Classification bayésienne non supervisée de données fonctionnelles en présence de covariables

Damien Juery

Collège Doctoral du Languedoc-Roussillon

THÈSE

Pour obtenir le grade de

Docteur

Délivré par **UNIVERSITE MONTPELLIER 2**

Préparée au sein de l'école doctorale I2S
Et de l'unité de recherche UMR MISTEA

Spécialité : **Biostatistique**

Présentée par : **Damien JUERY**

Classification bayésienne non supervisée de données fonctionnelles en présence de covariables

Soutenue le 18/12/2014 devant le jury composé de :

M. Christophe ABRAHAM, Professeur, Montpellier SupAgro	Directeur de thèse
Mme Bénédicte FONTEZ, MCF, Montpellier SupAgro	Co-encadrante de thèse
M. Nicolas CHOPIN, Professeur, ENSAE	Rapporteur
M. Pierre DRUILHET, Professeur, Université Blaise Pascal	Rapporteur
M. Jean-Michel MARIN, Professeur, Université Montpellier 2	Examinateur
M. Denys POMMERET, Professeur, Université d'Aix-Marseille	Examinateur

UNIVERSITÉ MONTPELLIER 2
SCIENCES ET TECHNIQUES

A mon épouse Lucie et toute ma vie.

A mes parents Norbert et Geneviève.

A ma sœur Séverine.

Résumé

Un des objectifs les plus importants en classification non supervisée est d'extraire des groupes de similarité depuis un jeu de données. Avec le développement actuel du phénotypage où les données sont recueillies en temps continu, de plus en plus d'utilisateurs ont besoin d'outils capables de classer des courbes.

Le travail présenté dans cette thèse se fonde sur la statistique bayésienne. Plus précisément, nous nous intéressons à la classification bayésienne non supervisée de données fonctionnelles. Les lois *a priori* bayésiennes non paramétriques permettent la construction de modèles flexibles et robustes.

Nous généralisons un modèle de classification (DPM), basé sur le processus de Dirichlet, au cadre fonctionnel. Contrairement aux méthodes actuelles qui utilisent la dimension finie en projetant les courbes dans des bases de fonctions, ou en considérant les courbes aux temps d'observation, la méthode proposée considère les courbes complètes, en dimension infinie. La théorie des espaces de Hilbert à noyau reproduisant (RKHS) nous permet de calculer, en dimension infinie, les densités de probabilité des courbes par rapport à une mesure gaussienne. De la même façon, nous explicitons un calcul de loi *a posteriori*, sachant les courbes complètes et non seulement les valeurs discrétisées. Nous proposons un algorithme qui généralise l'algorithme "Gibbs sampling with auxiliary parameters" de Neal (2000). L'implémentation numérique requiert le calcul de produits scalaires, qui sont approchés à partir de méthodes numériques. Quelques applications sur données réelles et simulées sont également présentées, puis discutées.

En dernier lieu, l'ajout d'une hiérarchie supplémentaire à notre modèle nous permet de pouvoir prendre en compte des covariables fonctionnelles. Nous verrons à cet effet qu'il est possible de définir plusieurs modèles. La méthode algorithmique proposée précédemment est ainsi étendue à chacun de ces nouveaux modèles. Quelques applications sur données simulées sont présentées.

Abstract

One of the major objectives of unsupervised clustering is to find similarity groups in a dataset. With the current development of phenotyping, in which continuous-time data are collected, more and more users require new efficient tools capable of clustering curves.

The work presented in this thesis is based on Bayesian statistics. Specifically, we are interested in unsupervised Bayesian clustering of functional data. Nonparametric Bayesian *priors* allow the construction of flexible and robust models.

We generalize a clustering model (DPM), founded on the Dirichlet process, to the functional framework. Unlike current methods which make use of the finite dimension, either by representing curves as linear combinations of basis functions or by regarding curves as data points, calculations are hereby carried out on complete curves, in the infinite dimension. The reproducing kernel Hilbert space (RKHS) theory allows us to derive, in the infinite dimension, probability density functions of curves with respect to a gaussian measure. In the same way, we make explicit a *posterior* distribution, given complete curves and not only data points. We suggest generalizing the algorithm "Gibbs sampling with auxiliary parameters" by Neal (2000). The numerical implementation requires the calculation of inner products, which are approximated from numerical methods. Some case studies on real and simulated data are also presented, then discussed.

Finally, the addition of an extra hierarchy in our model allows us to take functional covariates into account. For that purpose, we will show that it is possible to define several models. The previous algorithmic method is therefore extended to each of these models. Some case studies on simulated data are presented.

Remerciements

Je voudrais tout d'abord remercier mon directeur de thèse, Christophe ABRAHAM, pour toute son aide. Je me rappellerai toujours lorsqu'il est venu nous donner son cours en Master 2. Je me rappellerai de ses conseils et de tout ce qu'il m'a appris à propos du stage, de l'équipe et de la rigueur du travail scientifique. Par son aide, j'ai pu obtenir ce stage et cette thèse au sein de l'équipe GAMMA. Je le remercie de m'avoir guidé et conseillé dans le travail de recherche, dans les différents enseignements et formations, ainsi que dans la rédaction d'articles.

Un grand merci à ma co-encadrante de thèse, Bénédicte FONTEZ, qui a accepté d'encadrer ce travail. C'est grâce à son écoute, sa confiance, ses idées et ses connaissances que cette thèse a été une expérience enrichissante. Je la remercie également pour son aide dans la recherche et dans la rédaction d'articles.

Je remercie l'ensemble des personnes de l'équipe GAMMA, et en particulier Mathias CHOUET, Nicolas SUTTON-CHARANI, Alexandre MAIRIN, Meïli BARAGATTI, Patrice LOISEL, Nicolas VERZELEN, Véronique SALS-VETTOREL et Maria TROUCHE qui, d'une manière ou d'une autre, ont contribué au bon déroulement de cette thèse.

Je remercie également l'ensemble des personnes avec qui nous avons eu des échanges fructueux. Grâce à ces échanges avec Bernard BARTHES, Michaël CLAIROTTE, Martin ECARNOT, Pierre ROUMET et Sébastien ROUX, nous avons pu améliorer la qualité des résultats de cette thèse.

Mes remerciements vont aussi à Jean-Michel MARIN et Denys POMMERET qui m'ont fait l'honneur de participer au jury de thèse, à mes rapporteurs Nicolas CHOPIN et Pierre DRUILHET pour l'intérêt qu'ils ont porté à ce travail.

J'exprime les derniers de mes remerciements à mes parents Norbert et Geneviève, ma sœur Séverine ainsi que Mathieu qui m'ont toujours soutenu. Je remercie également ma belle-famille pour leurs encouragements. Et enfin merci à tous les amis qui nous ont entourés et en particulier Gilles, Guillaume, Chrystelle, Marcelino, Rouba et Joyce.

Enfin, et non des moindres, je remercie du fond du cœur la personne sans qui je ne suis rien. Je remercie mon épouse pour son soutien, sa présence et son aide la plus précieuse tout au long de ces années. C'est grâce à elle que tout a abouti et que j'ai réussi à tout surmonter. Je lui dédicace ce travail.

Table des matières

Liste des figures

Notations

Abréviation	**Signification**
AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
BOA	Bayesian Output Analysis
CMC	Classification des Moindres Carrés
CRP	Chinese Restaurant Process
DP	Dirichlet Process
DPM	Dirichlet Process Mixture
DPMF	Dirichlet Process Mixture Fonctionnel
DPMFc	DPMF avec covariable
DPMFcs	DPMF avec covariable (simple)
DPMFcr	DPMF ajusté d'une covariable
EM	Expectation-Maximisation
fda	Functional Data Analysis
FDP	Functional Dirichlet Process
GEM	Griffiths, Engen, McCloskey
HDDC	High Dimensional Data Clustering
ICL	Integrated Completed Likelihood
kCFC	k-Centres Functional Clustering
MAP	Maximum A Posteriori
MCMC	Markov Chain Monte Carlo
MV	Modèle Vrai
PG	Processus Gaussien
RI	Rand Index
RKHS	Reproducing Kernel Hilbert Space
SUGS	Sequential Updating and Greedy Search
TCC	Taux de Classification Correcte
VI	Variation of Information

Symbole	**Signification**
argmax	Argument maximum d'une fonction
argmin	Argument minimum d'une fonction
$\mathscr{B}eta(\bullet,\bullet)$	Loi bêta avec deux paramètres de forme
$\mathscr{C}$	Classification d'un jeu de données
C_k	k^e classe d'une classification
$\mathscr{C}_{[0,\mathrm{T}]}$	Espace des fonctions continues sur $[0,\mathrm{T}]$
$\mathscr{D}$	Jeu de données
$\mathrm{C}ov(\bullet,\bullet)$	Covariance entre deux variables aléatoires
$\mathscr{D}ir$	Distribution de Dirichlet
$\mathrm{DP}(\alpha_0,\mathrm{G}_0)$	Processus de Dirichlet de paramètre de concentration α_0 et de distribution de base G_0
δ_θ	Mesure de Dirac au point θ
$\mathbb{E}(\bullet)$	Espérance d'une variable aléatoire
$\mathscr{G}amma(\bullet,\bullet)$	Loi gamma avec paramètre de forme et paramètre d'échelle
$\Gamma(\bullet)$	Fonction Gamma
H(K)	Espace de Hilbert à noyau reproduisant et de noyau K
$\mathrm{I}(\bullet,\bullet)$	Information mutuelle entre deux classifications
$\mathbb{I}_\bullet$	Fonction indicatrice d'un ensemble
j_i	Nombre de temps d'observation pour la i^e observation
$\mathscr{L}(\bullet)$	Loi d'une variable aléatoire ou d'un processus
log	Logarithme népérien
$\mathrm{L}^2([0,\mathrm{T}])$	Espace des fonctions de carré intégrable sur $[0,\mathrm{T}]$
$\mathscr{N}(\bullet,\bullet)$	Loi gaussienne multivariée avec moyenne et matrice de variance-covariance
n	Nombre d'observations
n_k	Nombre d'observations dans la k^e classe
$\mathbb{P}(\bullet)$	Probabilité d'un événement
$p(\bullet)$	Densité de probabilité
$\mathrm{P}_{m,\mathrm{K}}$	Processus gaussien de fonction moyenne m et de fonction de covariance K
$\mathrm{rg}\bullet$	Rang d'une matrice
$\mathrm{RI}(\bullet,\bullet)$	Index de Rand entre deux classifications
S_k	Simplexe de dimension k
t_i	Vecteur des temps d'observation pour la i^e observation
$\mathbb{V}(\bullet)$	Variance d'une variable aléatoire
$\mathrm{VI}(\bullet,\bullet)$	Variation d'information entre deux classifications
Y_i	i^e observation
$\bullet!$	Factorielle d'un entier naturel
$\lvert\bullet\rvert$	Cardinal d'un ensemble
$\bullet^{-i}$	Vecteur privé de sa i^e composante
$\bullet^{-1}$	Inverse d'une matrice
$\bullet^\top$	Transposée d'une matrice
$\bullet \overset{\mathscr{L}}{=} \bullet$	Egalité en loi

$\bullet << \bullet$	Absolue continuité entre deux mesures
$\bullet \propto \bullet$	Quantités proportionnelles
$\bullet \sim \bullet$	Simulation suivant une loi de probabilité
$\bullet \overset{ind}{\sim} \bullet$	Simulations indépendantes suivant une loi de probabilité
$(\bullet, \bullet)_K$	Produit scalaire dans le RKHS H(K)

Introduction

Beaucoup de domaines d'application tels que la météorologie, l'agronomie ou encore l'informatique font appel à des signaux, et donc à des courbes. Cela a conduit, depuis de nombreuses années, au développement d'une nouvelle branche des statistiques : la statistique de données fonctionnelles [102].

Cette thèse se concentre sur la classification bayésienne non supervisée de données fonctionnelles. Si les techniques usuelles multivariées peuvent être employées, des méthodes adaptées aux données fonctionnelles ont été proposées, prenant en compte l'aspect temporel. En statistique non bayésienne, différentes méthodes ont été proposées [10, 47, 54, 55, 68–70, 116, 137]. En statistique bayésienne, les méthodes de classification utilisent couramment le processus de Dirichlet [11, 23, 37, 38, 43, 51, 53, 64, 73, 87, 105], mais peu de travaux ont été réalisés avec les données fonctionnelles [11, 23, 43, 53, 105]. Le processus de Dirichlet est utile en classification car il permet de choisir le nombre de classes de manière automatique. Nous nous intéressons à la généralisation de ces approches bayésiennes aux courbes.

Les approches actuelles de classification de données fonctionnelles peuvent être divisées en deux parties : celles qui traitent les courbes de manière multivariée en les considérant aux temps d'observation [11, 53], et celles qui font usage de la décomposition dans des bases de fonctions, typiquement les fonctions splines et les ondelettes [23, 43, 105]. Par exemple, Ray & Mallick [105] proposent l'utilisation d'une base d'ondelettes, l'inférence étant alors réalisée sur les coefficients de décomposition. Une approche similaire est proposée par Gelfand, Kottas & MacEachern [43], mais dans un objectif de prédiction de données spatiales. Plus récemment, Jackson et al. [53] ont proposé un modèle faisant intervenir les processus gaussiens, calculés aux temps d'observation.

Les méthodes dans lesquelles les courbes sont discrétisées aux temps d'observation font intervenir le déterminant de matrices de variance-covariance de lois normales de très grande dimension, ce qui peut induire une instabilité numérique. De plus, les calculs nécessaires dans ces algorithmes seront d'autant plus longs que le nombre de temps d'observation augmente. Dans le cas de la décomposition dans une base de fonctions, l'utilisateur est soumis au problème du choix de la base et à l'adéquation entre modèle d'approximation et données.

Dans un premier temps, nous travaillons sans aucune covariable. Nous proposons de généraliser le modèle (DPM) (*Dirichlet Process Mixture*) [4], couramment employé pour classer des observations modélisées par des lois normales multivariées, à des observations fonctionnelles

modélisées par des processus gaussiens. Sur le plan théorique, notre approche se distingue des méthodes précédentes en considérant les courbes complètes en dimension infinie. Notre méthodologie requiert le calcul de densités de processus gaussiens, calculées à l'aide de la théorie des espaces de Hilbert à noyau reproduisant (RKHS) [93]. Ainsi, le passage à la dimension finie, inévitable pour toute implémentation est relégué à un problème numérique consistant simplement en un calcul d'intégrale intervenant dans le produit scalaire de deux fonctions. Ce calcul est numériquement facile à réaliser. En contrepartie, notre méthode nécessite d'exprimer des densités de processus gaussiens, de paramètres différents et relativement à une même mesure de référence. L'implémentation numérique est réalisée grâce à une méthode MCMC selon l'algorithme *Gibbs sampling with Auxiliary Parameters* de Neal [87].

Dans un second temps, nous souhaitons prendre en compte des covariables fonctionnelles liées aux courbes. Pour ce faire, nous ajoutons une hiérarchie supplémentaire à notre modèle bayésien. Plusieurs façons de faire sont possibles, conduisant à différents modèles. A notre connaissance, les approches prenant en compte des covariables sont peu nombreuses [27, 36, 116, 137] et le plus souvent, les covariables ne sont pas des fonctions. Nous généralisons les résultats obtenus pour le modèle sans covariable à ce cadre, ainsi que l'algorithme *Gibbs sampling with Auxiliary Parameters* de Neal [87].

Dans le chapitre 1, nous donnons une vue d'ensemble sur les méthodes multivariées de classification non supervisée et nous expliquons le modèle (DPM). Ce chapitre inclut également une présentation générale du processus de Dirichlet, nécessaire dans la compréhension de cette thèse.

Le chapitre 2 traite plus particulièrement des résultats théoriques nécessaires. En particulier, nous souhaitons calculer une loi *a posteriori* dans un modèle où loi *a priori* et vraisemblance sont des processus gaussiens. A cet effet, nous verrons comment la théorie RKHS nous permet de calculer des densités de processus gaussiens.

Dans le chapitre 3, nous adaptons le modèle (DPM) au cadre des données fonctionnelles, ce qui donne naissance au modèle (DPMF). En plus des difficultés propres aux données fonctionnelles, comme la dimension infinie, il nous faut également généraliser un algorithme de simulation. Nous terminons par une analyse des performances de notre méthode sur données simulées et données réelles. Dans ce cadre, nous comparons l'approche proposée à des approches plus habituelles.

Enfin, dans le chapitre 4, nous étudions comment l'ajout d'une hiérarchie au modèle (DPMF) permet la prise en compte de covariables fonctionnelles. La même méthode algorithmique que celle du (DPMF) est utilisée pour l'implémentation numérique, et les performances sont analysées sur données simulées.

CHAPITRE

1 Classification de données multivariées

Résumé

Au préalable, nous présentons le problème de la classification non supervisée. La notion de processus de Dirichlet comme loi de probabilité sur un ensemble de lois de probabilité, est ensuite introduite afin de développer un modèle bayésien utilisé en classification. Le processus de Dirichlet est fonction de deux paramètres : un paramètre de dispersion, qui est un réel positif, et un paramètre de position, qui est une distribution de probabilité. Cette distribution est d'abord supposée discrète, puis continue. Quelques algorithmes d'implémentation sont également présentés. Nous concluons par une étude de simulations permettant d'étudier l'influence des paramètres du modèle sur la classification.

1.1 Classification non supervisée

1.1.1 Objectifs et intérêts

La classification est une technique qui a pour but de trouver des groupes de similarité dans un jeu de données, que l'on appelle classes. Les données d'une même classe sont plus apparentées entre elles qu'avec des données d'autres classes. La classification de données est appliquée à de nombreux domaines biologiques, économiques ou encore informatiques, allant du regroupement de séquences génétiques à la segmentation d'images. Dans tout domaine scientifique ayant recours à des données empiriques, les utilisateurs cherchent à identifier dans les données des groupes aux comportements similaires. L'intérêt pour le partitionnement de données est donc croissant.

Supposons que nous voulions classer n observations $\mathrm{Y}_1, \ldots, \mathrm{Y}_n$ et notons $\mathscr{D}$ ce jeu de données :

$$\mathscr{D} = (\mathrm{Y}_1, \ldots, \mathrm{Y}_n),$$

où $\mathrm{Y}_i = (\mathrm{Y}_{i1}, \ldots, \mathrm{Y}_{im})$ est un vecteur de l'espace $\mathbb{R}^m$ et m est un entier naturel non nul. Une classification de $\mathscr{D}$ en K classes est une partition de $\mathscr{D}$ en K sous-ensembles disjoints $\mathrm{C}_1, \ldots, \mathrm{C}_\mathrm{K}$ non vides. On définit ainsi la classification $\mathscr{C} = \{\mathrm{C}_1, \ldots, \mathrm{C}_\mathrm{K}\}$ et on note respectivement $n_1, \ldots, n_\mathrm{K}$ le nombre d'observations dans chacune des classes $\mathrm{C}_1, \ldots, \mathrm{C}_\mathrm{K}$.

En classification non supervisée, les données sont brutes et on ne dispose d'aucune information préalable sur les classes. En classification supervisée au contraire, l'appartenance des observations aux différentes classes est connue et l'objectif est de construire une règle de classement pour prédire la classe d'une nouvelle observation. Dans cette thèse, seule la classification non supervisée est abordée.

1.1.2 Techniques usuelles

Parmi les méthodes les plus classiques de classification non supervisée, nous pouvons citer la classification hiérarchique [59], l'algorithme des K-means [65, 74], l'utilisation de mélanges gaussiens [97] ou encore la classification spectrale [30, 78, 90, 115]. Ces méthodes étant employées dans la suite de cette thèse, nous les présentons ici brièvement.

Classification hiérarchique

Les algorithmes de classification hiérarchique [20, 59, 117] produisent une hiérarchie de classes, représentée sous la forme d'un dendrogramme. Au bas de la hiérarchie se trouve la partition la plus fine, ne comportant qu'une seule observation par classe, tandis qu'en haut de la hiérarchie se trouve la partition la plus grossière pour laquelle toutes les observations sont dans une même classe.

On distingue les algorithmes ascendants et descendants. Dans le premier cas, il s'agit de regrouper les observations deux à deux en construisant le dendrogramme, jusqu'à ne former qu'une seule classe. Dans le second cas, le dendrogramme est construit à partir de la partition constituée d'une seule classe. Dans les deux cas, la classification hiérarchique suppose de savoir calculer, à chaque étape, une distance entre classes, appelée lien, ainsi qu'une mesure de dissimilarité entre observations.

Une fois ces distances choisies, le principe d'une classification ascendante hiérarchique est simple. Une partition de n classes contenant chacune une seule observation est formée. L'algorithme commence par calculer la matrice de dissimilarité dont l'élément générique d_{ij} est la distance entre les observations Y_i et Y_j. L'algorithme forme alors une classe par agrégation des deux observations les plus proches. La distance entre cette nouvelle classe et les autres classes est déterminée par le lien. Ce processus est alors réitéré jusqu'à l'obtention d'une seule classe. Une méthode algorithmique similaire s'applique dans le cas d'une classification descendante hiérarchique.

Ces algorithmes présentent l'avantage de pouvoir choisir les distances en fonction de la nature des données. En contrepartie, il est facile de vérifier qu'une classification hiérarchique est très sensible à ce choix. De plus, les algorithmes de classification hiérarchique ne sont pas adaptés aux jeux de données de grande dimension, en raison de leur complexité élevée, et si l'on souhaite rajouter une observation au jeu de données à classer, il est nécessaire de répéter l'algorithme depuis le début.

Algorithme des K-means

Initialement proposé en 1957 par Lloyd [65], puis repris en 1967 par MacQueen [74], l'algorithme des K-means est une méthode itérative qui, quelle que soit la configuration initiale, converge vers une solution. Elle consiste à regrouper les observations par minimisation de la distance entre chaque observation et le centre de sa classe, appelé centroïde.

Etant donné un nombre de classes K, la première étape consiste à choisir aléatoirement K observations comme centroïdes $\mu_1, \ldots, \mu_K \in \mathbb{R}^m$, où $\mu_k = (\mu_{k1}, \ldots, \mu_{km})$. Puis, chaque observation Y_i est affectée au centroïde dont elle est la plus proche en ce qui concerne la distance euclidienne :

$$d(Y_i, \mu_k) = \sqrt{\sum_{j=1}^{m} (Y_{ij} - \mu_{kj})^2}.$$

Chaque centroïde est alors recalculé à l'aide de la formule suivante :

$$\mu_k = \frac{1}{n_k} \sum_{Y \in C_k} Y.$$

Ces étapes sont ainsi répétées jusqu'à ce qu'un critère de convergence soit atteint. En pratique, l'algorithme est répété jusqu'à ce que les affectations des observations aux classes ne changent plus. Le partitionnement $\mathscr{C} = \{C_1, \ldots, C_K\}$ obtenu à partir de l'algorithme des K-means dépend uniquement des centroïdes obtenus et il est construit de sorte à minimiser la fonction suivante :

$$(C_1, \ldots, C_K) \mapsto \sum_{k=1}^{K} \sum_{Y \in C_k} d(Y, \mu_k)^2.$$

Notons que minimiser cette fonction directement est un problème NP-difficile [33]. Ainsi, un algorithme des K-means ne peut converger que vers un minimum local, même si de récentes études [79] ont montré qu'avec grande probabilité, un K-means converge vers un minimum global lorsque les classes sont bien séparées.

L'algorithme des K-means est facile à implémenter et s'exécute rapidement, ce qui en fait un algorithme de classification non supervisée très populaire. Cependant, il est nécessaire de choisir le nombre de classes K. Même si de nombreuses méthodes existent pour le calculer [126], il n'est pas possible de s'abstenir de ce choix. Enfin, l'algorithme des K-means peut peiner à retrouver les bonnes classes lorsque celles-ci sont imbriquées les unes dans les autres. Une nouvelle méthode remplaçant la distance euclidienne $d(\bullet, \bullet)$ par la distance de Mahalanobis a été proposée par Mao et al. [75], mais au prix d'un temps de calcul élevé.

Utilisation de mélanges gaussiens

Un modèle de mélange [97] est une loi dont la densité est une combinaison convexe de plusieurs densités de probabilité. Considérons une famille de réels positifs $\pi_1, \ldots, \pi_K$ telle que $\sum_{k=1}^{K} \pi_k = 1$ et une famille de densités $p(\bullet; \theta_k)$ paramétrées par θ_k. La densité d'une loi de mélange à K composantes est définie par :

$$p(Y; \pi_1, \ldots, \pi_K, \theta_1, \ldots, \theta_K) = \sum_{k=1}^{K} \pi_k p(Y; \theta_k).$$

Chaque composante θ_k du mélange caractérise une classe et π_k est la proportion de la k^e classe. Les modèles de mélange gaussiens sont couramment employés et consistent à choisir pour densités de mélanges $p(\bullet; \theta_k)$ des densités de lois normales multivariées.

En pratique, l'étape la plus délicate est l'estimation des paramètres θ_k et π_k. En considérant un échantillon $Y_1, \ldots, Y_n$ suivant le modèle de mélange gaussien, il est courant de chercher à

déterminer les paramètres qui maximisent la log-vraisemblance :

$$\mathrm{L}(\mathrm{Y}_1,\ldots,\mathrm{Y}_n;\pi_1,\ldots,\pi_\mathrm{K},\theta_1,\ldots,\theta_\mathrm{K}) = \sum_{i=1}^{n} \log\ p(\mathrm{Y}_i;\pi_1,\ldots,\pi_\mathrm{K},\theta_1,\ldots,\theta_\mathrm{K}).$$

Les paramètres sont estimés numériquement par approximations successives à l'aide d'un algorithme EM [29], abréviation de *Expectation-Maximization* en anglais. Différents critères permettant de choisir le nombre de composantes K du mélange existent, tels que les critères AIC (*Akaike Information Criterion*) [2], BIC (*Bayesian Information Criterion*) [110] ou encore ICL (*Integrated Classification Likelihood*) [15]. La valeur de K qui minimise le critère de sélection choisi est retenue. Il est également possible de faire appel à la validation croisée pour choisir K.

Une fois le modèle ajusté, la classification est généralement obtenue en affectant l'observation Y_i à la classe C_k dont le degré d'appartenance est le plus élevé. Ce degré d'appartenance est défini par :

$$\frac{\pi_k p(\mathrm{Y}_i;\theta_k)}{\sum_{k=1}^{\mathrm{K}} \pi_k p(\mathrm{Y}_i;\theta_k)}.$$

La flexibilité des modèles de mélange gaussiens permet de modéliser un grand nombre de phénomènes. Cependant, leur utilisation requiert de déterminer le nombre de composantes K du mélange et l'utilisation d'un algorithme EM peut parfois se révéler laborieuse.

Classification spectrale

La classification spectrale [30] est une technique de classification simple à implémenter et qui consiste à transformer un problème de classification en un problème de théorie des graphes. Dans la suite de cette thèse, cette technique ne sera pas exploitée, cependant le principe nous ayant semblé judicieux dans certains cas, nous avons pris l'initiative d'en donner ici une brève introduction.

Chaque observation Y_i est perçue comme le sommet d'un graphe, où la longueur des arêtes correspond à la proximité entre les sommets. Le choix de cette mesure de proximité, que l'on appelle aussi similarité, est laissé à l'initiative de l'utilisateur. La plus courante est la fonction de similarité gaussienne :

$$\mathrm{S}(\mathrm{Y}_i,\mathrm{Y}_j) = e^{-\frac{\sum_{l=1}^{m}(\mathrm{Y}_{il}-\mathrm{Y}_{jl})^2}{2\sigma^2}},$$

où σ est un réel strictement positif. Il existe également des méthodes d'estimation de la similarité [78].

Les sommets sont alors reliés entre eux afin de former un graphe pondéré sur les arêtes. Pour cela, différentes méthodes existent, la plus courante étant de relier les sommets pour lesquels la similarité est supérieure à un seuil fixé ϵ.

En notant alors S la matrice de similarité $\mathrm{S} = \left(\mathrm{S}(\mathrm{Y}_i,\mathrm{Y}_j)\right)_{1\leq i,j\leq n}$, l'outil principal à présent est la matrice de Laplacien de graphe. Cette matrice peut être définie de plusieurs façons [89, 115] mais la matrice Laplacienne L = S est le choix le plus courant dans la littérature.

Supposons que l'on souhaite obtenir une classification en K classes. Le principe de la classification spectrale consiste à calculer les K premiers vecteurs propres $v_1,\ldots,v_\mathrm{K} \in \mathbb{R}^n$ de la matrice de

Laplacien de graphe L. Une nouvelle matrice V est construite à partir de la donnée des v_k :

$$V = \begin{pmatrix} v_{11} & \dots & v_{K1} \\ \vdots & & \vdots \\ v_{1n} & \dots & v_{Kn} \end{pmatrix}.$$

Chaque ligne de la matrice V peut alors être interprétée comme la donnée d'une nouvelle observation $\widetilde{Y}_i = (v_{1i}, \dots, v_{Ki}) \in \mathbb{R}^K$. Le jeu de données initial $(Y_1, \dots, Y_n)$ est ainsi transformé en un nouveau jeu de données $\left(\widetilde{Y_1}, \dots, \widetilde{Y_n}\right)$ dont les observations sont classées, par exemple, à l'aide de l'algorithme des K-means.

En pratique, la classification spectrale est très sensible à la fonction de similarité. De même que pour d'autres techniques de classification, il est nécessaire de déterminer le nombre de classes K. Cependant, les observations sont projetées dans un espace de dimension inférieure où très souvent, les classes sont plus facilement identifiables. Notons qu'il existe de nombreuses techniques dérivées de cette méthode très générale.

Choix d'une technique de classification non supervisée

Chaque technique possède ses propres avantages et inconvénients, et choisir la "meilleure" méthode constitue un véritable défi. Dans toutes ces techniques, il est nécessaire de déterminer le nombre de classes. A cause de ces difficultés, une pratique courante consiste à répéter plusieurs fois un algorithme pour différentes valeurs de paramètres et à comparer les résultats obtenus. Retenons toutefois qu'une classification est fortement dépendante de l'application et que le choix d'une technique particulière se fait de façon subjective.

1.1.3 Comparaison et validation de classifications

La qualité d'une classification est difficile à évaluer, car dans la pratique on ne connaît pas les vraies classes. Une solution simple est de construire un jeu de données simulées pour lequel on connaît la vraie classification $\mathscr{C}^\star$. Imaginons alors que l'on dispose de deux classifications $\mathscr{C}$ et $\mathscr{C}'$ de ce même jeu de données. Il est légitime de se poser la question suivante : quelle classification est "la plus proche" de $\mathscr{C}^\star$?

Répondre à cette question nécessite de définir une distance entre classifications. Beaucoup de distances ont été proposées dans la littérature [16, 77, 80, 86, 90, 103, 125]. Les avantages et inconvénients de chacune dépendent fortement du jeu de données considéré.

Notons $\mathscr{C} = \{C_1, \dots, C_K\}$ et $\mathscr{C}' = \{C'_1, \dots, C'_{K'}\}$. Dans cette notation, K et K′ désignent respectivement le nombre de classes pour chacune des deux classifications et il est possible d'avoir $K \neq K'$. Les notations de base utilisées sont données dans la figure 1.1. L'ensemble des valeurs $(n_{kk'})_{1\leq k\leq K, 1\leq k'\leq K'}$ est généralement regroupé sous forme de matrice que l'on appelle matrice de confusion.

Il est courant de distinguer trois types de critères. Le premier est basé sur les paires d'observations $\left(Y_i, Y_j\right)$. A l'instar de nombreux auteurs, notons :
- N_{11} : nombre de paires d'observations classées ensemble dans $\mathscr{C}$ et $\mathscr{C}'$,
- N_{00} : nombre de paires d'observations pas dans une même classe ni dans $\mathscr{C}$ ni dans $\mathscr{C}'$,
- N_{10} : nombre de paires d'observations classées ensemble dans $\mathscr{C}$ mais pas dans $\mathscr{C}'$,
- N_{01} : nombre de paires d'observations classées ensemble dans $\mathscr{C}'$ mais pas dans $\mathscr{C}$.

$\mathscr{C}'/\mathscr{C}$	C_1	C_2	$\cdots$	C_K	
C'_1	n_{11}	n_{21}		n_{K1}	n'_1
C'_2	n_{12}	n_{22}		n_{K2}	n'_2
$\vdots$					
$C'_{K'}$	$n_{1K'}$	$n_{2K'}$		$n_{KK'}$	$n'_{K'}$
	n_1	n_2		n_K	n

FIGURE 1.1 – $\mathscr{C} = \{C_1, \ldots, C_K\}$ *et* $\mathscr{C}' = \{C'_1, \ldots, C'_{K'}\}$ *sont deux classifications d'un même jeu de données.* n_k *et* $n'_{k'}$ *désignent respectivement le nombre d'observations dans* C_k *et* $C'_{k'}$*. Enfin,* $n_{kk'} = |C_k \cap C'_{k'}|$ *correspond au nombre d'observations à la fois dans les classes* C_k *et* $C'_{k'}$*.*

Historiquement, Rand [103] a proposé en 1971 un indice appelé indice de Rand, ou *Rand Index* en anglais, permettant de comparer deux classifications :

$$\mathrm{RI}(\mathscr{C}, \mathscr{C}') = \frac{N_{00} + N_{11}}{\binom{n}{2}}.$$

$\binom{n}{2}$ correspond au nombre de combinaisons possibles de paires d'observations. Les valeurs de l'indice de Rand sont comprises entre 0 et 1, où 1 signifie que les deux classifications sont identiques. A titre d'exemple, choisissons $n = 3$ et les deux classifications $\mathscr{C} = \{\{Y_1, Y_2\}, \{Y_3\}\}$ et $\mathscr{C}' = \{\{Y_1, Y_3\}, \{Y_2\}\}$; l'indice de Rand vaut 0.33 dans ce cas. Il s'agit du critère le plus souvent rencontré, mais il ne s'agit pas d'une métrique mathématique. Pour cette raison, Hubert & Arabie [48] ont proposé en 1985 une version ajustée de l'indice de Rand. D'autres auteurs ont ensuite proposé leur propre version ; notons le fameux indice de Fowlkes-Mallows [40] ainsi que sa version ajustée, l'indice de Jaccard [52] et la distance de Mirkin [84]. Bien souvent, tous ces critères sont des modifications directes de l'indice de Rand et pour cela, nous ne prendrons pas la peine de les détailler ici.

Le second type de critères pour la comparaison de classifications est basé directement sur la matrice de confusion (figure 1.1). Si l'on suppose $\mathscr{C}'$ une classification de référence à laquelle on souhaite comparer $\mathscr{C}$, de nombreux auteurs dans la littérature proposent de calculer le taux de classification correcte, abrégé TCC, défini comme le rapport entre le nombre maximal possible d'observations bien classées (par rapport à $\mathscr{C}'$) sur le nombre total d'observations. Formellement, il s'écrit :

$$\mathrm{TCC} = \begin{cases} \frac{1}{n} \times \sum_{k'=1}^{K'} \max_{1 \leq k \leq K} n_{kk'}, & K' \leq K, \\ \\ \frac{1}{n} \times \sum_{k=1}^{K} \max_{1 \leq k' \leq K'} n_{kk'}, & K' > K. \end{cases}$$

Le troisième et dernier type de critères fait appel à la théorie de l'information, l'idée étant de mesurer la quantité d'information que l'on dispose dans chacune des classifications, ainsi que la quantité d'information apportée par une classification à l'autre. Meilă [80] introduit en 2007 la variation d'information. Supposons que l'on choisisse au hasard une observation de $\mathscr{D}$: quelle incertitude y a-t-il sur sa classe dans $\mathscr{C}$? Si chaque observation a la même probabilité d'être tirée au sort, la probabilité pour cette observation d'appartenir à la classe C_k peut se noter $P(k) = \frac{n_k}{n}$.

On définit par ce biais une variable aléatoire discrète de support $\{1, \ldots, K\}$, qui représente la classe de l'observation choisie. En choisissant alors au hasard une observation de $\mathscr{D}$, l'incertitude sur

sa classe correspond à l'entropie de cette variable aléatoire et vaut :

$$\mathrm{H}(\mathscr{C}) = -\sum_{k=1}^{\mathrm{K}} \mathrm{P}(k)\log(\mathrm{P}(k)).$$

En notant de même $\mathrm{P}(k,k') = \frac{n_{kk'}}{n}$ la probabilité pour qu'une observation soit à la fois dans les classes C_k et $\mathrm{C}'_{k'}$, et $\mathrm{P}'(k) = \frac{n'_{k'}}{n}$ la probabilité pour une observation d'être dans la classe $\mathrm{C}'_{k'}$, l'information mutuelle entre $\mathscr{C}$ et $\mathscr{C}'$ se définit de la manière suivante :

$$\mathrm{I}(\mathscr{C},\mathscr{C}') = \sum_{k=1}^{\mathrm{K}}\sum_{k'=1}^{\mathrm{K}'} \mathrm{P}(k,k')\log\left(\frac{\mathrm{P}(k,k')}{\mathrm{P}(k)\mathrm{P}'(k')}\right).$$

L'information mutuelle mesure la corrélation entre la répartition des observations dans $\mathscr{C}$ et la répartition des observations dans $\mathscr{C}'$. Il existe également des versions normalisées de l'information mutuelle, dues entre autres à Strehl & Ghosh [122] ou encore Vinh et al. [129], mais nous ne les évoquerons pas ici. Pour conclure, Meilă définit la variation d'information entre $\mathscr{C}$ et $\mathscr{C}'$:

$$\mathrm{VI}(\mathscr{C},\mathscr{C}') = \mathrm{H}(\mathscr{C}) + \mathrm{H}(\mathscr{C}') - 2\mathrm{I}(\mathscr{C},\mathscr{C}').$$

La variation d'information est une métrique mathématique et en particulier, $\mathrm{VI}(\mathscr{C},\mathscr{C}') = 0$ si, et seulement si, $\mathscr{C} = \mathscr{C}'$. Intuitivement, dans le tableau de la figure 1.1, plus il y a de zéros à l'extérieur de la diagonale, plus les classifications sont proches.

De la même façon que nous ne pouvons parler de "meilleure" technique de classification, nous ne pouvons dire qu'il existe de "meilleur" critère de comparaison. En classification non supervisée, comme l'on ne dispose presque jamais de la vraie classification, il est assez délicat d'utiliser de tels critères. Sur un jeu de données réelles, la classification obtenue par un algorithme est le plus souvent jugée par l'utilisateur lui-même. Un algorithme produira de bons résultats dans la mesure où il est possible d'en extraire une information utile et pertinente sur les données. Sur un jeu de données simulées, lorsque l'on connaît les vraies classes, la qualité d'une classification réside dans son aptitude à retrouver la vraie classification.

1.1.4 Bilan

Après avoir rappelé quelques notions élémentaires de classification non supervisée, nous proposons dans la suite de cette thèse d'étudier la classification au travers des modèles bayésiens et en particulier au travers des processus de Dirichlet.

Certaines définitions ou propriétés usuelles de statistique bayésienne seront supposées connues ; le lecteur, s'il le souhaite, peut se référer à l'annexe A.1 pour plus de précisions sur ces résultats.

1.2 Présentation et choix d'un modèle bayésien

1.2.1 Les processus de Dirichlet

Les processus de Dirichlet sont très utiles en classification car ils permettent de choisir le nombre de classes de façon automatique. Formellement, le processus de Dirichlet est une distribution définie sur un espace de distributions. Cela implique qu'un tirage suivant un processus

de Dirichlet produit une loi de probabilité. Le processus de Dirichlet est donc de plus en plus utilisé en tant que loi *a priori* dans les modèles non paramétriques bayésiens. Définissons-le et étudions-en quelques propriétés fondamentales.

Dans toute la suite, $\mathscr{D}ir$ désignera la distribution de Dirichlet ; le lecteur, s'il le souhaite, peut se référer à l'annexe A.2 pour plus de précisions concernant cette distribution et ses propriétés usuelles.

Définition 1.1 (Processus de Dirichlet)
Soient G_0 *une distribution sur un ensemble* Θ *et* α_0 *un réel strictement positif. On dit que la mesure aléatoire* G *suit un processus de Dirichlet de distribution de base* G_0 *et de paramètre de concentration* α_0*, et on écrit* $G \sim DP(\alpha_0, G_0)$*, si l'on a :*

$$\Big(G(A_1), \ldots, G(A_r)\Big) \sim \mathscr{D}ir\Big(\alpha_0 G_0(A_1), \ldots, \alpha_0 G_0(A_r)\Big), \tag{1.2.1}$$

pour toute partition mesurable $\{A_1, \ldots, A_r\}$ *de* Θ.

Remarque 1.1
Cette définition implique que G *a le même support* Θ *que* G_0.

Dans cette définition, les rôles de G_0 et de α_0 n'apparaissent pas clairement, mais ce qui suit nous en donne l'intuition. Il est en effet facile de déduire de l'expression (1.2.1) que l'on a pour tout sous-ensemble mesurable A de Θ :

$$\mathbb{E}\Big(G(A)\Big) = G_0(A),$$

$$\mathbb{V}\Big(G(A)\Big) = \frac{G_0(A)(1 - G_0(A))}{\alpha_0 + 1}.$$

D'une part, on s'aperçoit que G_0 agit comme la moyenne du processus du Dirichlet et d'autre part que le paramètre de concentration α_0 est relié à l'inverse d'une variance. α_0 est donc un paramètre d'échelle du processus ; plus celui-ci est grand, plus le processus de Dirichlet se concentre autour de sa moyenne. Nous verrons par la suite comment ces paramètres agissent plus spécifiquement dans une optique de classification.

Dans toute la suite, nous serons parfois amenés à distinguer le cas où G_0 est une distribution discrète de celui où G_0 est une distribution continue. Sauf mention explicite, les résultats énoncés seront valables dans les deux cas.

Représentation de Sethuraman

Si les processus de Dirichlet sont autant utilisés en classification, c'est aussi parce qu'un tirage suivant un processus de Dirichlet est une probabilité discrète. En effet, Rolin [106] en 1993 et Sethuraman [112] en 1994 donnent une définition constructive de ces processus à l'aide de masses de Dirac. Enonçons d'abord la loi du bâton cassé, ou encore *stick-breaking* en anglais.

