, notons

$$\lambda_h = \frac{1}{h}(\tau_h \lambda - \lambda) \in \Lambda_l.$$

Définition 36. Une IRF(l) $F_{\mathrm{G}}(\lambda)$ est dite *dérivable en moyenne quadratique* en $\lambda \in \Lambda_l$ si $F_{\mathrm{G}}(\lambda_h)$ converge en moyenne quadratique quand h tend vers zéro. Lorsque la limite existe, elle est notée $F_{\mathrm{G}}^{(1)}(\lambda)$.

Proposition 48. *Soit $F_{\mathrm{G}}(\lambda)$ une IRF(l). Si $k^{(2)}(h)$ existe pour tout h, $F_{\mathrm{G}}(\lambda)$ est dérivable en moyenne quadratique pour tout λ appartenant à Λ_l, auquel cas $F_{\mathrm{G}}^{(1)}(\lambda)$ est une IRF(l) et sa fonction de covariance généralisée est $-k^{(2)}(h)$.*

Démonstration. Si $\lambda = \sum_{i=1}^{n} \lambda_i \delta_{x_i}$ appartient à Λ_l,

$$\begin{aligned} \|F_{\mathrm{G}}(\lambda_h)\|^2 &= \frac{1}{h^2}\left\| F_{\mathrm{G}}\Big(\sum_{i=1}^{n} \lambda_i (\delta_{x_i+h} - \delta_{x_i}) \Big) \right\|^2 \\ &= \frac{1}{h^2} \sum_{i,j=1}^{n} \lambda_i \lambda_j (2k(x_i - x_j) - k(x_i - x_j + h) - k(x_i - x_j - h)) \,. \end{aligned}$$

Si $k(h)$ est deux fois dérivable pour tout $h \in \mathbb{R}$, $\|F_{\mathrm{G}}(\lambda_h)\|$ converge vers une limite finie quand h tend vers zéro et

$$\lim_{h\to 0} \|F_{\mathrm{G}}(\lambda_h)\|^2 = -\sum_{i,j=1}^{n} \lambda_i \lambda_j k^{(2)}(x_i - x_j) \,.$$

Par suite, $F_{\mathrm{G}}^{(1)}(\lambda)$ est un processus aléatoire généralisé défini sur Λ_l à moyenne nulle et de covariance généralisée $-k^{(2)}(h)$. □

Remarquons que la convergence de $F_{\mathrm{G}}(\lambda_h)$ dans $L^2(\Omega, \mathcal{A}, \mathsf{P})$ quand h tend vers zéro équivaut à celle de λ_h dans Λ_l. Posons $\lambda^{(1)} = \lim_{h\to 0} \lambda_h$ lorsque la limite existe. Nous identifierons alors $F_{\mathrm{G}}^{(1)}(\lambda)$ et $F_{\mathrm{G}}(\lambda^{(1)})$.

Proposition 49. *Soient $F_{\mathrm{G}}(\lambda)$ une IRF(l) et $F(\boldsymbol{x})$ un représentant de cette IRF. Alors $F^{(1)}(\boldsymbol{x})$ est un représentant de $F_{\mathrm{G}}^{(1)}(\lambda)$.*

Démonstration. Pour tout $\lambda \in \Lambda_l$,

$$F_{\mathrm{G}}^{(1)}(\lambda) = \lim_{h\to 0} F_{\mathrm{G}}(\lambda_h) = \lim_{h\to 0} F(\lambda_h) = F(\lambda^{(1)}) = F^{(1)}(\lambda) \,.$$

□

Les dérivées d'ordre supérieur sont notées $F_{\mathrm{G}}^{(r)}(\lambda)$. Étant donnés $\lambda = \sum_i \lambda_i \delta_{x_i}^{(q_i)}$ et $\mu = \sum_j \mu_j \delta_{y_j}^{(r_j)}$ appartenant à Λ_l, on vérifie aisément que

$$\mathrm{Cov}[F_{\mathrm{G}}(\lambda), F_{\mathrm{G}}(\mu)] = \sum_{i,j} (-1)^{r_j} \lambda_i \mu_j k^{(q_i + r_j)}(x_i - y_j) \,.$$

La seconde partie de cette section est consacrée à la prédiction des dérivées des représentants d'une IRF(l) $F_{\mathrm{G}}(\lambda)$. Nous étudions directement le cas où les observations sont corrompues par un bruit blanc additif. Les valeurs observées de la sortie du système correspondent aux réalisations des variables aléatoires $F^{\mathrm{obs}}(x_i) = F(x_i) + N_i$, $i = 1, \ldots, n$, où $F(x)$ est un représentant inconnu de $F_{\mathrm{G}}(\lambda)$, et les N_i sont des variables aléatoires gaussiennes indépendantes, de moyenne nulle et de variance σ_N^2. Nous cherchons un estimateur linéaire $\widehat{F^{(r)}}(x)$ de $F^{(r)}(x)$, qui s'écrit donc sous la forme

$$\widehat{F^{(r)}}(x) = \sum_i \widehat{\lambda}_{i,x} F^{\mathrm{obs}}(x_i) \,.$$

Bien que le krigeage intrinsèque ne soit pas couramment employé dans la littérature pour prédire des dérivées, la généralisation à ce cas est aisée. Quand on ne s'intéresse pas aux dérivées, la

variance de l'erreur de prédiction $F(x)-\widehat{F}(x)$ est minimisée sous la contrainte $\delta_x-\sum\widehat{\lambda}_{i,x}\delta_{x_i}\in\Lambda_l$. Pour traiter le cas des dérivées, il suffit de minimiser $\mathrm{Var}[F^{(r)}(x)-\widehat{F^{(r)}}(x)]$ sous la contrainte

$$\delta_x^{(r)}-\sum_i\widehat{\lambda}_{i,x}\delta_{x_i}\in\Lambda_l\,. \tag{4.21}$$

La solution de ce problème s'obtient de manière analogue à la solution du krigeage intrinsèque en utilisant $\mathrm{Var}[F(\delta_x^{(r)}-\sum_i\widehat{\lambda}_{i,x}\delta_{x_i})]=\mathrm{Var}[F_{\mathrm{G}}(\delta_x^{(r)}-\sum_i\widehat{\lambda}_{i,x}\delta_{x_i})]$. On vérifie que les coefficients $\widehat{\lambda}_{i,x}$, $i=1,\ldots,n$ sont solutions d'un système d'équations linéaires, exprimé sous forme matricielle par :

$$\begin{pmatrix}\boldsymbol{K}+\sigma_N^2\boldsymbol{I_d} & \boldsymbol{P}^{\mathsf{T}}\\ \boldsymbol{P} & \boldsymbol{0}\end{pmatrix}\begin{pmatrix}\widehat{\boldsymbol{\lambda}}_x\\ \boldsymbol{\mu}\end{pmatrix}=\begin{pmatrix}\boldsymbol{k}_x^{(r)}\\ \boldsymbol{p}_x^{(r)}\end{pmatrix}, \tag{4.22}$$

où $\boldsymbol{K}$ est la matrice $n\times n$ des valeurs de la covariance généralisée $k(x_i-x_j)$, $\boldsymbol{P}=(x_j^i)_{i=0,\,j=1}^{l,n}$ est une matrice $(l+1)\times n$, $\boldsymbol{\mu}$ est un vecteur de coefficients de Lagrange, $\boldsymbol{k}_x^{(r)}$ est le vecteur de taille n des éléments $k^{(r)}(x-x_i)$ et enfin, $\boldsymbol{p}_x^{(r)}$ est le vecteur de taille $l+1$ des éléments $(x^i)^{(r)}$, $i=0,\ldots,l$. Notons que

$$(x^i)^{(r)}=\begin{cases}0 & \text{pour } i<r\\ \frac{i!}{(i-r)!}x^{i-r} & \text{pour } i\geq r\end{cases}$$

La variance de l'erreur de prédiction est donnée par

$$\mathrm{Var}[F^{(r)}(x)-\widehat{F^{(r)}}(x)]=-k^{(2r)}(0)-\begin{pmatrix}\widehat{\boldsymbol{\lambda}}\\ \boldsymbol{\mu}\end{pmatrix}^{\mathsf{T}}\begin{pmatrix}\boldsymbol{k}_x^{(r)}\\ \boldsymbol{p}_x^{(r)}\end{pmatrix}.$$

Comme précédemment, si le système linéaire est mal conditionné, des méthodes numériques adéquates doivent être utilisées.

4.7 Conclusions

Prédire un processus aléatoire à moyenne inconnue nécessite d'ajouter des contraintes sur l'estimateur. On demande par exemple que le prédicteur soit non-biaisé. Une façon élégante de traiter ce problème d'indétermination consiste à transformer linéairement le processus aléatoire de manière à éliminer les termes inconnus. Le processus transformé est appelé processus aléatoire généralisé. Ce dernier n'est plus paramétré par l'espace des facteurs mais par des mesures sur cette espace. Le même type de transformation est utilisé dans la méthode du maximum de vraisemblance lorsque la moyenne du processus est inconnue (cette méthode sera présentée dans la section 5.4.5). La notion de processus aléatoire généralisé est liée à celle d'espaces hilbertiens à noyau conditionnellement positif en analyse fonctionnelle.

La théorie des processus aléatoires intrinsèques a des conséquences pratiques importantes pour la modélisation boîte noire. Les processus aléatoires intrinsèques permettent de modéliser la sortie du système en distinguant deux effets. L'un constitue une tendance, par exemple une fonction polynomiale en les facteurs qui modélise des variations lentes de la sortie ; l'autre est un processus aléatoire du second ordre qui modélise les variations locales, plus rapides. Le krigeage intrinsèque réalise une prédiction linéaire de ces processus aléatoires généralisés en résolvant un système d'équations linéaires similaire à celui du krigeage.

Dans le krigeage intrinsèque, le terme de tendance n'est pas régularisé. Ceci permet d'inclure des comportements connus a priori dans le modèle du système.

Au chapitre 6 nous verrons des applications de cette méthode. Nous illustrerons en particulier la dérivation de processus aléatoire intrinsèques dans la section 6.1.2.

Chapitre 5

Choix d'un noyau

Résumé — Le choix d'un noyau est une étape fondamentale pour construire des modèles boîte noire à l'aide de méthodes d'approximation à noyau (krigeage ou régression régularisée dans un espace hilbertien à noyau reproduisant). Ce choix est délicat en pratique, surtout dans les espaces de dimension élevée. Nous explorons dans ce chapitre différents points de vue sur ce problème et tentons de présenter de façon synthétique les approches classiques utilisées dans différentes communautés. Des expériences numériques viennent montrer la pertinence pratique de résultats théoriques importants de la littérature.

5.1 Introduction

La présentation qui suit est centrée sur deux questions :

- Quel noyau est adapté à une classe de fonctions à approximer donnée ?
- Comment estimer un noyau à partir d'un nombre fini d'observations ?

Compte tenu de la première question, la seconde pourrait être reformulée en « Comment déterminer une classe de fonctions à partir d'un nombre fini d'échantillons de l'un de ses représentants ? ».

Le chapitre est organisé en trois parties. Nous décrivons tout d'abord les grandes classes de covariances paramétrées utilisées dans les applications (section 5.2). Nous présentons ensuite une synthèse des résultats asymptotiques concernant la prédiction linéaire sous échantillonnage non-uniforme (section 5.3). L'analyse asymptotique est en effet un outil privilégié pour comprendre l'influence du choix du noyau sur la prédiction. Enfin, nous rappelons les principales méthodes d'estimation des paramètres des covariances (section 5.4).

5.2 Structures de corrélation usuelles

Dans cette section, nous présentons les familles de covariances paramétrées que nous avons utilisées dans les applications et passons rapidement en revue leurs principales propriétés. Pour plus de détails, le lecteur pourra se reporter aux nombreux ouvrages traitant de ce sujet dans le

domaine des statistiques, notamment (Yadrenko, 1983 ; Yaglom, 1986 ; Matérn, 1986), et de la géostatistique (Cressie, 1993 ; Wackernagel, 1995 ; Chilès et Delfiner, 1999 ; Stein, 1999).

5.2.1 Principales propriétés des covariances

Combinaisons de covariances

Nous utiliserons les propriétés suivantes, en particulier dans les sections 5.4.4 et 6.5. Soient $F_i(\boldsymbol{x})$, $i = 1, \ldots, q$, des processus aléatoires indépendants, de moyenne nulle et de covariances $k_i(\boldsymbol{x}, \boldsymbol{y})$, et soit $F(\boldsymbol{x}) = \sum_{i=1}^{q} \alpha_i F_i(\boldsymbol{x})$. Alors $F(\boldsymbol{x})$ est de moyenne nulle et sa covariance s'écrit $k(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{q} \alpha_i^2 k_i(\boldsymbol{x}, \boldsymbol{y})$. Toute combinaison linéaire *positive* de covariances est donc une covariance *admissible*.

De même, si l'on considère une famille de covariances $k_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{y})$ paramétrées par le vecteur $\boldsymbol{\theta} \in \mathbb{R}^p$, α une mesure *positive* sur $\mathbb{R}^p$ et un domaine $\mathbb{A} \subset \mathbb{R}^p$ tel que $\alpha(\mathbb{A}) < \infty$, alors $\int_{\mathbb{A}} k_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{y}) d\alpha(\boldsymbol{\theta})$ est une covariance admissible.

Isotropie

Rappelons qu'un processus aléatoire $F(\boldsymbol{x})$, $\boldsymbol{x} \in \mathbb{R}^d$, est *isotrope au second ordre* si sa moyenne est constante et s'il admet une fonction de covariance telle que $\mathrm{Cov}[F(\boldsymbol{x}), F(\boldsymbol{y})] = k(\|\boldsymbol{x} - \boldsymbol{y}\|_2)$, où $\|.\|_2$ est la norme euclidienne de $\mathbb{R}^d$. Il s'agit donc d'une propriété d'invariance par rotation. Un processus isotrope au second ordre est implicitement stationnaire au second ordre.

L'utilisation de fonctions de covariance isotropes est intéressante en pratique car celles-ci ne nécessitent pas l'estimation d'un nombre de paramètres dépendant de la dimension de l'espace des facteurs. Toutefois, les facteurs intervenant dans la modélisation d'un système quelconque sont souvent de nature physique différente. De plus, la sensibilité de la sortie d'un système peut varier beaucoup d'un facteur à l'autre. Une solution classique consiste à *normaliser* les facteurs, ce qui permet parfois de modéliser le système par un processus aléatoire isotrope alors que cela n'était pas possible au départ. Plus généralement, il est possible d'effectuer une transformation linéaire de l'espace des facteurs $\mathbb{X} \subset \mathbb{R}^d$. Nous cherchons alors une matrice $d \times d$ inversible $\boldsymbol{V}$ telle que les nouveaux facteurs s'expriment sous la forme $\boldsymbol{V}\boldsymbol{x}$. La nouvelle métrique devient $d(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{(\boldsymbol{x} - \boldsymbol{y})^\top \boldsymbol{V}^\top \boldsymbol{V} (\boldsymbol{x} - \boldsymbol{y})} = \sqrt{(\boldsymbol{x} - \boldsymbol{y}) \boldsymbol{A} (\boldsymbol{x} - \boldsymbol{y})}$, où $\boldsymbol{A}$ est une matrice symétrique définie positive. Toutefois, effectuer une telle transformation nécessite généralement d'estimer au préalable les paramètres de $\boldsymbol{A}$ à partir des observations de la sortie du système, ce qui se révèle souvent coûteux. Une méthode de normalisation empirique consiste à choisir $\boldsymbol{A}$ comme l'inverse d'une estimée $\hat{\boldsymbol{C}}$ de la matrice de covariance des facteurs (ce qu'on peut interpréter comme un *blanchiment* de l'espace des facteurs).

Si toute fonction $k(h)$ de type positif peut être utilisée comme covariance d'un processus $F(x)$, $x \in \mathbb{R}$, nous verrons dans les paragraphes suivants qu'il faut imposer des conditions supplémentaires pour que $k(\|\boldsymbol{x} - \boldsymbol{y}\|_2)$ soit également une covariance isotrope admissible en dimension d. Notons $\mathcal{D}_d$ la classe des fonctions de covariances continues isotropes admissibles en dimension d et soit $\mathcal{D}_\infty = \bigcap_i \mathcal{D}_i$. Comme tout processus aléatoire isotrope sur $\mathbb{R}^d$ l'est également sur tout sous-espace vectoriel

$$\mathcal{D}_\infty \subset \cdots \subset \mathcal{D}_2 \subset \mathcal{D}_1.$$

Théorème 22. *Une fonction réelle $k(h)$ est une fonction de covariance isotrope admissible sur $\mathbb{R}^d$ si et seulement si elle peut être représentée sous la forme*

$$k(h) = 2^{(d-2)/2}\Gamma(d/2)\int_0^\infty (hu)^{-(d-2)/2} J_{(d-2)/2}(hu) dG(u)\,, \tag{5.1}$$

où J_k désigne la fonction de Bessel de première espèce d'ordre k et où la fonction $G(u)$, $u \in [0,\infty[$, est une fonction de répartition (non-décroissante, bornée et telle que $G(0) = 0$).

Démonstration. On trouvera ce résultat dans de nombreux ouvrages de la littérature, par exemple (Yaglom, 1986). □

Notons que G est liée à la mesure spectrale $\xi(\boldsymbol{u})$ du processus par la relation

$$G(u) = \int_{\|\boldsymbol{u}\|<u} d\xi(\boldsymbol{u})\,.$$

La représentation (5.1) permet d'établir quelques propriétés importantes des covariances isotropes.

Proposition 50. *Toute fonction de covariance isotrope $k(h)$ dans $\mathbb{R}^d$ est au moins $\lfloor \frac{1}{2}(d-1) \rfloor$ dérivable sur $[0,\infty[$.*

Démonstration. Voir (Schoenberg, 1938) et (Stein, 1999). □

Les fonctions de corrélation isotropes en dimension d admettent également la représentation (5.1), à condition d'ajouter la contrainte $\int_0^\infty dG(u) = 1$. Explicitons la représentation (5.1) de la corrélation pour quelques dimensions particulières :

$$\begin{aligned}
\rho(h) &= \int_0^\infty \cos(hu) dG(u), \qquad \forall \rho(h) \in \mathcal{D}_1,\\
\rho(h) &= \int_0^\infty J_0(hu) dG(u), \qquad \forall \rho(h) \in \mathcal{D}_2,\\
\rho(h) &= \int_0^\infty \frac{\sin(hu)}{hu} dG(u), \qquad \forall \rho(h) \in \mathcal{D}_3,\\
\rho(h) &= \int_0^\infty \exp(-h^2u^2) dG(u), \qquad \forall \rho(h) \in \mathcal{D}_\infty.
\end{aligned}$$

La première de ces expressions est la représentation spectrale bien connue d'une fonction de corrélation stationnaire. La dernière ne peut pas être obtenue directement à partir de (5.1) (voir (Schoenberg, 1938)). Enfin, des bornes inférieures des fonctions de corrélation peuvent être obtenues en utilisant la relation

$$\rho(h) \geq \inf_{t\geq 0} 2^{(d-2)/2}\Gamma(d/2) t^{-(d-2)/2} J_{(d-2)/2}(t)\,.$$

Ainsi, pour $d = 2$,

$$\rho(h) \geq \inf J_0(t) \approx -0.403,$$

pour $d = 3$,

$$\rho(h) \geq \inf \frac{\sin t}{t} \approx -0.218,$$

et pour $d = \infty$, $\rho(h) \geq 0$. Par conséquent, une covariance $k(h)$ admissible quelle que soit la dimension de l'espace des facteurs est nécessairement positive.

Covariances séparables

Définition 37. Une fonction de covariance stationnaire $k(\boldsymbol{h})$, $\boldsymbol{h} \in \mathbb{R}^d$ est dite *séparable* s'il existe des fonctions de covariances stationnaires $k_{[1]}(h), \ldots, k_{[d]}(h)$, $h \in \mathbb{R}$, telles que

$$k(\boldsymbol{h}) = k_{[1]}(h_{[1]}) \cdots k_{[d]}(h_{[d]})\,.$$

Une fonction de covariance stationnaire $k(\boldsymbol{h})$, $\boldsymbol{h} \in \mathbb{R}^d$, est dite *partiellement séparable* si existe des fonctions de covariances stationnaires $k_{[1]}(\boldsymbol{h}_1), \ldots, k_{[m]}(\boldsymbol{h}_m)$, $\boldsymbol{h}_1 \in \mathbb{R}^{d_1}$, $\ldots, \boldsymbol{h}_m \in \mathbb{R}^{d_m}$, et $\sum_{i=1}^m d_i = d$, telles que

$$k(\boldsymbol{h}) = k_{[1]}(\boldsymbol{h}_{[1]}) \cdots k_{[m]}(\boldsymbol{h}_{[m]})\,,$$

avec $\boldsymbol{h} = (\boldsymbol{h}_{[1]}^\mathsf{T}, \ldots, \boldsymbol{h}_{[m]}^\mathsf{T})^\mathsf{T}$.

Exemple 1. La fonction de covariance gaussienne

$$k(\boldsymbol{h}) = \sigma_0^2 \exp\big(-\frac{1}{2}\boldsymbol{h}^\mathsf{T}\boldsymbol{W}\boldsymbol{h}\big)\,,$$

avec $\boldsymbol{W} = \mathrm{diag}[w_1, \ldots, w_d]$ est séparable.

Exemple 2. Les covariances utilisées en géostatistique pour modéliser des processus spatiaux dépendant du temps $F(\boldsymbol{x}, t)$, $\boldsymbol{x} \in \mathbb{R}^2$, $t \in \mathbb{R}$, s'écrivant sous la forme

$$k(\boldsymbol{h}, \Delta t) = k_s(\boldsymbol{h}) k_t(\Delta t)\,,$$

sont partiellement séparables.

5.2.2 Familles classiques de covariances paramétrées

Nous considérons seulement des covariances stationnaires.

Fonctions sphériques

Les fonctions sphériques s'obtiennent par auto-convolution de l'indicatrice d'une sphère de rayon $a/2$ en dimension $\mathbb{R}^d$

$$k(\boldsymbol{h}) \propto \int_{\mathbb{R}^d} \mathbb{1}_{B(0,a/2)}(\boldsymbol{h} - \boldsymbol{y}) \mathbb{1}_{B(0,a/2)}(\boldsymbol{y}) d\boldsymbol{y}.$$

Nous pouvons considérer une telle covariance comme fonction de $h = \|\boldsymbol{h}\|$ seulement. Les covariances sphériques s'annulent pour $h > a$ et sont donc à support borné, ce qui peut faciliter la résolution du système linéaire du krigeage. En géostatistique, le paramètre d'échelle a s'appelle la *portée* de la covariance. Pour $h \leq a$, les expressions des fonctions de corrélation pour $d = 1, 2, 3$

et 5 sont respectivement

$$\begin{aligned}
\rho_1(h) &= 1 - \frac{h}{a}\,, \\
\rho_2(h) &= 1 - \frac{2}{\pi}\left(\frac{h}{a}\sqrt{1-\left(\frac{h}{a}\right)^2} + \arcsin\left(\frac{h}{a}\right)\right)\,, \\
\rho_3(h) &= 1 - \frac{3}{2}\frac{h}{a} + \frac{1}{2}\left(\frac{h}{a}\right)^3\,, \\
\rho_5(h) &= 1 - \frac{15}{8}\frac{h}{a} + \frac{5}{4}\left(\frac{h}{a}\right)^3 - \frac{3}{8}\left(\frac{h}{a}\right)^5\,.
\end{aligned}$$

Bien que ces covariances soient d'usage courant en géostatistique, nous ne les utiliserons pas dans les applications en raison de leur dépendance non triviale avec la dimension de l'espace des facteurs.

Fonctions cubiques

Le modèle cubique possède un terme principal irrégulier en $|h|^3$. Une covariance cubique peut donc être utilisée pour modéliser un processus aléatoire *différentiable* en moyenne quadratique, mais n'est admissible qu'en dimension au plus égale à trois. Comme pour les covariances sphériques, la covariance cubique est nulle pour $h > a$. Pour $h \leq a$, la fonction de corrélation cubique est

$$\rho(h) = 1 - 7\left(\frac{h}{a}\right)^2 + \frac{35}{4}\left(\frac{h}{a}\right)^3 - \frac{7}{2}\left(\frac{h}{a}\right)^5 + \frac{3}{4}\left(\frac{h}{a}\right)^7\,.$$

On peut construire sur ce modèle des covariances dérivables aux ordres supérieurs avec des termes principaux irréguliers en $|h|^{2p+1}$, $p \in \mathbb{N}$. Nous n'utiliserons pas cette covariance en pratique car elle n'est utilisable qu'en dimension trois au plus.

Fonctions exponentielles

Les covariances exponentielles s'écrivent

$$k(h) = \sigma_0^2 \exp\left(-(h\rho)^\alpha\right)\,, \quad \text{pour } 0 < \alpha \leq 2\,.$$

ρ est un paramètre d'échelle qui permet d'ajuster la largeur caractéristique de la covariance (la *portée*). α permet d'ajuster le comportement de la covariance au voisinage de 0. Toutefois, pour $\alpha < 2$, la covariance n'est jamais dérivable à l'origine. Pour $\alpha = 2$, la covariance devient analytique ($k(h) \in C^\infty$), et on parle de *covariance gaussienne*. Par conséquent, la famille des covariances exponentielles ne permet pas d'ajuster la régularité de la covariance à l'origine.

Notons que des fonctions de covariance du type produit de polynômes et d'exponentielles peuvent également être construites.

Covariances de Matérn

Nous tendons à privilégier l'utilisation de cette covariance dans les applications. Dans la littérature, la covariance de Matérn (appellation utilisée par M.L. Stein) est aussi appelée covariance $\mathcal{K}$-Bessel (Chilès et Delfiner, 1999), covariance *ARMA* spatiale (Jones et Vecchia, 1993) ou noyau

de Sobolev (Schaback, 1999). Les covariances de Matérn possèdent la particularité très intéressante que leur régularité (dérivabilité à l'origine) peut s'ajuster à l'aide d'un seul paramètre (Stein, 1999). De plus, les covariances de Matérn sont toutes dans $\mathcal{D}_\infty$ et sont donc utilisables en dimension quelconque. Cette classe de covariances est obtenue (pour $d = 1$) en considérant les densités spectrales rationnelles du type

$$\frac{d\xi}{du} = \frac{\phi}{(\alpha^2 + u^2)^{-\nu-1/2}}$$

où les constantes ν, ϕ et α sont positives (u désigne la fréquence). Le paramètre permettant d'ajuster la régularité est donc ν. Plus ν est grand, plus la covariance correspondante est régulière. En particulier, $F(x)$ sera r fois dérivable en moyenne quadratique si et seulement si $\nu > r$.

En dimension d quelconque, la classe des covariances de Matérn correspond aux fonctions s'écrivant sous la forme

$$k(h) = \phi(\alpha h)^\nu \mathcal{K}_\nu(\alpha h)\,, \quad h \geq 0\,,$$

où $\mathcal{K}_\nu$ est la fonction de Bessel de seconde espèce modifiée d'ordre ν. Notons la propriété des fonctions de Bessel modifiées qui permet de dériver la covariance de Matérn (Abramowitz et Stegun, 1965) :

$$\mathcal{K}'_\nu(x) = -\mathcal{K}_{\nu-1}(x) - \frac{\nu}{x}\mathcal{K}_\nu(x)\,.$$

(Il n'est toutefois pas facile d'obtenir les expressions analytiques des dérivées de $\mathcal{K}_\nu$ par rapport à ν.) Pour certaines valeurs de ν, la covariance de Matérn prend la forme d'une covariance polynôme–exponentielle en h. Ainsi, pour $\nu = 1/2$, $k(h) = \pi\phi\alpha^{-1}e^{-\alpha h}$ et pour $\nu = 3/2$, $k(h) = \frac{1}{2}\pi\phi\alpha^{-3}(1 + \alpha h)e^{-\alpha h}$.

Après reparamétrisation (Handcock et Wallis, 1994 ; Stein, 1999), la covariance de Matérn peut s'écrire sous la forme

$$k(h) = \frac{\sigma_0^2}{2^{\nu-1}\Gamma(\nu)} r^\nu \mathcal{K}_\nu(r)\,,$$

avec $r = \frac{2\nu^{1/2}h}{\rho}$. Le paramètre d'échelle ρ caractérise la largeur de la covariance (typiquement $\rho = 1$ implique une portée approximativement égale à 1) et σ_0^2 est la valeur de la covariance à l'origine. Nous utiliserons cette paramétrisation en pratique.

5.2.3 Covariances bandes

Nous avons déjà eu l'occasion de présenter la méthode des bandes tournantes (Matheron, 1973) dans la section 2.1.3. Soit $F_s(x)$, $x \in \mathbb{R}$, un processus aléatoire du second ordre, de moyenne nulle et de covariance stationnaire $k_s(x - y)$. Soit de plus un vecteur aléatoire $\boldsymbol{U} \in \mathbb{R}^d$, indépendant de $F_s(\mathbb{R})$, suivant une loi à densité uniforme sur la sphère unité. Considérons ensuite le processus aléatoire $F_d(\boldsymbol{x}) = F_s((\boldsymbol{x}, \boldsymbol{U})_{\mathbb{R}^d})$, $\boldsymbol{x} \in \mathbb{R}^d$, où $(.,.)_{\mathbb{R}^d}$ désigne le produit scalaire dans $\mathbb{R}^d$.

F_d est isotrope (Matheron, 1973) et sa covariance s'écrit en fonction de celle de F_s sous la forme

$$k_d(h) = \frac{2\Gamma(\frac{1}{2}d)}{\pi^{\frac{1}{2}}\Gamma(\frac{1}{2}(d-1))} \int_0^1 k_s(vh)(1 - v^2)^{\frac{1}{2}(d-3)} dv\,. \tag{5.2}$$

Cette représentation peut être reliée à la représentation spectrale (5.1) en choisissant la mesure G telle que

$$k_s(h) = \int_0^\infty \cos(hu) dG(u)\,.$$

Il existe donc une correspondance bijective entre une fonction de covariance k_d isotrope en dimension d et une fonction de covariance k_s admissible en dimension un. Pour $d = 3$ cette correspondance prend une forme particulièrement simple puisque

$$k_3(h) = \int_0^1 k_s(vh)dh\,.$$

La méthode des bandes tournantes permet de simuler un processus aléatoire isotrope en dimension d. Soient $F_i(x)$, $i = 1, \ldots, q$, $x \in \mathbb{R}$, des processus aléatoires indépendants, de moyenne nulle, de même covariance $k_s(h)$, et soit

$$F(\boldsymbol{x}) = \frac{1}{q^{1/2}} \sum_{i=1}^{q} F_i((\boldsymbol{x}, \boldsymbol{U}_i)_{\mathbb{R}^d}), \tag{5.3}$$

où les $\boldsymbol{U}_i$ sont des vecteurs aléatoires unitaires de $\mathbb{R}^d$, indépendants des F_i et de distribution uniforme sur la sphère unité. Alors $F(\boldsymbol{x})$ admet une fonction de covariance k_d donnée par (5.2). Conditionnellement à des réalisations $\boldsymbol{u}_i$ des vecteurs aléatoires $\boldsymbol{U}_i$, $i = 1, \ldots, q$, $F(\boldsymbol{x})$ est une combinaison linéaire de processus aléatoires F_i de covariance $k_i(\boldsymbol{h}) = k_s((\boldsymbol{h}, \boldsymbol{u}_i)_{\mathbb{R}^d})$. Les covariances k_i seront appelées *covariances bandes*. Lorsque l'on considère un nombre suffisant de termes dans (5.3), la méthode des bandes tournantes permet d'obtenir une réalisation de $F(\boldsymbol{x})$ approximativement isotrope en dimension d à partir de q réalisations indépendantes d'un processus en dimension un (Chilès et Delfiner, 1999).

5.2.4 Compléments sur les covariances généralisées

Nous revenons dans cette section sur les covariances généralisées des processus aléatoires intrinsèques. Rappelons qu'une covariance généralisée $k(\boldsymbol{h})$, $\boldsymbol{h} \in \mathbb{X}$, d'une *IRF*($l$) $F_{\mathrm{G}}(\lambda)$ est une fonction non-linéaire symétrique (paire) telle que $\forall \lambda, \mu \in \Lambda_l$

$$\mathrm{E}[F_{\mathrm{G}}(\lambda)F_{\mathrm{G}}(\mu)] = \iint k(\boldsymbol{x} - \boldsymbol{y})d\lambda(\boldsymbol{x})d\mu(\boldsymbol{y})\,.$$

Remarquons que la propriété

$$\mathrm{E}[F_{\mathrm{G}}(\lambda)^2] = \iint k(\boldsymbol{x} - \boldsymbol{y})d\lambda(\boldsymbol{x})d\lambda(\boldsymbol{y})\,, \quad \forall \lambda \in \Lambda_k\,,$$

suffit pour définir une covariance généralisée, comme on peut s'en convaincre en remplaçant λ par $\lambda + \mu$. Les covariances généralisées des *IRF* d'ordre l s'identifient à l'ensemble des fonctions conditionnellement positives par rapport aux polynômes de degré $\leq l$. Rappelons également que les covariances généralisées d'une *IRF*(l) particulière constituent une *classe d'équivalence*, obtenue en ajoutant à l'une quelconque d'entre elles un polynôme pair de degré inférieur ou égal à $2l$.

Lorsque l augmente, on impose aux mesures λ d'appartenir à des espaces de plus en plus petits. Corrélativement, l'ensemble des covariances généralisées devient de plus en plus grand. Par exemple, pour $\alpha > 0$, $\|\boldsymbol{h}\|^\alpha$ n'est jamais une covariance stationnaire, puisqu'elle est non bornée, mais constitue un *variogramme* admissible si $\alpha < 2$.

Par la suite, nous présentons plus spécifiquement une classe de covariances généralisées isotropes.

Proposition 51. *La fonction*

$$k(\|\boldsymbol{h}\|) = k(h) = \Gamma(-\alpha/2)h^{\alpha}\,, \tag{5.4}$$

est une covariance généralisée admissible d'une IRF(l) *pour* $\alpha \in]0, 2l+2[$. *Lorsque* α *est un entier impair, (5.4) s'écrit* $(-1)^{p+1}h^{2p+1}$, *qui est donc une covariance généralisée admissible d'une IRF d'ordre* l, *lorsque* $p \leq l$.

Démonstration. Voir (Matheron, 1973). □

Sans entrer dans les détails de la théorie spectrale des processus aléatoires généralisés intrinsèques, il est utile de retenir que la densité spectrale (dans un sens généralisé) correspondant à la covariance (5.4) est proportionnelle à $\|\boldsymbol{u}\|^{-\alpha-d}$ (Matheron, 1973) (rappelons que d est la dimension de l'espace des facteurs). Cette propriété permet de relier l'ordre α de la covariance généralisée (5.4) à la régularité des trajectoires du processus aléatoire.

Plus généralement, la fonction

$$k(h) = \sum_{p=0}^{l}(-1)^{p+1}b_p h^{2p+1}\,, \tag{5.5}$$

où les coefficients b_p sont des réels positifs, est une covariance généralisée d'ordre l. (La condition de positivité des coefficients b_p garantit que $k(h)$ est une covariance généralisée isotrope admissible quelle que soit la dimension de l'espace des facteurs.) Cette covariance généralisée, appelée *covariance polynomiale*, est d'utilisation très pratique pour au moins deux raisons. Tout d'abord, les paramètres b_p interviennent de manière linéaire. Par conséquent, la forme analytique des dérivées de la covariance par rapport à ces paramètres s'écrit immédiatement. Ensuite, la régularité de la covariance est fixée par la valuation de la covariance (degré du terme de plus petit degré) et par conséquent, les covariances polynomiales présentent un avantage similaire à celui de la covariance de Matérn.

5.3 Théorie asymptotique de la prédiction linéaire

Cette section présente des résultats théoriques de la littérature aux conséquences pratiques importantes pour les applications. Soit un processus aléatoire $F(\boldsymbol{x})$ de moyenne $m(\boldsymbol{x})$ et de covariance $k(\boldsymbol{x}, \boldsymbol{y})$. Si l'on effectue une prédiction de ce processus par krigeage en utilisant un modèle du second ordre incorrect (moyenne ou covariance incorrecte), il n'est plus possible d'affirmer que le prédicteur possède toutes les propriétés agréables qui seraient vérifiées si le modèle correct était utilisé (estimateur non biaisé et variance d'erreur minimale). Le rôle et l'influence de l'erreur d'estimation du modèle sur la prédiction ne sont que partiellement compris aujourd'hui, et les résultats de la littérature sont essentiellement asymptotiques. Ces résultats reposent souvent sur des preuves difficiles que nous nous abstiendrons alors de reproduire. Nous nous bornerons à illustrer ces propriétés à l'aide d'expériences numériques présentées au fur et à mesure.

5.3.1 Convergence des prédicteurs linéaires

Notion de consistance

Soit une suite $(\boldsymbol{x}_i)_{i\in\mathbb{N}}$ de points de $\mathbb{R}^d$, telle que $\boldsymbol{x}$ soit un point d'adhérence. Nous considérons un processus aléatoire $F(\boldsymbol{x})$ à moyenne nulle et nous nous intéressons à la prédiction linéaire $\hat{F}_n(\boldsymbol{x})$ de $F(\boldsymbol{x})$ en fonction des $F(\boldsymbol{x}_1),\ldots,F(\boldsymbol{x}_n)$. Nous dirons que l'estimateur $\hat{F}_n(\boldsymbol{x})$ est *consistant* si

$$\lim_{n\to\infty} \mathrm{Var}[F(\boldsymbol{x}) - \hat{F}_n(\boldsymbol{x})] = 0$$

Cette notion de consistance est pertinente quand il est possible de faire des observations avec une densité arbitraire sur un voisinage de l'espace des facteurs contenant $\boldsymbol{x}$. Notons qu'une telle notion n'existe pas dans le cas des séries chronologiques pour lesquelles les études asymptotiques conduisent à faire grandir l'horizon d'observation.

Rappelons que la variance de l'erreur de prédiction s'exprime sous forme matricielle par

$$\mathrm{Var}[F(\boldsymbol{x}) - \hat{F}(\boldsymbol{x})] = k(\boldsymbol{x},\boldsymbol{x}) - 2\boldsymbol{\lambda}_{\boldsymbol{x}}^{\mathsf{T}}\boldsymbol{k}_{\boldsymbol{x}} + \boldsymbol{\lambda}_{\boldsymbol{x}}^{\mathsf{T}}\boldsymbol{K}\boldsymbol{\lambda}_{\boldsymbol{x}} \tag{5.6}$$

Si $\boldsymbol{\lambda}_{\boldsymbol{x}}$ correspond au meilleur prédicteur linéaire, alors la variance de l'erreur peut se mettre sous la forme plus simple $\mathrm{Var}[F(\boldsymbol{x}) - \hat{F}(\boldsymbol{x})] = k(\boldsymbol{x},\boldsymbol{x}) - \boldsymbol{\lambda}_{\boldsymbol{x}}^{\mathsf{T}}\boldsymbol{k}_{\boldsymbol{x}}$.

Proposition 52. *Si la covariance de $F(\boldsymbol{x})$ est continue sur sa diagonale, le meilleur prédicteur linéaire est consistant.*

Démonstration. Soit $\boldsymbol{x}_{\varphi(n)}$ une suite extraite convergente vers $\boldsymbol{x}$, où $(\varphi(n))_n$ est une suite non décroissante telle que $\varphi(n) \leq n$,

$$\begin{aligned}\mathrm{Var}[F(\boldsymbol{x}) - \hat{F}_n(\boldsymbol{x})] &\leq \mathrm{Var}[F(\boldsymbol{x}) - F(\boldsymbol{x}_{\varphi(n)})] \\ &= k(\boldsymbol{x},\boldsymbol{x}) + k(\boldsymbol{x}_{\varphi(n)},\boldsymbol{x}_{\varphi(n)}) - 2k(\boldsymbol{x},\boldsymbol{x}_{\varphi(n)}) \\ &\overset{n\to\infty}{\longrightarrow} 0\,,\end{aligned}$$

lorsque $k(\boldsymbol{x},\boldsymbol{y})$ est continue sur sa diagonale. □

Dans les sections suivantes, des résultats plus précis sont présentés sur la rapidité de convergence.

Résultats classiques de convergence

Les résultats classiques de convergence des prédicteurs linéaires appartiennent à la littérature de la théorie de l'approximation et de l'interpolation et sont souvent particulièrement difficile à établir. Le premier de ces résultats est présenté par Duchon (1978) dans le cadre des splines. D'autres bornes d'erreur ont été obtenues dans le contexte plus général des espaces de Sobolev (Madych et Nelson, 1988, 1990). Nous présenterons le résultat encore plus général décrit dans (Wu et Schaback, 1993). Mentionnons que des résultats similaires ont été établis par d'autres auteurs (Light et Wayne, 1998 ; Narcowich et al., 2003).

Avant de rappeler ces résultats, donnons quelques indications pour comprendre de manière intuitive leur origine. Dans le cas de processus aléatoires stationnaires, nous avons déjà vu dans

la section 2.5.1 que la variance d'erreur de prédiction pouvait être majorée par la relation

$$\mathrm{Var}[F(\boldsymbol{x}) - \hat{F}_n(\boldsymbol{x})] \leq k(\mathbf{0}) - \frac{k(\boldsymbol{x} - \boldsymbol{x}_{\varphi(n)})^2}{k(\mathbf{0})} \leq 2(k(\mathbf{0}) - k(\boldsymbol{x} - \boldsymbol{x}_{\varphi(n)}))\,.$$

Cette majoration suggère que le comportement de la covariance à l'origine joue un rôle essentiel dans la convergence de l'erreur de prédiction. Il est traditionnel de définir la régularité d'une fonction à l'aide de la définition suivante.

Définition 38. Une fonction f est *ponctuellement Lipschitz* $\alpha > 0$ en $\boldsymbol{x}_0$ s'il existe une constante $C > 0$ et un polynôme $p_{\boldsymbol{x}_0}$ de degré $l = \lfloor \alpha \rfloor$ tels que

$$\forall \boldsymbol{x} \in \mathbb{R}^d, |f(\boldsymbol{x}) - p_{\boldsymbol{x}_0}(\boldsymbol{x})| \leq C\|\boldsymbol{x} - \boldsymbol{x}_0\|^\alpha. \tag{5.7}$$

Une fonction est *uniformément Lipschitz* α sur $\mathbb{X}$ si elle vérifie (5.7) pour tout $\boldsymbol{x} \in \mathbb{X}$, avec une constante C indépendante de $\boldsymbol{x}$. La *régularité* (lipschitzienne) de f en $\boldsymbol{x}_0$ est le supremum des α pour lesquels f est Lipschitz α en $\boldsymbol{x}_0$.

Si f est uniformément Lipschitz α au voisinage de $\boldsymbol{x}$ avec $\alpha > r$, $r \in \mathbb{N}$, alors on peut vérifier que f est nécessairement r fois dérivable sur ce voisinage (le polynôme $p_{\boldsymbol{x}_0}$ est déterminé de manière unique). Il est possible de montrer que si la covariance stationnaire $k(\boldsymbol{h})$ est de régularité α au voisinage de $\mathbf{0}$, alors

$$\mathrm{Var}[F(\boldsymbol{x}) - \hat{F}_n(\boldsymbol{x})] \leq C\|\boldsymbol{x} - \boldsymbol{x}_{\varphi(n)}\|^\beta\,,$$

où la constante β est égale à la valuation du polynôme $p_{\mathbf{0}}(\boldsymbol{x}) - k(\mathbf{0})$. L'erreur de prédiction décroît donc lorsque $\boldsymbol{x}$ est entouré par un nombre croissant de points de plus en plus proches et lorsque la régularité de la covariance augmente. Les résultats présentés dans les paragraphes suivants sont cependant nettement plus forts, puisque l'on montre que la décroissance de la variance de l'erreur est l'ordre de h^α, où h est une quantité définie au paragraphe suivant.

Dans la littérature, il est usuel d'introduire une notion de densité d'échantillonnage et de majorer l'erreur de prédiction en fonction de celle-ci. La ρ-densité d'un échantillonnage $S = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ autour d'un point $\boldsymbol{x} \in \mathbb{X}$ est définie (Wu et Schaback, 1993) par

$$h_{\rho,S}(\boldsymbol{x}) = \sup_{\|\boldsymbol{x}-\boldsymbol{y}\|_2 \leq \rho} \min_{\boldsymbol{z} \in S} \|\boldsymbol{y} - \boldsymbol{z}\|_2\,.$$

Elle permet donc de quantifier la façon dont S couvre une boule de rayon ρ autour de $\boldsymbol{x}$. Pour quantifier la qualité de l'approximation en un point $\boldsymbol{x}$, il est pertinent de s'intéresser à l'ensemble des échantillonnages S tels que

$$h_{\rho,S}(\boldsymbol{x}) \leq h_{\rho,d},$$

où $h_{\rho,d}$ dépend de ρ et la dimension de l'espace des facteurs. Par conséquent, le paramètre d'échelle $h_{\rho,d}$ peut servir à quantifier la qualité de l'approximation en terme de densité de points autour de $\boldsymbol{x}$. Dans les bornes d'erreurs présentées ci-dessous, la dimension de l'espace des facteurs joue un rôle important. Typiquement, la rapidité de décroissance de l'erreur avec le nombre de points observés dans un voisinage de $\boldsymbol{x}$ diminue lorsque la dimension d augmente.

Ce fait s'explique notamment par le phénomène portant le nom de *concentration de la mesure*. Par exemple, si l'on tire des points avec une densité uniforme dans une boule de dimension d

élevée, ceux-ci auront tendance à s'accumuler près de la surface de la boule au sens de la distance euclidienne. Si l'on tire aléatoirement et uniformément des points dans un ensemble $\mathbb{X}$ borné, il faut, quand d augmente, de plus en plus de tirages pour s'approcher d'un point quelconque $\boldsymbol{x}$ de $\mathbb{X}$ parce que les distances euclidiennes entre les points augmentent, comme le montre par exemple le théorème suivant (l'écart type de la norme d'un vecteur aléatoire à valeurs dans $\mathbb{R}^d$ converge vers une constante lorsque d augmente, alors que sa moyenne augmente proportionnellement à la racine carrée de d).

Théorème 23. *Soit un vecteur aléatoire $\boldsymbol{X}$ à valeurs dans $\mathbb{R}^d$. Sous l'hypothèse que les moments d'ordre 8 des variables $X_{[i]}$, $i = 1, \ldots, d$, existent et soient finis,*

$$\begin{aligned} \mathrm{E}[\|\boldsymbol{X}\|_2] &= \sqrt{ad - b} + \mathrm{O}(1/d)\,, \\ \mathrm{Var}[\|\boldsymbol{X}\|_2] &= b + \mathrm{O}(1/\sqrt{d})\,, \end{aligned}$$

où a et b sont des paramètres qui dépendent uiquement des moments centrés d'ordre 1, 2, 3 et 4 des $X_{[i]}$.

Démonstration. Voir (Demartines, 1994). □

Un autre phénomène rend la prédiction difficile en dimension élevée. Étant donné un point $\boldsymbol{x}$ quelconque du domaine $\mathbb{X}$, sous certaines conditions que nous ne précisons pas, $(D^n_{\max} - D^n_{\min})/D^n_{\min}$ tend en probabilité vers zéro quand d tend vers l'infini, où $D^n_{\min}$ et $D^n_{\max}$ sont les plus petite et plus grande distances euclidiennes de $\boldsymbol{x}$ à un ensemble S de n points distribués aléatoirement dans $\mathbb{X}$ (Beyer et al., 1999 ; Aggarwal et al., 2001). Autrement dit, lorsque la dimension de l'espace des facteurs augmente, la différence entre la plus grande et la plus petite distance à un point $\boldsymbol{x}$ croît moins rapidement que la plus petite distance. Par conséquent, la notion de *plus proche voisin* devient moins pertinente en grande dimension. Par suite, le principe consistant à prédire en fonction des valeurs prises par les observations les plus proches (l'un des fondements de la géostatistique) devient également moins pertinent. Nous constaterons expérimentalement ce phénomène dans la section 6.4.2.

Théorème 24. *Soient une fonction f^* de $H^l(\mathbb{X})$, $\mathbb{X} \in \mathbb{R}^d$, et la D^l-spline $\hat{f}$ qui interpole f^* sur un ensemble S de $\bar{\mathbb{X}}$ en minimisant $\|D^l f\|_{L^2(\mathbb{R}^d)}$[1]. L'erreur d'estimation $f^* - \hat{f}$ en fonction de la distance $h = \sup_{\boldsymbol{x} \in \mathbb{X}} \inf_{\boldsymbol{z} \in S} \|\boldsymbol{z} - \boldsymbol{x}\|_2$ de S à $\mathbb{X}$ est telle que*

$$\|D^r(f^* - \hat{f})\|_{L^p(\mathbb{X})} = \mathrm{o}(h^{l-r-(d/2)+(d/p)}).$$

Démonstration. La démonstration, très technique, est présentée dans (Duchon, 1978). □

[1] D^l désigne ici l'opérateur de dérivation d'ordre l et $H^l(\mathbb{X})$ l'espace de Sobolev d'ordre l, c'est-à-dire le sous-espace de $L^2(\mathbb{X})$ défini par

$$H^l(\mathbb{X}) = \{f \in L^2(\mathbb{X}) \text{ tels que } D^p f \in L^2(\mathbb{X}) \text{ pour } 1 \leq p \leq l\}$$

et muni du produit scalaire

$$(f, g)_{H^l} = \sum_{k=0}^{l} (D^k f, D^k g)_{L^2}\,.$$

Notons que l'ordre l de l'espace de Sobolev conditionne la régularité de $\hat{f}$. Le théorème 24 met donc en relation la régularité d'une classe de fonctions donnée avec la vitesse de décroissance de l'erreur d'approximation.

Madych et Nelson (1988, 1990) généralisent le résultat de Duchon aux noyaux conditionnellement positifs. Wu et Schaback (1993) présentent un résultat analogue (et encore un peu plus général) que nous mentionnons maintenant.

Théorème 25. *Soit $k(\boldsymbol{h})$, $\boldsymbol{h} \in \mathbb{R}^d$, un noyau conditionnellement positif d'ordre l, admettant une transformée de Fourier généralisée $\tilde{k}$, telle que $\tilde{k}(\boldsymbol{u}) = \mathrm{O}(\|\boldsymbol{u}\|^{-d-s_\infty})$ lorsque $\|\boldsymbol{u}\| \to \infty$. Sous des hypothèses que nous ne précisons pas,*

$$|D^r(f^* - \hat{f})(\boldsymbol{x})| = \mathrm{O}(h_{\rho,S}(\boldsymbol{x})^{s_\infty/2-r})$$

pour $0 \leq r \leq s_\infty/2$.

Démonstration. Une démonstration technique mais relativement accessible est présentée dans (Wu et Schaback, 1993). □

Ces théorèmes (Duchon, 1978 ; Wu et Schaback, 1993) expriment essentiellement les mêmes propriétés. L'erreur de prédiction d'une fonction ou de ses dérivées diminue lorsque la densité d'échantillonnage ou la régularité de la covariance augmente, et augmente avec la dimension de l'espace des facteurs ou avec l'ordre de dérivation. Ces comportements seront vérifiés numériquement dans la section suivante.

Expériences numériques sur la convergence des prédicteurs linéaires

Examinons d'un point de vue numérique l'influence du choix de la covariance sur la décroissance de l'erreur de prédiction. Le protocole expérimental est le suivant. Nous sélectionnons une covariance isotrope $k(h)$, de variance unité à l'origine.

Phase 1

1. Une séquence finie de points $\boldsymbol{x}_i$, $i = 1, \ldots, n$ est tirée aléatoirement, avec une distribution uniforme sur la boule unité de $\mathbb{R}^d$ [2].
2. Le vecteur $\boldsymbol{\lambda}^0$ des coefficients de la meilleure prédiction linéaire au point $\mathbf{0}$ est calculé.
3. La variance σ^2 de l'erreur de prédiction en ce point est calculée.

Phase 2

1. La phase 1 est répétée N fois pour obtenir une suite $\sigma_1^2, \ldots, \sigma_N^2$. N est choisi suffisamment grand pour que l'histogramme des σ_i^2 soit stable.
2. Les quantiles à 10% et à 90% de la suite (σ_n) sont calculés.

Phase 3

[2] Le choix d'un volume particulier de $\mathbb{R}^d$ a des implications non évidentes. Le volume d'un cube de coté 2 en dimension d est 2^d et tend vers l'infini quand la dimension de l'espace augmente. Cependant, le volume de la boule unité (tangente aux $2d$ faces du cube précédent) est $\frac{\pi^p}{p!}$ si $d = 2p$ et $\frac{2^{p+1}\pi^p}{1.3.5\ldots(2p+1)}$ si $d = 2p+1$; il tend donc vers 0 quand la dimension augmente. Les phénomènes en dimension élevée ne sont jamais très intuitifs. Notons toutefois que quel que soit le volume, les points d'une distribution uniforme ont tendance, en probabilité, à s'accumuler à sa surface.

1. La phase 2 est répétée en faisant varier n.

Les figures 5.1 à 5.5 présentent les résultats obtenus. La figure 5.1 représente l'évolution typique en échelle logarithmique des quantiles à 10% et à 90% de la variance de l'erreur en fonction du nombre de points. Nous constatons que la pente a tendance à être constante lorsque le régime asymptotique est atteint. Cela suggère que les bornes polynomiales présentées dans le théorème 25 sont optimales.

Nous constatons que la décroissance de l'erreur prédiction dépend de deux facteurs principalement, à savoir la régularité de la covariance et la dimension de l'espace des facteurs.

Régularité de la covariance. Nous utilisons dans un premier temps la covariance de Matérn (qui a été présentée dans la section 5.2.2). Ce type de covariance possède trois paramètres, σ_0^2, ρ et ν, correspondant respectivement à la variance à l'origine, à la distance caractéristique de corrélation et à la régularité de la covariance (ν permet d'ajuster la dérivabilité de la covariance à l'origine). La figure 5.2 montre que la décroissance asymptotique de l'erreur est plus rapide lorsque la régularité augmente. Nous constatons le même type de résultat, illustré par la figure 5.3, avec des covariances généralisées polynomiales (l'expérience est menée en regardant la variance de l'erreur de prédiction du krigeage intrinsèque d'ordre un).