Définition 1.2 (Loi du bâton cassé)
On dit qu'une suite de nombres $(\pi_k)_{k\geq 1}$ *suit la loi du bâton cassé de paramètre* α_0*, et on note* $\pi \sim GEM(\alpha_0)$*, du nom de ses auteurs Griffiths, Engen et McCloskey [99], si elle est définie comme suit :*

$$\pi'_k \overset{ind}{\sim} \mathscr{B}eta(1, \alpha_0), k \geq 1,$$

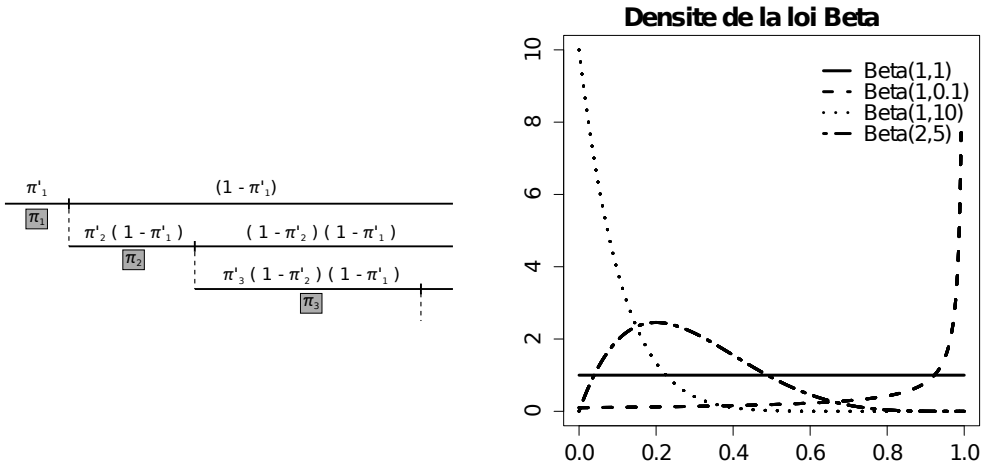

FIGURE 1.2 – *Illustration de la loi du bâton cassé. A gauche : une suite de nombres générée suivant la loi du bâton cassé. A droite : quelques densités de lois de probabilité Beta.*

et

$$\pi_1 = \pi'_1,$$
$$\pi_k = \pi'_k \prod_{j=1}^{k-1} (1-\pi'_j), k \geq 2.$$

D'un point de vue intuitif, on peut comprendre la construction de π comme suit ; on choisit un bâton de longueur 1 que l'on casse au point π'_1. π_1 désigne la longueur de la tige que l'on vient de casser. On procède alors de même afin d'obtenir les nombres π_2, π_3, ainsi de suite, ce qui est illustré sur la figure 1.2.

Proposition 1.1 (Sethuraman, 1994)
Soient une suite de nombres $\pi \sim \mathrm{GEM}(\alpha_0)$ *et une suite de variables aléatoires* $(\phi_k)_{k\geq 1}$ *indépendantes et identiquement distribuées suivant* G_0. *Supposons que les suites* $(\phi_k)_{k\geq 1}$ *et* $(\pi_k)_{k\geq 1}$ *sont indépendantes.*
Alors, la mesure aléatoire G *suit le processus de Dirichlet* $\mathrm{DP}(\alpha_0, \mathrm{G}_0)$ *si, et seulement si,* G *s'écrit sous la forme* $\mathrm{G} = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}$, *où* δ_ϕ *désigne la masse de Dirac au point* ϕ.

Cette représentation permet de voir qu'une probabilité G engendrée par un processus de Dirichlet est une loi discrète et à support infini dénombrable. Si l'on simule à nouveau suivant G, il existe donc une probabilité non nulle d'obtenir des valeurs ex-æquo.

Remarquons de plus qu'à partir d'un certain entier $k \geq 1$, les nombres π_k deviennent négligeables car très proches de 0. Les poids les plus conséquents correspondent donc aux faibles valeurs de k, c'est-à-dire à π_1, π_2, etc. Ainsi, seul un faible nombre d'atomes sont potentiellement représentés, les autres poids étant négligeables. Il est donc courant d'approcher la mesure G par la mesure $\mathrm{G_K}$ définie par :

$$\mathrm{G_K} = \frac{1}{\sum_{k=1}^{\mathrm{K}} \pi_k} \sum_{k=1}^{\mathrm{K}} \pi_k \delta_{\phi_k}.$$

La représentation de Sethuraman nécessitant une infinité de tirages, certains auteurs [50, 51, 63]

préfèrent utiliser cette représentation approchée plus simple. Dans cette thèse, il n'est pas fait usage de cette approximation.

1.2.2 Mélange suivant un processus de Dirichlet

Supposons à présent que l'on souhaite classer un nombre n de paramètres θ_i. Un modèle couramment employé est le suivant :

$$\begin{cases} \theta_i|\mathrm{G} & \overset{ind}{\sim} & \mathrm{G}, \\ \mathrm{G} & \sim & \mathrm{DP}(\alpha_0, \mathrm{G}_0), \end{cases} \tag{1.2.2}$$

où $\mathrm{DP}(\alpha_0, \mathrm{G}_0)$ désigne la loi du processus de Dirichlet de paramètre de concentration $\alpha_0 > 0$ et de distribution de base G_0. La notation ind signifie que les variables sont indépendantes entre elles.

Urne de Pólya

Dans une optique de classification, le modèle (1.2.2) est adéquat grâce à la représentation par urne de Pólya de la loi jointe du vecteur $(\theta_1, \ldots, \theta_n)$ [8] :
– $\theta_1 \sim \mathrm{G}_0$,
– θ_2 est égal à θ_1 avec probabilité $\frac{1}{\alpha_0+1}$, et tiré suivant G_0 avec probabilité $\frac{\alpha_0}{\alpha_0+1}$,
– θ_3 est égal à θ_j, $1 \leq j \leq 2$, avec probabilité $\frac{1}{\alpha_0+2}$, et tiré suivant G_0 avec probabilité $\frac{\alpha_0}{\alpha_0+2}$,
– ainsi de suite.
Plus formellement, ce résultat a été publié en 1973 par Blackwell & MacQueen [8].

Proposition 1.2 (Blackwell & MacQueen, 1973)
Soient des variables $\theta_1, \ldots, \theta_n|\mathrm{G} \overset{ind}{\sim} \mathrm{G}$ *avec* $\mathrm{G} \sim \mathrm{DP}(\alpha_0, \mathrm{G}_0)$. *Pour chaque entier* $i \in \{1, \ldots, n\}$, *désignons par* θ^{-i} *le vecteur* $(\theta_1, \ldots, \theta_n)$ *privé de sa* i^e *composante. Intégrer* G *mène alors à la loi conditionnelle suivante :*

$$\mathrm{P}(d\theta_i|\theta^{-i}) = \sum_{j \neq i} \frac{1}{\alpha_0 + n - 1} \delta_{\theta_j}(d\theta_i) + \frac{\alpha_0}{\alpha_0 + n - 1} \mathrm{G}_0(d\theta_i), \tag{1.2.3}$$

De plus, la suite $(\theta_1, \ldots, \theta_n)$ *est échangeable, c'est-à-dire que la loi du n-uplet est invariante pour toute permutation des indices. Dit autrement, on a pour toute permutation* σ *de* $\{1, \ldots, n\}$ *l'égalité en loi* $\left(\theta_{\sigma(1)}, \ldots, \theta_{\sigma(n)}\right) \overset{\mathscr{L}}{=} (\theta_1, \ldots, \theta_n)$.

D'après cette représentation, la probabilité pour que θ_j soit égal à une ancienne valeur $\theta_1, \ldots, \theta_{j-1}$ ne dépend pas des valeurs des $\theta_1, \ldots, \theta_{j-1}$. Définissons alors le vecteur $(c_1, \ldots, c_n)$ de loi jointe :
– $c_1 = 1$,
– c_2 est égal à c_1 avec probabilité $\frac{1}{\alpha_0+1}$, et est égal à $k_1 + 1 (= 2)$ avec probabilité $\frac{\alpha_0}{\alpha_0+1}$,
– c_3 est égal à c_j, $1 \leq j \leq 2$, avec probabilité $\frac{1}{\alpha_0+2}$, et est égal à $k_2 + 1$ avec probabilité $\frac{\alpha_0}{\alpha_0+2}$,
– ainsi de suite,
où k_i désigne le nombre de valeurs distinctes dans le vecteur $(c_1, \ldots, c_i)$. Par la suite, notons $\mathrm{CRP}(\alpha_0)$ la loi du vecteur $(c_1, \ldots, c_n)$. MacEachern & Müller [73] ont montré en 1998 que le vecteur $(c_1, \ldots, c_n)$ est échangeable.

Dans le modèle (1.2.2) et le modèle :

$$\begin{cases} (c_1,\dots,c_n) & \sim & \mathrm{CRP}(\alpha_0), \\ \phi_c & \overset{ind}{\sim} & \mathrm{G}_0, \\ \theta_i & = & \phi_{c_i}, \end{cases} \tag{1.2.4}$$

on peut vérifier facilement que la loi jointe du vecteur $(\theta_1,\dots,\theta_n)$ est la même. Dans ce dernier modèle, les ϕ_c correspondent aux valeurs distinctes des θ_i, et $(c_1,\dots,c_n)$ est le vecteur d'assignation des classes pour chaque observation.

Métaphore du restaurant chinois

La loi CRP (*Chinese Restaurant Process* en anglais) est souvent expliquée à partir de la métaphore du restaurant chinois [3]. Dans cette métaphore, on considère un restaurant avec une infinité de tables, chacune pouvant accueillir une infinité de clients et servant le même plat à tous les clients qui y sont installés. De plus, les tables sont circulaires, de manière à ce que l'ordre des clients installés importe peu. Un premier client arrive alors dans le restaurant, s'assied à la table 1 ($c_1 = 1$) et commande un plat θ_1. Lorsque le deuxième client arrive, il peut soit s'asseoir à la table du premier client ($c_2 = 1$) et commander le même plat ($\theta_2 = \theta_1$), soit s'installer à la table suivante ($c_2 = 2$) et commander un nouveau plat θ_2. De manière générale, le i^e client s'assied soit à une table déjà occupée, avec probabilité proportionnelle au nombre de clients qui y sont installés, soit à une table vide, avec probabilité proportionnelle à α_0. La figure 1.3 illustre ce procédé.

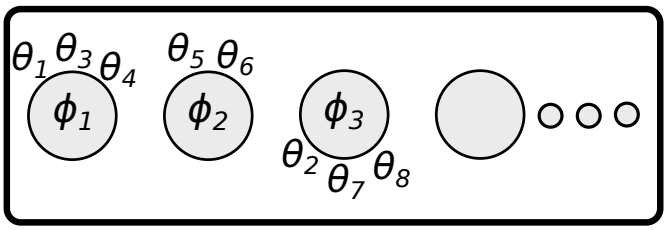

FIGURE 1.3 – *Illustration de la métaphore du restaurant chinois.*

1.2.3 Présentation du modèle (DPM)

L'effet de regroupement des θ_i par le processus de Dirichlet a été utilisé par Escobar [37] dans le cadre des modèles de mélange afin de classer des données quelconques Y_i. Ces données Y_i sont supposées provenir indépendamment d'une distribution $\mathscr{F}(\theta_i)$ de paramètre θ_i, où la loi de $(\theta_1,\dots,\theta_n)$ est donnée par le modèle (1.2.2). Ceci conduit au modèle (DPM) (*Dirichlet Process Mixture*) [1] :

$$(\mathrm{DPM}) \begin{cases} \mathrm{Y}_i|\theta_i & \overset{ind}{\sim} & \mathscr{F}(\theta_i), \\ \theta_i|\mathrm{G} & \overset{ind}{\sim} & \mathrm{G}, \\ \mathrm{G} & \sim & \mathrm{DP}(\alpha_0, \mathrm{G}_0). \end{cases} \tag{1.2.5}$$

D'après la représentation par urne de Pólya, ce modèle implique la présence de paramètres θ_i ex-æquo. En regroupant alors les données Y_i dont les paramètres θ_i sont égaux, on obtient une

1. Egalement appelé *Mixture of Dirichlet Process models*, dénomination due à Antoniak [4].

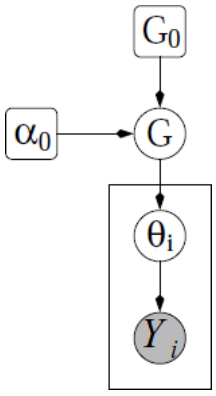

FIGURE 1.4 – *Illustration du modèle (DPM).* G_0 *est connue et* G *est un processus de Dirichlet. Les* Y_i *sont les données observées et les* θ_i *les paramètres du modèle.* α_0 *est le paramètre de concentration.*

classification de ces données. A l'aide du modèle (1.2.4), nous sommes en mesure d'exprimer un modèle équivalent au modèle (1.2.5) :

$$(\text{DPM}) \begin{cases} Y_i | c_i, \phi_c & \overset{ind}{\sim} & \mathscr{F}(\phi_{c_i}), \\ (c_1, \ldots, c_n) & \sim & \text{CRP}(\alpha_0), \\ \phi_c & \overset{ind}{\sim} & G_0. \end{cases} \tag{1.2.6}$$

Dans la définition de ce modèle, la distribution G_0 n'est pas spécifiée explicitement ; on préfèrera cependant choisir la loi G_0 conjuguée à la loi des observations afin de pouvoir expliciter analytiquement certains calculs.

1.3 Lien avec les modèles de mélange

Dans cette section, nous allons faire le lien entre le modèle (DPM) et les modèles de mélange.

1.3.1 Cas discret

Supposons que G_0 soit une distribution finie sur l'espace $\Theta = \{1, \ldots, J\}$, où $J \neq 0$ est un entier naturel quelconque. Nous pouvons alors écrire $G_0 = \sum_{j=1}^{J} \frac{\alpha_j}{\alpha_0} \delta_j$ avec $\alpha_0 = \sum_{j=1}^{J} \alpha_j > 0$. Dans cette écriture, les réels α_j et α_0 sont arbitraires.

Commençons par montrer que la loi de G est la même dans les deux modèles suivants :

$$G \sim \text{DP}(\alpha_0, G_0),$$

et

$$\begin{cases} G|\sigma_1, \ldots, \sigma_J & = & \sum_{j=1}^{J} \sigma_j \delta_j, \\ (\sigma_1, \ldots, \sigma_J) & \sim & \mathscr{D}ir(\alpha_1, \ldots, \alpha_J). \end{cases}$$

Supposons $G \sim \text{DP}(\alpha_0, G_0)$. D'après la remarque 1.1, G a pour support $\{1, \ldots, J\}$. Ainsi, on peut écrire $G = \sum_{j=1}^{J} G(\{j\}) \delta_j$. En notant pour chaque entier j, $\sigma_j = G(\{j\})$, on a alors $(\sigma_1, \ldots, \sigma_J) \sim \mathscr{D}ir(\alpha_1, \ldots, \alpha_J)$ et de plus, $G|\sigma_1, \ldots, \sigma_J = \sum_{j=1}^{J} \sigma_j \delta_j$.

Réciproquement, si $\left\{ \begin{array}{rcl} \mathrm{G}|\sigma_1,\ldots,\sigma_\mathrm{J} & = & \sum_{j=1}^{\mathrm{J}} \sigma_j \delta_j, \\ (\sigma_1,\ldots,\sigma_\mathrm{J}) & \sim & \mathscr{D}ir(\alpha_1,\ldots,\alpha_\mathrm{J}), \end{array} \right.$ alors pour toute partition mesurable $\mathrm{A}_1,\ldots,\mathrm{A}_r$ de $\{1,\ldots,\mathrm{J}\}$ on a :

$$(\mathrm{G}(\mathrm{A}_1),\ldots,\mathrm{G}(\mathrm{A}_r))\,|\sigma_1,\ldots,\sigma_\mathrm{J} = (\sigma_{\mathrm{A}_1},\ldots,\sigma_{\mathrm{A}_r}),$$

où pour chaque entier j, $\sigma_{\mathrm{A}_j} = \sum_{i\in\mathrm{A}_j}\sigma_i$. Comme $\sigma \sim \mathscr{D}ir(\alpha_1,\ldots,\alpha_\mathrm{J})$, nous obtenons également $(\sigma_{\mathrm{A}_1},\ldots,\sigma_{\mathrm{A}_r}) \sim \mathscr{D}ir(\alpha_{\mathrm{A}_1},\ldots,\alpha_{\mathrm{A}_r})$ où à nouveau, $\alpha_{\mathrm{A}_j} = \sum_{i\in\mathrm{A}_j}\alpha_i$. Ainsi $(\mathrm{G}(\mathrm{A}_1),\ldots,\mathrm{G}(\mathrm{A}_r))\,|\sigma_1,\ldots,\sigma_\mathrm{J} \sim \mathscr{D}ir(\alpha_{\mathrm{A}_1},\ldots,\alpha_{\mathrm{A}_r})$, c'est-à-dire $\mathrm{G} \sim \mathrm{DP}(\alpha_0,\mathrm{G}_0)$ par définition, d'où l'équivalence entre les deux modèles.

Pour le cas du modèle (DPM) (1.2.5) :

$$(\mathrm{DPM})\left\{ \begin{array}{rcl} \mathrm{Y}_i|\theta_i & \overset{ind}{\sim} & \mathscr{F}(\theta_i), \\ \theta_i|\mathrm{G} & \overset{ind}{\sim} & \mathrm{G}, \\ \mathrm{G} & \sim & \mathrm{DP}(\alpha_0,\mathrm{G}_0), \end{array} \right.$$

il est équivalent d'écrire :

$$\left\{ \begin{array}{rcl} \mathrm{Y}_i|\theta_i & \overset{ind}{\sim} & \mathscr{F}(\theta_i), \\ \theta_i|\mathrm{G} & \overset{ind}{\sim} & \mathrm{G}, \\ \mathrm{G}|\sigma_1,\ldots,\sigma_\mathrm{J} & = & \sum_{j=1}^{\mathrm{J}} \sigma_j \delta_j, \\ (\sigma_1,\ldots,\sigma_\mathrm{J}) & \sim & \mathscr{D}ir(\alpha_1,\ldots,\alpha_\mathrm{J}). \end{array} \right.$$

En intégrant G, il vient :

$$\left\{ \begin{array}{rcl} \mathrm{Y}_i|\theta_i & \overset{ind}{\sim} & \mathscr{F}(\theta_i), \\ \theta_i|\sigma_1,\ldots,\sigma_\mathrm{J} & \overset{ind}{\sim} & \sum_{j=1}^{\mathrm{J}} \sigma_j \delta_j, \\ (\sigma_1,\ldots,\sigma_\mathrm{J}) & \sim & \mathscr{D}ir(\alpha_1,\ldots,\alpha_\mathrm{J}). \end{array} \right. \tag{1.3.1}$$

Enfin, nous pouvons écrire :

$$\begin{array}{rcl} p(\mathrm{Y}_i|\sigma_1,\ldots,\sigma_\mathrm{J}) & = & \int p(\mathrm{Y}_i|\theta_i) \times \left(\sum_{j=1}^{\mathrm{J}} \sigma_j \delta_j(d\theta_i)\right), \\ & = & \sum_{j=1}^{\mathrm{J}} \sigma_j\, p(\mathrm{Y}_i|j), \end{array}$$

où $p(\bullet)$ est la notation générique d'une densité de probabilité. En intégrant alors θ_i dans le modèle (1.3.1), le modèle (DPM) est équivalent dans ce cas à :

$$\left\{ \begin{array}{rcl} \mathrm{Y}_i|\sigma_1,\ldots,\sigma_\mathrm{J} & \overset{ind}{\sim} & \sum_{j=1}^{\mathrm{J}} \sigma_j \mathscr{F}(j), \\ (\sigma_1,\ldots,\sigma_\mathrm{J}) & \sim & \mathscr{D}ir(\alpha_1,\ldots,\alpha_\mathrm{J}), \end{array} \right.$$

et revient à un modèle de mélange bayésien classique où les poids du mélange sont simulés suivant la distribution de Dirichlet.

1.3.2 Cas continu

Lorsque G_0 est une distribution continue, le modèle de mélange est plus complexe. En effet, d'après la représentation de Sethuraman (proposition 1.1), nous pouvons écrire $\mathrm{G} = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}$

avec $\pi \sim \text{GEM}(\alpha_0)$ et $\left(\phi_k\right)_{k\geq 1}$ i.i.d. de loi G_0. Nous pouvons ainsi affirmer l'équivalence du modèle (DPM) (1.2.5) avec le modèle :

$$\left\{\begin{array}{ccc} \text{Y}_i|\theta_i & \overset{ind}{\sim} & \mathscr{F}(\theta_i), \\ \theta_i|\pi,\phi & \overset{ind}{\sim} & \sum_{k=1}^{\infty}\pi_k\delta_{\phi_k}, \\ \pi & \sim & \text{GEM}(\alpha_0), \\ \phi_k & \overset{ind}{\sim} & \text{G}_0. \end{array}\right. \tag{1.3.2}$$

De cette écriture, on peut déduire pour tout ensemble mesurable A :

$$\begin{array}{rcl} \mathbb{P}(\text{Y}_i \in \text{A}|\pi,\phi) & = & \mathbb{E}(\mathbb{P}(\text{Y}_i \in \text{A}|\pi,\phi,\theta_i)|\pi,\phi), \\ & = & \mathbb{E}(\mathscr{F}(\theta_i)(\text{A})|\pi,\phi), \\ & = & \sum_{k=1}^{\infty}\pi_k\mathscr{F}(\phi_k)(\text{A}). \end{array}$$

En intégrant alors θ_i dans le modèle (1.3.2), nous déduisons l'équivalence suivante :

$$\left\{\begin{array}{ccc} \text{Y}_i|\pi,\phi & \overset{ind}{\sim} & \sum_{k=1}^{\infty}\pi_k\mathscr{F}(\phi_k), \\ \pi & \sim & \text{GEM}(\alpha_0), \\ \phi_k & \overset{ind}{\sim} & \text{G}_0. \end{array}\right.$$

Le modèle (DPM) peut donc s'interpréter comme un modèle de mélange avec une infinité de classes *a priori*.

1.4 Analyse de la consistance du modèle

La question de la consistance est importante en classification. Supposons en effet que les données $\text{Y}_1,\ldots,\text{Y}_n$ soient indépendantes et identiquement distribuées suivant une certaine distribution de probabilité P sur un espace probabilisé $\mathscr{X}$. La question de la consistance est la suivante : si nous simulons de plus en plus de données, est-ce que les résultats de classification convergent vers une bonne partition de l'espace $\mathscr{X}$?

Dans cette section, nous considérerons toujours deux modèles. Le premier est le modèle "vrai" (MV) et le second est celui que l'on utilise pour classer les données Y_i, et il s'agit du modèle (DPM) dans notre cas.

1.4.1 Cas discret

Résultats bibliographiques

En 2008, Abraham & Cadre [1] ont étudié la concentration des lois *a posteriori* dans le cadre de modèles mal-spécifiés. Plus précisément, les données proviennent de manière i.i.d. d'une loi $\mathscr{Q}$, mais les auteurs vont les supposer provenir d'une famille paramétrique de densités $\{h_\sigma(x), \sigma \in \Theta\}$. De plus, le modèle peut être mal-spécifié, c'est-à-dire qu'il n'existe pas forcément $\sigma_\text{V} \in \Theta$ pour lequel $q = h_{\sigma_\text{V}}$, où q est la densité de la loi $\mathscr{Q}$ par rapport à une certaine mesure. Ainsi, le modèle (MV) est le suivant :

$$(\text{MV})\left\{\begin{array}{ccc} \text{Y}_i & \overset{ind}{\sim} & \mathscr{Q}. \end{array}\right.$$

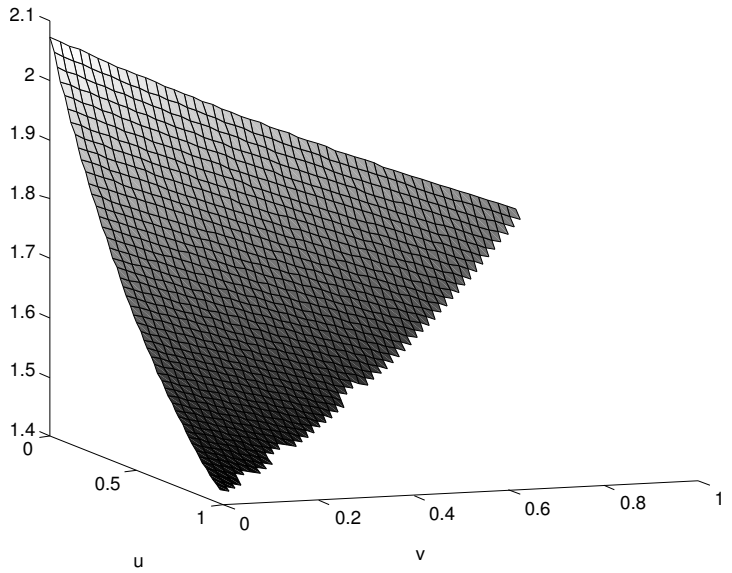

FIGURE 1.5 – *Représentation de* $(u,v) \mapsto -\int_{\mathbb{R}} log\big(up(x|1)+vp(x|2)+(1-u-v)p(x|3)\big)\,q(x)dx$. *Nous choisissons* $J=3$, $\mathscr{F}(j)=\mathscr{N}(1,j)$ *et* $\mathscr{Q}=\mathscr{N}(1,1)$. *Le modèle n'est pas mal-spécifié.*

Soit donc G_0 une distribution finie sur l'espace $\Theta=\{1,\dots,J\}$, où $J\neq 0$ est un entier naturel quelconque. Ecrivons $G_0=\sum_{j=1}^{J}\frac{\alpha_j}{\alpha_0}\delta_j$ avec $\alpha_0=\sum_{j=1}^{J}\alpha_j>0$. Dans cette écriture, les réels α_j et α_0 sont arbitraires. D'après la section précédente, le modèle (DPM) devient dans ce cas :

$$\text{(DPM)}\begin{cases} Y_i|\sigma=(\sigma_1,\dots,\sigma_J) & \overset{ind}{\sim} & \sum_{j=1}^{J}\sigma_j\mathscr{F}(j), \\ \sigma & \sim & \mathscr{D}ir(\alpha_1,\dots,\alpha_J), \end{cases}$$

Clairement, nous pouvons donc identifier Θ avec le simplexe de dimension $J-1$ et $h_\sigma(Y)=\sum_{j=1}^{J}\sigma_j p(Y|j)$, où $p(\bullet|j)$ est la densité de la loi $\mathscr{F}(j)$ par rapport à la mesure de Lebesgue. Si l'on écrit pour toute fonction g $\mathscr{Q}$-intégrable sur $\mathbb{R}$:

$$Q(g)=\int_{\mathbb{R}} g(x)\mathscr{Q}(dx),$$

et si l'on note pour tout $\sigma\in\Theta$, $f_\sigma(Y)=-\log(h_\sigma(Y))$, une condition suffisante pour que σ converge *a posteriori* est l'existence d'un $\theta\in\mathring{\Theta}$, intérieur de Θ, tel que pour tout $\sigma\in\overline{\Theta}$, adhérence de Θ, avec $\theta\neq\sigma$, $Q(f_\theta)<Q(f_\sigma)$.

En pratique, il peut être intéressant de voir comment se vérifie cette hypothèse. Pour cela, imaginons que $\mathscr{Q}$ soit la loi normale univariée de centre 1 et de variance 1. Si $J=3$, on obtient :

$$Q(f_\sigma)=-\int_{\mathbb{R}}\log\left(\sum_{j=1}^{3}\sigma_j p(x|j)\right)q(x)dx,$$

où $\sigma_3=1-\sigma_1-\sigma_2$ car $(\sigma_1,\sigma_2,\sigma_3)\in\Theta$ simplexe de dimension $J-1=2$. Supposons enfin que $\mathscr{F}(j)=\mathscr{N}(1,j)$ soit la loi normale univariée de centre 1 et de variance j. Dans ce cas, le modèle n'est pas mal-spécifié et il est clair que $\sigma_V=(1,0,0)$. Le graphe de $\sigma\mapsto Q(f_\sigma)$ est donné sur la figure 1.5.

D'après la figure 1.5, on note l'existence un paramètre $\theta\in\Theta$ minimisant la fonction $\sigma\mapsto Q(f_\sigma)$, atteint sur le bord de l'espace Θ et non dans l'intérieur de celui-ci. Après une lecture attentive de

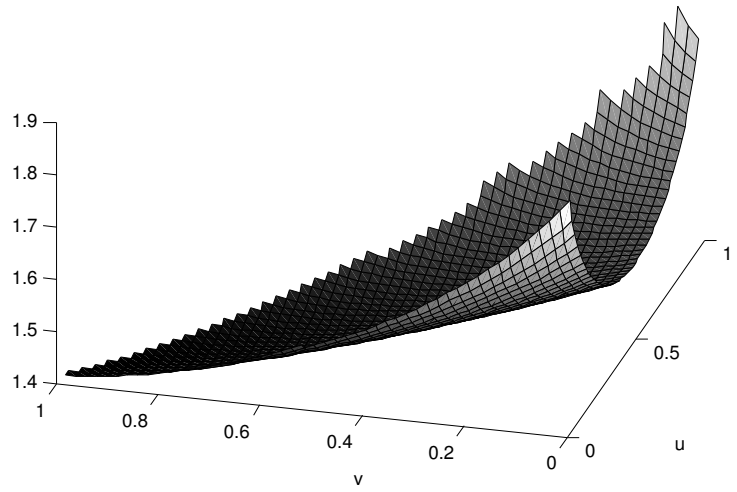

FIGURE 1.6 – *Représentation de* $(u,v) \mapsto -\int_{\mathbb{R}} log\big(up(x|1)+vp(x|2)+(1-u-v)p(x|3)\big)\,q(x)\,dx$. *Nous choisissons* $J=3$, $\mathscr{F}(j)=\mathscr{N}(j,1)$ *et* $\mathscr{Q}=\mathscr{N}(1,1)$. *Le modèle n'est pas mal-spécifié.*

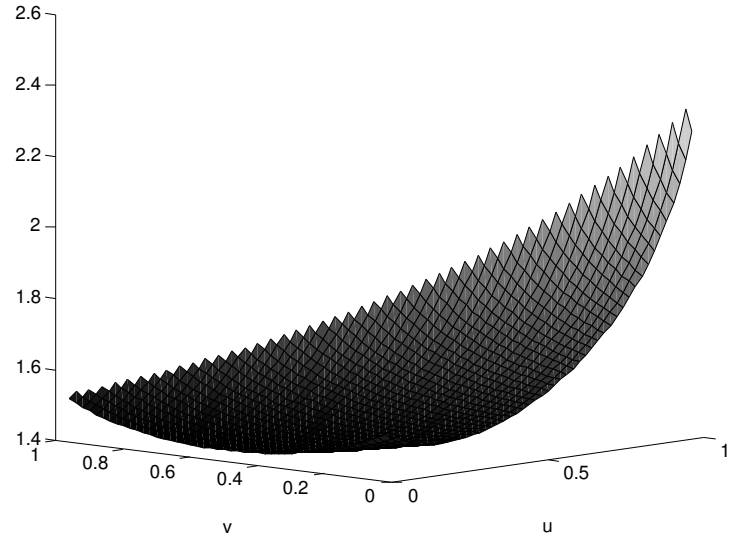

FIGURE 1.7 – *Représentation de* $(u,v) \mapsto -\int_{\mathbb{R}} log\big(up(x|1)+vp(x|2)+(1-u-v)p(x|3)\big)\,q(x)\,dx$. *Nous choisissons* $J=3$, $\mathscr{F}(j)=\mathscr{N}(j,1)$ *et* $\mathscr{Q}=\mathscr{N}(1.5,1)$. *Le modèle est mal-spécifié.*

l'article [1], il se trouve qu'il n'est pas nécessaire de restreindre θ à être dans l'intérieur $\mathring{\Theta}$, tout du moins pour le premier théorème de convergence *a posteriori*. D'après le graphe, on voit bien que $\theta=(1,0,0)=\sigma_V$, et nous sommes donc assurés que le modèle (DPM) est consistant dans ce cas-là.

Nous avons également représenté la fonction $\sigma \mapsto Q(f_\sigma)$ pour deux autres cas. Dans le premier cas, $\mathscr{F}(j)=\mathscr{N}(j,1)$ et $\mathscr{Q}=\mathscr{N}(1,1)$. Le modèle n'est donc pas mal-spécifié. Dans le second cas, $\mathscr{F}(j)=\mathscr{N}(j,1)$ et $\mathscr{Q}=\mathscr{N}(1.5,1)$. Le modèle est cette fois-ci mal-spécifié. Les graphes sont visibles sur les figures 1.6 et 1.7.

Dans le premier cas, $\theta=(1,0,0)=\sigma_V$. Le modèle (DPM) est donc consistant. Dans le second cas, le modèle étant mal-spécifié, on ne peut avoir $\theta=\sigma_V$, mais on observe que $\theta=(0,0.5,0.5)$. Les

données étant simulées suivant $\mathcal{Q} = \mathcal{N}(1.5,1)$, le modèle (DPM) attribue la moitié des observations à la classe de centre 1 et l'autre moitié à la classe de centre 2.

Ainsi, dans le cas fini, la loi *a posteriori* de σ semble toujours se concentrer vers le paramètre θ minimisant la fonction $\sigma \mapsto \mathrm{Q}(f_\sigma)$ précédente. Plus récemment, Rousseau & Mengersen [107] ont étudié la question de la consistance de la loi *a posteriori* dans le cadre de modèles de mélange finis. Les auteurs étudient la consistance dans le cas où le vrai modèle est un modèle de mélange à k_0 composantes et où le modèle approché est un modèle de mélange à k composantes, avec $k < k_0$. Dans notre cas, nous avons la même chose, mis à part que nous n'estimons pas les centres des classes. Lorsqu'il est nécessaire d'estimer également les centres des classes en plus des proportions, une condition sur les α_j apparaît, faisant intervenir la dimension d des centres à estimer. Lorsque les α_j sont suffisamment petits, le modèle est consistant.

Enfin, si l'on note K le nombre de classes, il est possible de montrer le résultat suivant :

$$\mathbb{P}(\mathrm{K}=k|\mathrm{Y}_1,\dots,\mathrm{Y}_n) \propto \sum\sum \alpha_{i_1}^{[|g_{i_1}|]} \dots \alpha_{i_k}^{[|g_{i_k}|]} \times \prod_{i'_1 \in g_{i_1}} f(x'_{i_1}|i_1) \dots \prod_{i'_k \in g_{i_k}} f(x'_{i_k}|i_k),$$

où $\prod_{i=1}^{n}(\alpha_0+i-1) = \alpha_0^{[n]}$ et la première somme porte sur les sous-ensembles $\{i_1,\dots,i_k\}$ de $\{1,\dots,\mathrm{J}\}$ en k parties, avec $i_1 < \dots < i_k$ et la seconde somme sur les partitions $\{g_{i_1},\dots,g_{i_k}\}$ de $\{1,\dots,n\}$ en k parties, où $g_j = \{i : \theta_i = j\}$. Cette formule est cependant numériquement impossible à mettre en œuvre, comme nous le verrons plus loin dans le cas continu.

Données simulées

Afin d'étudier le comportement du modèle (DPM), n données ont été simulées indépendamment suivant l'unique loi $\mathcal{N}(1,1)$. Le modèle (DPM) étant un modèle de mélange fini, nous inférons le paramètre σ *a posteriori* à l'aide d'un échantillonneur de Gibbs classique. Cela nous permet de simuler suivant la loi de $\sigma|\mathrm{Y}$. Nous choisissons un temps de chauffe (*burn-in*) de 20000 et nous conservons 1 itération sur 10 (*thinning*) jusqu'à la 50000^e itération, pour un total de 3000 itérations retenues. Nous souhaitons simplement analyser le comportement des proportions $\sigma_1,\dots,\sigma_\mathrm{J}$ *a posteriori* de Y.

Les paramètres sont fixés à $\mathrm{J}=5$ et $\alpha_1 = \dots = \alpha_5 = 5$. Les simulations ont été réalisées pour différents nombres d'observations, à savoir $n \in \{10,50,100,500,1000,5000\}$. Afin de justifier le choix du *burn-in* et du *thinning*, les graphiques de convergence du paramètre $\sigma_1|\mathrm{Y}$, ainsi que de la fonction d'auto-corrélation avec thinning de 10, ont été reportés figures 1.8 et 1.9.

Résultats de simulation

A chaque itération, nous obtenons les proportions *a posteriori*. Nous avons représenté sur la figure 1.10, pour chaque valeur $n \in \{10,50,100,500,1000,5000\}$, le diagramme en boîte des valeurs *a posteriori* de la proportion σ_1, ainsi que l'histogramme du nombre de classes *a posteriori*, basés sur les 3000 itérations retenues.

Une première constatation est que l'on ne peut pas inférer le nombre de classes à l'aide de l'histogramme. En effet, l'algorithme peut reconnaître un nombre erroné de classes (par exemple, une majorité de 3 classes) alors que 98% des observations sont classées dans la même classe 1, qui est la bonne classe. On remarque aussi que même si le nombre de classes observées augmente lorsque n augmente, il y a convergence de la proportion $\sigma_1|\mathrm{Y}$ vers 1.

Cela illustre le comportement asymptotique du modèle (DPM) dans ce cas.

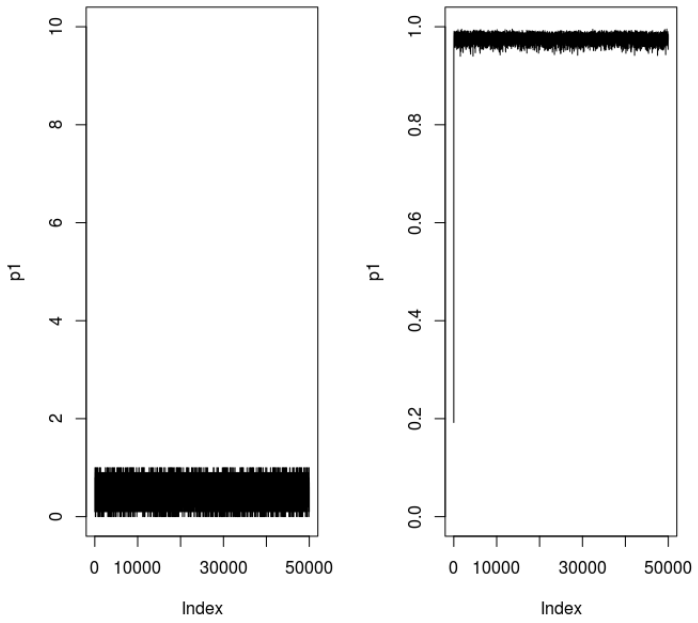

FIGURE 1.8 – *Convergence du paramètre* $\sigma_1|Y$. *Le burn-in est choisi de valeur 20000. A gauche pour* $n = 10$ *et à droite pour* $n = 5000$.

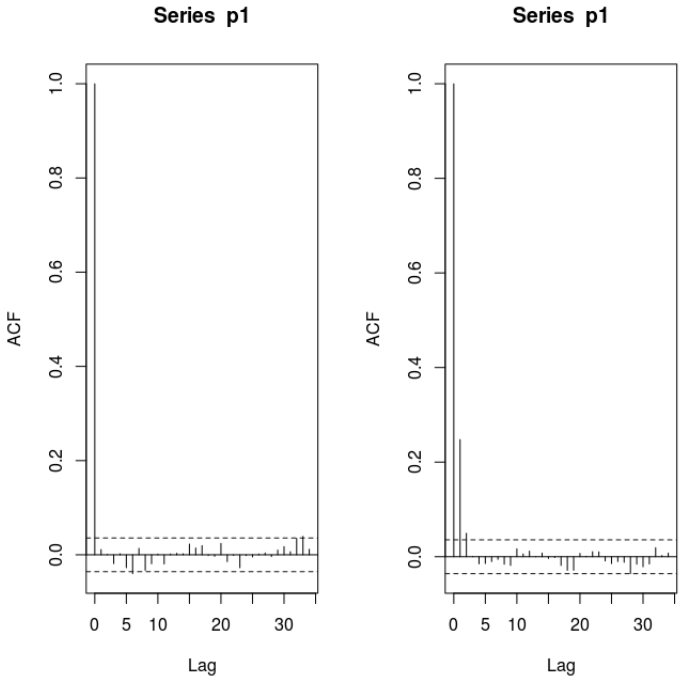

FIGURE 1.9 – *Fonction d'auto-corrélation pour les valeurs des proportions a posteriori* $\sigma_1|Y$, *lorsque l'on choisit un thinning de 10. A gauche pour* $n = 10$ *et à droite pour* $n = 5000$.

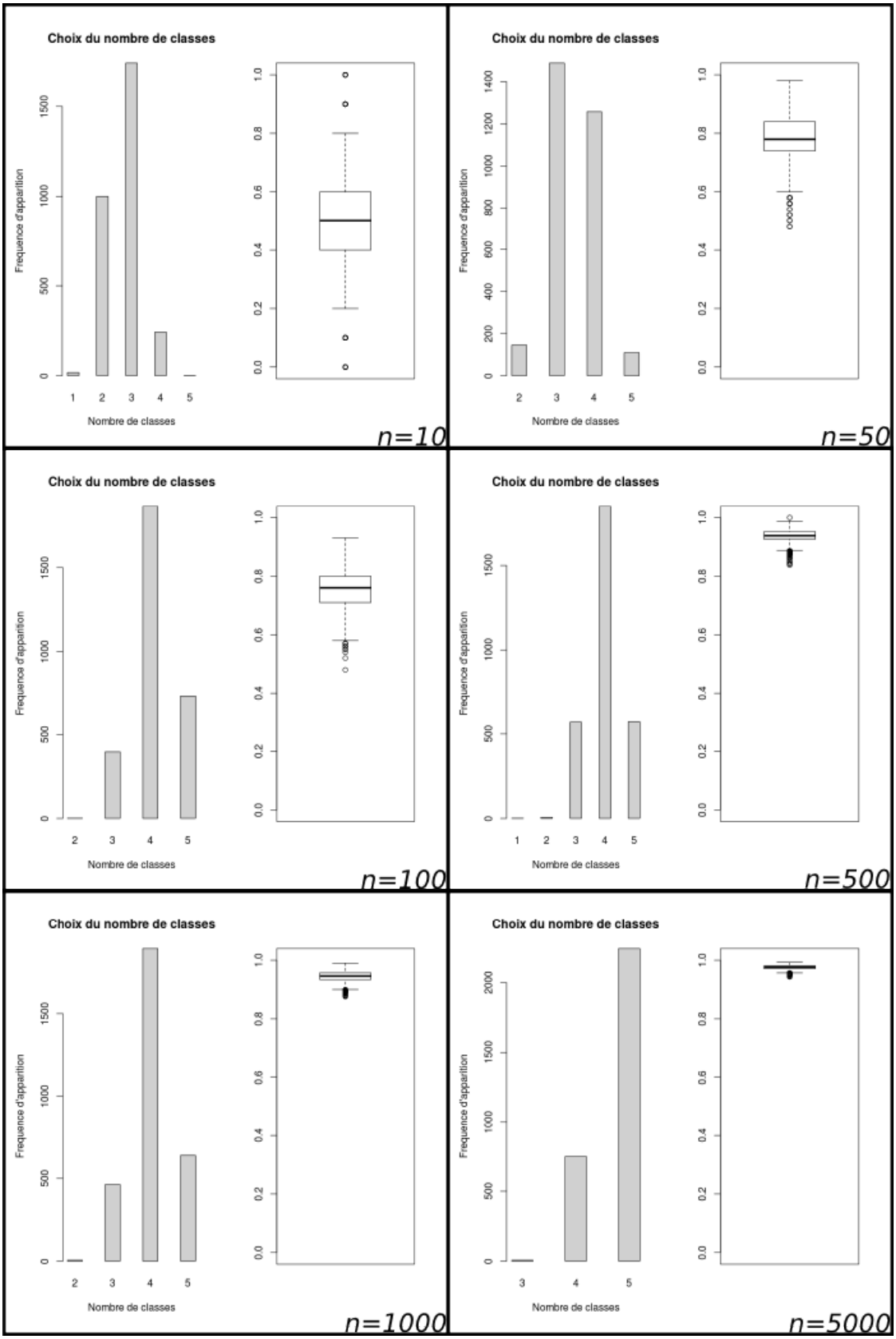

FIGURE 1.10 – *Histogrammes du nombre de classes a posteriori et diagrammes en boîte des valeurs* $\sigma_1|Y$.