Dimension de l'espace des facteurs. Lorsque la dimension de l'espace augmente, l'erreur de prédiction augmente pour un nombre de points échantillonnés fixé. La figure 5.4 illustre le fait qu'en dimension 15, la prédiction utilisant la covariance de Matérn devient complètement inefficace parce que les points échantillonnés se concentrent à une distance supérieure à la distance caractéristique de corrélation (la *portée*). En revanche, la prédiction avec une covariance généralisée polynomiale est meilleure (voir la figure 5.5) probablement parce que ce type de covariance ne possède pas de portée.

5.3.2 Consistance avec un modèle incorrect

Soit un processus aléatoire $F(\boldsymbol{x})$ de moyenne nulle et de covariance $k_0(\boldsymbol{x}, \boldsymbol{y})$. Pour prédire $F(\boldsymbol{x})$, nous utilisons cette fois une covariance $k_1(\boldsymbol{x}, \boldsymbol{y})$. L'objectif est de montrer que même si la covariance est incorrecte, la prédiction reste consistante (sous des hypothèses faibles) lorsque le nombre d'observations croît de manière dense sur un domaine borné de l'espace des facteurs.

Exemple élémentaire

Supposons $k_1 = \gamma k_0$ où γ est une constante multiplicative positive. Dans ce cas, les coefficients $\lambda_{i,\boldsymbol{x}}$ du krigeage calculés avec la covariance k_1 sont identiques à ceux calculés avec la covariance k_0. (Il est immédiat en effet de vérifier que la solution du système d'équations linéaires du krigeage est invariante par multiplication de la covariance par une constante.) La multiplication de la covariance par une constante n'affecte que la variance de l'erreur de prédiction. Nous devons cependant distinguer la variance de l'erreur calculée avec la covariance k_1 de celle calculée avec la vraie covariance k_0. Puisque les coefficients du krigeage sont inchangés la variance de l'erreur sous k_0 est également inchangée et par conséquent, le prédicteur obtenu avec la covariance incorrecte possède la même propriété de consistance que celui obtenu à partir de la vraie covariance. En

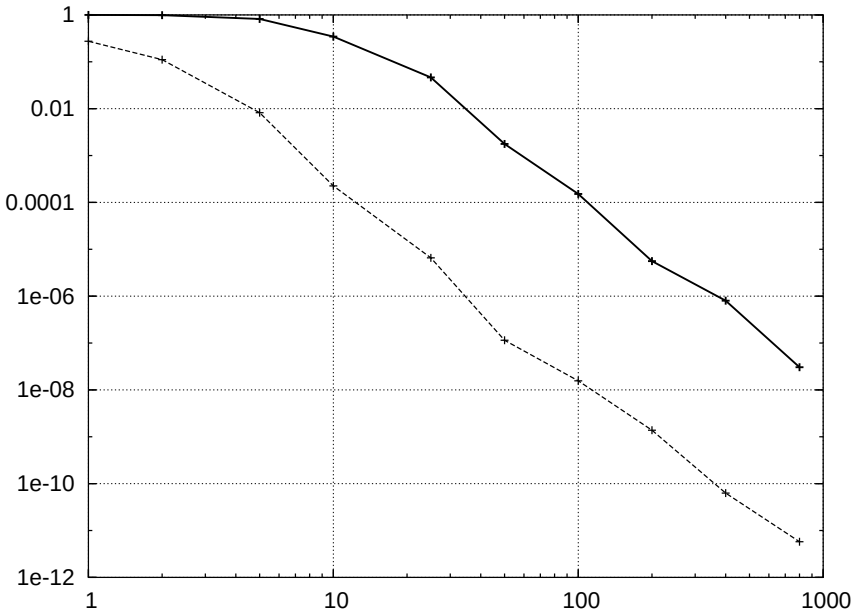


FIG. 5.1 – Comportement des quantiles à 10% (trait pointillé) et à 90% (trait continu) des variances d'erreur de prédiction en fonction du nombre de points tirés aléatoirement. L'échelle est logarithmique selon les deux axes. Expérience en dimension 1. Covariance de Matérn de paramètres $\nu = 2.0$, $\rho = 0.4$, $\sigma_0^2 = 1.0$.

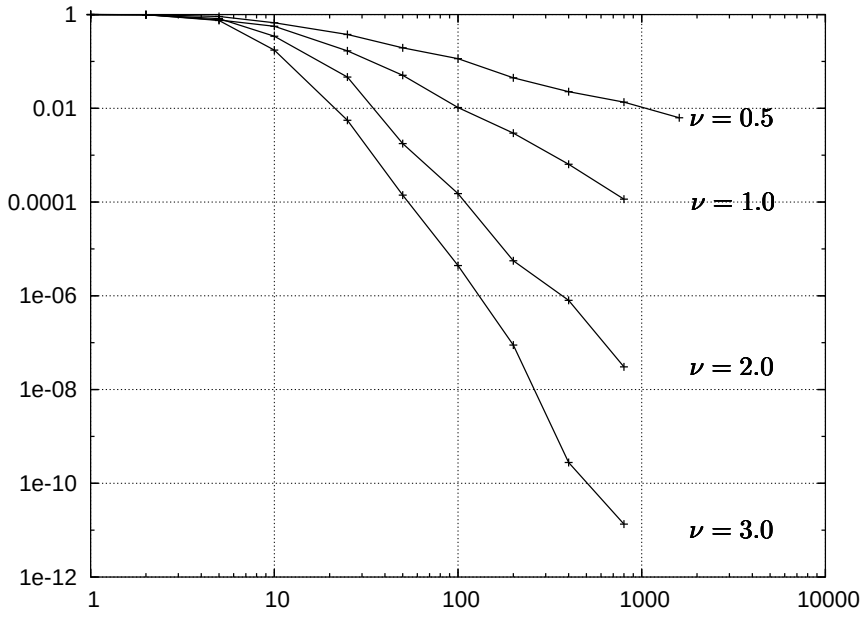


FIG. 5.2 – Comportement des quantiles à 90% des variances d'erreur de prédiction en fonction du nombre de points tirés aléatoirement, pour différentes régularités : nous faisons varier le paramètre ν de la covariance de Matérn en gardant $\rho = 0.4$ et $\sigma_0^2 = 1.0$. L'expérience est en dimension 1.

conclusion, certains types d'erreur sur la covariance n'ont aucune influence sur la qualité des prédictions.

Généralisation

Même si la covariance utilisée n'est pas correcte, nous souhaitons que le prédicteur du krigeage converge vers $F(\boldsymbol{x})$ quand le nombre des observations tend vers l'infini. Si $F(\boldsymbol{x}) - \hat{F}(\boldsymbol{x})$ converge vers zéro en moyenne quadratique pour une covariance donnée $k_1(\boldsymbol{x}, \boldsymbol{y})$, il s'agit de chercher des conditions pour que cette quantité converge aussi vers zéro en moyenne quadratique sous une autre covariance $k_0(\boldsymbol{x}, \boldsymbol{y})$ (considérée comme la vraie covariance). Notons que les conditions cherchées sur les covariances k_0 et k_1 sont plus faibles que celles qu'il faudrait vérifier pour que toute suite d'éléments convergeant vers zéro en moyenne quadratique sous une covariance converge également vers zéro sous l'autre. En effet, le meilleur prédicteur linéaire de $F(\boldsymbol{x})$ possède des propriétés spécifiques. Cette remarque est justifiée par le résultat suivant, fondé sur la théorie des distributions.

Lemme 3. *Soit $k_1(\boldsymbol{h})$ une covariance continue* stationnaire *admettant une densité spectrale $d\xi(\boldsymbol{u})/d\boldsymbol{u}$. Supposons qu'il existe des constantes positives q et C telles que*

$$\liminf_{\|\boldsymbol{u}\| \to \infty} \quad \|\boldsymbol{u}\|^q \left\| \frac{d\xi(\boldsymbol{u})}{d\boldsymbol{u}} \right\| > C. \tag{5.8}$$

Soient un domaine compact $\mathbb{X}$ de $\mathbb{R}^d$, une suite $(\boldsymbol{x}_i)_{i \in \mathbb{N}} \in \mathbb{X}^{\mathbb{N}}$, et un point d'adhérence $\boldsymbol{x}$ de $(\boldsymbol{x}_i)$.

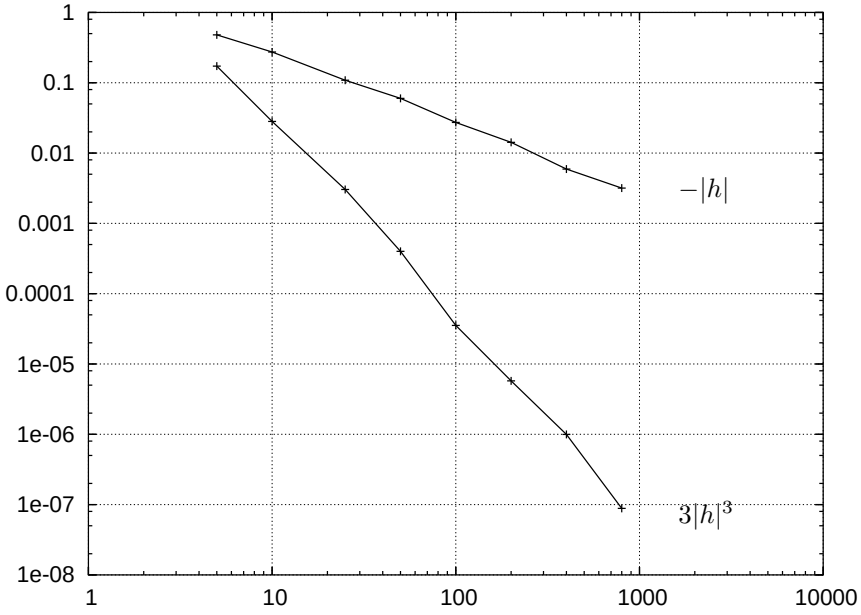


FIG. 5.3 – Comportement des quantiles à 90% des variances d'erreur de prédiction en fonction du nombre de points tirés aléatoirement pour différentes régularités : nous utilisons un krigeage intrinsèque d'ordre 1 et des covariances généralisées polynomiales d'ordre zéro et un. Expérience en dimension 1.

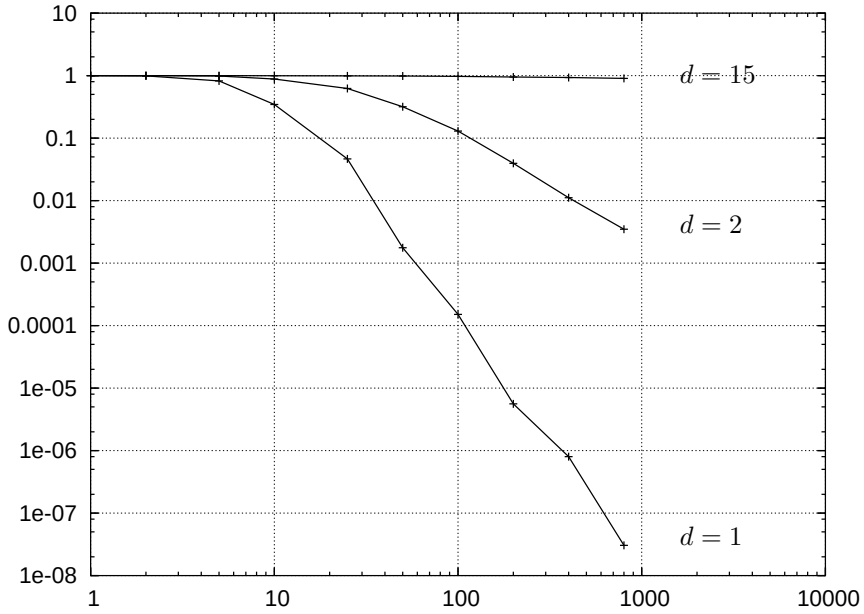


FIG. 5.4 – Comportement des quantiles à 90% des variances d'erreur de prédiction en fonction du nombre de points tirés aléatoirement, pour différentes dimensions d de l'espace des facteurs. Expérience menée avec une covariance de Matérn pour $\nu = 2.0$, $\rho = 0.4$ et $\sigma_0^2 = 1.0$.

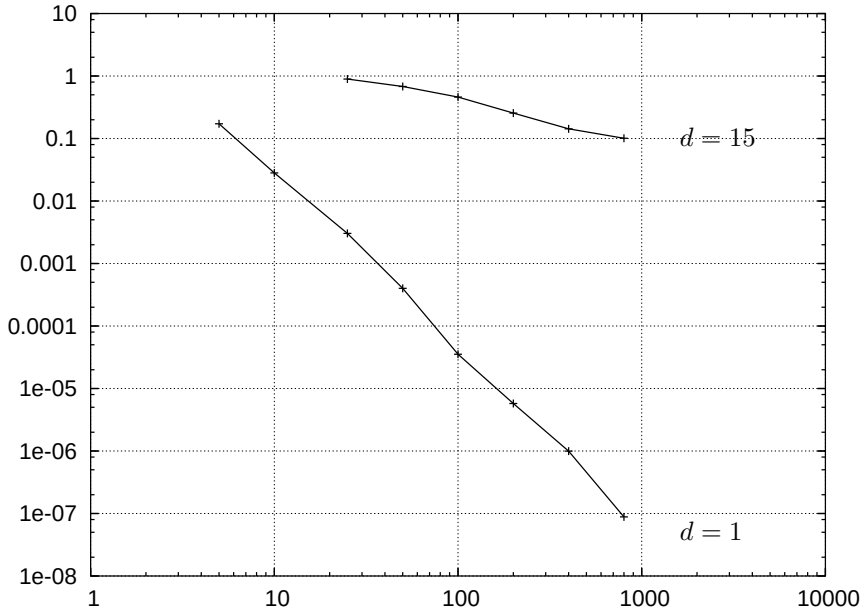


FIG. 5.5 – Comportement des quantiles à 90% des variances d'erreur de prédiction en fonction du nombre de points tirés aléatoirement, pour différentes dimensions de l'espace des facteurs. Expérience menée avec une covariance généralisée polynomiale $|h|^3$.

Considérons la mesure à support fini $\lambda_{S_n,\boldsymbol{x}} = \sum_{i=1}^{n} \lambda_{i,\boldsymbol{x}} \delta_{\boldsymbol{x}_i}$*, où les* $\lambda_{i,\boldsymbol{x}}$ *sont les coefficients de la meilleure prédiction linéaire de* $F(\boldsymbol{x})$ *à partir des* $F(\boldsymbol{x}_i)$*,* $i = 1, \ldots, n$*, obtenue avec la covariance* k_1*. Alors,* $\lambda_{S_n,\boldsymbol{x}}$ *converge au sens faible vers* $\delta_{\boldsymbol{x}}$ *dans l'espace des distributions* $\mathcal{S}'(\mathbb{R}^d)$*, où* $\mathcal{S}(\mathbb{R}^d)$ *désigne conventionnellement l'espace des fonctions à décroissance rapide.*

Démonstration. Voir Yakowitz et Szidarovszky (1985). □

Ce lemme indique que, sous certaines hypothèses, les poids affectés par le prédicteur $\hat{F}(\boldsymbol{x})$ aux variables aléatoires $F(\boldsymbol{x}_i)$ se concentrent asymptotiquement autour du point $\boldsymbol{x}$, lorsque les observations deviennent denses dans un voisinage de $\boldsymbol{x}$. Si nous effectuons des expériences numériques, par exemple avec des covariances de Matérn stationnaires, nous constatons effectivement que les poids du prédicteur linéaire se concentrent sur les observations les plus proches de $\boldsymbol{x}$, ce qui est d'ailleurs le comportement que l'on attend intuitivement. Remarquons que l'hypothèse (5.8) s'applique aux covariances usuelles, sauf les covariances gaussiennes. Que se passe-t-il lorsqu'on utilise une covariance gaussienne ? Dans ce cas, nous constaterons numériquement au paragraphe suivant que $\delta_{S_n,\boldsymbol{x}}$ ne converge pas vers $\delta_{\boldsymbol{x}}$ (voir la figure 5.8). Concluons cette section par la proposition suivante (Yakowitz et Szidarovszky, 1985).

Proposition 53. *Soit* $F(\boldsymbol{x})$*,* $\boldsymbol{x} \in \mathbb{R}^d$*, un processus aléatoire du second ordre, de moyenne nulle, de covariance* $k_0(\boldsymbol{x}, \boldsymbol{y})$ *continue et possédant des trajectoires continues avec probabilité 1. Soient un domaine compact* $\mathbb{X}$ *de* $\mathbb{R}^d$*, une suite* $(\boldsymbol{x}_i) \in \mathbb{X}^{\mathbb{N}}$*, et un point d'adhérence* $\boldsymbol{x}$ *de* $(\boldsymbol{x}_i)$*. Soit* $k_1(\boldsymbol{x}, \boldsymbol{y})$ *une seconde covariance continue satisfaisant la condition (5.8). Alors le prédicteur linéaire* $\hat{F}_n(\boldsymbol{x})$ *de* $F(\boldsymbol{x})$ *à partir des* $F(\boldsymbol{x}_i)$*,* $i = 1, \ldots, n$*, optimal sous la covariance* $k_1(\boldsymbol{x}, \boldsymbol{y})$ *est consistant.*

Démonstration (différente de celle de Yakowitz et Szidarovszky (1985)). Soit $F(\omega, \boldsymbol{x})$ une trajectoire quelconque. Montrons que $\hat{F}_n(\omega, \boldsymbol{x}) = \langle \lambda_{S_n,\boldsymbol{x}}, F(\omega, \boldsymbol{x}) \rangle$ tend vers $F(\omega, \boldsymbol{x})$ lorsque n tend vers l'infini. Soit $\tilde{f}$ une fonction de $C_c(\mathbb{R}^d)$, l'espace des fonctions continues à support compact, coïncidant avec $F(\omega, \boldsymbol{x})$ sur $\mathbb{X}$. Comme $\mathcal{S}(\mathbb{R}^d)$ est dense dans $C_c(\mathbb{R}^d)$, on a d'après le lemme 3 $\lim_{n\to\infty} \hat{F}_n(\omega, \boldsymbol{x}) = \lim_{n\to\infty} \langle \lambda_{S_n,\boldsymbol{x}}, \tilde{f} \rangle = \tilde{f}(\boldsymbol{x}) = F(\omega, \boldsymbol{x})$. □

Cette proposition montre donc qu'une prédiction linéaire est consistante pour n'importe quel choix de noyau satisfaisant la condition (5.8). Cela signifie-t-il que l'on ne devrait pas se préoccuper du noyau utilisé ? En pratique, on constate effectivement que le choix du noyau n'est pas fondamental lorsque l'on dispose de beaucoup d'observations. Il vaut alors mieux privilégier des noyaux conduisant à des algorithmes de prédiction rapide, comme par exemple des noyaux à support compact. Cependant, lorsque l'on construit un modèle boîte noire d'un système en dimension élevée, il est rare de disposer d'un nombre de données suffisamment important pour que ceci soit vrai. Dans la plupart des applications, les variances des erreurs de prédiction *ne sont pas négligeables*. Il est alors essentiel d'obtenir des prédicteurs optimaux, et donc de ne pas sous-évaluer l'importance du choix du noyau.

Inconsistances numériques

Nous présentons une série d'expériences montrant que l'utilisation d'une covariance gaussienne peut conduire à des prédictions peu satisfaisantes en terme de consistance. Soit un processus aléatoire gaussien $F(x)$, $x \in \mathbb{R}$, de moyenne nulle et de covariance stationnaire $k_0(h)$, où $k_0(h)$

est une fonction de Matérn. Ici, nous prenons $\nu = 2$, ce qui implique que la covariance est quatre fois continûment dérivable. Les autres paramètres sont arbitrairement choisis unitaires. Soit une seconde covariance stationnaire $k_1(h) = \sigma_0^2 \exp(-(h/\rho)^2)$ (covariance dite gaussienne). Ces deux covariances sont représentées à la figure 5.6. La covariance k_1 est utilisée pour prédire $F(x)$. La figure 5.7 montre un résultat surprenant : la prédiction obtenue ne se comporte pas comme une interpolation. Plus précisément, les valeurs prédites sont éloignées des valeurs observées alors que l'échantillonnage est dense par rapport à la distance caractéristique des variations de $F(x)$. Les résultats suivants suggèrent que ce phénomène est lié au conditionnement numérique de la matrice de covariance issue de k_1, qui n'est pas favorable en raison de l'échantillonnage choisi. Ajoutons à la covariance $k_1(h)$ une composante de bruit blanc, ou suivant la terminologie des géostatisticiens un *effet de pépite*, de sorte que

$$k_1(h) = \sigma_0^2 \exp\left(- \left(\frac{h}{\rho} \right)^2 \right) + \varepsilon \delta(h) \,, \tag{5.9}$$

où la fonction $\delta(h) = 1$ en $h = 0$ et vaut zéro ailleurs. En utilisant $k_1(h)$ pour prédire $F(0)$ et en faisant varier ε, regardons les variances sous les modèles k_0 et k_1 de l'erreur de prédiction. D'après les résultats présentés dans la table 5.1, la matrice de covariance gaussienne est mal conditionnée lorsque ε est petit. La variance d'erreur sous le modèle k_1 décroît très rapidement avec N pour atteindre la variance ε du bruit. La variance sous le vrai modèle k_0 tend à décroître lorsque N augmente mais reste importante quelle que soit la valeur de ε, ce qui explique les résultats de la figure 5.7. La prédiction est *numériquement inconsistante*. On pourrait même parler d'un phénomène d'amplification du bruit du fait de l'utilisation de la covariance gaussienne à la place de la covariance de Matérn. La figure 5.8 représente le vecteur $\boldsymbol{\lambda}_0$ en fonction des points échantillonnés pour $\varepsilon = 10^{-12}$. Nous constatons que la prédiction réalise un moyennage des valeurs observées sur un voisinage largement *étendu* autour du point zéro. Cela met en évidence le fait que la prédiction n'est pas consistante. Si elle l'était, on devrait en effet constater que la valeur prédite est une moyenne des observations les *plus proches* lorsque le bruit est négligeable.

5.3.3 Notion d'efficacité asymptotique

Dans une série d'articles (Stein, 1988, 1990a,b), M.L. Stein introduit et étudie la notion d'efficacité asymptotique. L'objectif est de comparer les vitesses de décroissance asymptotique des erreurs de prédiction lorsque la densité des observations augmente et que l'on utilise une covariance incorrecte. Les résultats obtenus montrent que la notion d'efficacité asymptotique est liée à une notion de *compatibilité* entre les covariances en jeu.

Compatibilité entre covariances

Définition 39. Soit $F(\boldsymbol{x})$, $\boldsymbol{x} \in \mathbb{X}$, un processus aléatoire du second ordre, de moyenne $m_0(\boldsymbol{x})$ et de covariance $k_0(\boldsymbol{x}, \boldsymbol{y})$ continues. Soit $m_1(\boldsymbol{x})$ et $k_1(\boldsymbol{x}, \boldsymbol{y})$ une autre structure de second ordre. $e_i(F(\boldsymbol{x}), n) = F(\boldsymbol{x}) - \hat{F}_n^i(\boldsymbol{x})$ désigne l'erreur de prédiction linéaire optimale de $F(\boldsymbol{x})$ en fonction des variables aléatoires $F(\boldsymbol{x}_1), \ldots, F(\boldsymbol{x}_n)$, calculée sous le modèle du second ordre (m_i, k_i), $i \in \{0, 1\}$. E_0 désigne l'espérance sous le modèle (m_0, k_0). Soient $(\boldsymbol{x}_i)_{i \in \mathbb{N}}$ une suite de $\mathbb{X}$, et $\boldsymbol{x}$ un point d'adhérence de cette suite. Le prédicteur linéaire optimal $\hat{F}_n^1(\boldsymbol{x})$ calculé sous (m_1, k_1) à partir des

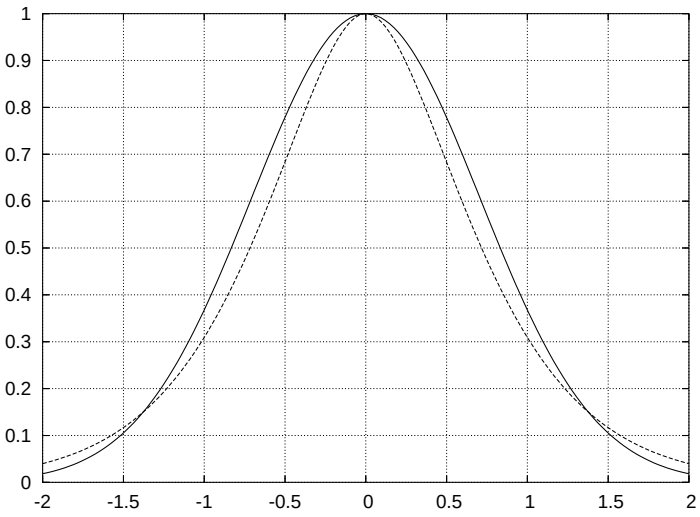


FIG. 5.6 – Covariances de Matérn (en trait discontinu) et gaussienne (en trait plein).

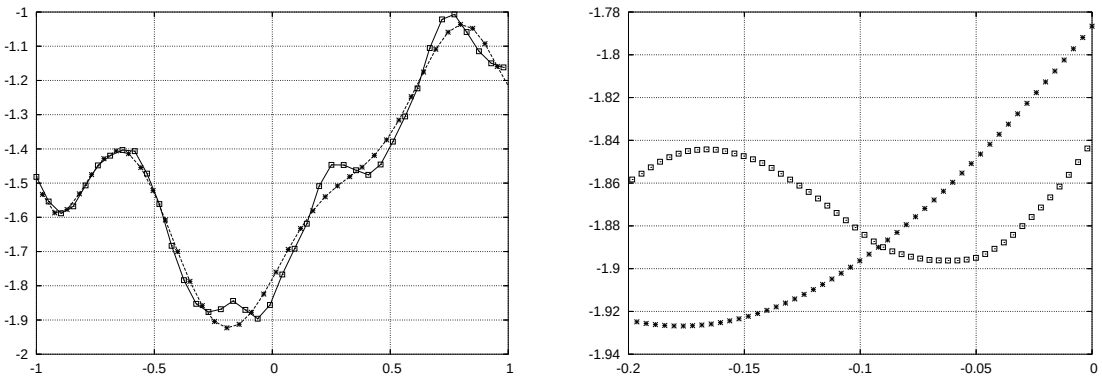


FIG. 5.7 – Une réalisation de $F(x)$ est simulée sur l'intervalle $[-1, 1]$ en choisissant un échantillonnage régulier $x_i = (2/n) \cdot (i - n/2)$, $i = 1, \ldots, n$. Les observations simulées sont représentées par des carrés reliés en trait plein. Les valeurs de $F(x)$ entre les échantillons observés sont prédites aux points $\frac{x_{i+1}+x_i}{2}$ à l'aide la covariance gaussienne. Les valeurs prédites sont représentées par des croix reliées en trait discontinu. Sur la figure de gauche, $n = 40$. Le conditionnement de la matrice de covariance est de l'ordre de 10^{19}. Sur la figure de droite, $n = 500$ (l'échantillonnage est plus dense mais nous avons utilisé la même simulation que dans le cas $n = 40$). Pour $n = 500$, le conditionnement de la matrice de covariance est de l'ordre de 10^{20}. Nous constatons dans les deux cas que les valeurs prédites sont aberrantes pour une interpolation.

$\varepsilon = 10^{-16}$

$n = 10$	$n = 25$	$n = 50$	$n = 100$
$1.1 \cdot 10^{9}$	$2.0 \cdot 10^{18}$	$3.1 \cdot 10^{18}$	$2.7 \cdot 10^{19}$
$1.7 \cdot 10^{-11}$	$2.2 \cdot 10^{-16}$	$1.5 \cdot 10^{-15}$	$4.4 \cdot 10^{-16}$
$1.8 \cdot 10^{-3}$	$2.5 \cdot 10^{-4}$	$2.5 \cdot 10^{-4}$	$2.5 \cdot 10^{-4}$

$\varepsilon = 10^{-12}$

$n = 10$	$n = 25$	$n = 50$	$n = 100$	$n = 200$	$n = 400$	$n = 800$
$1.1 \cdot 10^{9}$	$1.6 \cdot 10^{13}$	$3.2 \cdot 10^{13}$	$6.5 \cdot 10^{13}$	$1.4 \cdot 10^{14}$	$3.0 \cdot 10^{14}$	$9.5 \cdot 10^{14}$
$1.9 \cdot 10^{-11}$	$1.4 \cdot 10^{-12}$	$1.2 \cdot 10^{-12}$	$1.1 \cdot 10^{-12}$	$1.0 \cdot 10^{-12}$	$1.0 \cdot 10^{-12}$	$1.0 \cdot 10^{-12}$
$1.8 \cdot 10^{-3}$	$4.5 \cdot 10^{-4}$	$4.5 \cdot 10^{-4}$	$4.4 \cdot 10^{-4}$	$4.2 \cdot 10^{-4}$	$4.1 \cdot 10^{-4}$	$3.9 \cdot 10^{-4}$

$\varepsilon = 10^{-10}$

$n = 10$	$n = 25$	$n = 50$	$n = 100$	$n = 200$	$n = 400$	$n = 800$
$1.1 \cdot 10^{9}$	$1.6 \cdot 10^{11}$	$3.2 \cdot 10^{11}$	$6.5 \cdot 10^{11}$	$1.3 \cdot 10^{12}$	$2.6 \cdot 10^{12}$	$5.2 \cdot 10^{12}$
$2.0 \cdot 10^{-10}$	$1.3 \cdot 10^{-10}$	$1.2 \cdot 10^{-10}$	$1.1 \cdot 10^{-10}$	$1.0 \cdot 10^{-10}$	$1.0 \cdot 10^{-10}$	$1.0 \cdot 10^{-10}$
$1.8 \cdot 10^{-3}$	$7.6 \cdot 10^{-4}$	$7.3 \cdot 10^{-4}$	$6.7 \cdot 10^{-4}$	$6.1 \cdot 10^{-4}$	$5.5 \cdot 10^{-4}$	$5.0 \cdot 10^{-4}$

TAB. 5.1 – Trois tableaux présentant le conditionnement de la matrice de covariance (première ligne), et les variances d'erreur de prédiction au point 0 calculées sous la covariance k_1 (deuxième ligne) et k_0 (troisième ligne). Nous faisons varier le nombre n de points utilisés pour la prédiction ainsi que ε ($\varepsilon = 10^{-16}$ dans le premier tableau, $\varepsilon = 10^{-12}$ dans le deuxième, $\varepsilon = 10^{-10}$ dans le dernier). La variance d'erreur de prédiction calculée sous k_1 tend vers ε lorsque $n \to \infty$ mais la variance d'erreur sous k_0 est très supérieure en raison du mauvais conditionnement des matrices de covariance

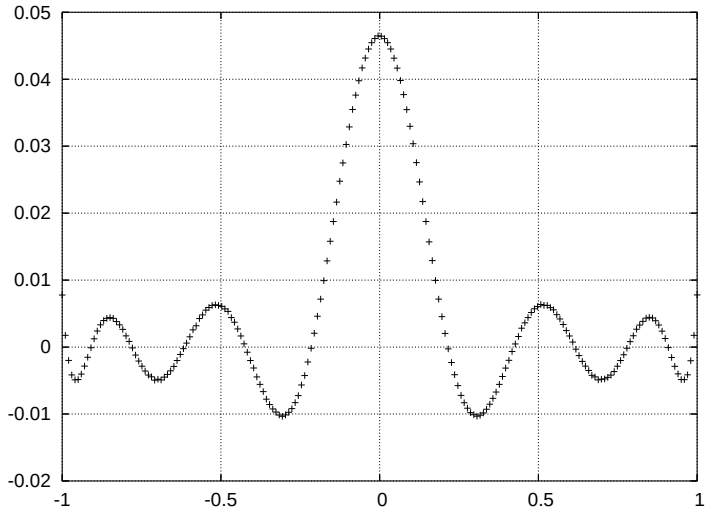


FIG. 5.8 – Coefficients du krigeage $\lambda_{i,0}$ en fonction des points d'observation $x_i \in [-1, 1]$, $i = 1, \ldots, 200$. $\varepsilon = 10^{-12}$. Remarque : $\sum_i \lambda_{i,0} = 1$

variables aléatoires $F(\boldsymbol{x}_1), \ldots, F(\boldsymbol{x}_n)$ est dit *asymptotiquement efficace* si

$$\lim_{n\to\infty} \frac{\mathrm{E}_0[e_1(F(\boldsymbol{x}), n)^2]}{\mathrm{E}_0[e_0(F(\boldsymbol{x}), n)^2]} = 1\,, \tag{5.10}$$

Par la suite, nous supposons les moyennes nulles et regardons seulement l'influence de la covariance. Les notions d'équivalence et d'orthogonalité de mesures sont fondamentales pour définir la notion de compatibilité entre covariances.

Définition 40. Soit P_0 et P_1 deux mesures de probabilité sur une σ-algèbre $\mathcal{A}$. Si $\mathsf{P}_1(B) = 0$ pour tout $B \in \mathcal{B}$ tel que $\mathsf{P}_0(B) = 0$, P_1 est *absolument continue* par rapport à P_0, et nous écrivons alors $\mathsf{P}_0 \leq \mathsf{P}_1$. P_0 et P_1 sont *équivalentes* si $\mathsf{P}_1 \leq \mathsf{P}_0$ et $\mathsf{P}_0 \leq \mathsf{P}_1$, et nous écrivons alors $\mathsf{P}_0 \equiv \mathsf{P}_1$. P_0 et P_1 sont *mutuellement singulières* ou bien *orthogonales*, s'il existe deux ensembles disjoints $B_0, B_1 \in \mathcal{A}$ tels que P_0 soit portée par B_0 ($\mathsf{P}_0(B_0) = 1$) et P_1 soit portée par B_1 ($\mathsf{P}_1(B_1) = 1$), et nous écrivons alors $\mathsf{P}_0 \perp \mathsf{P}_1$.

Si deux variables aléatoires suivent des lois P_0 et P_1 telles que $\mathsf{P}_0 \perp \mathsf{P}_1$, il existe un événement de $\mathcal{A}$ permettant de déterminer avec certitude quelle variable aléatoire a été observée.

Par la suite, $\mathbb{X}$ est un domaine *borné* de $\mathbb{R}^d$. Soit $F(\omega, \boldsymbol{x})$, $\boldsymbol{x} \in \mathbb{X}$, un processus aléatoire *gaussien* défini sur l'espace de probabilités $(\Omega, \mathcal{A}, \mathsf{P}_0)$. P_0 est alors une probabilité *gaussienne* sur la σ-algèbre $\sigma(F(\mathbb{X}))$. Soit P_1 une autre probabilité gaussienne définie sur $\sigma(F(\mathbb{X}))$. Comme un processus aléatoire gaussien est uniquement déterminé par sa structure du second ordre, les mesures de probabilité gaussiennes P_0 et P_1 sont déterminées sur $\sigma(F(\mathbb{X}))$ par des structures $(m_0(\boldsymbol{x}), k_0(\boldsymbol{x}, \boldsymbol{y}))$ et $(m_1(\boldsymbol{x}), k_1(\boldsymbol{x}, \boldsymbol{y}))$. Si $\mathsf{P}_0 \perp \mathsf{P}_1$, comment décrire explicitement les événements disjoints B_0 et B_1, supports des mesures P_0 et P_1 ? (Ibragimov et Rozanov, 1978) montre que $\mathsf{P}_0 \perp \mathsf{P}_1$ s'il existe une suite d'évènements $B_1, B_2, \ldots \in \mathcal{A}$ telle que

$$\lim_{n\to\infty} \mathsf{P}_0(B_n) = 0 \text{ et } \lim_{n\to\infty} \mathsf{P}_1(B_n) = 1 \tag{5.11}$$

Ce résultat sert à caractériser l'orthogonalité en termes des structures du second ordre possibles de $F(\boldsymbol{x})$. S'il existe une suite $F(\lambda_n) = \sum_{i=1}^{l_n} \lambda_{i,n} F(\boldsymbol{x}_{i,n})$ telle que

$$\lim_{n\to\infty} \frac{\mathrm{Var}_1[F(\lambda_n)]}{\mathrm{Var}_0[F(\lambda_n)]} = 0 \quad (\text{ou } \infty),$$

ou encore

$$\lim_{n\to\infty} \frac{\mathrm{E}_1[F(\lambda_n)] - \mathrm{E}_0[F(\lambda_n)]}{\mathrm{Var}_0[F(\lambda_n)]} = \infty,$$

alors la propriété (5.11) permet d'établir $\mathsf{P}_0 \perp \mathsf{P}_1$ (Ibragimov et Rozanov, 1978). Par exemple, si $F(\boldsymbol{x})$ est stationnaire et dérivable en moyenne quadratique sous k_0 mais pas sous k_1, $\lim_{\boldsymbol{x}_n \to \boldsymbol{x}} \mathrm{Var}[(F(\boldsymbol{x}) - F(\boldsymbol{x}_n))/\|\boldsymbol{x} - \boldsymbol{x}_n\|]$ est finie sous k_0 mais infinie sous k_1, ce qui implique $\mathsf{P}_0 \perp \mathsf{P}_1$. Nous présentons ci-dessous la notion de compatibilité entre covariances introduite par M.L. Stein.

Définition 41. Deux covariances k_0 et k_1 sont *compatibles sur* $\mathbb{X}$ si les mesures de probabilité gaussiennes correspondantes aux structures du second ordre $(0, k_0)$ et $(0, k_1)$ sont équivalentes.

Deux probabilités gaussiennes sont équivalentes ou orthogonales (Ibragimov et Rozanov, 1978). Par conséquent, dans le cas gaussien, les conditions suffisantes d'orthogonalité sont des conditions

nécessaires d'équivalence. Ainsi, si deux covariances stationnaires ont des ordres de dérivabilité différents, elles ne peuvent pas être compatibles. Il est assez difficile d'établir des conditions suffisantes pour garantir la compatibilité de deux covariances sur un domaine $\mathbb{X}$ donné (voir à ce propos la discussion dans (Stein, 1999)). Certaines ont été données dans (Ibragimov et Rozanov, 1978). Nous nous contenterons de retenir le concept intuitif suivant : pour deux covariances soient compatibles, il faut qu'elles se comportent de la même façon à l'origine, ce que nous vérifions plus loin à l'aide d'expériences numériques.

Résultat principal

Soit $F(\omega, \boldsymbol{x})$, $\omega \in \Omega$, $\boldsymbol{x} \in \mathbb{R}^d$, un processus aléatoire gaussien de moyenne nulle (pour simplifier) et de covariance $k_0(\boldsymbol{x}, \boldsymbol{y})$. Remarquons que $F(\boldsymbol{x})$ est généralement défini sur $\mathbb{R}^d$ mais le résultat établi par M.L. Stein concerne un domaine d'étude *borné* $\mathbb{X} \subset \mathbb{R}^d$. Notons $\mathcal{H}_0$ l'espace généré par $F(\boldsymbol{x})$ lorsque $\boldsymbol{x}$ parcourt $\mathbb{X}$ borné, complété par rapport au produit scalaire k_0. Notons également $\mathcal{S}_0$ la σ-algèbre engendrée sur Ω par les éléments de $\mathcal{H}_0$. Supposons qu'il existe une suite de variables aléatoires linéairement indépendantes $(F(\boldsymbol{x}_i))_{i\in\mathbb{N}}$, dense dans $\mathcal{H}_0$. Notons $\mathcal{H}_{0,n}$ le sous-espace vectoriel $\mathrm{vect}\{F(\boldsymbol{x}_i), i = 1, \ldots, n\}$ et $\mathcal{H}_{0,-n}$ le sous-espace $\mathcal{H}_0 \backslash \mathcal{H}_{0,n}$. Tout élément $X \in \mathcal{H}_0$ peut donc s'écrire comme la somme d'un élément $X_{|n}$ de $\mathcal{H}_{0,n}$ (le passé) et $X_{|-n}$ de $\mathcal{H}_{0,-n}$ (le futur). Les éléments du passé $\mathcal{H}_{0,n}$ engendrent la σ-algèbre $\mathcal{S}_{0,n}$ sur Ω, et les éléments de futur $\mathcal{H}_{0,-n}$ engendrent la σ-algèbre $\mathcal{S}_{0,-n}$.

Définition 42. Une probabilité P est *prédictive* sur $(\Omega, \mathcal{S})$ si pour tout $n \geq 1$, il existe une probabilité conditionnelle P_{-n} sur le futur $\mathcal{H}_{-n}$ étant donné le passé. Plus précisément, il existe une fonction $\mathsf{P}_{-n}(X_{|n})(B)$, où $X_{|n} \in \mathcal{H}_n$ et $B \in \mathcal{S}_{-n}$ avec les propriétés suivantes :

- $\mathsf{P}_{-n}(X_{|n})(B)$ est $\mathcal{S}_n$-mesurable pour B fixé ;
- $\mathsf{P}_{-n}(X_{|n}(\omega))(B)$ est une probabilité de distribution sur $\mathcal{S}_{-n}$ lorsque ω est fixé ;
- pour toute variable aléatoire $X \in \mathcal{H}_0$,

$$\int X d\mathsf{P} = \int_\omega \left(\int_{\omega'} X d\mathsf{P}_{-n}(X_{|n}(\omega))(\omega') \right) d\mathsf{P}_n(\omega) \,,$$

où P_n est la probabilité marginale de P sur $\mathcal{S}_n$.

Le résultat principal de M.L. Stein sur l'efficacité asymptotique est une conséquence du théorème fondamental suivant :

Théorème 26 (« Main Theorem » de Blackwell et Dubins). *Soit P_0 une probabilité prédictive. Considérons une probabilité P_1 absolument continue par rapport à P_0. Alors pour toute probabilité $\mathsf{P}_{0,-n}$ sur le futur conditionnellement au passé,* il existe *une probabilité conditionnelle $\mathsf{P}_{1,-n}$ définie par rapport à P_1, telle qu'à l'exception d'un sous-ensemble de Ω de P_1-probabilité nulle, la distance*[3] *entre $\mathsf{P}_{0,-n}$ et $\mathsf{P}_{1,-n}$ converge vers 0 quand n tend vers l'infini.*

Démonstration. Voir (Blackwell et Dubins, 1962). □

[3] La distance entre deux probabilités P_0 et P_1 est définie ici comme le $\sup_B |\mathsf{P}_1(B) - \mathsf{P}_0(B)|$

Théorème 27 (Efficacité asymptotique, (Stein, 1990a,b, 1999)). *Soient k_0 et k_1 deux covariances compatibles. Notons que $\mathcal{H}_{0,-n}$ est tel que $\forall X \in \mathcal{H}_{0,-n}$, $\mathrm{E}_0[e_0(X,n)^2] > 0$. Alors*

$$\lim_{n\to\infty} \sup_{X\in\mathcal{H}_{0,-n}} \left| \frac{\mathrm{E}_1[e_0(X,n)^2] - \mathrm{E}_0[e_0(X,n)^2]}{\mathrm{E}_0[e_0(X,n)^2]} \right| = 0 \tag{5.12}$$

$$\lim_{n\to\infty} \sup_{X\in\mathcal{H}_{0,-n}} \left| \frac{\mathrm{E}_0[e_1(X,n)^2] - \mathrm{E}_1[e_1(X,n)^2]}{\mathrm{E}_1[e_1(X,n)^2]} \right| = 0 \tag{5.13}$$

$$\lim_{n\to\infty} \sup_{X\in\mathcal{H}_{0,-n}} \frac{\mathrm{E}_0[e_1(X,n)^2] - \mathrm{E}_0[e_0(X,n)^2]}{\mathrm{E}_0[e_0(X,n)^2]} = 0 \tag{5.14}$$

$$\lim_{n\to\infty} \sup_{X\in\mathcal{H}_{0,-n}} \frac{\mathrm{E}_1[e_0(X,n)^2] - \mathrm{E}_1[e_1(X,n)^2]}{\mathrm{E}_1[e_1(X,n)^2]} = 0 \tag{5.15}$$

Notons que (5.14) est équivalente à

$$\lim_{n\to\infty} \sup_{X\in\mathcal{H}_{0,-n}} \frac{\mathrm{E}_0[e_1(X,n)^2]}{\mathrm{E}_0[e_0(X,n)^2]} = 1\,,$$

qui implique la propriété d'efficacité asymptotique définie ci-dessus en prenant $X = F(\boldsymbol{x})$.

Démonstration partielle. L'application du théorème de Blackwell et Dubins implique que

$$\mathsf{P}_1\left\{ \lim_{n\to\infty} \sup_{X\in\mathcal{H}_0, s\in\mathbb{R}} \left| \mathsf{P}_{0,-n}(X_{|n})\{X_{|-n} \le s\} - \mathsf{P}_{1,-n}(X_{|n})\{X_{|-n} \le s\} \right| = 0 \right\} = 1 \tag{5.16}$$

En utilisant le fait que la distribution de X est gaussienne aussi bien sous P_0 que sous P_1, il est possible de montrer que (5.16) est vérifiée si et seulement si

$$\lim_{n\to\infty} \sup_{X\in\mathcal{H}_{0,-n}} \frac{\mathrm{Var}_0[X \mid \mathcal{S}_{0,n}]}{\mathrm{Var}_1[X \mid \mathcal{S}_{0,n}]} = 0 \tag{5.17}$$

et

$$\lim_{n\to\infty} \sup_{X\in\mathcal{H}_{0,-n}} \frac{\mathrm{E}_1[(\mathrm{E}_0[X \mid \mathcal{S}_{0,n}] - \mathrm{E}_1[X \mid \mathcal{S}_{0,n}])^2]}{\mathrm{Var}_1[X \mid \mathcal{S}_{0,n}]} = 0\,. \tag{5.18}$$

En prenant $X = e_0(X,n)$ dans (5.18), on obtient,

$$\lim_{n\to\infty} \sup_{X\in\mathcal{H}_{0,-n}} \frac{\mathrm{E}_1[(e_0(X,n) - e_1(X,n))^2]}{\mathrm{E}_1[e_1(X,n)^2]} = 0\,,$$

ce qui démontre (5.13). En combinant (5.18) et (5.16), on obtient

$$\lim_{n\to\infty} \sup_{X\in\mathcal{H}_{0,-n}} \frac{\mathrm{E}_0[e_0(X,n)^2] - \mathrm{E}_1[e_1(X,n)^2]}{\mathrm{E}_1[e_0(X,n)^2]} = 0\,. \tag{5.19}$$

□

Le message que nous pouvons retenir est donc le suivant. Si les observations deviennent asymptotiquement denses sur le domaine d'étude, il est inutile de distinguer deux covariances compatibles pour effectuer une prédiction linéaire optimale. Si nous ajoutons que la notion de compatibilité porte *approximativement* sur le comportement de la covariance à l'origine et sur sa régularité,

nous pouvons dire qu'il est important de bien estimer la régularité de la covariance et d'utiliser des modèles de covariance permettant d'ajuster facilement cette régularité. Par la suite, et notamment dans les applications, nous avons *privilégié l'utilisation des covariances de Matérn et des covariances généralisées polynomiales*. À l'inverse, nous avons évité les covariances exponentielles et gaussiennes, ainsi que d'autres covariances classiques utilisées en géostatistique qui ne permettent pas l'ajustement de la régularité. Dans la section suivante, nous présentons des expériences numériques visant à confirmer l'importance de la régularité de la covariance sur les comportements asymptotiques. Ces expériences sont également intéressantes afin de regarder les régimes *non-asymptotiques*.

Expériences numériques

Nous regardons l'influence d'un choix incorrect de covariance sur la variance de l'erreur de prédiction d'un point de vue numérique. Les résultats sont présentés dans les figures 5.9 à 5.14. Le protocole expérimental est le suivant. (Seul le cas d'espace des facteurs de dimension un sera considéré ici.) Nous notons k_0 la vraie covariance et k_1 la covariance incorrecte.
Phase 1

1. Tirage aléatoire d'une séquence finie de points $\boldsymbol{x}_i$, $i = 1, \ldots, n$, avec une distribution uniforme sur l'intervalle $[-1, 1]$.
2. Calcul des vecteurs $\boldsymbol{\lambda}^0$ et $\boldsymbol{\lambda}^1$ des coefficients de la meilleure prédiction linéaire au point $\mathbf{0}$ sous les covariances k_0 et k_1.
3. Calcul des variances σ_0^2 et σ_1^2 des erreurs de prédiction sous k_0 pour $\boldsymbol{\lambda}^0$ et $\boldsymbol{\lambda}^1$.
4. Calcul du rapport d'efficacité $\kappa = \sigma_1^2/\sigma_0^2$.

Phase 2

1. La phase 1 est répétée N fois pour obtenir une suite $\kappa_1, \ldots \kappa_N$. N est choisi suffisamment grand pour obtenir un histogramme des κ_i stable.
2. Les quantiles à 10% et à 90% de la suite (κ_n) sont calculés.

Phase 3

1. La phase 2 est répétée en faisant varier n.

Erreur sur la distance caractéristique de la corrélation. La figure 5.9 décrit l'influence d'une erreur sur la portée de la covariance (comportement basse fréquence). Asymptotiquement, on constate qu'une telle erreur est sans conséquence.

Erreur sur la régularité. Les figures 5.10 à 5.14 décrivent l'influence d'une erreur sur le type de covariance. Asymptotiquement, une telle erreur se révèle sans conséquence quand les covariances ont la même régularité à l'origine, et seulement dans ce cas.