1.4.2 Cas continu

A présent, le modèle de mélange possède une infinité de classes possibles. A ce jour, la consistance du (DPM) en classification ne semble pas avoir été résolue entièrement dans le cadre continu. Plusieurs articles [81, 82] soulignent l'inconsistance des processus de Dirichlet pour inférer sur le nombre de classes, notamment lorsque le paramètre α_0 est fixe. Aussi, pour limiter ces problèmes, et comme nous le verrons plus loin, nous poserons une loi *a priori* sur ce paramètre et il sera inféré *a posteriori*. Notons que Kimura et al. [61] obtiennent de bons résultats de classification en inférant sur α_0 à partir d'un algorithme EM.

Tout comme dans le cas discret, certaines classes peuvent être isolées et ne contenir qu'un très faible nombre d'observations. Nous avons également remarqué, par le biais de simulations, que même si deux observations se trouvent dans deux classes distinctes, les centres des classes inférés peuvent néanmoins être très proches.

Nous allons à présent trouver une formule théorique du nombre de classes *a posteriori* inféré par le (DPM). Notons $\mathrm{K}(\theta)$ le nombre de classes (qui dépend des valeurs de $\theta = (\theta_1, \ldots, \theta_n)$). Nous pouvons écrire :

$$\begin{aligned} p(\mathrm{Y}_1, \ldots, \mathrm{Y}_n) &= \int p(\mathrm{Y}_1, \ldots, \mathrm{Y}_n | \theta_1, \ldots, \theta_n) p(d\theta_1, \ldots, d\theta_n), \\ &= \int \prod_{i=1}^n \mathrm{P}(\mathrm{Y}_i|\theta_i) \prod_{i=1}^n \left(\frac{\alpha_0 g_0 + \sum_{j=1}^{i-1} \delta_{\theta_j}}{\alpha_0 + i - 1} \right) (d\theta_i), \end{aligned}$$

où g_0 est la densité de probabilité de G_0. En notant pour chaque indice i, $g_i(\theta_i) = \mathrm{P}(\mathrm{Y}_i|\theta_i)$, il vient d'autre part :

$$p(\mathrm{Y}_1, \ldots, \mathrm{Y}_n) = \frac{1}{\alpha_0^{[n]}} \int \prod_{i=1}^n g_i(\theta_i) \prod_{i=1}^n \left(\alpha_0 g_0 + \sum_{j=1}^{i-1} \delta_{\theta_j} \right) (d\theta_i).$$

Par application du lemme 2 de Lo [66], nous obtenons alors :

$$p(\mathrm{Y}_1, \ldots, \mathrm{Y}_n) = \frac{1}{\alpha_0^{[n]}} \sum_{\mathrm{P}} \prod_{i=1}^{\mathrm{N(P)}} (e_i - 1)! \int \prod_{l \in \mathrm{C}_i} p(\mathrm{Y}_l|u) \alpha_0 g_0(u) du,$$

où P désigne l'ensemble des partitions possibles de l'ensemble $\{1, \ldots, n\}$, $\mathrm{N(P)}$ le nombre de sous-ensembles de la partition $\mathrm{P} = \{\mathrm{C}_1, \ldots, \mathrm{C}_{\mathrm{N(P)}}\}$ de $\{1, \ldots, n\}$ et e_i le nombre d'éléments dans C_i. Ecrivons ce résultat de la manière suivante :

$$\begin{aligned} p(\mathrm{Y}_1, \ldots, \mathrm{Y}_n) &= \frac{1}{\alpha_0^{[n]}} \sum_{m=1}^n \sum_{\mathrm{P}:|\mathrm{P}|=m} \left(\prod_{i=1}^m (e_i - 1)! \int \prod_{l \in \mathrm{C}_i} p(\mathrm{Y}_l|u) \alpha_0 g_0(u) du \right), \\ &= \frac{1}{\alpha_0^{[n]}} \sum_{m=1}^n \alpha_0^m \sum_{\mathrm{P}:|\mathrm{P}|=m} \left(\prod_{i=1}^m (e_i - 1)! \int \prod_{l \in \mathrm{C}_i} p(\mathrm{Y}_l|u) g_0(u) du \right), \\ &= \frac{1}{\alpha_0^{[n]}} \sum_{m=1}^n \alpha_0^m \sum_{\mathrm{P}:|\mathrm{P}|=m} \left(\prod_{i=1}^m (e_i - 1)! \mathrm{H}(\mathrm{C}_i) \right), \end{aligned}$$

avec $\mathrm{H}(\mathrm{C}_i) = \int \prod_{l \in \mathrm{C}_i} p(\mathrm{Y}_l|u) g_0(u) du$. A l'instar de Liu [64], nous pouvons écrire ce résultat sous la forme suivante :

$$p(\mathrm{Y}_1, \ldots, \mathrm{Y}_n) = \frac{1}{\alpha_0^{[n]}} \sum_{m=1}^n \alpha_0^m \mathscr{L}_m(\mathrm{Y}),$$

où $\mathscr{L}_m(\mathrm{Y}) = \sum_{\mathrm{P}:|\mathrm{P}|=m} \left(\prod_{i=1}^{m}(e_i-1)!\mathrm{H}(\mathrm{C}_i)\right)$. Remarquons alors que $\mathbb{P}(\mathrm{K}(\theta)=m|\mathrm{Y}) = \frac{p(\mathrm{Y},\mathrm{K}(\theta)=m)}{p(\mathrm{Y})}$, or $p(\mathrm{Y},\theta) = \prod_{i=1}^{n} p(\mathrm{Y}_i|\theta_i)\left(\alpha_0 g_0 + \sum_{j=1}^{i-1}\delta_{\theta_j}\right)(\theta_i) \times \frac{1}{\alpha_0^{[n]}}$. Ainsi :

$$\begin{aligned} p(\mathrm{Y},\mathrm{K}(\theta)=m) &= \int_{\{\theta:\mathrm{K}(\theta)=m\}} \prod_{i=1}^{n} p(\mathrm{Y}_i|\theta_i)\left(\alpha_0 g_0 + \sum_{j=1}^{i-1}\delta_{\theta_j}\right)(\theta_i) \times \frac{1}{\alpha_0^{[n]}} \\ &= \frac{\alpha_0^m}{\alpha_0^{[n]}} \sum_{\mathrm{P}:|\mathrm{P}|=m} \left(\prod_{i=1}^{m}(e_i-1)!\mathrm{H}(\mathrm{C}_i)\right) \\ &= \frac{\alpha_0^m}{\alpha_0^{[n]}} \mathscr{L}_m(\mathrm{Y}). \end{aligned}$$

Le résultat final s'ensuit :

$$\mathbb{P}(\mathrm{K}(\theta)=m|\mathrm{Y}) = \frac{\alpha_0^m \mathscr{L}_m(\mathrm{Y})}{\sum_{j=1}^{n} \alpha_0^j \mathscr{L}_j(\mathrm{Y})},$$

c'est-à-dire :

$$\mathbb{P}(\mathrm{K}(\theta)=m|\mathrm{Y}) \propto \alpha_0^m \mathscr{L}_m(\mathrm{Y}).$$

Cependant, comme dans le cas discret, cette somme ne peut pas être explicitée. Le temps de calcul est en effet trop élevé, à cause de la recherche de toutes les partitions possibles de l'ensemble $\{1,\ldots,n\}$. Le nombre de partitions en k sous-ensembles d'un ensemble à n éléments est égal au nombre de Stirling de seconde espèce $\left\{{n \atop k}\right\}$. Pour n fixé, la somme des nombres de Stirling de seconde espèce, $\sum_{k=1}^{n}\left\{{n \atop k}\right\}$ désigne le n^e nombre de Bell noté B_n. La complexité est telle que pour $n=10$, le calcul est instantané, alors que pour $n=20$, plusieurs jours sont nécessaires. En effet pour $n=10$ on dénombre 115975 partitions possibles et pour $n=20$ il y en a plus d'un milliard.

1.5 Implémentation algorithmique

En réalisant l'inférence *a posteriori* des paramètres du (DPM), nous inférons automatiquement le nombre de classes présentes dans le jeu de données ainsi que l'assignation des observations aux classes. La plupart des auteurs proposent des échantillonneurs de Gibbs, utilisant différentes représentations des processus de Dirichlet. Historiquement, le premier algorithme MCMC pour le modèle (DPM) a été proposé par Escobar en 1988 dans un travail non publié. Cet algorithme sera plus tard publié en 1994 [37].

D'autres auteurs [12, 38, 71, 87] ont à leur tour proposé des améliorations à cet algorithme. Grâce à un développement numérique important, plusieurs algorithmes ont été proposés dans le cas de modèles non conjugués [44, 45, 72, 73, 87, 132]. Dans ces derniers, le processus de Dirichlet est intégré.

D'autres méthodes algorithmiques existent, comprenant entre autres le *slice sampling* [58, 130, 131], le *sequential importance sampling* [28, 39, 64], l'*expectation propagation* [61, 83], le *fast search* et le recuit-simulé [49, 62, 133], la *predictive recursion* [18, 88] ou encore les méthodes variationnelles [9].

Dans cette partie, nous détaillons simplement l'algorithme de base tel que présenté par Escobar en 1994 [37]. Dans ce qui suit, nous notons $\theta = (\theta_1,\ldots,\theta_n)$ et $\mathrm{Y} = (\mathrm{Y}_1,\ldots,\mathrm{Y}_n)$. Le paramètre α_0 ainsi que la distribution de base G_0 sont supposés fixes. Si $i \in \{1,\ldots,n\}$, nous notons θ^{-i} (respectivement Y^{-i}) le vecteur θ (respectivement Y) privé de sa i^e composante.

1.5.1 Algorithme de base

Pour simuler suivant la loi *a posteriori* de θ dans le modèle (DPM) (1.2.5), le plus simple est d'utiliser la représentation par urne de Pólya, que l'on peut écrire de manière formelle d'après l'égalité (1.2.3) :

$$\mathrm{P}(d\theta_i|\theta^{-i}) = \sum_{j\neq i} \frac{1}{\alpha_0+n-1}\delta_{\theta_j}(d\theta_i) + \frac{\alpha_0}{\alpha_0+n-1}\mathrm{G}_0(d\theta_i).$$

Or, d'après le théorème de Bayes, la loi de $\theta_i|\mathrm{Y},\theta^{-i}$ est proportionnelle à :

$$\mathrm{P}(d\theta_i|\mathrm{Y},\theta^{-i}) \propto p(\mathrm{Y}|\theta_i,\theta^{-i})\mathrm{P}(d\theta_i|\theta^{-i}).$$

Il nous faut donc calculer la vraisemblance du modèle. Celle-ci est donnée par définition par :

$$p(\mathrm{Y}|\theta_i,\theta^{-i}) = \prod_{j=1}^{n} p(\mathrm{Y}_j|\theta_j) = \left[\prod_{j\neq i}^{n} p(\mathrm{Y}_j|\theta_j)\right] p(\mathrm{Y}_i|\theta_i).$$

Ainsi, nous pouvons écrire :

$$\begin{array}{rcl}
\mathrm{P}(d\theta_i|\mathrm{Y},\theta^{-i}) & \propto & p(\mathrm{Y}|\theta_i,\theta^{-i})\mathrm{P}(d\theta_i|\theta^{-i}), \\
& \propto & \left[\left(\prod_{j\neq i}^{n} p(\mathrm{Y}_j|\theta_j)\right) p(\mathrm{Y}_i|\theta_i)\right] \times \left(\sum_{j\neq i} \frac{1}{\alpha_0+n-1}\delta_{\theta_j}(d\theta_i) + \frac{\alpha_0}{\alpha_0+n-1}\mathrm{G}_0(d\theta_i)\right), \\
& \propto & p(\mathrm{Y}_i|\theta_i) \times \left(\sum_{j\neq i} \frac{1}{\alpha_0+n-1}\delta_{\theta_j}(d\theta_i) + \frac{\alpha_0}{\alpha_0+n-1}\mathrm{G}_0(d\theta_i)\right), \\
& \propto & \sum_{j\neq i} \frac{1}{\alpha_0+n-1} p(\mathrm{Y}_i|\theta_i)\delta_{\theta_j}(d\theta_i) + \frac{\alpha_0}{\alpha_0+n-1} p(\mathrm{Y}_i|\theta_i)\mathrm{G}_0(d\theta_i).
\end{array}$$

Remarquons alors que $p(\mathrm{Y}_i|\theta_i)\delta_{\theta_j}(d\theta_i) = p(\mathrm{Y}_i|\theta_j)\delta_{\theta_j}(d\theta_i)$ et concluons que la loi de $\theta_i|\mathrm{Y},\theta^{-i}$ est proportionnelle à :

$$\mathrm{P}(d\theta_i|\mathrm{Y},\theta^{-i}) \propto \sum_{j\neq i} \frac{1}{\alpha_0+n-1} p(\mathrm{Y}_i|\theta_j)\delta_{\theta_j}(d\theta_i) + \frac{\alpha_0}{\alpha_0+n-1} p(\mathrm{Y}_i|\theta_i)\mathrm{G}_0(d\theta_i). \tag{1.5.1}$$

Cet algorithme de base peut donc se résumer de la façon suivante :

Algorithme 1: Echantillonneur de Gibbs basé sur l'urne de Pólya

Résultat : Loi a posteriori de $\theta|\mathrm{Y}$
pour *i=1,...,n* **faire**
  Mettre à jour θ_i via l'équation (1.5.1);

L'implémentation de cet algorithme n'est guère difficile. La convergence de cet algorithme est toutefois très lente car la méthode nécessite de générer chaque θ_i. Les performances de cet algorithme de base peuvent donc être améliorées en utilisant la représentation équivalente (1.2.6) du (DPM). Notre objectif n'étant pas la description des algorithmes existants, le lecteur pourra se référer aux précédentes références pour plus de précisions.

1.6 Sélection d'une classification dans un cadre bayésien

Plaçons-nous dans le cas où un algorithme MCMC aurait permis de produire un grand nombre d'itérations. Chaque itération correspond à une classification *a posteriori* des observations ; pour chaque $s \in \{1,\ldots,\mathrm{N}\}$, $c^{(s)} = (c_1^{(s)},\ldots,c_n^{(s)})$ désigne le vecteur d'assignation des classes pour chaque

observation à l'itération s. Il est alors souvent préférable de résumer cet échantillon en une seule estimation $\widehat{c}$.

Une solution intuitive consiste à classer chaque observation *a posteriori* marginalement, conduisant à la classification dite du *Maximum A Posteriori* marginal, ou *MAP* marginal. Cependant, cette tâche est complexe car l'estimateur naturel ne conduit presque jamais à des résultats convenables. Cela est dû au problème du "label switching", qui signifie que durant une itération MCMC les labels associés aux classes peuvent changer. Le *MAP* marginal requiert la mise en place de pivots afin de contourner ce problème [14, 41, 42, 118]. Cependant, les méthodes à base de pivots sont au prix de complications numériques importantes.

Une seconde possibilité est de classer les observations conjointement. Pour cela, un choix classique dans la littérature est la classification dite du *Maximum A Posteriori* global, ou encore *MAP* global, et correspond à la classification dont la probabilité *a posteriori* est la plus grande. Un estimateur naturel de la classification du *MAP* global est donné par la classification qui apparaît le plus souvent dans les itérations. Ce choix est, entre autres, suggéré dans les articles de De la Cruz-Mesía & Quintana [26] ou encore de Kim et al. [60]. En 2007, Druilhet & Marin [76] ont montré que le *MAP* global n'est pas invariant par reparamétrisation. Les auteurs proposent de remplacer l'estimateur *MAP* par l'estimateur Jeffreys *MAP*, dans lequel la mesure de Lebesgue utilisée pour l'écriture de la densité de la loi *a posteriori* est remplacée par la mesure de Jeffreys. Précisons tout de même que l'article de Druilhet & Marin a été écrit dans le cas où l'on cherche à calculer le *MAP* d'un paramètre de dimension infini.

Dahl [25] a proposé une autre méthode, qu'il nomme *Classification des Moindres Carrés*. Pour chaque classification $c^{(s)}$ dont on dispose, il définit la matrice $\delta^{(s)}$ dont l'élément générique est :

$$\delta_{ij}^{(s)} = \begin{cases} 1 & \text{si} \quad c_i^{(s)} = c_j^{(s)}, \\ 0 & \text{sinon.} \end{cases}$$

Il définit alors une seconde matrice $\widehat{\pi}$ d'élément générique :

$$\widehat{\pi}_{ij} = \frac{1}{\mathrm{N}} \sum_{s=1}^{\mathrm{N}} \delta_{ij}\left(c^{(s)}\right),$$

qui est une estimation de la probabilité *a posteriori* pour que la paire d'observations $(\mathrm{Y}_i, \mathrm{Y}_j)$ se trouve dans la même classe. $\widehat{\pi}$ est la matrice dite de probabilités par paires. La proposition de Dahl est de choisir la classification qui minimise la somme des écarts quadratiques $\left(\delta_{ij}^{(s)} - \widehat{\pi}_{ij}\right)^2$. Cette classification est définie par :

$$\underset{s=1,\ldots,\mathrm{N}}{\arg\min} \sum_{i=1}^{n} \sum_{j=1}^{n} \left(\delta_{ij}^{(s)} - \widehat{\pi}_{ij}\right)^2.$$

1.7 Prior sur le paramètre de concentration α_0

Nous allons, dans cette dernière section, nous consacrer à l'étude de l'influence de l'hyperparamètre α_0 sur une classification *a posteriori*.

1.7.1 Conséquences d'un paramètre α_0 fixé

Commençons par étudier le comportement du modèle (DPM) lorsque α_0 est fixé. Pour cela, nous avons simulé n données indépendamment suivant l'unique loi normale $\mathcal{N}(0,1)$. Nous sou-

haitons vérifier que le (DPM) puisse distinguer un nombre fini de classes et analyser le comportement du nombre de classes.

Pour cela, nous décidons de mettre en œuvre l'algorithme de la section 1.5.1. Nous choisissons le cas le plus courant dans la littérature, à savoir $\mathscr{F}(\theta_i) = \mathscr{N}(\theta_i, \sigma^2)$ et $G_0 = \mathscr{N}(\mu, \sigma_0^2)$. Nous initialisons l'algorithme avec des valeurs $\sigma = \sigma_0 = 1$ et $\mu = 0$. L'hyperparamètre σ_0 est fixé de sorte que la variabilité des données générées à partir de G_0 soit de l'ordre de celle des données. Différentes valeurs de σ_0 ont été testées, avec pour seule conséquence un temps de convergence de l'algorithme vers la loi stationnaire plus ou moins grand. La moyenne μ est fixée à zéro car celle-ci intervient peu dans les résultats de classification.

Les résultats ont été obtenus en produisant 30000 itérations, avec un temps de chauffe de 15000 et en retenant 1 itération sur 10. La classification retenue est celle qui apparaît le plus souvent dans les simulations *a posteriori*. Pour chaque valeur du couple $(\alpha_0, n) \in \{0.01, 0.1, 1, 10\} \times \{10, 50, 100, 500, 1000, 5000\}$, nous avons répété l'algorithme 50 fois, ce qui nous a permis d'estimer le nombre de classes *a posteriori*.

Sans présenter graphiquement les observations et les résultats, qui ne nous seraient d'aucune utilité ici, nous avons remarqué que :
– lorsque n est fixé, plus α_0 est grand et plus le nombre de classes trouvées augmente,
– lorsque n augmente, il faut diminuer la valeur de α_0 pour conserver le même nombre de classes.

Le nombre de classes dans le modèle (DPM) dépendant fortement du paramètre α_0, on conclut qu'il vaut mieux faire de l'inférence sur ce paramètre plutôt que de le fixer. Dans la littérature, plusieurs méthodes ont été proposées pour cela.

1.7.2 Prior sur le paramètre α_0

Une première méthode, proposée par Escobar en 1994 [37], consiste à discrétiser l'espace des valeurs possibles pour α_0, puis à calculer la loi *a posteriori* complète correspondante.

Plus tard en 1995, Escobar & West [38] ont proposé de placer une loi *a priori* $\mathscr{G}amma(a, b)$ sur α_0, où a est le paramètre de forme et b le paramètre d'échelle. Les lois gamma sont de support réel positif. Les auteurs introduisent de plus une variable auxiliaire η de loi conditionnelle :

$$\eta|\alpha_0 \sim \mathscr{B}eta(\alpha_0 + 1, n).$$

Le problème consiste alors à estimer α_0 *a posteriori* dans le modèle suivant :

$$\begin{aligned} \alpha_0 &\sim \mathscr{G}amma(a, b), \\ \eta|\alpha_0 &\sim \mathscr{B}eta(\alpha_0 + 1, n). \end{aligned}$$

En posant :

$$\tilde{\pi} = \frac{a + K - 1}{a + K - 1 + n\left(b - \log \eta\right)},$$

où K désigne le nombre de classes à l'instant courant de l'algorithme, il est possible de déduire la loi *a posteriori* complète suivante :

$$\alpha_0|\eta \sim \tilde{\pi}\, \mathscr{G}amma\left(a + K, b - \log \eta\right) + (1 - \tilde{\pi})\, \mathscr{G}amma\left(a + K - 1, b - \log \eta\right).$$

De plus, conditionnellement à α_0, la loi conditionnelle complète de $(\theta_1,\ldots,\theta_n)$ (1.5.1) ne change pas. Ainsi, on peut obtenir à chaque itération une nouvelle valeur de α_0 en simulant d'abord une variable η, basée sur la valeur courante de α_0, puis en simulant α_0 selon sa loi *a posteriori*.

Dans ce dernier article cependant, les auteurs n'ont pas mentionné de méthode particulière pour fixer les hyperparamètres a et b de la loi gamma. Une première idée est de fixer les valeurs des hyperparamètres de sorte à obtenir une loi peu informative $\mathcal{G}amma(1,0.5)$ pour α_0. En 2009, Dorazio [32] propose une méthode basée sur la minimisation de la distance de Kullback-Leibler entre la loi du nombre de classes *a priori* et une distribution connue du nombre de classes qui serait apportée par un expert.

D'après un résultat d'Antoniak [4], un *prior* de type Dirichlet implique :

$$\mathbb{P}(\mathrm{K}(\theta)=k|\alpha_0)=\mathrm{S}(n,k)\alpha_0^k\frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0+n)}.$$

Dans cette expression, $\mathrm{S}(n,k)$ désigne un nombre de Stirling de première espèce non signé. En intégrant alors sur α_0, on obtient :

$$\mathbb{P}(\mathrm{K}(\theta)=k)=\frac{b^a\mathrm{S}(n,k)}{\Gamma(a)}\int_0^\infty\frac{\alpha_0^{k+a-1}e^{-b\alpha_0}\Gamma(\alpha_0)}{\Gamma(\alpha_0+n)}d\alpha_0.$$

Supposons alors que l'on dispose d'une fonction de masse $g(k)$ mesurant la probabilité d'avoir k classes, apportée par exemple par un savoir d'expert. Dorazio [32] propose de minimiser la distance de Kullback-Leibler entre $\mathbb{P}(\mathrm{K}(\theta)=k)$ et $g(k)$:

$$\mathrm{D}=\sum_{k=1}^n g(k)\log\left(\frac{g(k)}{\mathbb{P}(\mathrm{K}(\theta)=k)}\right).$$

Lorsqu'aucune information *a priori* sur les classes n'est connue, on peut choisir pour g la fonction de masse d'une loi uniforme sur l'ensemble $\{1,\ldots,n\}$, ce qui donne dans ce cas :

$$\mathrm{D}=-\log n-\frac{1}{n}\sum_{k=1}^n\log\mathbb{P}(\mathrm{K}(\theta)=k).$$

Néanmoins, nous noterons que cette méthode est difficile à mettre en œuvre car l'implémentation numérique de l'intégrale et la recherche des valeurs de a et b minimisant la distance D est fastidieuse en pratique.

CHAPITRE

2 Processus gaussien a posteriori pour données fonctionnelles

Résumé

Dans ce chapitre, nous démontrons l'ensemble des résultats que nous utiliserons dans les chapitres suivants. En particulier, nous souhaitons calculer une loi a posteriori dans un modèle où loi a priori et vraisemblance sont des processus gaussiens. Notre méthodologie requiert le calcul de densités de processus gaussiens. Ces densités sont calculées à l'aide de la théorie des espaces de Hilbert à noyau reproduisant. Dans un premier temps, nous commençons par définir les données fonctionnelles et les processus gaussiens, avant d'introduire les concepts dont nous aurons besoin. Dans un second temps, nous démontrons le résultat principal dans le cas d'une seule observation, avant de le généraliser au cas d'observations multiples.

2.1 Les données fonctionnelles

2.1.1 Généralités

L'analyse de données fonctionnelles [102] consiste à mettre en place des méthodes statistiques dans lesquelles les observations sont des fonctions, c'est-à-dire des courbes. Beaucoup de domaines d'application font appel à des courbes et des signaux, comme la spectrométrie ou lors de l'étude de courbes de croissance. Avec le développement actuel du phénotypage où les données sont recueillies en temps continu, de plus en plus d'utilisateurs ont besoin d'outils capables de classer des courbes.

Une variable aléatoire est dite fonctionnelle si ses valeurs sont dans un espace de dimension infinie. Une observation d'une variable fonctionnelle est appelée donnée fonctionnelle. Le plus souvent, une donnée fonctionnelle est définie comme la trajectoire d'un processus stochastique $Y = (Y_t)_{t\in[0,T]}$. Sauf mention explicite, dans toute la suite de cette thèse, nous considérerons uniquement des processus dont les trajectoires appartiennent à l'espace $L^2([0,T])$, espace des fonctions de carré intégrable sur $[0,T]$. Cet espace étant polonais [1], il nous garantit l'existence des probabilités conditionnelles d'après Dudley [35].

En pratique, une donnée fonctionnelle n'est jamais observée continûment, mais en un nombre

1. Un espace polonais est un espace métrique, complet et séparable

fini de temps d'observation. Il est donc possible de résumer chaque courbe comme un vecteur et ainsi de transformer le problème en dimension finie, mais cette approche néglige l'aspect fonctionnel. L'utilisation de modèles fonctionnels présente l'avantage de pouvoir prendre en compte la corrélation temporelle des données. Une des spécificités des données fonctionnelles est également la possibilité d'utiliser l'information contenue dans les dérivées. Certains auteurs [19] ont montré qu'elles pouvaient révéler des caractéristiques importantes des jeux de données.

2.1.2 Les processus gaussiens

Les processus gaussiens [104, 114] jouent un rôle crucial dans la théorie des processus stochastiques car :

1. beaucoup de processus stochastiques peuvent être approchés par des processus gaussiens,
2. beaucoup de calculs sont facilités dans le cadre des processus gaussiens.

Rappelons que les processus gaussiens sont la généralisation des lois normales multivariées aux espaces de dimension infinie et qu'un processus est gaussien si, et seulement si, toutes ses lois fini-dimensionnelles sont des lois normales multivariées. Un processus gaussien de loi $\mathrm{P}_{m,\mathrm{K}}$ est défini par le biais de deux fonctions qui sont sa fonction moyenne m et sa fonction de covariance K, qui est symétrique et définie positive (voir annexe A.3). Rappelons qu'une fonction de deux variables K définie sur $[0,\mathrm{T}]\times[0,\mathrm{T}]$ est dite :

(i) définie positive si pour tout entier non nul $n\in\mathbb{N}\backslash\{0\}$, tous $t_1,\dots,t_n\in[0,\mathrm{T}]$ et $a_1,\dots,a_n\in\mathbb{R}$,

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \mathrm{K}(t_i,t_j)\geq 0, \tag{2.1.1}$$

(ii) symétrique si pour tous $s,t\in[0,\mathrm{T}]$,

$$\mathrm{K}(s,t)=\mathrm{K}(t,s), \tag{2.1.2}$$

Dans la littérature, les processus gaussiens sont souvent notés $\mathrm{GP}(m,\mathrm{K})$, mais pour des raisons de commodité, nous les noterons $\mathrm{P}_{m,\mathrm{K}}$ dans toute la suite. La fonction de covariance influe sur la régularité des trajectoires du processus. Le lecteur pourra se référer par exemple à Cramér & Leadbetter [22] ou encore Shi & Choi [114] concernant le choix de cette dernière.

2.2 Densité d'un processus gaussien

2.2.1 Objectif

Notons $\mathrm{Y}=(\mathrm{Y}_t)_{t\in[0,\mathrm{T}]}$ un processus gaussien quelconque et P_Y sa mesure de probabilité associée (voir annexe A.3). Notre objectif est de trouver une mesure de référence P pour laquelle on puisse exprimer la dérivée de Radon-Nikodym $\frac{d\mathrm{P}_\mathrm{Y}}{d\mathrm{P}}$, c'est-à-dire une expression de la densité de probabilité de P_Y par rapport à P.

Un problème récurrent en traitement du signal, et qui rejoint notre but initial, est de pouvoir extraire un signal depuis une observation bruitée. Formellement, cela revient à considérer deux processus stochastiques $\mathrm{X}=(\mathrm{X}_t)_{t\in[0,\mathrm{T}]}$ et $\epsilon=(\epsilon_t)_{t\in[0,\mathrm{T}]}$, le premier étant appelé "signal" et le second "bruit". En notant $\mathrm{Y}=\mathrm{X}+\epsilon$ le processus observé, l'un des objectifs en traitement du signal est de déterminer $\frac{d\mathrm{P}_{\mathrm{X}+\epsilon}}{d\mathrm{P}_\epsilon}$.

Ce problème est en particulier équivalent au problème de test d'hypothèse suivant :

(H0) $Y = \epsilon$,

(H1) $Y = X + \epsilon$,

dans lequel le rapport de vraisemblance est égal à $\frac{dP_{X+\epsilon}}{dP_\epsilon}(Y)$, que nous noterons $L(Y)$ dans toute la suite et que nous appellerons processus de vraisemblance [2].

Pour la bonne compréhension de la suite, le lecteur peut se référer à l'annexe A.4 pour une brève définition d'une intégrale d'un processus.

2.2.2 Travaux précurseurs pour un bruit blanc gaussien

Parmi les travaux précurseurs, Price [100, 101] est le pionnier dans le cas où ϵ est un bruit blanc gaussien $P_{0,R}$, pour lequel la fonction de covariance R est donnée par :

$$R(s,t) = \delta_s(t).$$

Supposons de plus $X \sim P_{0,K}$ avec X et ϵ indépendants. Historiquement, à condition que :

1. la fonction $K(\bullet,\bullet)$ soit continue sur $[0,T] \times [0,T]$, (2.2.1)
2. $\int_0^T K(t,t)dt < \infty$, (2.2.2)

Price a montré que l'on pouvait écrire :

$$L(Y) = \frac{1}{\sqrt{B(Y)}} e^{\frac{1}{2}\int_0^T \int_0^T H(s,t)Y_s Y_t ds dt},$$

où $B(Y)$ est un terme déterministe de biais et $H(\bullet,\bullet)$ est une fonction appelée résolvant de Fredholm de K, solution de l'équation intégrale :

$$H(s,t) + \int_0^T H(\tau,t)K(s,\tau)d\tau = K(s,t).$$

Dans le courant des années 1960, Stratanovich & Sosulin [119–121] et Schweppe [111] ont voulu généraliser cette formule en introduisant un processus stochastique noté $\widehat{X_1}$, qui est une fonction du passé de Y, c'est-à-dire que $\widehat{X_1}(t)$ dépend uniquement des valeurs $\{Y_s, s < t\}$. Cependant, le calcul de $\widehat{X_1}$ n'est la plupart du temps pas réalisable hormis par approximations, et ces approches souffrent d'un problème pratique.

Il est possible de s'affranchir de la condition d'indépendance entre le signal et le bruit. Dans ce cas, on écrira $Cov(Y_s, Y_t) = \delta_s(t) + K(s,t)$, où :

$$K(s,t) = \mathbb{E}(X_s X_t) + \mathbb{E}(X_s \epsilon_t) + \mathbb{E}(\epsilon_s X_t).$$

Dans ce cas, la fonction K reste symétrique mais n'est plus forcément définie positive au sens de l'équation (2.1.1). Pour pallier ce problème, il nous faut supposer que :

1. la fonction $(s,t) \mapsto Cov(Y_s, Y_t)$ est définie positive sur $[0,T] \times [0,T]$, (2.2.3)
2. $\int_0^T \int_0^T K^2(s,t)dsdt < \infty$. (2.2.4)

2. Il s'agit bien d'un processus, en tant que dérivée de Radon-Nikodym de deux processus.

Remarquons que les conditions (2.2.1) et (2.2.2) ci-dessus impliquent les conditions (2.2.3) et (2.2.4). Sous ces conditions, une nouvelle expression du rapport de vraisemblance a été donnée par Shepp [113] en 1966, sous la forme :

$$\mathrm{L}(\mathrm{Y}) = \frac{1}{\sqrt{\mathrm{C}(\mathrm{Y})}} e^{\mathrm{J}(\mathrm{Y})} e^{\int_0^{\mathrm{T}} \int_0^{\mathrm{T}} \mathrm{H}(s,t)\mathrm{K}(s,t) ds dt},$$

où C(Y) est une fonction déterministe et J(Y) fait intervenir l'intégrale double de Wiener centrée. Dans un rapport de 1969, Kailath [56] a montré l'expression suivante :

$$\mathrm{L}(\mathrm{Y}) = e^{\int_0^{\mathrm{T}} \widehat{\mathrm{X}_1}(t)\mathrm{Y}_t dt - \frac{1}{2}\int_0^{\mathrm{T}} \widehat{\mathrm{X}_1}(t)^2 dt},$$

où $\int$ désigne l'intégrale d'Itô et $\widehat{\mathrm{X}_1}(t) = \mathbb{E}(\mathrm{X}_t | \{\mathrm{Y}_s, s < t\})$. Des détails sur l'intégrale d'Itô se trouvent par exemple dans le livre de Doob [31].

Toutes ces formules ne sont pas facilement explicitables et restent donc très peu utilisées en pratique. Remarquons simplement que dans cette dernière formule, si X est déterministe et égal à la fonction m, alors $\widehat{\mathrm{X}_1} = m$ et on retrouve la formule du rapport de vraisemblance pour le problème de test suivant :

(H0) $\mathrm{Y} = \epsilon$,

(H1) $\mathrm{Y} = m + \epsilon$.

2.2.3 Travaux précurseurs pour un bruit gaussien quelconque

Supposons dans cette sous-section que ϵ est le processus gaussien quelconque $\mathrm{P}_{0,\mathrm{R}}$ et que X est déterministe et égal à la fonction continue m. Dans la littérature, une première approche due à Grenander [46] a été de considérer les développements de Karhunen-Loève. A condition que la fonction R soit continue sur $[0,\mathrm{T}] \times [0,\mathrm{T}]$ et que $\int_0^{\mathrm{T}} \int_0^{\mathrm{T}} \mathrm{R}^2(s,t) ds dt < \infty$, on sait que l'on peut trouver des valeurs propres $(\lambda_k)_{k\geq 1}$ et des fonctions propres $(\psi_k)_{k\geq 1}$ vérifiant pour tout entier $k \geq 1$:

$$\int_0^{\mathrm{T}} \mathrm{R}(s,t)\psi_k(s) ds = \lambda_k \psi_k(t),$$

et on a alors la décomposition suivante :

$$\epsilon_t = \sum_{k=1}^{\infty} \epsilon_k \psi_k(t),$$

$$\epsilon_k = \int_0^{\mathrm{T}} \epsilon_t \psi_k(t) dt.$$

La convergence ci-dessus est une convergence en moyenne quadratique pour chaque $t \in [0,\mathrm{T}]$. Les coefficients ϵ_k sont des variables aléatoires non corrélées, de moyenne nulle et de variance λ_k. Il est alors possible de déduire :

$$\mathrm{L}(\mathrm{Y}) = e^{\sum_{k=1}^{\infty} \frac{m_k \mathrm{Y}_k}{\lambda_k} - \frac{1}{2}\sum_{k=1}^{\infty} \frac{m_k^2}{\lambda_k}},$$

où $m_k = \int_0^{\mathrm{T}} m(t)\psi_k(t) dt$ et $\mathrm{Y}_k = \int_0^{\mathrm{T}} \mathrm{Y}_t \psi_k(t) dt$. En posant de plus $a(t) = \sum_{k=1}^{\infty} \frac{m_k}{\lambda_k}\psi_k(t)$, l'expression peut également s'écrire :

$$\mathrm{L}(\mathrm{Y}) = e^{\int_0^{\mathrm{T}} a(t)\mathrm{Y}_t dt - \frac{1}{2}\int_0^{\mathrm{T}} a(t)m(t) dt}.$$

En revanche, même si la fonction a est solution d'une équation difficile, celle-ci est souvent peu explicitable. La théorie des espaces de Hilbert à noyau reproduisant (RKHS) permet de contourner de nombreux problèmes. Nous allons voir en quoi cette théorie offre un cadre de travail plus simple pour expliciter des densités de processus gaussiens.

2.2.4 Espaces de Hilbert à noyau reproduisant

Commençons par donner quelques définitions et propriétés générales sur les espaces RKHS, abréviation de *Reproducing Kernel Hilbert Space* en anglais. Rappelons avant tout qu'un espace vectoriel H, muni d'un produit scalaire $(\bullet,\bullet)_{\mathrm{H}}$, est un espace de Hilbert si c'est de plus un espace complet pour la norme induite par le produit scalaire, c'est-à-dire la norme définie par $||u||_{\mathrm{H}} = \sqrt{(u,u)_{\mathrm{H}}}$.

Un théorème énoncé par Moore [85] en 1935 et démontré, entre autres, par Aronszajn [5] en 1950 justifie l'existence des espaces RKHS et peut se résumer de la façon suivante :

Théorème 2.1 (Moore, 1935)
Soit K *la fonction de covariance d'un processus gaussien sur* $[0,\mathrm{T}]$. *Alors il existe un unique espace de Hilbert, que l'on note* H(K) *et que l'on appelle espace de Hilbert à noyau reproduisant et de noyau* K, *défini comme l'espace des fonctions réelles* f *sur* $[0,\mathrm{T}]$ *vérifiant :*

(i) $\forall s \in [0,\mathrm{T}]$, $\mathrm{K}(\bullet,s) \in \mathrm{H}(\mathrm{K})$, *où* $\mathrm{K}(\bullet,s)$ *est la fonction* $s' \longmapsto \mathrm{K}(s',s)$,

(ii) $\forall t \in [0,\mathrm{T}]$, $\forall f \in \mathrm{H}(\mathrm{K})$, $f(t) = (f, \mathrm{K}(\bullet,t))_{\mathrm{K}}$.

$(\bullet,\bullet)_{\mathrm{K}}$ *désigne le produit scalaire dans l'espace* H(K).

Remarque 2.1
La propriété (ii) est dite propriété reproduisante et justifie le nom de la théorie RKHS.

Une question naturelle est alors la suivante : comment expliciter le produit scalaire ? Pour trouver un produit scalaire, le plus simple reste de le construire. A chaque fonction de covariance étant associé un unique espace RKHS, on peut dans certains cas proposer un produit scalaire et montrer qu'il vérifie les propriétés (i) et (ii) précédentes ; l'espace étant unique, il s'agira alors du produit scalaire de H(K). Kailath, Geesey & Weinert [57] ou encore Weinert [135] proposent des écritures de produits scalaires pour différentes fonctions de covariance. Les auteurs y décrivent également de nombreux espaces RKHS. Précisons qu'il est aussi possible d'approcher numériquement un produit scalaire. Nous le verrons à la fin de cette sous-section.

Revenons à présent à la théorie de traitement du signal. Parzen [96] a démontré le résultat suivant :

Théorème 2.2 (Parzen, 1963)
$\frac{d\mathrm{P}_{m,\mathrm{K}}}{d\mathrm{P}_{0,\mathrm{K}}}$ *existe si, et seulement si,* $m \in \mathrm{H}(\mathrm{K})$. *Dans ce cas, il est possible de montrer au travers l'utilisation des lois fini-dimensionnelles que l'on aboutit à l'expression suivante :*

$$\frac{d\mathrm{P}_{m,\mathrm{K}}}{d\mathrm{P}_{0,\mathrm{K}}}(\mathrm{Y}) = e^{(\mathrm{Y},m)_{\mathrm{K}} - \frac{1}{2}(m,m)_{\mathrm{K}}},$$

en supposant que K *soit une fonction faiblement continue sur* $[0,\mathrm{T}]\times[0,\mathrm{T}]$. *Une fonction* K *est dite faiblement continue sur* $[0,\mathrm{T}]\times[0,\mathrm{T}]$ *si :*

$$\forall t \in [0,\mathrm{T}],\ \mathrm{K}(\bullet,t) \text{ est continue sur } [0,\mathrm{T}], \tag{2.2.5a}$$

$$\begin{aligned}\forall\, t \in [0,\mathrm{T}],\ &\textit{il existe une boule ouverte } \mathrm{S}(t) \textit{ contenant } t \\ &\textit{et une constante } \mathrm{M}(t) \textit{ telle que pour tout } t' \in \mathrm{S}(t),\ \mathrm{K}(t',t') \leq \mathrm{M}(t).\end{aligned} \tag{2.2.5b}$$

On remarque dans cette expression que $(\mathrm{Y}, m)_{\mathrm{K}}$ définit un produit scalaire entre le processus stochastique Y et une fonction connue de l'espace H(K). La notation $(\mathrm{Y}, m)_{\mathrm{K}}$ ne définit donc pas réellement un produit scalaire et cache l'expression d'une intégrale stochastique. Pour définir une telle notation, il est nécessaire d'établir une correspondance entre les éléments de H(K) et ceux d'un autre espace de Hilbert que nous allons définir.

Définition 2.1
Soit $\mathrm{Y} = (\mathrm{Y}_t)_{t\in[0,\mathrm{T}]}$ *un processus stochastique. On définit* $\mathrm{L}_2(\mathrm{Y})$ *comme l'espace des variables aléatoires qui sont des combinaisons linéaires finies des variables aléatoires* Y_t *ou bien des limites en moyenne quadratique de telles combinaisons linéaires.*

Intuitivement, on peut interpréter $\mathrm{L}_2(\mathrm{Y})$ *comme l'espace de toutes les fonctions linéaires du processus* Y.