5.3.4 Conclusions

Nous pouvons tirer de cette étude les conclusions suivantes, classées par ordre d'importance croissante :

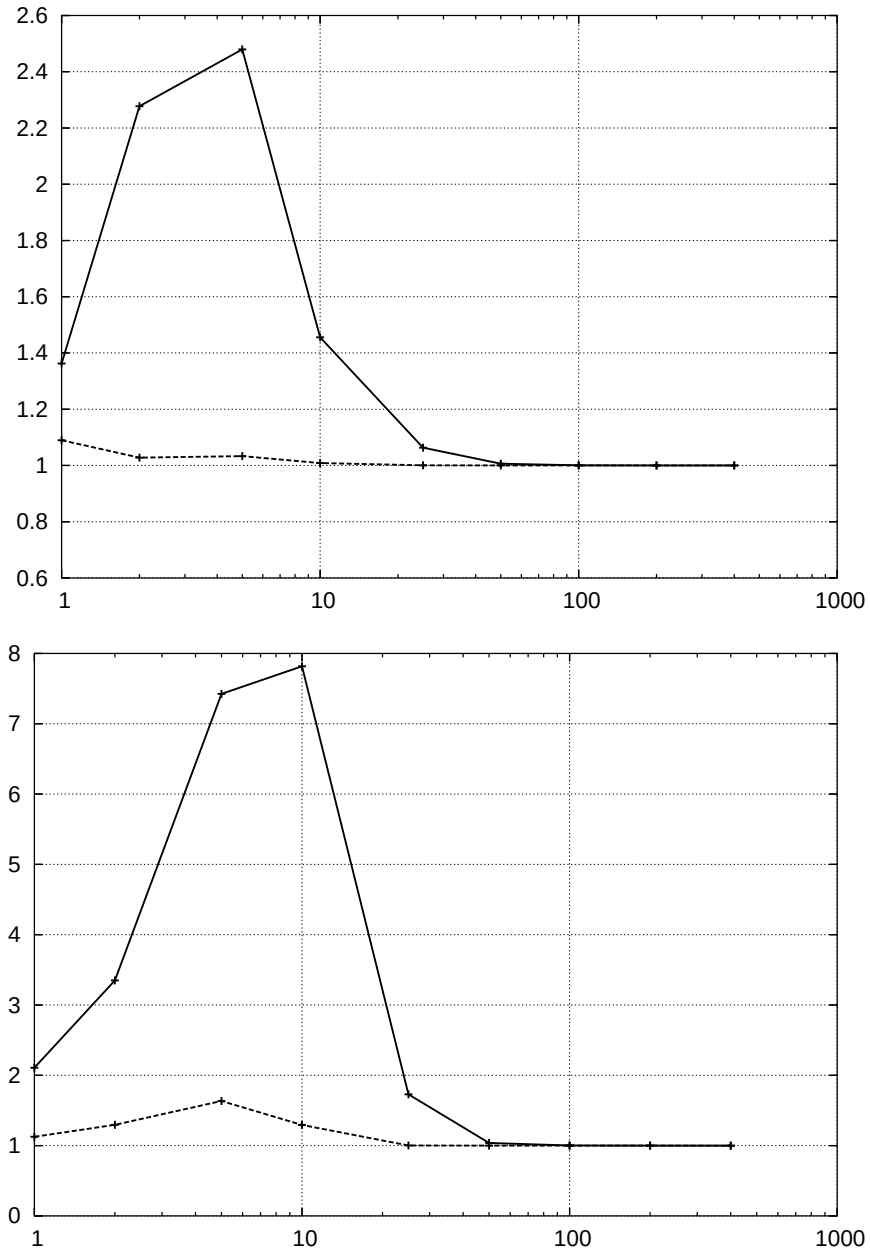


FIG. 5.9 – Influence de la portée de la covariance ; quantiles à 10% et à 90% du rapport d'efficacité d'erreur de prédiction en fonction du nombre de points tirés aléatoirement. L'échelle est logarithmique selon l'axe des abscisses. k_0 et k_1 sont des covariances de Matérn de paramètre $\nu_0 = \nu_1 = 2.0$. Sur le graphique du haut, $\rho_0 = 0.3$ et $\rho_1 = 1.0$. Sur le graphique du bas, $\rho_0 = 1.0$ et $\rho_1 = 0.3$

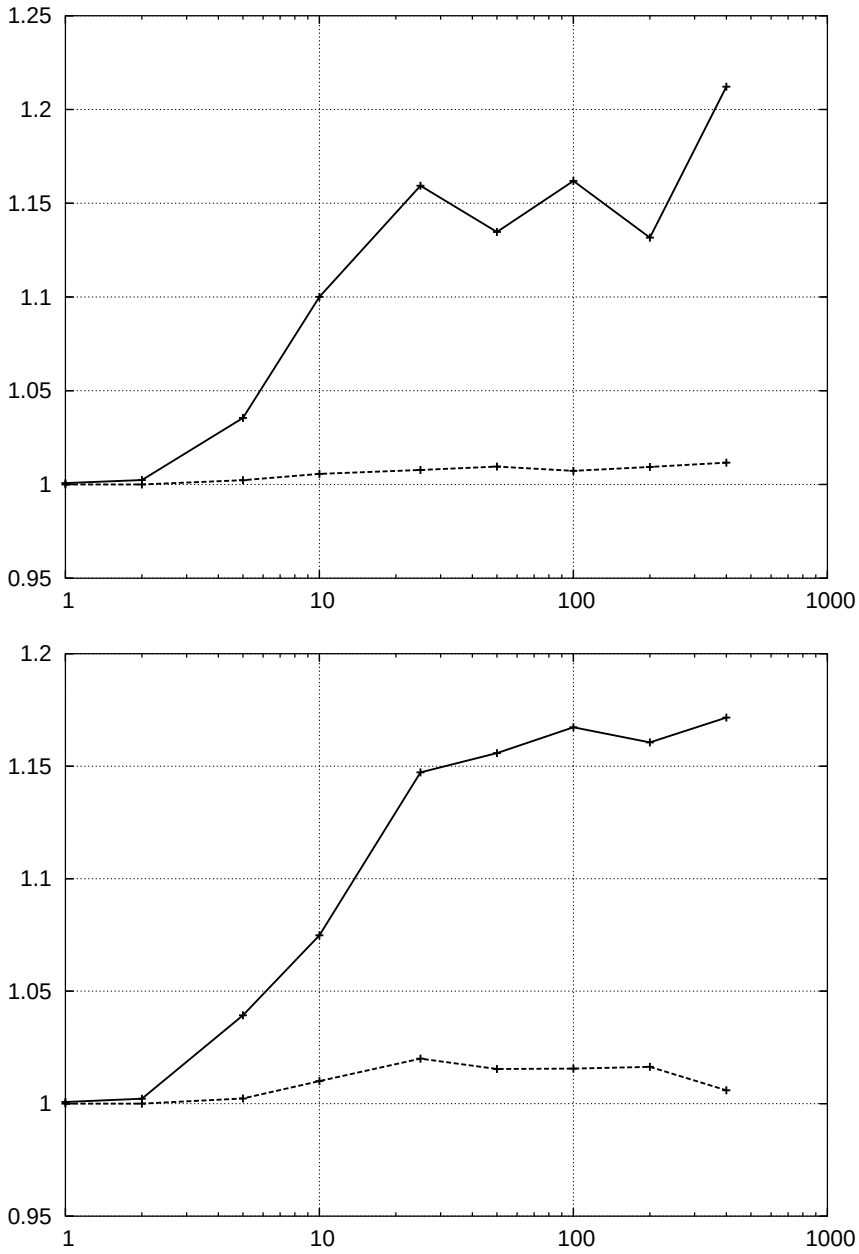


FIG. 5.10 – Influence de la régularité de la covariance ; quantiles à 10% et à 90% du rapport d'efficacité d'erreur de prédiction en fonction du nombre de points tirés aléatoirement. k_0 et k_1 sont des covariances de Matérn de paramètre $\rho_0 = \rho_1 = 0.5$. Sur le graphique du haut, $\nu_0 = 2.0$ et $\nu_1 = 2.5$. Sur le graphique du bas $\nu_0 = 2.5$ et $\nu_1 = 2.0$

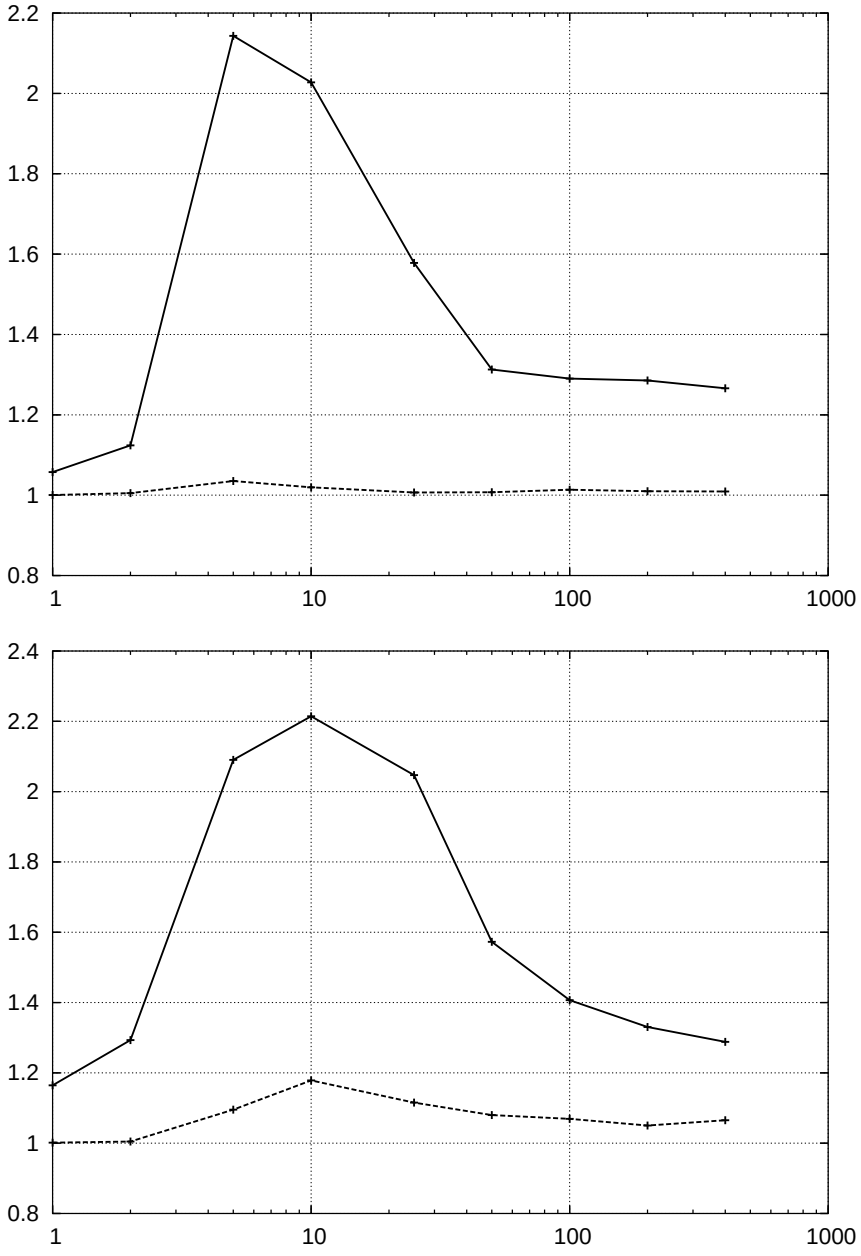


FIG. 5.11 – Influence de la régularité de la covariance ; quantiles à 10% et à 90% du rapport d'efficacité d'erreur de prédiction en fonction du nombre de points tirés aléatoirement. Sur le graphique du haut, k_0 est une covariance de Matérn ($\nu_0 = 0.5$ et $\rho_0 = 0.5$) et k_1 est une covariance exponentielle $\exp(-1.43(h/0.5)^{1.8})$. Sur le graphique du bas, les covariances k_0 et k_1 échangent leurs rôles.

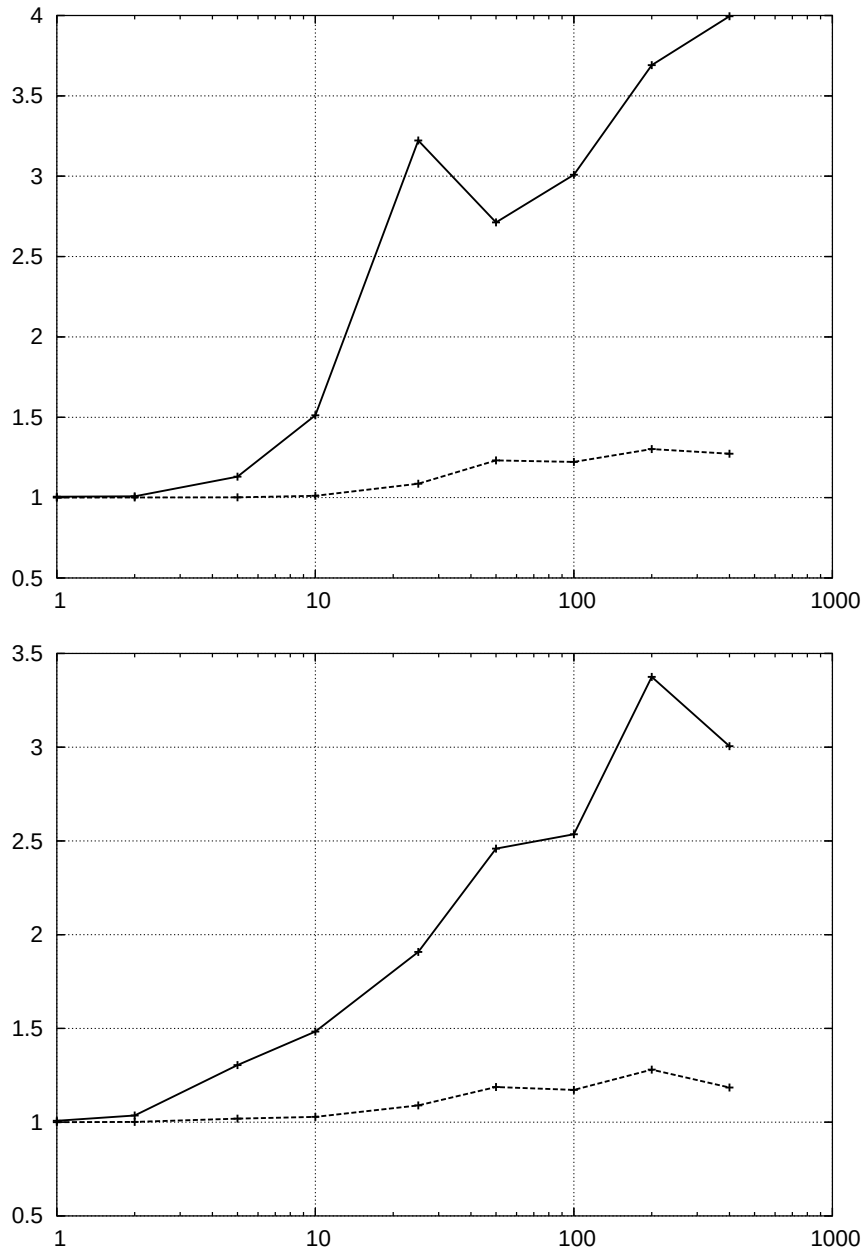


FIG. 5.12 – Influence de la régularité de la covariance; quantiles à 10% et à 90% du rapport d'efficacité d'erreur de prédiction en fonction du nombre de points tirés aléatoirement. k_0 est une covariance de Matérn ($\nu_0 = 2.0$ et $\rho_0 = 0.5$). Sur le graphique du haut k_1 est une covariance exponentielle $\exp(-1.43(h/0.5)^{1.79})$. Sur le graphique du bas, $k_1(h) = \exp(-1.43(h/0.5)^{1.97})$. Bien que k_1 semble très proche de k_0, le rapport d'efficacité asymptotique ne tend pas vers un.

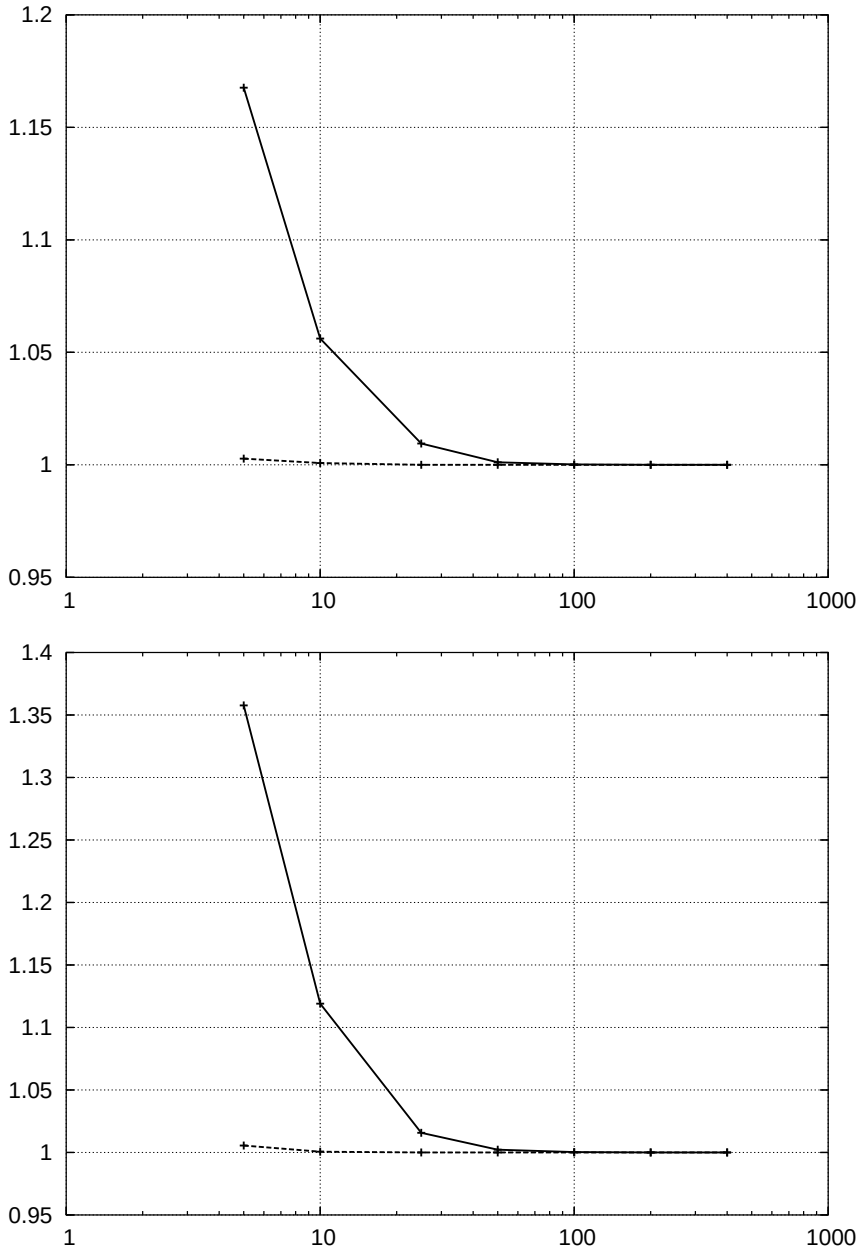


FIG. 5.13 – Influence de la régularité de la covariance; quantiles à 10% et à 90% du rapport d'efficacité d'erreur de prédiction en fonction du nombre de points tirés aléatoirement. k_0 est une covariance de Matérn ($\nu_0 = 0.5$ et $\rho_0 = 0.5$). Sur le graphique du haut, nous utilisons la covariance généralisée polynomiale $k_1(h) = -|h|$. Sur le graphique du bas, nous utilisons la covariance généralisée polynomiale $k_1(h) = -|h| + |h|^3$.

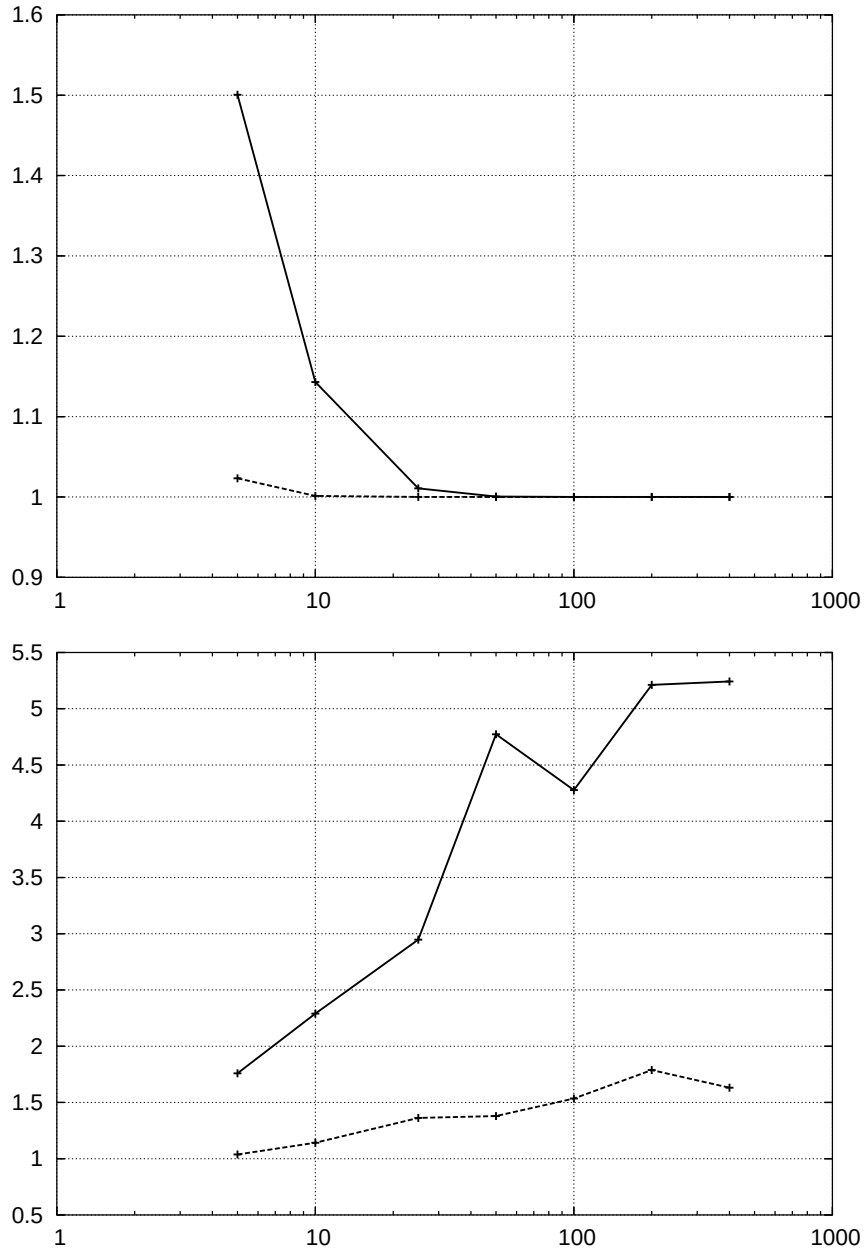


FIG. 5.14 – Influence de la régularité de la covariance; quantiles à 10% et à 90% du rapport d'efficacité d'erreur de prédiction en fonction du nombre de points tirés aléatoirement. k_0 est une covariance de Matérn ($\nu_0 = 1.5$ et $\rho_0 = 0.5$). Sur le graphique du haut, nous utilisons la covariance généralisée polynomiale $k_1(h) = |h|^3$. Sur le graphique du bas, nous utilisons la covariance généralisée polynomiale $k_1(h) = -|h| + |h|^3$.

- une estimation incorrecte de la portée de la covariance n'a asymptotiquement aucune conséquence en terme d'efficacité de la prédiction (voir par exemple la figure 5.9) ;
- une estimation incorrecte de la régularité de la covariance provoque asymptotiquement une perte d'efficacité de la prédiction (voir par exemple les figures 5.10, 5.12 et 5.14) ;
- une estimation incorrecte de la covariance n'a asymptotiquement aucune conséquence, dans la mesure où la variance d'erreur de prédiction tend vers zéro (sauf en présence de problèmes numériques, comme dans la figure 5.7) ;
- dans les régimes non asymptotiques, et même en dimension un, les conséquences d'une mauvaise estimation de la covariance peuvent être importantes (voir par exemple les figures 5.7, 5.9, 5.11 et 5.13) ;
- l'augmentation de la dimension de l'espace des facteurs implique une augmentation non négligeable de l'erreur de prédiction (voir par exemple les figures 5.4 et 5.5).

5.4 Analyse des données

L'analyse de l'influence d'un choix incorrect de covariance sur les prédictions linéaires optimales a mis en évidence la nécessité de choisir une covariance de manière appropriée. L'objectif est de déterminer les caractéristiques importantes (régularité, échelle caractéristique de variations, etc) d'un processus aléatoire susceptible d'avoir généré des observations. L'information contenue dans les données est nécessairement limitée : les observations sont en nombre fini, donc pas denses dans le domaine d'étude. Elles peuvent ne pas être uniformément réparties. De plus, elles correspondent souvent à une unique réalisation du processus aléatoire. Il est alors nécessaire de faire des hypothèses simplificatrices sur le modèle afin de faciliter l'inférence, comme par exemple celle de la stationnarité du processus ou de ses accroissements, qui permet de palier l'unicité de la réalisation du processus aléatoire par une forme de « répétition » spatiale. Une autre difficulté est d'éviter d'introduire dans le modèle des informations qui ne seraient pas contenues dans les données observées (principe du maximum d'entropie) ou connues a priori. Nous privilégions donc les covariances possédant un petit nombre de paramètres permettant d'ajuster facilement les caractéristiques importantes d'un processus aléatoire. Dans les sections suivantes, nous passons en revue trois grandes familles de procédures d'estimation des paramètres de covariance.

5.4.1 Analyse de la variation des échantillons observés

Généralités sur les variogrammes

Nous avons déjà rencontré la notion de variogramme à plusieurs reprises dans ce mémoire. Rappelons que si $F(\boldsymbol{x})$ est un processus aléatoire de moyenne constante et si la variance des accroissements de $F(\boldsymbol{x})$ ne dépend que des différences $\boldsymbol{x} - \boldsymbol{y}$, c'est-à-dire si on peut écrire

$$\mathrm{Var}[F(\boldsymbol{x}) - F(\boldsymbol{y})] = 2\gamma(\boldsymbol{x} - \boldsymbol{y}) \,,$$

alors $F(\boldsymbol{x})$ est *intrinsèquement stationnaire*. Cette notion est plus faible que la notion de stationnarité au second ordre, comme nous l'avons vu au chapitre 4. La fonction $\gamma(\boldsymbol{h})$ s'appelle *variogramme* (le terme exact est *semi-variogramme*). Si $F(\boldsymbol{x})$ est stationnaire au second ordre et

admet la covariance $k(\boldsymbol{h})$ alors

$$\gamma(\boldsymbol{h}) = k(\mathbf{0}) - k(\boldsymbol{h})\,.$$

Inversement, supposons que l'on souhaite trouver la fonction de covariance à partir du variogramme. Dans le cas d'un processus aléatoire stationnaire ergodique[4], il suffit de remarquer que $k(\boldsymbol{h})$ tend vers zéro lorsque $\|\boldsymbol{h}\|$ tend vers l'infini (une conséquence du lemme de Lebesgue-Riemann). Par conséquent, $k(\mathbf{0})$ s'obtient comme la limite de $\gamma(\boldsymbol{h})$ quand $\|\boldsymbol{h}\|$ tend vers l'infini. Cependant, si le processus aléatoire est intrinsèquement stationnaire, la limite du variogramme n'est pas nécessairement définie lorsque $\|\boldsymbol{h}\|$ tend vers l'infini. Si elle n'est pas définie, $F(\boldsymbol{x})$ n'est pas un processus stationnaire.

Le variogramme à l'origine comporte les mêmes informations sur la régularité du processus aléatoire que la covariance à l'origine. Lorsque l'on s'intéresse à la sortie d'un système, il faut prendre en compte la présence éventuelle d'un bruit d'observation. Nous avons vu qu'il pouvait être utile de modéliser ce bruit d'observation par un effet de pépite, ou bruit « blanc » structurel. Quatre grands types de propriétés du variogramme des observations méritent un examen particulier :

- une discontinuité à l'origine traduit la présence d'un bruit d'observation, avec une amplitude qui correspond à la variance du bruit ;
- la pente à l'origine est caractéristique de la régularité de la sortie du système ;
- il existe souvent une distance de décorrélation caractéristique, appelée la *portée*, au delà de laquelle le variogramme devient approximativement constant ;
- la valeur maximale du variogramme correspond à la variance des observations.

La figure 5.15 illustre ces propriétés.

Méthodes d'estimation des paramètres fondées sur le variogramme

En géostatistique, la méthode d'estimation de la covariance la plus utilisée consiste à ajuster les paramètres d'une covariance admissible (Matérn, gaussienne, etc.) pour optimiser un critère d'adéquation à un variogramme empirique construit à partir des données observées.

Estimation d'un variogramme. Les variogrammes empiriques sont construits à partir d'une estimation de la variance des accroissements

$$2\hat{\gamma}(h) = \mathrm{Var}_{\mathrm{estim}}\left\{F^{\mathrm{obs}}(\boldsymbol{x}_i) - F^{\mathrm{obs}}(\boldsymbol{x}_j)\,,\quad i > j,\, h \le d(\boldsymbol{x}_i, \boldsymbol{x}_j) < h + \Delta h\right\}\,.$$

Le paramètre Δh s'ajuste expérimentalement en fonction du nombre de données observées et de leur densité. Les estimateurs classiques (biaisés ou non) de la variance peuvent être utilisés. Typiquement, si l'on considère n variables aléatoires X_i de moyenne nulle, l'estimateur standard (aussi appelé empirique) de la variance est

$$\mathrm{Var}_{\mathrm{estim}}\{X_i, i = 1, \ldots, n\} = \frac{1}{n-1}\sum_{i=1}^{n} X_i^2\,. \tag{5.20}$$

Les estimateurs robustes classiques peuvent également être utilisés, comme ceux utilisant les quantiles (la médiane notamment). Si les X_i sont des variables aléatoires gaussiennes de variance σ^2,

[4] Il s'agit d'une condition suffisante qui n'est pas nécessaire.

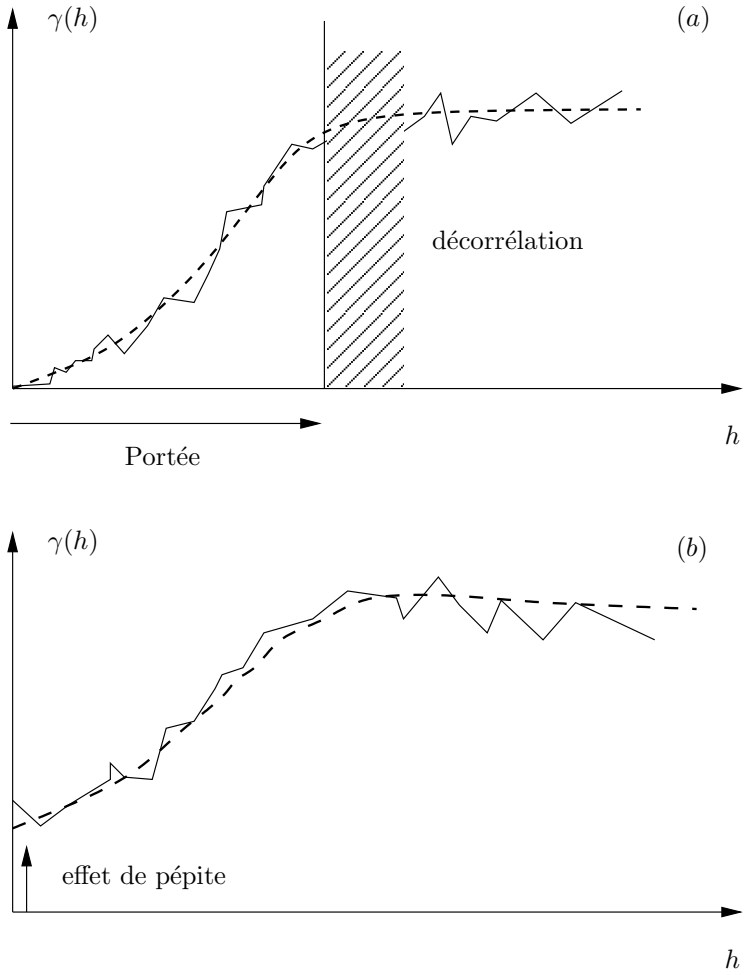


FIG. 5.15 – Propriétés d'un variogramme ; (a) illustration de la notion de portée ; (b) illustration de l'effet de pépite.

les X_i^2/σ^2 suivent une loi χ_1^2. Il en résulte que les quantiles sont proportionnels à la variance σ^2. Un autre estimateur possible de la variance est ainsi

$$\mathrm{Var}_{\mathrm{estim}}\{X_i, i=1,\ldots,n\} = \frac{1}{0.457}\,\mathrm{mediane}\{X_i^2\,,\quad i=1,\ldots,n\}\,.$$

D'autres formes sont également conseillées dans la littérature géostatistique, comme celle proposée par Cressie et Hawkins (1980) :

$$\mathrm{Var}_{\mathrm{estim}}\{X_i, i=1,\ldots,n\} = \frac{1}{0.457+0.494/n}\left\{\frac{1}{n}\sum_{i=1}^{n}|X_i|^{1/2}\right\}^4\,.$$

Pour un traitement plus général des estimateurs robustes, on pourra se référer à (Huber, 1981).

Estimation d'une covariance paramétrée à partir du variogramme. Ce type de méthode ne s'applique en pratique qu'à la détermination de covariances paramétrées *isotropes* $k_{\boldsymbol{\theta}}(h)$, où $\boldsymbol{\theta}$ est un vecteur de paramètres. Il s'agit d'estimer le variogramme paramétré $\gamma_{\boldsymbol{\theta}}(h) = k_{\boldsymbol{\theta}}(0) - k_{\boldsymbol{\theta}}(h)$ en formant seulement des accroissements des données observées. Deux voies peuvent être suivies pour ajuster un variogramme paramétré. La première consiste à optimiser un critère d'adéquation de $\gamma_{\boldsymbol{\theta}}$ à la nuée variographique, c'est-à-dire à l'ensemble de points $\left\{\left(\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_2, \frac{1}{2}(f^{\mathrm{obs}}_{\boldsymbol{x}_i}-f^{\mathrm{obs}}_{\boldsymbol{x}_j})^2\right), i>j\right\}$. Nous présentons ici rapidement la seconde possibilité, qui consiste à obtenir une estimée non-paramétrique du variogramme (paragraphe précédent) et puis à approximer cette estimée par le variogramme paramétré $\gamma_{\boldsymbol{\theta}}$.

Notons $\hat{\boldsymbol{\gamma}}$ le vecteur des valeurs du variogramme empirique aux distances $h_1,\ldots,h_N$ (par l'une des méthodes décrites ci-dessus) et $\boldsymbol{\gamma}(\boldsymbol{\theta})$ le vecteur des valeurs du variogramme paramétré aux distances $h_1,\ldots,h_N$. Les trois méthodes classiques d'estimation de $\boldsymbol{\theta}$ sont

– la méthode des moindres carrés ordinaires (*OLS*), dans laquelle on minimise

$$(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}(\boldsymbol{\theta}))^{\mathsf{T}}(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}(\boldsymbol{\theta}))\,;$$

– la méthode des moindres carrés pondérés (*WLS*), dans laquelle on minimise

$$(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}(\boldsymbol{\theta}))^{\mathsf{T}}\boldsymbol{W}_{\boldsymbol{\theta}}^{-1}(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}(\boldsymbol{\theta}))\,,$$

où $\boldsymbol{W}_{\boldsymbol{\theta}}$ est la matrice diagonale des variances du vecteur $\hat{\boldsymbol{\gamma}}$ obtenues en considérant que le vrai variogramme est $\gamma_\theta(h)$;

– la méthode des moindres carrés généralisés (*GLS*), dans laquelle on minimise

$$(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}(\boldsymbol{\theta}))^{\mathsf{T}}\boldsymbol{V}_{\boldsymbol{\theta}}^{-1}(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}(\boldsymbol{\theta}))\,,$$

où $\boldsymbol{V}_{\boldsymbol{\theta}}$ est la matrice de covariance du vecteur $\hat{\boldsymbol{\gamma}}$ en considérant que le vrai variogramme est $\gamma_\theta(h)$.

Les estimateurs *OLS*, *WLS* et *GLS* ont une efficacité croissante mais une facilité d'utilisation décroissante. Plus précisément, l'estimateur *OLS* requiert l'utilisation d'un algorithme de moindres carrés non-linéaire mais les estimateurs *WLS* et *GLS* nécessitent de déterminer en plus les matrices $\boldsymbol{W}_{\boldsymbol{\theta}}$ et $\boldsymbol{V}_{\boldsymbol{\theta}}$. En pratique, la variance des éléments du vecteur $\hat{\boldsymbol{\gamma}}$ est approximée par

$$\mathrm{Var}[\hat{\gamma}(h)] \approx \frac{8\gamma^2(h)}{n} \tag{5.21}$$

où n est le nombre d'échantillons dans l'estimateur de la variance (par exemple dans (5.20)). Notons que l'approximation (5.21) s'obtient à partir de (5.20) et de la relation vérifiée lorsque $F(\boldsymbol{x})$ est gaussien :

$$\mathrm{Var}[\frac{1}{2}(F(\boldsymbol{x}+\boldsymbol{h})-F(\boldsymbol{x}))^2] = 2\gamma(\|\boldsymbol{h}\|)^2\,.$$

On trouve alors une approximation du critère de l'estimateur *WLS* (Cressie, 1993) sous la forme

$$\sum_j n_{h_j}\left(\frac{\hat{\gamma}_j}{\gamma_{\boldsymbol{\theta}(h_j)}}-1\right)^2\,,$$

où n_{h_j} est le nombre d'accroissements utilisés pour estimer le variogramme à la distance h_j.

5.4.2 Validation croisée

Dans le cadre des méthodes de régression régularisée à noyau reproduisant, les méthodes de validation croisée sont souvent utilisées. Wahba (1990) consacre un chapitre sur l'estimation du paramètre réglant le compromis entre fidélité aux données et régularité des régressions par splines.

Rappelons très brièvement le principe. Étant donné un ensemble S de n observations, on s'intéresse à l'influence de la suppression d'une observation sur la fonction d'attache aux données (en anglais, on parle d'une méthode *leave-one-out*). Plus précisément, en notant $\hat{f}_S$ l'approximation obtenue en utilisant toutes les données et $\hat{f}_{S\setminus k}$ celle obtenue en supprimant l'observation k, on s'intéresse au critère

$$V_0(\boldsymbol{\theta}) = \sum_{i=1}^{n}\left(f^{\mathrm{obs}}_{\boldsymbol{x}_i} - \hat{f}_{S\setminus i}(\boldsymbol{x}_i)\right)^2$$

qui dépend du vecteur $\boldsymbol{\theta}$ des paramètres de la régression (paramètres du noyau, bruit d'observation, etc.). La méthode de *validation croisée ordinaire* consiste à choisir le minimiseur $\hat{\boldsymbol{\theta}}$ de $V_0(\boldsymbol{\theta})$. Autrement dit, le critère consiste à retenir les paramètres du modèle minimisant une estimée de l'erreur de prédiction sur la base des données observées. Il est possible de décliner l'idée du *leave-one-out* en calculant l'erreur de prédiction sur plusieurs observations simultanément, ou plus généralement, en partitionnant l'ensemble des données afin de calculer l'erreur de prédiction d'un sous-ensemble des données à partir d'un autre sous-ensemble des données.

La validation croisée ordinaire possède des défauts qui motivent l'utilisation de la *validation croisée généralisée* (Wahba, 1990). D'autres solutions plus avancées ont été proposées, comme celle de la *validation croisée approximée généralisée* ou *GACV* (Wahba, 1998). Certaines de ces méthodes ont des fondements théoriques pertinents, et en pratique, les méthodes de validation croisée semblent satisfaisantes (nous n'avons pas utilisé ces méthodes dans notre travail).

5.4.3 Maximum de vraisemblance

Nous trouvons dans (Kitadinis, 1983 ; Mardia et Marshall, 1984 ; Jones et Vecchia, 1993) les premières propositions d'estimation de la covariance d'un processus aléatoire *spatial* par maximum de vraisemblance. Si le processus observé est gaussien, de *moyenne connue* et de covariance paramétrée $k_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{y})$, il est immédiat d'écrire la vraisemblance des données en fonction des paramètres de la covariance et par suite, de maximiser cette vraisemblance en fonction de $\boldsymbol{\theta}$. Notons $\boldsymbol{K}_S(\boldsymbol{\theta})$ la matrice de covariance de $\boldsymbol{F}_S = [F(\boldsymbol{x}_1),\ldots,F(\boldsymbol{x}_n)]^{\mathsf{T}}$ et $\boldsymbol{K}_N(\boldsymbol{\theta}')$ la matrice de covariance du vecteur

aléatoire $\boldsymbol{N}$ du bruit d'observation, supposé gaussien, de moyenne nulle et de covariance dépendant d'un vecteur de paramètres $\boldsymbol{\theta}'$. Pour simplifier, posons $\bar{\boldsymbol{\theta}} = (\boldsymbol{\theta}^\mathsf{T}, \boldsymbol{\theta}'^\mathsf{T})^\mathsf{T}$ et $\boldsymbol{K}(\bar{\boldsymbol{\theta}}) = \boldsymbol{K}_S(\boldsymbol{\theta}) + \boldsymbol{K}_N(\boldsymbol{\theta}')$. Le vecteur aléatoire $\boldsymbol{F}^{\text{obs}} = \boldsymbol{F}_S + \boldsymbol{N}$ admet la densité de probabilité

$$p(\boldsymbol{f}^{\text{obs}} \mid \bar{\boldsymbol{\theta}}) = \frac{1}{(2\pi)^{n/2}(\det \boldsymbol{K}(\bar{\boldsymbol{\theta}}))^{1/2}} \exp\left(-\frac{1}{2}\boldsymbol{f}^{\text{obs}\mathsf{T}} \boldsymbol{K}(\bar{\boldsymbol{\theta}})^{-1} \boldsymbol{f}^{\text{obs}}\right).$$

Par conséquent, la log–vraisemblance s'écrit

$$L(\boldsymbol{f}^{\text{obs}}, \bar{\boldsymbol{\theta}}) = -\frac{n}{2}\log 2\pi - \frac{1}{2}\log\det \boldsymbol{K}(\bar{\boldsymbol{\theta}}) - \frac{1}{2}\boldsymbol{f}^{\text{obs}\mathsf{T}} \boldsymbol{K}(\bar{\boldsymbol{\theta}})^{-1} \boldsymbol{f}^{\text{obs}}. \tag{5.22}$$

(Nous avons déjà rencontré cette expression dans la section 3.6.2.) La log–vraisemblance se maximise numériquement à l'aide d'algorithmes de programmation non-linéaire. Nous pouvons évaluer (5.22) via la décomposition de Cholesky $\boldsymbol{C}\boldsymbol{C}^\mathsf{T}$ de la matrice de covariance $\boldsymbol{K}(\bar{\boldsymbol{\theta}})$, où $\boldsymbol{C}$ est une matrice triangulaire inférieure (une telle décomposition existe si la matrice de covariance est positive (Ciarlet, 1998)). Le déterminant de $\boldsymbol{K}(\bar{\boldsymbol{\theta}})$ s'obtient alors comme le produit des éléments diagonaux de $\boldsymbol{C}$ au carré. Le calcul de $\boldsymbol{K}(\bar{\boldsymbol{\theta}})^{-1}\boldsymbol{f}^{\text{obs}}$ s'effectue en résolvant le système linéaire $\boldsymbol{C}\boldsymbol{C}^\mathsf{T}\boldsymbol{u} = \boldsymbol{f}^{\text{obs}}$ ce qui peut se faire en exploitant la structure triangulaire de $\boldsymbol{C}$.

Pour faciliter la maximisation de la vraisemblance, il est souvent nécessaire d'évaluer le gradient de la log–vraisemblance (5.22). Ce gradient s'obtient analytiquement par

$$-2\frac{\partial L(\boldsymbol{f}^{\text{obs}}, \bar{\boldsymbol{\theta}})}{\partial \bar{\boldsymbol{\theta}}} = \operatorname{tr}\left[\boldsymbol{K}(\bar{\boldsymbol{\theta}})^{-1}\frac{\partial \boldsymbol{K}(\bar{\boldsymbol{\theta}})}{\partial \bar{\boldsymbol{\theta}}}\right] - (\boldsymbol{K}(\bar{\boldsymbol{\theta}})^{-1}\boldsymbol{f}^{\text{obs}})^\mathsf{T}\frac{\partial \boldsymbol{K}(\bar{\boldsymbol{\theta}})}{\partial \bar{\boldsymbol{\theta}}}(\boldsymbol{K}(\bar{\boldsymbol{\theta}})^{-1}\boldsymbol{f}^{\text{obs}}),$$

ce qui nécessite d'inverser explicitement la matrice $\boldsymbol{K}(\bar{\boldsymbol{\theta}})$ et d'être capable de dériver la covariance par rapport aux paramètres[5].

Rappelons quelques concepts classiques. Pour alléger par la suite, nous utiliserons la notation $\frac{\partial}{\partial \bar{\boldsymbol{\theta}}} = \partial_{\bar{\boldsymbol{\theta}}}$.

Définition 43. Supposons la densité $p(\boldsymbol{f} \mid \bar{\boldsymbol{\theta}})$ dérivable par rapport à $\bar{\boldsymbol{\theta}}$. Le *score*, noté $V_{\bar{\boldsymbol{\theta}}}(\boldsymbol{F}^{\text{obs}})$, est le vecteur aléatoire définie par

$$\boldsymbol{V}_{\bar{\boldsymbol{\theta}}}(\boldsymbol{F}^{\text{obs}}) = \partial_{\bar{\boldsymbol{\theta}}} \log p(\boldsymbol{F}^{\text{obs}} \mid \bar{\boldsymbol{\theta}}) = \frac{\partial_{\bar{\boldsymbol{\theta}}} p(\boldsymbol{F}^{\text{obs}} \mid \bar{\boldsymbol{\theta}})}{p(\boldsymbol{F}^{\text{obs}} \mid \bar{\boldsymbol{\theta}})}.$$

Le score représente la dépendance de la vraisemblance vis-à-vis des paramètres qui forment le vecteur $\bar{\boldsymbol{\theta}}$. Dans la procédure du maximum de vraisemblance, cette quantité s'interprète comme la dépendance des paramètres estimés vis-à-vis des données observées.

Proposition 54.

$$\mathrm{E}_{p(\boldsymbol{f}\mid\bar{\boldsymbol{\theta}})}[\boldsymbol{V}_{\bar{\boldsymbol{\theta}}}(\boldsymbol{F}^{\text{obs}})] = \boldsymbol{0}. \tag{5.23}$$

Démonstration. Sans difficulté. □

L'interprétation classique de (5.23) est que l'estimateur du maximum de vraisemblance est généralement *non-biaisé* : en moyenne, le « vrai » paramètre $\bar{\boldsymbol{\theta}}$ maximise la (log–)vraisemblance. Toutefois, la fonction de vraisemblance des données peut ne pas être unimodale en fonction des paramètres.

[5]Lorsque la covariance ne se dérive pas facilement, des méthodes de *différentiation automatique* peuvent être envisagées. Nous n'avons cependant pas eu le temps de mettre en œuvre cette démarche dans notre travail.

Définition 44. La *matrice d'information de Fisher*, notée $\boldsymbol{I}_{\bar{\boldsymbol{\theta}}}$, est la matrice de covariance du score définie par

$$\boldsymbol{I}_{\bar{\boldsymbol{\theta}},[i,j]} = \mathrm{E}_{p(\boldsymbol{f}|\bar{\boldsymbol{\theta}})}\left[\partial_{\bar{\boldsymbol{\theta}}_{[i]}} \log p(\boldsymbol{F}^{\mathrm{obs}} \mid \bar{\boldsymbol{\theta}}) \cdot \partial_{\bar{\boldsymbol{\theta}}_{[j]}} \log p(\boldsymbol{F}^{\mathrm{obs}} \mid \bar{\boldsymbol{\theta}})\right]$$

Proposition 55. *$\boldsymbol{I}_{\bar{\boldsymbol{\theta}}}$ est telle que*

$$\boldsymbol{I}_{\bar{\boldsymbol{\theta}},[i,j]} = -\,\mathrm{E}_{p(\boldsymbol{f}|\bar{\boldsymbol{\theta}})}\left[\partial^2_{\bar{\boldsymbol{\theta}}_{[i]},\bar{\boldsymbol{\theta}}_{[j]}} \log p(\boldsymbol{F}^{\mathrm{obs}} \mid \bar{\boldsymbol{\theta}})\right] .$$

Démonstration. Sans difficulté. □

L'estimateur $\hat{\bar{\boldsymbol{\theta}}}$ est un vecteur aléatoire, en tant que fonction de $\boldsymbol{F}^{\mathrm{obs}}$. Notons $\boldsymbol{B}_{\hat{\bar{\boldsymbol{\theta}}}}$ la matrice de covariance de $\hat{\bar{\boldsymbol{\theta}}}$ définie par

$$\boldsymbol{B}_{\hat{\bar{\boldsymbol{\theta}}},[i,j]} = \mathrm{E}_{p(\boldsymbol{f}|\bar{\boldsymbol{\theta}})}\left[\left(\hat{\bar{\boldsymbol{\theta}}}_{[i]} - \mathrm{E}_{p(\boldsymbol{f}|\bar{\boldsymbol{\theta}})}[\hat{\bar{\boldsymbol{\theta}}}_{[i]}]\right)\left(\hat{\bar{\boldsymbol{\theta}}}_{[j]} - \mathrm{E}_{p(\boldsymbol{f}|\bar{\boldsymbol{\theta}})}[\hat{\bar{\boldsymbol{\theta}}}_{[j]}]\right)\right] .$$

La borne inférieure de la variance d'un estimateur non biaisé $\hat{\bar{\boldsymbol{\theta}}}$ est établie grâce au théorème classique suivant.

Théorème 28 (Borne de Cramér et Rao). *Tout estimateur non biaisé $\hat{\bar{\boldsymbol{\theta}}}$ satisfait la relation*

$$\det(\boldsymbol{I}_{\bar{\boldsymbol{\theta}}}\boldsymbol{B}_{\hat{\bar{\boldsymbol{\theta}}}}) \geq 1 .$$

Démonstration. Ce résultat s'obtient comme conséquence de l'inégalité de Cauchy–Schwarz. Voir par exemple (Cover et Thomas, 1991). □

Définition 45. L'efficacité d'un estimateur $\bar{\boldsymbol{\theta}}$ est définie par $e = 1/\det(\boldsymbol{I}_{\bar{\boldsymbol{\theta}}}\boldsymbol{B}_{\hat{\bar{\boldsymbol{\theta}}}})$. (On a donc $e \leq 1$.)

Plus e est proche de 1, plus la variance de l'estimateur est petite. On s'attend souvent à ce que l'estimateur du maximum de vraisemblance soit asymptotiquement efficace. Il est cependant difficile d'établir les propriétés de la matrice de covariance $\boldsymbol{B}_{\hat{\bar{\boldsymbol{\theta}}}}$ de l'estimateur du maximum de vraisemblance de données non uniformément échantillonnées[6]. (Mardia et Marshall, 1984) examine l'efficacité asymptotique de cet estimateur dans le cas d'observations effectuées sur un domaine croissant de l'espace des facteurs. Plus précisément, le résultat est établi en considérant un certain $h_0 > 0$ et une suite de point $(\boldsymbol{x}_i)_{i\in\mathbb{N}}$ telle que $\|\boldsymbol{x}_i - \boldsymbol{x}_j\| \geq h_0$, $\forall i, j$. Les résultats présentés dans cet article, même s'ils nécessitent des observations sur un domaine non borné, plaident en faveur de l'utilisation de l'estimateur du maximum de vraisemblance. Même si nous ne sommes pas en mesure de prouver l'efficacité asymptotique de l'estimateur, il est utile d'évaluer le Hessien de la fonction de vraisemblance pour le vecteur des paramètres estimé. En regardant les racines carrées des éléments diagonaux de l'inverse du Hessien, nous pouvons en effet nous faire une idée des écarts type des erreurs d'estimation sur chacun des paramètres.

[6] (Schölkopf et Smola, 2002), p. 74, citent un lemme de Murata et al. (1994) semblant fournir une réponse agréable à ce problème. Cependant, un examen attentif de la preuve montre que le lemme ne s'applique pas parce qu'il suppose implicitement l'indépendance des données observées $(\boldsymbol{x}_i, \boldsymbol{f}^{\mathrm{obs}}_{\boldsymbol{x}_i})$ pour utiliser la loi des grands nombres.

Remarques.

1. Si l'on utilise des algorithmes d'optimisation itératifs locaux, il faut fournir à ces algorithmes une valeur initiale suffisamment proche de l'optimum car il peut y avoir des optima locaux. Mentionnons toutefois l'existence d'algorithmes issus de techniques de calcul par intervalles permettant des optimisations globales (Hansen, 1992).
2. Il faut parfois tenir compte de contraintes (de signe, par exemple) sur les paramètres.
3. Le coût algorithmique de la méthode du maximum de vraisemblance est élevé puisque le calcul de la vraisemblance est typiquement en $\mathrm{O}(n^3)$. Le calcul du gradient est en $\mathrm{O}(n^4)$. Lorsque n est grand (disons à partir de $n = 1000$), il est possible d'approximer la vraisemblance à l'aide d'algorithmes n'utilisant qu'une partie des données. De tels algorithmes sont proposés par (Vecchia, 1988 ; Stein et al., 2004) ;
4. L'utilisation d'une méthode efficace d'optimisation est souhaitable. Par exemple, nous utilisons
   - la méthode du simplexe de Nelder–Mead (optimisation sans contrainte et sans calcul de gradient),
   - les méthodes de type gradient conjugué de Polack–Ribiere (optimisation sans contrainte, avec calcul du gradient).
5. Il n'est pas facile de comparer le maximum de vraisemblance aux autres approches ; (Zimmerman et Zimmerman, 1991) compare numériquement le maximum de vraisemblance aux estimateurs par variogramme ; (Wahba, 1990) développe des arguments théoriques en faveur des méthodes de validation croisée. Résumons de manière synthétique quelques éléments de comparaison.
   - Le maximum de vraisemblance possède un degré de généralité supérieur aux autres méthodes, puisqu'il permet en principe d'estimer n'importe quelle forme de covariance paramétrée. D'autre part, le maximum de vraisemblance n'est pas restreint aux processus gaussiens (la méthode est applicable tant que l'on est capable d'exprimer la loi du vecteur des observations).
   - La forme (5.22) doit en principe être réservée à l'estimation de paramètres de processus gaussiens. En pratique cependant nous l'utilisons sans nous assurer que les données suivent une loi gaussienne. Wahba (1990) déconseille l'utilisation du maximum de vraisemblance en raison de son manque de robustesse vis-à-vis des écarts possibles de la vraie distribution des données au modèle gaussien. Notons que ce problème s'apparente à celui de l'estimation robuste d'un paramètre de position (Huber, 1981).
   - Le maximum de vraisemblance, permet de comparer différents modèles et d'estimer l'incertitude sur les paramètres estimés.

Nous verrons dans la section 5.4.5 l'extension du maximum de vraisemblance au cas de processus à moyenne inconnue (méthode *REML*).

5.4.4 Combinaisons linéaires de noyaux

Cette section est une introduction à la méthode que nous proposons dans l'annexe de ce chapitre. Lorsque la dimension de l'espace des facteurs est faible (un ou deux), le variogramme

empirique est un outil adapté pour comprendre la structure des données. Cependant, dans des espaces de facteurs de dimension plus grande ceci devient difficile. Les anisotropies sont par exemple très difficiles à estimer et le choix d'une covariance paramétrée est empirique (voir par exemple l'application présentée dans la section 6.4.2). Nous ne pouvons qu'être partiellement satisfaits des méthodes existantes d'adaptation d'un noyau aux données. L'approche traditionnelle consistant à sélectionner un noyau paramétré parmi les familles classiques de fonctions de type positif, puis ajuster les paramètres de ce noyau pour optimiser un critère d'adéquation aux données, nous semble limitée et trop rigide.

Nous aimerions être capables de construire un noyau à partir des données, par exemple en combinant plusieurs noyaux possédant chacun des caractéristiques spécifiques simples. Nous présentons dans l'annexe un algorithme itératif dont l'idée est d'extraire des informations d'un ensemble d'observations puis de combiner linéairement des noyaux portant ces informations. Dans les paragraphes suivants, nous donnons seulement les idées principales de cette approche.

Choisir un noyau approprié signifie que les données observées sont admissibles comme échantillons d'une fonction dans l'espace de Hilbert généré par ce noyau. Plus précisément, soient $F(\boldsymbol{x})$ un processus aléatoire du second ordre centré et de covariance $k(\boldsymbol{x},\boldsymbol{y})$ et $k_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{y})$ une famille de noyaux reproduisants paramétrée par le vecteur $\boldsymbol{\theta}$. Nous considérons aussi les processus aléatoires gaussiens centrés $F_{\boldsymbol{\theta}}$ associés aux covariances $k_{\boldsymbol{\theta}}$. Posons $\Phi_{\boldsymbol{\theta}}(\boldsymbol{f}) = \boldsymbol{f}^{\mathsf{T}}\boldsymbol{K}^{-1}\boldsymbol{K}_{\boldsymbol{\theta}}\boldsymbol{K}^{-1}\boldsymbol{f}$, $\boldsymbol{f}\in\mathbb{R}^n$. La statistique $\Phi_{\boldsymbol{\theta}}(\boldsymbol{f}^{\mathrm{obs}})$ s'interprète comme une norme de $\boldsymbol{f}^{\mathrm{obs}}$ dans l'espace à noyau reproduisant $k_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{y})$. En effet, la projection orthogonale de $\boldsymbol{F} = (F(\boldsymbol{x}_1),\ldots,F(\boldsymbol{x}_n))^{\mathsf{T}}$ sur l'espace

$$\mathcal{H}_{S_n,\boldsymbol{\theta}} = \mathrm{span}\{F_{\boldsymbol{\theta}}(\boldsymbol{x}_1),\ldots,F_{\boldsymbol{\theta}}(\boldsymbol{x}_n)\}$$

est $\hat{\boldsymbol{F}}_{\boldsymbol{\theta}} = \boldsymbol{K}_{\boldsymbol{\theta}}\boldsymbol{K}^{-1}\boldsymbol{F}$. Par conséquent

$$\hat{\boldsymbol{f}}_{\boldsymbol{\theta}}{}^{\mathsf{T}}\boldsymbol{K}_{\boldsymbol{\theta}}^{-}\hat{\boldsymbol{f}}_{\boldsymbol{\theta}} = \boldsymbol{f}^{\mathrm{obs}\,\mathsf{T}}\boldsymbol{K}^{-1}\boldsymbol{K}_{\boldsymbol{\theta}}\boldsymbol{K}^{-1}\boldsymbol{f}^{\mathrm{obs}}\,,$$

où $\boldsymbol{K}_{\boldsymbol{\theta}}^{-}$ est la matrice pseudo-inverse de $\boldsymbol{K}_{\boldsymbol{\theta}}$, et $\hat{\boldsymbol{f}}_{\boldsymbol{\theta}}$ est la réalisation de $\hat{\boldsymbol{F}}_{\boldsymbol{\theta}}$ correspondant à la réalisation $\boldsymbol{f}^{\mathrm{obs}}$ de $\boldsymbol{F}$. Si $\hat{\boldsymbol{f}}_{\boldsymbol{\theta}}$ est un vecteur propre de $\boldsymbol{K}_{\boldsymbol{\theta}}$, c'est-à-dire, $\boldsymbol{K}_{\boldsymbol{\theta}}\hat{\boldsymbol{f}}_{\boldsymbol{\theta}} = \gamma\hat{\boldsymbol{f}}_{\boldsymbol{\theta}}$, $\gamma\in\mathbb{R}$, alors $\hat{\boldsymbol{f}}_{\boldsymbol{\theta}}{}^{\mathsf{T}}\boldsymbol{K}_{\boldsymbol{\theta}}^{-}\hat{\boldsymbol{f}}_{\boldsymbol{\theta}} = \gamma^{-1}\|\hat{\boldsymbol{f}}_{\boldsymbol{\theta}}\|_2^2$. Par suite, les vecteurs propres de $\boldsymbol{K}_{\boldsymbol{\theta}}$ associés à de grandes valeurs propres γ possèdent des normes petites dans l'espace à noyau reproduisant $k_{\boldsymbol{\theta}}$, et correspondent corrélativement à des réalisations de $\boldsymbol{F}_{\boldsymbol{\theta}}$ ayant une densité de probabilité grande.