Définition 2.2
Soit $f = \left(f_t\right)_{t\in[0,\mathrm{T}]}$ *une famille de vecteurs d'un espace de Hilbert* H. *On définit* $\mathrm{L}_2(f)$ *comme l'espace des vecteurs qui sont des combinaisons linéaires finies des vecteurs* f_t *ou bien des limites de telles combinaisons linéaires.*

Une propriété importante reprise par Parzen [94] est la suivante :

Théorème 2.3 (Parzen, 1961)
Soient $\mathrm{Y} = (\mathrm{Y}_t)_{t\in[0,\mathrm{T}]}$ *un processus stochastique de fonction de covariance* K *et* $f = \left(f_t\right)_{t\in[0,\mathrm{T}]}$ *une famille de vecteurs d'un espace de Hilbert* H. *On dit que la famille* f *est une représentation du processus stochastique* Y *si pour tous* $s, t \in [0,\mathrm{T}]$, *on a :*

$$(f_s, f_t)_{\mathrm{H}} = \mathrm{K}(s,t).$$

Il existe alors une congruence ψ *de* $\mathrm{L}_2(f)$ *sur* $\mathrm{L}_2(\mathrm{Y})$ *vérifiant pour tout* $t \in [0,\mathrm{T}]$, $\psi(f_t) = \mathrm{Y}_t$, *et toute variable aléatoire* $\mathrm{U} \in \mathrm{L}_2(\mathrm{Y})$ *peut s'écrire sous la forme* $\mathrm{U} = \psi(g)$, *pour un unique vecteur* $g \in \mathrm{L}_2(f)$. *Cette congruence est un isomorphisme et préserve les produits scalaires :*

$$(f_s, f_t)_{\mathrm{L}_2(f)} = (\psi(f_s), \psi(f_t))_{\mathrm{L}_2(\mathrm{Y})}.$$

En particulier, nous remarquons que $\mathrm{K}(s,t) = (\mathrm{K}(\bullet, s), \mathrm{K}(\bullet, t))_{\mathrm{K}}$ d'après la propriété reproduisante (ii) du théorème 2.1. Ainsi, la famille de fonctions $(\mathrm{K}(\bullet, t))_{t\in[0,\mathrm{T}]}$ est une représentation du processus stochastique Y. D'après le théorème 2.3, il existe une congruence ψ de $\mathrm{L}_2\left((\mathrm{K}(\bullet, t))_{t\in[0,\mathrm{T}]}\right)$ sur $\mathrm{L}_2(\mathrm{Y})$. De plus, il est connu que $\mathrm{L}_2\left((\mathrm{K}(\bullet, t))_{t\in[0,\mathrm{T}]}\right) = \mathrm{H}(\mathrm{K})$ d'après Parzen [94]. Ainsi, nous déduisons que la congruence ψ est à valeurs de H(K) dans $\mathrm{L}_2(\mathrm{Y})$. Si $g \in \mathrm{H}(\mathrm{K})$, alors $(\mathrm{Y}, g)_{\mathrm{K}}$, produit scalaire entre un élément $g \in \mathrm{H}(\mathrm{K})$ et le processus stochastique Y, est l'image $\mathrm{U} \in \mathrm{L}_2(\mathrm{Y})$ correspondant à la congruence définie précédemment, c'est-à-dire $\mathrm{U} = \psi(g)$. Il s'agit donc d'une variable aléatoire de l'espace $\mathrm{L}_2(\mathrm{Y})$. On obtient également les deux propriétés suivantes pour tous réel $t \in [0,\mathrm{T}]$ et fonctions $f, g \in \mathrm{H}(\mathrm{K})$:

(i) $(\mathrm{Y}, \mathrm{K}(\bullet, t))_{\mathrm{K}} = \mathrm{Y}_t$,

(ii) $(f, g)_{\mathrm{K}} = \mathbb{E}((\mathrm{Y}, f)_{\mathrm{K}}(\mathrm{Y}, g)_{\mathrm{K}})$,

Un autre théorème intéressant est le théorème de représentation intégrale.

Définition 2.3 (Famille aléatoire orthogonale)
Soit $(\mathrm{Q}, \mathscr{B}, \mu)$ *un espace mesuré. On dit qu'une famille de variables aléatoires* $(\mathrm{Z}(\mathrm{B}))_{\mathrm{B}\in\mathscr{B}}$ *de* Q *est une famille aléatoire orthogonale de noyau de covariance* μ *si pour tous* $\mathrm{B}_1, \mathrm{B}_2 \in \mathscr{B}$*, on a :*

$$\mathbb{E}(\mathrm{Z}(\mathrm{B}_1)\mathrm{Z}(\mathrm{B}_2)) = \mu(\mathrm{B}_1 \cap \mathrm{B}_2).$$

Définition 2.4
Soit $(\mathrm{Q}, \mathscr{B}, \mu)$ *un espace mesuré. On définit* $\mathrm{L}_2(\mathrm{Q})$ *comme l'espace de Hilbert de toutes les fonctions* f $\mathscr{B}$*-mesurables définies sur* Q*, et vérifiant de plus :*

$$\int_{\mathrm{Q}} f^2 d\mu < \infty.$$

Le produit scalaire sur l'espace $\mathrm{L}_2(\mathrm{Q})$ *est défini par :*

$$\big(f, g\big)_{\mathrm{L}_2(\mathrm{Q})} = \int_{\mathrm{Q}} f g d\mu.$$

Théorème 2.4 (Parzen, 1961)
Soit $\mathrm{Y} = (\mathrm{Y}_t)_{t\in[0,\mathrm{T}]}$ *un processus stochastique de fonction de covariance* K*. S'il existe un espace mesuré* $(\mathrm{Q}, \mathscr{B}, \mu)$ *et une famille de fonctions* $f = \big(f_t\big)_{t\in[0,\mathrm{T}]}$ *de* Q*, vérifiant de plus :*

$$\mathrm{K}(s,t) = \int_{\mathrm{Q}} f_s f_t d\mu,$$

alors la famille f *est une représentation du processus stochastique* Y*.*

Si de plus, la famille f *engendre l'espace* $\mathrm{L}_2(\mathrm{Q})$*, alors il existe une famille aléatoire orthogonale* $(\mathrm{Z}(\mathrm{B}))_{\mathrm{B}\in\mathscr{B}}$ *de* Q *et de noyau de covariance* μ*, telle que :*

$$\mathrm{Y}_t = \int_{\mathrm{Q}} f_t d\mathrm{Z},$$

et toute variable aléatoire $\mathrm{U} \in \mathrm{L}_2(\mathrm{Y})$ *peut s'écrire sous la forme :*

$$\mathrm{U} = \int_{\mathrm{Q}} g d\mathrm{Z},$$

pour un unique vecteur $g \in \mathrm{L}_2(\mathrm{Q})$*.*

D'après ce théorème, nous voyons bien que l'expression $(\mathrm{Y}, g)_{\mathrm{K}}$ est une variable aléatoire pouvant s'écrire comme une intégrale stochastique. Terminons cette sous-section en citant deux résultats importants, correspondant aux théorèmes 6E et 9B de Parzen [93] :

Définition 2.5 (Fonction non singulière)
Une fonction K *définie sur* $[0,\mathrm{T}] \times [0,\mathrm{T}]$ *est dite non singulière sur tout sous-ensemble de* $[0,\mathrm{T}] \times [0,\mathrm{T}]$ *si quels que soient l'entier naturel non nul* $n \in \mathbb{N}$ *et les nombres réels* $t_1, \ldots, t_n \in [0,\mathrm{T}]$*, la matrice* $\big[\mathrm{K}(t_i, t_j)\big]_{1\leq i,j\leq n}$ *est inversible.*

Théorème 2.5 (Parzen, 1959)
Soit K *une fonction de covariance d'un processus stochastique. Supposons que* K *est une fonction faiblement continue sur* $[0,\mathrm{T}] \times [0,\mathrm{T}]$ *et qu'elle est non singulière sur tout sous-ensemble de*

$[0,\mathrm{T}]\times[0,\mathrm{T}]$. *Nous avons alors pour toutes fonctions* $f,g\in \mathrm{H}(\mathrm{K})$ *et toute suite* $(t_i)_{i\geq 1}$ *de* $[0,\mathrm{T}]$ *dense dans* $[0,\mathrm{T}]$ *:*

$$\lim_{n\to\infty} f_n^\top \mathrm{K}_n^{-1} g_n = \big(f,g\big)_{\mathrm{K}},$$

où $f_n=\big(f(t_1),\ldots,f(t_n)\big)$, $\mathrm{K}_n=\big[\mathrm{K}\big(t_i,t_j\big)\big]_{1\leq i,j\leq n}$ *et* $g_n=\big(g(t_1),\ldots,g(t_n)\big)$.

Théorème 2.6 (Parzen, 1959)
Soit $\mathrm{Y}=(\mathrm{Y}_t)_{t\in[0,\mathrm{T}]}$ *un processus stochastique de fonction moyenne nulle et de fonction de covariance* K. *Supposons que* K *est une fonction faiblement continue sur* $[0,\mathrm{T}]\times[0,\mathrm{T}]$ *et qu'elle est non singulière sur tout sous-ensemble de* $[0,\mathrm{T}]\times[0,\mathrm{T}]$. *Nous avons alors pour toute fonction* $g\in\mathrm{H}(\mathrm{K})$ *et toute suite* $(t_i)_{i\geq 1}$ *de* $[0,\mathrm{T}]$ *dense dans* $[0,\mathrm{T}]$ *:*

$$\lim_{n\to\infty} \mathrm{Y}_n^\top \mathrm{K}_n^{-1} g_n = \big(\mathrm{Y},g\big)_{\mathrm{K}} = \psi(g),$$

où $\mathrm{Y}_n=(\mathrm{Y}(t_1),\ldots,\mathrm{Y}(t_n))$, $\mathrm{K}_n=\big[\mathrm{K}\big(t_i,t_j\big)\big]_{1\leq i,j\leq n}$ *et* $g_n=\big(g(t_1),\ldots,g(t_n)\big)$ *et* ψ *est la congruence définie précédemment. La convergence existe en moyenne quadratique mais aussi presque sûrement.*

Ces théorèmes sont très importants en pratique car ils fournissent une méthode d'approximation des produits scalaires. A défaut de pouvoir expliciter le produit scalaire, la méthode d'approximation la plus courante consiste à approcher $(f,g)_{\mathrm{K}}$ par $f_n^\top \mathrm{K}_n^{-1} g_n$, où f_n, K_n et g_n sont des versions discrétisées des fonctions f, K et g en des temps d'observations donnés. En 1961, Parzen [94] propose une méthode itérative pour estimer les produits scalaires, lorsque la fonction de covariance est connue analytiquement ou seulement numériquement. Sa méthode consiste en une première estimation H_0, puis à construire une suite de fonctions H_n par récurrence. Cette suite H_n est alors telle que :

$$\lim_{n\to+\infty} \mathbb{E}\left(\left|(\mathrm{Y},g)_{\mathrm{K}} - \int_0^{\mathrm{T}} \mathrm{H}_n(t)\mathrm{X}_t\,dt\right|^2\right)=0,$$

ce qui permet d'obtenir une estimation de $(\mathrm{Y},g)_{\mathrm{K}}$. En 1965, Weiner [134] développe une autre méthode itérative, encore utilisée à l'heure actuelle. Cependant, la complexité de la méthode est importante. Finalement, c'est plus récemment en 2009 que Oya et al. [92] proposent une nouvelle approche pour évaluer numériquement un produit scalaire. Dans un premier temps, leur méthode requiert de résoudre un problème aux valeurs propres "généralisé", puis à faire usage des valeurs et vecteurs propres trouvés afin d'obtenir l'estimation souhaitée.

En comparaison aux développements de Karhunen-Loève et aux travaux précurseurs, la théorie RKHS offre donc un cadre de travail plus souple tout en faisant intervenir des quantités facilement mesurables numériquement.

2.3 Processus gaussien a posteriori

2.3.1 Cas d'une seule observation

Dans toute cette section, nous adoptons les notations suivantes :

$t^n=(t_1,\ldots,t_n)$ lorsque $(t_1,\ldots,t_n)$ est un vecteur,

$\mathrm{X}(t^n)=(\mathrm{X}(t_1),\ldots,\mathrm{X}(t_n))$ lorsque X est une fonction d'une seule variable,

$\mathrm{X}(t^n)=(\mathrm{X}_{t_1},\ldots,\mathrm{X}_{t_n})$ lorsque X est un processus stochastique.

Etant donnés deux processus gaussiens indépendants X et ϵ, supposons que l'on observe le processus $Y = X + \epsilon$. Nous souhaitons calculer la loi *a posteriori* de X sachant Y.

Dans la littérature, et O'Hagan [91] en particulier, le processus X est supposé être observé en différents temps d'observation. Seule la loi *a priori* de X est infini-dimensionnelle. De plus, il n'est pas possible de généraliser la preuve de O'Hagan dans le cas d'une vraisemblance infini-dimensionnelle. En effet, l'utilisation de vecteurs n'étant plus possible, on ne peut espérer écrire de densités à cause de l'absence de mesure de Lebesgue en dimension infinie.

Nous généralisons dans la suite le théorème de O'Hagan au cas où la loi *a priori* ainsi que la vraisemblance sont de dimension infinie. D'un point de vue théorique, notre travail requiert des densités de processus gaussien, que nous calculons à l'aide de la théorie RKHS.

Dans cette section, nous notons $\mathcal{T}$ l'ensemble de définition de tous les processus, de sorte que $X = (X_t)_{t\in\mathcal{T}}$, $Y = (Y_t)_{t\in\mathcal{T}}$ et $\epsilon = (\epsilon_t)_{t\in\mathcal{T}}$. On suppose que $\mathcal{T}$ est un espace métrique séparable et que W et C sont deux fonctions réelles de deux variables définies sur $\mathcal{T}\times\mathcal{T}$, définies positives, symétriques et faiblement continues sur $\mathcal{T}\times\mathcal{T}$.

Nous supposons également que les fonctions W et C sont non singulières sur tout sous-ensemble de $\mathcal{T}\times\mathcal{T}$.

Lemme 2.1
La fonction $W + C$ *est symétrique, définie positive et faiblement continue sur* $\mathcal{T}\times\mathcal{T}$. *De plus, elle est non singulière sur tout sous-ensemble fini de* $\mathcal{T}\times\mathcal{T}$.

PREUVE. D'après les équations (2.1.1), (2.1.2) et (2.2.5), la fonction $W + C$ est clairement symétrique, définie positive et faiblement continue sur $\mathcal{T}\times\mathcal{T}$.

Soient un entier naturel non nul $n \in \mathbb{N}$ et $t_1,\dots,t_n \in \mathcal{T}$. Par hypothèse, les matrices $\left[W(t_i,t_j)\right]_{1\le i,j\le n}$ et $\left[C(t_i,t_j)\right]_{1\le i,j\le n}$ sont symétriques, positives et inversibles. Dans un premier temps, montrons que pour toutes matrices symétriques et positives X et Y, on a $\mathrm{rg}(X+Y) \ge \mathrm{rg}Y$.

Considérons d'abord le cas où Y est une matrice diagonale. Sans perte de généralités, nous pouvons écrire pour des réels $\alpha_1,\dots,\alpha_r \in \mathbb{R}$:

$$Y = \begin{pmatrix} \alpha_1 & & & & & \\ & \ddots & & & 0 & \\ & & \alpha_r & & & \\ & & & 0 & & \\ & 0 & & & \ddots & \\ & & & & & 0 \end{pmatrix},$$

de sorte que $\mathrm{rg}Y = r$. En conséquence, nous pouvons supposer que tous les α_i sont strictement positifs. Clairement, la matrice $\left[X_{ij}\right]_{1\le i,j\le r} + \begin{pmatrix} \alpha_1 & & 0 \\ & \ddots & \\ 0 & & \alpha_r \end{pmatrix}$ est définie positive et ainsi nous avons :

$$\mathrm{rg}\left(\left[X_{ij}\right]_{1\le i,j\le r} + \begin{pmatrix} \alpha_1 & & 0 \\ & \ddots & \\ 0 & & \alpha_r \end{pmatrix}\right) = r,$$

donc

$$\operatorname{rg}(\mathrm{X}+\mathrm{Y}) \geq \operatorname{rg}\mathrm{Y} = r.$$

Finalement, considérons le cas où Y est une matrice symétrique et positive quelconque. Comme elle est positive, il existe une matrice orthogonale P telle que $\mathrm{P}^\top\mathrm{YP}$ est une matrice diagonale. Ainsi, comme $\mathrm{P}^\top$ et P sont inversibles, on obtient :

$$\operatorname{rg}(\mathrm{X}+\mathrm{Y}) = \operatorname{rg}(\mathrm{P}^\top(\mathrm{X}+\mathrm{Y})\mathrm{P}) = \operatorname{rg}(\mathrm{P}^\top\mathrm{XP} + \mathrm{P}^\top\mathrm{YP}).$$

Comme la matrice $\mathrm{P}^\top\mathrm{XP}$ est à nouveau positive et que la matrice $\mathrm{P}^\top\mathrm{YP}$ est diagonale, on peut faire usage de ce qui précède.

En conclusion, $\operatorname{rg}\left[(\mathrm{W}+\mathrm{C})(t_i,t_j)\right]_{1\leq i,j\leq n} \geq n$ donc $\left[(\mathrm{W}+\mathrm{C})(t_i,t_j)\right]_{1\leq i,j\leq n}$ est inversible. ■

Théorème 2.7
Soient $\mathrm{X} \sim \mathrm{P}_{m,\mathrm{W}}$ *et* $\epsilon \sim \mathrm{P}_{0,\mathrm{C}}$. *Supposons que* X *et* ϵ *sont indépendants et notons* $\mathrm{Y} = \mathrm{X} + \epsilon$. *Enfin, supposons que les trajectoires de* Y *sont continues p.s.. Alors, la loi a posteriori* X|Y *est également un processus gaussien* $\mathrm{P}_{m^\star,\mathrm{W}^\star}$, *donné par :*

$$m^\star(t) = m(t) + (\mathrm{W}(\bullet,t), \mathrm{Y}-m)_{\mathrm{W}+\mathrm{C}}, \tag{2.3.1a}$$

$$\mathrm{W}^\star(t,t') = \mathrm{W}(t,t') - \left(\mathrm{W}(\bullet,t), \mathrm{W}(\bullet,t')\right)_{\mathrm{W}+\mathrm{C}}, \tag{2.3.1b}$$

où $(\bullet,\bullet)_{\mathrm{W}+\mathrm{C}}$ *désigne le produit scalaire dans l'espace* $\mathrm{H}(\mathrm{W}+\mathrm{C})$.

Remarque 2.2
Supposer Y *à avoir des trajectoires continues presque sûrement n'est pas une hypothèse restrictive en soi. En effet, c'est le cas des processus Brownien, d'Ornstein-Uhlenbeck et de nombreux autres processus intéressants en pratique.*

Remarque 2.3
Lorsque nous avons débuté ce travail de thèse, nous ne connaissions pas l'existence de résultats similaires [21, 34, 128]. Cependant, la preuve ne nous ayant pas parue triviale, nous avons pris la peine de démontrer notre théorème ex nihilo. Notons toutefois que, par chance, notre théorème diffère des résultats existants, en ce sens que :
- *Cox [21] se place directement dans un espace de Banach séparable et Van der Vaart [128] dans un espace de Hilbert, alors que nous ne faisons aucune hypothèse a priori sur l'espace des trajectoires de nos processus,*
- *Driscoll [34] ne considère pas de répétitions, se plaçant uniquement dans une optique de traitement du signal.*

PREUVE. Fixons $t^n = (t_1,\ldots,t_n) \in \mathcal{T}^n$. Pour commencer, cherchons la loi conditionnelle de :

$$\mathrm{X}(t^n)|\mathrm{Y}. \tag{2.3.2}$$

La loi de (2.3.2) est entièrement caractérisée par sa fonction caractéristique, c'est-à-dire la fonction $\psi : u \mapsto \mathbb{E}(\exp(iu^\top \mathrm{X}(t^n))|\mathrm{Y})$, où $u = (u_1,\ldots,u_n) \in \mathbb{R}^n$. On écrit $\mathbb{E}(\bullet|\mathrm{Y}) = \mathbb{E}(\bullet|\mathscr{B}_\mathrm{Y})$ avec $\mathscr{B}_\mathrm{Y} = \sigma(\mathrm{Y}_t, t \in \mathcal{T}) = \sigma(\mathrm{Y})$.

Soit $\tau^m = (\tau_1,\ldots,\tau_m) \in \mathcal{T}^m$, où $(\tau_m)_{m\geq 1}$ est une suite quelconque de $\mathcal{T}$ qui est dense dans $\mathcal{T}$. D'après le résultat de O'Hagan [91], il est connu que :

$$\mathrm{X}(t^n)|\mathrm{Y}(\tau^m) \sim \mathcal{N}\left(m_m^\star(t^n), \mathrm{W}_m^\star(t^n,t^n)\right), \tag{2.3.3}$$

avec :

$$m^\star_m(t) = m(t) + \mathrm{W}_m(t)^\top (\mathrm{W}_m + \mathrm{C}_m)^{-1}(\mathrm{Y}(\tau^m) - m(\tau^m)), \tag{2.3.4a}$$

$$\mathrm{W}^\star_m(t,t') = \mathrm{W}(t,t') - \mathrm{W}_m(t)^\top (\mathrm{W}_m + \mathrm{C}_m)^{-1}\mathrm{W}_m(t'), \tag{2.3.4b}$$

où nous écrivons $\mathrm{W}_m(t) = (\mathrm{W}(\tau_1,t),\ldots,\mathrm{W}(\tau_m,t))$, $\mathrm{W}_m = \left[\mathrm{W}(\tau_i,\tau_j)\right]_{1\le i,j\le m}$ et de même $\mathrm{C}_m = \left[\mathrm{C}(\tau_i,\tau_j)\right]_{1\le i,j\le m}$. En conséquence, comme la loi de (2.3.3) est une normale multivariée, nous pouvons calculer sa fonction caractéristique comme suit :

$$\psi_m(u) = \mathbb{E}(\exp(iu^\top \mathrm{X}(t^n))|\mathrm{Y}(\tau^m)) = \exp\left(iu^\top m^\star_m(t^n) - \frac{1}{2}u^\top \mathrm{W}^\star_m(t^n,t^n)u\right). \tag{2.3.5}$$

A partir de maintenant, étudions la limite $\lim_{m\to\infty}\psi_m(u)$ dans l'équation (2.3.5) et montrons qu'elle est égale à $\psi(u)$ p.s..

D'une part, d'après les résultats de la théorie RKHS, nous savons que $\mathrm{W}_m(t)^\top(\mathrm{W}_m+\mathrm{C}_m)^{-1}\mathrm{W}_m(t')$ converge lorsque $m\to\infty$ si, et seulement si, $\mathrm{W}(\bullet,t)\in \mathrm{H}_{\mathrm{W}+\mathrm{C}}$ et $\mathrm{W}(\bullet,t')\in \mathrm{H}_{\mathrm{W}+\mathrm{C}}$. D'après le théorème 12 de Berlinet et al. [6], nous avons toujours $\mathrm{H}_{\mathrm{W}}\subset \mathrm{H}_{\mathrm{W}+\mathrm{C}}$, car $(\mathrm{W}+\mathrm{C})-\mathrm{W}=\mathrm{C}$ est une fonction de covariance. Ainsi, comme pour chaque $t\in\mathscr{T}$ on a $\mathrm{W}(\bullet,t)\in\mathrm{H}_{\mathrm{W}}$, on obtient $\mathrm{W}(\bullet,t)\in\mathrm{H}_{\mathrm{W}+\mathrm{C}}$. D'après le lemme 2.1 et le théorème 2.5,

$$\mathrm{W}_m(t)^\top(\mathrm{W}_m+\mathrm{C}_m)^{-1}\mathrm{W}_m(t') \underset{m\to\infty}{\longrightarrow} (\mathrm{W}(\bullet,t),\mathrm{W}(\bullet,t'))_{\mathrm{W}+\mathrm{C}}.$$

Donc $\mathrm{W}^\star_m(t^n,t^n)$ converge vers $\mathrm{W}^\star(t^n,t^n)$ lorsque $m\to\infty$, et qui est donné par l'équation (2.3.1).

D'autre part, étudions à présent la limite $\lim_{m\to\infty}\mathrm{W}_m(t)^\top(\mathrm{W}_m+\mathrm{C}_m)^{-1}(\mathrm{Y}(\tau^m)-m(\tau^m))$. En premier lieu, $\mathrm{Y}-m\sim \mathrm{P}_{0,\mathrm{W}+\mathrm{C}}$. Il est facile de le vérifier car nous pouvons écrire pour chaque entier non nul n et tout $t^n=(t_1,\ldots,t_n)\in\mathscr{T}^n$, $((\mathrm{Y}-m)(t_1),\ldots,(\mathrm{Y}-m)(t_n)) = (\mathrm{X}(t^n)+\epsilon(t^n)-m(t^n))$, qui est une normale multivariée $\mathscr{N}(0,\left[(\mathrm{W}+\mathrm{C})(t_i,t_j)\right]_{1\le i,j\le n})$ en raison de l'indépendance entre X et ϵ.

En faisant usage à nouveau du lemme 2.1 et du théorème 2.6, nous savons que $\phi_m = \mathrm{W}_m(t)^\top(\mathrm{W}_m+\mathrm{C}_m)^{-1}(\mathrm{Y}(\tau^m)-m(\tau^m))$ converge vers $\phi(\mathrm{W}(\bullet,t))$, où ϕ désigne la congruence de $\mathrm{H}_{\mathrm{W}+\mathrm{C}}$ sur $\mathrm{L}_2((\mathrm{Y}_t-m_t)_{t\in\mathscr{T}})$ telle que $\phi((\mathrm{W}+\mathrm{C})(\bullet,t)) = \mathrm{Y}_t - m_t$. Cette convergence existe en tant que limite en moyenne quadratique mais aussi presque sûrement, ce qui signifie que pour presque toute réalisation de Y, on a $\phi_m \underset{m\to\infty}{\longrightarrow} \phi(\mathrm{W}(\bullet,t))$.

Ecrivons maintenant $(\mathrm{W}(\bullet,t),\mathrm{Y}-m)_{\mathrm{W}+\mathrm{C}}$ au lieu de $\phi(\mathrm{W}(\bullet,t))$. Alors $m^\star_m(t^n)\underset{m\to\infty}{\longrightarrow} m^\star(t^n)$ p. s. et

$$\mathbb{E}(\exp(iu^\top \mathrm{X}(t^n))|\mathrm{Y}(\tau^m)) \underset{m\to\infty}{\longrightarrow} \exp\left(iu^\top m^\star(t^n) - \frac{1}{2}u^\top \mathrm{W}^\star(t^n,t^n)u\right) \text{ p.s..} \tag{2.3.6}$$

Cela conclut la première partie de la preuve. Si nous montrons maintenant que :

$$\mathbb{E}(\exp(iu^\top \mathrm{X}(t^n))|\mathrm{Y}(\tau^m)) \underset{m\to\infty}{\longrightarrow} \mathbb{E}(\exp(iu^\top \mathrm{X}(t^n))|\mathrm{Y}) \text{ p.s.}, \tag{2.3.7}$$

nous déduirons que la loi de $\mathrm{X}(t^n)|\mathrm{Y}$ est une normale multivariée, de moyenne $m^\star(t^n)$ et de covariance $\mathrm{W}^\star(t^n,t^n)$ et ainsi la loi de X|Y sera le processus gaussien $\mathrm{P}_{m^\star,\mathrm{W}^\star}$. En effet, nous savons d'après le théorème d'extension de Kolmogorov que le processus stochastique est entièrement défini par la famille de ses lois fini-dimensionnelles.

Soient $Z_m = \mathbb{E}(\exp(iu^\top X(t^n))|Y(\tau^m))$ et $Z = \mathbb{E}(\exp(iu^\top X(t^n))|Y)$. Z est intégrable et, si nous écrivons $\mathscr{B}_m = \sigma(Y(\tau_1),\ldots,Y(\tau_m))$, Z_m est une martingale par rapport à $\mathscr{B}_m$ d'après la suite d'égalités suivantes :

$$\begin{aligned} \forall k \geq m, \quad \mathbb{E}(Z_k|\mathscr{B}_m) &= \mathbb{E}(\mathbb{E}(\exp(iu^\top X(t^n))|\mathscr{B}_k)|\mathscr{B}_m), \\ &= \mathbb{E}(\exp(iu^\top X(t^n))|\mathscr{B}_m), \\ &= Z_m. \end{aligned} \tag{2.3.8}$$

De plus, Z est une martingale fermée dans le sens où :

$$\begin{aligned} \mathbb{E}(Z|\mathscr{B}_m) &= \mathbb{E}(\mathbb{E}(\exp(iu^\top X(t^n))|Y)|\mathscr{B}_m), \\ &= \mathbb{E}(\exp(iu^\top X(t^n))|\mathscr{B}_m), \\ &= Z_m, \end{aligned} \tag{2.3.9}$$

car $\mathscr{B}_m \subset \mathscr{B}_Y$. Ainsi, d'après le théorème 10.5.1 de Dudley [35], on obtient :

$$Z_n \underset{n\to\infty}{\longrightarrow} \mathbb{E}(Z|\mathscr{B}_\infty) \text{ p.s.},$$

où $\mathscr{B}_\infty$ désigne la plus petite tribu contenant tous les $\mathscr{B}_m$, c'est-à-dire $\sigma(Y_{\tau_i}, i \in \mathbb{N}\backslash\{0\})$.

Montrons à présent que $\mathscr{B}_\infty = \sigma(Y)$. D'une part, il est clair que $\mathscr{B}_\infty \subset \sigma(Y)$. D'autre part, il nous faut montrer que $\sigma(Y) \subset \mathscr{B}_\infty$. Comme $\sigma(Y)$ est générée par la famille d'ensembles $\{Y_t \in B\}$ pour tout $t \in \mathscr{T}$ et tout $B \in \mathscr{B}_\mathbb{R}$, il suffit de montrer que chaque $\{Y_t \in B\} \in \mathscr{B}_\infty$.

Soit $(\tau_i^*)_{i\geq 1}$ une sous-suite de $(\tau_i)_{i\geq 1}$ telle que :

$$\tau_i^* \underset{i\to\infty}{\longrightarrow} t.$$

D'après la définition d'une tribu, $Y_{\tau_i^*}$ est mesurable pour la tribu $\mathscr{B}_\infty$ et il en est de même pour $\lim_{i\to\infty} Y_{\tau_i^*}$. Cependant, Y ayant par hypothèse des trajectoires continues p.s., nous avons $\lim_{i\to\infty} Y_{\tau_i^*} = Y_t$ donc $Y_t \in \mathscr{B}_\infty$.

En conclusion, les expressions (2.3.6) et (2.3.7) sont identiques. Toutes les convergences ci-dessus étant presque sûres, nous concluons la preuve. ■

2.3.2 Cas de multiples observations

La seconde partie de cette section est consacrée à la généralisation du théorème 2.7 pour des processus gaussiens i.i.d. $\epsilon_1,\ldots,\epsilon_n$ centrés et de fonction de covariance C.

Lemme 2.2
Soient X, Y, Z *des processus stochastiques, à valeurs d'un espace probabilisé* $(\Omega, \mathscr{A}, \mathbb{P})$ *dans un espace polonais muni de sa tribu borélienne* $(\mathscr{X}, \mathscr{B})$. *Supposons* Z *indépendant de* X *et de* Y *(signifiant que la tribu* $\sigma(Z)$ *est indépendante de* $\sigma(X,Y)$*). Alors, il existe* $N \in \mathscr{A}$ *tel que* $\mathbb{P}(N) = 0$ *et*

$$\forall w \notin N, \mathbb{P}_X(\bullet|Y,Z)(w) = \mathbb{P}_X(\bullet|Y)(w),$$

où $\mathbb{P}_X(\bullet|Y,Z)$ *et* $\mathbb{P}_X(\bullet|Y)$ *désignent respectivement la loi conditionnelle de* X *sachant* (Y, Z) *et de* X *sachant* Y.

PREUVE. Le fait que l'espace d'arrivée des processus soit polonais assure de l'existence de $\mathbb{P}_X(\bullet|Y,Z)$ et $\mathbb{P}_X(\bullet|Y)$. Comme $\mathscr{X}$ est séparable, nous appliquons la proposition 2.1.4 de Dudley [35], ce qui permet d'affirmer qu'il existe une base d'ouverts $\theta = (\theta_i)_{i\in\mathbb{N}}$ qui génère la tribu : $\sigma(\theta) = \mathscr{B}$.

Ecrivons alors $\tau = \{\theta_{i_1} \cap \cdots \cap \theta_{i_k}, k \in \mathbb{N}, \theta_{i_j} \in \theta\}$, qui est un ensemble dénombrable. Puis, écrivons :

$$N_{i_1,\ldots,i_k} = \{w \in \Omega : \mathbb{P}_X(\theta_{i_1} \cap \cdots \cap \theta_{i_k}|Y,Z)(w) \neq \mathbb{P}_X(\theta_{i_1} \cap \cdots \cap \theta_{i_k}|Y)(w)\}.$$

D'après la propriété 9.7(k) de Williams [136], comme $\mathbb{P}_X(A|Y,Z)$ est une version de $\mathbb{E}(\mathbb{I}_X(A)|Y,Z)$, $\mathbb{P}(N_{i_1,\ldots,i_k}) = 0$. Ainsi $\mathbb{P}\left(\cup_{k\in\mathbb{N}} \cup_{i_1,\ldots,i_k} N_{i_1,\ldots,i_k}\right) = 0$.

En conséquence, en écrivant $N = \cup_{k\in\mathbb{N}} \cup_{i_1,\ldots,i_k} N_{i_1,\ldots,i_k}$, on a pour tout $w \notin N$ l'égalité $\mathbb{P}_X(D|Y,Z)(w) = \mathbb{P}_X(D|Y)(w)$ pour chaque $D \in \tau$. Or, deux mesures finies sur un π-système (qui est ici τ) étant égales sur $\sigma(\tau) = \mathscr{B}$, on obtient finalement :

$$\mathbb{P}_X(\bullet|Y,Z)(w) = \mathbb{P}_X(\bullet|Y)(w).$$

■

Ainsi, en notant respectivement $\mathscr{L}(X|Y,Z)$ et $\mathscr{L}(X|Y)$ une version de la loi conditionnelle de X sachant (Y,Z) et de X sachant Y, nous avons l'égalité :

$$\mathscr{L}(X|Y,Z) = \mathscr{L}(X|Y),$$

lorsque $Z \perp\!\!\!\perp (X,Y)$.

Lemme 2.3
Soient $\epsilon_1,\ldots,\epsilon_n$ des processus gaussiens indépendants $P_{0,C}$. Alors $(\epsilon_1 - \bar{\epsilon},\ldots,\epsilon_n - \bar{\epsilon})$ et $\bar{\epsilon}$ sont indépendants.

PREUVE. Prouvons dans un premier temps que ce résultat est vrai pour des normales multivariées indépendantes $\epsilon_1,\ldots,\epsilon_n$ suivant $\mathcal{N}(0,\widetilde{C})$. Le vecteur $(\epsilon_1 - \bar{\epsilon},\ldots,\epsilon_n - \bar{\epsilon},\bar{\epsilon})$ est une combinaison linéaire des composantes du vecteur $(\epsilon_1,\ldots,\epsilon_n)$. Par indépendance, $(\epsilon_1,\ldots,\epsilon_n)$ a une loi normale multivariée, et il en est de même pour toute combinaison linéaire des composantes du vecteur, donc $(\epsilon_1 - \bar{\epsilon},\ldots,\epsilon_n - \bar{\epsilon},\bar{\epsilon})$ a une loi normale multivariée.

En conclusion, une condition nécessaire et suffisante pour prouver que $(\epsilon_1 - \bar{\epsilon},\ldots,\epsilon_n - \bar{\epsilon})$ est indépendant de $\bar{\epsilon}$ est que pour chaque entier $i \in \{1,\ldots,n\}$, $\epsilon_i - \bar{\epsilon}$ et $\bar{\epsilon}$ sont non corrélés.

Clairement, $\mathbb{E}(\epsilon_i - \bar{\epsilon}) = 0$ et $\mathbb{E}(\bar{\epsilon}) = 0$. On en déduit ainsi les covariances suivantes :

$$\begin{aligned}
Cov(\epsilon_i - \bar{\epsilon},\bar{\epsilon}) &= \mathbb{E}((\epsilon_i - \bar{\epsilon})\bar{\epsilon}^\top) \\
&= \mathbb{E}((\epsilon_i - \bar{\epsilon})\bar{\epsilon}^\top) \\
&= \tfrac{1}{n}\textstyle\sum_{j=1}^n \mathbb{E}(\epsilon_i\epsilon_j^\top) - \tfrac{1}{n^2}\sum_{j,k=1}^n \mathbb{E}(\epsilon_j\epsilon_k^\top) \\
&= \tfrac{1}{n}\widetilde{C} - \tfrac{1}{n^2}\textstyle\sum_{i=1}^n \widetilde{C} \\
&= 0.
\end{aligned}$$

Les variables étant non corrélées, nous venons de prouver le lemme 2.3 dans le cas de lois fini-dimensionnelles. A présent, il nous faut généraliser ce résultat à la tribu entière. En réalité, il nous faut montrer que $\sigma(\epsilon_1 - \bar{\epsilon},\ldots,\epsilon_n - \bar{\epsilon})$ et $\sigma(\bar{\epsilon})$ sont des tribus indépendantes. Or, d'après ce qui précède, les lois fini-dimensionnelles de $(\epsilon_1 - \bar{\epsilon},\ldots,\epsilon_n - \bar{\epsilon})$ et $\bar{\epsilon}$ sont indépendantes. Le résultat s'ensuit donc. ■

Corollaire 2.1
Soient X *et* $\epsilon_1, \dots, \epsilon_n$ *des processus gaussiens avec* X *indépendant de chaque* ϵ_i. *Pour chaque* $i \in \{1, \dots, n\}$, *supposons que* $\epsilon_i \sim \mathrm{P}_{0,\mathrm{C}}$. *Ecrivons de même* $\mathrm{Y}_i = \mathrm{X} + \epsilon_i$, $\mathrm{Y} = (\mathrm{Y}_1, \dots, \mathrm{Y}_n)$ *et* $\epsilon = (\epsilon_1, \dots, \epsilon_n)$. *Alors on a* $\mathscr{L}(\mathrm{X}|\mathrm{Y}) = \mathscr{L}(\mathrm{X}|\overline{\mathrm{Y}})$.

PREUVE. Il est clair que $\sigma(\mathrm{Y}) = \sigma(\mathrm{X} + \bar{\epsilon}, \epsilon - \bar{\epsilon})$, où $\epsilon - \bar{\epsilon} = (\epsilon_1 - \bar{\epsilon}, \dots, \epsilon_n - \bar{\epsilon})$.

Comme X est indépendant de ϵ, X est également indépendant de $\epsilon - \bar{\epsilon}$. D'après le lemme 2.3, $\bar{\epsilon}$ est indépendant de $\epsilon - \bar{\epsilon}$. Ainsi, la somme $\mathrm{X} + \bar{\epsilon}$ est indépendante de $\epsilon - \bar{\epsilon}$.

Nous déduisons donc notre résultat de la suite d'égalités suivantes et du lemme 2.2 :

$$\begin{aligned} \mathscr{L}(\mathrm{X}|\mathrm{Y}) &= \mathscr{L}(\mathrm{X}|\mathrm{X} + \bar{\epsilon}, \epsilon - \bar{\epsilon}) \\ &= \mathscr{L}(\mathrm{X}|\mathrm{X} + \bar{\epsilon}) \\ &= \mathscr{L}(\mathrm{X}|\overline{\mathrm{Y}}). \end{aligned}$$

■

Nous sommes maintenant capables de généraliser les résultats du théorème 2.7 au cas de multiples observations.

Théorème 2.8
Soient $\mathrm{X} \sim \mathrm{P}_{m,\mathrm{W}}$ *et* $\epsilon_1, \dots, \epsilon_n$ *des processus gaussiens* $\mathrm{P}_{0,\mathrm{C}}$. *Supposons que* X *et chaque* ϵ_i *sont indépendants et notons* $\mathrm{Y}_i = \mathrm{X} + \epsilon_i$ *pour tout* $i \in \{1, \dots, n\}$. *Enfin, supposons que les trajectoires de* Y *sont continues p.s.. Alors, la loi a posteriori* $\mathrm{X}|\mathrm{Y}_1, \dots, \mathrm{Y}_n$ *est également un processus gaussien* $\mathrm{P}_{m^\star,\mathrm{W}^\star}$, *donné par :*

$$m^\star(t) = m(t) + \left(\mathrm{W}(\bullet, t), \bar{\mathrm{Y}} - m\right)_{\mathrm{W}+\frac{\mathrm{C}}{n}}, \tag{2.3.10a}$$

$$\mathrm{W}^\star(t, t') = \mathrm{W}(t, t') - \left(\mathrm{W}(\bullet, t), \mathrm{W}(\bullet, t')\right)_{\mathrm{W}+\frac{\mathrm{C}}{n}}, \tag{2.3.10b}$$

où $(\bullet, \bullet)_{\mathrm{W}+\frac{\mathrm{C}}{n}}$ *désigne le produit scalaire dans l'espace* $\mathrm{H}\left(\mathrm{W} + \frac{\mathrm{C}}{n}\right)$.

PREUVE. Observons avant tout que $\overline{\mathrm{Y}} = \frac{1}{n}\sum_{i=1}^n \mathrm{Y}_i = \mathrm{X} + \bar{\epsilon}$. Comme les processus ϵ_i sont des processus gaussiens indépendants $\mathrm{P}_{0,\mathrm{C}}$, nous savons que $\bar{\epsilon} = \mathrm{P}_{0,\frac{\mathrm{C}}{n}}$.

De plus, d'après le corollaire 2.1, on a $\mathscr{L}(\mathrm{X}|\mathrm{Y}) = \mathscr{L}(\mathrm{X}|\overline{\mathrm{Y}})$. Ainsi, en appliquant le théorème 2.7 et en remplaçant Y par $\bar{\mathrm{Y}}$ et ϵ par $\bar{\epsilon}$, nous obtenons le résultat escompté. ■

CHAPITRE

3 Classification de données fonctionnelles

Résumé

Dans le premier chapitre, nous avons présenté le modèle (DPM). Ce modèle a fourni une solution au problème de classification dans le cas où les observations sont de dimension finie. Nous souhaitons désormais adapter un tel modèle à des données fonctionnelles. En plus des difficultés propres à ces données, comme la dimension infinie, il faudra également généraliser un algorithme de simulation. Dans ce chapitre, nous passons en revue les approches de classification spécifiques à ce type de données, avant de généraliser le modèle basé sur le processus de Dirichlet.

3.1 Classification de données fonctionnelles

Beaucoup d'approches d'analyse de données fonctionnelles se sont concentrées sur la classification. Si les techniques usuelles multivariées (classification hiérarchique, K-means, mélanges gaussiens) peuvent être employées, des méthodes adaptées aux données fonctionnelles ont été proposées, prenant en compte l'aspect temporel. Dans ce qui suit, nous présentons les approches fonctionnelles de classification en deux sous-sections : d'une part les méthodes non bayésiennes et d'autre part les méthodes bayésiennes.