Nous pouvons calculer les statistiques $\Phi_{\boldsymbol{\theta}}(\boldsymbol{f}^{\mathrm{obs}})$ pour différents noyaux $k_{\boldsymbol{\theta}}$ et évaluer la présence des caractéristiques portées par le noyau $k_{\boldsymbol{\theta}}$ dans le vecteur des observations. L'objectif est ainsi de jauger les caractéristiques essentielles du processus aléatoire ayant généré les données. Nous avons essayé de trouver un algorithme fondé sur ce principe. Les résultats présentés en annexe, sans être totalement satisfaisants, laissent penser que l'idée est pertinente.

5.4.5 Estimation des paramètres d'une covariance généralisée polynomiale

Cette section présente deux méthodes d'estimation des paramètres des covariances généralisées.

Méthode du variogramme

La méthode présentée dans ce paragraphe pour estimer les paramètres du modèle de la fonction aléatoire intrinsèque d'ordre l est dérivée de la méthode du variogramme (Delfiner et Matheron, 1980). Nous nous intéressons spécifiquement aux covariances généralisées polynomiales qui s'expriment sous la forme

$$k(h) = b_{-1}\delta(h) + \sum_{p=0}^{l}(-1)^{p+1}b_p|h|^{2p+1}\,. \tag{5.24}$$

Le premier terme de (5.24) permet de prendre en compte un bruit d'observation. Une méthode possible pour estimer les paramètres b_p, $p = -1, \ldots, l$, est d'estimer statistiquement à partir des données la variance de la fonction aléatoire intrinsèque (en fait il s'agit plus précisément de la variance d'une combinaison linéaire de l'un des représentants de l'*IRF*). Compte tenu, de (5.24), la variance de $F_{\mathrm{G}}(\lambda)$, avec $\lambda = \sum_{i=1}^{n}\lambda_i\delta_{\boldsymbol{x}_i}$, s'exprime comme

$$\mathrm{E}[F_{\mathrm{G}}(\lambda)^2] = b_{-1}\sum_i \lambda_i^2 + \sum_{p=0}^{l}(-1)^{p+1}b_p\sum_{i,j}\lambda_i\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^{2p+1}\lambda_j\,.$$

Si l'on pose

$$\left\{\begin{array}{rcl} T_{-1} & = & \sum_i \lambda_i^2\,, \\ T_{2p+1} & = & (-1)^{p+1}\sum_{i,j}\lambda_i\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^{2p+1}\lambda_j\,, \end{array}\right.$$

cette variance s'écrit

$$\mathrm{E}[F_{\mathrm{G}}(\lambda)^2] = \sum_{p=-1}^{l} b_p T_{2p+1}\,,$$

où les T_{2p+1} peuvent être calculés.

Pour déterminer les coefficients b_p, l'idée est de construire à partir des observations un grand nombre d'accroissements $\Delta_m = \langle \lambda_m f^{\mathrm{obs}}(\boldsymbol{x})\rangle$ et de minimiser le coût quadratique

$$\begin{array}{rcl} Q(b_{-1}, \cdots, b_l) & = & \displaystyle\sum_m \left[\Delta_m^2 - \mathrm{E}[F_{\mathrm{G}}(\lambda_m)^2]\right]^2 \\ & = & \displaystyle\sum_m \left[\Delta_m^2 - \sum_{p=-1}^{l} b_p T_{2p+1}^m\right]^2\,. \end{array} \tag{5.25}$$

En pratique, on a toutefois intérêt à privilégier les mesures λ_m *concentrées dans des petits voisinages* de l'espace des facteurs parce qu'une connaissance suffisamment précise du comportement de la covariance généralisée quand $h \to 0$ est très importante.

Pour ce faire, on peut se servir d'une fonction de coût pondérée :

$$Q(b_{-1}, \cdots, b_l) = \sum_m w_m^2\left[\Delta_m^2 - \mathrm{E}[F_{\mathrm{G}}(\lambda_m)^2]\right]^2\,. \tag{5.26}$$

Les coefficients de pondération w_m^2 accordé aux différentes variables $F_{\mathrm{G}}(\lambda_m)^2$ devraient être égaux en principe aux inverses de $\mathrm{Var}[F_{\mathrm{G}}(\lambda_m)^2]$. Ces quantités font intervenir les moments d'ordre quatre de $F_{\mathrm{G}}(\lambda_m)$. Si $F_{\mathrm{G}}(\lambda)$ est une *IRF* gaussienne, le moment d'ordre quatre $\mathrm{E}[F_{\mathrm{G}}(\lambda)^4]$ est égal à $3\,\mathrm{E}[F_{\mathrm{G}}(\lambda)^2]^2$. Pour l'estimation des coefficients w_m, on peut alors envisager une procédure itérative dans laquelle on utilise les w_m calculés en utilisant les estimées des paramètres à l'étape n pour estimer les paramètres à l'étape $n+1$.

Maximum de vraisemblance restreint (*REML*)

L'estimation au sens du méthode du maximum de vraisemblance des caractéristiques d'une *IRF* peut être menée selon deux approches. La première consiste à estimer le modèle à partir de la forme générale des représentants de l'*IRF*. La seconde consiste à écrire non pas la fonction de vraisemblance des données observées, mais celle des accroissements (ou accroissements généralisés) de ces données. La méthode résultante s'appelle méthode du maximum de vraisemblance restreint (*restricted maximum likelihood*, abrégé par *REML*, en anglais). En statistique, ces accroissements s'appellent des *contrastes*. Nous présentons cette méthode dans les paragraphes suivants.

Soit $F_{\mathrm{G}}(\lambda)$ une *IRF*(l) gaussienne. Notons $\boldsymbol{F}^{\mathrm{obs}}$ le vecteur aléatoire des observations, correspondant à la somme du vecteur aléatoire $\boldsymbol{F}_S = (F(\boldsymbol{x}_1), \ldots, F(\boldsymbol{x}_n))^{\mathsf{T}}$, où $F(\boldsymbol{x})$ est un représentant de $F_{\mathrm{G}}(\lambda)$, et d'un vecteur de bruit d'observation supposé de moyenne nulle. Notons également $\boldsymbol{P}_S = (\boldsymbol{x}_j^{\boldsymbol{l}_i})_{i,j=1}^{q,n}$ la matrice $q \times n$ de fonctions de bases de $\mathcal{N}_l$ évaluées sur S. Comme $\mathcal{N}_l$ est de dimension q et comme l'espace des mesures à support S est de dimension n, l'espace des mesures à support S annulant les fonctions de $\mathcal{N}_l$ est de dimension $n - q$. Supposons trouvée une matrice $\boldsymbol{W}$ de taille $n \times (n - q)$ et de rang $n - q$, telle que

$$\boldsymbol{P}_S \boldsymbol{W} = \boldsymbol{0}\,.$$

(Les colonnes de $\boldsymbol{W}$ sont dans le noyau de $\boldsymbol{P}_S$.) Notons que les colonnes de $\boldsymbol{W}$ sont donc les coefficients de mesures à support S, $\sum_{j=1}^{n} \boldsymbol{W}_{[i,j]} \delta_{\boldsymbol{x}_j} \in \Lambda_l$. Alors $\boldsymbol{Z} = \boldsymbol{W}^{\mathsf{T}}(\boldsymbol{F}_S + \boldsymbol{N})$ est un vecteur aléatoire gaussien à valeurs dans $\mathbb{R}^{n-q}$, de moyenne nulle et de matrice de covariance $\boldsymbol{W}^{\mathsf{T}}(\boldsymbol{K}_S(\boldsymbol{\theta}) + \boldsymbol{K}_N(\boldsymbol{\theta}'))\boldsymbol{W} = \boldsymbol{W}^{\mathsf{T}}\boldsymbol{K}(\bar{\boldsymbol{\theta}})\boldsymbol{W}$, où $\boldsymbol{K}_S(\boldsymbol{\theta})$ est la matrice symétrique de covariance généralisée ayant comme éléments les scalaires $k_{\boldsymbol{\theta}}(\boldsymbol{x}_i - \boldsymbol{x}_j)$ et où $\boldsymbol{K}_N(\boldsymbol{\theta}')$ est la matrice de covariance du bruit d'observation. Le vecteur aléatoire $\boldsymbol{Z}$ est un *vecteur de contrastes*. La log–vraisemblance des contrastes s'écrit

$$L(\boldsymbol{z}, \bar{\boldsymbol{\theta}}) = -\frac{n-q}{2} \log 2\pi - \frac{1}{2} \log \det(\boldsymbol{W}^{\mathsf{T}}\boldsymbol{K}(\bar{\boldsymbol{\theta}})\boldsymbol{W}) - \frac{1}{2}\boldsymbol{z}^{\mathsf{T}}(\boldsymbol{W}^{\mathsf{T}}\boldsymbol{K}(\bar{\boldsymbol{\theta}})\boldsymbol{W})^{-1}\boldsymbol{z}\,.$$

Plusieurs méthodes peuvent être envisagées pour calculer la matrice $\boldsymbol{W}$. Les approches proposées par (Patterson et Thompson, 1971 ; Harville, 1974 ; Stein, 1999) sont considérées comme classiques. Nous préférons utiliser la décomposition QR de $\boldsymbol{P}_S^{\mathsf{T}}$

$$\boldsymbol{P}_S^{\mathsf{T}} = (\boldsymbol{Q}_1 \mid \boldsymbol{Q}_2) \begin{pmatrix} \boldsymbol{R} \\ \boldsymbol{0} \end{pmatrix},$$

où $(\boldsymbol{Q}_1 \mid \boldsymbol{Q}_2)$ est une matrice orthogonale de taille $n \times n$ et $\boldsymbol{R}$ une matrice triangulaire supérieure de taille $q \times q$. Il est immédiat de vérifier que les colonnes de la matrice $\boldsymbol{Q}_2$ forment une base du noyau de $\boldsymbol{P}_S$ et nous pouvons donc choisir $\boldsymbol{W} = \boldsymbol{Q}_2$. Notons que $\boldsymbol{W}^{\mathsf{T}}\boldsymbol{W} = \boldsymbol{I}_{n-q}$. Le coût algorithmique du calcul de la vraisemblance est dominé par celui du calcul de $\boldsymbol{W}^{\mathsf{T}}\boldsymbol{K}(\bar{\boldsymbol{\theta}})\boldsymbol{W}$. Nous avons trouvé que l'approche QR était conceptuellement très satisfaisante et numériquement moins sensible aux problèmes de conditionnement de la matrice de covariance que les méthodes mentionnées dans la littérature.

5.5 Conclusions

Le choix d'un noyau ou d'une covariance de processus aléatoire permet de traduire l'hypothèse que la sortie du système appartient à un espace de fonctions particulier. Ce choix présente des aspects arbitraires car une fonction quelconque appartient à une infinité d'espaces de fonctions. Il convient alors de s'interroger sur la robustesse des procédures de prédiction en fonction du choix du noyau ou de la covariance. Le point de vue probabiliste permet d'affirmer que l'erreur de prédiction linéaire est asymptotiquement décroissante pour une large gamme de fonctions de covariance (qui exclut cependant la covariance gaussienne). De plus, cette décroissance est optimale lorsque la covariance choisie est compatible en un certain sens avec celle du processus aléatoire ayant généré les données (Stein, 1988, 1990b,a). Dans les régimes non asymptotiques, il existe très peu voire pas de résultats[7]. Nos expériences numériques présentées dans la section 5.3 confirment cependant que les écarts entre les erreurs de prédiction de différents modèles peuvent être significatifs.

Par conséquent, le choix de la covariance est l'un des aspect les plus importants de la modélisation comportementale d'un système. Ce choix est constitué de deux étapes. Il faut décider d'une forme de covariance paramétrée et ensuite estimer ses paramètres. Nous avons présenté les covariances classiques de la littérature avant d'effectuer une synthèse des principales méthodes d'estimation des paramètres de la covariance. Dans les applications, nous avons privilégié les méthodes du maximum de vraisemblance et du maximum de vraisemblance restreint malgré leur coût algorithmique assez élevé.

Il nous semble enfin que les covariances classiques sont parfois trop limitées pour obtenir de bons modèles. En petite dimension, cette limitation n'est pas forcément pénalisante dans la mesure où un nombre d'observations peu élevé suffit souvent à couvrir le domaine d'étude, ce qui permet de bénéficier de la propriété de décroissance asymptotique de l'erreur de prédiction. Dans la section 6.5, nous présenterons cependant un exemple de prédiction d'une série chronologique qui tend à montrer que même en petite dimension, l'utilisation de familles de covariances plus flexibles est avantageux. En dimension élevée, le problème du choix de la covariance semble particulièrement délicat, comme l'illustrent nos expériences numériques. Nous avons proposé dans la section 5.6 une méthode de construction parcimonieuse de la covariance, fondée sur un principe de capture des caractéristiques essentielles des données. Cette méthode en est encore à un stade préliminaire, et nous semble mériter des recherches ultérieures.

[7] Nous n'avons pas abordé ici l'apport récent de la théorie statistique de l'apprentissage (Smola, 1998).

5.6 Annexe : combinaisons linéaires de noyaux, hypernoyaux, et poursuite de caractéristiques par maximum de vraisemblance

5.6.1 Introduction

Les propriétés d'une approximation obtenue par régression régularisée dans un espace de Hilbert à noyau reproduisant dépendent fortement, comme nous l'avons dit, du noyau retenu. Le choix d'un noyau doit être effectué en fonctions des connaissances a priori et des données. Au lieu de choisir le noyau dans une famille de fonctions de type positif, nous proposons de le construire en combinant linéairement des noyaux sélectionnés dans une base où chaque noyau comporte des caractéristiques différentes. Le noyau que nous souhaitons construire s'écrit alors sous la forme

$$k_{\boldsymbol{\alpha}}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{l} \alpha_i^2 k_i(\boldsymbol{x}, \boldsymbol{y}) \,. \tag{5.27}$$

Les paramètres α_i interviennent par leur carré dans la combinaison linéaire (5.27) pour garantir la positivité de $k_{\boldsymbol{\alpha}}$. L'utilisation d'une combinaison linéaire de noyaux est d'abord motivée par la possibilité d'estimer facilement les paramètres du vecteur $\boldsymbol{\alpha} = [\alpha_1 \cdots \alpha_l]^\mathsf{T}$. Comme deuxième objectif, nous voulons obtenir des combinaisons linéaires du type (5.27) susceptibles de rendre correctement compte des différentes caractéristiques du processus aléatoire ayant généré les données, mais comportant le moins de termes possible. Pour satisfaire ces objectifs, nous proposons dans la section 5.6.3 un algorithme itératif du type *poursuite de caractéristiques*.

Avant de présenter cet algorithme, nous rappelons une approche alternative proposée dans (Ong et al., submitted), fondée sur la notion d'hypernoyau. Nous montrons comment cette notion conduit également à la construction de combinaisons linéaires de noyaux.

5.6.2 Régularisation avec des hypernoyaux

Afin de simplifier notre présentation, nous revenons aux modèles paramétriques linéaires, utilisés au chapitre 3 pour introduire la notion de noyau reproduisant. Nous cherchons donc à approximer une fonction $f^* : \mathbb{X} \subset \mathbb{R}^d \to \mathbb{R}$ par le modèle linéairement paramétré

$$f(\boldsymbol{x}) = \boldsymbol{b}^\mathsf{T} \boldsymbol{r}(\boldsymbol{x}) \,, \tag{5.28}$$

où $\boldsymbol{b} = (b_{[1]} \dots b_{[l]})^\mathsf{T}$ est un vecteur de paramètres et $\boldsymbol{r}(\boldsymbol{x}) = (r_{[1]}(\boldsymbol{x}) \dots r_{[l]}(\boldsymbol{x}))^\mathsf{T}$ un vecteur de fonctions, que nous appelons vecteur des caractéristiques. Pour estimer le vecteur $\boldsymbol{b}$, nous nous concentrons sur les méthodes de régression régularisée qui consistent à minimiser un critère s'exprimant par exemple sous la forme

$$J(\boldsymbol{b}) = \|\boldsymbol{b}\|_2^2 + C \sum_{i=1}^{n} (f_{\boldsymbol{x}_i}^{\text{obs}} - \boldsymbol{b}^\mathsf{T} \boldsymbol{r}(\boldsymbol{x}_i))^2 \,. \tag{5.29}$$

Le critère (5.29) pénalise les grandes valeurs des éléments de $\boldsymbol{b}$. Le paramètre $C \in \mathbb{R}^+$ permet d'ajuster le compromis entre régularisation et adéquation aux données. Nous avons constaté dans

la section 3.4.1 que ce problème de régression régularisée pouvait s'écrire sous une forme duale. Si $\widehat{\boldsymbol{b}}$ est le minimiseur du critère (5.29), il existe des scalaires $\widehat{\lambda}_i$ tels que

$$\widehat{\boldsymbol{b}} = \sum_{i=1}^{n} \widehat{\lambda}_i \boldsymbol{r}(\boldsymbol{x}_i)\,.$$

On peut alors réécrire (5.29) sous la forme

$$J(\lambda_i, i = 1, \cdots, n) = \left\| \sum_{i=1}^{n} \lambda_i \boldsymbol{r}(\boldsymbol{x}_i) \right\|^2 + C \sum_{i=1}^{n} (f^{\text{obs}}_{\boldsymbol{x}_i} - \widehat{f}(\boldsymbol{x}_i))^2\,.$$

Soit $\mathcal{F}$ l'espace de dimension finie à noyau reproduisant $\boldsymbol{r}(\boldsymbol{x})$ défini par vect$\{\boldsymbol{r}(\boldsymbol{x}) \in \mathbb{R}^l, \boldsymbol{x} \in \mathbb{X}\} \subseteq \mathbb{R}^l$ (voir la section 3.2.1). Tout élément $\boldsymbol{b} = \sum_{i=1}^{n} \lambda_i \boldsymbol{r}(\boldsymbol{x}_i) \in \mathcal{F}$ s'identifie à une fonction f définie par

$$f(\boldsymbol{x}) = (\boldsymbol{r}(\boldsymbol{x}), \sum_{i=1}^{n} \lambda_i \boldsymbol{r}(\boldsymbol{x}_i))\,, \quad \forall \boldsymbol{x} \in \mathbb{X}\,.$$

(Nous utilisons la notation $f \equiv \boldsymbol{b}$ pour rappeler la propriété d'identification.) Rappelons que la norme de $f \equiv \boldsymbol{b} \in \mathcal{F}$ s'exprime sous la forme

$$\|f\|^2_{\mathcal{F}} = \sum_{i,j=1}^{n} \lambda_i (\boldsymbol{r}(\boldsymbol{x}_i), \boldsymbol{r}(\boldsymbol{x}_j)) \lambda_j = \sum_{i=1}^{l} b^2_{[i]}\,.$$

L'espace $\mathcal{F}$ est donc muni d'une norme qui pénalise de la même manière tous les termes $r_{[i]}(\boldsymbol{x})$ du vecteur $\boldsymbol{r}(\boldsymbol{x})$, ce qui n'est pas forcément souhaitable. Nous ne revenons pas sur la nécessité de choisir correctement le type de régularisation, ce problème ayant déjà été abordé dans la section 3.4.2.

L'idée développée dans (Ong et al., submitted) consiste à introduire dans le critère (5.29) un terme de régularisation sur le noyau même. Soient $\boldsymbol{r}_p(\boldsymbol{x})$ un vecteur de caractéristiques et $\boldsymbol{\alpha}$ un vecteur de poids tel que le noyau utilisé pour la régression s'exprime sous la forme $\boldsymbol{r}(\boldsymbol{x}) = \boldsymbol{\alpha} \odot \boldsymbol{r}_p(\boldsymbol{x})$ ($\odot$ désigne l'opérateur de multiplication terme à terme). Pour régulariser le noyau, un critère est dérivé de (5.29) sous la forme

$$J(\boldsymbol{b}, \boldsymbol{\alpha}) = \|\boldsymbol{b}\|_2^2 + C_{\text{k}} \|\boldsymbol{\alpha}\|_4^4 + C_{\text{d}} \sum_{i=1}^{n} (f^{\text{obs}}_{\boldsymbol{x}_i} - \boldsymbol{b}^{\mathsf{T}} \boldsymbol{r}(\boldsymbol{x}_i))^2\,, \tag{5.30}$$

où la raison du choix de la norme $\|.\|_4$ apparaîtra plus loin, et où

$$\begin{aligned}
\|\boldsymbol{b}\|_2^2 &= \sum_{i,j} \lambda_i (\boldsymbol{r}(\boldsymbol{x}_i), \boldsymbol{r}(\boldsymbol{x}_j)) \lambda_j \\
&= \sum_{i,j} \lambda_i (\boldsymbol{\alpha} \odot \boldsymbol{r}_p(\boldsymbol{x}_i), \boldsymbol{\alpha} \odot \boldsymbol{r}_p(\boldsymbol{x}_j)) \lambda_j \\
&= \sum_{i,j} \lambda_i (\boldsymbol{\alpha} \odot \boldsymbol{\alpha}, \boldsymbol{r}_p(\boldsymbol{x}_i) \odot \boldsymbol{r}_p(\boldsymbol{x}_j)) \lambda_j\,.
\end{aligned}$$

Le noyau reproduisant s'écrit maintenant sous la forme

$$k(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{r}(\boldsymbol{x}), \boldsymbol{r}(\boldsymbol{y})) = (\boldsymbol{\alpha} \odot \boldsymbol{\alpha}, \boldsymbol{r}_p(\boldsymbol{x}) \odot \boldsymbol{r}_p(\boldsymbol{y}))\,, \tag{5.31}$$

qui s'apparente à une combinaison linéaire du type (5.27).

En examinant (5.31), l'ensemble des objets $\boldsymbol{\alpha} \odot \boldsymbol{\alpha}$ peut être identifié à un ensemble de noyaux reproduisants

$$\begin{array}{rccc} k: & \underline{\mathbb{X}} \stackrel{\Delta}{=} \mathbb{X} \times \mathbb{X} & \rightarrow & \mathbb{R} \\ & (\boldsymbol{x}, \boldsymbol{y}) & \mapsto & k(\boldsymbol{x}, \boldsymbol{y})\,. \end{array}$$

En effet, soit l'espace de fonctions

$$\underline{\mathcal{F}} = \{g : \underline{\mathbb{X}} \rightarrow \mathbb{R},\ g = (\boldsymbol{\mu}, \boldsymbol{r}_p(\boldsymbol{x}) \odot \boldsymbol{r}_p(\boldsymbol{y})), \boldsymbol{\mu} \in \mathbb{R}^l\},$$

muni du produit scalaire canonique de $\mathbb{R}^l$. Le sous-ensemble des fonctions de type positif

$$\underline{\mathcal{F}}^+ = \left\{\boldsymbol{\mu} \in \underline{\mathcal{F}}, \quad \forall i \in \{1, \cdots, l\},\ \mu_{[i]} \geq 0\right\}$$

définit un espace de noyaux reproduisants.

Remarquons que $\boldsymbol{r}_p(\boldsymbol{x}) \odot \boldsymbol{r}_p(\boldsymbol{y})$ agit comme un opérateur d'évaluation aux points de $\underline{\mathbb{X}}$ sur les éléments de $\underline{\mathcal{F}}$. Pour cette raison, $\boldsymbol{r}_p(\boldsymbol{x}) \odot \boldsymbol{r}_p(\boldsymbol{y})$ est appelé *hypernoyau*. Dans (Ong et al., submitted) l'objet $\boldsymbol{r}_p(\boldsymbol{x}) \odot \boldsymbol{r}_p(\boldsymbol{y})$ est noté $\underline{k}((\boldsymbol{x}, \boldsymbol{y}), \cdot)$.

Il existe un *théorème du représentant* établissant que le minimiseur $\widehat{\boldsymbol{\alpha} \odot \boldsymbol{\alpha}}$ de (5.30) s'écrit sous la forme

$$\boldsymbol{\alpha} \odot \boldsymbol{\alpha} = \sum_{i,j=1}^{n} \gamma_{i,j} \boldsymbol{r}_p(\boldsymbol{x}_i) \odot \boldsymbol{r}_p(\boldsymbol{x}_j)\,, \tag{5.32}$$

c'est-à-dire comme une combinaison linéaire d'hypernoyaux. En pratique, on accepte de perdre en généralité en ajoutant les contraintes $\gamma_{i,j} \geq 0$ pour garantir la positivité de $\boldsymbol{\alpha} \odot \boldsymbol{\alpha}$ plus facilement d'un point de vue numérique. L'équation (5.32) implique que le terme de régularisation du noyau s'écrit sous la forme

$$\|\boldsymbol{\alpha}\|_4^4 = \sum_{i,j,k,l=1}^{n} \gamma_{i,j} (\boldsymbol{r}_p(\boldsymbol{x}_i) \odot \boldsymbol{r}_p(\boldsymbol{x}_j), \boldsymbol{r}_p(\boldsymbol{x}_k) \odot \boldsymbol{r}_p(\boldsymbol{x}_l)) \gamma_{k,l} = \boldsymbol{\gamma}^\mathsf{T} \underline{\boldsymbol{K}} \boldsymbol{\gamma}\,,$$

où $\boldsymbol{\gamma}$ est le vecteur des $\gamma_{i,j}$ de dimension n^2, et $\underline{\boldsymbol{K}}$ est une matrice $n^2 \times n^2$ dont les éléments s'écrivent $(\boldsymbol{r}_p(\boldsymbol{x}_i) \odot \boldsymbol{r}_p(\boldsymbol{x}_j), \boldsymbol{r}_p(\boldsymbol{x}_k) \odot \boldsymbol{r}_p(\boldsymbol{x}_l))$. Il est possible de montrer comme dans (Ong et al., submitted ; Lanckriet et al., 2004) que minimiser (5.30) conduit à résoudre le problème suivant.

$$\begin{array}{l} \underset{\boldsymbol{\gamma}}{\text{minimiser}}\ \underset{\boldsymbol{\lambda}}{\text{maximiser}} \quad C_\mathrm{k} \boldsymbol{\gamma}^\mathsf{T} \underline{\boldsymbol{K}} \boldsymbol{\gamma} - \boldsymbol{\lambda}^\mathsf{T} (\boldsymbol{K}_{\boldsymbol{\gamma}} + \frac{1}{C_\mathrm{d}} \boldsymbol{I}) \boldsymbol{\lambda} + 2 \boldsymbol{\lambda}^\mathsf{T} \boldsymbol{f}^\mathrm{obs}\,, \\ \text{sous les contraintes } \boldsymbol{\gamma} \geq 0\,, \end{array} \tag{5.33}$$

où $\boldsymbol{K}_{\boldsymbol{\gamma}}$ est une matrice de taille $n \times n$ obtenue en réécrivant le vecteur $\underline{\boldsymbol{K}} \boldsymbol{\gamma}$. Ce problème peut être résolu par programmation semi-definie positive (Ong et al., submitted). La complexité numérique de ce problème est cependant élevée, parce que des objets de dimensions n^4 doivent être calculés. Il peut donc être nécessaire en pratique d'utiliser des approximations permettant de réduire la dimension de ces objets.

L'approche des hypernoyaux est potentiellement intéressante pour le problème du choix de noyau sous forme de combinaison linéaire d'autres noyaux écrites sous la forme (5.31). Notre présentation de cette approche s'est volontairement limitée à des noyaux de dimension finie construits

à partir de vecteurs de caractéristiques $\boldsymbol{r}_p(\boldsymbol{x})$, mais il est possible de considérer des combinaisons linéaires de noyaux tels que des covariances (Ong et al., submitted ; Lanckriet et al., 2004).

Dans la section suivante, nous reprenons l'idée de combiner linéairement des noyaux en empruntant la voie des processus aléatoires. Cette exploration nous conduit à proposer une autre méthode dont l'intérêt devra être examiné dans de futurs travaux.

5.6.3 Sélection de caractéristiques par maximum de vraisemblance

Nous modélisons $f^*(\boldsymbol{x})$ par un processus aléatoire $F(\boldsymbol{x})$ de covariance k. Considérons des combinaisons linéaires $F_{\boldsymbol{\alpha}}(\boldsymbol{x}) = \sum_{i=1}^{l} \alpha_i F_i(\boldsymbol{x})$ de processus aléatoires indépendants $F_i(\boldsymbol{x})$, gaussiens, centrés et de covariance k_i. La covariance de $F_{\boldsymbol{\alpha}}$ s'écrit $k_{\boldsymbol{\alpha}}(\boldsymbol{x},\boldsymbol{y}) = \sum_{i=1}^{l} \alpha_i^2 k_i(\boldsymbol{x},\boldsymbol{y})$, comme dans (5.27). Nous proposons une procédure dont l'objectif est d'obtenir une approximation de la covariance k sous la forme (5.27), obtenue en choisissant des covariances k_i dans un dictionnaire $\mathcal{K}$, comprenant par exemple des covariances anisotropes. Il s'agit d'une procédure itérative où de nouveaux noyaux sélectionnés dans le dictionnaire $\mathcal{K}$ sont inclus dans la combinaison linéaire (5.27) s'ils sont jugés pertinents. Le critère de pertinence est obtenu à partir de la vraisemblance des données. Cet algorithme s'apparente donc à un algorithme de poursuite de caractéristiques.

Soit $F_{\boldsymbol{\alpha}}(\boldsymbol{x}) = \sum_{i=1}^{l} \alpha_i F_i(\boldsymbol{x})$. L'estimée du maximum de vraisemblance du vecteur $\boldsymbol{\alpha} \in \mathbb{R}^l$ des paramètres de la covariance $k_{\boldsymbol{\alpha}}$ s'obtient en maximisant la log-vraisemblance du vecteur des observations $\boldsymbol{f}^{\text{obs}} = (f(\boldsymbol{x}_1), \cdots, f(\boldsymbol{x}_n))^\mathsf{T}$ qui s'écrit

$$L(\boldsymbol{f}^{\text{obs}}, \boldsymbol{\alpha}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log\det \boldsymbol{K}_{\boldsymbol{\alpha}} - \frac{1}{2}\boldsymbol{f}^{\text{obs}\mathsf{T}} \boldsymbol{K}_{\boldsymbol{\alpha}}^{-1} \boldsymbol{f}^{\text{obs}}\,, \tag{5.34}$$

où $\boldsymbol{K}_{\boldsymbol{\alpha}}$ est la matrice de covariance de $\boldsymbol{F}_{\boldsymbol{\alpha}} = (F_{\boldsymbol{\alpha}}(\boldsymbol{x}_1), \cdots, F_{\boldsymbol{\alpha}}(\boldsymbol{x}_n))^\mathsf{T}$. $\boldsymbol{K}_{\boldsymbol{\alpha}}$ s'exprime en fonction des matrices de covariances $\boldsymbol{K}_i$ des vecteurs aléatoires $\boldsymbol{F}_i = (F_i(\boldsymbol{x}_1), \cdots, F_i(\boldsymbol{x}_n))^\mathsf{T}$ sous la forme

$$\boldsymbol{K}_{\boldsymbol{\alpha}} = \sum_{i=1}^{l} \alpha_i^2 \boldsymbol{K}_i\,. \tag{5.35}$$

Le calcul du gradient de $L(\boldsymbol{f}^{\text{obs}}, \boldsymbol{\alpha})$ est particulièrement simple en raison de la forme de $k_{\boldsymbol{\alpha}}$ et on obtient

$$\frac{\partial L(\boldsymbol{f}^{\text{obs}}, \boldsymbol{\alpha})}{\partial \alpha_i} = \alpha_i \left[\boldsymbol{f}^{\text{obs}\mathsf{T}} \boldsymbol{K}_{\boldsymbol{\alpha}}^{-1} \boldsymbol{K}_i \boldsymbol{K}_{\boldsymbol{\alpha}}^{-1} \boldsymbol{f}^{\text{obs}} - \text{trace}(\boldsymbol{K}_{\boldsymbol{\alpha}}^{-1}\boldsymbol{K}_i) \right]. \tag{5.36}$$

L'estimée $\hat{\boldsymbol{\alpha}}$ du maximum de vraisemblance satisfait la condition $\frac{\partial L}{\partial \boldsymbol{\alpha}}(\hat{\boldsymbol{\alpha}}) = \boldsymbol{0}$. Par conséquent, la densité de probabilité estimée de $\boldsymbol{F}_{\boldsymbol{\alpha}}$, que l'on peut écrire sous la forme

$$p_{\hat{\boldsymbol{\alpha}}}(\boldsymbol{f}) = \frac{1}{Z(\hat{\boldsymbol{\alpha}})} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{l} \hat{\alpha_i}^2 \boldsymbol{f}^\mathsf{T} \boldsymbol{K}_{\hat{\boldsymbol{\alpha}}}^{-1} \boldsymbol{K}_i \boldsymbol{K}_{\hat{\boldsymbol{\alpha}}}^{-1} \boldsymbol{f} \right\}, \tag{5.37}$$

satisfait la relation

$$\text{E}_{p_{\hat{\boldsymbol{\alpha}}}}[\Phi_{i,\hat{\boldsymbol{\alpha}}}(\boldsymbol{F}_{\hat{\boldsymbol{\alpha}}})] = \Phi_{i,\hat{\boldsymbol{\alpha}}}(\boldsymbol{f}^{\text{obs}}) \quad \forall i \in \{1, \cdots, l\}\,, \tag{5.38}$$

où $\Phi_{i,\boldsymbol{\alpha}}(\boldsymbol{f}) = \boldsymbol{f}^\mathsf{T} \boldsymbol{K}_{\boldsymbol{\alpha}}^{-1} \boldsymbol{K}_i \boldsymbol{K}_{\boldsymbol{\alpha}}^{-1} \boldsymbol{f}$, $\boldsymbol{f} \in \mathbb{R}^n$. En effet, pour tout $\boldsymbol{\alpha} \in \mathbb{R}^l$

$$\begin{aligned} \text{E}_{p_{\boldsymbol{\alpha}}} \left[\boldsymbol{F}_{\boldsymbol{\alpha}}^\mathsf{T} \boldsymbol{K}_{\boldsymbol{\alpha}}^{-1} \boldsymbol{K}_i \boldsymbol{K}_{\boldsymbol{\alpha}}^{-1} \boldsymbol{F}_{\boldsymbol{\alpha}} \right] &= \text{E}\left[\text{trace}(\boldsymbol{K}_{\boldsymbol{\alpha}}^{-1} \boldsymbol{K}_i \boldsymbol{K}_{\boldsymbol{\alpha}}^{-1} \boldsymbol{F}_{\boldsymbol{\alpha}} \boldsymbol{F}_{\boldsymbol{\alpha}}^\mathsf{T}) \right] \\ &= \text{trace}(\boldsymbol{K}_{\boldsymbol{\alpha}}^{-1} \boldsymbol{K}_i)\,. \end{aligned} \tag{5.39}$$

Comme nous l'avons mentionné dans la section 5.4.4, les statistiques $\Phi_{i,\boldsymbol{\alpha}}(\boldsymbol{f}^{\text{obs}})$, $i \in \{1, \cdots, l\}$, peuvent être interprétées comme des normes des vecteurs $\boldsymbol{f}^{\text{obs}}$ dans les espaces à noyau reproduisant $k_i(\boldsymbol{x}, \boldsymbol{y})$. En calculant les statistiques $\Phi_{i,\boldsymbol{\alpha}}(\boldsymbol{f}^{\text{obs}})$ nous pouvons évaluer les caractéristiques du processus aléatoire $F(\boldsymbol{x})$ qui a généré les données. Les fonctions $\Phi_{i,\boldsymbol{\alpha}}$ s'interprètent donc comme des fonctions qui extraient de l'information des données. La densité (5.37) estimée par maximum de vraisemblance est celle d'entropie maximale parmi toutes les densités satisfaisant les contraintes (5.38). Autrement dit, (5.37) est la densité la moins informative parmi toutes les densités qui reproduisent les statistiques observées $\Phi_{i,\boldsymbol{\alpha}}(\boldsymbol{f}^{\text{obs}})$. Le principe du maximum d'entropie est rappelé plus en détails dans la section 5.6.4 (nous y expliquons également pourquoi nous préférons l'approche du maximum de vraisemblance à celle du maximum d'entropie).

Pour juger de la pertinence de l'ajout d'un noyau k_{l+1} dans une combinaison (5.27) comportant l termes, nous pouvons comparer les statistiques observées ${\boldsymbol{f}^{\text{obs}}}^{\mathsf{T}} \boldsymbol{K}_{\boldsymbol{\alpha}}^{-1} \boldsymbol{K}_{l+1} \boldsymbol{K}_{\boldsymbol{\alpha}}^{-1} \boldsymbol{f}^{\text{obs}}$, où $\boldsymbol{K}_{l+1}$ est la matrice de covariance correspondant au noyau k_{l+1} avec l'espérance de ces statistiques calculée par la relation (5.39) sous le modèle courant (comportant l termes). Si $\boldsymbol{K}_{\boldsymbol{\alpha}}^{-1}$ est préalablement calculée, chaque comparaison nécessite de l'ordre de $O(n^2)$ opérations, ce qui permet d'envisager de tester un grand nombre de noyaux. Une telle procédure nous paraît d'un coût moins élevé que celui qui consisterait à estimer tous les coefficients α_i d'une combinaison linéaire d'un très grand nombre de termes, correspondant par exemple à tous les noyaux d'un dictionnaire $\mathcal{K}$. Dans les paragraphes suivants, nous justifions cette comparaison à l'aide d'un calcul sur la vraisemblance.

Poursuite de caractéristiques

Posons $\boldsymbol{\alpha}_- = (\hat{\boldsymbol{\alpha}}^{\mathsf{T}}, 0)^{\mathsf{T}} \in \mathbb{R}^{l+1}$, où $\hat{\boldsymbol{\alpha}}$ désigne le vecteur des paramètres α_i, $i = 1, \ldots, l$, de la combinaison (5.27), estimés par maximum de vraisemblance. Considérons le développement de Taylor à l'ordre deux de la log–vraisemblance des données sous le modèle de densité (5.37) comportant $l+1$ termes paramétriques. Pour tout $\boldsymbol{\alpha}_+ \in \mathbb{R}^{l+1}$, on a

$$L(\boldsymbol{f}, \boldsymbol{\alpha}_+) = L(\boldsymbol{f}, \boldsymbol{\alpha}_-) + (\boldsymbol{\alpha}_+ - \boldsymbol{\alpha}_-)^{\mathsf{T}} \frac{\partial L}{\partial \boldsymbol{\alpha}}(\boldsymbol{f}, \boldsymbol{\alpha}_-) + \frac{1}{2}(\boldsymbol{\alpha}_+ - \boldsymbol{\alpha}_-)^{\mathsf{T}} \frac{\partial^2 L}{\partial \boldsymbol{\alpha}^2}(\boldsymbol{f}, \boldsymbol{\alpha}_-)(\boldsymbol{\alpha}_+ - \boldsymbol{\alpha}_-) + \mathrm{o}(\|\boldsymbol{\alpha}_+ - \boldsymbol{\alpha}_-\|_2^2)\,. \tag{5.40}$$

Nous avons déjà constaté que les dérivées premières de la log–vraisemblance pouvaient s'écrire sous la forme

$$\begin{aligned} \frac{\partial L}{\partial \alpha_i}(\boldsymbol{f}, \boldsymbol{\alpha}) &= \alpha_i \Phi_i(\boldsymbol{f}) - \alpha_i \mathrm{trace}(\boldsymbol{K}_{\boldsymbol{\alpha}}^{-1} \boldsymbol{K}_i) \\ &= \alpha_i (\Phi_i(\boldsymbol{f}) - \mathrm{E}_{p_{\boldsymbol{\alpha}}}[\Phi_i(\boldsymbol{F}_{\boldsymbol{\alpha}})])\,. \end{aligned}$$

On a donc $\frac{\partial L}{\partial \boldsymbol{\alpha}}(\boldsymbol{f}, \boldsymbol{\alpha}_-) = \mathbf{0}$ puisque $\forall i \in \{1, \cdots, l\}$, $\Phi_i(\boldsymbol{f}) = \mathrm{E}_{p_{\boldsymbol{\alpha}_-}}[\Phi_i(\boldsymbol{F})]$, et $\alpha_{-[l+1]} = 0$. Les dérivées secondes peuvent se mettre sous la forme

$$\begin{aligned} \frac{\partial^2 L}{\partial \alpha_i^2}(\boldsymbol{f}, \boldsymbol{\alpha}) &= \Phi_i(\boldsymbol{f}) - \mathrm{E}_{p_{\boldsymbol{\alpha}}}[\Phi_i(\boldsymbol{F})] + \mathrm{Var}_{p_{\boldsymbol{\alpha}}}[\alpha_i \Phi_i(\boldsymbol{F})] \\ &\quad -4\alpha_i^2 \boldsymbol{f}^{\mathsf{T}} (\boldsymbol{K}_{\boldsymbol{\alpha}}^{-1} \boldsymbol{K}_i)^2 \boldsymbol{K}_{\boldsymbol{\alpha}}^{-1} \boldsymbol{f} \end{aligned} \tag{5.41}$$

et, pour $i \neq j$,

$$\begin{aligned}\frac{\partial^2 L}{\partial\alpha_i\partial\alpha_j}(\boldsymbol{f},\boldsymbol{\alpha}) &= \mathrm{Cov}_{p_{\boldsymbol{\alpha}}}[\alpha_i\Phi_i(\boldsymbol{F}),\alpha_j\Phi_j(\boldsymbol{F})] \\ &\quad -2\alpha_i\alpha_j\boldsymbol{f}^{\mathsf{T}}\boldsymbol{K}_{\boldsymbol{\alpha}}^{-1}(\boldsymbol{K}_i\boldsymbol{K}_{\boldsymbol{\alpha}}^{-1}\boldsymbol{K}_j+\boldsymbol{K}_j\boldsymbol{K}_{\boldsymbol{\alpha}}^{-1}\boldsymbol{K}_i)\boldsymbol{K}_{\boldsymbol{\alpha}}^{-1}\boldsymbol{f}\,. \qquad (5.42)\end{aligned}$$

Pour simplifier, nous choisissons d'approximer le développement de Taylor en ne retenant que l'influence de la nouvelle statistique Φ_{l+1}. D'après (5.41), nous obtenons alors l'approximation empirique

$$L(\boldsymbol{f},\boldsymbol{\alpha}_+)-L(\boldsymbol{f},\boldsymbol{\alpha}_-)\approx\Phi_{l+1}(\boldsymbol{f})-\mathrm{E}_{p_{\boldsymbol{\alpha}}}[\Phi_{l+1}(\boldsymbol{F})]=\boldsymbol{f}^{\mathsf{T}}\boldsymbol{K}_{\boldsymbol{\alpha}}^{-1}\boldsymbol{K}_{l+1}\boldsymbol{K}_{\boldsymbol{\alpha}}^{-1}\boldsymbol{f}-\mathrm{trace}(\boldsymbol{K}_{\boldsymbol{\alpha}}^{-1}\boldsymbol{K}_{l+1})\,. \quad (5.43)$$

L'approximation (5.43) nous permet ainsi de justifier la comparaison des statistiques observées avec l'espérance de ces statistiques sous le modèle courant. En effet, le calcul de cette différence permet de choisir un noyau qui conduira potentiellement à l'augmentation maximale de la vraisemblance des observations. Nous pouvons envisager une procédure itérative, qui partant d'une densité de probabilité a priori (celle d'un bruit blanc, par exemple), ajoute à étape l un nouveau noyau choisi dans un dictionnaire $\mathcal{K}$ selon le principe

$$k_{l+1}=\arg\max_{k_i\in\mathcal{K}}\quad\Phi_i(\boldsymbol{f}^{\mathrm{obs}})-\mathrm{E}_{p_{\boldsymbol{\alpha}}}[\Phi_i(\boldsymbol{F})]\,. \qquad (5.44)$$

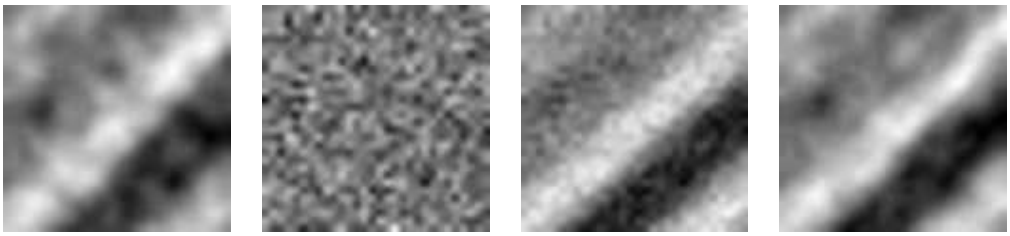
L'écart (5.43) peut encore s'interpréter comme un critère au sens de la théorie de l'information. En effet, si cet écart est petit, alors l'ajout du noyau k_{l+1} dans (5.27) est probablement redondant. À l'inverse, si l'écart est grand, alors le modèle courant n'est pas capable d'expliquer les données de manière correcte et ce modèle est susceptible d'être enrichi en ajoutant k_{l+1}. Nous invitons le lecteur à comparer notre approche avec celle présentée dans (Zhu et al., 1997), rappelée dans la section 5.6.4, où un principe minimax d'entropie est utilisé pour modéliser des textures dans une image (les similarités entre les deux approches viennent du fait que la méthode du maximum de vraisemblance obéit au principe du maximum d'entropie).

Exemples numériques

Dans les paragraphes suivants, l'algorithme de poursuite de caractéristiques proposé ci-dessus est appliqué à deux exemples où les covariances à estimer sont caractérisées par des anisotropies importantes. Ces exemples permettent de juger de la pertinence du critère retenu pour l'ajout de noyau.

Commençons par un exemple en dimension deux, qui permettra de présenter facilement des figures. Soient k_1 et k_2 des covariances de Matérn de portée 0.1 et 0.3 respectivement (les autres paramètres sont choisis unitaires pour simplifier). Posons $k(\boldsymbol{x},\boldsymbol{y})=0.4^2\,k_1(\|\boldsymbol{x}-\boldsymbol{y}\|_2)+k_2(\boldsymbol{u}^{\mathsf{T}}(\boldsymbol{x}-\boldsymbol{y}))$, où le deuxième terme constitue une covariance bande orientée dans une direction formant un angle de 45° avec l'axe des abscisses ($\boldsymbol{u}^{\mathsf{T}}=[1/\sqrt{2}\quad 1/\sqrt{2}]$). Soit $F(\boldsymbol{x})$ un processus aléatoire gaussien centré et de covariance $k(\boldsymbol{x},\boldsymbol{y})$. Une simulation de $F(\boldsymbol{x})$ sur $[0,1]^2$ est représentée à la figure 5.16-(a).

Nous considérons un dictionnaire $\mathcal{K}$ comprenant des covariances anisotropes de type bande, construites à partir de fonctions de Matérn, en choisissant des portées et des angles régulièrement répartis. Notons que ce dictionnaire est très simple mais possède un caractère très empirique, ce qui



(a) (b) (c) (d)

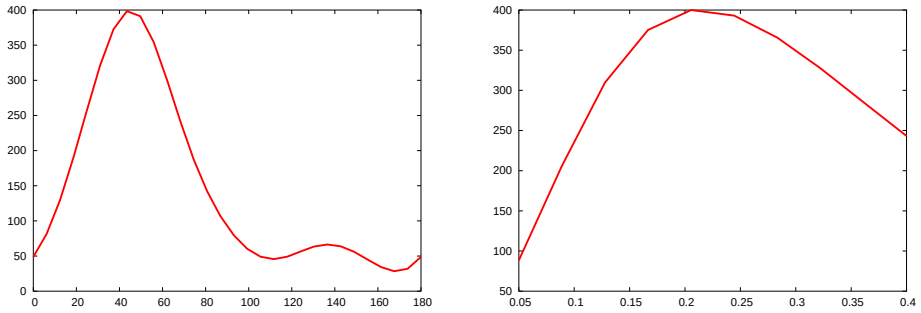
FIG. 5.16 – Réalisations d'un processus aléatoire en dimension deux obtenues par la méthode de décomposition de Cholesky de la matrice de covariance utilisant (a) : la vraie covariance, (b) : la covariance de l'« effet de pépites », (c) : le mélange d'une covariance bande et de l'« effet de pépites » obtenu après la première itération de l'algorithme de poursuite, (d) : la combinaison linéaire de huit covariances obtenue par le même algorithme. Pour permettre une comparaison visuelle, la même réalisation de bruit blanc a été utilisée pour générer les différentes réalisations des processus aléatoires.

pourrait être amélioré. Un tel dictionnaire ne constitue pas une base de fonctions de covariances (avec les défauts que cela entraîne, comme par exemple le fait que la meilleure approximation de la vraie covariance peut ne pas être unique).

Pour estimer k, nous sélectionnons un ensemble de 40 observations d'une réalisation de $F(\boldsymbol{x})$, réparties irrégulièrement sur $[0,1]^2$. L'algorithme d'approximation de la covariance k est initialisé avec une covariance de type « effet de pépites », dont la variance est estimée par maximum de vraisemblance. Ensuite, l'écart (5.43) est calculé pour toutes les covariances de $\mathcal{K}$. La figure 5.17 représente l'évolution de cet écart avec l'orientation et la portée d'une covariance bande. Notons que l'anisotropie est détectée de manière efficace par le critère de pertinence (5.43). Des réalisations des processus aléatoires approximant $F(\boldsymbol{x})$ sont présentées à différentes étapes de l'algorithme de poursuite dans les figures 5.16-(b-d). La figure 5.16-(b) correspond à l'étape d'initialisation, où l'on a seulement l'« effet de pépite ». La figure 5.16-(c) est obtenue après la première itération et on constate que la partie anisotrope a été correctement capturée. Dans les étapes suivantes, l'algorithme incorpore successivement des covariances bandes de portée 0.1 et d'angles 74°, 172°, 41°, 147°, 16° et enfin 123°. Ces angles sont distribués approximativement uniformément sur $[0°, 180°]$, et on constate que l'algorithme essaye d'approximer la composante isotrope de la vrai covariance. La figure 5.16-(d) montre une réalisation du processus caractérisé par une combinaison linéaire de huit covariances de $\mathcal{K}$. Notons que le nombre de covariances nécessaires pour approximer $k(\boldsymbol{x}, \boldsymbol{y})$ de manière satisfaisante dépend du choix de $\mathcal{K}$.

Comme second exemple, considérons un processus aléatoire $F(\boldsymbol{x})$ gaussien, centré, de covariance $k(\boldsymbol{x}, \boldsymbol{y})$, défini sur un espace de facteurs de dimension dix. La procédure d'estimation utilise un ensemble $\mathcal{S}$ de 200 points répartis aléatoirement sur $\mathbb{X} = [0,1]^{10}$. La covariance k est une combinaison de deux covariances bande $k_i(\boldsymbol{x}, \boldsymbol{y}) = \exp(-0.5(\boldsymbol{u}_i^\mathsf{T}(\boldsymbol{x} - \boldsymbol{y}))^2/\sigma_i^2)$, $i = 1, 2$ (les angles d'orientation sont aléatoires). Le dictionnaire $\mathcal{K}$ comporte 400 covariances bandes orientées aléatoirement (le dictionnaire est reconstruit à chaque étape de l'algorithme de poursuite).

La qualité du modèle de covariance est jugée par validation croisée en utilisant la procédure suivante.



(a) (b)

FIG. 5.17 – Critère de pertinence (5.43) calculé avec des covariances bandes. (a) : en fonction de l'angle en degré pour une portée de 0.2. (b) : en fonction de la portée pour un angle de 45°.

1. L'ensemble $\mathcal{S}$ de taille 200 est divisé en deux sous-ensembles $\mathcal{S}^{\text{obs}}$ et $\mathcal{S}^{\text{test}}$ de tailles 100.
2. Nous calculons ensuite les prédictions par krigeage des données de $\mathcal{S}^{\text{test}}$ en utilisant les données de $\mathcal{S}^{\text{obs}}$.
3. Nous évaluons enfin les moyennes quadratiques des écarts entre les prédictions et les vrais valeurs.

Le tableau 5.2 présente les moyennes et les écarts type de ces erreurs lorsque l'on répète les étapes 1 à 3 ci-dessus pour différentes dichotomies de $\mathcal{S}$.

Nous constatons que l'erreur de prédiction est très petite si l'on utilise la vraie covariance. Ceci s'explique par le fait que la dimension intrinsèque de l'espace des facteurs est seulement deux. Cependant si l'on utilise une version légèrement perturbée de la vraie covariance, en remplaçant $\boldsymbol{u}_2$ par $(\boldsymbol{u}_2 + 0.01\boldsymbol{u}_1)/\|\boldsymbol{u}_2 + 0.01\boldsymbol{u}_1\|$, l'erreur de prédiction est considérablement dégradée, comme on peut le voir dans la troisième colonne. Ceci illustre l'importance d'estimer correctement l'anisotropie des données. L'algorithme de poursuite est arrêté après avoir obtenu une combinaison de 25 covariances. Nous atteignons l'erreur de prédiction indiquée dans la quatrième colonne. Même si ce résultat reste sous-optimal en comparaison de ce que l'on obtient avec la vraie covariance, nous constatons une amélioration sensible par rapport à la version perturbée de la vraie covariance. Dans la dernière colonne, nous présentons le résultat obtenu avec une covariance séparable gaussienne (un modèle classique) $k_{\boldsymbol{w}}(\boldsymbol{x}, \boldsymbol{y}) = \exp(-0.5 \sum_{i=1}^{d} w_{[i]}(x_{[i]} - y_{[i]})^2)$, où le vecteur des paramètres $w_{[i]}$ a été estimé par maximum de vraisemblance. L'algorithme de poursuite obtient une covariance presque deux fois meilleure que la covariance gaussienne séparable. Ce résultat illustre encore une fois l'importance d'estimer correctement le noyau avant d'effectuer une prédiction.

covariance	vraie	perturbée	poursuite de caractéristiques	gaussienne
moyenne (écart type)	0.05 (0.02)	0.48 (0.35)	0.29 (0.03)	0.56 (0.03)

TAB. 5.2 – Moyenne et écart type empirique de la moyenne quadratique de l'erreur de prédiction pour différentes covariances.