3.1.1 Méthodes non bayésiennes

Une première classe de méthodes se base sur la réduction de la dimension. Les approches actuelles traitent les courbes de manière multivariée en les considérant aux temps d'observation, par décomposition dans des bases de fonctions telles que les splines cubiques [47, 55, 70] ou les *truncated power splines* [47]. En décomposant chaque fonction Y_i dans une base, on peut alors écrire :

$$\mathrm{Y}_i(t) = \sum_{l=1}^{\mathrm{L}} \beta_{il}\Phi_l(t),$$

avec $\beta_i = (\beta_{i1}, \ldots, \beta_{i\mathrm{L}})$, L désignant le nombre de fonctions de base et $(\Phi_l(t))_{l=1}^{\mathrm{L}}$ la base de fonctions. La base de fonctions étant fixée arbitrairement, le vecteur β_i des coefficients aux nœuds est calculé. Une fois les observations résumées en dimension finie, les techniques de classification usuelles multivariées sont appliquées afin de classer les coefficients de décomposition β_i.

Une seconde classe de méthodes est basée sur l'utilisation de la vraisemblance d'un modèle. Comme il est difficile de parler de densité de probabilité en dimension infinie, les auteurs calculent la vraisemblance de paramètres de dimension finie. Le plus souvent, cette vraisemblance est alors pénalisée puis les paramètres du modèle sont déterminés par maximum de vraisemblance ; l'algorithme EM est le plus souvent employé à cet effet. Un point commun à toutes ces méthodes est de fixer le nombre de classes à l'avance, le plus souvent à l'aide des critères *BIC* ([110]), *AIC* ([2]) ou encore *ICL* ([15]).

En général, les méthodes se basent à nouveau sur une décomposition dans une base de fonctions, mais contrairement à la première classe de méthodes dans laquelle les coefficients de décomposition étaient fixés, ils sont ici considérés aléatoires. Par exemple, James & Sugar [55] considèrent que les coefficients de décomposition β_i sont distribués suivant une loi normale multivariée dont les moyennes μ_k sont spécifiques à chaque classe et les matrices de variance-covariance Σ sont identiques :

$$\beta_i \sim \mathcal{N}\left(\mu_{c_i}, \Sigma\right).$$

Rappelons que c_i désigne le label de classe de la i^e observation. Lorsque les méthodes ne décomposent pas les fonctions dans une base, ils font appel à la théorie de l'analyse en composantes principales fonctionnelle. Bouveyron & Jacques [10] ou encore Jacques & Preda [54] supposent que les composantes principales suivent une loi normale. Ils proposent alors un modèle de mélange sur celles-ci et estiment les paramètres du modèle à nouveau à l'aide d'un algorithme EM.

D'autres auteurs comme Ma et al. [69] proposent un modèle de mélange gaussien à effet mixte. Leur algorithme consiste à mettre à jour les courbes moyennes et les paramètres de classes par maximisation d'une vraisemblance pénalisée, et à mettre à jour certains autres paramètres par validation croisée. La probabilité d'appartenir à une classe est alors estimée et les affectations des courbes également.

3.1.2 Méthodes bayésiennes

Les méthodes bayésiennes procèdent à l'inférence *a posteriori* des paramètres d'un modèle. Dans la littérature, le processus de Dirichlet est le plus souvent utilisé dans les modèles bayésiens. Parmi les nombreuses algorithmes d'implémentation, l'échantillonneur de Gibbs est le plus souvent employé. Le plus souvent, les méthodes consistent à travailler directement sur les données discrétisées et à faire usage du modèle (DPM) que nous avons étudié dans le premier chapitre. Eventuellement, chaque courbe est décomposée dans une base de fonctions et les coefficients de décomposition β_i sont classés à l'aide d'un modèle (DPM).

Ray & Mallick [105] utilisent un modèle (DPM) pour classer les coefficients de décomposition dans une base d'ondelettes. Les fonctions ondelettes présentent l'avantage d'avoir de bonnes propriétés d'approximation sur une large classe d'espaces fonctionnels et peuvent prendre la forme de presque toutes les fonctions existantes en pratique. En décomposant alors chaque fonction Y_i dans une telle base, les auteurs proposent le modèle suivant :

$$\begin{array}{rcl} \mathrm{Y}_i|\theta_i = (\beta_i, \sigma_i^2) & \sim & \mathcal{N}(\phi^\top \beta_i, \sigma_i^2 \mathrm{I}), \\ \theta_i|\mathrm{G} & \sim & \mathrm{G}, \\ \mathrm{G} & \sim & \mathrm{DP}(\alpha_0, \mathrm{G}_0), \end{array}$$

où I est la matrice identité. Une approche similaire à celle de Ray & Mallick, plus ancienne, est celle de Gelfand, Kottas & MacEachern [43], mais dans un objectif de prédiction de données spatiales.

Des auteurs comme Brown [11], souhaitant faire de la classification de séquences de protéines, proposent une adaptation d'un modèle (DPM) multivarié. L'algorithme d'implémentation est basé sur l'algorithme *split-merge* de Dahl [24]. Jackson et al. [53] proposent également un modèle (DPM) multivarié. Ils considèrent les courbes discrétisées aux temps d'observation. Leur modèle de classification opère à la fois sur la moyenne et le noyau de covariance, avec des covariances entre classes pouvant être différentes.

D'autres approches permettent de développer un modèle hiérarchique comme c'est le cas avec Scarpa & Dunson [108]. Le modèle possède une partie paramétrique et une partie non-paramétrique à base de processus de Dirichlet ; les processus gaussiens sont discrétisés aux temps d'observation.

Parmi les extensions possibles de la classification de courbes, Petrone et al. [98] ont introduit le *Hybrid Dirichlet Process*, permettant de faire de la classification non pas seulement à partir des centres de courbes, mais de résumer chaque centre comme un mélange de "courbes canoniques" et d'associer en chaque point un indicateur de classe. Ainsi, une même courbe peut avoir différentes parties dans différentes classes et à chaque courbe est associé un label $c_i(t)$ en chaque point ; on appelle cela le *local clustering*.

3.1.3 Spécificités de la thèse

Les méthodes dans lesquelles les courbes sont discrétisées aux temps d'observation font intervenir le déterminant de matrices de variance-covariance de lois normales de très grande dimension, ce qui peut induire une instabilité numérique. De plus, les calculs nécessaires dans ces algorithmes seront d'autant plus longs que le nombre de temps d'observation augmente. Dans le cas de la décomposition dans une base de fonctions, l'utilisateur est soumis au problème du choix de la base et à l'adéquation entre modèle d'approximation et données. Afin de contourner ces limitations, nous proposons de généraliser le modèle (DPM) à des observations fonctionnelles modélisées par des processus gaussiens.

3.2 Présentation du modèle fonctionnel

Nous modélisons les observations $\mathrm{Y}_1, \ldots, \mathrm{Y}_n$ par des processus gaussiens indépendants et nous notons $t_i = (t_{i1}, \ldots, t_{ij_i})$ le vecteur des temps d'observation pour la i^e observation. Nous proposons une généralisation du modèle (DPM) défini par (1.2.5) et (1.2.6) au cas infini-dimensionnel. Ce modèle, que nous notons DPMF (DPM fonctionnel), est défini de la façon suivante :

$$(\mathrm{DPMF}) \begin{cases} \mathrm{Y}_i|\theta_i & \overset{ind}{\sim} & \mathrm{P}_{\theta_i,\Sigma}, \\ \theta_i|\mathrm{G} & \overset{ind}{\sim} & \mathrm{G}, \\ \mathrm{G} & \sim & \mathrm{DP}(\alpha_0, \mathrm{G}_0), \end{cases} \tag{3.2.1}$$

équivalent d'après le premier chapitre au modèle suivant :

$$\begin{cases} \mathrm{Y}_i|c_i, \phi_c & \overset{ind}{\sim} & \mathrm{P}_{\phi_{c_i},\Sigma}, \\ (c_1, \ldots, c_n) & \sim & \mathrm{CRP}(\alpha_0), \\ \phi_c & \overset{ind}{\sim} & \mathrm{G}_0. \end{cases} \tag{3.2.2}$$

Pour des raisons pratiques, nous supposons que $\mathrm{G}_0 = \mathrm{P}_{\mu,\Sigma_0}$. Dans un premier temps, nous considérons fixes les paramètres μ, Σ, Σ_0 ainsi que α_0. Dans les études numériques, nous envisagerons un modèle plus complet nous permettant de les estimer.

3.3 Traitement multivarié

3.3.1 Modèle fini-dimensionnel et implémentation algorithmique

Rappelons que nous travaillons avec des processus dont les trajectoires sont de carré intégrable sur [0,T]. Définissons l'opérateur suivant pour chaque vecteur $t = (t_1, \dots, t_d) \in \mathbb{R}^d$ avec $d \geq 1$:

$$\begin{array}{rccc} \pi_t & : & \mathrm{L}^2([0,\mathrm{T}]) & \rightarrow & \mathbb{R}^d \\ & & f & \mapsto & \pi_t(f) = (f(t_1), \dots, f(t_d)) \end{array}$$

Étudions les processus gaussiens sur l'ensemble de leurs temps d'observation. D'une part, par définition des processus gaussiens dans le modèle (3.2.1), nous pouvons écrire :

$$\pi_{t_i}(\mathrm{Y}_i)|\pi_{t_i}(\theta_i) \overset{ind}{\sim} \mathcal{N}\left(\pi_{t_i}(\theta_i), \Sigma_{j_i}\right), \tag{3.3.1}$$

où $\Sigma_{j_i} = [\Sigma(t_{ik}, t_{il})]_{1\leq k,l\leq j_i}$. Considérons à présent un vecteur t contenant au moins tous les vecteurs t_i, c'est-à-dire tel que $\cup_{i=1}^{n} \cup_{j=1}^{j_i} \{t_{ij}\} \subseteq \{t\}$, et que l'on note $t = (t_1, \dots, t_d)$. La loi conditionnelle de (3.3.1) ne dépendant des $\pi_t(\theta_i)$ qu'au travers des valeurs $\pi_{t_i}(\theta_i)$, nous pouvons également écrire :

$$\pi_{t_i}(\mathrm{Y}_i)|\pi_t(\theta_i) \overset{ind}{\sim} \mathcal{N}\left(\pi_{t_i}(\theta_i), \Sigma_{j_i}\right).$$

Le vecteur t contenant tous les t_i, on peut poser X_i comme la matrice de 0 et de 1 telle que $\mathrm{X}_i\pi_t(\theta_i) = \pi_{t_i}(\theta_i)$, c'est-à-dire la matrice à j_i lignes et d colonnes constituée des éléments $\mathrm{X}_{i_{k,l}} = \delta_{t_{ik}}(t_l)$. Ainsi :

$$\pi_{t_i}(\mathrm{Y}_i)|\pi_t(\theta_i) \overset{ind}{\sim} \mathcal{N}\left(\mathrm{X}_i\pi_t(\theta_i), \Sigma_{j_i}\right).$$

D'autre part, démontrons la proposition suivante :

Proposition 3.1
Si le vecteur $(\theta_1, \dots, \theta_n)$ admet une représentation par urne de Pólya de paramètres α_0 et $\mathrm{G}_0 = \mathrm{P}_{\mu,\Sigma_0}$, alors $(\pi_t(\theta_1), \dots, \pi_t(\theta_n))$ admet une représentation par urne de Pólya de paramètres α_0 et $\mathrm{G}_0^{(d)} = \mathcal{N}(\pi_t(\mu), \Sigma_{0d})$, où $\Sigma_{0d} = \left[\Sigma_0(t_i, t_j)\right]_{1\leq i,j\leq d}$.

PREUVE. Par définition d'une suite de Pólya on a $\theta_1 \sim \mathrm{G}_0$ donc $\pi_t(\theta_1) \sim \mathcal{N}(\pi_t(\mu), \Sigma_{0d}) = \mathrm{G}_0^{(d)}$. De plus, la loi conditionnelle de $\theta_2|\theta_1$ est définie de la manière suivante d'après la représentation par urne de Pólya (1.2.3) :

$$\mathrm{P}(d\theta_2|\theta_1) = \frac{1}{\alpha_0+1}\delta_{\theta_1}(d\theta_2) + \frac{\alpha_0}{\alpha_0+1}\mathrm{G}_0(d\theta_2).$$

Si $\theta_2 = \theta_1$, il est clair que $\pi_t(\theta_2) = \pi_t(\theta_1)$ et si $\theta_2 \neq \theta_1$, alors $\pi_t(\theta_2) \sim \mathrm{G}_0^{(d)}$. Cela permet d'écrire :

$$\mathrm{P}(d\pi_t(\theta_2)|\theta_1) = \frac{1}{\alpha_0+1}\delta_{\pi_t(\theta_1)}(d\pi_t(\theta_2)) + \frac{\alpha_0}{\alpha_0+1}\mathrm{G}_0^{(d)}(d\pi_t(\theta_2)).$$

Or, la loi de $\pi_t(\theta_2)|\theta_1$ ne dépend de θ_1 qu'au travers des valeurs prises $\pi_t(\theta_1)$, donc la loi de $\pi_t(\theta_2)|\pi_t(\theta_1)$ est la même que la loi de $\pi_t(\theta_2)|\theta_1$.

Le raisonnement par récurrence permet de conclure que $(\pi_t(\theta_1), \dots, \pi_t(\theta_n))$ admet une représentation par urne de Pólya de paramètres α_0 et $\mathrm{G}_0^{(d)}$. ■

Le modèle fini-dimensionnel impliqué par le modèle fonctionnel (3.2.1) est donc le suivant :

$$\begin{cases} \pi_{t_i}(\mathrm{Y}_i)|\pi_t(\theta_i) & \overset{ind}{\sim} & \mathcal{N}\left(\mathrm{X}_i\pi_t(\theta_i), \Sigma_{j_i}\right), \\ \pi_t(\theta_i)|\mathrm{G}^{(d)} & \overset{ind}{\sim} & \mathrm{G}^{(d)}, \\ \mathrm{G}^{(d)} & \sim & \mathrm{DP}(\alpha_0, \mathrm{G}_0^{(d)}), \\ \mathrm{G}_0^{(d)} & = & \mathcal{N}(\pi_t(\mu), \Sigma_{0d}). \end{cases}$$

De manière équivalente, ce modèle s'écrit :

$$\begin{cases} \pi_{t_i}(\mathrm{Y}_i)|c_i, \pi_t(\phi_c) & \overset{ind}{\sim} & \mathrm{P}_{\mathrm{X}_i\pi_t(\phi_{c_i}),\Sigma_{j_i}}, \\ (c_1, \ldots, c_n) & \sim & \mathrm{CRP}(\alpha_0), \\ \pi_t(\phi_c) & \overset{ind}{\sim} & \mathrm{G}_0^{(d)}, \\ \mathrm{G}_0^{(d)} & = & \mathcal{N}(\pi_t(\mu), \Sigma_{0d}). \end{cases}$$

Appliquer l'algorithme de base page 44 requiert de générer chaque $\pi_t(\theta_i)$. Lorsque l'on travaille en grande dimension, cela nécessite trop de temps et l'algorithme peut ne pas s'exécuter correctement. L'idéal serait de ne pas générer ces valeurs $\pi_t(\theta_i)$, ce qui revient à intégrer les $\pi_t(\phi_c)$ dans le modèle ci-dessus et inférer uniquement sur les c_i. Pour cela, un algorithme de référence a été proposé par MacEachern [71] et consiste à simuler les c_i suivant leur loi conditionnelle complète :

$$\mathbb{P}(c_i = c|c_j, j \neq i, \pi_{t_1}(\mathrm{Y}_1), \ldots, \pi_{t_n}(\mathrm{Y}_n)) \propto \begin{cases} \#\{j \neq i : c_j = c\} p(\pi_{t_i}(\mathrm{Y}_i)|c, \mathrm{Y}^{-i}), & c = c_j, j \neq i, \\ \alpha_0 p(\pi_{t_i}(\mathrm{Y}_i)), & c \neq c_j, j \neq i, \end{cases}$$

où Y^{-i} désigne le vecteur $(\pi_{t_1}(\mathrm{Y}_1), \ldots, \pi_{t_n}(\mathrm{Y}_n))$ privé de sa i^e composante. Les quantités $p(\pi_{t_i}(\mathrm{Y}_i)|c, \mathrm{Y}^{-i})$ et $p(\pi_{t_i}(\mathrm{Y}_i))$ correspondent respectivement aux densités de loi prédictive *a posteriori* sachant la classe et prédictive *a priori*.

Proposition 3.2
Nous avons les résultats suivants :

$$\begin{aligned} \pi_{t_i}(\mathrm{Y}_i) &\sim \mathcal{N}(\mathrm{X}_i\pi_t(\mu), \Sigma_{j_i} + \mathrm{X}_i\Sigma_{0d}\mathrm{X}_i^\top), \\ \pi_{t_i}(\mathrm{Y}_i)|c, \mathrm{Y}^{-i} &\sim \mathcal{N}(\mathrm{X}_i\mathrm{S}, \Sigma_{j_i} + \mathrm{X}_i\mathrm{T}\mathrm{X}_i^\top), \end{aligned}$$

où $\mathrm{T} = (\sum_{k\neq i:c_k=c} \mathrm{X}_k^\top \Sigma_{j_k}^{-1}\mathrm{X}_k + \Sigma_{0d}^{-1})^{-1}$ *et* $\mathrm{S} = \mathrm{T}(\sum_{k\neq i:c_k=c} \mathrm{X}_k^\top \Sigma_{j_k}^{-1}\pi_{t_k}(\mathrm{Y}_k) + \Sigma_{0d}^{-1}\pi_t(\mu))$.

PREUVE. Il faut calculer la loi prédictive *a priori* dans le modèle suivant :

$$\begin{cases} \pi_{t_i}(\mathrm{Y}_i)|\pi_t(\phi) & \sim & \mathcal{N}\left(\mathrm{X}_i\pi_t(\phi), \Sigma_{j_i}\right), \\ \pi_t(\phi) & \sim & \mathrm{G}_0^{(d)}. \end{cases} \tag{3.3.2}$$

Un calcul bien connu en statistique bayésienne permet de montrer que la loi prédictive $\pi_{t_i}(\mathrm{Y}_i)$ dans ce modèle est une normale $\mathcal{N}(\mathrm{X}_i\pi_t(\mu), \Sigma_{j_i} + \mathrm{X}_i\Sigma_{0d}\mathrm{X}_i^\top)$. En particulier, $p(\pi_{t_i}(\mathrm{Y}_i))$ n'est autre que la densité de cette loi par rapport à la mesure de Lebesgue.

Calculons en second lieu la loi de $\pi_t(\phi)|c, \mathrm{Y}^{-i}$, loi *a posteriori* de $\pi_t(\phi)$ en se basant sur le prior $\mathrm{G}_0^{(d)}$ et toutes les observations $\pi_{t_k}(\mathrm{Y}_k)$ pour lesquelles $k \neq i$ et $c_k = c$. On sait que la densité de cette loi par rapport à la mesure de Lebesgue est proportionnelle à :

$$\prod_{k\neq i:c_k=c} p(\pi_{t_k}(\mathrm{Y}_k)|\pi_t(\phi)) g_0^{(d)}(\pi_t(\phi)),$$

où $g_0^{(d)}$ désigne la densité de $\mathrm{G}_0^{(d)}$ par rapport à la mesure de Lebesgue. Ecrivons alors :

$$\begin{aligned}\prod_{k\neq i:c_k=c} p(\pi_{t_k}(\mathrm{Y}_k)|\pi_t(\phi))g_0^{(d)}(\pi_t(\phi)) &\propto \prod_{k\neq i:c_k=c} e^{-\frac{1}{2}(\pi_{t_k}(\mathrm{Y}_k)-\mathrm{X}_k\pi_t(\phi))^\top\Sigma_{j_k}^{-1}(\pi_{t_k}(\mathrm{Y}_k)-\mathrm{X}_k\pi_t(\phi))} \\ &\quad\times e^{-\frac{1}{2}(\pi_t(\phi)-\pi_t(\mu))^\top\Sigma_{0d}^{-1}(\pi_t(\phi)-\pi_t(\mu))}, \\ &\propto e^{-\frac{\mathrm{Q}}{2}},\end{aligned}$$

où :

$$\begin{aligned}\mathrm{Q} &= \sum_{k\neq i:c_k=c}(\pi_{t_k}(\mathrm{Y}_k)-\mathrm{X}_k\pi_t(\phi))^\top\Sigma_{j_k}^{-1}(\pi_{t_k}(\mathrm{Y}_k)-\mathrm{X}_k\pi_t(\phi))+(\pi_t(\phi)-\pi_t(\mu))^\top\Sigma_{0d}^{-1}(\pi_t(\phi)-\pi_t(\mu)), \\ &= \phi^\top(\sum_{k\neq i:c_k=c}\mathrm{X}_k^\top\Sigma_{j_k}^{-1}\mathrm{X}_k+\Sigma_{0d}^{-1})\pi_t(\phi)-2\pi_t(\phi)^\top(\sum_{k\neq i:c_k=c}\mathrm{X}_k^\top\Sigma_{j_k}^{-1}\pi_{t_k}(\mathrm{Y}_k)+\Sigma_{0d}^{-1}\pi_t(\mu))+\ldots,\end{aligned}$$

ce qui montre que $\pi_t(\phi)|c,\mathrm{Y}^{-i}$ suit la loi normale $\mathcal{N}(\mathrm{S},\mathrm{T})$, où :

$$\begin{aligned}\mathrm{T} &= (\sum_{k\neq i:c_k=c}\mathrm{X}_k^\top\Sigma_{j_k}^{-1}\mathrm{X}_k+\Sigma_{0d}^{-1})^{-1}, \\ \mathrm{S} &= \mathrm{T}(\sum_{k\neq i:c_k=c}\mathrm{X}_k^\top\Sigma_{j_k}^{-1}\pi_{t_k}(\mathrm{Y}_k)+\Sigma_{0d}^{-1}\pi_t(\mu)).\end{aligned}$$

Il suffit enfin d'utiliser les résultats précédents pour la loi prédictive dans le modèle (3.3.2) afin d'obtenir la seconde égalité en loi. ■

3.3.2 Limitations de la version multivariée

Le passage en dimension finie permet de contourner certaines limitations théoriques et d'envisager un algorithme portant sur les labels c_i uniquement. Cependant, il est à noter que l'implémentation de cet algorithme est très lente en pratique car elle demande à chaque fois l'inversion de matrices dont la dimension est celle des données ainsi que les déterminants de ces matrices. A cause des problèmes soulevés dans la sous-section 3.1.3, il n'est pas possible d'implémenter cet algorithme. De plus, les matrices à inverser sont susceptibles de changer à chaque itération et il n'est pas possible de stocker leur inverse au préalable en amont de l'algorithme.

En conclusion, les nombreux algorithmes existants pour inférer sur le (DPM) et portant uniquement sur les labels s'exécutent bien pour des données de faible dimension, mais ils sont insatisfaisants pour des données de grande dimension telles que les données fonctionnelles. Les algorithmes existants pour inférer sur le (DPM) dans la littérature ne sont donc, à notre connaissance, pas satisfaisants.

3.4 Vers un algorithme fonctionnel

Rappelons que nous étudions les modèles (3.2.1) et (3.2.2). Dans un cadre fonctionnel, l'idéal serait de ne pas générer de courbes, ce qui revient à intégrer sur les ϕ_c dans le modèle (3.2.2) et inférer uniquement sur les c_i. A cet effet, comme nous venons de le voir en dimension finie, nous disposons de l'algorithme de référence proposé par MacEachern [71] et qui consiste à simuler les c_i suivant leur loi conditionnelle complète :

$$\mathbb{P}(c_i=c|c_j,j\neq i,\mathrm{Y}_1,\ldots,\mathrm{Y}_n)\propto\begin{cases}\#\{j\neq i:c_j=c\}p(\mathrm{Y}_i|c,\mathrm{Y}^{-i}), & c=c_j,j\neq i,\\ \alpha_0 p(\mathrm{Y}_i), & c\neq c_j,j\neq i.\end{cases}\tag{3.4.1}$$

Généraliser cet algorithme dans notre cas infini-dimensionnel demanderait d'exprimer les densités $p(\mathrm{Y}_i|c,\mathrm{Y}^{-i})$ et $p(\mathrm{Y}_i)$ relativement à une même mesure de référence. En effet, une mesure

différente entraînerait des poids incohérents dans le choix des classes. Nous avons analysé dans l'annexe A.5 le cas où $\Sigma = \Sigma_0$, et même dans ce cas il nous a été impossible de trouver une même mesure de référence.

Plus généralement, il ne nous a pas été possible d'appliquer l'algorithme de MacEachern car nous ne sommes pas arrivés à exprimer les densités nécessaires relativement à une même mesure de référence. D'une part, il n'existe pas d'équivalent de la mesure de Lebesgue en dimension infinie. D'autre part, nous n'avons pas réussi à exprimer ces densités par rapport à un même processus gaussien car elles sont associées à des fonctions de covariance multiples. Le théorème 2.2 dont nous disposons considère une même fonction de covariance. Parmi les algorithmes existants pour inférer sur le (DPM), ceux inférant uniquement sur les labels c_i font intervenir les densités (b) et (c) ci-dessous et n'ont pas pu être implémentés :

(a) $p(\mathrm{Y}_i|c,\phi_c)$,

(b) $p(\mathrm{Y}_i)$,

(c) $p(\mathrm{Y}_i|c,\mathrm{Y}^{-i})$.

Quant aux autres algorithmes faisant intervenir les c_i et les ϕ_c, ils font tous appel à la densité (a) et parfois aux densités (b) ou (c) également. Seuls ceux ne faisant intervenir que la densité (a) sont implémentables dans notre cas, pour les raisons citées précédemment.

Parmi ces derniers, nous choisissons de généraliser l'algorithme 8 de Neal [87] qui est, d'après l'auteur, l'un des plus performants. Cet algorithme est généralisé en choisissant comme mesure de référence le processus gaussien $\mathrm{P}_{0,\Sigma}$. Il s'agit d'un algorithme exact de simulation *a posteriori* basé sur un échantillonneur de Gibbs. Cet algorithme est une amélioration d'un algorithme proposé par MacEachern & Müller [73] et qui se base sur les paramètres $(c_1,\dots,c_n)$ et $(\phi_1,\dots,\phi_n,\dots,\phi_{n+m})$. L'entier m est fixé arbitrairement et sert à générer m valeurs auxiliaires de paramètres. La convergence de cet algorithme est assurée par le même raisonnement que pour celui de MacEachern & Müller, quelle que soit la valeur de m. D'après certaines simulations de l'auteur, plus m est grand et plus l'algorithme converge rapidement vers sa loi stationnaire. En revanche, chaque itération exigera plus de temps. Étant donné l'entier $m \geq 1$, cet algorithme se présente de la façon suivante :

Algorithme 2: Gibbs sampling with Auxiliary Parameters (pour données fonctionnelles)

pour *i=1,...,n* **faire**

$k^- = \#\{c_j, j \neq i\}$;
$h = k^- + m$;
Numéroter les c_j, $j \neq i$, dans $\{1,\dots,k^-\}$;
Si $c_i = c_j$ pour un $j \neq i$, simuler $\phi_c \sim \mathrm{P}_{\mu,\Sigma_0}$ pour $k^- < c \leq h$;
Si $c_i \neq c_j$ pour tout $j \neq i$, simuler $\phi_c \sim \mathrm{P}_{\mu,\Sigma_0}$ pour $k^- + 1 < c \leq h$;
Générer c_i à l'aide des probabilités conditionnelles suivantes :

$$\mathbb{P}(c_i = c|c_j, j \neq i, \phi_1,\dots,\phi_h,\mathrm{Y}_1,\dots,\mathrm{Y}_n) \propto \begin{cases} \#\{j \neq i : c_j = c\} p(\mathrm{Y}_i|c,\phi_c), & 1 \leq c \leq k^-, \\ \frac{\alpha_0}{m} p(\mathrm{Y}_i|c,\phi_c), & k^- + 1 \leq c \leq h. \end{cases} \quad (3.4.2)$$

pour $c \in \{c_1,\dots,c_n\}$ **faire**

Simuler ϕ_c suivant la loi *a posteriori* de ϕ sachant les données Y_i telles que $c_i = c$

3.5 Résultats sur la vraisemblance et le processus a posteriori

L'implémentation de l'algorithme requiert :
– les densités $p(Y_i|c,\phi_c)$, pour tout i et pour tout c, relativement à une même mesure de référence,
– la simulation suivant la loi *a posteriori* de chaque ϕ_c sachant les données qui sont dans la classe c.

Pour le premier point, le calcul des densités revient à trouver une mesure de référence P qui permette d'exprimer les dérivées de Radon-Nikodym $\frac{dP_{\phi_c,\Sigma}}{dP}$. D'après le théorème 2.2, chaque processus gaussien $P_{\phi_c,\Sigma}$ admet une densité par rapport au processus $P_{0,\Sigma}$ si, et seulement si, $\phi_c \in H(\Sigma)$. La régularité des trajectoires des ϕ_c étant due, d'après le modèle équivalent (3.2.2), à la fonction de covariance Σ_0, nous considérerons dans toute la suite une fonction de covariance Σ_0 de sorte que la condition $\phi_c \in H(\Sigma)$ soit toujours remplie. En utilisant le théorème de Bayes, valable dans un cadre général [109], nous pouvons désormais expliciter la loi conditionnelle complète (3.4.2) de chaque c_i.

Pour le second point concernant la simulation suivant la loi *a posteriori* de chaque ϕ_c sachant les données qui sont dans la classe c, le théorème 2.8 permet à nouveau de montrer que la loi conditionnelle de ζ sachant $x_1,\dots,x_n$ dans le modèle suivant :

$$\begin{cases} x_i|\zeta & \overset{ind}{\sim} & P_{\zeta,\Sigma}, \\ \zeta & \sim & P_{\mu,\Sigma_0}, \end{cases} \tag{3.5.1}$$

est la loi $P_{m,K}$, où les fonctions m et K sont définies par :

$$\begin{aligned} m(t) &= \mu(t) + \left(\Sigma_0(\bullet,t), \left(\tfrac{1}{n}\textstyle\sum_{i=1}^n x_i\right) - \mu\right)_{\frac{\Sigma}{n}+\Sigma_0}, \\ K(s,t) &= \Sigma_0(s,t) - \left(\Sigma_0(\bullet,s), \Sigma_0(\bullet,t)\right)_{\frac{\Sigma}{n}+\Sigma_0}. \end{aligned}$$

En particulier, nous supposons que Σ et Σ_0 sont deux fonctions faiblement continues sur $[0,T]\times[0,T]$. A partir de ce résultat, nous déduisons que la loi *a posteriori* de ϕ_c sachant les données qui sont dans la classe c est la loi $P_{m,K}$ donnée par :

$$\begin{aligned} m(t) &= \mu(t) + \left(\Sigma_0(\bullet,t), \left(\tfrac{1}{n_c}\textstyle\sum_{j:c_j=c} Y_j\right) - \mu\right)_{\frac{\Sigma}{n_c}+\Sigma_0}, \\ K(s,t) &= \Sigma_0(s,t) - \left(\Sigma_0(\bullet,s), \Sigma_0(\bullet,t)\right)_{\frac{\Sigma}{n_c}+\Sigma_0}, \end{aligned} \tag{3.5.2}$$

où $n_c = \#\{j : c_j = c\}$.

3.6 Résultats et discussion

L'implémentation a été réalisée à partir du logiciel Matlab. Afin d'analyser les performances de notre modèle, nous avons appliqué notre méthode sur trois jeux de données. Comme il est toujours difficile d'évaluer la performance d'un algorithme de classification, nous avons fait le choix, lorsque c'est possible, de comparer le taux de classification correcte (voir sous-section 1.1.3) abrégé TCC. Enfin, nous donnons également à titre indicatif le temps moyen requis pour une itération, utile dans le cas du traitement de jeux de données en très grande dimension. Tous les résultats présentés ont été obtenus en produisant 10000 itérations, avec un temps de chauffe

de 1000 et en retenant 1 itération sur 5. L'implémentation numérique étant déjà conséquente, la classification retenue est celle qui apparaît le plus souvent dans les simulations *a posteriori* (voir discussion en fin de chapitre).

3.6.1 Spécification du modèle

Nous fixons la fonction moyenne μ à zéro et choisissons un noyau de covariance Σ d'un processus d'Ornstein-Uhlenbeck :

$$\Sigma(s,t) = \frac{\sigma^2}{2\beta} e^{-\beta|s-t|}, \tag{3.6.1}$$

où σ et β sont deux réels strictement positifs. Cela nous permet de travailler avec des produits scalaires simples. En effet, d'après certains résultats [6, 94], l'espace $\mathrm{H}(\Sigma)$ est formé des fonctions différentiables sur $[0,\mathrm{T}]$ et le produit scalaire sur cet espace est donné par :

$$(f,g)_\Sigma = \frac{1}{\sigma^2}\int_0^{\mathrm{T}} (f'(t)g'(t) + \beta^2 f(t)g(t))\,dt + \frac{\beta}{\sigma^2}(f(0)g(0) + f(\mathrm{T})g(\mathrm{T})). \tag{3.6.2}$$

Remarque 3.1
Dans cet exemple, le produit scalaire est explicite. Cependant, notre méthode est valable aussi si l'on ne dispose pas de forme explicite du produit scalaire. A défaut de pouvoir l'expliciter, les théorèmes 2.5 et 2.6 fournissent une solution d'approximation.

Rappelons que l'écriture des densités $p(\mathrm{Y}_i|c,\phi_c)$ impose que chaque $\phi_c \in \mathrm{H}(\Sigma)$. Pour assurer cette condition, nous proposons le noyau de covariance Σ_0 suivant :

$$\Sigma_0(s,t) = \frac{\sigma_0^2}{2\beta_0} e^{-\beta_0(s-t)^2}, \tag{3.6.3}$$

lequel garantit des trajectoires de classe C^∞.

Sachant que le nombre de classes dans notre modèle dépend fortement du paramètre α_0 (voir section 1.7), et afin d'éviter de fixer arbitrairement ce paramètre, nous choisissons d'utiliser la méthode d'Escobar & West [38], détaillée dans la sous-section 1.7.2. Conditionnellement à α_0, les lois conditionnelles complètes de $(c_1,\dots,c_n)$ (3.4.2) et de ϕ (3.5.2) ne changent pas. Ainsi, il est possible d'adapter notre algorithme 2 afin de rajouter l'inférence sur α_0 :

Algorithme 3: Algorithme 2 avec inférence sur α_0

pour *i=1,...,n* **faire**
  Simuler c_i suivant sa loi conditionnelle complète (3.4.2).
pour $c \in \{c_1,\dots,c_n\}$ **faire**
  Simuler ϕ_c suivant sa loi conditionnelle complète (3.5.2).
Simuler une variable η de loi conditionnelle $\eta|\alpha_0 \sim \mathrm{B}eta(\alpha_0+1, n)$.
Poser $\tilde{\pi} = \frac{a+\mathrm{K}-1}{a+\mathrm{K}-1+n(b-\log(\eta))}$, où K est le nombre de classes à l'instant courant.
Simuler α_0 suivant sa loi conditionnelle complète qui est :

$$\tilde{\pi}\mathcal{G}amma(a+\mathrm{K}, b-\log(\eta)) + (1-\tilde{\pi})\mathcal{G}amma(a+\mathrm{K}-1, b-\log(\eta)).$$

Le nombre de valeurs auxiliaires est fixé à $m=5$. Les hyperparamètres β_0 et σ_0 sont fixés de sorte que la variabilité des données générées à partir de Σ_0 soit de l'ordre de celle des données. Différentes valeurs de β_0 et σ_0 ont été testées, avec pour seule conséquence un temps de convergence

de l'algorithme vers la loi stationnaire plus ou moins grand. Concernant les hyperparamètres β et σ, nous proposons de les fixer de manière empirique. En effet, notre expérience montre qu'une modélisation bayésienne complète induirait des problèmes numériques sans améliorer nécessairement les résultats de classification. Ainsi, sachant que les courbes Y_i sont générées à partir de processus gaussiens de fonction de covariance Σ avec $\Sigma(s,t) = \frac{\sigma^2}{2\beta} e^{-\beta|s-t|}$, les hyperparamètres β et σ sont fixés à partir de l'estimation empirique de la matrice de variance-covariance intra-classe des courbes discrétisées en quelques points.

Enfin, le processus *a posteriori* $\mathrm{P}_{m,\mathrm{K}}$ de l'équation (3.5.2) peut être approché, en le discrétisant, par les lois normales multivariées $\mathcal{N}(\mathrm{S},\mathrm{T})$ à l'aide des deux matrices suivantes [93] :

$$\mathrm{S} = \mu_{\mathrm{N}} + \Sigma_{0\mathrm{N}}^{\top} \left(\frac{\Sigma_{\mathrm{N}}}{n_c} + \Sigma_{0\mathrm{N}} \right)^{-1} \left(\frac{\sum_{j:c_j=c} \mathrm{Y}_{j\mathrm{N}}}{n_c} - \mu_{\mathrm{N}} \right), \tag{3.6.4}$$

$$\mathrm{T} = \Sigma_{0\mathrm{N}} - \Sigma_{0\mathrm{N}}^{\top} \left(\Sigma_{0\mathrm{N}} + \frac{\Sigma_{\mathrm{N}}}{n_c} \right)^{-1} \Sigma_{0\mathrm{N}}, \tag{3.6.5}$$

où l'on note $\Sigma_{\mathrm{N}} = \left[\Sigma(t_i,t_j)\right]_{1\leq i,j\leq \mathrm{N}}$, $\Sigma_{0\mathrm{N}} = \left[\Sigma_0(t_i,t_j)\right]_{1\leq i,j\leq \mathrm{N}}$, $\mathrm{Y}_{j\mathrm{N}} = \left[\mathrm{Y}_j(t_i)\right]_{1\leq i\leq \mathrm{N}}$ et enfin $\mu_{\mathrm{N}} = \left[\mu(t_i)\right]_{1\leq i\leq \mathrm{N}}$. Ces matrices correspondent aux versions discrétisées des produits scalaires intervenant dans le processus $\mathrm{P}_{m,\mathrm{K}}$ de l'équation (3.5.2) d'après les théorèmes 2.5 et 2.6. Notons finalement que nous pouvons déterminer le nombre de points de discrétisation nécessaires au calcul d'un produit scalaire entre deux fonctions en comparant par exemple des calculs numériques suivant différents pas de discrétisation (voir discussion en fin de chapitre).

3.6.2 Jeu de données simulées

Le jeu de données simulées est constitué de 40 courbes observées uniformément sur 100 points de l'ensemble $[0,10]$. Sur cet ensemble, nous avons généré les quatre polynômes suivants :

$$\begin{cases} s_1(t) &= 0.011t^3 - 0.16t^2 + 0.5t, \\ s_2(t) &= -0.0075t^4 + 0.149t^3 - 0.91t^2 + 1.7t, \\ s_3(t) &= 0.00391t^5 - 0.0977t^4 + 0.854t^3 - 3.05t^2 + 3.7t, \\ s_4(t) &= -0.002009t^6 + 0.06026t^5 - 0.6822t^4 + 3.6t^3 - 8.71t^2 + 7.6t, \end{cases}$$

qui ont permis de créer quatre classes. Pour chacune de ces fonctions, nous avons simulé 10 processus gaussiens de moyenne s_i et de covariance donnée par celle du processus d'Ornstein-Uhlenbeck, de paramètres $\beta = 10$ et $\sigma = 2.5$. Les données sont générées de manière indépendante et sont présentées figures 3.1 et 3.2.

Nous initialisons notre algorithme avec des valeurs $\beta = \beta_0 = 15$ et $\sigma = \sigma_0 = 1$, fixées de manière arbitraire. Ceci nous permet d'obtenir une première classification, nous permettant alors d'estimer et fixer $\beta = \beta_0 = 14$ et $\sigma = \sigma_0 = 1.58$. La loi *a priori* gamma sur α_0 est telle que $a = 1$ et $b = 0.5$.

En répétant notre algorithme 50 fois, nous obtenons un TCC moyen de 77.90% et un temps moyen par itération de l'ordre de 6s. Les résultats complets sont présentés dans la figure 3.3 sous forme de diagramme en boîte pour le TCC.

Enfin, nous avons souhaité connaître le comportement de notre algorithme sur des données simulées avec un noyau de covariance différent de celui utilisé en estimation. Ainsi, nous avons simulé des données suivant un noyau de covariance de type Σ_0 ou encore un noyau de covariance exponentiel. Nous obtenons respectivement un TCC de l'ordre de 62% et 40%. Ces résultats sont

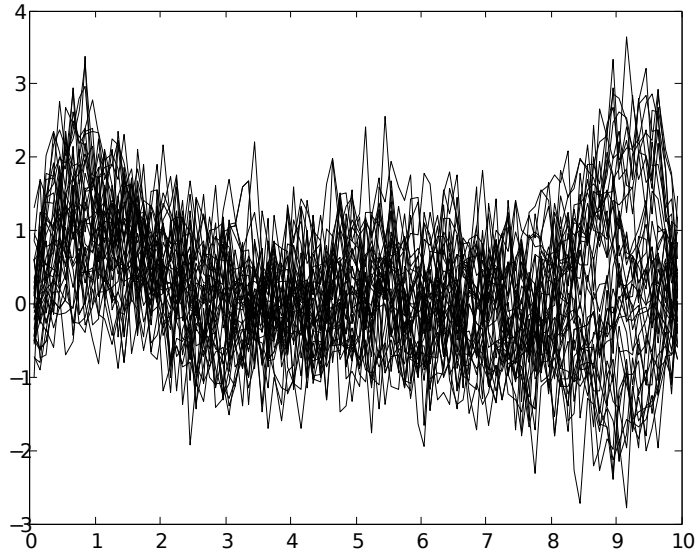

FIGURE 3.1 – *Représentation des courbes simulées, toutes classes confondues.*

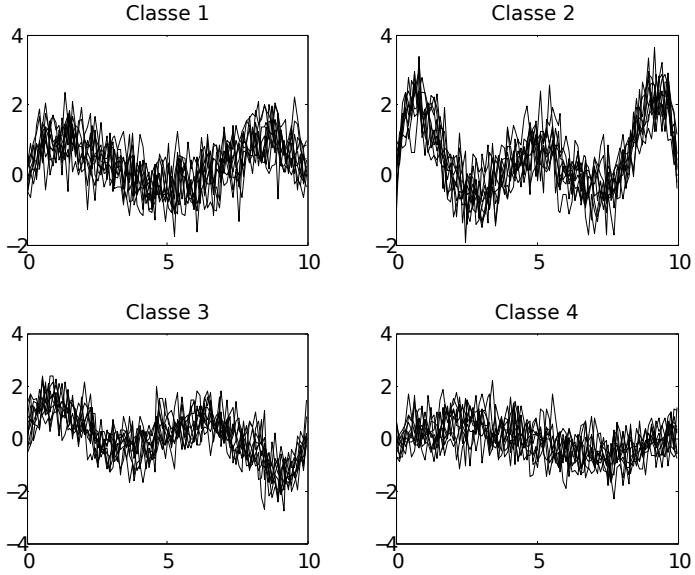

FIGURE 3.2 – *Représentation des courbes simulées, classes séparées.*

évidemment moins bons que lorsque les données sont simulées suivant une covariance de type Ornstein-Uhlenbeck. Ceci est du à la nature des données, dont la régularité des trajectoires ne correspond plus à celle du noyau utilisé en estimation.

3.6.3 Jeu de données de courbes de croissance

Notre second jeu de données est un jeu de données réelles provenant de l'étude Berkeley, reprise par Tuddenham & Snyder [127]. Cette étude a permis de produire des courbes de croissance de garçons et de filles, de la naissance à l'âge adulte. Les données sont disponibles dans le package *fda* de R. Ce jeu de données représente l'évolution temporelle de la taille de 54 filles et de 39 garçons, de 1 an à 18 ans et à 31 instants variables. Les courbes sont représentées figure 3.4.

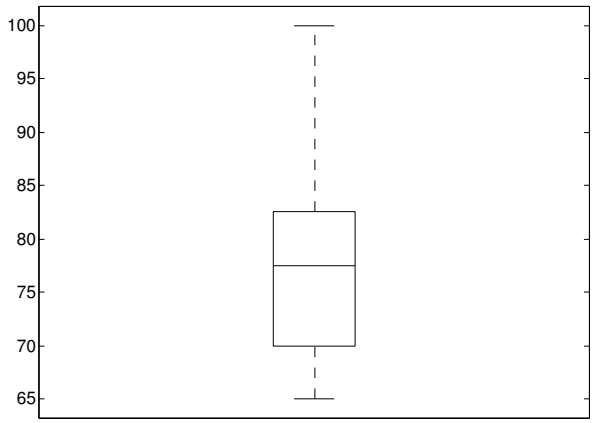

FIGURE 3.3 – *Diagramme en boîte du TCC, obtenu sur 50 répétitions de notre algorithme. La moyenne est de 77.90%.*

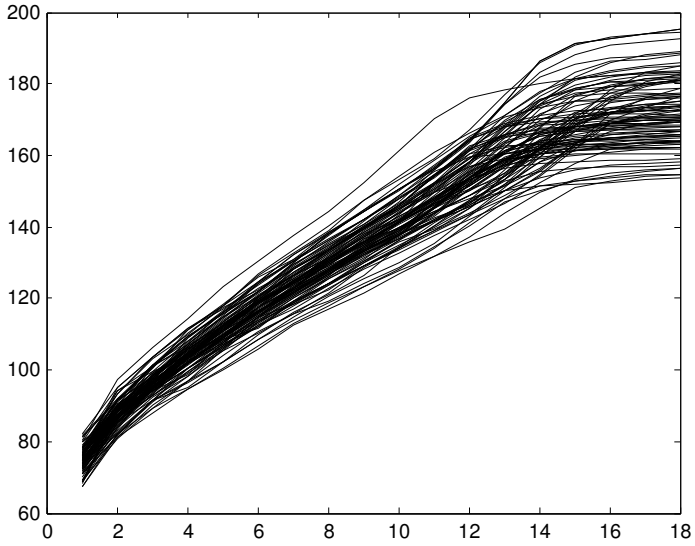

FIGURE 3.4 – *Représentation des courbes de croissance.*

Nous initialisons notre algorithme avec des valeurs $\beta = \beta_0 = 1$ et $\sigma = \sigma_0 = 5$ afin d'obtenir une première classification. Cette classification nous permet alors d'estimer, et de fixer les paramètres $\beta = \beta_0 = 0.7$ et $\sigma = \sigma_0 = 9.5$. La loi *a priori* gamma sur α_0 est telle que $a = 1$ et $b = 0.5$.

Nous souhaitons comparer la classification obtenue par notre algorithme à la classification naturelle induite par le genre de l'individu (garçon/fille). Sur 50 répétitions, notre algorithme affiche un TCC moyen de 70.97% (par rapport au genre) et a toujours su retrouver une classification constituée de 2 classes. Le temps moyen par itération est de 10s. Ce jeu de données ayant déjà été étudié précédemment en détails par Jacques & Preda [54], nous reportons à titre de comparaison les TCC d'autres méthodes dans le tableau 3.1. Dans leur étude, les auteurs comparent la méthode qu'ils proposent (funclust) à trois autres méthodes fonctionnelles qui sont celles de James & Sugar (fclust) [55], Chiou & Li (kCFC) [17] et Bouveyron & Jacques (funHDDC) [10]. Toutes ces méthodes sont parfaitement adaptées au cas de courbes. Les auteurs considèrent également des méthodes fini-dimensionnelles, qui ont été appliquées sur les scores d'une analyse en composantes principales fonctionnelle, sur les observations discrétisées et aussi sur les coefficients d'une décomposition en splines cubiques. Le lecteur pourra se référer à l'article de Jacques & Preda [54] pour plus de détails.

Taux de classification correcte (TCC)					
Méthodes fonctionnelles		**Méthodes non fonctionnelles**			
			Données discrétisées	**Splines cubiques**	**Scores ACPF**
DPMF	**70.97%**	HDDC	56.99%	50.51%	97.85%
fclust	69.89%	MixtPPCA	62.36%	50.53%	97.85%
kCFC	93.55%	GMM	65.59%	63.44%	95.70%
funHDDC	96.77%	k-means	65.59%	66.67%	64.52%
funclust	69.98%	hclust	51.61%	75.27%	68.81%

TABLE 3.1 – *TCC obtenus sur différentes méthodes pour le jeu de données de courbes de croissance. Hormis pour le (DPMF), les résultats ont été obtenus par Jacques & Preda [54].*

Il est intéressant de noter que les courbes mal classées (par rapport au genre) ont des formes similaires à la classe à laquelle elles appartiennent. Par exemple, les courbes de croissance des filles assignées à la classe "garçons" sont similaires aux courbes de croissance des garçons de cette classe.

Notons néanmoins que Jacques & Preda ont comparé leur méthode sur d'autres jeux de données et que les résultats sont parfois inversés selon les situations. Les méthodes kCFC et funHDDC peuvent afficher des taux très bons, mais l'inconvénient est qu'elles sont toutes basées sur des modèles qui demandent de choisir une base d'approximation, ce que notre méthode ne requiert pas. Rappelons en effet que nous n'avons ni à choisir le nombre de classes ni la base d'approximation. Cette spécificité du (DPMF) permet de le rendre indépendant de toute base d'approximation, ce qui n'est pas le cas des autres méthodes.

3.6.4 Jeu de données de spectrométrie

Ce dernier jeu de données réelles est issu du projet SpecBio, qui a pour objectif de caractériser des sols à l'aide de la spectrométrie infrarouge. Il est constitué de 78 courbes qui sont des indicateurs spectraux de caractéristiques biologiques de sol, mesurés de 350nm à 2500nm et avec un pas de 1nm. Tout l'intérêt ici consiste à produire des classes de courbes correspondant à des caractéristiques bien précises du sol. Les courbes sont représentées figure 3.5.

Nous initialisons notre algorithme avec des valeurs $\beta = \beta_0 = 0.5$ et $\sigma = \sigma_0 = 0.01$ afin d'obtenir une première classification, nous permettant d'estimer et fixer $\beta = \beta_0 = 0.7$ et $\sigma = \sigma_0 = 0.14$. La loi *a priori* gamma sur α_0 est telle que $a = 1$ et $b = 0.5$. La figure 3.6 présente les résultats de classification.

Pour ce jeu de données, nous ne pouvons pas donner de TCC car nous ne connaissons pas les vraies classes. Notre algorithme pour le (DPMF) a su trouver une classification constituée de 3 classes et le temps moyen par itération est de 63s. Afin d'étudier la pertinence de ces résultats, nous avons reporté la même classification sur les données de masse de carbone organique présente dans le sol. En effet, à chaque courbe est associée une valeur de masse en carbone organique. La figure 3.7 présente ces résultats sous forme de diagramme en boîte. La nuance de couleurs de chaque diagramme correspond à la classe associée. On constate qu'il existe un lien entre nos résultats et les données carbone. Des études plus approfondies pourraient être menées afin de démontrer le lien entre la classification obtenue et les caractéristiques attendues des sols.

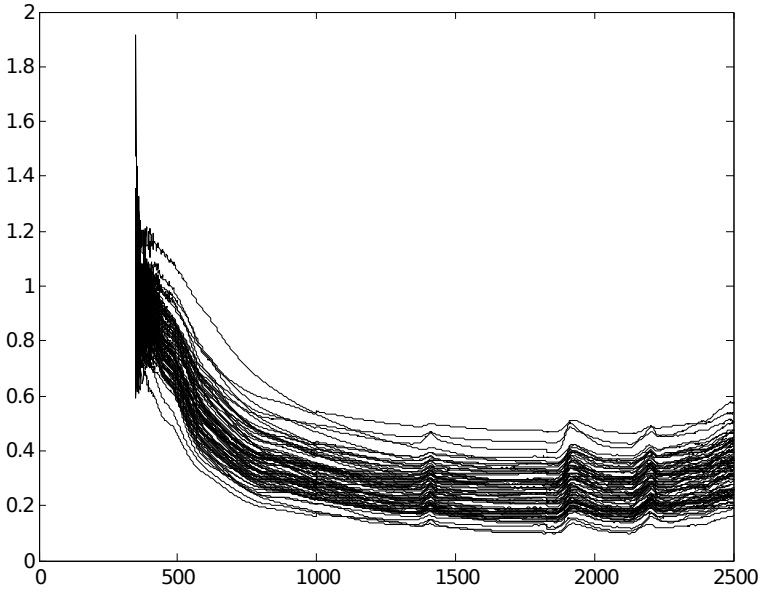

FIGURE 3.5 – *Représentation des courbes de spectrométrie.*

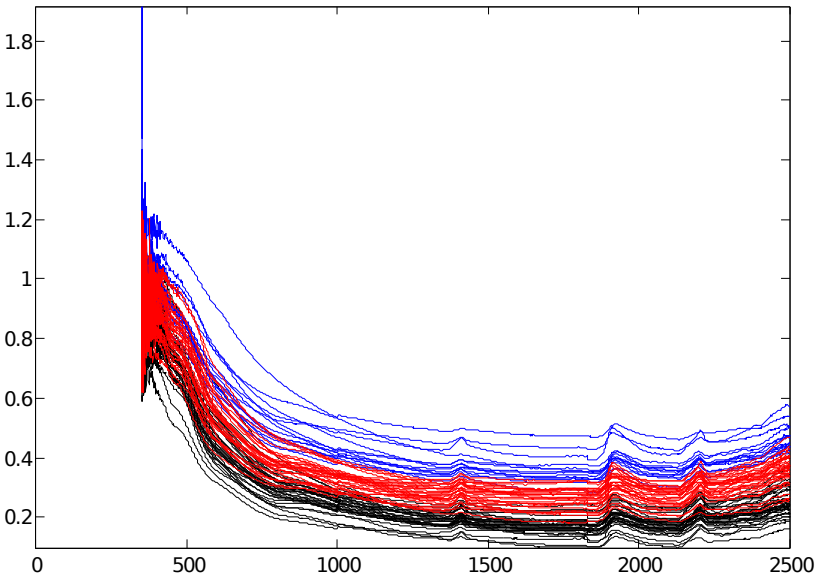

FIGURE 3.6 – *Résultats de classification, toutes classes confondues.*

3.6.5 Discussion

Le modèle (DPMF) que nous proposons offre une approche fonctionnelle de classification de courbes et s'est révélé capable de classer des courbes observées en un très grand nombre de points. Nous avons également voulu le comparer à la méthode de Jackson et al. [53]. Leur modèle est un (DPM) appliqué aux processus gaussiens, mais où les auteurs proposent un modèle finidimensionnel. En effet, ils ne considèrent pas les courbes en tant qu'objets de dimension infinie mais discrétisées aux temps d'observation. Leur modèle de classification opère à la fois sur la moyenne et le noyau de covariance, avec des covariances entre classes pouvant être différentes. Pour des raisons de temps, nous n'avons pu produire assez d'itérations pour conclure quant à

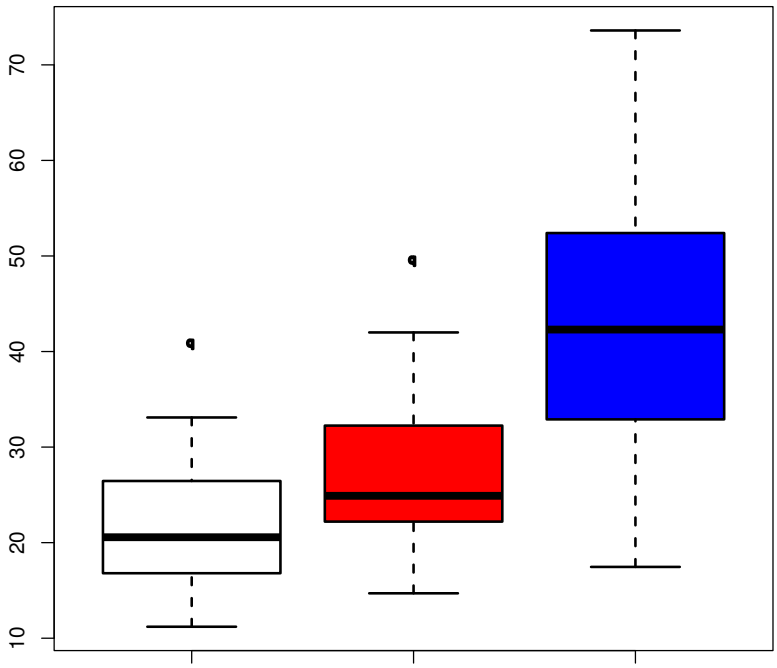

FIGURE 3.7 – *Résultats de la classification obtenue sur la masse de carbone organique (g/kg). Chaque diagramme correspond à une classe.*

leur méthode, car une itération demande environ 6h39 et presque 1 mois serait nécessaire pour produire une centaine d'itérations.

Notre modèle considère les courbes en dimension infinie. Une étape de discrétisation est nécessaire pour calculer les produits scalaires (3.6.2) entre les courbes. L'avantage de notre approche réside dans le fait que le pas de discrétisation peut être choisi relativement large, dans la mesure où le produit scalaire entre les courbes discrétisées donne une approximation satisfaisante. Notre expérience montre que ce pas est souvent beaucoup plus grossier que le pas des courbes réellement observées, d'où un gain substantiel de temps de calcul numérique. Par ailleurs dans notre cas, les densités des courbes sont exprimées relativement à une mesure gaussienne, car il n'existe pas d'analogue à la mesure de Lebesgue en dimension infinie. Ceci permet d'éviter quelques instabilités numériques, notamment dans le calcul du déterminant de matrice de variance-covariance en grande dimension.

L'utilisation du *MAP* global pour la classification retenue n'est pas le plus efficace, mais il présente l'avantage de ne pas nécessiter d'outils compliqués pour le calculer. En effet, à l'inverse du *MAP* marginal, ce dernier ne nécessite pas l'utilisation de pivot. Nous sommes conscients de l'inconvénient de l'utilisation du *MAP* global, mais tenons à rappeler que le travail réalisé dans cette thèse se veut davantage être un début d'application de la théorie RKHS en classification, plutôt qu'une mise en œuvre optimisée d'une méthode algorithmique. D'autant plus que l'implémentation numérique est déjà conséquente, en raison du calcul des produits scalaires et de la simulation suivant le processus *a posteriori* $\mathrm{P}_{m,\mathrm{K}}$ de l'équation (3.5.2).

Nous utilisons les processus de Dirichlet pour classer des données fonctionnelles, mais d'autres processus comme celui de Pitman-Yor sont aussi utilisés pour la classification non supervisée. Ces processus permettent un choix automatique du nombre de classes, mais ce choix est sensible aux hyperparamètres du processus. Plusieurs articles [81, 82] soulignent l'inconsistance des processus de Dirichlet et Pitman-Yor pour inférer sur le nombre de classes, lorsque les hy-

perparamètres sont fixes. Aussi, pour limiter ces problèmes, on pose une loi *a priori* sur ces paramètres et ils sont inférés *a posteriori*. Notons que Kimura et al. [61] obtiennent de bons résultats de classification en inférant sur α_0 à partir d'un algorithme EM.

Remarquons pour finir qu'il est possible de généraliser notre méthode au cas où chaque courbe Y_i disposerait de sa propre fonction de covariance Σ_i. En généralisant alors le (DPMF) pour inférer à la fois sur les moyennes et les paramètres du noyau de covariance, on obtiendrait des covariances Σ_i égales dans chaque classe. Notre méthode est alors directement généralisable, avec $\prod_{i=1}^{n} P_{0,\Sigma_i}$ comme mesure commune pour la vraisemblance, et un résultat similaire pour la loi décrite dans l'équation (3.5.1).

CHAPITRE

4

Classification de données fonctionnelles avec covariables

Résumé

Dans le chapitre précédent, nous avons vu comment classer des données fonctionnelles à partir d'un modèle basé sur le processus de Dirichlet. Ce dernier modèle, que nous avons appelé (DPMF), permettait de travailler en dimension infinie et de réaliser les calculs sur les courbes complètes, par le biais des processus gaussiens. Nous rajoutons ici une hiérarchie supplémentaire nous permettant de pouvoir prendre en compte différentes covariables. La même méthode algorithmique que celle du (DPMF) est utilisée pour l'implémentation, et des cas pratiques sur données simulées sont étudiés puis discutés.

4.1 Objectifs

A présent, nous souhaitons prendre en compte dans notre modèle de classification des covariables liées aux données. Ainsi, chaque courbe Y_i est désormais associée à une covariable fonctionnelle Z_i. Il s'agit donc de construire un modèle linéaire fonctionnel dont à la fois la réponse Y_i et la covariable Z_i sont des fonctions.

Comme cela a été fait dans le chapitre précédent, nous ne considérons que des processus dont les trajectoires appartiennent à $L^2([0,T])$.

4.2 Classification de données fonctionnelles avec covariables

A notre connaissance, la littérature sur la classification de données fonctionnelles prenant en compte des covariables est succincte. Certains auteurs [116, 137] se sont penchés sur le problème en statistique non bayésienne. Dans un premier temps, Shi & Wang [116] ont cherché à reconstruire une courbe à partir de covariables fonctionnelles. En notant $Z = (Z_1, \ldots, Z_Q)$ le vecteur de covariables, leur modèle est le suivant :

$$\begin{array}{rcl} Y_i(t,Z)|c_i = k & = & \mu_{ik}(t) + \tau_{ik}(Z) + \epsilon_i, \\ \mu_{ik}(t) & = & u_i^\top \beta_k(t), \\ \tau_{ik}(Z) & \sim & P_{Z,\theta_k}, \\ \epsilon_i & \sim & P_{0,C}. \end{array}$$

et les auteurs posent une loi logistique sur le vecteur $(c_1, \ldots, c_n)$. Les moyennes β_k sont décomposées dans une base de fonctions splines. Les paramètres du modèle sont inférés à l'aide d'un l'algorithme EM puis, dans une optique de classification, une courbe est affectée à la classe dont la probabilité *a posteriori* est la plus élevée.

Yi et al. [137] proposent de faire de la classification en même temps que de la sélection de covariables afin d'ajuster au mieux le modèle. S'alternent alors un procédé de validation croisée permettant de sélectionner les covariables, ainsi qu'une étape de minimisation d'une vraisemblance pénalisée pour estimer les paramètres du modèle. Les covariables sont finidimensionnelles.

En statistique bayésienne, la prise en compte de covariables a été étudiée par d'autres auteurs [27, 36]. Les covariables sont le plus souvent scalaires et non fonctionnelles. Dunson, Herring & Siega-Riz [36] proposent une application médicale pour classer des courbes de prise de poids Y_i pendant la grossesse en fonction du poids de l'enfant Z_i à la naissance. Cette méthode est très proche de celle de Ray & Mallick [105], mais afin de lier les données aux Z_i, une hiérarchie supplémentaire est proposée. L'implémentation numérique alterne des étapes SUGS (*Sequential Updating with Greedy Search*) avec un échantillonneur de Gibbs.

4.3 Résultats théoriques nécessaires

Avant de donner plus de détails sur la construction du modèle, nous devons démontrer un théorème important pour la suite, qui n'est autre que la généralisation du théorème 2.7. Nous souhaitons calculer la fonction moyenne ainsi que la fonction de covariance d'un processus gaussien transformé par une application linéaire. C'est cette application linéaire que nous retrouverons dans la suite de ce chapitre.

Théorème 4.1
Soient $\mathrm{X} \sim \mathrm{P}_{m,\mathrm{W}}$, $\epsilon \sim \mathrm{P}_{0,\mathrm{C}}$ *et* Z *une fonction déterministe, continue et non nulle sur* $[0,\mathrm{T}]$. *Supposons que* X *et* ϵ *sont indépendants et que* m *et* W *sont deux fonctions continues, respectivement sur* $[0,\mathrm{T}]$ *et* $[0,\mathrm{T}]\times[0,\mathrm{T}]$. *Supposons que le processus* X *est à trajectoires continues p.s.. Définissons pour chaque* $t \in [0,\mathrm{T}]$, $\mathscr{X}(t) = \int_0^t \mathrm{X}(u)\mathrm{Z}(u)\,du$ *et notons enfin* $\mathrm{Y} = \mathscr{X} + \epsilon$.

Alors, la loi a posteriori X|Y *est également un processus gaussien* $\mathrm{P}_{\overline{m},\overline{\mathrm{W}}}$, *donné par :*

$$\begin{array}{rcl} \overline{m}(t) & = & m(t) + (\mathrm{I}_t, \mathrm{Y} - \widetilde{m})_{\widetilde{\mathrm{W}}+\mathrm{C}}, \\ \overline{\mathrm{W}}(s,t) & = & \mathrm{W}(s,t) - (\mathrm{I}_t, \mathrm{I}_s)_{\widetilde{\mathrm{W}}+\mathrm{C}}, \\ \widetilde{m}(t) & = & \int_0^t m(u)\mathrm{Z}(u)\,du, \\ \widetilde{\mathrm{W}}(s,t) & = & \int_0^s \int_0^t \mathrm{W}(u,v)\mathrm{Z}(u)\mathrm{Z}(v)\,du\,dv, \end{array}$$

et où l'on définit la fonction $\mathrm{I}_t : x \mapsto \int_0^x \mathrm{W}(u,t)\mathrm{Z}(u)\,du$.

PREUVE. Commençons par déterminer la loi de $\mathscr{X}|\mathrm{Y}$. D'après le lemme A.2 en annexe A.6, nous savons que $\mathscr{X}$ est un processus gaussien de moyenne $\widetilde{m}$ et de fonction de covariance $\widetilde{\mathrm{W}}$. De plus, les processus $\mathscr{X}$ et ϵ sont indépendants car les processus X et ϵ le sont. En effet, écrivons pour tous réels $s, t \in [0,\mathrm{T}]$:

$$\mathrm{Cov}(\mathscr{X}(s), \epsilon(t)) = \mathbb{E}\mathscr{X}(s)\epsilon(t) - \mathbb{E}\mathscr{X}(s)\mathbb{E}\epsilon(t) = \mathbb{E}\mathscr{X}(s)\epsilon(t),$$

car ϵ est un processus centré. Ainsi, nous avons par définition de $\mathscr{X}$:

$$Cov(\mathscr{X}(s),\epsilon(t)) = \mathbb{E}\int_0^s \mathrm{X}(u)\mathrm{Z}(u)\epsilon(t)\,du.$$

Utilisons le théorème de Fubini afin de montrer que l'on peut intervertir l'espérance et l'intégrale. Pour cela, il suffit de montrer que pour tout réel $s \in [0,\mathrm{T}]$:

$$\int_0^s \mathbb{E}\Big|\mathrm{X}(u)\mathrm{Z}(u)\epsilon(t)\Big|\,du < \infty.$$

Soit donc $s \in [0,\mathrm{T}]$ fixé. En appliquant l'inégalité de Hölder avec la fonction Z fonction qui est non aléatoire, écrivons pour tout réel $u \in [0,s]$:

$$\mathbb{E}\Big|\mathrm{X}(u)\mathrm{Z}(u)\epsilon(t)\Big| \leq \Big|\mathrm{Z}(u)\Big|\sqrt{\mathbb{E}\mathrm{X}(u)^2}\sqrt{\mathbb{E}\epsilon(t)^2}.$$

La quantité $\mathbb{E}\mathrm{X}(u)^2$ existe car $\mathrm{X}(u)$ est une variable aléatoire gaussienne. En faisant usage des relations $\mathbb{E}\mathrm{X}(u)^2 = m^2(u) + \mathrm{W}(u,u)$ et $\mathbb{E}\epsilon(t)^2 = \mathrm{C}(t,t)$, nous pouvons écrire :

$$\mathbb{E}\Big|\mathrm{X}(u)\mathrm{Z}(u)\epsilon(t)\Big| \leq \Big|\mathrm{Z}(u)\Big|\sqrt{m^2(u)+\mathrm{W}(u,u)}\sqrt{\mathrm{C}(t,t)}.$$

Les quantités ci-dessus étant positives et l'inégalité étant valable pour tout réel $u \in [0,s]$, nous pouvons également écrire :

$$\int_0^s \mathbb{E}\Big|\mathrm{X}(u)\mathrm{Z}(u)\epsilon(t)\Big|\,du \leq \sqrt{\mathrm{C}(t,t)}\int_0^s \Big|\mathrm{Z}(u)\Big|\sqrt{m^2(u)+\mathrm{W}(u,u)}\,du.$$

Dans cette inégalité, les fonctions intervenant dans l'intégrale du membre de droite étant par hypothèse continues sur $[0,s]$, leur intégrale sur ce domaine est finie. D'où :

$$\int_0^s \mathbb{E}\Big|\mathrm{X}(u)\mathrm{Z}(u)\epsilon(t)\Big|\,du < \infty.$$

Ainsi, par application du théorème de Fubini :

$$\begin{array}{rcl} Cov(\mathscr{X}(s),\epsilon(t)) & = & \mathbb{E}\int_0^s \mathrm{X}(u)\mathrm{Z}(u)\epsilon(t)\,du, \\ & = & \int_0^s \mathbb{E}\mathrm{X}(u)\mathrm{Z}(u)\epsilon(t)\,du, \\ & = & \int_0^s \mathrm{Z}(u)\mathbb{E}\mathrm{X}(u)\epsilon(t)\,du, \\ & = & 0. \end{array}$$

Ainsi $Cov(\mathscr{X}(s),\epsilon(t)) = 0$ et il s'ensuit que les processus X et ϵ sont indépendants car il s'agit de processus gaussiens non corrélés. En utilisant alors le théorème 2.7, on déduit la loi *a posteriori* suivante :

$$\mathscr{X}|\mathrm{Y} \sim \mathrm{P}_{m^\star,\mathrm{W}^\star},$$

où pour tous réels $s,t \in [0,\mathrm{T}]$:

$$\begin{array}{rcl} m^\star(t) & = & \widetilde{m}(t) + \big(\widetilde{\mathrm{W}}(\bullet,t),\mathrm{Y}-\widetilde{m}\big)_{\widetilde{\mathrm{W}}+\mathrm{C}}, \\ \mathrm{W}^\star(s,t) & = & \widetilde{\mathrm{W}}(s,t) - \big(\widetilde{\mathrm{W}}(\bullet,s),\widetilde{\mathrm{W}}(\bullet,t)\big)_{\widetilde{\mathrm{W}}+\mathrm{C}}, \\ \widetilde{m}(t) & = & \int_0^t m(u)\mathrm{Z}(u)\,du, \\ \widetilde{\mathrm{W}}(s,t) & = & \int_0^s\int_0^t \mathrm{W}(u,v)\mathrm{Z}(u)\mathrm{Z}(v)\,du\,dv. \end{array}$$

Notons à présent $\mathscr{X}_\mathrm{Y}$ ce processus $\mathscr{X}|\mathrm{Y}$. D'après les résultats de Parzen [95] page 83, le processus dérivé $\frac{d\mathscr{X}_\mathrm{Y}}{dt}$ existe si, et seulement si :

– $m^\star$ est dérivable sur $[0,\mathrm{T}]$,
– $\mathrm{W}^\star$ est deux fois différentiable et $\frac{\partial^2}{\partial s \partial t}\mathrm{W}^\star$ est continue sur $[0,\mathrm{T}]\times[0,\mathrm{T}]$.
Pour le premier point concernant la dérivabilité de la fonction $m^\star$, nous avons de bonnes raisons de conjecturer que c'est le cas, d'après les résultats de Parzen [95] page 83.

D'une part, les fonctions $\widetilde{m}$ et $\widetilde{\mathrm{W}}$ sont clairement dérivables sur $[0,\mathrm{T}]$, car elles sont définies par le biais d'intégrales. D'autre part, pour montrer que $m^\star$ est dérivable, il nous faut dériver le produit scalaire $\left(\widetilde{\mathrm{W}}(\bullet,t),\mathrm{Y}-\widetilde{m}\right)_{\widetilde{\mathrm{W}}+\mathrm{C}}$. Comme la fonction $t\mapsto\widetilde{\mathrm{W}}(\bullet,t)$ est intégrée et que l'opérateur de dérivation est linéaire, nous conjecturons raisonnablement que l'on peut écrire :

$$\frac{\partial}{\partial t}\left(\widetilde{\mathrm{W}}(\bullet,t),\mathrm{Y}-\widetilde{m}\right)_{\widetilde{\mathrm{W}}+\mathrm{C}}=\left(\frac{\partial}{\partial t}\widetilde{\mathrm{W}}(\bullet,t),\mathrm{Y}-\widetilde{m}\right)_{\widetilde{\mathrm{W}}+\mathrm{C}}.$$

De plus, si nous définissons pour chaque $t\in[0,\mathrm{T}]$ la suite de fonctions $\left(f_n^t\right)_{n\geq 1}$ de la manière suivante :

$$\forall x\in[0,\mathrm{T}],\ f_n^t(x)=n\times\left(\widetilde{\mathrm{W}}(t+\frac{1}{n},x)-\widetilde{\mathrm{W}}(t,x)\right),$$

cette suite de fonctions est dans l'espace $\mathrm{H}(\widetilde{\mathrm{W}})$. Or, par application du théorème 12 de Berlinet et al. [6], $\mathrm{H}(\widetilde{\mathrm{W}})\subset\mathrm{H}(\widetilde{\mathrm{W}}+\mathrm{C})$ parce que $(\widetilde{\mathrm{W}}+\mathrm{C})-\mathrm{C}=\widetilde{\mathrm{W}}$ est une fonction de covariance. Il s'agit donc d'une suite de fonctions dans $\mathrm{H}(\widetilde{\mathrm{W}}+\mathrm{C})$. En notant pour une fonction f quelconque, $(\mathrm{D}^1 f)_t(\bullet)=\frac{\partial}{\partial t}f(t,\bullet)$, on a clairement $\lim_{n\longrightarrow+\infty}f_n^t=(\mathrm{D}^1\widetilde{\mathrm{W}})_t$.

Or, par application du lemme 1 de Sun & Zhou [124], comme $\widetilde{\mathrm{W}}$ est continue sur $[0,\mathrm{T}]\times[0,\mathrm{T}]$, on a $(\mathrm{D}^1\widetilde{\mathrm{W}})_t\in\mathrm{H}(\widetilde{\mathrm{W}})$. Une fois de plus, comme $\mathrm{H}(\widetilde{\mathrm{W}})\subset\mathrm{H}(\widetilde{\mathrm{W}}+\mathrm{C})$, cela implique $(\mathrm{D}^1\widetilde{\mathrm{W}})_t\in\mathrm{H}(\widetilde{\mathrm{W}}+\mathrm{C})$.

Ainsi, la convergence de la suite de fonctions $\left(f_n^t\right)_{n\geq 1}$ vers la fonction $(\mathrm{D}^1\widetilde{\mathrm{W}})_t$ est une convergence simple. Nous conjecturons que la convergence est également faible, ce qui permet d'introduire la limite à l'intérieur du produit scalaire et d'obtenir le résultat souhaité sur la dérivation.

Pour le second point, concernant la différentiabilité de $\mathrm{W}^\star$, le raisonnement est identique car la fonction $t\mapsto\widetilde{\mathrm{W}}(\bullet,t)$ est deux fois différentiable sur $[0,\mathrm{T}]\times[0,\mathrm{T}]$.

Nous pouvons ainsi définir le processus $\mathscr{X}^\star$ de la façon suivante :

$$\mathscr{X}^\star(t)=\frac{d\mathscr{X}_\mathrm{Y}}{dt}(t).$$

La limite $\mathscr{X}^\star(t)=\lim_{n\to+\infty}n\left(\mathscr{X}_\mathrm{Y}\left(t+\frac{1}{n}\right)-\mathscr{X}_\mathrm{Y}(t)\right)$ est définie en moyenne quadratique. D'autre part, nous remarquons que la variable aléatoire $\mathscr{X}^\star(t)$ est gaussienne. En effet, d'après le lemme A.1 en annexe A.6, il s'agit de la limite en moyenne quadratique de variables aléatoires gaussiennes. De même, pour tout n-uplet $(t_1,\ldots,t_n)\in[0,\mathrm{T}]^n$, le vecteur $\left(\mathscr{X}^\star(t_1),\ldots,\mathscr{X}^\star(t_n)\right)$ est gaussien et il en résulte que $\mathscr{X}^\star$ est un processus gaussien.

Ce processus étant entièrement défini par sa fonction moyenne et sa fonction covariance, nous calculons pour tous réels fixés $s,t\in[0,\mathrm{T}]$ d'après les résultats de Parzen [95] page 83 :

*1) sa moyenne :

$$\begin{aligned}
\mathbb{E}\mathscr{X}^\star(t) &= \mathbb{E}\left(\tfrac{d\mathscr{X}_\mathrm{Y}}{dt}(t)\right),\\
&= \tfrac{d}{dt}\mathbb{E}\left(\mathscr{X}(t)|\mathrm{Y}\right),\\
&= \tfrac{d}{dt}m^\star(t),\\
&= m(t)\mathrm{Z}(t)+\left(\tfrac{d}{dt}\widetilde{\mathrm{W}}(\bullet,t),\mathrm{Y}-\widetilde{m}\right)_{\widetilde{\mathrm{W}}+\mathrm{C}},\\
&= m(t)\mathrm{Z}(t)+(\mathrm{Z}(t)\mathrm{I}_t,\mathrm{Y}-\widetilde{m})_{\widetilde{\mathrm{W}}+\mathrm{C}},\\
&= \mathrm{Z}(t)\left[m(t)+(\mathrm{Z}(t)\mathrm{I}_t,\mathrm{Y}-\widetilde{m})_{\widetilde{\mathrm{W}}+\mathrm{C}}\right],
\end{aligned}$$

*2) sa covariance :

$$\begin{aligned} Cov(\mathscr{X}^\star(s),\mathscr{X}^\star(t)) &= \frac{\partial^2}{\partial s\partial t}\mathrm{W}^\star(s,t), \\ &= \mathrm{W}(s,t)\mathrm{Z}(s)\mathrm{Z}(t) - (\mathrm{Z}(t)\mathrm{I}_t,\mathrm{Z}(s)\mathrm{I}_s)_{\widetilde{\mathrm{W}}+\mathrm{C}}, \\ &= \mathrm{Z}(s)\mathrm{Z}(t)\Big[\mathrm{W}(s,t) - (\mathrm{I}_t,\mathrm{I}_s)_{\widetilde{\mathrm{W}}+\mathrm{C}}\Big]. \end{aligned}$$

De plus, comme nous avons la relation $\mathrm{Z}(t) \neq 0$ pour tout $t \in [0,\mathrm{T}]$, il vient l'égalité :

$$\mathrm{X}(t) = \frac{1}{\mathrm{Z}(t)}\frac{d\mathscr{X}}{dt}(t).$$

Nous pouvons donc conclure quant à la loi de X|Y :

$$\mathscr{L}(\mathrm{X}|\mathrm{Y}) = \mathscr{L}\left(\frac{1}{\mathrm{Z}}\mathscr{X}^\star\right) = \mathrm{P}_{\overline{m},\overline{\mathrm{W}}},$$

où :

$$\begin{aligned} \overline{m}(t) &= m(t) + (\mathrm{I}_t,\mathrm{Y}-\widetilde{m})_{\widetilde{\mathrm{W}}+\mathrm{C}}, \\ \overline{\mathrm{W}}(s,t) &= \mathrm{W}(s,t) - (\mathrm{I}_t,\mathrm{I}_s)_{\widetilde{\mathrm{W}}+\mathrm{C}}, \end{aligned}$$

avec les notations données précédemment. ■

4.4 Un premier modèle avec covariable et son implémentation

4.4.1 Présentation du modèle

A l'instar de Ramsay & Silverman [102], nous proposons de modéliser les observations Y_i sous la forme intégrée suivante :

$$\mathrm{Y}_i(t) = \int_0^{\mathrm{T}} \mathrm{Z}_i(u)\beta_i(u,t)du + \epsilon_i(t).$$

Cependant, le plus souvent la covariable Z_i intervient jusque l'instant t et jamais à "reculons" dans le temps. Ainsi, l'influence de Z_i au temps t n'intervient que jusque cet instant, et il est donc plus naturel de décomposer chaque fonction β_i sous la forme suivante :

$$\beta_i(u,t) = \beta_i(u)\mathbb{I}_{[0,t]}(u),$$

où $\mathbb{I}_\mathrm{A}$ désigne la fonction indicatrice d'un ensemble A. Ceci implique la modélisation suivante de la i^e observation :

$$\mathrm{Y}_i(t) = \int_0^t \beta_i(u)\mathrm{Z}_i(u)du + \epsilon_i(t).$$

Nous pouvons interpréter β_i comme un effet individuel pour l'observation i. Plus précisément, au temps t, $\beta_i(u)$ correspond au poids donné à la covariable Z_i à l'instant u dans la prédiction de $\mathrm{Y}_i(t)$. Souhaitant classer les observations Y_i au travers des fonctions β_i, nous proposons une généralisation de notre modèle (DPMF). Le nouveau modèle, que nous notons (DPMFc), est défini de la façon suivante :

$$(\mathrm{DPMF}c)\begin{cases} \mathrm{Y}_i(t) = \int_0^t \beta_i(u)\mathrm{Z}_i(u)du + \epsilon_i(t), \\ \epsilon_i \overset{ind}{\sim} \mathrm{P}_{0,\mathrm{C}}, \\ (\mathrm{DPMF})\begin{cases} \beta_i|\theta_i \overset{ind}{\sim} \mathrm{P}_{\theta_i,\Sigma}, \\ \theta_i|\mathrm{G} \overset{ind}{\sim} \mathrm{G}, \\ \mathrm{G} \sim \mathrm{DP}(\alpha_0,\mathrm{G}_0), \end{cases} \end{cases}$$

où nous rappelons que la notation $\mathrm{P}_{m,\mathrm{K}}$ désigne le processus gaussien de fonction moyenne m et de fonction de covariance K et où ϵ_i est un processus de bruit, indépendant de β_i. Dans ce modèle, on peut intégrer sur les fonctions β_i afin d'obtenir une représentation équivalente, similaire à celle du (DPMF). En effet, définissons pour chaque entier $i \in \{1,\dots,n\}$ le processus $\mathscr{X}_i$ de la manière suivante :

$$\mathscr{X}_i(t) = \int_0^t \beta_i(u)\mathrm{Z}_i(u)du.$$

Par application du lemme A.2 en annexe A.6, $\mathscr{X}_i$ est un processus gaussien $\mathrm{P}_{m_i,\mathrm{W}_i}$, où pour tous $s,t \in [0,\mathrm{T}]$:

$$\begin{array}{rcl} m_i(t) & = & \int_0^t \theta_i(u)\mathrm{Z}_i(u)du, \\ \mathrm{W}_i(s,t) & = & \int_0^s \int_0^t \Sigma(u,v)\mathrm{Z}_i(u)\mathrm{Z}_i(v)dudv. \end{array}$$

Les processus gaussiens $\mathscr{X}_i$ et ϵ_i sont de plus indépendants car non corrélés. En effet, pour $s,t \in [0,\mathrm{T}]$ quelconques, $Cov(\beta_i(s),\epsilon_i(t)) = 0$ implique $Cov(\mathscr{X}_i(s),\epsilon_i(t)) = 0$, car la covariance est une application bilinéaire. Ce modèle est donc équivalent au suivant :

$$\left\{ \begin{array}{ccc} \mathrm{Y}_i(t) & = & \mathscr{X}_i(t) + \epsilon_i(t), \\ \epsilon_i & \overset{ind}{\sim} & \mathrm{P}_{0,\mathrm{C}}, \\ \mathscr{X}_i|\theta_i & \overset{ind}{\sim} & \mathrm{P}_{m_i,\mathrm{W}_i}, \\ \theta_i|\mathrm{G} & \overset{ind}{\sim} & \mathrm{G}, \\ \mathrm{G} & \sim & \mathrm{DP}(\alpha_0,\mathrm{G}_0), \end{array} \right.$$

Une intégration sur les fonctions $\mathscr{X}_i$ nous permet alors d'aboutir au modèle équivalent suivant, par sommation de deux processus gaussiens indépendants :

$$\left\{ \begin{array}{ccc} \mathrm{Y}_i|\theta_i & \overset{ind}{\sim} & \mathrm{P}_{m_i,\mathrm{W}_i+\mathrm{C}}, \\ \theta_i|\mathrm{G} & \overset{ind}{\sim} & \mathrm{G}, \\ \mathrm{G} & \sim & \mathrm{DP}(\alpha_0,\mathrm{G}_0). \end{array} \right.$$

Finalement, d'après le premier chapitre, ce modèle est encore équivalent au suivant :

$$\left\{ \begin{array}{ccc} \mathrm{Y}_i|c_i,\phi_c & \overset{ind}{\sim} & \mathrm{P}_{m_i,\mathrm{W}_i+\mathrm{C}}, \\ (c_1,\dots,c_n) & \sim & \mathrm{CRP}(\alpha_0), \\ \phi_c|\mathrm{G}_0 & \overset{ind}{\sim} & \mathrm{G}_0, \end{array} \right.$$

où nous notons de même $m_i(t) = \int_0^t \phi_{c_i}(u)\mathrm{Z}_i(u)du$.

Pour des raisons pratiques, nous supposons que $\mathrm{G}_0 = \mathrm{P}_{0,\Sigma_0}$. En effet, d'après les conclusions expérimentales du chapitre précédent, la moyenne de G_0 ne joue aucun rôle dans la classification. Nous supposons que les paramètres α_0, C, Σ et Σ_0 sont fixés dans un premier temps. Nous verrons plus tard comment nous pouvons estimer ces paramètres.

4.4.2 Implémentation algorithmique

Nous souhaitons appliquer la méthode inférentielle du chapitre précédent, valable dans le cadre du (DPMF), au cas du modèle avec covariable (DPMFc). Rappelons que cet algorithme consiste à inférer sur les paramètres $(c_1,\dots,c_n)$ et $(\phi_1,\dots,\phi_{n+m})$, où $m \geq 1$ est un entier arbitrairement choisi. Nous l'adaptons à présent au cas où la mesure de référence est $\mathrm{P}_{0,\mathrm{W}_i+\mathrm{C}}$. Rappelons également que l'implémentation de cet algorithme requiert :

– le calcul numérique des densités $p(\mathrm{Y}_i|c,\phi_c)$, pour tout i et pour tout c, relativement à une même mesure de référence,
– la simulation suivant la loi *a posteriori* de chaque ϕ_c sachant les données qui sont dans la classe c.