5.6.4 Approche du maximum d'entropie

Cette section explique brièvement comment adapter la méthode d'estimation de textures par maximum d'entropie (Zhu et al., 1997) à la construction de noyaux. Une application directe de cette méthode ne conduit pas toutefois à des combinaisons linéaires de noyaux, et c'est la raison pour laquelle nous avons préféré la méthode présentée ci-dessus.

Soit $\mathcal{P}_\Phi$, l'ensemble de toutes les densités de probabilités p du vecteur aléatoire $\boldsymbol{F}$ satisfaisant les contraintes

$$\mathrm{E}_p[\Phi_i(\boldsymbol{F})] = \phi_i\,, \quad \forall i \in \{1, \cdots, l\}\,, \tag{5.45}$$

où les statistiques Φ_i sont des fonctions mesurables. La forme de la densité de probabilité p d'entropie $h(p)$ maximale est donnée par le théorème suivant (voir par exemple (Zhu et al., 1997)).

Théorème 29. *Soit $p_{\boldsymbol{\alpha}}(\boldsymbol{f}) = Z(\boldsymbol{\alpha})^{-1} \exp(-\sum_{i=1}^{l} \alpha_i \Phi_i(\boldsymbol{f}))$, où $Z(\boldsymbol{\alpha})$ est une constante de normalisation, et $\boldsymbol{\alpha}$ est choisi tel que $p_{\boldsymbol{\alpha}} \in \mathcal{P}_\Phi$. Alors, pour tous $p \in \mathcal{P}_\Phi$, la divergence de Kullback-Leibler $\mathrm{D}(p||p_{\boldsymbol{\alpha}}) \triangleq \mathrm{E}_p[\log p(\boldsymbol{F})] - \mathrm{E}_p[\log p_{\boldsymbol{\alpha}}(\boldsymbol{F})]$, peut être écrite sous la forme*

$$\mathrm{D}(p||p_{\boldsymbol{\alpha}}) = h(p_{\boldsymbol{\alpha}}) - h(p)\,.$$

Il s'ensuit que $p_{\boldsymbol{\alpha}}$ est la densité d'entropie maximale parmi toutes les densités de $\mathcal{P}_\Phi$.

La forme de la densité d'entropie maximale est donc déterminée par le choix des statistiques. Considérons des statistiques écrites sous la forme

$$\Phi_i(\boldsymbol{f}) = \boldsymbol{f}^\mathsf{T} \boldsymbol{K}_i^- \boldsymbol{f}\,.$$

Ces statistiques sont choisies telles que la densité d'entropie maximale qui en découle corresponde à la densité estimée par maximum de vraisemblance (5.35) lorsque les matrices de covariances $\boldsymbol{K}_i$ possèdent des espaces image orthogonaux.

La matrice de covariance de $\boldsymbol{F}$ estimée par maximum d'entropie s'écrit alors sous la forme

$$\boldsymbol{K}_{\boldsymbol{\alpha}} = \frac{1}{2}\left[\sum_{i=1}^{l} \alpha_i \boldsymbol{K}_i^-\right]^{-1}\,. \tag{5.46}$$

Par conséquent, le noyau obtenu n'est plus une combinaison linéaire de noyaux (sauf si les espaces images des $\boldsymbol{K}_i$ sont orthogonaux). On peut toutefois s'attendre à ce que les matrices (5.35) et (5.46) soient assez proches l'une de l'autre (elles sont associées à des normes dans des espaces à noyaux reproduisants construits à partir de caractéristiques identiques).

Afin de calculer $\hat{\boldsymbol{\alpha}}$ tel que

$$\mathrm{E}_{p_{\hat{\boldsymbol{\alpha}}}}[\Phi_i(\boldsymbol{F})] = \Phi_i(\boldsymbol{f}^{\mathrm{obs}}) \quad \forall i \in \{1, \cdots, l\}$$

il faut résoudre $G_i(\boldsymbol{\alpha}) = \mathrm{E}_{p_{\boldsymbol{\alpha}}}[\Phi_i(\boldsymbol{F})] - \Phi_i(\boldsymbol{f}^{\mathrm{obs}}) = \mathrm{trace}(\boldsymbol{K}_i^- \boldsymbol{K}_{\boldsymbol{\alpha}}) - \boldsymbol{f}^{\mathrm{obs}\mathsf{T}} \boldsymbol{K}_i^- \boldsymbol{f}^{\mathrm{obs}} = \boldsymbol{0}$, pour tout i.

Une fois la matrice de covariance estimée, l'étape suivante consiste à ajouter un nouveau noyau. Dans (Zhu et al., 1997), le principe de cette extension est fondé sur la différence entre l'espérance des statistiques Φ_i sous le modèle gouverné par la densité de probabilité actuelle et la valeur de $\Phi_i(\boldsymbol{f}^{\mathrm{obs}})$. Il est possible de montrer que cette différence est une approximation de la divergence de Kullback-Leibler $\mathrm{D}[p_{\boldsymbol{\alpha}_+}||p_{\hat{\boldsymbol{\alpha}_-}}]$ entre les densités de probabilité future et actuelle (Kullback, 1968, Théorème 4.1, p. 47). Cette divergence doit être maximisée si l'on veut minimiser l'entropie de la future densité, et ce faisant, améliorer le modèle.

5.6.5 Conclusions et perspectives

L'objectif poursuivi est de permettre plus de flexibilité dans le choix des noyaux par rapport aux méthodes traditionnelles (où ces derniers sont sélectionnés dans des familles de fonctions de type positif à faible nombre de paramètres). Les approches traditionnelles nous semblent adaptées dans des espaces de facteurs de dimension faible mais se révèlent en pratique limitées lorsque la dimension croît.

En particulier, l'estimation de structures anisotropes par les méthodes classiques se révèle difficile et coûteuse, comme nous le verrons dans l'application présentée à la section 6.4.2. L'utilisation de combinaisons de noyaux simples pour former des noyaux plus élaborés nous paraît être une approche prometteuse. Nous avons présenté une nouvelle méthode de construction de noyau exploitant cette idée. Cette méthode fournit une alternative à l'approche fondée sur les hypernoyaux avec laquelle une comparaison de performance reste à faire.

L'algorithme de poursuite a été appliqué à quelques exemples simples avec des structures anisotropes marquées. Ces exemples montrent qu'il est effectivement possible d'éviter l'optimisation d'un critère d'adéquation aux données dépendant d'un nombre très élevé de paramètres. Ces expériences numériques suggèrent que le critère de pertinence (5.43) donne en pratique les résultats attendus. Entre autres perspectives de cette étude, nous pourrions nous intéresser davantage aux propriétés des critères du type (5.43). Si ces premiers résultats paraissent satisfaisants, ils ne garantissent pas que la méthode se généralise à des cas plus difficiles. Il nous semble qu'il faudrait davantage réfléchir au choix des dictionnaires de covariances. Enfin, notons que la procédure pourrait également s'appliquer à l'estimation de covariances non stationnaires par combinaison de noyaux localement stationnaires.

Chapitre 6

Exemples

Résumé — Ce chapitre est constitué d'une série d'exemples d'application du krigeage à des problèmes de nature variée. Par ces exemples, nous souhaitons illustrer la pertinence des modèles boîte noire présentés aux chapitres précédents. Nous insistons sur les aspects méthodologiques. Les problèmes sont présentés selon un ordre croissant de la dimension de l'espace des facteurs. Plus cette dimension est élevée, plus le problème de prédiction est délicat. Ce chapitre illustre également les difficultés liées aux faibles nombres d'observations, les méthodes de prise en compte des connaissances a priori et le problème de la planification d'expériences.

6.1 Exemples élémentaires de krigeage et de cokrigeage

6.1.1 Caractères typiques d'une prédiction par krigeage

Les figures 6.1 à 6.3 représentent des interpolations en dimension un. Elles illustrent deux propriétés fondamentales du krigeage.

Propriété de lissage. La prédiction par krigeage varie plus doucement que les trajectoires du processus aléatoire prédit. En effet, le krigeage est une *moyenne* des trajectoires probables conditionnellement aux observations.

Intervalles de confiance. Dans les figures présentées, le modèle aléatoire générant les trajectoires est utilisé pour déterminer les intervalles de confiance des prédictions. Il n'est donc pas surprenant que ces intervalles de confiances soient pertinents. L'incertitude de prédiction augmente loin des observations mais d'autant moins vite que le modèle est plus régulier. On constate que l'incertitude de prédiction est plus grande dans les zones d'extrapolation que dans les zones d'interpolation.

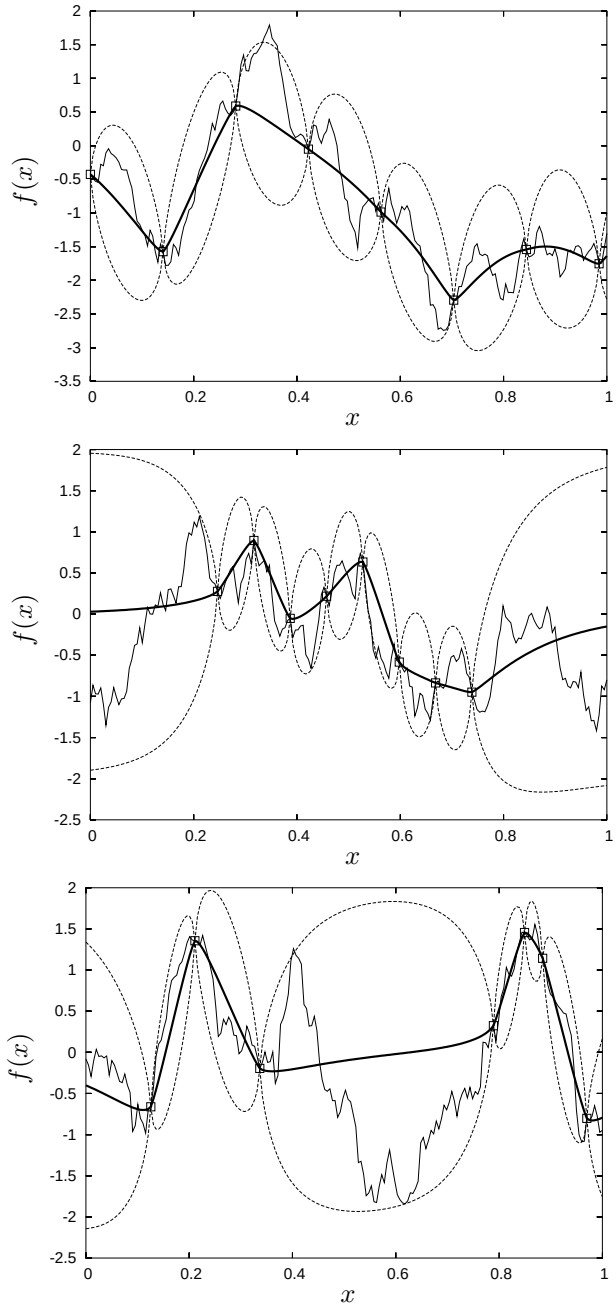


FIG. 6.1 – Exemples en dimension 1. On observe une trajectoire (en trait fin) d'un processus gaussien $F(x)$ de moyenne nulle et de covariance $k(h)$, où k est une covariance de Matérn de paramètres $\nu = 0.7$, $\rho = 0.2$ et $\sigma_0^2 = 1$. Les observations, matérialisées par les carrés, sont interpolées par la courbe en trait gras continu obtenue par krigeage. Dans la figure du haut, l'échantillonnage est régulier sur $[0, 1]$; dans celle du milieu l'échantillonnage est régulier sur $[0.25, 0.75]$; dans celle du bas l'échantillonnage est aléatoire. Les courbes en pointillés représentent les intervalles de confiance à 95% pour la prédiction.

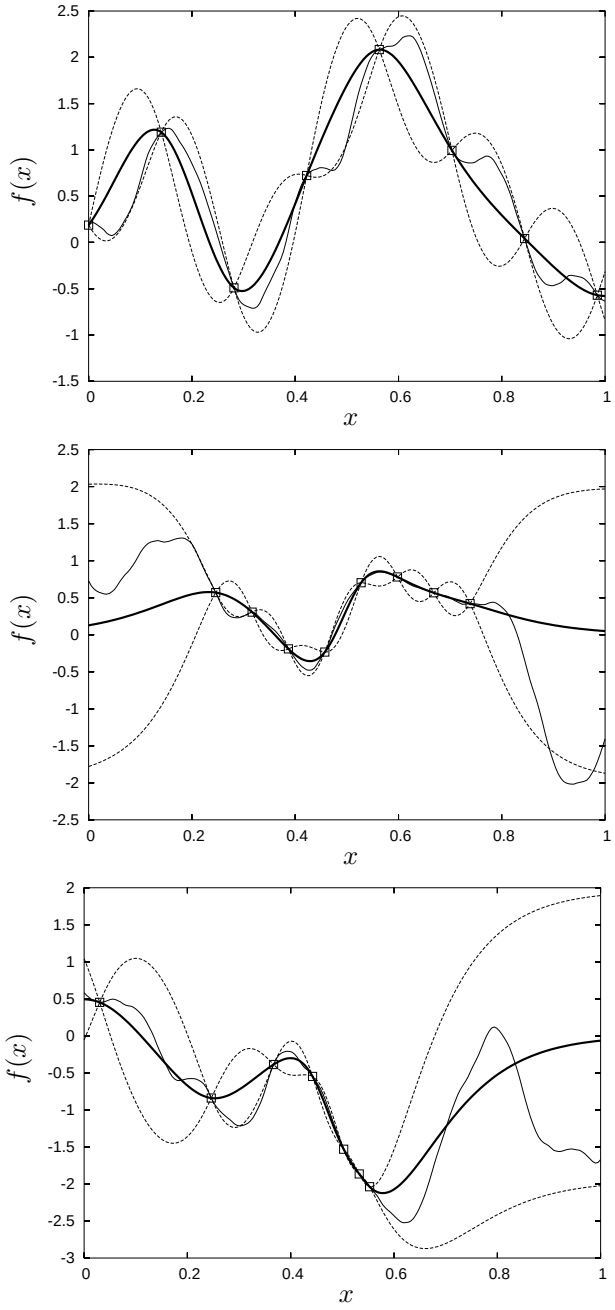


FIG. 6.2 – Exemples en dimension 1. On observe une trajectoire (en trait fin) d'un processus gaussien $F(x)$ de moyenne nulle et de covariance $k(h)$, où k est une covariance de Matérn de paramètres $\nu = 2$, $\rho = 0.2$ et $\sigma_0^2 = 1$. Les observations, matérialisées par les carrés, sont interpolées par la courbe en trait gras continu obtenue par krigeage. Dans la figure du haut, l'échantillonnage est régulier sur $[0, 1]$; dans celle du milieu l'échantillonnage est régulier sur $[0.25, 0.75]$; dans celle du bas l'échantillonnage est aléatoire. Les courbes en pointillés représentent les intervalles de confiance à 95% pour la prédiction.

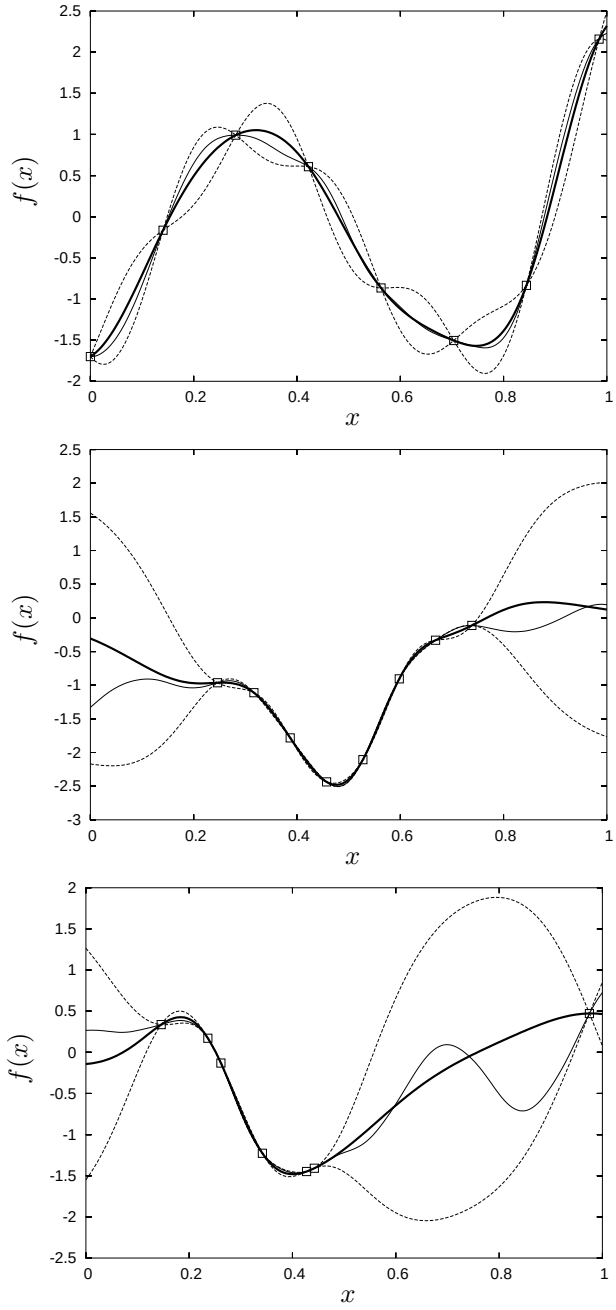


FIG. 6.3 – Exemples en dimension 1. On observe une trajectoire (en trait fin), d'un processus gaussien $F(x)$ de moyenne nulle et de covariance $k(h)$, où k est une covariance de Matérn de paramètres $\nu = 5$, $\rho = 0.2$ et $\sigma_0^2 = 1$. Les observations, matérialisées par les carrés, sont interpolées par la courbe en trait gras continu obtenue par krigeage. Dans la figure du haut, l'échantillonnage est régulier sur $[0, 1]$; dans celle du milieu l'échantillonnage est régulier sur $[0.25, 0.75]$; dans celle du bas l'échantillonnage est aléatoire. Les courbes en pointillés représentent les intervalles de confiance à 95% pour la prédiction.

6.1.2 Cokrigeage et application à la prédiction de dérivées

Le principe du cokrigeage, c'est-à-dire la prédiction linéaire à partir de plusieurs processus aléatoires, a été rappelé dans la section 2.5.2. Il est pertinent d'utiliser cette méthode lorsque le système considéré possède plusieurs sorties. En effet, il existe en général des corrélations entre ces sorties et la connaissance de ces corrélations est susceptible d'améliorer les prédictions obtenues à partir d'observations effectuées sur les différentes sorties. Dans cette section, nous considérons l'application du cokrigeage à l'estimation d'une dérivée d'une sortie.

Prédiction de la dérivée d'une sortie.

La figure 6.4 représente une sortie et sa dérivée. À partir d'un nombre d'observations sans bruit irrégulièrement espacées de cette sortie, nous constatons que la prédiction de la dérivée de la sortie est satisfaisante. Nous pouvons également donner des intervalles d'incertitude de prédiction, ce qui est un avantage de cette méthode. Nous constatons que les intervalles d'incertitudes sont plus resserrés dans les zones où les observations ont été effectuées mais, fait marquant, lorsque deux observations sont suffisamment proches, la prédiction est meilleure *entre* les observations.

Nous répétons l'expérience dans la figure 6.5, en ajoutant cette fois du bruit sur les observations. Ces expériences montrent qu'il est possible d'estimer de manière très simple et satisfaisante les dérivées d'une fonction, éventuellement bruitée. Les possibilités d'application de tels prédicteurs nous semblent très nombreuses.

Prédiction d'une dérivée avec modèle connu a priori

Dans la section 4.6.1, nous avons montré comment prendre en compte de l'information connue a priori en utilisant des modèles semi-paramétriques, dans lesquels la partie paramétrique porte l'information connue. La figure 6.6 illustre cette possibilité dans le cas de la prédiction d'une dérivée. La fonction à dériver est la somme de deux fonctions dont l'une est supposée parfaitement connue.

Prédiction d'une sortie avec informations sur la dérivée.

Les exemples précédents n'épuisent pas toutes les possibilités du cokrigeage appliqué à la prédiction des dérivées d'une sortie. Dans la figure 6.7, nous utilisons à la fois des observations de la fonction et de sa dérivée. Ceci permet d'améliorer une prédiction par exemple lorsque l'on connaît des comportements du système aux limites de son domaine de définition, telles que des gradients nuls. Il s'agit donc d'une autre possibilité pour incorporer de l'information connue a priori dans un modèle boîte noire. Dans (Morris et al., 1993), ce problème est abordé du point de vue de la modélisation de résultats de simulation et de la planification d'expériences.

Approximation d'une fonction à partir d'observations sur sa dérivée. Intégration.

Pour conclure cette section, nous appliquons le cokrigeage à l'intégration. Formellement, nous pouvons nous ramener au cas précédent. Il s'agit en effet de prédire une fonction $f(x)$, en considérant des observations effectuées sur sa dérivée. Notons qu'il faut au moins une observation sur

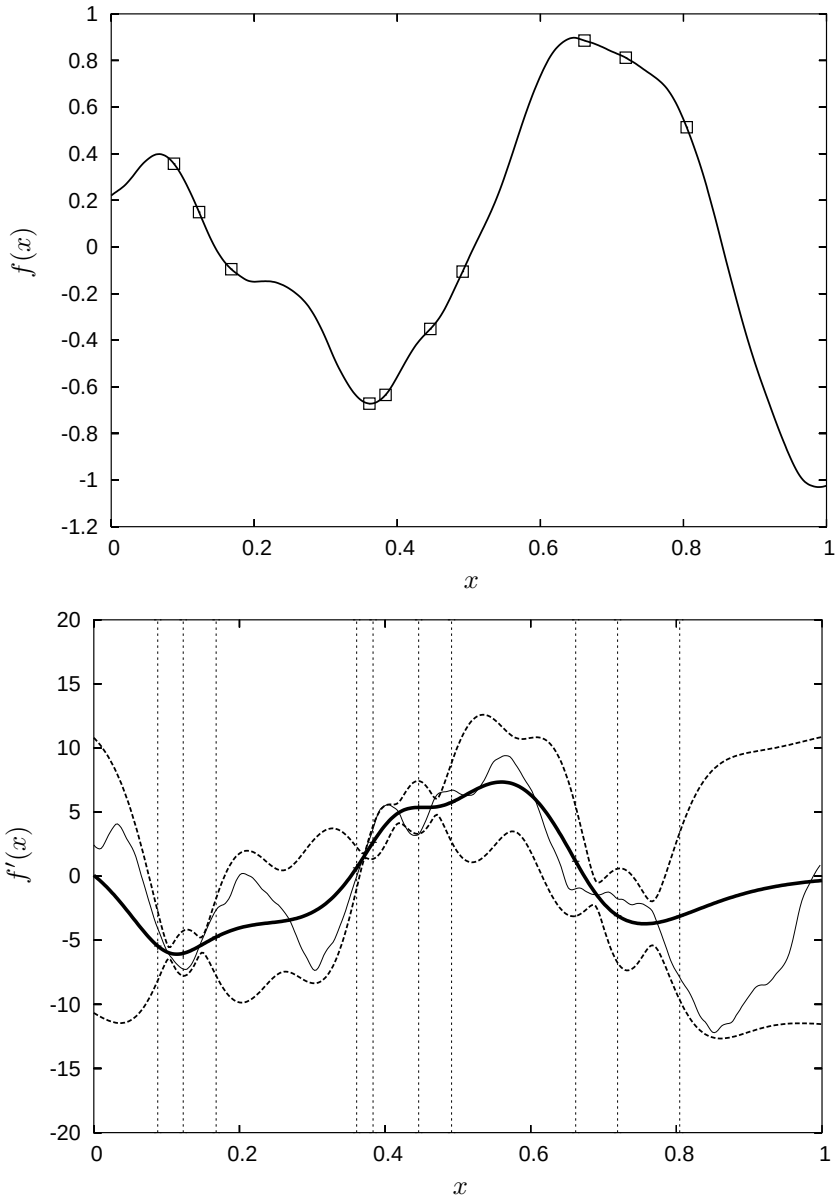


FIG. 6.4 – Figure du haut : 10 observations (matérialisées par des carrés) non bruitées irrégulièrement échantillonnées d'une fonction $f(x)$, $x \in [0, 1]$. Figure du bas : estimée (en trait gras continu) de la dérivée de $f(x)$ à partir des observations précédentes. La dérivée exacte est représentée en trait continu fin et les intervalles d'incertitude à 95% sont indiqués en traits interrompus. Les barres verticales rappellent la position des observations.

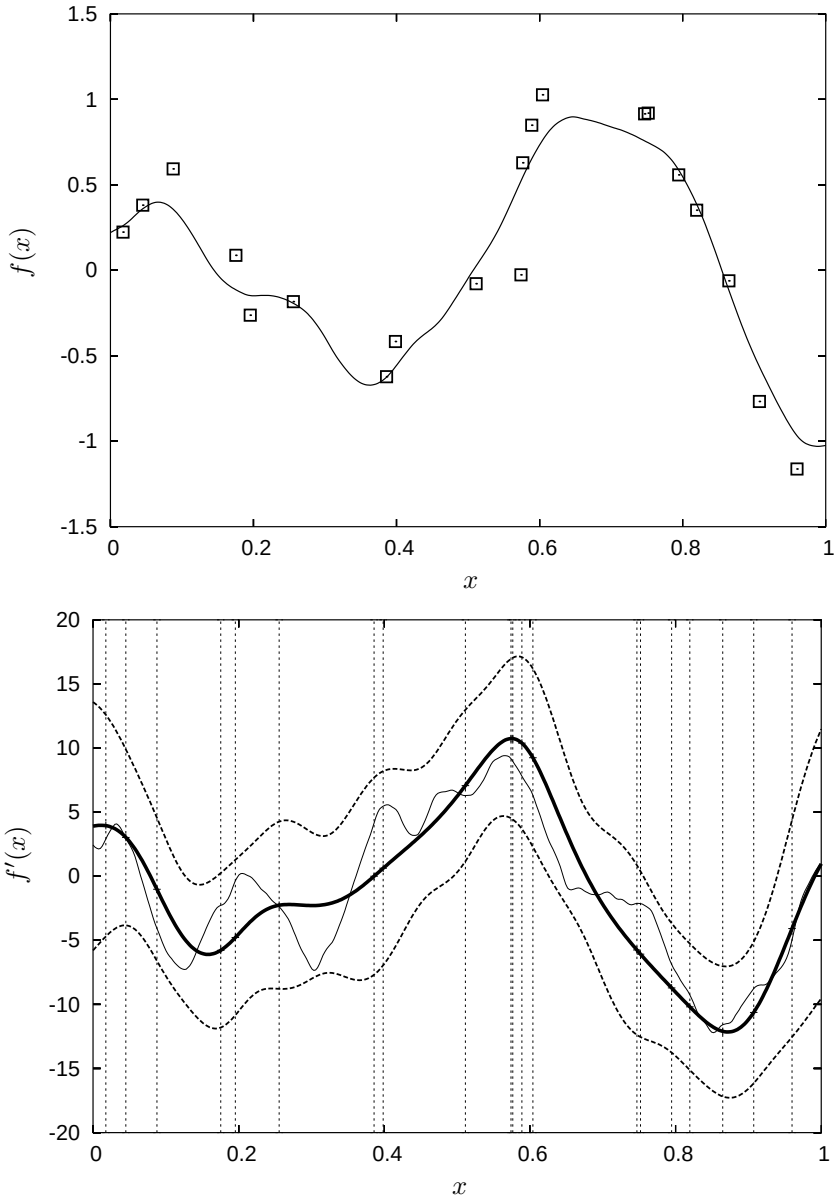


FIG. 6.5 – Figure du haut : 20 observations (matérialisées par des carrés) bruitées et irrégulièrement échantillonnées d'une fonction $f(x)$, $x \in [0,1]$. L'écart type du bruit d'observation vaut 0.2. Figure du bas : estimée (en trait gras continu) de la dérivée de $f(x)$ à partir des observations précédentes. La dérivée exacte est représentée en trait continu fin et les intervalles d'incertitude à 95% sont indiqués en traits interrompus. Les barres verticales rappellent la position des observations.

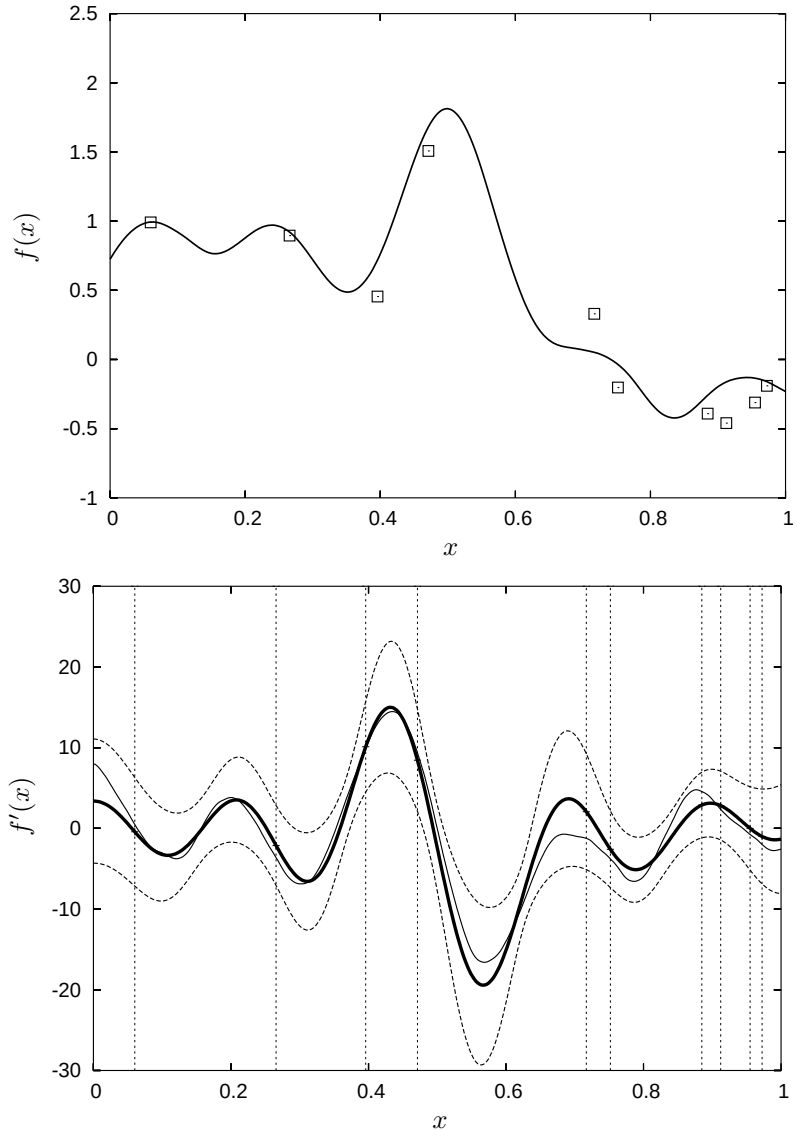


FIG. 6.6 – Figure du haut : 10 observations (matérialisées par des carrés) bruitées et irrégulièrement échantillonnées d'une fonction $f(x)$, $x \in [0,1]$. L'écart type du bruit d'observation est 0.2. $f(x)$ est la somme d'un sinus cardinal et d'une trajectoire d'un processus aléatoire centré. Figure du bas : estimée (en trait gras continu) de la dérivée de $f(x)$ à partir des observations précédentes. La dérivée exacte est représentée en trait continu fin et les intervalles d'incertitude à 95% sont indiqués en traits interrompus. Les barres verticales rappellent la position des observations. L'estimation utilise l'information disponible sur $f(x)$.

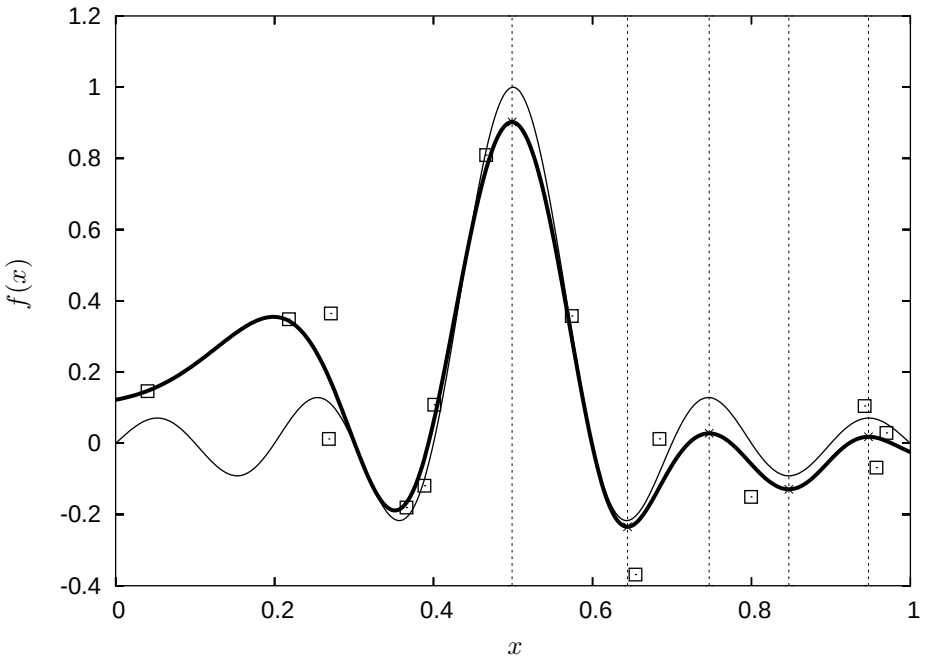


FIG. 6.7 – Approximation (en trait continu gras) d'un sinus cardinal (en trait continu fin) à partir d'observations bruitées (matérialisée par des carrés) et de la connaissance des valeurs des dérivées du sinus cardinal aux positions indiquées par les barres verticales.

f pour fixer une condition initiale. Cette méthode d'intégration peut être généralisée à la résolution numérique d'équations différentielles partielles linéaires. Elle est également utilisable pour modéliser des systèmes dynamiques non-linéaires sous la forme $\dot{\boldsymbol{x}} = f(\boldsymbol{x}, \boldsymbol{u})$, ce qui constitue une perspective de travail très intéressante.

6.2 Problème en compatibilité électromagnétique caractérisé par un nombre limité d'observations

Cette première étude de problème réel illustre les difficultés qui surviennent lorsque le nombre d'observations est limité (en ingénierie, effectuer une observation peut se révéler coûteux en moyens expérimentaux ou en temps de calcul). Nous ne disposons pas non plus de connaissance a priori sur le système étudié. En raison de la faible quantité d'information disponible, la validation des hypothèses (gaussianité du processus, régularité, etc.) et des modèles obtenus est délicate. Nous sommes donc confrontés à la difficulté d'établir un modèle boîte noire pertinent.

Cette étude a été réalisée en collaboration avec le Département de Recherche en Électromagnétisme. Nous avions à notre disposition des données et des programmes informatiques utilisés dans la thèse (Rannou, 2001), dans laquelle des modèles par krigeage avaient aussi été étudiés. Notons que notre collaboration avec le Département de Recherche en Électromagnétisme fait également suite à la thèse (Lefèbvre, 1997), qui avait mis en avant l'intérêt des modèles boîte noire par krigeage dans les problèmes complexes que l'on rencontre fréquemment dans le domaine de la compatibilité électromagnétique (voir aussi (Lefèbvre et al., 1996)). Nous poursuivons donc l'étude (Rannou, 2001) en nous concentrant davantage sur des aspects méthodologiques.

6.2.1 Description du problème

Nous considérons une ligne de transmission éclairée par une onde électromagnétique, comme représentée par le schéma de la figure 6.9. L'objectif est de caractériser l'influence d'appareils électromagnétiques tels que les téléphones portables et autres systèmes de communication sur les dispositifs électroniques sensibles (ordinateurs, appareils de mesure, etc.). Une ligne de transmission est caractérisée par des paramètres technologiques, comme par exemple

- son nombre de conducteurs,
- sa distance au plan de masse,
- son diamètre,
- sa longueur,
- l'impédance du générateur,
- l'impédance de la charge, etc.

La grandeur que l'on veut modéliser est l'intensité des courants induits par les perturbations électromagnétiques de l'environnement dans une ligne de transmission. Cette intensité dépend des paramètres technologiques de la ligne ainsi que de nombreux autres facteurs, comme par exemple

- la géométrie de l'environnement,
- la nature de la source émettrice (dipôle, forme d'onde, fréquence),
- la puissance de la source,
- la distance et l'angle de la source, etc.

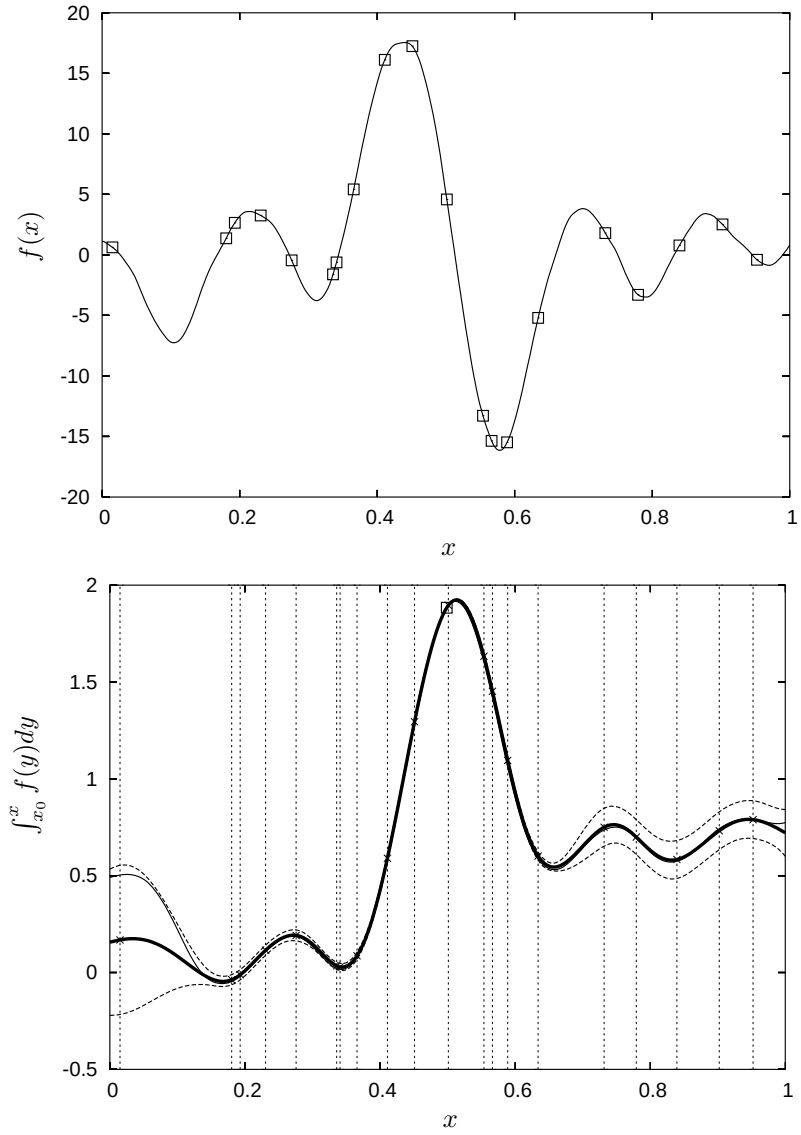


FIG. 6.8 – Exemple d'intégration d'une fonction. Figure du haut : la fonction $f(x)$ et 20 observations non uniformément échantillonnées (matérialisées par des carrés). Figure du bas : en gras, approximation d'une primitive de $f(x)$ à partir des observations précédentes et sachant la valeur de cette primitive en $x = 0.5$ (point indiqué par le carré). La primitive exacte est représentée en trait continu fin et les intervalles de confiances sont indiqués en trait discontinu. Les barres verticales rappellent la position des observations.

Afin de prendre en compte la diversité des perturbations possibles, le module du courant induit dans une ligne de paramètres donnés est modélisé par une variable aléatoire $I = f(\boldsymbol{x}, \boldsymbol{W})$, où $\boldsymbol{x} \in \mathbb{R}^d$ représente le vecteur des paramètres technologiques de la ligne et $\boldsymbol{W}$ est un vecteur aléatoire modélisant les perturbations de l'environnement. La loi de cette variable aléatoire est estimée par une méthode de Monte-Carlo, qui consiste à répéter des observations effectuées en laboratoire ou en simulation pour différentes réalisations du vecteur $\boldsymbol{W}$. La figure 6.10 représente un histogramme des modules du courant induit dans la ligne lorsque la nature de la perturbation varie. En observant plusieurs histogrammes de ce type, on constate que la loi du module du courant est (bien) une fonction des paramètres technologiques de la ligne. Comme un tel histogramme des courants induits s'avère coûteux à obtenir, nous souhaitons être capable de prédire les caractéristiques de la loi (comme la moyenne, la variance, etc.) du (module du) courant à l'aide d'un modèle boîte noire admettant comme entrées les paramètres technologiques de la ligne.

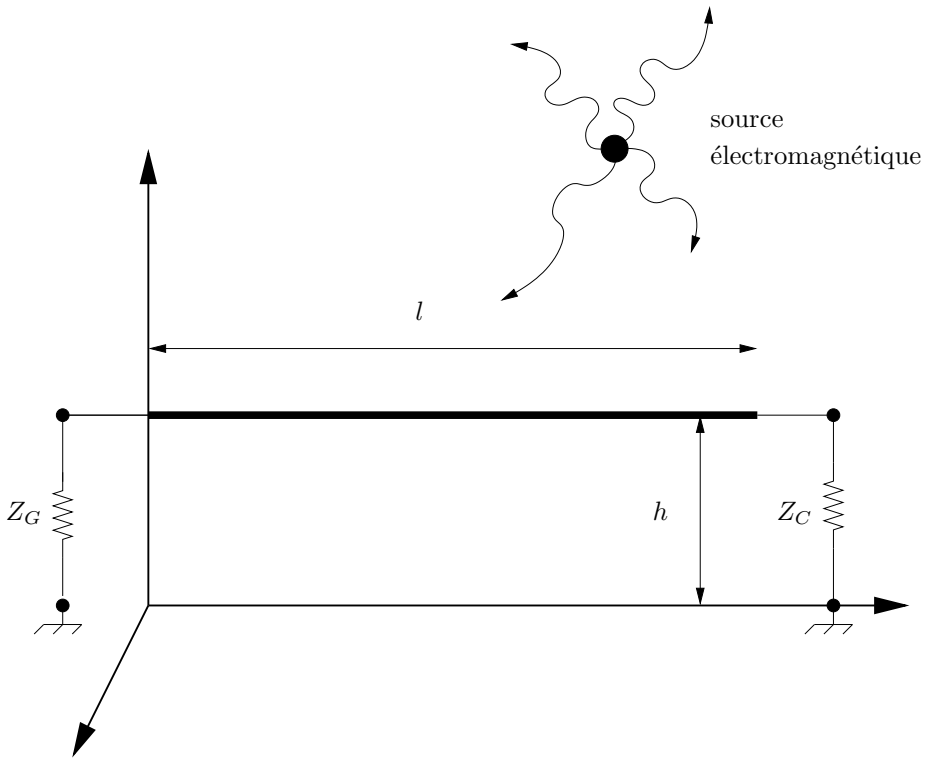


FIG. 6.9 – Représentation schématique d'une ligne de transmission et des perturbations électromagnétiques émises par une source. Z_G et Z_C sont les impédances du générateur et de la charge, l est la longueur de la ligne, h sa distance au plan de masse.

6.2.2 Système considéré

Un histogramme du courant est obtenu soit à partir de mesures effectuées sur un dispositif en laboratoire, soit à l'aide d'un programme informatique simulant le courant induit dans une ligne

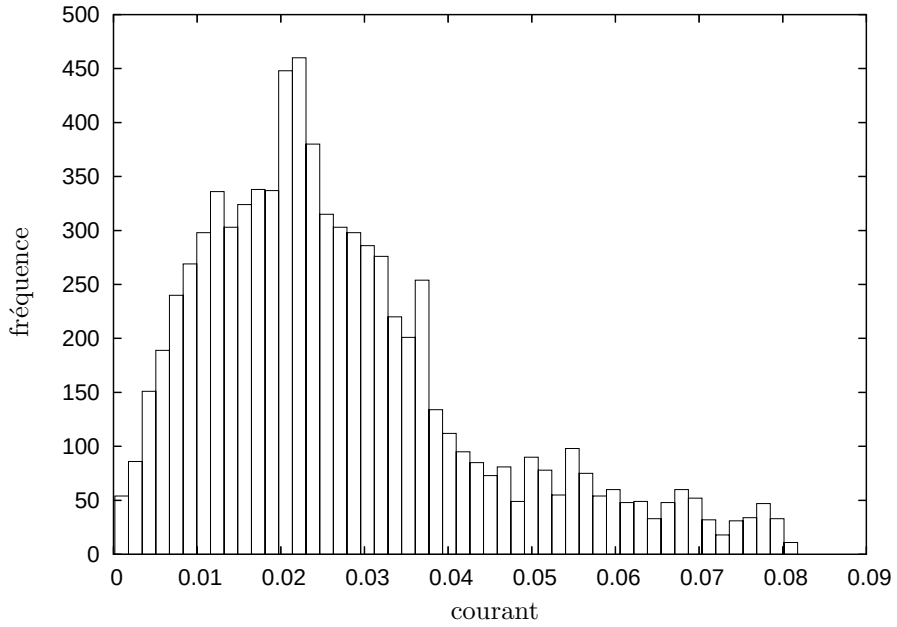


FIG. 6.10 – Histogramme du module du courant dans la ligne pour une configuration donnée.

de paramètres technologiques donnés soumise à une perturbation électromagnétique donnée. Dans les deux cas, obtenir l'histogramme du courant a un coût important.

Dans cette présentation, nous nous restreignons à l'étude de la moyenne de la loi du courant en fonction des trois paramètres de la ligne listés ci-dessous :

- le diamètre de la ligne d, $1 \cdot 10^{-4}$ m $\leq d \leq 1.98 \cdot 10^{-3}$ m,
- sa distance au plan de masse h, $1 \cdot 10^{-3}$ m $\leq h \leq 1.37 \cdot 10^{-1}$ m,
- sa longueur l, 0.75 m, $\leq l \leq 1.24$ m.

(Les autres paramètres sont fixés à leur valeur nominale.) Du point de vue de la modélisation boîte noire, nous considérons donc un système ayant comme entrées les paramètres d, h et l de la ligne, et comme sortie la moyenne i_m du courant induit. Effectuer une observation consiste donc à déterminer i_m pour un jeu fixé de paramètres de la ligne.

Par la suite, nous simplifions davantage le problème en ne considérant l'évolution du courant qu'en fonction d'un ou deux facteurs à la fois, les facteurs restants étant fixés. Le seul objectif de cette simplification est de pouvoir présenter plus facilement des figures.

6.2.3 Résultats et conclusions

Choix du modèle

Les figures 6.11, 6.13 et 6.15 présentent l'évolution de la moyenne du courant (obtenue à partir de simulations informatiques des phénomènes électromagnétiques dans la ligne) en fonction de d, h et l. Comme mentionné ci-dessus, nous ne considérons l'influence que d'un ou deux facteurs à la fois. Plus précisément, nous nous intéressons aux systèmes

1. $i_m = f_1(d)$,
2. $i_m = f_2(l)$,
3. $i_m = f_3(l, h)$.

Deux possibilités sont envisageables pour choisir la structure du second ordre des processus aléatoires modélisant ces trois systèmes. Dans la première, la totalité des données simulées est utilisée pour effectuer le choix. Une telle approche est justifiée si l'objectif est d'ajuster les modèles en simulation pour les utiliser ensuite dans des cas réels plus complexes et plus coûteux à mettre en œuvre. Dans la seconde, nous limitons le nombre des observations et choisissons le modèle sans utiliser de connaissances supplémentaires. Cette approche peut être suivie afin de simuler des conditions réelles où nous voudrions établir un modèle boîte noire sans connaissances a priori, où chaque observation est coûteuse, et pour tester la pertinence de nos procédures de choix de modèle. La première approche est bien sûr plus confortable. Par la suite, nous allons suivre la seconde. Un faible nombre d'observations signifie que nous avons peu d'information pour établir un modèle boîte noire du système. Il convient donc de se poser les questions suivantes.

– Est-il possible d'estimer un modèle avec très peu de données ?
– Reste-t-il possible de prédire la grandeur d'intérêt ?

Avec très peu d'observations, le choix d'une structure du second ordre est certainement arbitraire. Précisément, les données observées sont susceptibles d'avoir été générées par de nombreux processus aléatoires et il n'est pas possible de juger la pertinence d'un modèle par rapport à un autre. Toutefois, cela ne signifie pas qu'il n'y a pas d'information et que rien ne peut être estimé. Même à partir de très peu de données, il est possible d'estimer une moyenne et une variance des observations, et peut-être même la vitesse de variation de la sortie en fonction des facteurs. Pour cela, nous avons besoin de procédures d'estimation simples à mettre en œuvre, consistantes et efficaces (au moins asymptotiquement). (Nous utiliserons l'estimateur du maximum de vraisemblance.) D'autre part, même si un modèle précis ne peut être établi, nous devons pourtant en choisir un, et ce faisant, introduire une forme d'information a priori.

Dans cette application, nous faisons ainsi deux hypothèses (que l'on pourrait appeler également des paris en l'absence de connaissances a priori). Nous supposons d'abord que la sortie peut être modélisée par un processus gaussien[1]. Cette hypothèse est contestable, parce que le module du courant est une quantité positive, par exemple. Toutefois, nous avons vu dans la section 2.2.1 que l'hypothèse gaussienne correspond à un critère de simplicité (au sens du maximum d'entropie). Nous supposons ensuite que les tendances à croître ou décroître du courant peuvent être modélisées par des polynômes de faible degré.

Une modélisation par *IRF* semble donc appropriée, ce qui suppose d'introduire une hypothèse de stationnarité des accroissements (généralisés). Plus précisément, nous choisissons une *IRF* d'ordre deux admettant une covariance polynomiale d'ordre deux, qui s'écrit donc sous la forme

$$k(h) = -b_0|h| + b_1|h|^3 - b_2|h|^5,$$

où $b_0, b_1, b_2 \geq 0$. Notons que le système à deux facteurs (numéroté 3, ci-dessus), présente une anisotropie très marquée comme on peut le constater sur la figure 6.15. Nous traitons cette anisotropie en choisissant une normalisation adaptée des facteurs. Ce changement d'échelle peut être effectué

[1] Si une autre hypothèse devait être choisie, il faudrait utiliser le krigeage non-linéaire (Chilès et Delfiner, 1999)

sans information a priori dans la mesure où il est possible[2] de découvrir ce type d'anisotropie à partir d'un petit nombre d'observations[3].

Forte incertitude sur le modèle

Nous pouvons évaluer la qualité du modèle lors de deux étapes essentielles. Lors de l'estimation du modèle, nous utilisons les profils de vraisemblance pour nous renseigner sur la variance des paramètres estimés. Si le maximum de vraisemblance est très marqué, alors nous pouvons avoir une relative confiance dans les estimées des paramètres (faible variance de l'estimateur). À l'inverse, un profil très plat autour du maximum nous indique que le modèle reste mal déterminé malgré les données observées. Dans l'étape de prédiction, les intervalles de confiance nous renseignent sur la qualité des prédictions. Bien sûr, ces intervalles de confiance dépendent de l'estimation du modèle. Le maximum de vraisemblance obéissant à un principe de maximum d'entropie, nous espérons que les variances (estimées) d'erreur de prédiction ne seront pas trop optimistes.

Pour les trois systèmes considérés, l'examen des profils de vraisemblance présentés sur les figures 6.12, 6.14 et 6.18 met en évidence ce à quoi nous nous attendions : les modèle estimés restent très incertains avec un nombre faible de données observées. Plus précisément, nous voyons que la régularité de la covariance est sans doute arbitraire parce que la vraisemblance reste à peu près constante le long de lignes joignant des points des axes $b_0 = 0$ et $b_1 = 0$ de l'espace des paramètres (b_0, b_1).

Notons que nous avons choisi dans les trois cas un échantillonnage non uniforme afin d'essayer de maximiser l'information que nous pouvons retirer sur les variations de la sortie à différentes échelles de distance (cette approche est justifiée de manière empirique par l'expérience numérique présentée dans la section 6.6.2). Nous commentons brièvement ci-dessous les résultats obtenus pour les trois systèmes étudiés.

Système 1. La prédiction est satisfaisante. Les intervalles de confiances sont pessimistes (figure 6.11).

Système 2. La prédiction n'est pas de bonne qualité. Les intervalles de confiances sont optimistes. Le module du courant en fonction de la longueur de la ligne est caractérisé par des pics périodiques d'intensité. Nous pouvons dire que ce système n'est pas correctement modélisé par un processus gaussien (figure 6.13).

Système 3. La prédiction suit les tendances du courant mais reste de qualité médiocre (figure 6.16). Les intervalles de confiance sont assez satisfaisants (figure 6.17).

En conclusion, établir un modèle boîte noire d'un système nécessite un nombre relativement grand d'observations, ce qui n'est pas toujours possible dans les problèmes réels. Dans les systèmes considérés ci-dessus, le nombre d'observations est insuffisant pour effectuer des prédictions

[2]Du moins en dimension deux, car pour un nombre de facteurs plus élevé, la détermination d'anisotropies est délicate

[3]Le problème de la planification des observations pour découvrir les caractéristiques d'un processus aléatoire n'est pas abordée dans notre travail. Dans la section 6.6.5, nous avons réfléchi à la façon dont nous devons choisir les observations pour minimiser l'erreur de prédiction, ce qui est normalement un problème différent du précédent (et également plus facile)

de bonne qualité. Cependant de tels modèles ne sont pas inutiles en pratique. Par exemple, les intervalles d'erreur de prédiction peuvent servir à effectuer des dimensionnements. Ceci met en avant la nécessité de définir correctement l'objectif de la prédiction par modèles boîte noire.