Pour le premier point, d'après le théorème 2.2, nous savons que chaque processus gaussien $\mathrm{P}_{m_i,\mathrm{W}_i+\mathrm{C}}$ admet une densité par rapport au processus $\mathrm{P}_{0,\mathrm{W}_i+\mathrm{C}}$, si, et seulement si, $m_i \in \mathrm{H}(\mathrm{W}_i+\mathrm{C})$. Or, par application du théorème 12 de Berlinet et al. [6], $\mathrm{H}(\mathrm{C}) \subset \mathrm{H}(\mathrm{W}_i+\mathrm{C})$ parce que $(\mathrm{W}_i+\mathrm{C})-\mathrm{C}=\mathrm{W}_i$ est une fonction de covariance. Par exemple, si C est la fonction de covariance d'un processus Ornstein-Uhlenbeck, l'espace H(C) est formé des fonctions différentiables sur $[0,\mathrm{T}]$, et comme m_i est une application linéaire intégrale, elle est différentiable sur $[0,\mathrm{T}]$ donc $m_i \in \mathrm{H}(\mathrm{C}) \subset \mathrm{H}(\mathrm{W}_i+\mathrm{C})$. Ainsi, nous pouvons choisir $\mathrm{P}_{0,\mathrm{W}_i+\mathrm{C}}$ comme mesure de référence et nous avons alors :

$$\frac{d\mathrm{P}_{m_i,\mathrm{W}_i+\mathrm{C}}}{d\mathrm{P}_{0,\mathrm{W}_i+\mathrm{C}}}(\mathrm{Y}) = e^{(\mathrm{Y},m_i)_{\mathrm{W}_i+\mathrm{C}}-\frac{1}{2}(m_i,m_i)_{\mathrm{W}_i+\mathrm{C}}}.$$

De plus, nous pouvons calculer numériquement la moyenne m_i ainsi que la covariance $\mathrm{W}_i+\mathrm{C}$. Il ne reste qu'à calculer les différents produits scalaires dans le RKHS lié au noyau $\mathrm{W}_i+\mathrm{C}$. A défaut de pouvoir expliciter le produit scalaire dans l'espace $\mathrm{H}(\mathrm{W}_i+\mathrm{C})$, dans les simulations numériques qui suivent, nous choisissons l'approximation apportée par les théorèmes 2.5 et 2.6 afin de ne pas surcharger l'implémentation algorithmique, qui est déjà conséquente.

Pour le second point, nous avons besoin de simuler suivant la loi *a posteriori* de chaque ϕ_c sachant les données qui sont dans la classe c, loi *a posteriori* associée au modèle suivant :

$$\left\{\begin{array}{ccc} \mathrm{Y}_i(t) & = & \int_0^t \phi(u)\mathrm{Z}_i(u)du+\epsilon_i(t), \\ \phi & \sim & \mathrm{P}_{0,\Sigma_0}, \\ \epsilon_i & \overset{ind}{\sim} & \mathrm{P}_{0,\mathrm{W}_i+\mathrm{C}}. \end{array}\right.$$

En notant $f_i(\phi)$ la fonction $t \mapsto \int_0^t \phi(u)\mathrm{Z}_i(u)du$ et $\mathrm{F}(\phi)=\frac{1}{n_c}\sum_{i:c_i=c} f_i(\phi)$, où $n_c=\#\{j : c_j=c\}$, nous constatons que la fonction F est linéaire en ϕ, et ce d'après l'écriture suivante, valable quel que soit le réel $t \in [0,\mathrm{T}]$:

$$\mathrm{F}(\phi)(t)=\frac{1}{n_c}\sum_{i:c_i=c} f_i(\phi)(t)=\int_0^t\left(\frac{1}{n_c}\sum_{i:c_i=c}\mathrm{Z}_i(u)\right)\phi(u)du.$$

En posant $\overline{\mathrm{Z}}(u)=\frac{1}{n_c}\sum_{i:c_i=c}\mathrm{Z}_i(u)$, on a $\mathrm{F}(\phi)(t)=\int_0^t\overline{\mathrm{Z}}(u)\phi(u)du$. Ainsi, $\mathrm{F}(\phi)$ se définit comme la fonction $t\mapsto\int_0^t\overline{\mathrm{Z}}(u)\phi(u)du$. En utilisant la notation $\overline{\bullet}(t)=\frac{1}{n_c}\sum_{i:c_i=c}\bullet_i(t)$ pour une fonction f quelconque, notre modèle peut également s'écrire :

$$\left\{\begin{array}{ccc} \overline{\mathrm{Y}}(t) & = & \mathrm{F}(\phi)(t)+\overline{\epsilon}(t), \\ \phi & \sim & \mathrm{P}_{0,\Sigma_0}, \\ \overline{\epsilon} & \sim & \mathrm{P}_{0,\overline{\mathrm{W}+\mathrm{C}}}. \end{array}\right.$$

D'après le corollaire 2.1, nous déduisons que la loi de ϕ sachant les données dans la classe c est la même que la loi de ϕ sachant $\overline{\mathrm{Y}}$, et par application du théorème 4.1 :

$$\mathscr{L}\left(\phi|\{\mathrm{Y}_i\}_{i:c_i=c}\right)=\mathrm{GP}(\overline{m},\overline{\mathrm{W}}),$$

avec :

$$\begin{array}{rcl}\overline{m}(t) & = & \left(\mathrm{I}_t, \overline{\mathrm{Y}}\right)_{\widetilde{\mathrm{W}}+\overline{\mathrm{W}+\mathrm{C}}}, \\ \overline{\mathrm{W}}(s,t) & = & \Sigma_0(s,t) - (\mathrm{I}_t, \mathrm{I}_s)_{\widetilde{\mathrm{W}}+\overline{\mathrm{W}+\mathrm{C}}}, \\ \widetilde{\mathrm{W}}(s,t) & = & \int_0^s \int_0^t \Sigma_0(u,v)\overline{\mathrm{Z}}(u)\overline{\mathrm{Z}}(v)du\,dv, \\ \overline{\mathrm{W}+\mathrm{C}} & = & \frac{1}{n_c}\sum_{i:c_i=c}(\mathrm{W}_i+\mathrm{C}),\end{array}$$

où cette fois-ci, $\mathrm{I}_t(x) = \int_0^x \Sigma_0(u,t)\overline{\mathrm{Z}}(u)du$.

4.5 Vers un modèle plus simple

4.5.1 Présentation du modèle

Dans le modèle précédent (DPMFc), on voit bien que les observations sont générées à partir d'une fonction de covariance de la forme $\mathrm{W}_i+\mathrm{C}$, ce qui est du à la présence d'une hiérarchie supplémentaire dans le modèle (DPMF). Or, il pourrait être intéressant de proposer un modèle de classification avec covariable qui soit plus simple. Pour cela, nous choisissons dans cette section le modèle de classification suivant, nommé (DPMFcs) :

$$(\mathrm{DPMF}cs)\begin{cases} \mathrm{Y}_i(t) & = & \int_0^t \theta_i(s)\mathrm{Z}_i(s)ds + \epsilon_i(t), \\ \theta_i|\mathrm{G} & \overset{ind}{\sim} & \mathrm{G}, \\ \mathrm{G} & \sim & \mathrm{DP}(\alpha_0, \mathrm{G}_0). \end{cases}$$

Ce modèle est alors équivalent au suivant, par simple écriture :

$$\begin{cases} \mathrm{Y}_i|\theta_i & \overset{ind}{\sim} & \mathrm{P}_{m_i,\mathrm{C}}, \\ \theta_i|\mathrm{G} & \overset{ind}{\sim} & \mathrm{G}, \\ \mathrm{G} & \sim & \mathrm{DP}(\alpha_0, \mathrm{G}_0), \end{cases}$$

avec $m_i(t) = \int_0^t \theta_i(u)\mathrm{Z}_i(u)du$. D'après le premier chapitre, ce modèle est encore équivalent au modèle suivant :

$$\begin{cases} \mathrm{Y}_i|c_i,\phi_c & \overset{ind}{\sim} & \mathrm{P}_{m_i,\mathrm{C}}, \\ (c_1,\ldots,c_n) & \sim & \mathrm{CRP}(\alpha_0), \\ \phi_c|\mathrm{G}_0 & \overset{ind}{\sim} & \mathrm{G}_0, \end{cases}$$

où l'on note de même $m_i(t) = \int_0^t \phi_{c_i}(u)\mathrm{Z}_i(u)du$. De la même façon que pour le modèle (DPMFc), nous choisissons $\mathrm{G}_0 = \mathrm{P}_{0,\Sigma_0}$. Nous supposons à nouveau que les paramètres α_0, C et Σ_0 sont fixés dans un premier temps. Nous verrons plus tard comment nous pouvons estimer ces paramètres.

Les différences dans les deux modèles tiennent seulement dans la modélisation des observations. Dans le premier cas, la covariance des observations modélisées agit de façon individuelle. Dans le second cas, nous supposons une même covariance pour toutes les observations, conduisant ainsi à un modèle plus simple. Dans les études numériques, nous verrons comment ces modèles se comportent face à des jeux de données.

4.5.2 Implémentation algorithmique

Dans ce modèle plus simple, l'écriture des densités $p(\mathrm{Y}_i|c,\phi_c)$ ne pose pas de problème. Sous les mêmes conditions que pour le modèle (DPMFc), à savoir que chaque $m_i \in \mathrm{H}(\mathrm{C})$, nous savons

que chaque processus gaussien $\mathrm{P}_{m_i,\mathrm{C}}$ admet une densité par rapport au processus gaussien $\mathrm{P}_{0,\mathrm{C}}$. Selon la fonction de covariance C choisie, les produits scalaires sont désormais approchés ou bien calculés explicitement.

Concernant la loi *a posteriori* de chaque ϕ_c sachant les données qui sont dans la classe c, en faisant usage des résultats énoncés dans la sous-section 4.4.2, nous savons que cette loi est un processus gaussien $\mathrm{GP}(\overline{m},\overline{\mathrm{W}})$, donné par :

$$\begin{array}{rcl} \overline{m}(t) & = & \left(\mathrm{I}_t,\overline{\mathrm{Y}}\right)_{\widetilde{\mathrm{W}}+\mathrm{C}}, \\ \overline{\mathrm{W}}(s,t) & = & \Sigma_0(s,t)-(\mathrm{I}_t,\mathrm{I}_s)_{\widetilde{\mathrm{W}}+\mathrm{C}}, \\ \widetilde{\mathrm{W}}(s,t) & = & \int_0^s\int_0^t \Sigma_0(u,v)\overline{\mathrm{Z}}(u)\overline{\mathrm{Z}}(v)dudv. \end{array}$$

Comme nous le constatons dans ces équations, la complexité est la même que dans le cas du (DPMFc), mais au lieu d'avoir un RKHS lié au noyau $\widetilde{\mathrm{W}}+\overline{\mathrm{W}+\mathrm{C}}$, nous avons à présent un RKHS lié au noyau $\widetilde{\mathrm{W}}+\mathrm{C}$.

4.6 Astuces numériques pour les calculs d'intégrales

Dans le modèle (DPMFc), nous devons calculer les fonctions m_i, W_i et $\widetilde{\mathrm{W}}$ et dans le modèle (DPMFcs), nous devons calculer les fonctions m_i et $\widetilde{\mathrm{W}}$. Dans tous les cas, ces calculs font appel à des intégrales simples et doubles dont l'évaluation numérique augmente avec le nombre de temps d'observation.

Si la grille d'observation est très fine, le temps de calcul de ces intégrales peut devenir très important, voire ingérable. Si au contraire la grille d'observation est grossière, les calculs seront réalisables mais l'approximation du produit scalaire $(f,g)_{\mathrm{K}}$ apportée par les théorèmes 2.5 et 2.6 deviendra inexacte. Dans cette section, nous allons donc évoquer quelques astuces numériques permettant de calculer rapidement ces intégrales.

Supposons que nous disposions des fonctions aux temps $x_1,\ldots,x_{\mathrm{N}}$ et que l'on souhaite également calculer les intégrales aux temps $x_1,\ldots,x_{\mathrm{N}}$, c'est-à-dire aux temps d'observation. Supposons que l'on observe deux fonctions f et g aux temps d'observation $x_1,\ldots,x_{\mathrm{N}}$. Nous souhaitons à présent calculer les intégrales suivantes, pour tous entiers $i,j\in\{1,\ldots,n\}$:

$$\begin{array}{rcl} \mathrm{S}_i(t) & = & \int_{x_1}^{x_i} f(u,t)g(u)du, \\ \mathrm{D}_{ij} & = & \int_{x_1}^{x_i}\int_{x_1}^{x_j} f(u,v)g(u)g(v)dudv. \end{array}$$

Nous savons qu'alors nous pouvons approcher respectivement les intégrales $\mathrm{S}_i(t)$ et D_{ij} par :

$$\left\{\begin{array}{rcl} \widetilde{\mathrm{S}_1}(t) & = & 0, \\ \widetilde{\mathrm{S}_i}(t) & = & \sum_{k=1}^{i-1} f(x_k,t)g(x_k)(x_{k+1}-x_k), i>1, \end{array}\right.$$

et

$$\left\{\begin{array}{rcl} \widetilde{\mathrm{D}_{1j}} & = & 0,\forall j, \\ \widetilde{\mathrm{D}_{i1}} & = & 0,\forall i, \\ \widetilde{\mathrm{D}_{ij}} & = & \sum_{k=1}^{i-1}\sum_{l=1}^{j-1} f(x_k,x_l)g(x_k)g(x_l)(x_{k+1}-x_k)(x_{l+1}-x_l), i,j>1. \end{array}\right.$$

La qualité d'approximation sera d'autant meilleure que la grille des temps d'observation est fine. On observe bien ce phénomène sur les figures 4.1 et 4.2.

Le lemme suivant va nous permettre de faire le lien entre les intégrales simples $\mathrm{S}_i(t)$ et les intégrales doubles D_{ij}, afin d'aboutir à de meilleures performances numériques.

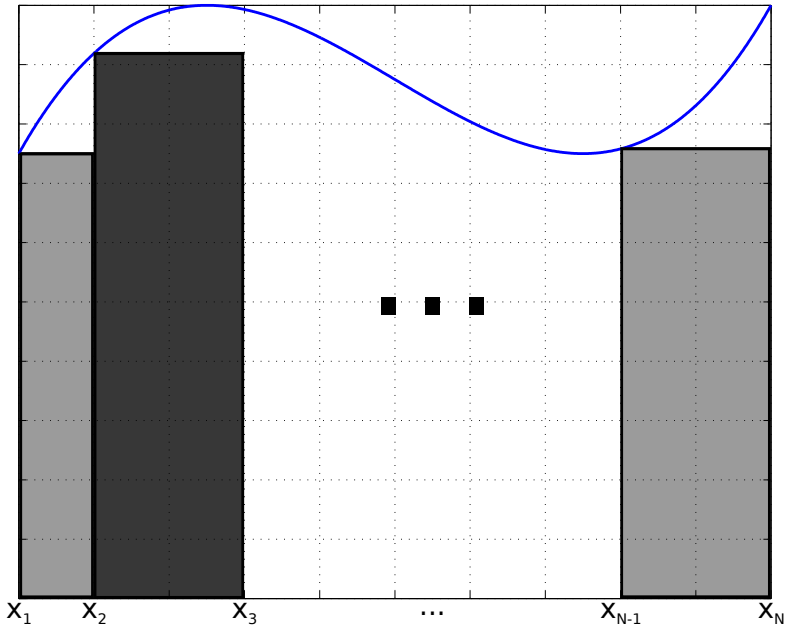

FIGURE 4.1 – *Approximation d'une intégrale simple par la méthode des rectangles. On peut déterminer la valeur approchée d'une intégrale par la somme de toutes les aires des rectangles.*

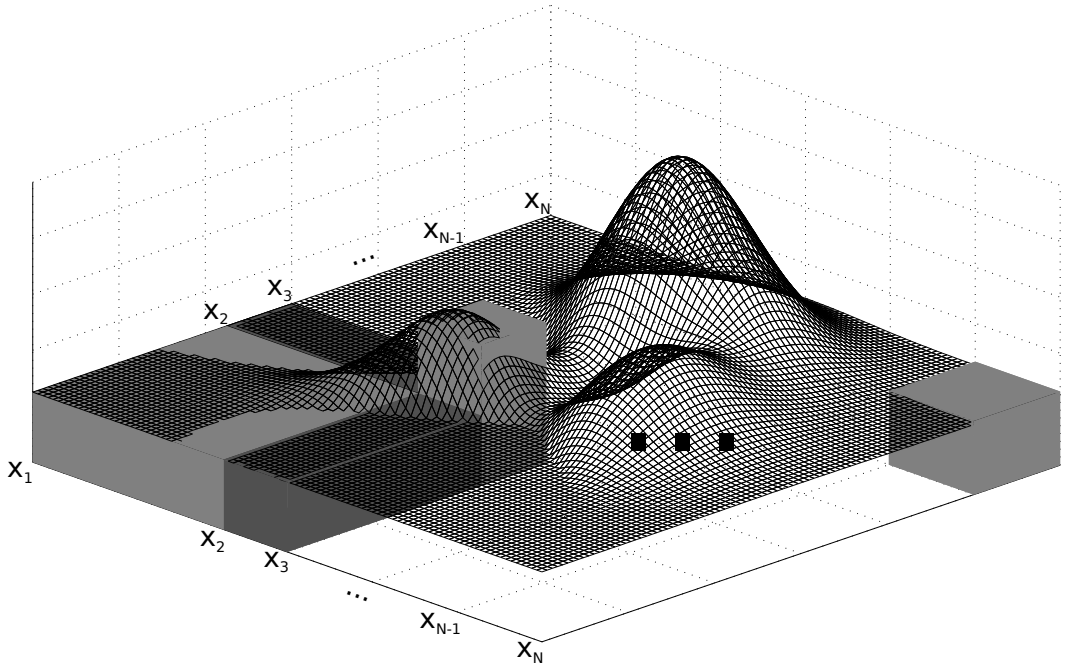

FIGURE 4.2 – *Approximation d'une intégrale double par la méthode des rectangles. On peut déterminer la valeur approchée d'une double intégrale par la somme de tous les volumes des pavés droits.*

Lemme 4.1
Avec les notations précédentes et en notant $\widetilde{\mathrm{S}}(t)$ *le vecteur* $\left(\widetilde{\mathrm{S}_1}(t),\ldots,\widetilde{\mathrm{S}_N}(t)\right)$*, on a le résultat suivant :*

$$\widetilde{\mathrm{S}}(t) = \frac{\widetilde{\mathrm{D}_{\bullet(j+1)}} - \widetilde{\mathrm{D}_{\bullet j}}}{g(x_j)(x_{j+1}-x_j)},$$

où l'entier j *est tel que* $t = x_j$ *et* $\widetilde{\mathrm{D}_{\bullet j}}$ *désigne la* j^e *colonne de la matrice* $\widetilde{\mathrm{D}} = (\widetilde{\mathrm{D}_{ij}})_{1\leq i,j\leq n}$.

PREUVE. Il nous suffit d'écrire pour tout entier $i \in \{1,\dots,n\}$:

$$\begin{aligned}\frac{\widetilde{\mathrm{D}_{i(j+1)}}-\widetilde{\mathrm{D}_{ij}}}{g(x_j)(x_{j+1}-x_j)} &= \frac{1}{g(x_j)(x_{j+1}-x_j)}\times \\ &\quad \Big(\sum_{k=1}^{i-1}\sum_{l=1}^{j} f(x_k,x_l)g(x_k)g(x_l)(x_{k+1}-x_k)(x_{l+1}-x_l) \\ &\quad -\sum_{k=1}^{i-1}\sum_{l=1}^{j-1} f(x_k,x_l)g(x_k)g(x_l)(x_{k+1}-x_k)(x_{l+1}-x_l)\Big), \\ &= \widetilde{\mathrm{S}_i}(x_j).\end{aligned}$$

Ceci conclut le lemme. ■

4.7 Estimation des hyperparamètres du modèle

Supposons dans cette section des temps d'observation $x_1,\dots,x_{\mathrm{N}}$ communs à toutes les observations. Qu'il s'agisse du modèle (DPMFc) ou (DPMFcs), il est nécessaire d'estimer les hyperparamètres des modèles. Dans cette section, nous allons voir comment déterminer les paramètres des différentes fonctions de covariance intervenant dans les équations.

Nous choisissons à nouveau pour fonctions de covariance C et Σ celles de processus d'Ornstein-Uhlenbeck :

$$\mathrm{C}(s,t)=\frac{a^2}{2b}e^{-b|s-t|}, \Sigma(s,t)=\frac{\sigma^2}{2\beta}e^{-\beta|s-t|},$$

où a, b, σ et β sont des réels strictement positifs. Cela nous permet de travailler avec des produits scalaires simples, de la même manière que pour le chapitre précédent. Rappelons à cet effet que certains résultats [6, 94] ont permis de montrer que les espaces H(C) et H(Σ) sont identiques et formés des fonctions différentiables sur $[0,\mathrm{T}]$. Nous proposons également le noyau de covariance Σ_0 suivant :

$$\Sigma_0(s,t)=\frac{\sigma_0^2}{2\beta_0}e^{-\beta_0(s-t)^2}, \tag{4.7.1}$$

lequel garantit des trajectoires de classe C^∞.

Les hyperparamètres β_0 et σ_0 sont fixés de sorte que la variabilité des données générées à partir de Σ_0 soit de l'ordre de celle des données observées. Concernant les autres hyperparamètres, nous proposons de les fixer de manière empirique. En effet, notre expérience montre qu'une modélisation bayésienne complète induirait des problèmes numériques sans améliorer nécessairement les résultats de classification.

4.7.1 Cas du modèle plus simple

Dans le modèle (DPMFcs), sachant que les courbes Y_i sont générées à partir des processus gaussiens $\mathrm{P}_{m_i,\mathrm{C}}$ avec de plus $\mathrm{C}(s,t)=\frac{a^2}{2b}e^{-b|s-t|}$, les paramètres a et b sont fixés à partir de l'estimation empirique de la matrice de variance-covariance intra-classe des courbes discrétisées en quelques points.

Plus précisément, nous appliquons notre algorithme une première fois, avec des paramètres fixés de manière *ad hoc,* et ce afin d'obtenir une première classification. Soient Y_{0i}^k la notation générique d'une observation dans la classe k au temps x_i, Y_{1i}^k la notation générique d'une observation dans la classe k au temps x_{i+1}, $\overline{\mathrm{Y}_0^k}$ la moyenne des observations dans de la classe k au temps x_i et $\overline{\mathrm{Y}_1^k}$ la moyenne des observations dans la classe k au temps x_{i+1}. Nous pouvons

estimer la variance intra-classe ainsi que la covariance intra-classe, par les formules d'approximations respectives

$$\mathrm{V_I} = \frac{\sum_k \sum_i \left(\left(\mathrm{Y}_{0i}^k - \overline{\mathrm{Y}_0^k} \right)^2 + \left(\mathrm{Y}_{1i}^k - \overline{\mathrm{Y}_1^k} \right)^2 \right)}{2(n-\mathrm{K})},$$

et

$$\mathrm{C_I} = \frac{\sum_k \sum_i \left(\mathrm{Y}_{0i}^k - \overline{\mathrm{Y}_0^k} \right) \times \left(\mathrm{Y}_{1i}^k - \overline{\mathrm{Y}_1^k} \right)}{2(n-\mathrm{K})},$$

où K désigne le nombre de classes. Or, d'un point de vue théorique, nous savons que la matrice de variance-covariance du processus $\mathrm{P}_{m_i,\mathrm{C}}$ aux temps x_i et x_{i+1} est égale à :

$$\begin{pmatrix} \mathrm{C}(x_i,x_i) & \mathrm{C}(x_i,x_{i+1}) \\ \mathrm{C}(x_{i+1},x_i) & \mathrm{C}(x_{i+1},x_{i+1}) \end{pmatrix} = \begin{pmatrix} \frac{a^2}{2b} & \frac{a^2}{2b} e^{-b|x_{i+1}-x_i|} \\ \frac{a^2}{2b} e^{-b|x_{i+1}-x_i|} & \frac{a^2}{2b} \end{pmatrix}.$$

En égalisant alors l'expression théorique et son approximation numérique, nous obtenons les estimations suivantes pour les paramètres a et b :

$$\begin{aligned} a &= \frac{1}{x_{i+1}-x_i} \log\left(\frac{\mathrm{V_I}}{\mathrm{C_I}}\right), \\ b &= \sqrt{\frac{2}{x_{i+1}-x_i} \mathrm{V_I} \log\left(\frac{\mathrm{V_I}}{\mathrm{C_I}}\right)}. \end{aligned}$$

Afin d'être encore plus précis, nous pouvons même faire une moyenne sur l'ensemble des temps d'observation x_i, x_{i+1}.

4.7.2 Cas du premier modèle

Dans le modèle (DPMFc), les courbes Y_i sont générées à partir des processus gaussiens $\mathrm{P}_{m_i,\mathrm{W}_i+\mathrm{C}}$ avec de nouveau $\mathrm{C}(s,t) = \frac{a^2}{2b} e^{-b|s-t|}$ et de plus, $\mathrm{W}_i(s,t) = \int_0^s \int_0^t \Sigma(u,v) \mathrm{Z}_i(u) \mathrm{Z}_i(v) du dv$ et $\Sigma(s,t) = \frac{\sigma^2}{2\beta} e^{-\beta|s-t|}$. Les hyperparamètres sont donc à présent a, b, σ et β, fixés à partir de l'estimation empirique de la matrice de variance-covariance intra-classe des courbes discrétisées en quelques points.

Pour ce faire, nous appliquons notre algorithme une première fois avec des paramètres fixés de manière *ad hoc*, et ce afin d'obtenir une première classification. D'un point de vue théorique, nous savons que la matrice de variance-covariance du processus $\mathrm{P}_{m_i,\mathrm{W}_i+\mathrm{C}}$ aux temps d'observation 0 et x_i est égale à :

$$\begin{pmatrix} (\mathrm{W}_i+\mathrm{C})(0,0) & (\mathrm{W}_i+\mathrm{C})(0,x_i) \\ (\mathrm{W}_i+\mathrm{C})(x_i,0) & (\mathrm{W}_i+\mathrm{C})(x_i,x_i) \end{pmatrix} = \begin{pmatrix} \frac{a^2}{2b} & \frac{a^2}{2b} e^{-b|x_i|} \\ \frac{a^2}{2b} e^{-b|x_i|} & \int_0^{x_i} \int_0^{x_i} \Sigma(u,v) \mathrm{Z}_i(u) \mathrm{Z}_i(v) du dv + \frac{a^2}{2b} \end{pmatrix}.$$

Ainsi, par rapport au modèle (DPMFcs), vient se rajouter sur la diagonale un terme lié à la double intégrale W_i. A l'instar de la méthode précédente pour le modèle plus simple (DPMFcs), nous pouvons alors poser :

$$\mathrm{V_{I0}} = \frac{\sum_k \sum_i \left(\left(\mathrm{Y}_{0i}^k - \overline{\mathrm{Y}_0^k} \right)^2 \right)}{(n-\mathrm{K})},$$

et

$$\mathrm{C_I} = \frac{\sum_k \sum_i \left(\mathrm{Y}_{0i}^k - \overline{\mathrm{Y}_0^k}\right) \times \left(\mathrm{Y}_{1i}^k - \overline{\mathrm{Y}_1^k}\right)}{2(n-\mathrm{K})},$$

où K désigne le nombre de classes. Cela nous permet d'estimer a et b de la façon suivante :

$$\begin{aligned} a &= \frac{1}{x_i}\log\left(\frac{\mathrm{V_{I0}}}{\mathrm{C_I}}\right), \\ b &= \sqrt{\frac{2}{x_i}\mathrm{V_{I0}}\log\left(\frac{\mathrm{V_{I0}}}{\mathrm{C_I}}\right)}. \end{aligned}$$

Il suffit alors de minimiser la fonction de paramètres σ^2 et β $\mathrm{W}(\sigma^2, \beta) = ((\mathrm{W}_i + \mathrm{C})(x_i, x_i) - \mathrm{V_{I1}})^2$, avec de plus :

$$\mathrm{V_{I1}} = \frac{\sum_k \sum_i \left(\left(\mathrm{Y}_{1i}^k - \overline{\mathrm{Y}_1^k}\right)^2\right)}{(n-\mathrm{K})}.$$

4.8 Résultats et discussion

L'implémentation a été réalisée à partir du logiciel Matlab. Afin d'analyser les performances de notre modèle, nous avons appliqué notre méthode sur deux jeux de données. Comme il est toujours difficile d'évaluer la performance d'un algorithme de classification, nous avons fait le choix, lorsque c'est possible, de comparer le taux de classification correcte (voir sous-section 1.1.3) abrégé TCC. Enfin, nous donnons également à titre indicatif le temps moyen requis pour une itération, utile dans le cas du traitement de jeux de données en très grande dimension. Tous les résultats présentés ont été obtenus en produisant 10000 itérations, avec un temps de chauffe de 1000 et en retenant 1 itération sur 5. L'implémentation numérique étant déjà conséquente, la classification retenue est celle qui apparaît le plus souvent dans les simulations *a posteriori* (voir discussion en fin de chapitre).

4.8.1 Spécification des modèles

Nous choisissons pour C la fonction de covariance suivante, associée à un processus d'Ornstein-Uhlenbeck :

$$\mathrm{C}(s,t) = \frac{a^2}{2b} e^{-b|s-t|}, \tag{4.8.1}$$

où a et b sont des réels strictement positifs. Lorsque nous utilisons le modèle général (DPMFc), nous choisissons de plus pour Σ :

$$\Sigma(s,t) = \frac{\sigma^2}{2\beta} e^{-\beta|s-t|}, \tag{4.8.2}$$

avec β et σ deux réels strictement positifs. D'après les résultats du chapitre 3, nous connaissons les espaces RKHS associés à ces fonctions de covariance ainsi que les produits scalaires.

Nous choisissons enfin la fonction de covariance Σ_0 suivante :

$$\Sigma_0(s,t) = \frac{\sigma_0^2}{2\beta_0} e^{-\beta_0(s-t)^2}, \tag{4.8.3}$$

laquelle garantit des trajectoires de classe C^∞.

Le nombre de valeurs auxiliaires est fixé à $m = 5$. Les hyperparamètres β_0 et σ_0 sont fixés de sorte que la variabilité des données générées à partir de Σ_0 soit de l'ordre de celle des données observées. Différentes valeurs de β_0 et σ_0 ont été testées, avec pour seule conséquence un temps de convergence de l'algorithme vers la loi stationnaire plus ou moins grand. Les hyperparamètres a et b, ainsi que β et σ le cas échéant, sont estimés empiriquement (voir section 4.7).

Enfin, nous approchons les processus *a posteriori* nécessaires pour l'implémentation par leurs lois normales fini-dimensionnelles, de la même manière que dans le chapitre 3.

4.8.2 Premier jeu de données simulées

Ce premier jeu de données simulées est constitué de 60 courbes observées uniformément sur 100 points de l'ensemble $[0,100]$. Pour construire ce jeu de données, nous avons fixé une fonction covariable Z et deux fonctions θ de la manière suivante :

$$\begin{cases} \mathrm{Z}_1(t) &= 1, \\ \theta_1(t) &= -100t, \\ \theta_2(t) &= 100t, \end{cases}$$

et qui ont permis de créer deux classes. Pour chacun des couples de fonctions $\left(\mathrm{Z}_1, \theta_j\right)_{1\leq j\leq 2}$, nous avons simulé 30 processus gaussiens de fonction moyenne $t \mapsto \int_0^t \theta_j(u)\mathrm{Z}_1(u)du$ et de fonction covariance donnée par celle du processus d'Ornstein-Uhlenbeck, de paramètres $a = 50$ et $b = 5$. Les données sont générées de manière indépendante et la figure 4.3 en présente un agrandissement sur $[0,10]$.

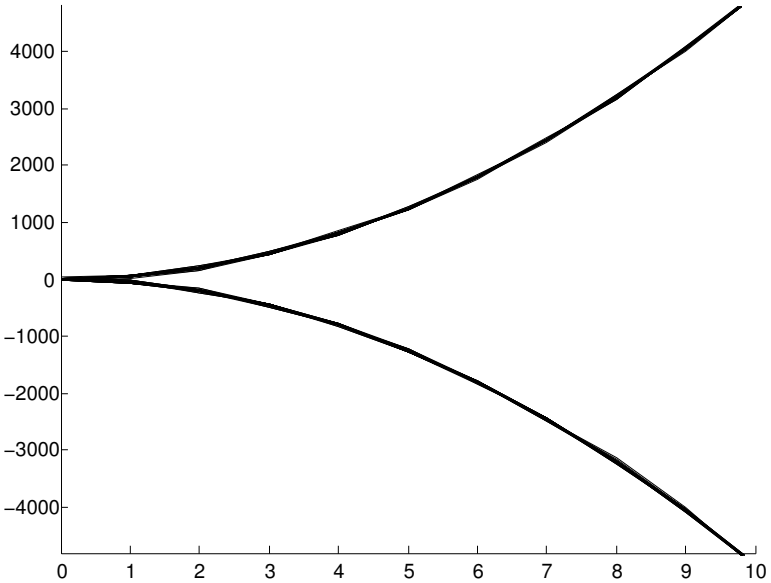

FIGURE 4.3 – *Représentation des courbes simulées, toutes classes confondues.*

Nous initialisons chaque algorithme avec des valeurs $b = \beta_0 = 1(= \beta)$ et $a = \sigma_0 = 10(= \sigma)$, fixées de manière arbitraire. Ceci nous permet d'obtenir une première classification, nous permettant alors d'estimer et fixer $b = \beta_0 = 73(= \beta)$ et $a = \sigma_0 = 2.5(= \sigma)$. La loi *a priori* gamma sur α_0 est une $\mathscr{G}amma(1, 0.5)$.

Dans toutes nos simulations, l'algorithme appliqué aux deux modèles proposées (DPMFc) et (DPMFcs) retrouve toujours la bonne classification. Le temps moyen par itération est de l'ordre

de 0.6s. Ce jeu de données simulées est bien évidemment un exemple jouet, pour lequel nous souhaitions simplement vérifier le comportement de nos algorithmes dans le cas d'une seule fonction covariable commune à toutes les observations. A titre indicatif, nous avons également appliqué l'algorithme du modèle (DPMF) sur ce jeu de données. Comme attendu, les résultats de classification sont identiques. Ainsi, en présence d'une seule covariable, nos modèles (DPMFc) et (DPMFcs) se comportent numériquement de la même façon que le modèle (DPMF).

Enfin, nous avons également souhaité connaître le comportement des modèles (DPMFc) et (DPMFcs) pour un niveau de bruit plus ou moins important. Ainsi nous avons simulé des données suivant plusieurs valeurs de a et b. Que ce soit avec le modèle complet (DPMFc) ou le modèle simple (DPMFcs), l'approche proposée permet de retrouver la vraie classification.

4.8.3 Second jeu de données simulées

Le second jeu de données simulées est constitué de 60 courbes observées uniformément sur 100 points de l'ensemble $[0,100]$. Cette fois-ci, nous avons fixé deux fonctions covariable Z et une seule fonction θ de la manière suivante :

$$\begin{cases} Z_1(t) &= -100, \\ Z_2(t) &= 100, \\ \theta_1(t) &= t, \end{cases}$$

et qui ont permis de créer une seule classe. Pour chacun des couples de fonctions $(Z_i,\theta_1)_{1\leq i\leq 2}$, nous avons simulé 30 processus gaussiens de fonction moyenne $t \mapsto \int_0^t \theta_1(u)Z_i(u)du$ et de fonction covariance donnée par celle du processus d'Ornstein-Uhlenbeck, de paramètres $a = 50$ et $b = 5$. Les données sont générées de manière indépendante et sont similaires aux données de la figure 4.3.

Nous initialisons chaque algorithme avec des valeurs $b = \beta_0 = 1(= \beta)$ et $a = \sigma_0 = 10(= \sigma)$, fixées de manière arbitraire. Ceci nous permet d'obtenir une première classification, nous permettant alors d'estimer et fixer $b = \beta_0 = 73(= \beta)$ et $a = \sigma_0 = 2.5(= \sigma)$. La loi *a priori* gamma sur α_0 est une $\mathcal{G}amma(1,0.5)$.

En répétant chaque algorithme 50 fois, nous obtenons pour les modèles (DPMFc) et (DPMFcs) un TCC moyen de 80% et un temps moyen par itération de l'ordre de 0.6s. Les résultats complets sont présentés dans la figure 4.4 sous forme de diagramme en boîte pour le TCC.

Ce jeu de données, en apparences similaire au premier jeu de données, nous permet de vérifier le comportement de nos algorithmes dans le cas où une seule classe est présente, mais avec deux fonctions covariables. D'après le TCC, 80% des observations sont bien classées (suivant une seule classe) en moyenne. Contrairement au jeu de données précédent où les données étaient à coup sûr classées en 2 classes, la donnée des covariables a ici permet de retrouver la bonne classification, à une dizaine de courbes près en moyenne. Plus précisément, les covariables semblent avoir une influence sur la classification obtenue uniquement sur l'intervalle $[0,1]$, là où les courbes du jeu de données se croisent. Bien que les covariables semblent interférer dans la classification, nos modèles de classification ne classent pas suivant les covariables.

4.8.4 Jeu de données réelles

Afin d'étudier l'apport en eau et ses conséquences sur le stress hydrique dans le cas de vignes enherbées, le modèle Walis [13] permet, connaissant des caractéristiques climatiques telles que

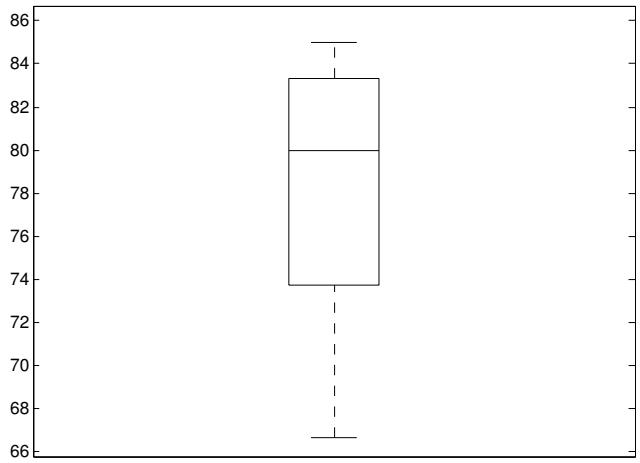

FIGURE 4.4 – *Diagramme en boîte du TCC, obtenu sur 50 répétitions de notre algorithme. La moyenne est de 80%.*

la pluie, la température et l'évapo-transpiration potentielle et des caractéristiques d'une parcelle de vigne enherbée, comme les paramètres de sol et de croissance, de calculer l'évolution temporelle d'un indicateur de stress hydrique appelé FTSW : la fraction d'eau du sol transpirable. Elle indique le pourcentage d'eau disponible pour une vigne. Il s'agit de l'abréviation *Fraction of Transpirable Soil Water* en anglais.

Ce jeu de données est constitué de 504 observations de courbes FTSW, réparties en 7 années climatiques, de 2004 à 2010, au domaine du Chapitre, implanté sur la commune de Villeneuve-Les-Maguelone dans l'Hérault. Chaque année climatique comporte 72 observations de courbes. Entre autres, nous disposons également, pour chaque année climatique, des courbes de pluviométrie journalière. Chaque courbe est observée sur une période de 157 jours et avec un pas de 1 jour. L'intérêt ici consiste à mettre en relation les classes obtenues avec la qualité du vin produit, afin de définir des modes de conduite communs au sein des classes. A titre indicatif, les courbes FTSW ainsi que la courbe de pluviométrie journalière sont représentées figure 4.5 pour l'année climatique 2004.

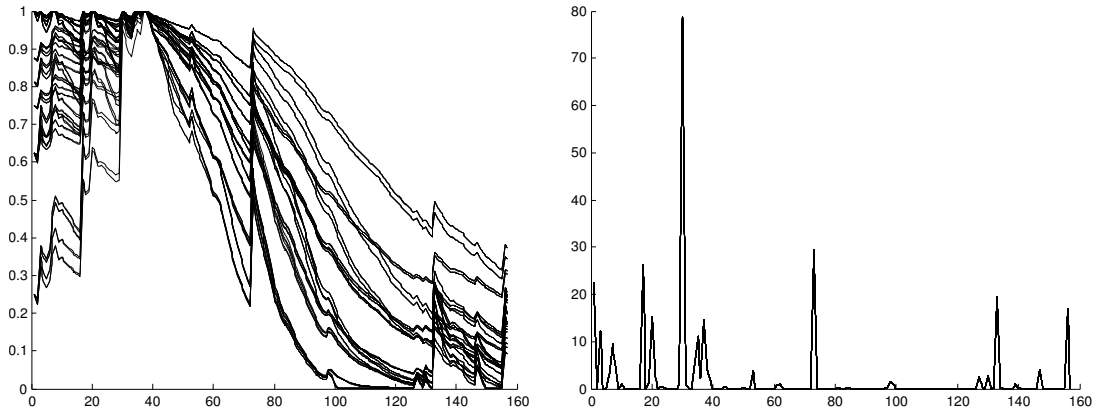

FIGURE 4.5 – *Représentation du jeu de données réelle pour l'année climatique 2004. A gauche : 72 courbes FTSW. A droite : courbe de pluviométrie journalière pour la même période.*

Nous initialisons notre algorithme avec des valeurs $b = \beta_0 = 1 (= \beta)$ et $a = \sigma_0 = 0.5 (= \sigma)$, fixées de manière arbitraire. Ceci nous permet d'obtenir une première classification, nous permettant

alors d'estimer et fixer $b = \beta_0 = 0.7(= \beta)$ et $a = \sigma_0 = 0.27(= \sigma)$. La loi *a priori* gamma sur α_0 est une $\mathcal{G}amma(1, 0.5)$.

Pour ce jeu de données, nous avons calculé le TCC de la classification obtenue sur 50 répétitions des algorithmes, en considérant comme "vraie classification" la classification de 7 classes suivant les années climatiques. Le temps moyen par itération est de l'ordre de 8s. D'après le modèle Walis [13] précédent, les courbes FTSW sont construites et non observées directement. Dans le procédé de construction, il est entre autres posé $\theta_i = Z_i$. Même si ces résultats peuvent sembler incohérents avec ceux des jeux de données simulées, pour lesquels les observations n'étaient jamais classées suivant leur covariable, ils sont en réalité attendus. En moyenne, 78% des courbes se retrouvent classées suivant l'année climatique, c'est-à-dire suivant leur covariable. Sans surprise, il est donc logique de retrouver une classification où 78% des courbes sont classées suivant leur covariable.

4.8.5 Discussion

Les modèles (DPMFc) et (DPMFcs) que nous proposons offrent une approche fonctionnelle de classification de courbes avec covariables, et se sont révélés capables de classer des courbes observées en un très grand nombre de points. Ces modèles considèrent toujours les courbes en dimension infinie et disposent donc des mêmes avantages que ceux du modèle (DPMF). En contrepartie, comme nous utilisons à nouveau les processus de Dirichlet, ils souffrent des mêmes inconvénients (voir discussion 3.6.5).