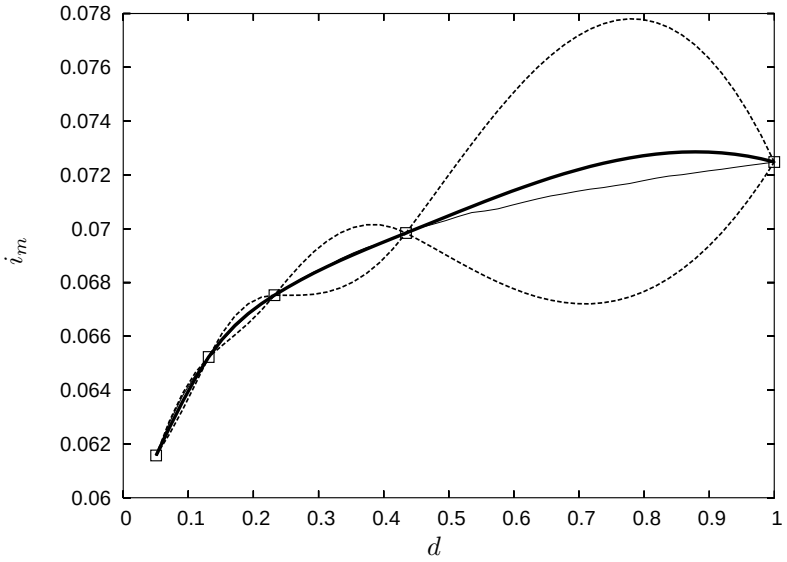


FIG. 6.11 – Évolution du module du courant en fonction du diamètre de la ligne en trait continu fin. Les observations sont indiquées par les carrés. La prédiction obtenue par krigeage est représentée en trait continu gras. Les intervalles de confiances à 95% sont tracés en trait interrompu.

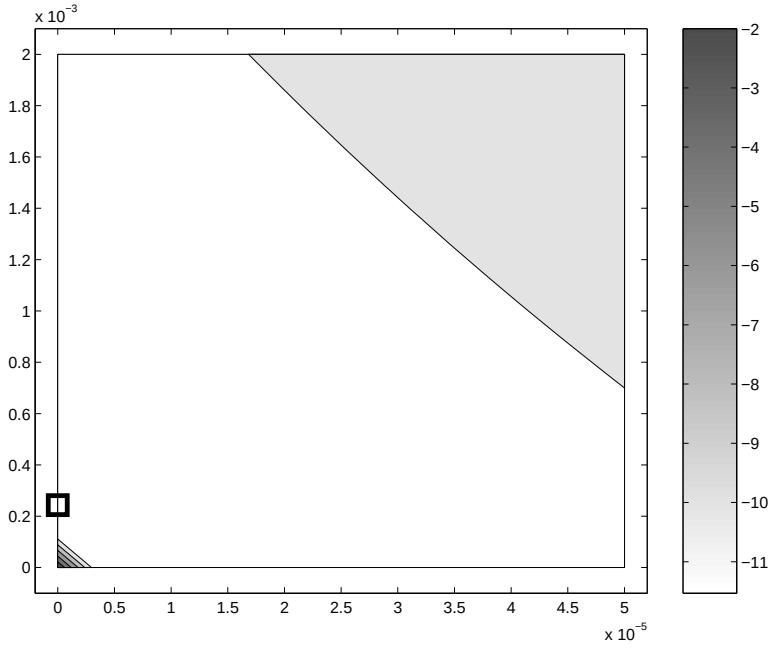


FIG. 6.12 – Profil de vraisemblance des données observées à la figure 6.11 en fonction des paramètres b_0 en abscisses et b_1 en ordonnées de la covariance polynomiale. Le carré indique la valeur des paramètres estimés.

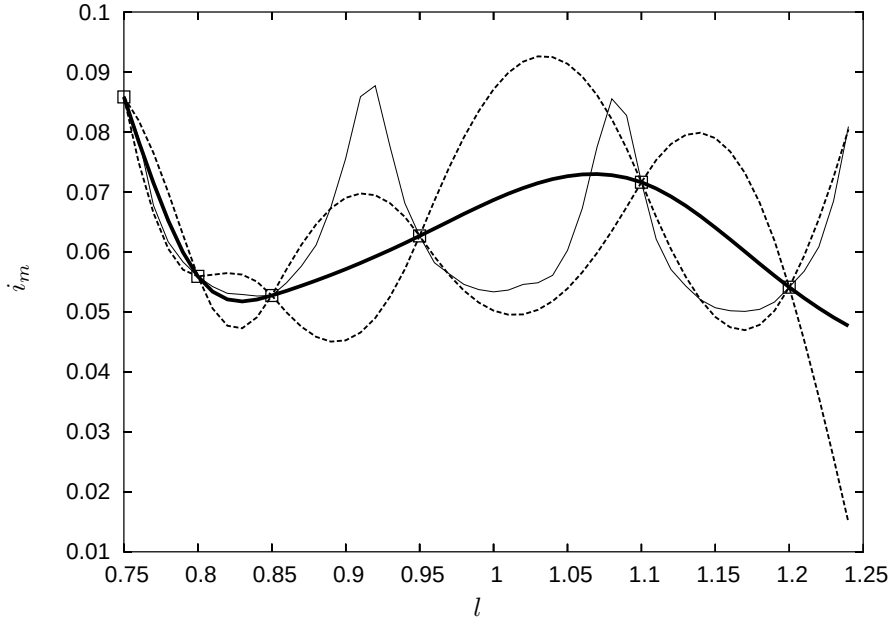


FIG. 6.13 – Évolution du module du courant en fonction de la longueur de la ligne (mêmes conventions graphiques que précédemment).

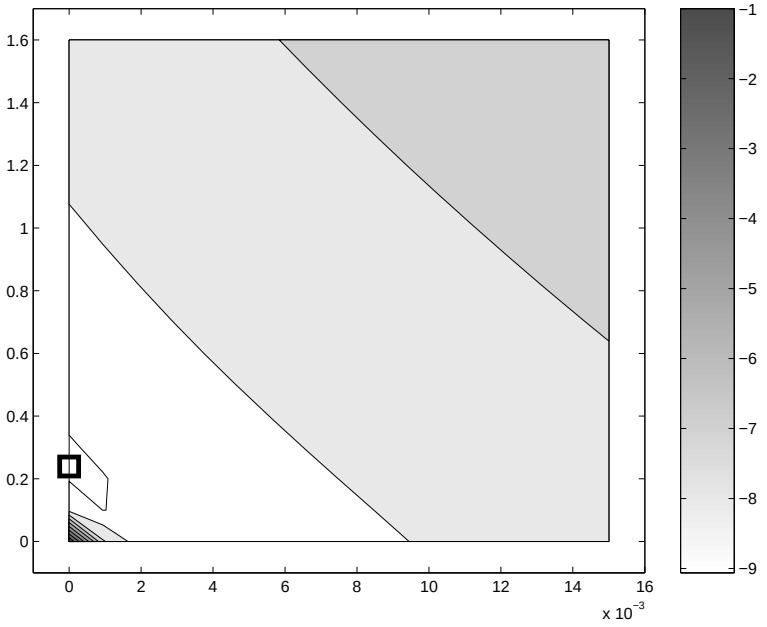


FIG. 6.14 – Profil de vraisemblance des données observées à la figure 6.13 en fonction des paramètres b_0 en abscisses et b_1 en ordonnées de la covariance polynomiale. Le carré indique la valeur des paramètres estimés.

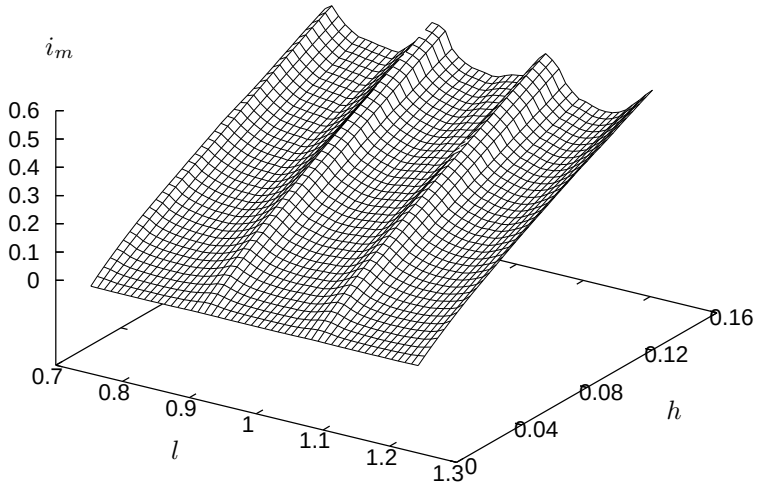


FIG. 6.15 – Représentation du module du courant en fonction de la longueur l et de la distance au plan de masse h de la ligne (pour $d \approx 1$ mm)

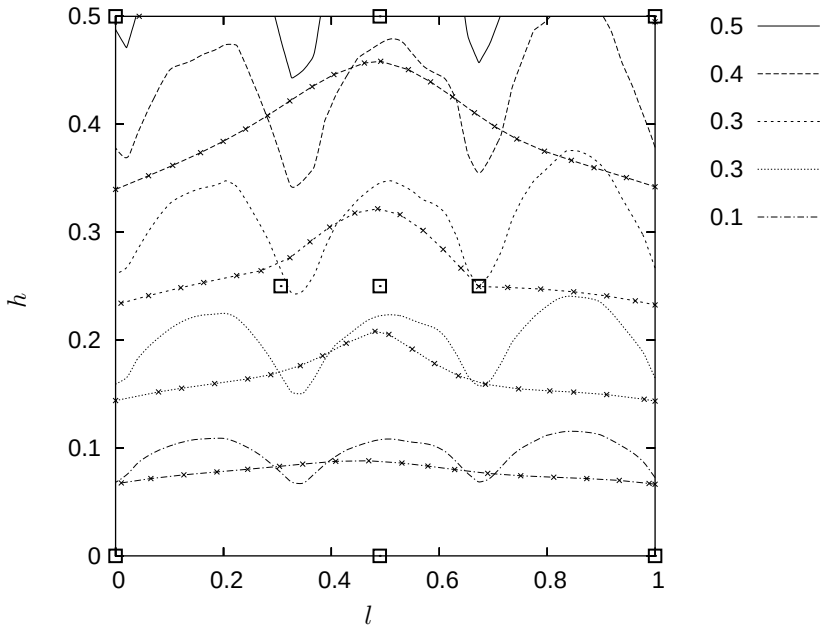


FIG. 6.16 – Lignes de niveau du module du courant en fonction de l et h (facteurs renormalisés empiriquement). Les observations sont indiquées par des carrés. Les lignes de niveau marquées par des croix correspondent à la prédiction obtenue qui est comparée au courant de référence représenté par les lignes de niveau non marquées.

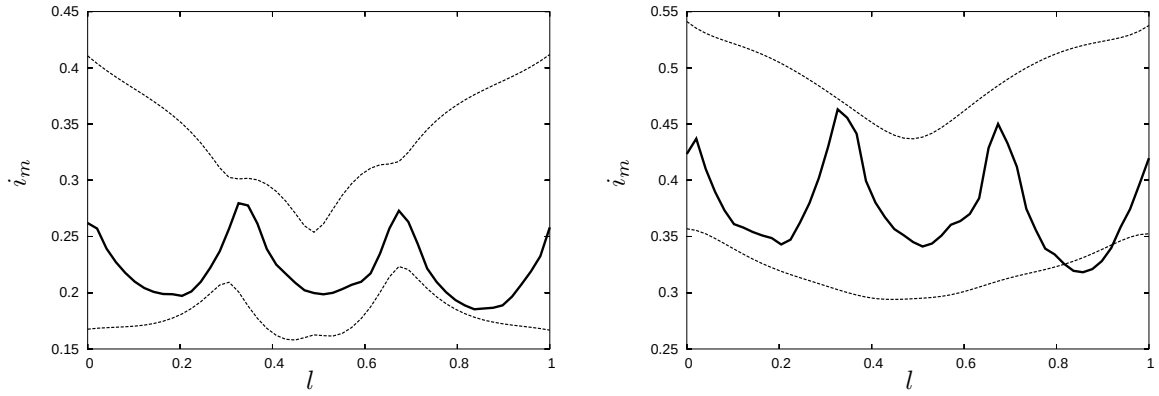


FIG. 6.17 – Intervalles de confiance (traits interrompu) et courant de référence (trait continu) en fonction de la longueur pour différentes sections à h constant de la figure 6.16. La figure de gauche correspond à $h = 0.22$. La figure de droite correspond à $h = 0.40$

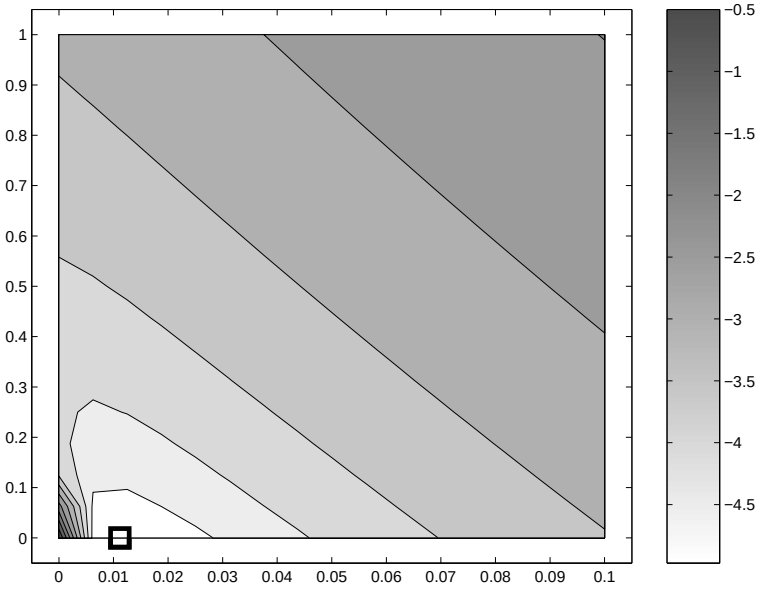


FIG. 6.18 – Profil de vraisemblance des données observées à la figure 6.16 en fonction des paramètres b_0 en abscisses et b_1 en ordonnées de la covariance polynomiale. Le carré indique la valeur des paramètres estimés.

6.3 Débitmétrie avec prise en compte d'a priori

Dans cet exemple, nous nous intéressons à un débitmètre, où se posent de nouveau les difficultés liées au manque d'observations. Il y a dans ce problème deux contraintes. Tout d'abord, chaque observation nécessite la mise en place d'un capteur, ce qui augmente le coût de l'appareil. D'autre part, il existe des contraintes physiques qui font que certaines zones de l'espace des facteurs ne peuvent pas être observées. Pour palier le manque d'observations, nous montrons comment prendre en compte dans un modèle boîte noire des connaissances sur la physique gouvernant le système pour en faire un modèle boîte grise.

6.3.1 Description du problème

Nous nous intéressons à l'estimation du *débit* d'un fluide dans une conduite. Il existe de nombreux principes pour mesurer un débit. Par exemple, il est possible d'effectuer des *observations ponctuelles de la vitesse du fluide* en des positions données d'une section de la conduite (voir la figure 6.19). Nous proposons une méthode boîte noire pour estimer le débit à partir de ces vitesses ponctuelles. Une première approche envisageable est de *reconstruire entièrement le profil* de vitesse du fluide sur une section à partir des observations, et d'intégrer ensuite la vitesse pour obtenir une estimée du débit. Nous verrons plus loin qu'une telle méthode n'est pas optimale mais permet d'aborder plusieurs aspects méthodologiques intéressants. La reconstruction d'un profil de vitesse est une tâche difficile en raison de la grande diversité des configurations envisageables : variations importantes possibles des débits, conduites divergentes ou convergentes, présence de coudes, etc. Les phénomènes physiques susceptibles d'influencer le profil de vitesse sont donc complexes et difficiles à modéliser. Il faudrait notamment tenir compte des turbulences, des frottements du fluide contre la paroi de la conduite, des modes secondaires apparaissant lorsque le fluide passe dans des coudes, etc.

Un code informatique utilisant la méthode des différences finies a permis de simuler plus de trois mille profils de vitesse obtenus en faisant varier le débit et la géométrie de la conduite. Ces simulations ont été effectuées par G. Fleury au Service des Mesures de Supélec. La figure 6.20 montre des exemples de profils de vitesse simulés. Notons qu'un profil de vitesse possède généralement une symétrie de révolution sauf dans les sections à l'abord des parties coudées de la conduite. Dans le dispositif de mesure réel, il n'est pas possible d'observer la vitesse du fluide proche de la paroi de la conduite. Nous avons donc choisi les positions des observations pour obtenir une répartition homogène sur la partie observable de la section. La figure 6.19 montre la distribution retenue[4] constituée de 16 points d'observation.

6.3.2 Modélisation boîte noire sans a priori

Lorsque le profil de vitesse a été obtenu, le débit est estimé par intégration de $\hat{f}(\boldsymbol{x})$, où $\boldsymbol{x}$ est la position sur la section $\mathbb{X} = \{\boldsymbol{x} \in \mathbb{R}^2, ||\boldsymbol{x}||_2 \leq r\}$. Cette intégration pourrait être faite en principe de manière analytique mais nous avons préféré ici une méthode de Monte-Carlo pour simplifier. L'étape principale est donc de parvenir à une bonne reconstruction du profil à partir

[4]Sans chercher à optimiser la position des points d'observation.

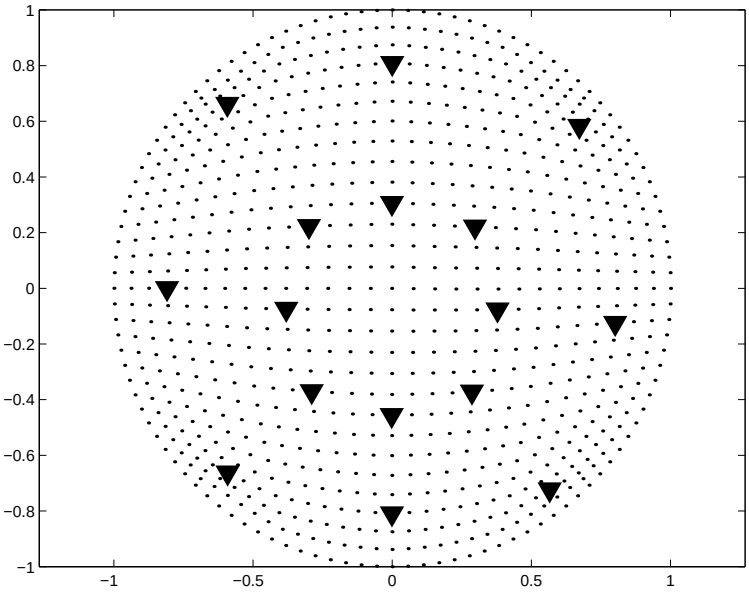


FIG. 6.19 – Section d'une conduite de fluide traitée dans le problème de débitmétrie. Les points indiquent les positions où la vitesse est calculée pendant les simulations. Les triangles matérialisent la position des observations dans le dispositif de mesure. Il n'y a que 16 observations.

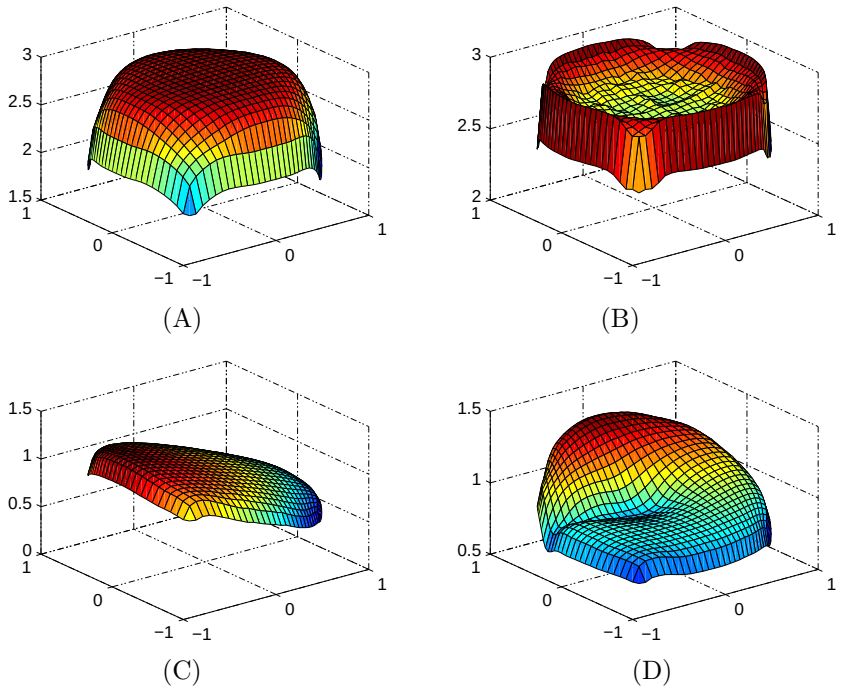


FIG. 6.20 – Profils de vitesse correspondants à différentes géométries. Par exemple, (A) montre le profil dans une conduite droite et (D) le profil dans une conduite coudée.

des observations ponctuelles. Il s'agit d'un problème d'interpolation classique. La vitesse en $\boldsymbol{x}$ est estimée par krigeage par

$$\hat{f}(\boldsymbol{x}) = \sum_{i=1}^{16} \lambda_i(\boldsymbol{x}) f^{\text{obs}}_{\boldsymbol{x}_i} , \tag{6.1}$$

où $\{f^{\text{obs}}_{\boldsymbol{x}_i}\}_{i=1}^{16}$ sont les vitesses observées.

Nous avons choisi une méthode de krigeage intrinsèque pour calculer les coefficients λ_i. Deux arguments permettent de justifier ce choix. Nous souhaitons modéliser les grandes tendances du profil de vitesse par un polynôme de degré deux (de deux variables). Nous pensons également au vu des variogrammes que l'échelle caractéristique des variations du profil de vitesse est de l'ordre du domaine d'étude. Les λ_i sont solution du système linéaire (4.16) où nous avons choisi une covariance généralisée polynomiale d'ordre deux, qui comprend des termes jusqu'à $|h|^5$. L'estimation des coefficients de la covariance est importante afin d'obtenir la régularité appropriée du processus aléatoire modélisant la vitesse. Nous avons utilisé pour ce faire la méthode REML. Cependant, si nous ne nous intéressons pas à la variance de l'erreur de prédiction, nous pouvons nous contenter d'une estimation approximative des paramètres, en nous rappelant en particulier que les coefficients du krigeage ne dépendent pas d'une multiplication de la covariance par une constante. Ainsi, il n'est pas nécessaire d'estimer les coefficients de la covariance pour chaque type de profil et chaque débit.

Les figures 6.21 montrent le résultat de l'interpolation obtenue. Nous constatons immédiatement que la vitesse près de la paroi de la conduite est surestimée, ce qui conduit à une prédiction peu satisfaisante du débit du fluide (typiquement, nous obtenons plus de 2% d'erreur relative). Cette surestimation est liée au fait que le modèle ne prend pas correctement en compte la chute de vitesse près de la paroi en raison de l'absence d'observations de la vitesse dans cette zone (due aux contraintes de l'appareil de mesure). En résumé, *un modèle totalement boîte noire de ce système ne donne pas des résultats satisfaisants.* Nous devons donc chercher à inclure de l'information a priori.

6.3.3 Incorporation de connaissances a priori

Comme souvent en ingénierie, les seules connaissances a priori disponibles sur le système sont sous forme de descriptions qualitatives du comportement de celui-ci. Ici, la connaissance que nous aimerions prendre en compte est la présence de frottements qui entraînent une chute de vitesse près de la paroi de la conduite.

Pour utiliser ce type de connaissance approximative, nous proposons d'utiliser une formulation semi-paramétrique du krigeage. Notons $f^*(\boldsymbol{x}) : \boldsymbol{x} \in \mathbb{X} \to \mathbb{R}$ un profil de vitesse de référence, censé représenter notre connaissance a priori. Il s'agit typiquement du profil standard de la vitesse sur la section d'une conduite. Nous supposons cette fonction connue en tout point de $\mathbb{X}$. Dans cette nouvelle approche, la vitesse estimée en un point quelconque de la section s'exprime encore sous la forme d'une combinaison linéaire des vitesses observées, comme dans (6.1). Cependant, le modèle du système étudié comporte maintenant un facteur supplémentaire, parce que la vitesse à prédire est supposée dépendre à la fois de la position $\boldsymbol{x} \in \mathbb{X}$ sur la section et du scalaire $f^*(\boldsymbol{x}) \in \mathbb{R}$. Nous utilisons alors le krigeage intrinsèque, avec les choix suivants :

– la covariance (généralisée) ne dépend pas du facteur $f^*(\boldsymbol{x})$, ce qui signifie que la corrélation suivant ce facteur est constante ;
– l'espace $\mathcal{N}$ contient la fonction

$$\begin{array}{rccc} f_1 : & \mathbb{X} \times \mathbb{R} & \to & \mathbb{R} \\ & (\boldsymbol{x}, f^*(\boldsymbol{x})) & \mapsto & f^*(\boldsymbol{x}) \,, \end{array} \tag{6.2}$$

(il s'agit donc d'une fonction linéaire du facteur $f^*(\boldsymbol{x})$).

Comme nous l'avons vu dans la section 4.6.1, ceci correspond à choisir un modèle paramétrique comportant le terme $b_* f^*(\boldsymbol{x})$. En fait, il est possible de considérer plusieurs profils de référence et d'introduire autant de termes paramétriques. Il y a cependant une limite et l'accroissement des degrés de liberté dans le modèle peut dégrader la qualité de la prédiction. Trop augmenter le nombre de profils de référence conduira à un problème mal posé. Rappelons en effet que les termes paramétriques n'étant pas régularisés, on ne contrôle pas leur contribution et il est possible avec un nombre relativement faible de termes ajoutés d'obtenir une phénomène de *sur-adéquation* aux données (*overfitting*). Par conséquent, il est nécessaire de sélectionner un nombre limité de profils de référence. Avec seulement deux profils de référence (nous avons retenu deux profils simulés dans une section de conduite droite correspondant à des débits faible et fort), l'utilisation de l'approche semi-paramétrique améliore considérablement la qualité de la prédiction du débit, comme le montre les histogrammes des erreurs de prédiction présentés à la figure 6.22.

Notons que les deux profils de vitesse utilisés dans l'approche semi-paramétrique ci-dessus correspondent à des situations simples pour lesquelles il serait possible de modéliser analytiquement la physique des écoulements. En conclusion, nous insistons sur le fait que l'approche semi-paramétrique est intéressante dès que l'on dispose d'un modèle *simple* et *approximatif* du système étudié. Lorsqu'un modèle boîte noire est ainsi rendu plus précis en utilisant la physique des phénomènes étudiés, on peut parler de *modèle boîte grise*.

méthode	moyenne	écart-type
modèle boîte noire	$1.16 \cdot 10^{-2}$	$1.13 \cdot 10^{-2}$
modèle boîte grise	$3.93 \cdot 10^{-4}$	$8.87 \cdot 10^{-3}$

TAB. 6.1 – Moyenne et écart type des erreurs relatives

6.4 Modèles boîte noire de systèmes à espace de facteurs de dimension élevée

6.4.1 Autre point de vue sur le problème de débitmétrie

Les méthodes de prédiction du débit dans une conduite de fluide présentées dans les sections 6.3.2 et 6.3.3 ne sont pas en fait à recommander. En effet, l'étape qui consiste à reconstruire le profil de vitesse n'est pas nécessaire. Il est préférable de prédire le débit directement en fonctions des données observées. Pour s'en convaincre, nous pouvons calculer la *régression linéaire* du débit en fonction des 16 vitesses observées. Nous trouvons que l'erreur résiduelle du modèle par régression linéaire obtenu à partir de la totalité des données est plus petite que l'erreur de prédiction obtenue

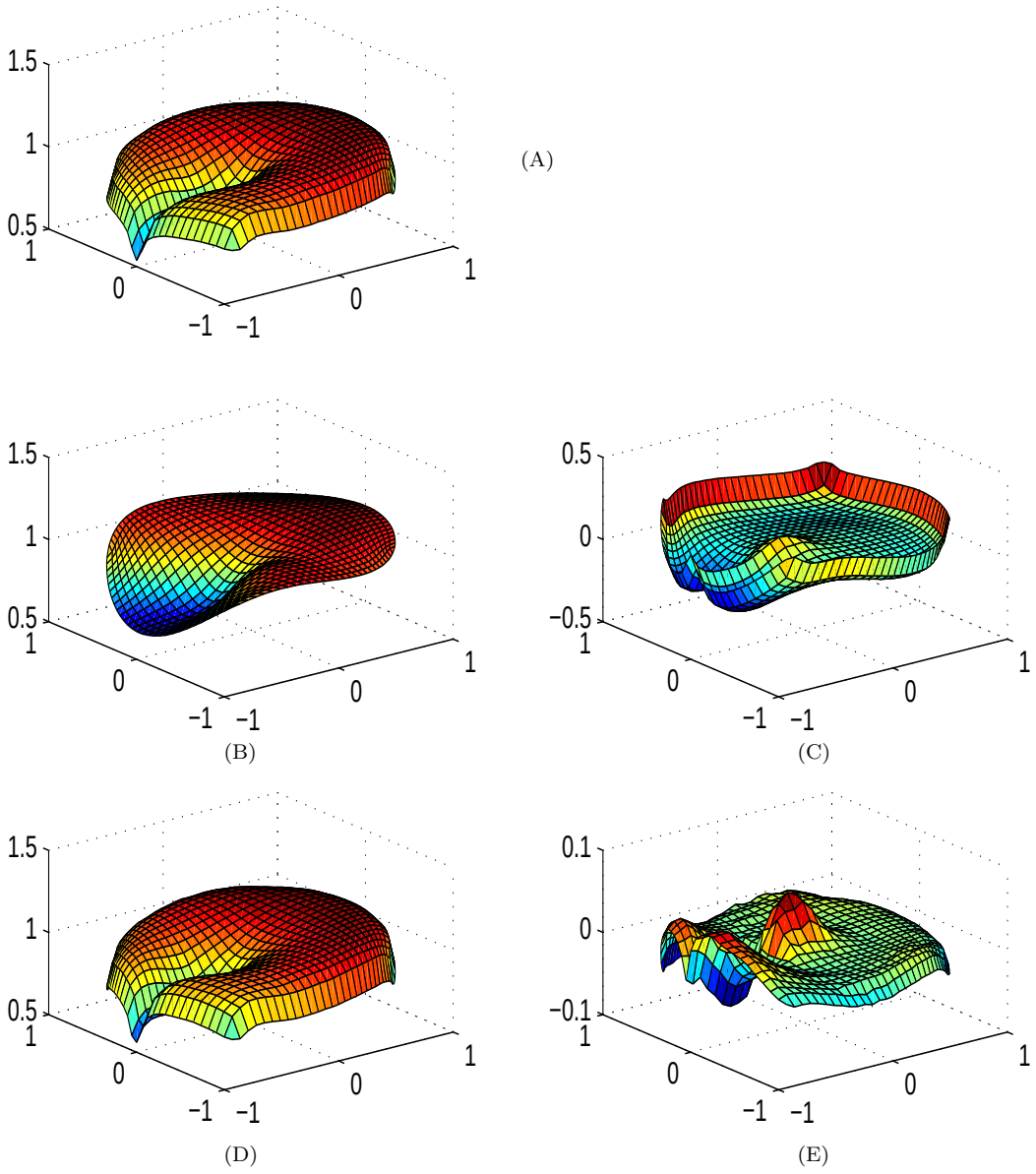


FIG. 6.21 – (A) vrai profil de vitesse, (B) reconstruction par krigeage, (C) erreur de prédiction correspondante, (D) interpolation avec incorporation d'a priori par krigeage intrinsèque, (E) erreur de prédiction ainsi améliorée.

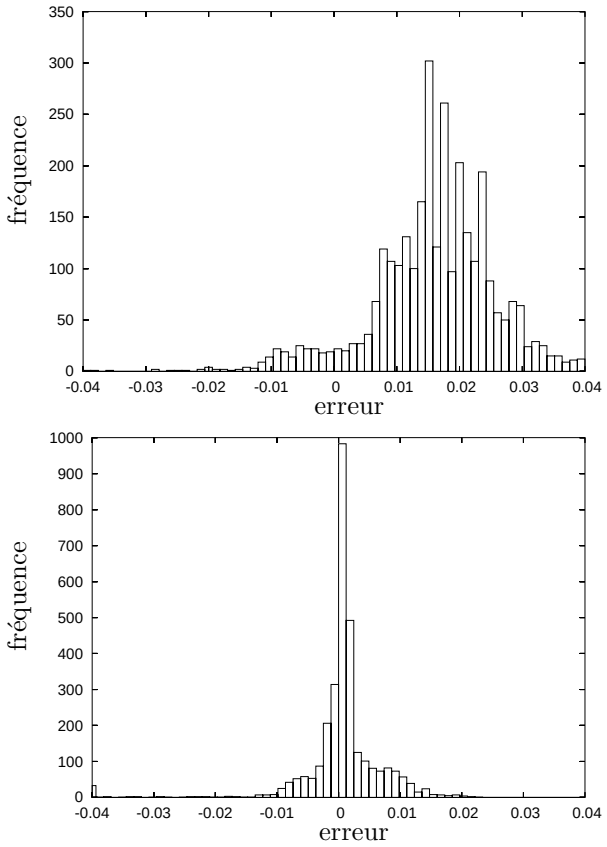


FIG. 6.22 – Histogramme des erreurs relatives. Figure du haut, modèle boîte noire simple. Figure du bas, prise en compte d'information a priori (modèle boîte grise).

précédemment (comparer les tables 6.1 et 6.2). Ceci suggère de considérer le système débit en fonction des vitesses observées. Dans cette approche, l'espace des facteurs est de dimension 16 et nous devons prédire le débit $f(\boldsymbol{x})$, $\boldsymbol{x} \in \mathbb{R}^{16}$, à partir de débits observés $f^{\text{obs}}_{\boldsymbol{x}_i}$, $i = 1, \ldots, n$, correspondants à des vecteurs de vitesses $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$. Afin de valider la méthode, il est nécessaire de partitionner la base des trois mille profils simulés en un ensemble de données d'apprentissage et un ensemble de données de test. Comme les débits observés proviennent d'une simulation, nous considérons le bruit d'observation négligeable et choisissons de prédire le débit en interpolant les observations. L'erreur de prédiction est par conséquent nulle sur la base d'apprentissage. Nous avons retenu une base d'apprentissage représentant environ 13% de la totalité de la base ($n = 400$).

L'interpolation est effectuée par krigeage intrinsèque en utilisant une covariance généralisée polynomiale d'ordre 1. Ceci permet d'inclure dans le modèle de prédiction un terme linéaire en les facteurs, ce qui est pertinent puisque l'on a vu plus haut que la régression linéaire du débit en fonction des vitesses observées était satisfaisante.

Les résultats sont présentés à la figure 6.23 sous forme d'histogramme et dans la table 6.2, où nous indiquons les moyennes et écarts types des erreurs relatives de prédiction sur les données de test. Nous voyons une amélioration par rapport à la régression linéaire. En conclusion, il semble préférable d'effectuer la prédiction de la grandeur d'intérêt directement en fonction des facteurs observés. Toutefois cette dernière méthode est sensible au choix de la base d'apprentissage. D'autre part, la prédiction du débit par reconstruction du profil présente l'avantage d'être plus proche de la physique du système.

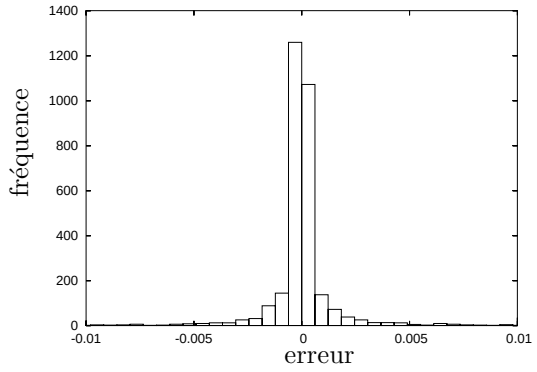


FIG. 6.23 – Histogramme des erreurs relatives.

méthode	moyenne	écart-type
régression linéaire	$5.06 \cdot 10^{-3}$	$5.94 \cdot 10^{-3}$
modèle boîte noire (16 facteurs, $n = 400$)	$2.39 \cdot 10^{-4}$	$4.85 \cdot 10^{-3}$

TAB. 6.2 – moyenne et écart type des erreurs relatives

6.4.2 Prédiction en climatologie

Nous avons comparé les performances d'un modèle boîte noire à noyau avec celles d'un *perceptron multicouche* ou *multilayer perceptron* (*MLP*) en anglais, qui est un cas particulier classique de réseau de neurones formels (Bishop, 1995). Cette étude nous a été proposée par le Laboratoire de Météorologie Dynamique (LMD) de l'École Polytechnique. Nous l'utilisons pour insister sur les problèmes spécifiques aux systèmes ayant des espaces de facteurs de dimension élevée.

Description du problème

L'objectif est la prédiction de la concentration d'un gaz[5] dans l'atmosphère terrestre à partir de données satellite. Ce problème est un thème de recherche du *LMD* qui nous a gracieusement offert des données. Le modèle du processus générant les données permet de calculer des *températures de brillance* de l'atmosphère à différentes fréquences du domaine spectral infrarouge et microonde en fonction de paramètres caractérisant une *situation atmosphérique* (température, pression, concentration des composants de l'atmosphère [H_2O, O_3, N_2O, CO, CO_2, etc.]). Ces températures de brillance correspondent aux données délivrées par des capteurs embarqués à bord d'un satellite, qui sont sensibles dans une bande limitée du domaine spectral infrarouge et microondes (ces bandes de fréquence sont également appelées *canaux*). Dans le modèle, la concentration du gaz étudié varie sur une plage d'environ 350 ppmv (parties par million et par volume) à 400 ppmv.

Les canaux sont plus ou moins sensibles à la présence du gaz étudié, mais la concentration en gaz ne peut pas être déterminée en utilisant un seul canal puisque d'autres facteurs interviennent dans les variations des températures de brillance. Les données fournies comportent environ mille concentrations du gaz étudié en fonction de trente-cinq températures de brillance dans des canaux susceptibles de comporter des informations permettant de prédire la concentration du gaz. La figure 6.24 présente des exemples de vecteurs de températures de brillance correspondant à deux concentrations différentes (les autres paramètres de l'atmosphère varient). Nous notons que les températures sont fortement corrélées.

Prédiction

Les données sont divisées en deux sous-ensembles, l'un constituant des observations disponibles pour la prédiction, l'autre constituant les données de tests (base d'apprentissage et base de validation). La méthode de prédiction utilisée par le *LMD* est fondée sur un réseau de neurones *MLP* (Bishop, 1995). Au moment où nous avons traité les données, ce *MLP* était capable de prédire les concentrations de gaz dans la base de validation avec un écart type de 0.25 ppmv environ. Nous avons repris cette étude avec une approche par krigeage.

Étude préliminaire. Nous commençons l'analyse des données en estimant le variogramme des données en fonction de la distance euclidienne (figure 6.25). La variance des accroissements s'avère ne dépendre que très peu de la distance euclidienne entre les facteurs (le variogramme $\hat{\gamma}(h)$ est approximativement constant quelle que soit la distance h entre les vecteurs de facteurs). Un modèle avec un processus aléatoire stationnaire isotrope ne semble donc pas pertinent. Notons que le

[5] Il n'est pas possible de révéler la nature de ce gaz dans le cadre de ce manuscrit.

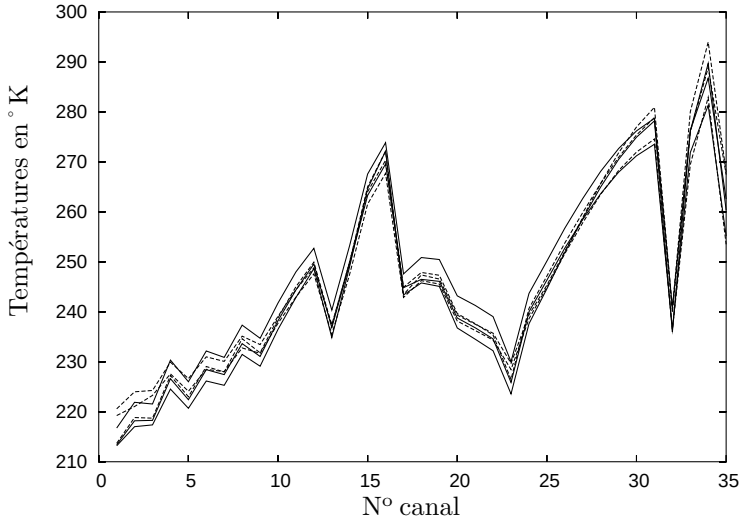


FIG. 6.24 – Vecteurs de températures de brillance pour une concentration de gaz égale à 352 ppmv (courbes en trait plein) et 392 ppmv (courbes en trait discontinu)

variogramme est calculé en fonction de la distance euclidienne mais le système peut n'être ni isotrope ni même intrinsèquement isotrope d'ordre 1 (isotropie des accroissements filtrant les tendances linéaires).

Nous effectuons ensuite la régression linéaire de la concentration du gaz par les facteurs. La variance d'erreur sur les données d'apprentissage est d'environ 1.29 ppmv et de 1.31 ppmv sur les données de validation, ce qui représente environ 2% de la variation totale de la sortie. Il y a donc une tendance linéaire. Nous estimons également le variogramme des résidus en fonction de la distance euclidienne (figure 6.25). Nous constatons que le variogramme des résidus possède une structure plus marquée que dans le cas précédent, ce qui permet d'envisager des résultats intéressants avec une *IRF* d'ordre un ou plus.

Pour terminer cette étude préliminaire, nous présentons à la figure 6.26 un histogramme des distances des points de la base d'apprentissage au centre de gravité de la base. Nous constatons que les distances ont tendance à se concentrer autour d'une distance moyenne. Il n'y pas de point au voisinage immédiat du centre de gravité. Nous interprétons ces résultats comme un phénomène de concentration de la mesure qui suggère que la prédiction par krigeage à partir de covariances isotropes sera difficile .

Première expérience. Un premier essai de prédiction avec une $IRF(1)$ et une covariance polynomiale d'ordre 1 isotrope donne un écart type d'erreur de prédiction sur les données de validation de 0.75 ppmv environ. Les paramètres de la covariance sont estimés par maximum de vraisemblance. L'estimation prend quelques minutes. Si nous utilisons une $IRF(2)$ en gardant une covariance polynomiale d'ordre 1, l'écart type d'erreur de prédiction sur les données de validation tombe à 0.445 ppmv. Lors de ce premier essai, nous n'avons donc pas atteint les performances du *MLP* mais l'erreur de prédiction est d'un ordre de grandeur acceptable.

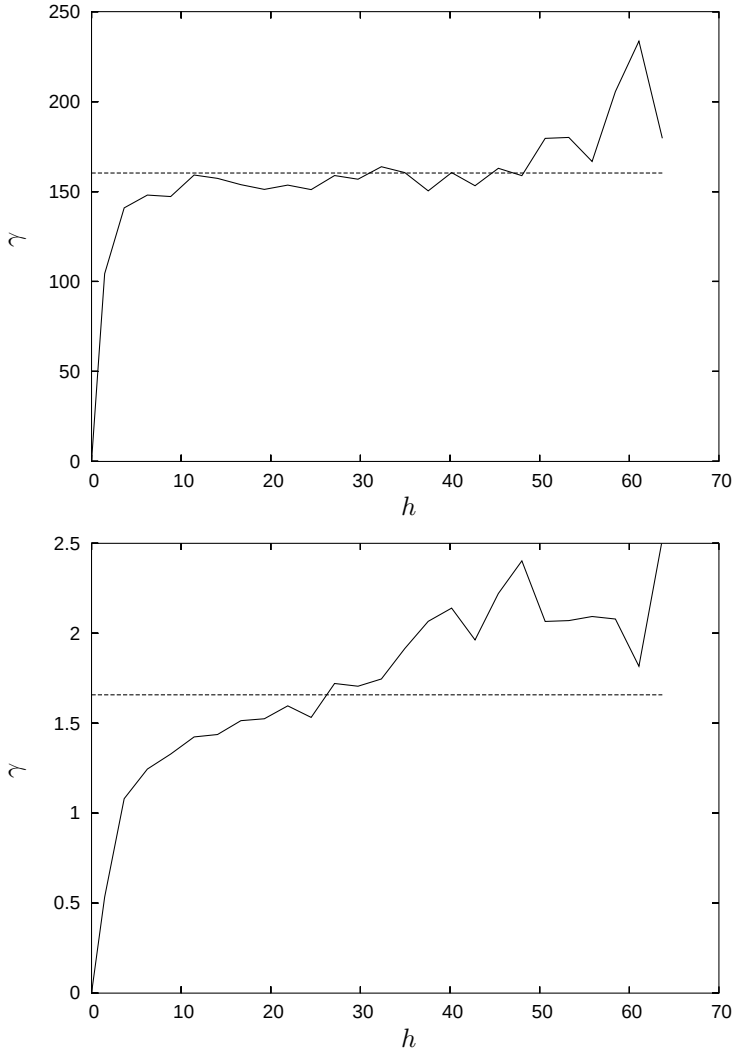


FIG. 6.25 – Figure du haut : variogramme des données. Valeur de la variance des données représentée en trait interrompu. Figure du bas : variogramme des résidus. Valeur de la variance des résidus représentée en trait interrompu

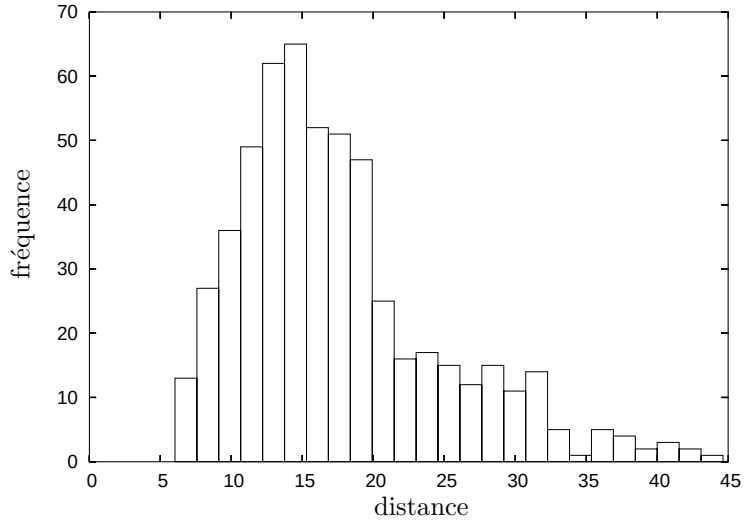


FIG. 6.26 – Histogramme des distances des données à leur centre de gravité

Deuxième expérience. Nous pensons que la principale limitation de l'approche précédente est de considérer une covariance isotrope. Or il est probable que le système ne soit pas du tout isotrope. Notamment, certains canaux présentent une grande variance de température mais sont peut-être très peu sensibles aux variations de la concentration du gaz. À l'inverse, certains canaux présentant une faible variance de température peuvent comporter beaucoup d'information sur la concentration. L'utilisation de la distance euclidienne peut donc faire perdre de l'information contenue dans les canaux sensibles au gaz étudié.

Pour essayer de contourner cette difficulté, nous opérons empiriquement une transformation linéaire de l'espace des facteurs en effectuant une dilatation dans une direction particulière. L'objectif est d'éloigner les points comportant des informations différentes sur la concentration. Nous choisissons la direction du vecteur $\boldsymbol{b}$ obtenu lors de la régression linéaire en fonction des facteurs, c'est-à-dire le vecteur $\boldsymbol{b}$ minimisant $\sum_{i=1}^{n}(f_{\boldsymbol{x}_i}^{\text{obs}} - \boldsymbol{b}^{\mathsf{T}}\boldsymbol{x}_i)^2$. La méthode donne des résultats intéressants puisqu'après transformation de l'espace des facteurs, nous obtenons un écart type d'erreur de prédiction sur les données de validation d'environ 0.411 ppmv.

Troisième expérience. Nous essayons de généraliser l'approche précédente en effectuant une transformation linéaire de l'espace des facteurs dont les paramètres sont estimés à l'aide d'un critère du maximum de vraisemblance (les paramètres de la covariance généralisée restent fixés). Pour ce faire, nous estimons les paramètres d'une matrice triangulaire supérieure $\boldsymbol{V}$ telle que les nouveaux facteurs s'expriment sous la forme $\boldsymbol{V}\boldsymbol{x}$. Une telle matrice comporte $n(n+1)/2 = 630$ éléments à estimer. Nous utilisons une procédure d'optimisation de type gradient conjugué nécessitant une évaluation coûteuse[6] du gradient. Nous initialisons la matrice $\boldsymbol{V}$ à la matrice identité et commençons l'optimisation vis-à-vis des seuls éléments diagonaux. Nous optimisons ensuite $\boldsymbol{V}$ sur

[6]Le calcul de chaque composante du gradient nécessite notamment d'évaluer une matrice de taille $n \times n$. Il serait très intéressant d'étudier l'apport des techniques de différentiation automatique pour effectuer de tels calculs.

une base d'apprentissage réduite. Nous utilisons enfin la totalité de la base d'apprentissage pour terminer l'optimisation. (La procédure complète dure quelques heures avec les moyens actuels de micro-informatique.) La figure 6.27 représente la matrice $\boldsymbol{V}^\mathsf{T}\boldsymbol{V}$ obtenue après optimisation. Nous constatons qu'elle est très différente de la matrice identité initiale, ce qui tend à confirmer le caractère anisotrope de l'espace des facteurs. L'écart type de l'erreur de prédiction linéaire sur la base de validation utilisant l'espace des facteurs transformé tombe à 0.36 ppmv. Autrement dit, le résultat reste moins bon que celui obtenu par *MLP*. Nous interprétons ce fait comme étant lié à la difficulté de l'optimisation d'une matrice $\boldsymbol{V}$ comportant beaucoup de paramètres. Cette interprétation est justifiée par la quatrième expérience.

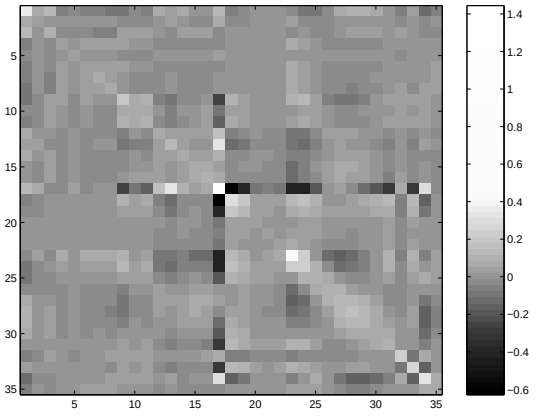


FIG. 6.27 – Représentation de la matrice $\boldsymbol{V}^\mathsf{T}\boldsymbol{V}$.

Quatrième expérience. Nous supprimons maintenant des facteurs. Ce faisant, l'erreur de prédiction diminue ou augmente selon les facteurs supprimés. Par une recherche systématique, nous atteignons une erreur de prédiction de 0.37 ppmv environ, c'est-à-dire mieux que dans la deuxième expérience. Nous en tirons deux conclusions.

- dans un problème où la dimension de l'espace des facteurs est élevée, si certains facteurs ne contribuent pas (ou très peu) aux variations de la sortie, ces facteurs « inutiles » perturbent la prédiction ;
- les méthodes d'estimation de la covariance ne sont pas très performantes dans le cas des espaces de facteurs de dimension élevée.

En conclusion, nous ne sommes pas parvenu à atteindre le résultat obtenu avec le *MLP*, même si les écarts type des erreurs de prédiction sur les données de validation sont du même ordre de grandeur (0.36 ppmv pour le krigeage, 0.25 ppmv pour le *MLP*). Nous n'avons pas cherché à aller plus loin dans cette étude pour le moment parce que notre travail est surtout motivé par les aspects méthodologiques et l'utilisation de méthodes systématiques (nous n'avons pas cherché à optimiser empiriquement la covariance, ce qui pourrait permettre éventuellement de battre le *MLP*). Nous tirons les conclusions suivantes.

Tout d'abord, nos modèles à noyaux se comparent ici défavorablement au meilleur réseau de neurones obtenu. Nuançons cette affirmation. En fait, nous n'avons pas réellement comparé une méthode de régression régularisée à noyau et un réseau de neurone. Nous avons plutôt comparé une

(ou deux) méthodes d'optimisation de la covariance à une procédure d'optimisation des poids d'un réseau de neurones. Par conséquent, une conclusion plus exacte serait que *nous ne disposons pas pour le moment de méthodes efficaces d'optimisation de la covariance lorsque l'espace des facteurs est de grande dimension.* Cette constatation nous a poussé a chercher d'autres possibilités pour choisir un noyau adapté à partir des données (voir nos propositions dans les sections 5.4.4, 6.5 et l'annexe A).

Par ailleurs, nous ne sommes pas des spécialistes de l'application traitée. Nous n'avons pas bénéficié de la même expertise que celle utilisée pour construire le réseau de neurones. Dans l'étude effectuée au *LMD*, la structure du *MLP* a été choisie pour intégrer le plus de connaissance a priori possible (prétraitementssur les facteurs, prise en compte d'a priori sur les sorties, etc.). Nous aurions probablement amélioré nos prédictions en prenant en compte ces informations.

D'autres possibilités auraient pu être envisagées. Citons, par exemple, celle qui consisterait à utiliser la corrélation spatiale (sur le globe terrestre) de la concentration du gaz étudié, ou celle qui utiliserait comme information a priori les concentrations prédites par des modèles climatologiques du transport du gaz. Il nous semble que ces dernières approches seraient très pertinentes mais difficiles à mettre en œuvre avec un *MLP*.

Nous pensons également que l'inconvénient majeur des *MLP* est qu'il est très difficile (car très empirique) de choisir la structure du réseau (nombre de couches, nombre de neurones dans les couches cachées, procédure d'optimisation). Finalement, le coût de la procédure d'optimisation du *MLP* (environ quatorze heures sur un ordinateur puissant) est beaucoup plus élevé que celui des optimisations du noyau que nous avons essayées (quelques heures sur un ordinateur personnel dans le cas le plus complexe, et ce pour un résultat comparable).

6.5 Prédiction de séries temporelles

Dans cette section, nous explorons le problème de la prédiction de séries temporelles en utilisant les méthodes de régression à noyau étudiées dans ce mémoire. Nous traitons une série temporelle comme des observations échantillonnées d'une réalisation d'un processus gaussien à *temps continu* et proposons une méthode pour choisir la covariance. L'avantage de ce point de vue est qu'il permet aussi de considérer des séries temporelles *non uniformément échantillonnées.*

Pour illustrer la méthode, considérons la série chronologique classique du nombre de lynx au Canada entre 1821 et 1934 (voir la figure 6.28). Elle est étudiée en détails dans (Brockwell et Davis, 1987), ce qui permet de comparer facilement les performances des modèles proposés. Nous la modélisons par un processus gaussien stationnaire $F(t)$, où $t \in \mathbb{R}$ est le facteur temps, de moyenne inconnue et de covariance $k(h)$, à choisir. La série observée correspond donc à une réalisation $\boldsymbol{f}^{\text{obs}} \in \mathbb{R}^n$ du vecteur aléatoire $(F(t_1), \ldots, F(t_n))^\mathsf{T}$ et la prédiction est un problème d'extrapolation, pour lequel nous avons vu des exemples dans la section 6.1.1. La difficulté du problème est de choisir une covariance appropriée pour $F(t)$.

Le caractère périodique de la série explique la forme oscillante de la fonction d'autocovariance empirique de la série, présentée à la figure 6.29. Cette remarque suggère de prendre $k(h)$ sous la forme d'une covariance avec des oscillations amorties. Toutefois, des expériences numériques avec des modèles de covariance simples[7] (c'est-à-dire comportant relativement peu de paramètres)

[7] Ces expériences numériques sont menées avec des covariances de type cosinus amorti.

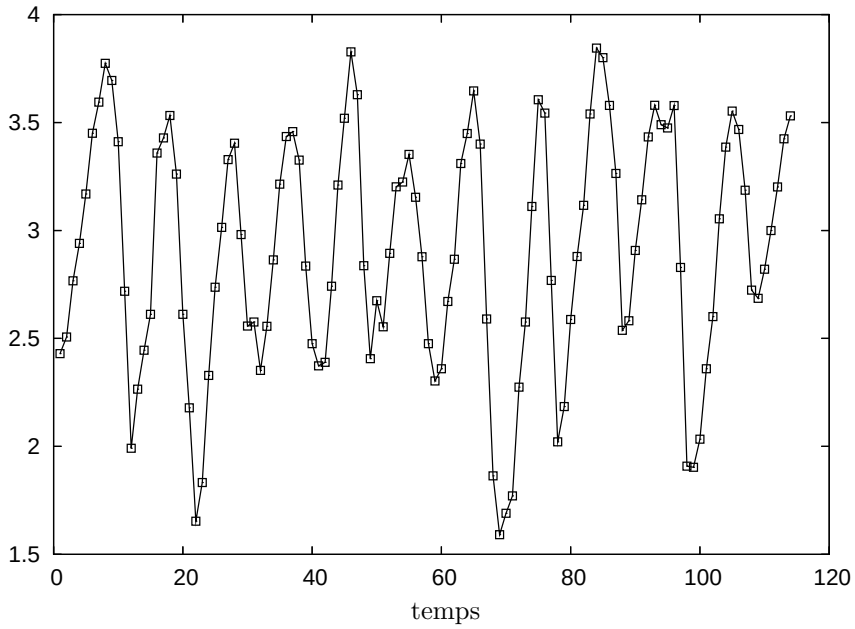


FIG. 6.28 – Évolution du logarithme en base 10 du nombre de lynx au Canada sur la période 1821–1934 (les abscisses correspondent à des pas de temps en années).

ne conduisent pas à des résultats satisfaisants. Notons que le modèle proposé par (Brockwell et Davis, 1987) est un AR d'ordre 12 (sélectionné d'après un critère d'Akaike), ce qui signifie qu'il y a typiquement six modes de résonance. Le modèle de covariance de cet $AR(12)$ possède donc une capacité d'adaptation aux données plus importante que les covariances à temps continu mentionnées ci-dessus. Ceci souligne la faiblesse des procédures classiques de choix de covariance où l'on sélectionne des noyaux dans des familles restreintes (voir la discussion de la section 5.4.4). Nous proposons ci-dessous une solution adaptée au cas d'espaces de facteurs unidimensionnels.

En se fondant sur les éléments de la section 3.4.1, nous cherchons à écrire la covariance de $F(t)$ sous la forme

$$k(t,s) = (\boldsymbol{r}(t), \boldsymbol{r}(s))_{\mathbb{R}^{2l}} ,$$

où $\boldsymbol{r}(t)$ est une fonction vectorielle de dimension $2l$ s'écrivant sous la forme

$$\boldsymbol{r}(t) = [\alpha_1 \cos u_1 t,\ \alpha_1 \sin u_1 t,\ \alpha_2 \cos u_2 t,\ \alpha_2 \sin u_2 t,\ \ldots,\ \alpha_l \cos u_l t,\ \alpha_l \sin u_l t]^\mathsf{T} .$$

Les pulsations u_i peuvent être choisies a priori et il y a donc alors l paramètres ajustables dans cette covariance, que nous réécrivons

$$k(t,s) = \sum_{i=1}^{l} \alpha_i^2 \cos u_i (t-s) ,$$

ce qui montre que k est invariante par translation. Pour des raisons de mise en œuvre, nous la reparamétrons sous la forme

$$k(h) = \sum_{i=1}^{l} e^{\alpha_i} \cos u_i h . \tag{6.3}$$

Notons que la périodicité de $k(h)$ n'est pas gênante si l'on prend soin de choisir une période plus grande que l'horizon d'étude de $F(t)$.

Nous souhaitons estimer les paramètres α_i lorsque l est grand (typiquement quelques centaines). La forme paramétrique (6.3) s'apparente alors à une série de Fourier. Une première idée consiste à estimer les α_i par maximum de vraisemblance restreint (*REML*). Cette méthode est préférée à celle du maximum de vraisemblance parce que la moyenne de $F(t)$ est inconnue. D'après la figure 6.29, la covariance à temps continu ainsi obtenue par cette estimation apparaît qualitativement proche de la covariance empirique.

Pour évaluer la qualité du modèle, nous utilisons le même critère S que dans (Brockwell et Davis, 1987), c'est-à-dire la moyenne quadratique des erreurs de prédictions à un pas sur les 14 dernières valeurs de la série. S vaut 0.138 dans le cas de l'$AR(12)$ proposé par Brockwell et Davis (1987) et environ 0.350 dans notre cas. Notre résultat de prédiction à un pas est donc très mauvais. En conclusion, même s'il existe une bonne fidélité entre le modèle de covariance et la covariance empirique, la capacité de prédiction (ou de généralisation) n'est pas satisfaisante. Il s'agit d'un phénomène de sur-adaptation aux données qui s'explique par le fait que le modèle de covariance comporte plus de paramètres qu'il n'y a de données.

Il apparaît nécessaire de régulariser les paramètres α_i, ce qui peut se faire par une méthode d'estimation du maximum a posteriori avec un a priori de régularité sur les paramètres. Comme ces paramètres s'apparentent à une estimée de la densité spectrale et que nous voulons que cette estimée soit relativement régulière, l'idée proposée est de modéliser les paramètres α_i par un processus gaussien $\alpha(u)$, tel que $\alpha_i = \alpha(u_i)$, $i = 1, \ldots, l$. Autrement dit, la densité de probabilité du vecteur aléatoire $\boldsymbol{\alpha} = (\alpha(u_1), \ldots, \alpha(u_l))^\mathsf{T}$ s'écrit sous la forme

$$p(\boldsymbol{\alpha}) = \frac{1}{Z} \exp\left(-\frac{1}{2} (\boldsymbol{\alpha} - \bar{\boldsymbol{\alpha}})^\mathsf{T} \boldsymbol{K}_{\boldsymbol{\alpha},\mathrm{reg}}^{-1} (\boldsymbol{\alpha} - \bar{\boldsymbol{\alpha}}) \right) ,$$

où Z est une constante de normalisation, où $\bar{\boldsymbol{\alpha}}$ est la moyenne du vecteur $\boldsymbol{\alpha}$, et où $\boldsymbol{K}_{\boldsymbol{\alpha},\mathrm{reg}}$ est sa matrice de covariance. Nous faisons l'hypothèse que la moyenne du processus $\alpha(u)$ est inconnue mais possède une forme linéairement paramétrée, par exemple du type $b_0 + b_1 u$. Il reste à effectuer le choix de la covariance de $\alpha(u)$. Nous avons choisi une covariance de type exponentielle en ajustant ses paramètres par une validation croisée rudimentaire.

Avec cet a priori, nous choisissons une estimée de $\boldsymbol{\alpha}$ au sens du MAP, c'est-à-dire en maximisant

$$\begin{aligned} J(\boldsymbol{\alpha}) = &- \frac{n-q}{2} \log 2\pi - \frac{1}{2} \log \det(\boldsymbol{W}^\mathsf{T} \boldsymbol{K}(\boldsymbol{\alpha}) \boldsymbol{W}) - \frac{1}{2} \boldsymbol{z}^\mathsf{T} (\boldsymbol{W}^\mathsf{T} \boldsymbol{K}(\boldsymbol{\alpha}) \boldsymbol{W})^{-1} \boldsymbol{z} \\ &- \frac{1}{2} \boldsymbol{\alpha}^\mathsf{T} \boldsymbol{W}_{\boldsymbol{\alpha}} (\boldsymbol{W}_{\boldsymbol{\alpha}}^\mathsf{T} \boldsymbol{K}_{\boldsymbol{\alpha},\mathrm{reg}} \boldsymbol{W}_{\boldsymbol{\alpha}})^{-1} \boldsymbol{W}_{\boldsymbol{\alpha}}^\mathsf{T} \boldsymbol{\alpha} \end{aligned} \tag{6.4}$$

par rapport à $\boldsymbol{\alpha}$. L'expression (6.4) est constituée de deux parties. La première correspond à la log-vraisemblance restreinte formée à partir d'une matrice de contrastes $\boldsymbol{W}$ de taille $n \times (n-q)$, où q est la dimension de l'espace $\mathcal{N}$ des fonctions polynomiales contenant la moyenne inconnue de $F(t)$. Le vecteur des contrastes $\boldsymbol{z}$ s'obtient donc par la transformation linéaire $\boldsymbol{z} = \boldsymbol{W}^\mathsf{T} \boldsymbol{f}^{\mathrm{obs}}$. La matrice $\boldsymbol{K}(\boldsymbol{\alpha})$ désigne la matrice de covariance du vecteur des observations, formée à partir de $k(h)$ paramétrée par $\boldsymbol{\alpha}$. (Voir la section 5.4.5 pour plus de détails concernant l'estimation *REML*.) Le dernier terme de (6.4) correspond à la log-densité restreinte du vecteur des paramètres formée

à partir d'une matrice de contrastes $\boldsymbol{W}_{\boldsymbol{\alpha}}$. Nous utilisons cette formulation pour ne pas avoir à prendre en compte la moyenne inconnue du processus $\alpha(u)$.

Il est aisé de maximiser (6.4) par rapport aux paramètres α_i parce que le gradient possède une expression analytique simple. Sur la série des lynx nous obtenons ainsi des valeurs S entre 0.120 et 0.125, selon le choix des paramètres de la covariance de $\alpha(u)$. Les performances de prédiction sont donc meilleures que celles obtenues avec le modèle $AR(12)$ de Brockwell et Davis (1987). La covariance estimée est présentée à la figure 6.29. Les paramètres $\hat{\alpha}_i$ estimés sont représentés en fonction des pulsations u_i sur la figure 6.30 où nous avons également représenté la densité spectrale du processus $AR(12)$ de Brockwell et Davis (1987). On constate la ressemblance entre les deux courbes. L'avantage du modèle proposé est que l'on contrôle de manière flexible l'adaptation aux données. En conclusion, cette méthode, qui peut s'appliquer aussi à des séries non uniformément échantillonnées, se révèle pertinente et pourrait constituer la base de futurs travaux. Les résultats présentés dans cette section restent toutefois préliminaires et d'autres possibilités pourraient être envisagées.

* * *

Nous avons également exploré une seconde méthode de prédiction, dans laquelle l'idée était de corriger un modèle autorégressif linéaire par un modèle autorégressif non-linéaire à l'aide de formulations semi-paramétriques. Nous mentionnons cette méthode mais nous pensons qu'elle reste très préliminaire. Dans ce contexte, une série chronologique $(x_t)_{t\in\mathbb{Z}}$ est modélisée par des variables aléatoires X_t, $t \in \mathbb{Z}$, telles que

$$X_t = F(X_{t-1}, X_{t-2}, \ldots, X_{t-p}) + N_t \,,$$

où $F(x_{[1]}, \ldots, x_{[p]})$ est un processus aléatoire gaussien paramétré par d variables (indépendant de la série (X_t)), et (N_t) est un bruit blanc centré. Nous supposons que la moyenne du processus est une fonction linéaire des paramètres $x_{[1]}, \ldots, x_{[p]}$. Il y a deux possibilités. Si nous supposons établi un modèle $AR(p)$ $\phi(B)X_t = Z_t$, où $\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p$ est un polynôme de degré p, B est l'opérateur retard et Z_t un bruit blanc centré, la moyenne de F peut être choisie sous la forme

$$\mathrm{E}[F(x_{[1]}, \ldots, x_{[p]})] = b_0 + b_1(\phi_1 x_{[1]} + \cdots + \phi_p x_{[p]}) \,,$$

où $b_0, b_1 \in \mathbb{R}$ sont des constantes inconnues. Si aucun modèle AR n'est déterminé à l'avance, la moyenne de F peut s'exprimer sous la forme

$$\mathrm{E}[F(x_{[1]}, \ldots, x_{[p]})] = b_0 + b_1 x_{[1]} + \cdots + b_p x_{[p]} \,,$$

avec des constantes $b_i \in \mathbb{R}$ inconnues. Ces deux formulations nous permettent d'introduire dans un modèle NAR un modèle AR, considéré comme modèle a priori. Nous avons appliqué cette méthode à la série des lynx en utilisant un ordre 12 d'autorégression et une covariance de Matérn, dont les paramètres sont estimés par la méthode *REML*. Les résultats obtenus sont décevants. Typiquement, il n'a pas été possible d'améliorer les performances de prédiction. Nous interprétons ce résultat comme étant lié à la difficulté de prédire efficacement dans des espaces de grande dimension et avec un faible nombre de données (dans notre cas, il s'agit de prédire en dimension 12 avec une centaine de données seulement).

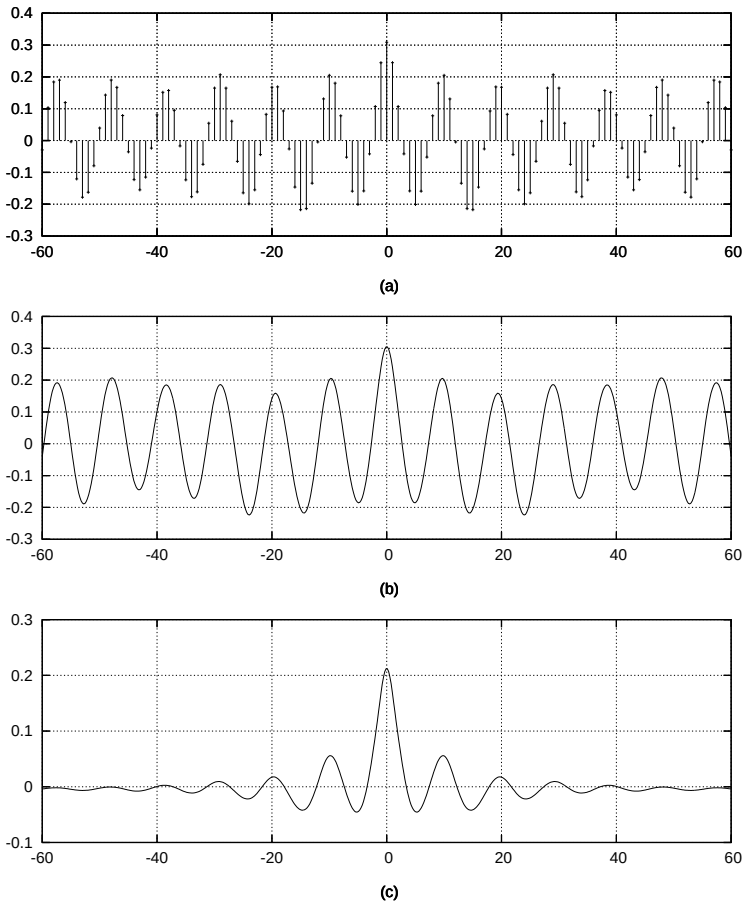


FIG. 6.29 – Fonctions de covariance de la série des lynx ; (a) covariance à temps discret estimée empiriquement (estimateur *non biaisé*) ; (b) covariance à temps continu estimée par maximum de vraisemblance ; (c) covariance à temps continu estimée par maximum a posteriori.

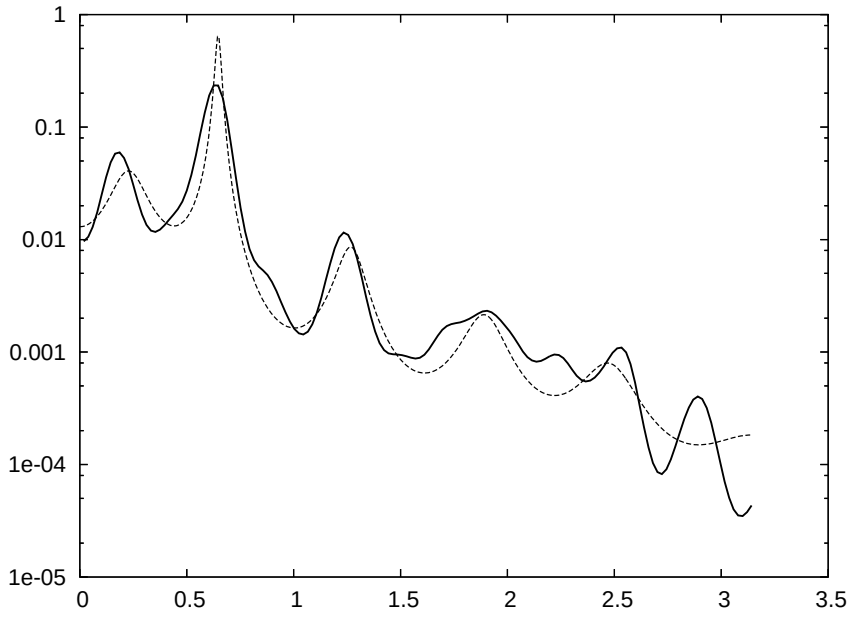


FIG. 6.30 – Représentation des paramètres estimés (et renormalisés) de la covariance en fonction de la pulsation (trait continu). Densité spectrale du modèle $AR(12)$ de Brockwell et Davis (1987) (trait interrompu).

6.6 Éléments de planification d'expérience : problème de construction d'éclateurs à gaz

Cette section présente nos résultats dans le cadre d'un étude industrielle pour la société THALES, pour le compte de la DGA/SPAÉ. Dans cette étude, nous posons le problème de la planification d'expériences, c'est-à-dire celui du choix d'un ensemble d'observations permettant de minimiser l'erreur de prédiction d'un modèle boîte noire. Nous présentons les éléments sur lesquels nous avons commencé à réfléchir et la démarche retenue (Sacks et al., 1989) pour cette application qui n'est qu'à ses débuts.

6.6.1 Présentation et formalisation du problème

Un *éclateur à gaz* est un composant fortement non-linéaire pouvant être simulé au moyen d'un modèle comportemental conçu par les experts du domaine. Ce *modèle de simulation* est un système d'équations différentielles représenté par un schéma électrique équivalent (voir la figure 6.31) comportant des composants passifs non-linéaires. Les *paramètres* du modèle doivent être adaptés à chaque type d'éclateur. Un type d'éclateur est spécifié par des *paramètres technologiques* définis par le fabricant, et qui diffèrent dans leur nature de ceux des paramètres du modèle de simulation. Pour adapter les paramètres du modèle de simulation à un type d'éclateur, l'expert procède à une série de tests sur un éclateur réel et en déduit la valeur des paramètres de simulation. La valeur

des paramètres ainsi estimés présente naturellement une incertitude liée au caractère expérimental des tests.

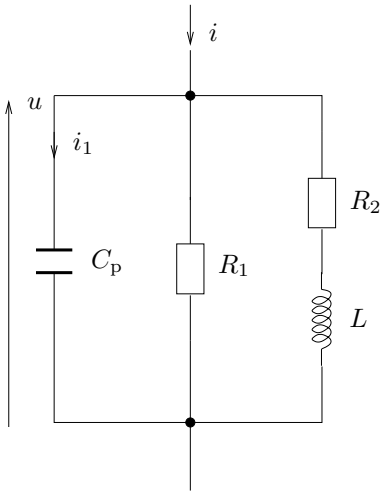


FIG. 6.31 – Schéma électrique équivalent d'un éclateur.

L'objectif de l'étude est la prédiction des paramètres du modèle de simulation d'un éclateur à partir de la donnée de ses paramètres technologiques. Il s'agit donc d'éviter la construction effective des éclateurs à gaz, suivie de la phase de tests. L'opération inverse, qui consiste à déterminer les paramètres technologiques d'un éclateur à partir d'un jeu de paramètres du modèle de simulation, doit aussi être envisagée. Le tableau 6.3 énumère les paramètres du modèle de simulation et les paramètres technologiques.

Paramètres de simulation	Paramètres technologiques
capacité parasite C_{p}	distance inter-électrode
tension d'arc V_{arc}	volume/surface des électrode
tension V_{stat}	pression partielle d'argon
puissance P_{max}	dilution de la poudre d'activation
inductance L_0	nature du métal d'électrode
constante K	type de traits en graphite
courant I_{seuil}	gaz d'addition
courant I_0	

TAB. 6.3 – Les deux types de paramètres d'un éclateur.

On remarque que les paramètres de simulation sont tous quantitatifs, contrairement à certains paramètres technologiques. On appellera ces paramètres non quantitatifs *facteurs catégoriels*. Du point de vue de la modélisation, les facteurs catégoriels sont difficiles à utiliser. Distinguons trois approches possibles pour incorporer des facteurs catégoriels. Nous pouvons tout d'abord chercher des grandeurs numériques pertinentes pour caractériser les catégories. Par exemple, dans le cas du paramètre « nature du métal », il pourrait s'agir de la conductivité électrique et thermique.

Notons que cela peut conduire à considérer des facteurs supplémentaires. Une autre approche conduisant aussi à une augmentation du nombre des facteurs consiste à coder un facteur catégoriel par des valeurs binaires, valant zéro ou un pour chaque catégorie. Une dernière approche consiste à attribuer à un facteur catégoriel une grandeur numérique, comme par exemple la valeur moyenne de la grandeur à prédire pour une catégorie.

Dans cette étude, le modèle boîte noire donnant la relation entre les paramètres technologiques et les paramètres de simulation sera obtenu à partir de $n = 22$ expériences (nombre limité par le cahier des charges), chaque expérience consistant à construire une série d'éclateurs supposés identiques, spécifiés à l'aide des d paramètres technologiques (paramètres qui joueront donc le rôle des facteur notés $\boldsymbol{x}$) et à effectuer les tests permettant d'en déduire les paramètres du modèle de simulation. Les q paramètres de simulation joueront donc le rôle du vecteur des sorties $\boldsymbol{f}(\boldsymbol{x})$.

6.6.2 Généralités sur la planification des expériences

Dans cette section, nous rappelons brièvement les différents types d'échantillonnage et quelques notions sur la planification d'expérience (voir par exemple Sacks et al., 1989 ; Walter et Pronzato, 1997).

Vocabulaire classique sur l'échantillonnage

Passons en revue quelques possibilités d'échantillonnage. Notons que le problème de l'échantillonnage a été très étudié en statistiques (Cochran, 1977).

Échantillonnage systématique. Il s'agit de l'échantillonnage régulier évoqué plus haut. L'inconvénient d'un échantillonnage effectué sur des grilles est qu'il peut demander beaucoup d'observations en dimension élevée.

Échantillonnage aléatoire. n expériences sont choisies au hasard. C'est la méthode la plus simple.

Échantillonnage aléatoire stratifié. L'échantillonnage aléatoire stratifié permet d'améliorer l'échantillonnage aléatoire dans certain cas, notamment quand on possède des informations sur le processus. Il consiste à déterminer des blocs (les strates) et d'effectuer un échantillonnage aléatoire dans ces blocs (éventuellement une seule observation par bloc). Par exemple, dans le cas des éclateurs à gaz, les experts possèdent des données sur des éclateurs qualifiés de « faible » et « robuste ». Il serait possible d'établir des strates selon la distance à ces observations, en supposant que les principales variations des grandeurs seront selon la direction « faible » – « robuste ». Il existe d'autres possibilités.

Échantillonnage probabiliste. L'échantillonnage dit probabiliste est le cadre théorique permettant de regrouper les échantillonnages précédents. Chaque site dans l'espace des facteurs se voit attribuer une probabilité d'être sélectionné.

Échantillonnage adaptatif. Une série d'observation est effectuée, et la stratégie pour les observations suivantes est décidée en conséquence pour satisfaire un critère. Cette façon de procéder nous semble être à privilégier compte tenu du faible nombre d'expériences autorisé et pour éviter de gaspiller des expériences.

Échantillonnage avec information a priori et facteurs explicatifs. Cet échantillonnage nous semble la stratégie la plus intéressante. Il s'agit de prendre en compte dans la stratégie d'échantillonnage l'effet sur la sortie de quelques facteurs seulement que l'on sait avoir une influence essentielle. Cette stratégie rejoint une remarque de Stein dans son commentaire faisant suite à (Sacks et al., 1989). Étudier la mise en œuvre systématique de cette démarche fait partie de nos perspectives.

Le protocole expérimental à mettre en place pour établir le modèle doit permettre de minimiser le nombre d'expériences à qualité donnée ou de maximiser la qualité du modèle à coût expérimental donné. Pour planifier une expérience, il faut commencer par définir un modèle ainsi qu'un critère de qualité et optimiser ce critère par rapport aux expériences, c'est-à-dire ici, au choix des vecteurs de facteur $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$.

Dans le cas de modèles paramétriques, l'approche classique est de choisir un critère relié à l'incertitude sur les paramètres. Ce critère est généralement une fonction scalaire de la matrice d'information de Fisher $\boldsymbol{I}_{\boldsymbol{\theta}}$. Rappelons que sous certaines hypothèses, l'estimateur du maximum de vraisemblance des paramètres tend asymptotiquement ($n \to \infty$) vers la distribution de loi normale $\mathcal{N}(\boldsymbol{\theta}^*, \boldsymbol{I}_{\boldsymbol{\theta}^*}^{-1})$, où $\boldsymbol{\theta}^*$ est le vecteur de paramètres inconnus à estimer. Le critère d'optimalité le plus utilisé correspond à minimiser $\det \boldsymbol{I}_{\boldsymbol{\theta}^*}^{-1}$, ou de manière équivalente maximiser $\det \boldsymbol{I}_{\boldsymbol{\theta}^*}$. Ce critère, dit de *D-optimalité* (D comme déterminant), revient à minimiser le volume des ellipsoïdes d'incertitude asymptotique sur les paramètres estimés. D'autres critères peuvent être utilisés.

Deux grands types de stratégie d'optimisation de ces critères existent. Le premier cherche l'optimum global du critère d'optimalité pour n expériences simultanément. Le second est séquentiel et conduit à optimiser itérativement le critère en ajoutant une nouvelle expérience à chaque étape. Dans le cas d'un modèle linéairement paramétré, la matrice de Fisher ne dépend pas des paramètres et il est aisé d'établir un protocole expérimental pour déterminer les paramètres du modèle de façon optimale.

Pour des modèles de type krigeage, le critère à optimiser naturel est plutôt fondé sur la variance de prédiction (voir Sacks et al., 1989). Reprenons l'exemple de la figure 2.3. On constate que la variance de prédiction est plus grande dans l'intervalle $[3, 5]$ où il n'y pas de données. Une observation supplémentaire semble donc intéressante dans la zone où la prédiction est la plus mauvaise (vers $x \approx 4$). Plus généralement, nous pouvons envisager différentes stratégies pour choisir un bon échantillonnage.

- maximiser la quantité d'information (minimiser l'entropie) des sites échantillonnés à propos des sites non observés à prédire (Shewry et Wynn, 1987, 1988). Si les données sont gaussiennes, cela correspond à maximiser la corrélation entre m sites à prédire et n sites échantillonnés.
- minimiser le maximum de la variance d'erreur de prédiction sur le domaine d'étude (*Maximum Mean Square Error* selon la terminologie de Sacks et al. (1989)). Comme les maxima de variance d'erreur de prédiction sont généralement observés pour les points éloignés des

sites observés, cela revient approximativement à minimiser la plus grande distance entre les sites observés.
- minimiser la moyenne de la variance de l'erreur de prédiction.
- minimiser des moyennes pondérées de la variance d'erreur de prédiction (*Integrated Mean Square Error* selon la terminologie de Sacks et al. (1989)).

Pour une structure de covariance isotrope, pour laquelle la covariance décroît de façon monotone en fonction de la distance, un échantillonnage régulier qui minimise la variance maximale d'erreur de prédiction peut être trouvé (Künsch et al., 2003). En dimension deux par exemple, il s'agit d'un échantillonnage régulier triangulaire. S'il existe des anisotropies, la grille d'échantillonnage optimale sera déformée. Cependant, ces considérations ne tiennent que si la fonction de corrélation est connue. Dans le cas contraire, il est souhaitable d'effectuer un échantillonnage préliminaire pour déterminer la structure des données. Dans ce cas, l'échantillonnage le plus adapté n'a aucune raison d'être régulier, ce que confirme l'expérience numérique ci-dessous.

6.6.3 Expérience numérique sur la planification d'expériences pour découvrir la structure des données

Dans ce paragraphe, nous nous intéressons à l'influence de l'échantillonnage sur l'estimation du modèle aléatoire par la méthode du maximum de vraisemblance. Notons que cette question est différente mais pas éloignée de celle de la planification d'expérience. L'expérience que nous proposons consiste à simuler N réalisations d'un processus aléatoire gaussien centré, stationnaire et de covariance de Matérn de paramètres $\nu = 2$, $\rho = 0.1$ et $\sigma^2 = 1$. Le processus aléatoire est échantillonné sur $[0, 1]$ de deux façons :

1. aux points $x_i = i/6$, $i = 0, \ldots, 6$;
2. aux points $x_i = 2^{-i}$, $i = -6, \ldots, 0$.

Le premier échantillonnage correspond donc à un échantillonnage régulier et le second à un échantillonnage logarithmique. Le premier échantillonnage sera donc plus satisfaisant que le second du point de vue de la prédiction. Cependant, si le critère considéré est non plus l'erreur de prédiction mais la variance d'estimation des paramètres de la covariance, nous pouvons nous attendre à ce que la situation soit renversée. Pour le vérifier, nous calculons pour chaque réalisation la dérivée seconde de la log–vraisemblance relative au paramètre ν de la covariance. La moyenne de ces dérivées secondes est, à un signe près, l'information de Fisher relative au paramètre ν. Un histogramme de ces dérivées secondes est présenté à la figure 6.32 et nous constatons d'après la table 6.4 que la variance estimée de l'estimateur du maximum de vraisemblance du paramètre ν est beaucoup plus grande dans le cas de l'échantillonnage régulier.

échantillonnage	écart type
N° 1	6.5
N° 2	0.7

TAB. 6.4 – Écarts type estimés pour $\hat{\nu}$.

Une autre situation peut conduire à ne pas considérer systématiquement l'échantillonnage régulier. Il s'agit du cas où certaines informations a priori sont connues. Si le comportement du

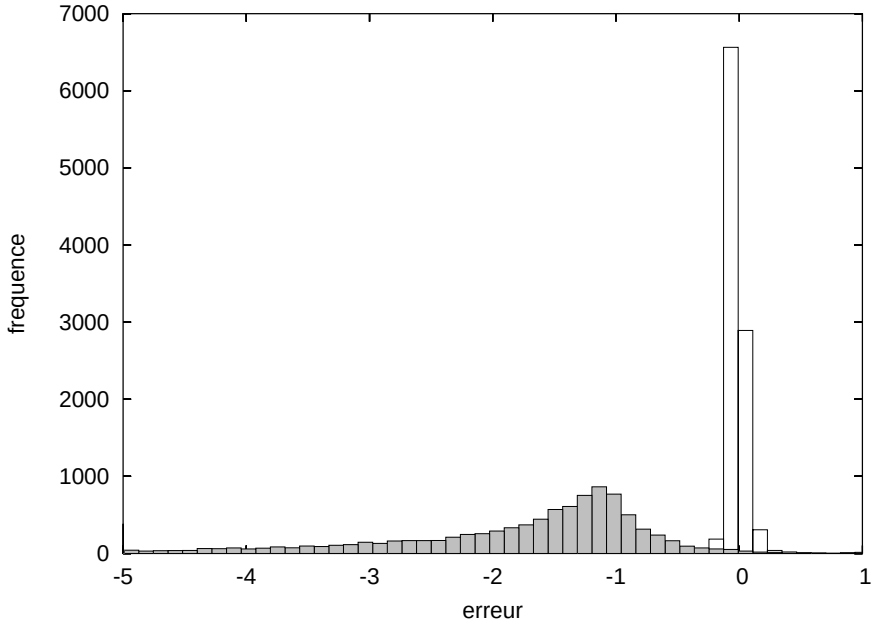


FIG. 6.32 – Histogrammes de la dérivée seconde de la log–vraisemblance pour 10000 réalisations et tableau des écarts type estimés. L'histogramme correspondant à l'échantillonnage N° 1 est représenté en blanc ; celui correspondant à l'échantillonnage N° 2 est en gris.

système est connu a priori dans une région donnée de l'espace des facteurs, il semble inutile de faire des observations dans cette région (voir le paragraphe 6.6.4).

6.6.4 Inclusion nécessaire d'information a priori

Comme nous l'avons noté au chapitre 5, les performances d'une modélisation de type boîte noire dépendent du nombre d'observations mais aussi, de la dimension de l'espace des facteurs, ce problème étant connu sous le terme anglais de *curse of dimensionality*. Si la dimension de l'espace des facteurs augmente il faudra naturellement plus de données pour explorer cet espace d'une façon satisfaisante. Dans cette application se pose donc de nouveau le problème de la limitation du nombre d'observations disponibles pour établir le modèle. D'une part, nous avons peu d'observations et d'autre part, l'espace des facteur est de grande dimension.

Par exemple, en dimension 2, 4 expériences suffisent pour parcourir les sommets d'un carré, mais si l'on retient 7 facteurs, correspondant aux paramètres de simulation par exemple, il faut $2^7 = 128$ observations pour parcourir les sommets d'un hypercube dans l'espace des facteurs. Notons qu'interpoler une fonction en dimension 2 avec 4 points seulement est déjà une situation délicate. Nous pouvons également nous intéresser au nombre de points (ou expériences) nécessaires pour définir un polynôme d'ordre k dans un espace de dimension n. Ce nombre est C_{n+k}^n (autant d'expériences que de paramètres). Par exemple, il faut $C_3^2 = 3$ points pour déterminer de manière unique un polynôme d'ordre 1 en dimension 2, c'est-à-dire du type $a_0 + a_1 x_{[1]} + a_2 x_{[2]}$. Il faut 6 points pour définir un polynôme d'ordre 2 du type $a_0 + a_1 x_{[1]} + a_2 x_{[2]} + a_3 x_{[1]} x_{[2]} + a_4 x_{[1]}^2 + a_5 x_{[2]}^2$. En dimension 7, il faut 7 points pour définir un polynôme d'ordre 1 et 28 pour définir un polynôme d'ordre 2. Ces considérations montrent qu'il n'est pas possible de déterminer un polynôme complet d'ordre supérieur à 1 en dimension 7 à partir de 22 expériences seulement, comme imposé dans le cahier des charges de cette étude. On pourrait donc penser a priori que l'on ne pourra pas proposer mieux qu'un modèle paramétrique linéaire avec éventuellement quelques termes mixtes (monômes en $x_{[i]} x_{[j]}$).

Pour remédier à ce problème, il y a deux types de solution. Nous pouvons essayer de réduire le nombre des facteurs, en faisant l'analyse des facteurs qui n'interviennent pas pour une grandeur à prédire. Le deuxième type de solution est d'introduire de l'information a priori sur la nature du processus étudié.

Nous avons demandé aux experts de nous fournir trois types d'informations. Nous cherchons d'abord les informations pouvant aider à déterminer un modèle de covariance. Ces informations sont par exemple les incertitudes sur les paramètres technologiques (les facteurs) et sur les paramètres de simulation déterminés à partir de ceux-ci (ces incertitudes correspondent pour nous à des bruits d'observation), ainsi que la sensibilité ou la rapidité des variations des grandeurs de sortie en fonction des facteurs (l'objectif est d'estimer qualitativement la portée de la covariance). Nous avons ensuite cherché à déterminer quels facteurs a priori sans influence sur certaines grandeurs de sortie. Ce type d'information permet alors d'envisager de diminuer le nombre d'expériences nécessaires en réduisant la dimension de l'espace des facteurs. Enfin, le comportement des grandeurs à prédire est parfois plus ou moins connu, même si cette information se réduit à des comportements au « premier ordre ». Nous avons particulièrement insisté sur ce dernier point. Par comportement au « premier ordre », nous entendons une idée, même vague, de l'influence des facteurs sur les

grandeurs de sorties, que l'expert du domaine est souvent capable de fournir. Comme dans la section 6.3, il est possible d'inclure cette information dans un modèle de type boîte noire sous forme de termes paramétriques. Par exemple, la capacité parasite C_p du modèle de simulation varie, au premier ordre, proportionnellement à S/d, avec S la surface des électrodes et d la distance inter-électrode. Notons qu'il n'est pas obligatoire de disposer d'un modèle analytique. Une information du type « telle grandeur augmente avec tel facteur » est déjà intéressante. Les termes paramétriques peuvent correspondre, par exemple, à des termes linéaires en les facteurs. Enfin, il peut s'agir de définir des comportements aux limites, ou des ruptures dans les grandeurs à prédire.

Mentionnons enfin que nous avons demandé les résultats des tests effectués sur des éclateurs existants. Cela pourrait permettre d'établir un premier modèle et faciliter la planification des expériences.

6.6.5 Méthode de planification retenue

Définition des facteurs et du domaine d'étude

Nous avons demandé des indications précises sur le domaine d'étude des paramètres technologiques. À la suite de cette demande, il a été jugé utile de restreindre le domaine d'étude. Cette réduction permet de supprimer les dépendances entre les paramètres technologiques. Cela signifie que l'on a pas de contrainte sur des paramètres liées aux valeurs prises par d'autres paramètres[8]. Le domaine d'étude a également été restreint de manière à limiter le nombre de facteurs catégoriels.

Nous avons finalement déterminé que deux facteurs étaient essentiels lors du choix des expériences : la pression partielle d'argon, $P_{\mathrm{Ar}} \in [30, 300]$, et la distance inter-électrodes, $d \in [0.4, 0.8]$. Les autres facteurs sont listés ci-dessous :

- nature du métal $\in$ {FN42, Cu},
- gaz d'addition, avec 4 possibilités exclusives : aucun gaz d'addition, néon dilué à 90%, azote dilué à 50% ou hydrogène dilué à 10%,
- poudre d'activation : PR $\in \{0\%, 30\%, 60\%, 85\%\}$,
- traits graphites (4 configurations possibles).

La démarche adoptée pour les facteurs de nature catégorielle est d'assigner des valeurs numériques aux catégories (voir les paragraphes suivants pour plus de détails).

Approche itérative de la planification

Nous savons que 22 expériences ne suffiront qu'à établir un modèle très approximatif des relations entre les paramètres. Le modèle que nous construirons sera un interpolateur et redonnera les valeurs des paramètres du modèle de simulation mesurées aux points d'expériences. Cependant, on sait bien que plus on s'éloigne des points expérimentaux et plus les prédictions deviennent incertaines. Si des zones ne nous intéressent pas, il faut donc éviter d'y gaspiller des expériences.

Comme le coût d'une expérience est élevé, nous souhaitons adopter une approche parcimonieuse. Nous proposons donc une approche séquentielle, exploitant les résultats des expériences précédentes pour le choix des expériences suivantes. À chaque étape de la procédure, la covariance

[8]Sur le domaine d'étude initialement prévu, il était nécessaire d'ajouter des traits graphites pour les grandes pressions d'argon. La plage de variations de la pression d'argon, initialement de [5 mbar, 500 mbar], a été réduite à [30 mbar, 300 mbar].

du modèle sera réestimée en maximisant la vraisemblance des données obtenues dans les expériences passées. Nous voyons deux avantages à cette approche. S'il apparaît a posteriori qu'un facteur n'influence pas la sortie du système ou bien que son influence est facilement prévisible, la décision de ne pas étudier ce facteur plus avant pourra être prise. Nous pourrons aussi étudier si l'information a priori est bien conforme aux résultats expérimentaux et adapter nos décisions dans le cas contraire.

Nous exposons dans les paragraphes suivants la démarche mise en œuvre pour obtenir les premières expériences à effectuer. Un nombre de quatre expériences initiales a été retenu. Nous établissons tout d'abord une structure de covariance a priori qui nous semble adaptée au problème de la modélisation d'éclateurs (cette structure est toutefois choisie en l'absence de toute donnée). Nous utilisons ensuite cette structure de covariance pour calculer les variances des erreurs de prédiction, obtenues après les quatre expériences. Nous choisissons quatre expériences qui minimisent la moyenne de la variance de l'erreur de prédiction sur le domaine d'étude.

Choix de la structure de covariance a priori

La difficulté du choix d'une structure de covariance pour le problème posé tient à la dimension de l'espace des facteurs, à la nature catégorielle de certains facteurs et à l'absence de données. Comme de plus, très peu d'information est disponible, le choix de cette structure initiale de covariance est largement arbitraire. Les paramètres de cette structure de covariance seront estimés seulement après les résultats des premières expériences. (La structure de covariance pourra bien évidemment être modifiée au cours des expériences et notamment, pour établir le modèle final.)

Pour simplifier la mise en œuvre nous choisissons une fonction de covariance radiale $k(h)$, $h \in \mathbb{R}^+$. Cette covariance peut par exemple être choisie du type Matérn et comporte alors trois paramètres. Notons que le paramètre correspondant à la variance n'a aucune incidence sur la planification des expériences. Bien évidemment, nous ne supposons pas que la structure du problème est radiale. Nous opérons une transformation des facteurs d'entrées $\boldsymbol{x}$ (les paramètres technologiques) visant à établir une distance pertinente entre les facteurs.

La démarche suivante a été adoptée.

1. Tous les facteurs sont à valeurs dans [0,1].
2. Les pressions d'argon et la distance inter-électrodes sont modifiées pour donner les éléments $x_{[1]}$ et $x_{[2]}$ du vecteur $\boldsymbol{x}$. Nous utilisons les transformations affines

$$x_{[1]} = \frac{P_{\mathrm{Ar}} - 30}{300 - 30}$$

et

$$x_{[2]} = \frac{d - 0.4}{0.8 - 0.4}\,.$$

3. Nous codons les deux possibilités pour la nature du métal par l'élément $x_{[3]} \in \{0, 1\}$.
4. Les quatre possibilités pour le gaz d'addition sont codées par l'élément $x_{[4]} \in \{0, 1/3, 2/3, 1\}$.
5. La dilution de la poudre d'activation (PR) est codé par l'élément $x_{[5]} \in \{0, 0.30, 0.60, 0.85\}$.
6. Les traits graphites sont codés par l'élément $x_{[6]} \in \{0, 1/3, 2/3, 1\}$.

Comme les facteurs $x_{[1]}$ et $x_{[2]}$ sont fondamentaux dans l'étude nous choisissons de calculer une distance à partir de ces deux facteurs. Cette distance est ensuite modifiée par des facteurs multiplicatifs. La formule empirique suivante est utilisée pour calculer la distance entre deux expériences $\boldsymbol{x}_i$ et $\boldsymbol{x}_j$:

$$h = \sqrt{(dx_{[1]})^2 + (dx_{[2]})^2}\rho_{\text{métal}}\rho_{\text{gaz}}\rho_{\text{PR}}\rho_{\text{traits}} \,,$$

avec $dx_{[1]} = x_{[1],i} - x_{[1],j}$ et $dx_{[2]} = x_{[2],i} - x_{[2],j}$. Les coefficients multiplicatifs $\rho_{\text{métal}}$, ρ_{gaz}, ρ_{PR}, ρ_{traits}, supérieurs ou égaux à un, sont calculés en fonction des facteurs $x_{[4]}$ à $x_{[6]}$ selon l'idée suivante. Si on compare deux mêmes configurations, avec par exemple les mêmes métaux, alors les variables ρ correspondantes sont égales à un. Si les configurations diffèrent, alors les sorties seront décorrélés dans une certaine mesure. Nous tenons compte de cette décorrélation en augmentant la distance de manière empirique. Notons que cette démarche conduit à considérer des paramètres supplémentaires dans la fonction de covariance. Il aurait été équivalent de calculer $k(h)$ avec $h = ((dx_{[1]})^2 + (dx_{[2]})^2)^{1/2}$ et de modifier $k(h)$ ensuite pour tenir compte des décorrélations, par exemple en multipliant $k(h)$ par des coefficients de décorrélation inférieurs ou égaux à un. Les deux démarches seront étudiées dans les futurs travaux.

Insistons encore sur le caractère très artificiel du choix de la distance et de la covariance. En raison de l'absence d'information avant le résultat des premières expériences, la démarche choisie nous semble cependant raisonnable, car elle privilégie le rôle des deux facteurs principaux tout en tenant compte des facteurs catégoriels par des coefficients d'ajustement.

Procédure d'optimisation

La démarche choisie pour sélectionner les quatre expériences initiales est présentée dans cette section. Pour choisir les expériences $\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3$ et $\boldsymbol{x}_4$, nous minimisons la moyenne de la variance d'erreur de prédiction sur le domaine d'étude. Pour évaluer cette moyenne, nous choisissons un maillage du domaine d'étude comportant 15488 points, incluant des grilles 11x11 pour les deux facteurs principaux. Le problème d'optimisation est non-linéaire et porte sur un vecteur de paramètres de dimension 24 (6 facteurs * 4 expériences = 24 paramètres). En raison de l'existence de facteurs catégoriels et du choix de la covariance, le coût à minimiser n'est pas continu. La procédure d'optimisation est donc coûteuse en temps de calcul et peut converger vers des minima locaux. L'algorithme retenu est une méthode du simplexe de Nelder–Mead. C'est une méthode de recherche directe (sans information sur le gradient) adapté au cas multidimensionnel et sans contrainte.

Répétons qu'il faut être prudent vis-à-vis de cette procédure. Même si un minimiseur global est trouvé, il ne faudrait pas en déduire que les expériences résultantes seront effectivement optimales au sens du minimum de la moyenne sur le domaine d'étude de la variance de l'erreur. En effet, la procédure dépend de la covariance choisie mais celle-ci est inconnue en l'absence de résultats sur les premières expériences. Pour palier au moins partiellement l'incertitude sur la covariance, nous pouvons toutefois répéter la procédure d'optimisation en choisissant différents paramètres de covariance.

Résultats de l'optimisation

Il est facile de constater que l'optimisation converge souvent vers des solutions différentes. L'existence de celles-ci s'explique en partie par les symétries du problème et par l'équivalence du critère par permutations des expériences. Elle pourrait aussi résulter de l'existence de minimiseurs locaux parasites. Nous devons donc prendre soin de choisir plusieurs conditions initiales, typiquement aléatoirement. On constate alors heureusement que la valeur de la fonction coût à l'optimum est à peu près la même et que les solutions trouvées sont toujours plus ou moins semblables, en particulier pour les pressions d'argon et la distance inter-électrodes.

Les résultats de plusieurs optimisations suggèrent finalement de réaliser les expériences caractérisées par les tableaux 1 et 2.

$x_{[1]}$	$x_{[2]}$	$x_{[3]}$	$x_{[4]}$	$x_{[5]}$	$x_{[6]}$
0.24	0.23	0	0	0.60	1/3
0.76	0.22	0	0	0.30	1/3
0.24	0.74	0	2/3	0.30	1/3
0.76	0.73	1	1	0.60	2/3

TAB. 6.5 – Premières expériences proposées (paramètres normalisés)

argon	distance	métal	gaz	PR	traits
95	0.50	cuivre	aucun	60%	2 traits (L=1.3mm)
235	0.49	cuivre	aucun	30%	2 traits (L=1.3mm)
95	0.70	cuivre	azote 50%	30%	2 traits (L=1.3mm)
235	0.70	FN42	hydrogène 10%	60%	2 traits (L=1.6mm)

TAB. 6.6 – Premières expériences proposées (paramètres technologiques)

Ces résultats semblent raisonnables même s'ils dépendent, comme mentionné plus haut, des conditions initiales de l'algorithme d'optimisation et du choix de la covariance. La figure 6.33 montre l'emplacement des expériences dans le plan pression d'argon – distance inter-électrodes. Nous suggérons donc d'utiliser ces points pour effectuer les premières expériences.

Nous n'avons pas encore eu les résultats des expériences au moment de l'écriture de ce mémoire. L'étape suivante consistera à estimer les paramètres de la covariance à partir des données obtenues. La procédure sera répétée jusqu'à obtenir 22 expériences.

6.7 Conclusions

Ce chapitre a présenté une série d'exemples d'application du krigeage à des problèmes de nature variée. Ces exemples montrent à la fois la simplicité d'utilisation et les performances des modèles boîte noire présentés aux chapitres précédents. D'autres exemple d'applications de modélisation boîte noire par krigeage pourront être trouvés dans (Lefèbvre et al., 1996 ; Costa et al., 1999,

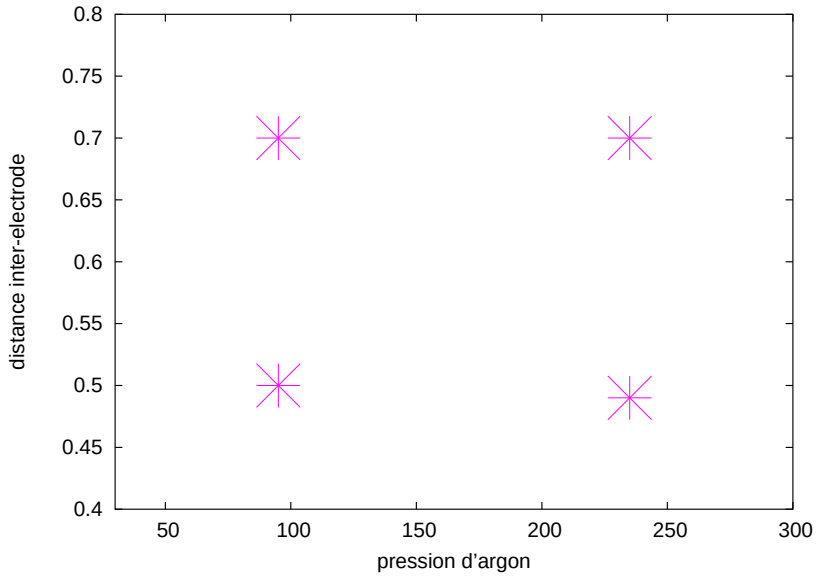


FIG. 6.33 – Projection des quatre premières expériences proposées dans le plan pression d'argon – distance inter-électrodes

2000) par exemple. Des exemples de modélisation de systèmes dynamiques pourront être trouvés dans (Girard, 2004).

Dans notre travail, nous avons surtout insisté sur les aspects méthodologiques et les difficultés rencontrées en pratique, notamment lorsque la dimension de l'espace des facteurs croît. Plus cette dimension est élevée, plus le problème de prédiction est délicat car il faut davantage d'observations pour couvrir correctement le domaine d'étude et la structure du processus générant les données peut devenir complexe. Ce chapitre illustre également les difficultés liées aux faibles nombres d'observations et la pertinence de la méthode proposée dans la section 4.6.1 pour incorporer des connaissances a priori. Nous recommandons son utilisation systématique chaque fois qu'il est possible de prédire a priori une évolution approximative de la sortie en fonctions des facteurs. L'application à la prédiction de séries chronologiques et la méthode que nous proposons pour estimer la covariance nous semblent prometteuses.

Enfin, nous avons appliqué le krigeage au problème de la planification d'expériences dans le cadre d'un contrat industriel. Cet exemple est d'un intérêt particulier parce qu'il permet de confronter réellement nos méthodes aux difficultés des problèmes industriels (nombre de facteurs relativement élevé, facteurs catégoriels, limitation du nombre d'expériences en raison de leur coût, etc.). D'un point de vue théorique, la planification d'expérience est encore relativement mal comprise. En effet, l'approche que nous avons retenue (Sacks et al., 1989) optimise les expériences en supposant connue la covariance du processus aléatoire, ce qui implique l'utilisation d'une procédure itérative alternant les phases d'estimation de la covariance et les phases de planification des expériences. Comprendre l'influence d'une erreur de modèle sur la planification est un problème ouvert et important. Il constitue un verrou sur la voie vers une procédure permettant d'optimiser les expériences en supposant la covariance inconnue.

6.8 Annexe : mise en œuvre et algorithmes

Les problèmes de mise en œuvre constituent un sujet d'étude en soi que nous n'avons pas spécifiquement abordé dans notre travail. Les méthodes que nous avons utilisées sont les méthodes classiques de résolution de systèmes linéaires et d'optimisation. Dans cette section, nous effectuons un survol des points importants.

6.8.1 Prédiction linéaire

Résolution de systèmes linéaires

L'étape de prédiction linéaire nécessite de résoudre un système d'équations linéaires avec une matrice de covariance $\boldsymbol{K}$ symétrique définie non-négative. Dans le cas le plus fréquent, $\boldsymbol{K}$ ne possède pas de structure particulière, contrairement à ce qui se passe dans le cas de la prédiction des séries chronologiques. Lorsque les observations ne sont pas répétées pour une même valeur du vecteur des facteurs, $\boldsymbol{K}$ est de rang plein mais peut être mal conditionnée.

La résolution d'un système linéaire nécessite généralement une capacité de stockage en mémoire en $\mathrm{O}(n^2)$ et un coût algorithmique en $\mathrm{O}(n^3)$. En pratique, il est préférable d'utiliser des factorisations de $\boldsymbol{K}$ adaptées (voir plus bas). Avec les outils d'informatique personnelle, des problèmes de taille allant jusqu'à cinq mille données peuvent être aisément considérés. Cependant, lorsque le nombre de données augmente, il est fréquent que la matrice de covariance devienne mal conditionnée (cela dépend de la régularité de la fonction de covariance et de la distance entre les observations dans l'espace des facteurs par rapport à la portée de la covariance). Lorsque $\boldsymbol{K}$ est mal conditionnée, une première solution est d'ajouter $\varepsilon \boldsymbol{I}_n$ à $\boldsymbol{K}$ afin de forcer le conditionnement à une valeur adéquate. Ceci revient à considérer un bruit d'observation de variance ε qui modifie les valeurs prédites de façon négligeable si ε est suffisamment petit. Une autre possibilité consiste à approximer $\boldsymbol{K}$ par une matrice de rang inférieur, ce qui présente un avantage en terme de coût algorithmique (voir plus bas).

Factoriser la matrice de covariance est une étape préliminaire importante pour diminuer le coût algorithmique de la résolution du système et améliorer la stabilité numérique lorsque $\boldsymbol{K}$ est presque singulière. En pratique, deux types de factorisation sont envisageables. Il s'agit de la factorisation de Cholesky et la de factorisation QR. Rappelons que la factorisation de Cholesky fournit une matrice triangulaire inférieure $\boldsymbol{C}$ telle que $\boldsymbol{K} = \boldsymbol{C}\boldsymbol{C}^\mathsf{T}$ et que la factorisation QR fournit une matrice triangulaire supérieure $\boldsymbol{R}$ et une matrice orthogonale $\boldsymbol{Q}$ telles que $\boldsymbol{K} = \boldsymbol{Q}\boldsymbol{R}$. La factorisation de Cholesky a l'avantage de la complexité algorithmique la moins élevée puisqu'elle nécessite $1/3n^3$ opérations alors que la factorisation QR en nécessite $2n^3$. Lorsque l'une ou l'autre des factorisations a été obtenue, le système linéaire peut être résolu en $\mathrm{O}(n^2)$ opérations seulement par une méthode de *remontée*. Cependant la factorisation de Cholesky est susceptible d'échouer lorsque $\boldsymbol{K}$ est mal conditionnée. La factorisation QR est alors préférable numériquement car elle permet d'obtenir une solution $\widehat{\boldsymbol{\lambda}}_{\boldsymbol{x}}$ au sens du minimum de $\|\boldsymbol{K}\widehat{\boldsymbol{\lambda}}_{\boldsymbol{x}} - \boldsymbol{k}_x\|$. La factorisation QR permet également de détecter et de supprimer les données redondantes. La méthode consiste à supprimer les colonnes de $\boldsymbol{R}$ correspondant aux éléments diagonaux inférieurs à un seuil (cela revient à supprimer des données). Cette méthode ne permet pas toutefois d'effectuer une sélection des données les plus représentatives, au sens de la régression à vecteurs de support par exemple.

Approximations par réduction de rang

Obtenir des approximations de rang inférieur présente plusieurs avantages importants. Il s'agit d'abord d'une solution envisageable lorsque la matrice de covariance est mal conditionnée. Ces approximations permettent aussi d'éliminer les données redondantes ou peu susceptibles d'améliorer l'erreur de prédiction, voire de sélectionner un sous-ensemble de données représentatives comme dans la régression à vecteurs de support. Enfin, ces approximations sont utiles pour résoudre des systèmes linéaires de grande dimension.

Dans la littérature, la réduction de rang est souvent traitée dans le cadre des systèmes linéaires à matrices singulières (*Rank-deficient problems* en anglais) ou lorsque l'on cherche des factorisations faisant apparaître le rang (*Rank-revealing factorizations*). Il existe de nombreuses méthodes. Les méthodes fondées sur une décomposition en valeurs singulières (Golub et Kahan, 1965) sont coûteuses et ne peuvent être considérées que pour des matrices de taille modeste (quelques centaines de colonnes). Les approximations envisageables en pratique sont fondées sur les décompositions UTV (Stewart, 1994 ; Mathias et Stewart, 1993 ; Fierro et Hansen, 1997), la factorisation QR, la factorisation QRP (Foster, 1986 ; Chan, 1987 ; Chan et Hansen, 1990 ; Bischof et Hansen, 1991) et la factorisation de Cholesky incomplète. Pour illustrer ces méthodes, prenons l'exemple d'une approximation utilisant une factorisation QRP. Cette méthode consiste à écrire $\boldsymbol{K}$ sous la forme

$$\begin{aligned}\boldsymbol{K} &= \boldsymbol{Q}\boldsymbol{R}\boldsymbol{P}^{\mathsf{T}} \\ &= \begin{pmatrix} \boldsymbol{Q}_1 & \boldsymbol{Q}_2 \end{pmatrix} \begin{pmatrix} \boldsymbol{R}_{1,1} & \boldsymbol{R}_{1,2} \\ \boldsymbol{0} & \boldsymbol{R}_{2,2} \end{pmatrix} \boldsymbol{P}^{\mathsf{T}},\end{aligned}$$

où $\boldsymbol{P}$ est une matrice de permutations. Une approximation de rang inférieur peut alors s'écrire sous la forme

$$\hat{\boldsymbol{K}} = \boldsymbol{Q}_1 \begin{pmatrix} \boldsymbol{R}_{1,1} & \boldsymbol{R}_{1,2} \end{pmatrix} \boldsymbol{P}^{\mathsf{T}}.$$

Pour résoudre le système linéaire, on considère la factorisation QR

$$\begin{pmatrix} \boldsymbol{R}_{1,1} & \boldsymbol{R}_{1,2} \end{pmatrix} = \begin{pmatrix} \boldsymbol{L} & \boldsymbol{0} \end{pmatrix} \boldsymbol{Q'}^{\mathsf{T}},$$

de manière à écrire $\hat{\boldsymbol{K}}$ sous la forme

$$\hat{\boldsymbol{K}} = \boldsymbol{Q}_1 \begin{pmatrix} \boldsymbol{L} & \boldsymbol{0} \end{pmatrix} \boldsymbol{Q'}^{\mathsf{T}} \boldsymbol{P}^{\mathsf{T}} = \boldsymbol{Q}_1 \begin{pmatrix} \boldsymbol{L} & \boldsymbol{0} \end{pmatrix} \boldsymbol{V}^{\mathsf{T}},$$

avec

$$\boldsymbol{V} = \boldsymbol{P}\boldsymbol{Q'} = \begin{pmatrix} \boldsymbol{V}_1 & \boldsymbol{V}_2 \end{pmatrix}.$$

Une solution approchée du système $\boldsymbol{K}\boldsymbol{\lambda}_{\boldsymbol{x}} = \boldsymbol{k}_x$ est donnée par

$$\boldsymbol{\lambda}_{\boldsymbol{x}} = \boldsymbol{V_1}\boldsymbol{L}^{-1}{\boldsymbol{Q}_1}^{\mathsf{T}}\boldsymbol{k}_{\boldsymbol{x}}.$$

Cette approche est illustrée par la figure 6.34.

Dans la littérature des méthodes à vecteurs de support, (Fine et Scheinberg, 2001) recommande la factorisation de Cholesky incomplète dans laquelle $\boldsymbol{K}$ se met sous une forme $\boldsymbol{C}\boldsymbol{C}^{\mathsf{T}}$, où $\boldsymbol{C}$ est une matrice triangulaire de même taille que $\boldsymbol{K}$ comportant éventuellement des colonnes égales à zéro. Cette factorisation s'obtient en considérant une version avec pivots de l'algorithme de

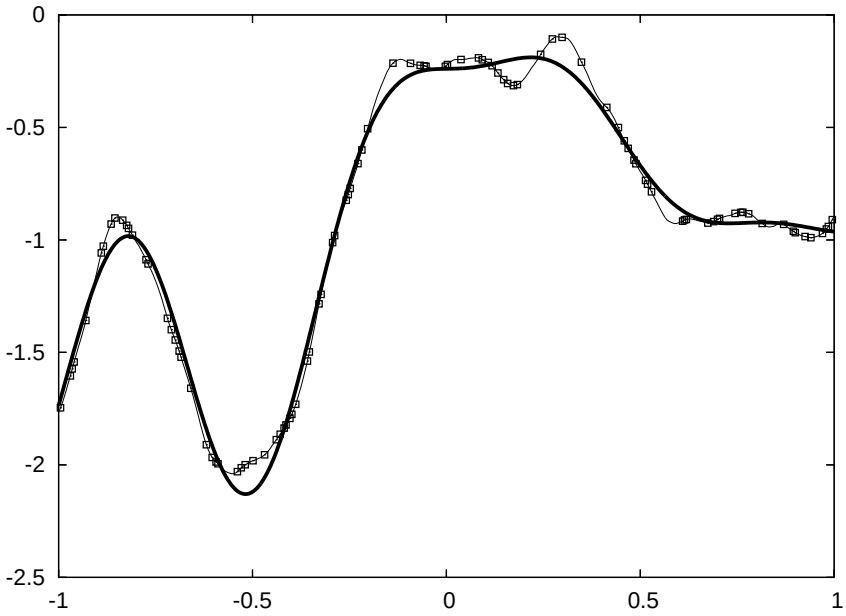


FIG. 6.34 – Réduction de rang par décomposition QRP tronquée de la matrice de covariance. La fonction générant les données (indiquées par les carrés) est représentée en trait fin. L'approximation par krigeage après réduction de rang de la matrice de covariance des données est tracée en trait épais (le rang est choisi égal à 10).

Cholesky et s'effectue en $\mathrm{O}(k^2 n)$ opérations, où k est le nombre de pivots non nuls utilisés lors de la factorisation, égal au rang de la matrice $\boldsymbol{K}$. Il est possible de ne pas prendre en compte les pivots en dessous d'une valeur seuil, ce qui permet de réduire le rang de la matrice $\boldsymbol{K}$ (Golub et Van Loan, 1983). À k fixé, on obtient un algorithme ayant un coût linéaire avec le nombre de données. Cette solution combine donc un avantage de faible complexité algorithmique et de robustesse.

Rappelons également l'existence de méthodes itératives efficaces (méthodes de gradient conjugué) pour résoudre des systèmes linéaires de grande dimension.

6.8.2 Estimation des paramètres

Le coût algorithmique des méthodes à noyaux est généralement dominé par l'étape d'estimation des paramètres de la covariance. Le calcul de la vraisemblance requiert $\mathrm{O}(n^3)$ opérations. Il est consiste à calculer le déterminant d'une matrice de taille $n \times n$ et à résoudre un système linéaire de même taille. Les problèmes numériques susceptibles d'être rencontrés sont essentiellement les mêmes que dans l'étape de prédiction linéaire. Notons que le calcul du déterminant s'effectue après avoir factorisé la matrice de covariance. Si l'on utilise par exemple la factorisation de Cholesky $\boldsymbol{C}\boldsymbol{C}^\mathsf{T}$ de $\boldsymbol{K}$, le déterminant s'obtient comme le carré du produit des éléments diagonaux de $\boldsymbol{C}$.

Les paramètres à estimer sont fréquemment des paramètres positifs. Afin d'utiliser des algorithmes d'optimisation sans contrainte, il est commode de prendre l'exponentielle des paramètres. Contrairement à ce que l'on pourrait penser, cette opération ne crée pas d'instabilité et tend parfois à faciliter la convergence vers les optimums globaux. Les algorithmes d'optimisation de type gradient conjugué sont les plus efficaces mais nécessitent le calcul du gradient de la log–vraisemblance, et donc l'inversion de la matrice de covariance avec un coût algorithmique en $\mathrm{O}(n^4)$[9]. Une autre difficulté est que le gradient ne s'obtient pas toujours analytiquement, comme dans le cas notamment des covariances de Matérn. On peut alors commencer par estimer les paramètres σ^2 et ρ d'une covariance exponentielle (en utilisant le gradient) pour choisir une valeur initiale pertinente des paramètres de la covariance de Matérn.

Au delà de plusieurs centaines de données, la recherche du maximum de vraisemblance devient coûteuse. Dans la plupart des cas, nous choisissons un sous-ensemble aléatoire des données pour effectuer l'optimisation. De nombreuses approches utilisent ce principe (par exemple Smola et Schölkopf, 2000 ; Stein et al., 2004).

[9] Nous pensons toutefois que l'utilisation de méthodes de différentiation automatique pourrait réduire ce coût car le calcul de la vraisemblance est seulement en $\mathrm{O}(n^3)$.

Chapitre 7

Conclusions et perspectives

7.1 Contributions

Une importante partie de notre travail a été consacrée à la synthèse de deux approches fondamentales de modélisation comportementale. La première est probabiliste. Il s'agit de la prédiction linéaire de processus aléatoires, ou krigeage. La seconde relève de l'analyse fonctionnelle. Elle regroupe des méthodes de régression régularisée par une norme d'espace hilbertien à noyau reproduisant (splines, *RBF*, *SVR*, etc.).

L'un de nos objectifs était de préciser les liens entre ces deux approches. Pour cela, nous avons puisé des résultats dans des domaines variés (géostatistique, modélisation des séries chronologiques, théorie des splines ou de l'apprentissage, etc.). Il ressort de cette étude que les liens entre processus aléatoires gaussiens et espaces hilbertiens à noyau reproduisant sont apparus très tôt (Parzen, 1962 ; Hajek, 1962 ; Kimeldorf et Wahba, 1970a). On constate cependant qu'ils ont été fort peu exploités. Or, ces deux points de vue permettent des interprétations complémentaires. Par exemple, le point de vue des processus aléatoires permet de comprendre comment formuler le problème de régression régularisée pour l'approximation de fonctions à valeurs vectorielles (cas des systèmes à plusieurs sorties). Le point de vue fonctionnel montre l'importance de la régularisation, qui est rarement mentionnée dans la littérature statistique ou géostatistique. Ce double point de vue nous a conduit à rappeler de manière formelle les définitions et les propriétés des processus aléatoires et des espaces hilbertiens à noyau reproduisant, afin de mieux saisir les relations subtiles qui existent entre ces deux notions (Lukic et Beder, 2001). Nous nous sommes ensuite intéressé au problème central du choix de noyau. Les résultats théoriques et expérimentaux suggèrent d'accorder une attention toute particulière à cette étape de la modélisation boîte noire. Là aussi, les méthodes diffèrent selon les domaines. Malheureusement, ce sujet est d'un abord relativement difficile d'un point de vue théorique et les principaux résultats en statistique sont seulement asymptotiques (Stein, 1999). Nous avons cependant proposé des expériences numériques originales qui permettent de mieux comprendre leur portée pratique. Notre synthèse reste toutefois incomplète. Il y manque par exemple l'apport récent de la théorie de l'information et de l'apprentissage sur la notion de bornes de risque (Smola, 1998). Il ne semble pas possible pour le moment de présenter les différentes approches de choix de noyau dans un cadre unifié. Nous espérons cependant que notre effort de

synthèse contribuera à établir une base solide pour de futurs travaux.

La mise en œuvre des méthodes de régression à noyaux est un sujet d'étude en soi, et notre travail n'était pas centré sur cet aspect. Le coût algorithmique principal est celui de l'étape d'estimation des paramètres du noyau, que l'on pourrait envisager de diminuer par des méthodes de différentiation automatique. Nous avons aussi suggéré l'utilisation de l'algorithme des innovations (section 2.5.4) pour la prédiction en ligne et l'utilisation d'une décomposition QR pour mettre en œuvre l'algorithme *REML*.

L'estimation de dérivées ou de primitives de fonctions à partir d'observations bruitées et échantillonnées irrégulièrement nous semble particulièrement intéressante. Nous avons illustré ces méthodes sans les appliquer à des problèmes réels pour le moment, mais nous pensons à leur utilisation pour la modélisation comportementale de systèmes dynamiques. Notre proposition d'extension de la méthode *SVR* à des systèmes à plusieurs sorties corrélées permet de bénéficier de la propriété de robustesse des *SVR* pour la prédiction de dérivées, et plus généralement, pour la modélisation de systèmes à sortie vectorielle.

Nous avons montré comment utiliser le krigeage intrinsèque (régression semi-régularisée) pour inclure de l'information a priori dans un modèle boîte noire. Cette méthode a donné des résultats très satisfaisants pour le problème de débitmétrie (section 6.3). De manière générale, elle nous semble d'un grand intérêt pour les applications.

Sur le problème de choix de noyau, nous avons présenté deux propositions visant à améliorer la qualité des modèles. Nous avons constaté que les noyaux classiques ne permettent de modéliser qu'une classe de systèmes restreinte. Nous avons donc cherché à construire des noyaux à partir de noyaux élémentaires, en envisageant deux possibilités. La première consiste à sélectionner un petit nombre de noyaux élémentaires représentant les données de manière pertinente. Dans la section 5.6, une méthode générale exploitant cette idée a été proposée, dont nous pensons cependant qu'elle reste encore à un stade préliminaire. La seconde possibilité consiste à assembler un noyau à partir d'un grand nombre de noyaux élémentaires. Ce faisant, il est nécessaire d'introduire un a priori sur la contribution de chacun afin de contrôler la capacité de généralisation du modèle. Dans la section 6.5 nous avons appliqué cette approche à la prédiction de séries chronologiques. Les résultats obtenus nous paraissent très encourageants.

Les applications présentées dans le chapitre 6 constituent une partie importante de notre travail. Nous avons essayé d'en dégager avant tout les aspects méthodologiques. Ces applications permettent de mieux saisir les avantages et les limites des méthodes utilisées, que nous détaillons dans la section suivante.

7.2 Modèles comportementaux par krigeage et méthodes à noyaux

(Pour ne pas alourdir cette discussion, nous emploierons ici le terme *krigeage* au sens large, c'est-à-dire pour désigner à la fois le krigeage, le krigeage intrinsèque et les méthodes de régression régularisée par une norme d'espace hilbertien à noyau reproduisant.)

Un modèle boîte noire est par définition construit à partir d'observations. Quelle que soit la méthode utilisée, la qualité d'un tel modèle dépend donc du nombre de ces observations. En

revanche, toutes les méthodes ne se comportent pas de la même façon. Un modèle satisfaisant doit posséder une bonne capacité d'approximation (capacité à décrire les données observées, par exemple au sens du risque empirique) tout en maximisant, à nombre d'observations constant, la capacité de généralisation (capacité à prédire des caractéristiques non observées, par exemple au sens du risque en espérance). Le krigeage possède ces deux propriétés lorsque les données ont été générées par un processus aléatoire dont la structure du second ordre est parfaitement connue. En effet, en tant qu'interpolateur redonnant les valeurs observées dans les expériences précédentes, le krigeage fournit un modèle qui représente au mieux les données (le risque empirique est nul). En tant que meilleure prédiction linéaire, le krigeage minimise l'erreur de généralisation lorsque le processus aléatoire générant les données est gaussien. Si le processus n'est pas gaussien, il existe de meilleurs modèles mais ces derniers sont plus difficiles à obtenir.

En réalité, le processus aléatoire générant les données est seulement un modèle de la sortie d'un système, qui exprime simplement l'hypothèse d'appartenance de la sortie à un espace de fonctions particulier (de manière équivalente, cette hypothèse revient à régulariser la sortie par la norme d'un espace hilbertien à noyau reproduisant). Le krigeage minimise l'erreur de généralisation conditionnellement à cette hypothèse d'appartenance, cette dernière ayant un caractère relativement arbitraire. Nous essayons donc en pratique de choisir une hypothèse telle que les données observées ne la contredisent pas. Il s'agit alors d'estimer la structure du second ordre du processus à partir des données. Deux questions naturelles peuvent être posées. Le krigeage reste-t-il robuste lorsque l'on change de modèle ? Existe-t-il des méthodes d'estimation de la structure du second ordre d'un processus aléatoire permettant de minimiser l'erreur de généralisation du modèle ?

Les résultats théoriques rassemblés dans ce mémoire montrent que l'erreur de généralisation est asymptotiquement décroissante pour une large gamme de fonctions de covariance (qui exclut cependant la covariance gaussienne). Cette propriété est déjà très satisfaisante par rapport à d'autres méthodes d'approximation. De plus, cette décroissance est optimale lorsque la covariance choisie est compatible en un certain sens avec celle du processus aléatoire ayant généré les données. Dans les régimes non asymptotiques il existe très peu voire pas de résultats. Nos expériences numériques présentées dans la section 5.3 indiquent cependant que les écarts entre les erreurs de généralisation de différents modèles sont significatifs.

Des réponses théoriques à la seconde question semblent encore plus difficiles à obtenir mais l'estimation au sens du maximum de vraisemblance nous paraît plutôt bien fondée. D'une part, il existe des résultats asymptotiques d'optimalité, bien que démontrés sous des hypothèses assez restrictives (Mardia et Marshall, 1984). D'autre part, la méthode du maximum de vraisemblance peut être vue comme l'application du principe de maximum d'entropie (voir la section 5.6). Cependant, l'estimation au sens du maximum de vraisemblance est loin de garantir la propriété de généralisation recherchée. En effet, le maximum de vraisemblance conduit à une sur-adaptation du modèle aux données observées s'il existe une valeur des paramètres de la covariance conduisant à une vraisemblance proche de un (par exemple, lorsque le nombre de paramètres est de l'ordre du nombre de données, comme dans la section 6.5). Cette sur-adaptation est généralement impossible avec les covariances classiques comportant peu de paramètres. Cependant, l'utilisation de covariances simples tend également à limiter la capacité de généralisation. Il nous semble donc préférable de privilégier des structures de noyau plus flexibles, tout en contrôlant le phénomène

de sur-adaptation. Dans la section 5.6, nous avons ainsi proposé une méthode de construction parcimonieuse de la covariance. Dans la section 6.5, nous avons proposé une méthode alternative fondée sur un estimateur au sens du maximum a posteriori. Notons que dans le domaine des séries chronologiques, de telles approches sont courantes (le critère d'Akaike, qui permet de limiter l'ordre d'un modèle autorégressif, est un exemple de contrôle de la sur-adaptation). Cette analyse rappelle que le choix du noyau constitue finalement le point essentiel de la méthode. Les problèmes dans ce domaine restent très ouverts.

Enfin, insistons encore sur l'extrême simplicité du krigeage, ce qui constitue à notre avis un de ses avantages les plus marquants. Le krigeage permet en effet de modéliser avec la même simplicité des systèmes observés avec ou sans bruit, des systèmes comportant des sorties corrélées, d'estimer des dérivées (ce qui est souvent considéré comme un problème difficile) ou de prendre en compte des connaissances a priori pour obtenir des modèles boîte grise. En résumé et en conclusion, l'utilisation du krigeage pour la modélisation comportementale de systèmes nous semble justifiée d'un point de vue tant théorique que pratique. Nous espérons avoir contribué par notre travail à rendre cette méthode mieux connue et nous pensons qu'elle offre de nombreuses perspectives d'étude.

7.3 Perspectives

Les fondements des méthodes utilisées dans ce mémoire sont aujourd'hui bien compris. Comme mentionné ci-dessus cependant, beaucoup de problèmes restent ouverts, comme par exemple celui des propriétés non asymptotiques des prédicteurs linéaires. À nos yeux, le point le plus délicat est bien sûr le choix du noyau. Nous pensons que les théories et les méthodes dont nous disposons aujourd'hui ne sont pas complètement satisfaisantes (probablement parce que les efforts des statisticiens se sont longtemps concentrés sur la modélisation des séries chronologiques échantillonnées régulièrement). Le problème devient particulièrement délicat lorsque l'espace des facteurs est de grande dimension car deux phénomènes jouent alors contre nous. D'une part, il faut davantage d'observations pour couvrir correctement le domaine d'étude et d'autre part, la structure du processus générant les données peut devenir complexe et par suite difficile à estimer. Dans ce cas, les outils traditionnels de la géostatistique sont plutôt inadaptés. D'autres approches doivent être envisagées, ce qui passe par une compréhension approfondie des concepts utilisés et de leur propriétés, et peut-être de nouvelles façons de considérer l'espace des facteurs (Adler, 2004). Une autre limite concerne l'utilisation de noyaux invariants par translation. Si cette pratique est systématique, elle n'est peut être pas toujours nécessaire. En effet, certaines méthodes de modélisation comportementale adaptent leurs caractéristiques à la densité d'échantillonnage (Bernard, 1999 ; Nelles et al., 2000). Cette adaptation de la régularité du noyau non pas en fonction des données mais en fonction de l'échantillonnage pourrait conduire à des procédures intéressantes.

Nos exemples de modélisation comportementale concernaient uniquement des systèmes statiques. Traditionnellement, les systèmes dynamiques sont traités en considérant des modèles *NARX* du type $y_{t+1} = f(y_t, y_{t-1}, \ldots, y_{t-m}, u_t, u_{t-1}, \ldots, u_{t-p})$ où y_t est la sortie du système au temps t et u_t est la commande du système (voir Girard et al. (2003) par exemple). Nous pensons toutefois que l'utilisation de tels modèles dans des applications réelles bénéficierait d'une réflexion théorique supplémentaire. Il y a principalement deux difficultés dans cette approche. La première tient à la

dimension de l'espace d'état qui peut être grande. La seconde est liée à la prise en compte du bruit d'observation (ou structurel) non seulement sur la sortie du modèle mais aussi sur les facteurs, et à la propagation de ces incertitudes dans le modèle dynamique à temps discret. Le problème d'incertitude sur les facteurs a été traité en géostatistique (Chilès, 1976 ; Chilès et Delfiner, 1999) et a également fait l'objet d'une thèse récente (Girard, 2004). La possibilité d'estimer des dérivées pourrait être exploitée pour établir des modèles de systèmes à temps continu.

Enfin, mentionnons qu'au moment de l'écriture de ces lignes une thèse que nous contribuons à encadrer a débuté sur la modélisation comportementale des valeurs extrêmes de la sortie d'un système. Le prolongement de nos travaux a donc déjà commencé.

Références

P. Abrahamsen. A review of Gaussian random fields and correlation functions. Technical report, Norwegian Computing Center, 1997.

M. Abramowitz et I. Stegun. *Handbook of Mathematical Functions*. Dover, New York, 1965.

R Adler. *An introduction to Continuity, Extrema, and Related Topics for General Gaussian Processes*, volume 12 of *Lecture Notes*. Institute of Mathematical Statistics, Hayward, CA, 1990.

R. Adler. Gaussian random fields on manifolds. Dans *Progress in Probability*, volume 58 of *Seminar on Stochastic Analysis, Random Fields and Applications IV*, pages 3–19, Basel, 2004. Birkhäuser.

R. J. Adler. *The Geometry of Random Fields*. Wiley, New York, 1981.

C. C. Aggarwal, A. Hinneburg, et D. A. Keim. On the surprinsing behavior of distance metrics in high dimensional spaces. Dans J. Van den Bussche et V. Vianu, éditeurs, *Proc. of Database Theory – ICDT 2001, 8th International Conference Lecture Notes in Computer Science*, volume 1973, pages 420–434, London, 2001. Springer.

M. Aizerman, E. Braverman, et L. Rozonoer. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, 25 :821–837, 1964.

N. Aronszajn. Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, 68 :337–404, 1950.

J.-P. Aubin. *Applied Functional Analysis*. Wiley-Interscience, New york, second edition, 2000.

R. J. Barnes et T. B. Johnson. Positive kriging. Dans G. Verly, M. David, A. G. Journel, et A. Maréchal, éditeurs, *Geostat. for Natural Resources Charact.*, volume 1, pages 231–244, Dordrecht, Holland, 1984. Reidel.

G. Bastin et M. Gevers. Identification and optimal estimation of random fields from scattered pointwise data. *Automatica*, 21(2) :139–155, 1985.

Yu. K. Belyaev. Continuity and Hölder's conditions for sample functions of stationary Gaussian processes. Dans *Proc Fourth Berk. Symp. on Math. Stat. and Probability*, volume 2, pages 23–33, 1961.

Yu. K. Belyaev. Point processes and first passage problems. Dans *Sixth Berkeley Symp. Math. Statist. Prob.*, volume 2, pages 1–17, Berkeley, 1972. Univ. of California Press.

C. Bernard. *Ondelettes et Problèmes mal posés : la mesure du flot optique et l'interpolation irrégulière*. Mémoire de thèse, École Polytechnique, Palaiseau, 1999.

K. S. Beyer, J. Goldstein, R. Ramakrishnan, et U. Shaft. When is "nearest neighbor" meaningful ? Dans C. Beeri et P. Buneman, éditeurs, *Proc. of Database Theory – ICDT 99, 7th International Conference Lecture Notes in Computer Science*, volume 1540, pages 217–235, Jerusalem, 1999. Springer.

P. Billingsley. *Probability and Measure*. Wiley, New York, 3rd edition, 1995.

C. H. Bischof et P. C. Hansen. Structure-preserving and rank-revealing qr-factorizations. *SIAM J. Sci. Statist. Comput.*, 12(6) :1332–1350, 1991.

C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, Oxford, 1995.

D. Blackwell et L. Dubins. Merging of opinions with increasing information. *Annals of Math. Stat.*, 38 :882–886, 1962.

B. E. Boser, I. M. Guyon, et V. N. Vapnik. A training algorithm for optimal margin classifiers. Dans D. Haussler, éditeur, *Proceedings of the Annual Conference on Computational Learning Theory*, pages 144–152, Pittsburgh, PA, 1992. ACM Press.

P. Brockwell et R. Davis. *Time Series : Theory and Methods*. Springer-Verlag, 1987.

T. F. Chan. Rank revealing *qr* factorizations. *Linear Algebra Appl.*, pages 67–82, 1987.

T. F. Chan et P. C. Hansen. Computing truncated singular value decomposition least squares solutions by rank revealing *qr*-factorizations. *SIAM J. Sci. Statist. Comput.*, 11(3) :519–530, 1990.

J.-P. Chilès. How to adapt kriging to non-classical problems : three case studies. Dans M. Guarascio, M. David, et C. Huijbregts, éditeurs, *Advanced Geostatistics in the Mining Industry*, pages 69–89, Dordrecht, 1976. Reidel.

J.-P. Chilès et P. Delfiner. *Geostatistics : Modeling Spatial Uncertainty*. Wiley, New York, 1999.

T. Chonavel. *Statistical Signal Processing, Modeling and Estimation*. Springer-Verlag, London, 2002.

P. Ciarlet. *Introduction à l'analyse numérique matricielle et à l'optimisation*. Dunod, Paris, 1998.

W. G. Cochran. *Sampling Techniques*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, New York-London-Sydney, third edition, 1977.

J.-P. Costa, L. Pronzato, et E. Thierry. Nonlinear filtering by kriging, with application to system inversion. Dans *Proc. ICASSP'99*, volume 3, pages 1313–1316, Phoenix, 1999.

J.-P. Costa, L. Pronzato, et E. Thierry. Nonlinear prediction by kriging, with application to noise cancellation. *Signal Processing*, 80 :553–566, 2000.

R. Courant et D. Hilbert. *Methods of Mathematical Physics*, volume 1. Interscience, 1965.

T. M. Cover et J. A Thomas. *Elements of Information Theory*. Wiley–Interscience, New York, 1991.

H. Cramér et M. R. Leadbetter. *Stationary and Related Stochastic Processes. Sample Function Properties and their Applications*. John Wiley & Sons, Inc., New York-London-Sydney, 1967.

N. Cressie. *Statistics for Spatial Data*. Wiley, New York, 1993.

N. Cressie et D. M. Hawkins. Robust estimation of the variogram, I. *J. Internat. Assoc. Math. Geol.*, 12 :115–125, 1980.

F. Cucker et S. Smale. On the mathematical foundations of learning. *Bull. Amer. Math. Soc.*, 39 (1) :1–49, 2001.

J. G. Daniell. Functions of limited variation in an infinite number of dimensions. *Annals of Mathematics (Series 2)*, 21 :30–38, 1919.

I. Daubechies. *Ten Lectures on Wavelets*. CBMS-NSF Regional Conference Series in Applied Mathematics. SIAM, Philadelphia, 1992.

P. Delfiner et G. Matheron. Les fonctions aléatoires intrinsèques d'ordre k. Technical report, ENSMP, Centre de géostatistique et de morphologie mathématique, Fontainebleau, France, 1980.

P Demartines. *Analyse de données par réseaux de neurones auto-organisés*. Mémoire de thèse, Institut National Polytechnique de Grenoble, France, 1994.

J. L. Doob. *Stochastic Processes*. Wiley, New York, 1953.

M. F. Driscoll. The reproducing kernel Hilbert space structure of the sample paths of a Gaussian process. *Z. Wahrscheinlichkeitstheorie verw. Geb.*, 26 :309–316, 1973.

J. Duchon. Sur l'erreur d'interpolation des fonctions de plusieurs variables par les d^m-splines. *R.A.I.R.O Analyse numérique*, 12(4) :325–334, 1978.

R. M. Dudley. Sample functions of the Gaussian process. *Annals of Probability*, 1(1) :66–103, 1973.

R. Duffin et A. Schaeffer. A class of nonharmonic Fourier series. *Trans. Amer. Math. Soc.*, 72 : 341–366, 1952.

J. Durbin. Efficient fitting of linear models for continuous stationary time series from discrete data. *Bull. Int. Statist. Inst.*, 38 :273–281, 1961.

R. D. Fierro et P. C. Hansen. Low-rank revealing utv decompositions. *Numer. Algorithms*, 15(1) : 37–55, 1997.

S. Fine et K. Scheinberg. Efficient SVM training using low-rank kernel representations. *J. Mach. Learn.*, 2 :243–264, 2001.

L. V. Foster. Rank and null space calculations using matrix decomposition without column interchanges. *Linear Algebra Appl.*, 74 :47–71, 1986.

J. Gao, C. Harris, et S. Gunn. On a class of support vector kernels based on frames in function hilbert space. *Neural computation*, 13(9) :1975–1994, 2001.

I. I. Gikhman et A. V. Skorohod. *The Theory of Stochastic Processes*. Springer-Verlag, Berlin, 1974.

A. Girard. *Approximate Methods for Propagation of Uncertainty with Gaussian Process Models*. Mémoire de thèse, Glasgow University, Glasgow, 2004.

A. Girard, C. E. Rasmussen, J. Quiñonero Candela, et R. Murray-Smith. Multiple-step ahead prediction for non linear dynamic systems – a Gaussian Process treatment with propagation of uncertainty. Dans S. Becker, S. Thrun, et K. Obermayer, éditeurs, *Advances in Neural Information Processing Systems*, volume 15, pages 529–536. MIT press, 2003.

G. Golub et W. Kahan. Calculating the singular values and pseudo-inverse of a matrix. *J. Soc. Indust. Appl. Math. Ser. B Numer. Anal.*, 2 :205–224, 1965.

G. H. Golub et C. F Van Loan. *Matrix Computations*. Johns Hopkins University Press, 1983.

J. Hajek. On linear statistical problems in stochastic processes. *Czech Math. J.*, 87 :404–444, 1962.

M. S. Handcock et J. R. Wallis. An approach to statistical spatial–temporal modelling of meteorological fields. *J. Amer. Statist.*, 89 :368–390, 1994.

E. R. Hansen. *Global Optimization using Interval Analysis*. Marcel Dekker, New York, NY, 1992.

D. A. Harville. Bayesian inference for variance components using only the error contrasts. *Biometrika*, 61 :383–385, 1974.

F. Hirsch et G. Lacombes. *Élements d'analyse fonctionnelle*. Dunod, Paris, 1999.

P. J. Huber. Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35 : 73–101, 1964.

P. J. Huber. *Robust Statistics*. Wiley, New York, 1981.

I. A. Ibragimov et Y. Rozanov. *Gaussian Random Processes*. Springer-Verlag, New York, 1978.

R. H. Jones. Fitting a continuous time autoregression to discrete data. Dans D.F. Findley, éditeur, *Applied Time Series Analysis II*, pages 651–682, New York, 1981. Academic Press.

R. H. Jones et A. Vecchia. Fitting continuous ARMA models to unequally spaces spatial data. *J. of the Amer. Stat. Assoc.*, 88(423) :947–954, 1993.

G. Kimeldorf et G. Wahba. Spline functions and stochastic processes. *Sankhyä : the Indian Journal of Statistics : Series A*, 32(2) :173–180, 1970a.

G. Kimeldorf et G. Wahba. Some results on Tchebycheffian spline functions. *J. Math. Anal. Applic.*, 33 :82–95, 1971.

G. S. Kimeldorf et G. Wahba. A correspondance between bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics*, 41(2) :495–502, 1970b.

P. K. Kitadinis. Statistical estimation of polynomial generalized covariance functions and hydrologic applications. *Water Resoures Research*, 19(909–921), 1983.

D. G. Krige. A statistical approach to some mine valuations and allied problems at the witwatersrand. Master's thesis, University of Witwatersrand, 1951. Unpublished.

S. Kullback. *Information Theory and Statistics*. Dover Publications, New York, 1997 edition, 1968.

H. R. Künsch, E. Agrell, et F. A. Hamprecht. Optimal lattices for interpolation of stationary random fields. Technical Report 119, ETH Zurich, 2003.

J. Kybic, T. Blu, et Unser M. Generalized sampling : A variational approach – part I : theory. *IEEE Trans. Sign. Proc.*, 50(8) :1965–1976, 2002.

G. Lanckriet, N. Cristianini, P. Barlett, L. El Ghaoui, et M. Jordan. Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research*, 5 :27–72, 2004.

V. Langlais. *Estimation sous contraintes d'inégalités*. Mémoire de thèse, E.N.S. des Mines de Paris, 1990.

C. Lantuejoul. *Geostatistical Simulation, Models and Algorithms*. Springer-Verlag, Berlin, 2002.

J. Lefèbvre. *Apport des statistiques aux études des phénomènes de couplage en compatibilité électromagnétique*. Mémoire de thèse, Univ. Paris VI, 1997.

J. Lefèbvre, H. Roussel, D. Lecointe, E. Walter, et W. Tabbara. Prediction from wrong models : the Kriging approach. *IEEE Antennas and propagation Magazine*, 38(4) :35–45, 1996.

W. Light et H. Wayne. On power functions and error estimates for radial basis functions interpolation. *J. Approx. Theory*, 92(2) :245–266, 1998.

M. N. Lukic et J. H. Beder. Stochastic processes with sample paths in reproducing kernel Hilbert spaces. *Trans. Amer. Math. Soc.*, 353(10) :3945–3969, 2001.

W. R. Madych et S. A Nelson. Multivariate interpolation and conditionally positive definite functions. *Approx. Theory and its Applications*, 4(4) :77–89, 1988.

W. R. Madych et S. A Nelson. Multivariate interpolation and conditionally positive definite functions. II. *Math. of computation*, 54(189) :211–230, 1990.

S. Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, San Diego, 2nd edition, 1999.

K. V. Mardia et R. J. Marshall. Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*, 73 :135–146, 1984.

B. Matérn. *Spatial Variation*. Springer-Verlag, 2nd edition, 1986.

G. Matheron. Principles of geostatistics. *Economic Geology*, 58 :1246–1266, 1963.

G. Matheron. La théorie des fonctions aléatoires intrinsèques généralisées. Note Geostatistiques 117, Centre de Géostatistique de l'Ecole des Mines, 1971a.

G. Matheron. *La Théorie des Variables Régionalisées et ses Applications*. Number 5 in Les cahiers du Centre de Morphologie Mathématique. Ecole Nationale Supérieure des Mines De Paris, 1971b. 212 p.

G. Matheron. The intrinsic random functions, and their applications. *Adv. Appl. Prob.*, 5 :439–468, 1973.

G. Matheron. Splines and Kriging : their formal equivalence. Dans Merriam D.F., éditeur, *Down-to-Earth Statistics : Solutions Looking for Geological Problems*, volume 8, pages 77–95. Academic Press, New York, Syracuse Univ. of geology contributions edition, 1981.

R. Mathias et G. W. Stewart. A block *qr* algorithm and the singular value decomposition. *Linear Algebra Appl.*, pages 91–100, 1993.

C. Micchelli. Interpolation of scattered data : distance matrices and conditionally positive definite functions. *Constructive Approximation*, 2 :11–22, 1986.

S. Mika, B. Schölkopf, A. J. Smola, K.-R. Müller, M. Scholz, et G. Rätsch. Kernel PCA and de-noising in feature spaces. Dans M.S. Kearns, S.A. Solla, et Cohn D.A., éditeurs, *Advances in Neural Information Processing Systems*, volume 11, pages 536–542. MIT Press, 1999.

A. Monfort. *Cours de Statistique mathématique*. Economica, Paris, 1997.

M. Morris, T. Mitchell, et D. Ylvisaker. Bayesian design and analysis of computer experiments : Use of derivatives in surface prediction. *Technometrics*, 35(3) :243–255, 1993.

K.-R. Müller, S. Mika, G. Rätsch, K. Tsuda, et B. Schölkopf. An introduction to kernel-based learning algorithms. *IEEE trans. Neur. Net.*, 12(2) :181–202, 2001.

N. Murata, S. Yoshizawa, et S. Amari. Network information criterion – determining the number of hidden units for artificial neural network model. *IEEE Tr. Neural Networks*, 5(6) :865–872, 1994.

F. J. Narcowich, J. D. Ward, et H. Wendland. Refined error estimates for radial basis function interpolation. *Constr. Approx.*, 19(4) :541–564, 2003.

Z. M. Nashed, éditeur. *Generalized Inverses and Applications (Proc. Sem., Math. Res. Center, Univ. Wisconsin, Madison, Wis., 1973)*. Academic Press, New York, 1976.

Z. M. Nashed et G. Wahba. Generalized inverses in reproducing kernel spaces : an approach to regularization of linear operator equations. *SIAM J. Math. Anal.*, 5 :974–987, 1974.

R. M. Neal. Monte carlo implementation of Gaussian process models for bayesian regression and classification. Technical Report 9702, Dept. of Statistics, University of Toronto, University of Toronto, 1997.

O. Nelles, A. Fink, et R. Isermann. Local linear model trees (LOLIMOT) toolbox for nonlinear system identification. Dans *12th IFAC Symposium on System Identification (SYSID)*, Santa Barbara, USA, 2000. IFAC.

C. S. Ong, A. Smola, et R. Williamson. Learning the Kernel with Hyperkernels. *Journal of Machine Learning Research*, submitted.

E. Parzen. An approach to time series analysis. *Ann. Math. Stat.*, 32 :951–989, 1962.

E. Parzen. Probability density functionals and reproducing kernel hilbert spaces. Dans M. Rozenblatt, éditeur, *Proc. Symposium on Time Series Analysis*, pages 155–169, New york, 1963. John Wiley.

E. Parzen. Statistical inference on time seris by rkhs methods. Dans R. Pyke, éditeur, *Proc. 12th Biennial Seminar*, pages 1–37, Montreal, Canada, 1970. Canadian Mathematical Congress.

H. D. Patterson et R. Thompson. Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58(3) :545–554, 1971.

M. J. D. Powell. Radial basis functions for multivariable interpolation : A review. Dans Mason J. C et Cox M. G., éditeurs, *Algorithms for Approximation of Functions and Data*, pages 143–167. Oxford University Press, 1987.

A. Rakotomamonjy et S. Canu. Frame, reproducing kernel, regularization and learning. Technical report, P.S.I INSA de Rouen, 2002.

V. Rannou. *Apport des probabilités et des statistiques à la prédiction des perturbations induites par un téléphone portable sur des câbles.* Mémoire de thèse, Université Paris VI, 2001.

F. Riesz et B. Nagy-Sz. *Leçons d'analyse fonctionnelle.* Gauthier-Villars, quatrième edition, 1965.

C. Robert et G. Casella. *Monte Carlo Statistical Methods.* Springer, New York, 1999.

J. Roll, A. Nazin, et L. Ljung. Non-linear system identification via Direct Weight Optimization. Dans *13th IFAC Symposium on System Identification, SYSID 2003*, pages 1554–1559, Rotterdam, 2003. IFAC.

Yu. A Rozanov. *Stationary Random Processes.* Holden-Day, Inc., San Francisco, 1967.

W. Rudin. *Real and Complex Analysis.* McGraw-Hill, New York, 3rd edition, 1987.

J. Sacks, W. J. Welch, T. J. Mitchell, et H. P. Wynn. Design and analysis of computer experiments. *Statistical Science*, 4(4) :409–435, 1989.

R. Schaback. Native Hilbert spaces for radial basis functions I. *Inter. Series of Numerical Mathematics*, 132 :255–282, 1999.

I. J. Schoenberg. Metric spaces and completely monotone functions. *Ann. Math.*, 39 :811–841, 1938.

I. J. Schoenberg. Spline functions and the problem of graduation. Dans *National Academy of Sciences*, volume 52, pages 947–950, USA, 1964.

B. Schölkopf, R. Herbrich, et A. Smola. A generalized representer theorem. Dans *Proceedings of the Annual Conference on Computational Learning Theory*, pages 416–426, 2001.

B. Schölkopf, A. J. Smola, et K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10 :1299–1319, 1998.

B. Schölkopf et A. Smola. *Learning with Kernels*. MIT Press, Cambridge, 2002.

M. C. Shewry et H. P. Wynn. Maximum entropy sampling. *J. Appl. Statist.*, 14 :165–170, 1987.

M. C. Shewry et H. P. Wynn. Maximum entropy sampling and simulation codes. Dans *12th World Congress on Scientific Computation, IMAC88*, volume 2, pages 517–519, 1988.

A. Smola, T. Friess, et B. Schölkopf. Semiparametric support vector and linear programming machines. Dans M. Kearns, S. Solla, et D. Cohn, éditeurs, *Advances in Neural Information Processing Systems*, 11, pages 585 – 591. MIT Press, Cambridge, 1999.

A. Smola, N. Murata, B. Schölkopf, et K.-R. M"uller. Asymptotically optimal choice of ε-loss function for support vector machines. Dans L. Niklasson, M. Bodén, et T. Ziemke, éditeurs, *Proceedings of the Internat. Conf. on Artific. Neural Net.*, pages 105–110, Berlin, 1998. Springer.

A. J. Smola. *Learning with Kernels*. Mémoire de thèse, Technische Universität Berlin, 1998.

A. J. Smola et B. Schölkopf. Sparse greedy matrix approximation for machine learning. Dans P. Langley, éditeur, *Proceedings of the International Conference on Machine Learning*, pages 911–918, San Francisco, 2000. Morgan Kaufmann.

M. L. Stein. Asymptotically efficient prediction of a random field with misspecified covariance function. *The Annals of Statistics*, 16(1) :55–63, 1988.

M. L. Stein. Bounds on the efficiency of linear predictions using an incorrect covariance function. *The Annals of Statistics*, 18(3) :1116–1138, 1990a.

M. L. Stein. Uniform asymptotic optimality of linear predictions of a random field using an incorrect second-order structure. *The Annals of Statistics*, 18(2) :850–872, 1990b.

M. L. Stein. *Interpolation of Spatial Data : Some Theory for Kriging*. Springer, New York, 1999.

M. L. Stein, Z. Chi, et L. J. Welty. Approximating likelihoods for large spatial data sets. *J. R. Statist. Soc. B*, 66 :275–296, 2004.

G. W. Stewart. Utv decompositions. Dans *Numerical analysis 1993 (Dundee, 1993)*, Pitman Res. Notes Math. Ser., pages 225–236, Harlow, 1994. Longman Sci. Tech.

M. Talagrand. Regularity of Gaussian processes. *Acta Math.*, 159(1-2) :99–149, 1987.

A. N. Tikhonov et V. Y. Arsenin. *Solutions of Ill-posed Problems*. Scripta Series in Mathematics. Winston & Sons, John Wiley & Sons, Washington, D. C, 1977. Translated from the Russian.

H. G. Tucker. *A Graduate Course in Probability*. Academic Press, New York, 1967.

V. Vapnik, S. Golowich, et A. Smola. Support vector method for function approximation, regression estimation, and signal processing. Dans M. Mozer, M. Jordan, et T. Petsche, éditeurs, *Advances in Neural Information Processing Systems 9*, pages 281–287, Cambridge, MA, 1997. MIT Press.

V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, Heidelberg, 1995.

E. Vazquez et E. Walter. Choix d'un noyau pour la régression à vecteurs de support par analyse structurelle : application à la régression multivariable. Dans *Actes électr. GRETSI 2003*, Paris, 2003a.

E. Vazquez et E. Walter. Multi–output support vector regression. Dans *13th IFAC Symposium on System Identification, SYSID 2003*, pages 1820–1825, Rotterdam, 2003b. IFAC.

E. Vazquez et E. Walter. Intrinsic kriging and prior information. *Applied Stochastic Models in Business and Industry*, 21(2) :215–226, 2005.

E. Vazquez et E. Walter. Kriging for indirect measurement, with application to flow measurement. *IEEE Trans. Instr. and Meas.*, soumis en 2003.

A. V. Vecchia. Estimation and model identification for continuous spatial processes. *J.R. Statist. Soc. B*, 50 :297–312, 1988.

H. Wackernagel. *Multivariate Geostatistics*. Springer, Berlin, 1995.

G. Wahba. *Spline Models for Observational Data*, volume 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. SIAM, Philadelphia, 1990.

G. Wahba. Support vector machines, reproducing kernel Hilbert spaces, and randomized GACV. Dans B. Schölkopf, C. J. C. Burges, et A. J. Smola, éditeurs, *Advances in Kernel Methods – Support Vector Learning*, chapitre 6, pages 69–87. MIT Press, Boston, 1998.

E. Walter et L. Pronzato. *Identification of Parametric Models from Experimental Data*. Communication and Control Engineering. Springer, London, 1997.

C. K. I. Williams. Regression with Gaussian processes. Dans S.W. Ellacott, J. C. Mason, et I. J. Anderson, éditeurs, *Mathematics of Neural Networks : Models, Algorithms and Applications*. Kluwer, 1997. Mathematics of Neural Networks and Applications Conference, Oxford, 1995.

C. K. I. Williams et C. E. Rasmussen. Gaussian processes for regression. Dans D. S. Touretzky, M. C. Mozer, et Hasselmo M. E., éditeurs, *Advances in Neural Information Processing Systems*, volume 8. MIT Press, 1996.

Z. Wu et R. Schaback. Local error estimates for radial basis function interpolation of scattered data. *IMA Journ. Numerical Anal.*, 13 :13–27, 1993.

M. I. Yadrenko. *Spectral Theory of Random Fields*. Optimization Software, New York, 1983.

A. M. Yaglom. *An Introduction to the Theory of Stationary Random Functions*. Dover, 2004 edition, 1973.

A. M. Yaglom. *Correlation Theory of Stationary and Related Random Functions I : Basic Results*. Springer Series in Statistics. Springer-Verlag, New york, 1986.

S. J. Yakowitz et F. Szidarovszky. A comparison of kriging with nonparametric regression methods. *J. Multivariate Analysis*, 16 :21–53, 1985.

K. Yosida. *Functional Analysis*. Springer, New-York, reprint of the 6th edition, 1980.

S. C. Zhu, Y. N. Wu, et D. Mumford. Minimax entropy principle and its application to texture modeling. *Neural computation*, 9(8) :1627–1660, 1997.

D. L. Zimmerman. Computationally exploitable structure of covariance matrices and generalized covariance matrices in spatial models. *J. Statist. Comput. Simulation*, 32 :1–15, 1989.

D. L. Zimmerman et M. B. Zimmerman. A Monte Carlo comparison of spatial semivariogram estimators and corresponding ordinary kriging predictors. *Technometrics*, pages 77–91, 1991.

Index

Résumé — *Les* méthodes de prédiction linéaire de processus aléatoires, *ou* krigeage, *et les* méthodes de régression régularisée par une norme d'espace hilbertien à noyau reproduisant *(*splines, *approximation par* fonctions de base radiales, régression à vecteurs de support, *etc.) constituent deux approches fondamentales de modélisation comportementale de systèmes non-linéaires. Les liens mathématiques entre ces deux approches ont été mentionnés à plusieurs reprises dans le passé. Ce travail de thèse présente une synthèse originale de ces liens, puisés dans la littérature de la théorie de l'approximation, de l'apprentissage, des séries chronologiques, de la géostatistique, etc. Fort peu exploités, ces liens n'en restent pas moins fondamentaux puisqu'ils permettent par exemple de comprendre comment formuler le problème de régression régularisée pour l'approximation de fonctions à valeurs vectorielles (cas des systèmes multivariables dits MIMO).*

Dans les deux approches, le choix du noyau est essentiel car il conditionne la qualité des modèles. Les principaux résultats théoriques sont issus de travaux en statistiques. Bien que de type asymptotique, ils ont des conséquences pratiques importantes rappelées et illustrées dans cette étude. Les noyaux considérés habituellement forment une famille restreinte offrant relativement peu de souplesse. Ceci nous a suggéré de développer des méthodes assemblant un noyau à partir d'un grand nombre de noyaux élémentaires. Elles ont permis d'obtenir des résultats satisfaisants notamment sur un problème test classique issu du domaine de la prédiction de séries chronologiques.

Enfin, ce travail s'attache à montrer comment utiliser les méthodes de régression à noyaux à travers la présentation de problèmes réels. Le choix de noyau est abordé en pratique. La prise en compte d'informations disponibles a priori par utilisation du krigeage intrinsèque *(régression semi-régularisée) est illustrée. Finalement, des éléments de planification d'expériences sont discutés.*

Abstract — *Two fundamental types of methods for non-linear black-box modeling are* linear prediction of random processes, *or* Kriging, *and* kernel-based regularized regression *(which includes* Splines, Radial Basis Functions *and* Support Vector Regression *as special cases). Mathematical links between these approaches had been noticed in the past, but this dissertation presents an original synthesis of these links, drawn from the literature on approximation, learning, time series, geostatistics, etc. Though quite confidential up to now, these links are nevertheless essential, for instance in order to understand how regularized regression via kernel-based methods should be formulated in the context of the approximation of vector-valued functions (for multivariable, or MIMO, systems).*

In all of these approaches, the choice of an adequate kernel is crucial since it has a direct impact on the quality of the resulting model. The main theoretical results are provided by statistics. Although mainly asymptotical in nature, these results have important practical consequences which are recalled and illustrated. Classical kernels constitute a relatively restricted family that does not offer much flexibility. This led us to propose methods that build up a kernel by combining a large number of elementary kernels. These methods made it possible to obtain promising results, for example on a classical benchmark of the literature on time-series prediction.

Finally, the question of how kernel-based regression methods can be applied is addressed, via the consideration of real problems. The choice of appropriate kernels for these problems is discussed. We show how prior knowledge can be taken into account using Intrinsic Kriging *(semi-regularized regression). Some contributions to experiment design are also presented.*