L'utilisation du *MAP* global pour la classification retenue n'est pas le plus efficace, mais il présente l'avantage de ne pas nécessiter d'outils compliqués pour le calculer. En effet, à l'inverse du *MAP* marginal, ce dernier ne nécessite pas l'utilisation de pivot. Nous sommes conscients de l'inconvénient de l'utilisation du *MAP* global, mais tenons à rappeler que le travail réalisé dans cette thèse se veut davantage être un début d'application de la théorie RKHS en classification, plutôt qu'une mise en œuvre optimisée d'une méthode algorithmique. D'autant plus que l'implémentation numérique est déjà conséquente.

Dans ces deux modèles, la classification opère toujours directement sur les fonctions β_i. Or, il pourrait être envisagé un dernier modèle de classification ajusté par les covariables. A cet effet, nous proposons en annexe A.7 un modèle de classification (DPMFcr). Il s'agit d'un modèle de classification avec covariable beaucoup plus général.

Conclusion

Le travail réalisé dans cette thèse a permis de mettre en évidence un modèle bayésien de classification non supervisée de données fonctionnelles. La prise en compte de covariables fonctionnelles est également assurée au travers d'une hiérarchie supplémentaire dans le modèle. Le chapitre 2 présente tout d'abord les résultats nécessaires pour les implémentations numériques de ces différents modèles. Le théorème principal nous permet de calculer une loi *a posteriori* dans un modèle où loi *a priori* et vraisemblance sont des processus gaussiens. Dans le chapitre 3, nous présentons un modèle original de classification de courbes (DPMF) qui s'avère être efficace pour pouvoir être convenablement appliqué sur des jeux de données en très grande dimension. Une étape de discrétisation est nécessaire pour calculer des produits scalaires entre courbes, mais l'avantage de notre approche réside dans le fait que le pas de discrétisation peut être choisi relativement large. Enfin, dans le chapitre 4, nous proposons plusieurs possibilités pour prendre en compte des covariables fonctionnelles dans le modèle (DPMF).

Il existe de nombreuses directions intéressantes pour le futur de cette thèse. Une première perspective est d'explorer davantage les techniques d'implémentation afin de trouver un algorithme encore plus performant et certainement moins corrélé qu'un échantillonneur de Gibbs. Au vu de l'intérêt croissant que suscite le modèle (DPM), il n'est pas impossible qu'une nouvelle méthode algorithmique soit proposée dans la littérature et permette une meilleure application de nos différents modèles. De manière générale, il serait intéressant de parfaire les différentes approches numériques que nous avons employées, notamment dans le calcul des produits scalaires ou dans le choix d'un estimateur MAP pour la classification à retenir. Nous avons donné plusieurs pistes pour cela dans le manuscrit.

L'algorithme de Gibbs que nous avons utilisé peut présenter une forte auto-corrélation dans ses simulations *a posteriori*, comme c'est le cas pour de nombreux algorithmes MCMC. Cependant, rappelons qu'à cause des conditions de mesure commune imposées dans notre modèle pour l'écriture des densités, cet algorithme était le seul algorithme généralisable. Notons qu'il existe des algorithmes de type recuit-simulé (*simulated annealing* en anglais) qui permettent de retrouver en quelques itérations la classification du *MAP*. Ces algorithmes sont une méthode d'optimisation globale et s'appuient sur l'algorithme de Metropolis-Hastings. Cependant, dans tous ces modèles, les ϕ sont intégrés et il est nécessaire de calculer les quantités de l'équation (3.4.1). Revient alors le problème de mesure commune dont nous ne sommes pas, à l'heure actuelle, arrivés à résoudre. Une orientation intéressante pour le futur ce cette thèse serait de creu-

ser dans cette direction, en généralisant un algorithme de recherche stochastique plutôt qu'un algorithme MCMC.

Il pourrait être également intéressant d'adapter les résultats de ce travail de thèse, notamment en ce qui concerne les densités de processus gaussiens, à la statistique non bayésienne, en proposant par exemple une méthode non bayésienne de classification de données fonctionnelles. Les techniques actuelles ne prenant pas en compte les courbes en dimension infinie et nos résultats permettant d'écrire des densités de processus gaussiens, nous pourrions construire un modèle de classification basé sur la vraisemblance.

Finalement, il serait intéressant de généraliser les modèles étudiés dans cette thèse à plusieurs dimensions. Cela correspond à de vraies attentes, par exemple lorsque les données de spectrométrie sont observées non seulement en fonction de la longueur d'onde, mais également en fonction du temps. Dans des études de vieillissement de feuilles ou de mûrissement de fruits, c'est une approche courante.

Bibliographie

[1] ABRAHAM, C. et CADRE, B. (2008). Concentration of Posterior Distributions with Misspecified Models. *In Pub. Inst. Stat. Univ. Paris LII, fasc. 3*, volume 52, pages 3–14. (Cité pages 36 et 38.)

[2] AKAIKE, H. (1974). A new look at the statistical model identification. *IEEE T. Aut. Control*, 19:716–723. (Cité pages 26 et 64.)

[3] ALDOUS, D. (1985). *Exchangeability and related topics*, pages 1–198. Springer. (Cité page 33.)

[4] ANTONIAK, C. E. (1974). Mixtures of Dirichlet Processes with applications to Bayesian nonparametric problems. *Ann. Stat.*, 2(6):1152–1174. (Cité pages 21, 33 et 47.)

[5] ARONSZAJN, N. (1950). Theory of reproducing kernels. *T. Am. Math. Soc.*, 68(3):337–404. (Cité page 53.)

[6] BERLINET, A. et THOMAS-AGNAN, C. (2004). *Reproducing kernel Hilbert spaces in Probability and Statistics*. Kluwer Academic Publishers. (Cité pages 59, 71, 82, 85 et 89.)

[7] BISGAARD, T. M. et SASVARI, Z. (2000). *Characteristic Functions and Moment Sequences : Positive Definiteness in Probability*. Nova Science Publishers, Inc. (Cité page 119.)

[8] BLACKWELL, D. et MACQUEEN, J. B. (1973). Ferguson Distributions via Pòlya urn schemes. *Ann. Stat.*, 1(2):353–355. (Cité page 32.)

[9] BLEI, D. M. et JORDAN, M. I. (2006). Variational Inference for Dirichlet Process Mixtures. *Bayesian Analysis*, 1(1):121–144. (Cité page 43.)

[10] BOUVEYRON, C. et JACQUES, J. (2011). Model-based clustering of time series in group-specific functional subspaces. *Adv. Data Anal. Classif.*, 5(4):281–300. (Cité pages 21, 64 et 74.)

[11] BROWN, D. P. (2008). Efficient functional clustering of protein sequences using the Dirichlet process. *Bioinformatics*, 24(16):1765–1771. (Cité pages 21 et 65.)

[12] BUSH, C. A. et MACEACHERN, S. N. (1996). A semiparametric Bayesian model for randomised block designs. *Biometrika*, 83(2):275–285. (Cité page 43.)

[13] CELETTE, F., GARY, C. et RIPOCHE, A. (2010). WaLIS-A simple model to simulate water partitioning in a crop association : The example of an intercropped vineyard. *Agr. Water Manage*, 97(11):1749–1759. (Cité pages 93 et 95.)

[14] CELEUX, G. (1998). Bayesian inference for Mixture : the label switching problem. *In Proceedings Compstat*, pages 227–232. (Cité page 45.)

[15] CELEUX, G., GOVAERT, G. et BIERNACKI, C. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE T. Pattern Anal.*, 22(4):719–725. (Cité pages 26 et 64.)

[16] CHERIAN, A., MORELLAS, V., PAPANIKOLOPOULOS, N. et BEDROS, S. J. (2011). Dirichlet Process Mixture Models on Symmetric Positive Definite Matrices for Appearance Clustering in Video Surveillance Applications. *In Proc. Cvpr. IEEE*, numéro 1, pages 3417–3424. (Cité page 27.)

[17] CHIOU, J. M. et LI, P. L. (2007). Functional clustering and identifying substructures of longitudinal data. *J. Roy. Stat. Soc. B*, 69(4):679–699. (Cité page 74.)

[18] CHOPIN, N. (2002). A sequential particle filter for static models. *Biometrika*, 89(3):539–552. (Cité page 43.)

[19] COFFEY, N. et HINDE, J. (2011). Analyzing Time-Course Microarray Data. Using Functional Data Analysis - A Review. *Stat. Appl. Genet. Mol.*, 10(1). (Cité page 50.)

[20] CORMACK, R. M. (1971). A review of classification. *J. Roy. Stat. Soc. A*, 134(3):321–367. (Cité page 24.)

[21] COX, D. D. (1993). An Analysis of Bayesian Inference for Nonparametric Regression. *Ann. Stat.*, 21(2):903–923. (Cité page 58.)

[22] CRAMÉR, H. et LEADBETTER, M. R. (2004). *Stationary and Related Stochastic Processes : Sample Function Properties and Their Applications*. Dover Publications. (Cité page 50.)

[23] CRANDELL, J. L. et DUNSON, D. B. (2011). Posterior simulation across nonparametric models for functional clustering. *Sankhya Ser. B*, (73):42–61. (Cité page 21.)

[24] DAHL, D. B. (2003). An improved merge-split sampler for conjugate Dirichlet process mixture models. Rapport technique, University of Wisconsin - Madison. (Cité page 65.)

[25] DAHL, D. B. (2006). *Model-Based Clustering for Expression Data via a Dirichlet Process Mixture Model*, chapitre 10, pages 201–218. Cambridge University Press. (Cité page 45.)

[26] DE LA CRUZ-MESIA, R. et QUINTANA, F. A. (2007a). A model-based approach to Bayesian classification with applications to predicting pregnancy outcomes from longitudinal β-hCG profiles. *Biostatistics*, 8(2):228–238. (Cité page 45.)

[27] DE LA CRUZ-MESIA, R. et QUINTANA, F. A. (2007b). Semiparametric Bayesian classification with longitudinal markers. *J. Roy. Stat. Soc. C - App.*, 56(2):119–137. (Cité pages 22 et 80.)

[28] DEL MORAL, P., DOUCET, A. et JASRA, A. (2006). Sequential Monte Carlo Samplers. *J. Roy. Stat. Soc. B*, 68(3):411–436. (Cité page 43.)

[29] DEMPSTER, A. P., LAIRD, N. M. et RUBIN, D. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *J. Roy. Stat. Soc. B*, 39(1):1–38. (Cité page 26.)

[30] DONATH, W. E. et HOFFMAN, A. J. (1973). Lower bounds for the partitioning of graphs. *IBM J. Res. Dev.*, 17(5):420–425. (Cité pages 24 et 26.)

[31] DOOB, J. L. (1953). *Stochastic Processes*. Wiley, New York. (Cité page 52.)

[32] DORAZIO, R. M. (2009). On selecting a prior for the precision parameter of Dirichlet process mixture models. *J. Stat. Plan. Infer.*, 139(9):3384–3390. (Cité page 47.)

[33] DRINEAS, P., FRIEZE, A., KANNAN, R., VEMPALA, S. et VINAY, V. (1999). Clustering large graphs via the singular value decomposition. *Mach. Learn.*, 56(1):9–33. (Cité page 25.)

[34] DRISCOLL, M. F. (1975). The signal-noise problem - A solution for the case that signal and noise are gaussian and independent. *J. Appl. Prob.*, 12:183–187. (Cité page 58.)

[35] DUDLEY, R. M. (1989). *Real Analysis and Probability*. Cambridge Studies in Advanced Mathematics. (Cité pages 49, 60 et 61.)

[36] DUNSON, D. B., HERRING, A. H. et SIEGA-RIZ, A. M. (2008). Bayesian Inference on Changes in Response Densities Over Predictor Clusters. *J. Am. Stat. Assoc.*, 103(484):1508–1517. (Cité pages 22 et 80.)

[37] ESCOBAR, M. D. (1994). Estimating Normal Means With a Dirichlet Process Prior. *J. Am. Stat. Assoc.*, 89(425):268–277. (Cité pages 21, 33, 43 et 46.)

[38] ESCOBAR, M. D. et WEST, M. (1995). Bayesian Density Estimation and Inference Using Mixtures. *J. Am. Stat. Assoc.*, 90(430):577–588. (Cité pages 21, 43, 46 et 71.)

[39] FEARNHEAD, P. (2004). Particle filters for mixture models with an unknown number of components. *Stat. Comput.*, 14(1):11–21. (Cité page 43.)

[40] FOWLKES, E. B. et MALLOWS, C. L. (1983). A Method for Comparing Two Hierarchical Clusterings. *J. Am. Stat. Assoc.*, 78(383):553–569. (Cité page 28.)

[41] FRÜHWIRTH-SCHNATTER, S. (2004). Estimating marginal likelihoods for mixture and Markov switching models using bridge sampling techniques. *Economet. J.*, 7(1):143–167. (Cité page 45.)

[42] FRÜHWIRTH-SCHNATTER, S. (2006). *Finite Mixture and Markov Switching Models*. Springer-Verlag, New York. (Cité page 45.)

[43] GELFAND, A. E., KOTTAS, A. et MACEACHERN, S. N. (2005). Bayesian Nonparametric Spatial Modeling With Dirichlet Process Mixing. *J. Am. Stat. Assoc.*, 100(471):1021–1035. (Cité pages 21 et 64.)

[44] GREEN, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732. (Cité page 43.)

[45] GREEN, P. J. et RICHARDSON, S. (2001). Modelling heterogeneity with and without the Dirichlet process. *Scand. J. Stat.*, 28(2):355–375. (Cité page 43.)

[46] GRENANDER, U. (1950). Stochastic processes and statistical inference. *Arkiv für Mat.*, 1(17): 195–277. (Cité page 52.)

[47] HEARD, N. A., HOLMES, C. C. et STEPHENS, D. A. (2006). A Quantitative Study of Gene Regulation Involved in the Immune Response of Anopheline Mosquitoes. *J. Am. Stat. Assoc.*, 101(473):18–29. (Cité pages 21 et 63.)

[48] HUBERT, L. et ARABIE, P. (1985). Comparing partitions. *J. Classif.*, 2(1):193–218. (Cité page 28.)

[49] III, H. D. (2007). Fast search for Dirichlet process mixture models. *In 11th International Conference on Artificial Intelligence and Statistics*, pages 1–8. (Cité page 43.)

[50] ISHWARAN, H. et JAMES, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *J. Am. Stat. Assoc.*, 90(453):161–173. (Cité page 31.)

[51] ISHWARAN, H. et ZAREPOUR, M. (2002). Exact and approximate sum representations for the Dirichlet process. *Can. J. Stat.*, 30(2):269–283. (Cité pages 21 et 31.)

[52] JACCARD, P. (1901). Etude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37:547–579. (Cité page 28.)

[53] JACKSON, E., DAVY, M., DOUCET, A. et FITZGERALD, W. J. (2007). Bayesian Unsupervised Signal Classification by Dirichlet Process Mixtures of Gaussian Processes. *In Int. Conf. Acoust. Spee.*, pages 1077–1080. (Cité pages 21, 65 et 76.)

[54] JACQUES, J. et PREDA, C. (2013). Funclust : a curves clustering method using functional random variable density approximation. *Neurocomputing*, 112:164–171. (Cité pages 21, 64, 74 et 75.)

[55] JAMES, G. M. et SUGAR, C. A. (2003). Clustering for sparsely sampled functional data. *J. Am. Stat. Assoc.*, 98(462):397–408. (Cité pages 21, 63, 64 et 74.)

[56] KAILATH, T. (1969). A General Likelihood-Ratio Formula for Random Signals in Gaussian Noise. *IEEE T. Inform. Theory*, 15(3):350–361. (Cité page 52.)

[57] KAILATH, T., GEESEY, R. T. et WEINERT, H. L. (1972). Some relations among RKHS norms, Fredholm equations, and innovations representations. *IEEE T. Inform. Theory*, 18(3):341–348. (Cité page 53.)

[58] KALLI, M., GRIFFIN, J. E. et WALKER, S. G. (2011). Slice sampling mixture models. *Stat. Comput.*, 21(1):93–105. (Cité page 43.)

[59] KAUFMAN, L. et ROUSSEEUW, P. J. (1990). *Finding Groups in Data : An Introduction to Cluster Analysis*. Wiley, New York. (Cité page 24.)

[60] KIM, S., TADESSE, M. G. et VANNUCCI, M. (2006). Variable selection in clustering via Dirichlet process mixture models. *Biometrika*, 93(4):877–893. (Cité page 45.)

[61] KIMURA, T., TOKUDA, T., NAKADA, Y., NOKAJIMA, T., MATSUMOTO, T. et DOUCET, A. (2013). Expectation-maximization algorithms for inference in Dirichlet processes mixture. *Pattern Anal. Appl.*, 16(1):55–67. (Cité pages 42, 43 et 78.)

[62] KIRKPATRICK, S., GELATT, C. D. et VECCHI, M. P. (1983). Optimization by simulated annealing. *Science*, 220(4598):671–680. (Cité page 43.)

[63] KOTTAS, A. et GELFAND, A. E. (2001). Bayesian semiparametric median regression modeling. *J. Am. Stat. Assoc.*, 96(456):1458—-1468. (Cité page 31.)

[64] LIU, J. S. (1996). Nonparametric hierarchical Bayes via sequential imputations. *Ann. Stat.*, 24(3):911–930. (Cité pages 21, 42 et 43.)

[65] LLOYD, S. P. (1957). Least square quantization in PCM. Rapport technique, Bell Telephone Laboratories. (Cité page 24.)

[66] LO, A. Y. (1984). On a class of bayesian nonparametric estimates : I. Density estimates. *Ann. Stat.*, 12(1):351–357. (Cité page 42.)

[67] LOÈVE, M. (1960). *Probability Theory*. D. Van Nostrand Co., London. (Cité page 120.)

[68] LUAN, Y. et LI, H. (2004). Model-based methods for identifying periodically expressed genes based on time course microarray gene expression data. *Bioinformatics*, 20(3):332–339. (Cité page 21.)

[69] MA, P., CASTILLO-DAVIS, C. I., ZHONG, W. et LIU, J. S. (2006). A data-driven clustering method for time course gene expression data. *Nucleic Acids Research*, 34(4):1261–1269. (Cité page 64.)

[70] MA, P., ZHONG, W., FENG, Y. et LIU, J. S. (2008). Bayesian Functional Data Clustering for Temporal Microarray Data. *International Journal of Plant Genomics*, 2008:8–11. (Cité pages 21 et 63.)

[71] MACEACHERN, S. N. (1994). Estimating Normal Means With a Conjugate Style Dirichlet Process Prior. *Commun. Stat. Simulat.*, (23):727–741. (Cité pages 43, 67 et 68.)

[72] MACEACHERN, S. N., CLYDE, M. et LIU, J. S. (1999). Sequential importance sampling for nonparametric Bayes models : The next generation. *Can. J. Stat.*, 27(2):251–267. (Cité page 43.)

[73] MACEACHERN, S. N. et MÜLLER, P. (1998). Estimating Mixture of Dirichlet Process Models. *J. Comp. Graph. Stat.*, 7(2):223–238. (Cité pages 21, 32, 43 et 69.)

[74] MACQUEEN, J. B. (1967). Some methods for classification and analysis of multivariate observations. *In 5th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297. (Cité page 24.)

[75] MAO, J. et JAIN, A. K. (1996). A self-organizing network for hyper-ellipsoidal clustering (hec). *IEEE T. Neur. Net.*, 7(1):16–29. (Cité page 25.)

[76] MARIN, J.-M. et DRUILHET, P. (2007). Invariant HPD credible sets and MAP estimators. *Bayesian Analysis*, 2(4):681–691. (Cité page 45.)

[77] MEILA, M. (2005a). Comparing Clusterings - An Axiomatic View. *In Proceedings of the 22nd International Conference on Machine Learning*, pages 577–584, Bonn. (Cité page 27.)

[78] MEILA, M. (2005b). Regularized spectral learning. *In Proceedings of the Artificial Intelligence and Statistics Workshop*, pages 1–8. (Cité pages 24 et 26.)

[79] MEILA, M. (2006). The uniqueness of a good optimum for k-means. *In Proceedings of the 23rd International Conference on Machine Learning*, pages 625–632. (Cité page 25.)

[80] MEILA, M. (2007). Comparing clusterings - an information based distance. *J. Multivariate Anal.*, 98(5):873–895. (Cité pages 27 et 28.)

[81] MILLER, J. W. et HARRISON, M. T. (2013a). A simple example of Dirichlet process mixture inconsistency for the number of components. Rapport technique \tt arXiv :1301.2708 [math.ST], Brown University. (Cité pages 42 et 77.)

[82] MILLER, J. W. et HARRISON, M. T. (2013b). Inconsistency of Pitman-Yor process mixtures for the number of components. Rapport technique \tt arXiv :1309.0024 [math.ST], Brown University. (Cité pages 42 et 77.)

[83] MINKA, T. et GHAHRAMANI, Z. (2003). Expectation propagation for infinite mixtures. *NIPS Workshop on Nonparametric Bayesian Methods and Infinite Models*, pages 1–6. (Cité page 43.)

[84] MIRKIN, B. (1996). *Mathematical classification and clustering*. Kluwer Academic Publishers. (Cité page 28.)

[85] MOORE, E. H. (1935). General analysis. *In Memoirs of the American Philosophical Society, Part I*. (Cité page 53.)

[86] MUKHOPADHYAY, S., BHATTACHARYA, S. et DIHIDAR, K. (2011). On Bayesian "central clustering" : Application to landscape classification of Western Ghats. *Ann. Appl. Stat.*, 5(3):1948–1977. (Cité page 27.)

[87] NEAL, R. M. (2000). Markov chain sampling methods for Dirichlet Process mixture models. *J. Comp. Graph. Stat.*, 9(2):249–265. (Cité pages 21, 22, 43 et 69.)

[88] NEWTON, M. A. et ZHANG, Y. (1999). A recursive algorithm for nonparametric analysis with missing data. *Biometrika*, 86(1):15–26. (Cité page 43.)

[89] NG, A., JORDAN, M. I. et WEISS, Y. (2002). On spectral clustering : analysis and an algorithm. *In Advances in Neural Information Processing Systems*, pages 849–856. MIT Press. (Cité page 26.)

[90] NGUYEN, M. H. (2011). *Segment-based SVMs for Time Series Analysis*. Thèse de doctorat, Carnegie Mellon University. (Cité pages 24 et 27.)

[91] O'HAGAN, A. (1978). Curve fitting and optimal design for prediction. *J. Roy. Stat. Soc. B*, 40(1):1–42. (Cité pages 57 et 58.)

[92] OYA, A., NAVARRO-MORENO, J. et RUIZ-MOLINA, J. C. (2009). Numerical Evaluation of Reproducing Kernel Hilbert Space Inner Products. *IEEE T. Signal Proces.*, 57(3):1227–1233. (Cité pages 56 et 124.)

[93] PARZEN, E. (1959). Statistical inference on time series by Hilbert space methods I. Rapport technique, Stanford University, California. (Cité pages 22, 55 et 72.)

[94] PARZEN, E. (1961). Regression Analysis of Continuous Parameter Time Series. *In Proc. Fourth Berkeley Symp. on Math. Statist. and Prob.*, pages 469–489, Stanford University. University of California Press. (Cité pages 54, 56, 71 et 89.)

[95] PARZEN, E. (1962). *Stochastic Processes*. Holden-Day, Inc. (Cité pages 81, 82 et 120.)

[96] PARZEN, E. (1963). Probability Density Functionals and Reproducing Kernel Hilbert Spaces. *In Time Series Analysis*, chapitre 11, pages 155–169. Wiley. (Cité page 53.)

[97] PEARSON, K. (1894). Contributions to the Mathematical Theory of Evolution. *Philos. T. Roy. Soc. A*, 185:71–110. (Cité pages 24 et 25.)

[98] PETRONE, S., GUINDANI, M. et GELFAND, A. E. (2009). Hybrid Dirichlet mixture models for functional data. *J. Roy. Stat. Soc. B*, 71(4):755–782. (Cité page 65.)

[99] PITMAN, J. (2002). Combinatorial stochastic processes. Rapport technique, University of California at Berkeley. (Cité page 30.)

[100] PRICE, R. (1956). Optimum detection of random signals in noise, with application to scatter-multipath communication–I. *IRE T. Inform. Theor.*, 2(4):125–135. (Cité page 51.)

[101] PRICE, R. et GREEN, P. E. (1958). A communication technique for multipath channels. *Proc. IRE*, 46(3):555–570. (Cité page 51.)

[102] RAMSAY, J. O. et SILVERMAN, B. W. (2005). *Functional Data Analysis*. Springer-Verlag. (Cité pages 21, 49 et 83.)

[103] RAND, W. M. (1971). Objective Criteria for the Evaluation of Clustering Methods. *J. Am. Stat. Assoc.*, 66(336):846– 850. (Cité pages 27 et 28.)

[104] RASMUSSEN, C. E. et WILLIAMS, C. K. I. (2006). *Gaussian Processes for Machine Learning*. The MIT Press. (Cité page 50.)

[105] RAY, S. et MALLICK, B. (2006). Functional clustering by Bayesian wavelet methods. *J. Roy. Stat. Soc. B*, 68(2):305–332. (Cité pages 21, 64 et 80.)

[106] ROLIN, J. M. (1993). On the Distribution of Jumps of the Dirichlet Process. Rapport technique, Université Catholique de Louvain - Institut de statistique, Louvain-la-Neuve. (Cité page 30.)

[107] ROUSSEAU, J. et MENGERSEN, K. (2011). Asymptotic behaviour of the posterior distribution in overfitted mixture models. *J. Roy. Stat. Soc. B*, 73(5):689–710. (Cité page 39.)

[108] SCARPA, B. et DUNSON, D. B. (2009). Bayesian Hierarchical Functional Data Analysis Via Contaminated Informative Priors. *Biometrics*, 65(3):772–780. (Cité page 65.)

[109] SCHERVISH, M. J. (1995). *Theory of Statistics*. Springer-Verlag. (Cité page 70.)

[110] SCHWARZ, G. E. (1978). Estimating the dimension of a model. *Ann. Stat.*, 6(2):461–464. (Cité pages 26 et 64.)

[111] SCHWEPPE, F. C. (1965). Evaluation of likelihood functions for Gaussian signals. *IEEE T. Inform. Theory*, 11(1):61–70. (Cité page 51.)

[112] SETHURAMAN, J. (1994). A constructive definition of Dirichlet priors. *Stat. Sinica*, 4:639–650. (Cité page 30.)

[113] SHEPP, L. A. (1966). Radon-Nikodym Derivatives of Gaussian Measures. *Ann. Math. Stat.*, 37(2):321–354. (Cité page 52.)

[114] SHI, J. Q. et CHOI, T. (2011). *Gaussian Processes Regression Analysis for Functional Data*. CRC Press. (Cité page 50.)

[115] SHI, J. Q. et MALIK, J. (2000). Normalized cuts and image segmentation. *IEEE T. Pattern Anal.*, 22(8):888–905. (Cité pages 24 et 26.)

[116] SHI, J. Q. et WANG, B. (2008). Curve prediction and clustering with mixtures of Gaussian process functional regression models. *Stat. Comput.*, 18(3):267–283. (Cité pages 21, 22 et 79.)

[117] SOKAL, R. R. et SNEATH, P. H. A. (1963). *Principles of numerical taxonomy*. Freeman and Co., San Francisco. (Cité page 24.)

[118] STEPHENS, M. (2000). Dealing with label switching in mixture models. *J. Roy. Stat. Soc. B*, 62(4):795–809. (Cité page 45.)

[119] STRATONOVICH, R. L. et SOSULIN, Y. G. (1964). Optimal Detection of a Markov Process in Noise. *Eng. Cybern.*, 6:7–19. (Cité page 51.)

[120] STRATONOVICH, R. L. et SOSULIN, Y. G. (1965). Optimal detection of a diffusion process in white noise. *Radio Eng. Electron. P.*, 10:704–713.

[121] STRATONOVICH, R. L. et SOSULIN, Y. G. (1966). Optimum reception of signals in non-Gaussian noise. *Radio Eng. Electron. P.*, 11:497–507. (Cité page 51.)

[122] STREHL, A. et GHOSH, J. K. (2002). Cluster ensembles – a knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.*, 3:583–617. (Cité page 29.)

[123] STRIEBEL, C. T. (1959). Densities for Stochastic Processes. *Ann. Math. Stat.*, 30(2):559–567. (Cité pages 115 et 118.)

[124] SUN, H. W. et ZHOU, D. X. (2008). Reproducing Kernel Hilbert Spaces Associated with Analytic Translation-Invariant Mercer Kernels. *J. Fourier Anal. Appl.*, 14(1):89–101. (Cité page 82.)

[125] THALAMUTHU, A., MUKHOPADHYAY, I., ZHENG, X. et TSENG, G. C. (2006). Evaluation and comparison of gene clustering methods in microarray analysis. *Bioinformatics*, 22(19):2405–2412. (Cité page 27.)

[126] TIBSHIRANI, R., WALTHER, G. et HASTIE, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *J. Roy. Stat. Soc. B*, 63(2):411–423. (Cité page 25.)

[127] TUDDENHAM, R. D. et SNYDER, M. M. (1954). *Physical Growth of California Boys and Girls from Birth to Eighteen years*, pages 183–364. University of California Press. (Cité page 73.)

[128] VAN DER VAART, A. W. et VAN ZANTEN, J. H. (2008). Reproducing kernel Hilbert spaces of Gaussian priors. *In Institute of Mathematical Statistics Collections Vol. 3*, pages 200–222. (Cité page 58.)

[129] VINH, N. X., EPPS, J. et BAILEY, J. (2010). Information Theoretic Measures for Clusterings Comparison : Variants, Properties, Normalization and Correction for Chance. *J. Mach. Learn. Res.*, 11:2837–2854. (Cité page 29.)

[130] WALKER, S. G. (2006). On rates of convergence for posterior distributions in infinite-dimensional models. *Ann. Stat.*, 35(2):738–746. (Cité page 43.)

[131] WALKER, S. G. (2007). Sampling the Dirichlet Mixture Model with Slices. *Commun. Stat. Simulat.*, 36(1):45–54. (Cité page 43.)

[132] WALKER, S. G., DAMIEN, P., LAUD, P. W. et SMITH, A. F. M. (1999). Bayesian nonparametric inference for random distributions and related functions. *J. Roy. Stat. Soc. B*, 61(3):485–527. (Cité page 43.)

[133] WANG, L. (2011). Fast Bayesian Inference in Dirichlet Process Mixture Models. *J. Comp. Graph. Stat.*, 20(1):196–216. (Cité page 43.)

[134] WEINER, H. J. (1965). The gradient iteration in time series analysis. *Soc. Ind. Appl. Math. J.*, 13(4):1096–1101. (Cité page 56.)

[135] WEINERT, H. L. (1982). *Reproducing Kernel Hilbert Spaces : applications in statistical signal processing*. Weinert eds, New York. (Cité page 53.)

[136] WILLIAMS, D. (1991). *Probability with Martingales*. Cambridge University Press. (Cité page 61.)

[137] YI, G., SHI, J. Q. et CHOI, T. (2011). Penalized Gaussian Process Regression and Classification for High-Dimensional Nonlinear Data. *Biometrics*, 67(4):1285–1294. (Cité pages 21, 22, 79 et 80.)

A

Annexes

A.1 Cadre théorique de la statistique bayésienne

Le principe bayésien consiste à modéliser par une distribution de probabilité la connaissance que l'on a d'un phénomène, et à mettre à jour cette connaissance après observation des résultats d'une expérience.

Définition A.1 (Loi a posteriori)
Soient deux applications mesurables $\mathrm{X} : \Omega \to \mathscr{X}$ *et* $\Theta : \Omega \to \mathrm{T}$.
Nous appelons X *l'observation de loi* P_X, $\mathscr{X}$ *l'espace des observations,* Θ *le paramètre de loi* Π *et enfin* T *l'espace des paramètres.*
La distribution Π *est appelée loi a priori,* P_X *la distribution prédictive a priori ou encore distribution marginale de* X.
Notons enfin P_θ *la loi de* $\mathrm{X}|\Theta = \theta$.

Sous certaines conditions théoriques[1]*, on peut montrer qu'il existe alors la distribution de* Θ *sachant* $\mathrm{X} = x$*, que l'on appelle loi a posteriori.*

Proposition A.1 (Théorème de Bayes)
Avec les notations précédentes, si P_θ *est absolument continue par rapport à une mesure* ν *pour tout* $\theta \in \mathrm{T}$*, alors la loi conditionnelle de* Θ *sachant* $\mathrm{X} = x$*, notée* Π_x*, est absolument continue par rapport à* Π *pour* P_X*-presque tout* x*. Sa densité par rapport à* Π *est alors :*

$$p(\theta|x) = \frac{p(x|\theta)}{\int_\mathrm{T} p(x|\theta)\Pi(d\theta)},$$

pour tout $x \in \mathscr{X}$ *tel que* $\int_\mathrm{T} p(x|\theta)\Pi(d\theta) \notin \{0, \infty\}$.
De plus, en notant $\mathscr{C}$ *l'ensemble sur lequel l'intégrale* $\int_\mathrm{T} p(x|\theta)\Pi(d\theta)$ *est nulle ou infini,* $\mathrm{P}_\mathrm{X}(\mathscr{C}) = 0$ *et ainsi la densité conditionnelle de* Θ *sachant* $\mathrm{X} = x$ *est bien définie par le quotient ci-dessus.*

Le théorème de Bayes est fondamental dans la statistique bayésienne. De plus, on peut montrer que si Π admet une densité par rapport à une mesure μ, que l'on note $p(\theta)$, alors la loi *a posteriori* admet une densité par rapport à μ, donnée par :

$$p(\theta|x) \propto p(\theta)\, p(x|\theta).$$

En statistique bayésienne, il est souvent fait usage de cette formule pour trouver une loi *a posteriori*.

1. L'espace T muni de sa tribu borélienne doit être un espace polonais, c'est-à-dire un espace métrisable complet et séparable.

A.2 La distribution de Dirichlet

La distribution de Dirichlet est définie sur le simplexe S_k de $\mathbb{R}^k$:

$$S_k = \left\{ (p_1, ..., p_k) \in \mathbb{R}^k \ : \ p_j \geq 0 \text{ pour tout } \ j \in \{1, ..., k\}, \text{ avec } \sum_{j=1}^{k} p_j = 1 \right\}.$$

La figure A.1 ci-jointe illustre le simplexe de dimension 2.

Définition A.2 (Distribution de Dirichlet)
On note $\mathcal{D}ir(\alpha)$ la distribution de Dirichlet d'ordre k et de paramètre $\alpha = (\alpha_1, ..., \alpha_k)$, avec chaque $\alpha_j > 0$, caractérisée sur le simplexe S_{k-1} par la densité de probabilité suivante par rapport à la mesure de Lebesgue :

$$p(P) = \frac{\Gamma\left(\sum_{j=1}^{k} \alpha_j\right)}{\prod_{j=1}^{k} \Gamma(\alpha_j)} \left(\prod_{j=1}^{k-1} p_j^{\alpha_j - 1} \right) \left(1 - \sum_{j=1}^{k-1} p_j \right)^{\alpha_k - 1} \mathbb{I}_{S_{k-1}}(P),$$

où l'on note $P = (p_1, ..., p_{k-1})$ et où $\Gamma(\bullet)$ est la fonction définie pour tout $\alpha > 0$ par :

$$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha - 1} e^{-x} dx.$$

La figure A.2 ci-jointe montre quelques exemples de densité d'une distribution de Dirichlet d'ordre 3, pour différentes valeurs du paramètre α. En particulier, on se rend compte qu'il s'agit d'une loi uniforme sur le simplexe lorsque $\alpha = (1, 1, 1)$.

Dans le cas où les α_j sont seulement positifs ou nuls avec $\sum_j \alpha_j > 0$, cette définition peut être simplifiée à l'aide de la caractérisation par les lois gamma :

Proposition A.2
Soient $Z_1, ..., Z_k$ des variables aléatoires indépendantes et de loi $Z_j \sim \mathcal{G}amma(\alpha_j, 1)$, où $\alpha_j \geq 0$ avec $\sum_j \alpha_j > 0$. Si $\alpha_j = 0$, on pose par convention $\mathcal{G}amma(0, 1) \overset{\mathcal{L}}{=} \delta_0$. Alors le vecteur $\left(\frac{Z_1}{\sum_{j=1}^{k} Z_j}, ..., \frac{Z_k}{\sum_{j=1}^{k} Z_j} \right)$ est distribué suivant la loi $\mathcal{D}ir(\alpha_1, ..., \alpha_k)$. On note aussi ce vecteur $\left(\frac{Z_1}{\sum_{j=1}^{k} Z_j}, ..., \frac{Z_{k-1}}{\sum_{j=1}^{k} Z_j} \right)$.
De plus, $\left(\frac{Z_1}{\sum_{j=1}^{k} Z_j}, ..., \frac{Z_k}{\sum_{j=1}^{k} Z_j} \right)$ est un vecteur aléatoire indépendant de $\sum_{j=1}^{k} Z_j$.

Cette proposition est particulièrement importante car elle implique un procédé de simulation de variables aléatoires suivant la loi de Dirichlet ; il suffit de générer des lois $\mathcal{G}amma(\alpha_j, 1)$ et de faire les quotients ci-dessus.

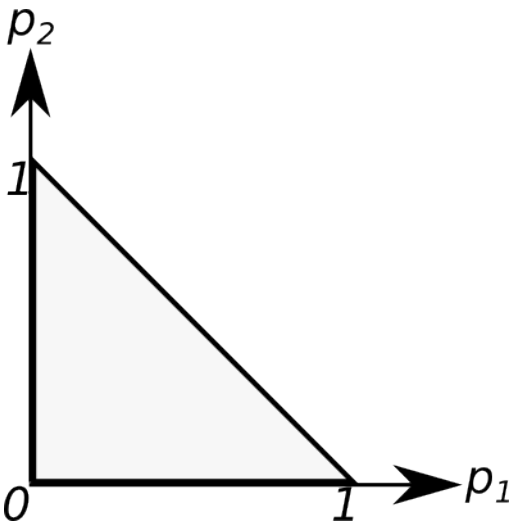

FIGURE A.1 – *Simplexe sur* $\mathbb{R}^2$.

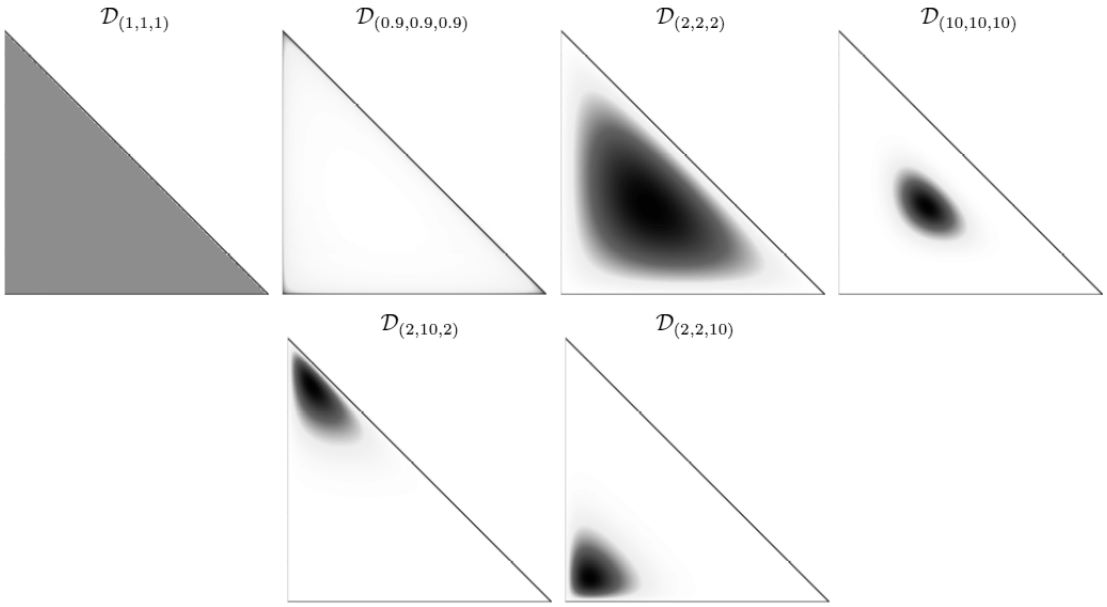

FIGURE A.2 – *Densité de la distribution de Dirichlet pour différentes valeurs du paramètre* α.

A.3 Généralités sur les processus stochastiques

Commençons par rappeler brièvement les définitions de mesures équivalentes et orthogonales.

Définition A.3 (Mesures équivalentes et orthogonales)
Soit $(\mathrm{E}, \mathscr{A})$ *un espace mesurable. Soient* μ *et* ν *deux mesures sur cet espace. Rappelons que* μ *est dite absolument continue par rapport à* ν*, et l'on note* $\mu << \nu$*, si pour tout ensemble* $\mathrm{A} \in \mathscr{A}$ *on a l'implication* $\nu(\mathrm{A}) = 0 \Rightarrow \mu(\mathrm{A}) = 0$*. On dit alors que les mesures* μ *et* ν *sont :*

1. *équivalentes, et on note* $\mu \sim \nu$*, si* $\mu << \nu$ *et* $\nu << \mu$*,*
2. *orthogonales, et on note* $\mu \perp \nu$*, s'il existe* $\mathrm{A} \in \mathscr{A}$ *tel que* $\mu(\mathrm{A}) = 0$ *et* $\nu(\mathrm{A}) = 1$*.*

Considérons à présent un processus stochastique $(\mathrm{X}_t)_{t\in[0,\mathrm{T}]}$ sur l'espace des fonctions continues sur $[0,\mathrm{T}]$, avec $\mathrm{T} \leq \infty$, muni de sa tribu borélienne $\mathscr{B}$.

Définition A.4 (Fonction moyenne et fonction de covariance)
On appelle fonction moyenne du processus la fonction définie sur $[0,\mathrm{T}]$ *par* $m(t) = \mathbb{E}(\mathrm{X}_t)$ *et fonction de covariance la fonction définie sur* $[0,\mathrm{T}] \times [0,\mathrm{T}]$ *par* $\mathrm{K}(s,t) = \mathrm{C}ov(\mathrm{X}_s, \mathrm{X}_t)$*.*

Dans le cas d'un processus stochastique de moyenne nulle, on remarque que la fonction de covariance est simplement donnée par $\mathrm{K}(s,t) = \mathbb{E}(\mathrm{X}_s\mathrm{X}_t)$*.*

Définition A.5 (Mesure de probabilité associée à un processus stochastique)
On appelle mesure de probabilité associée au processus stochastique $(\mathrm{X}_t)_{t\in[0,\mathrm{T}]}$ *la mesure définie pour tout* $\mathrm{B} \in \mathscr{B}$ *par :*

$$\mathrm{P}_\mathrm{X}(\mathrm{B}) = \mathbb{P}((\mathrm{X}_t)_{t\in[0,\mathrm{T}]} \in \mathrm{B}).$$

Définition A.6 (Fonction symétrique et définie positive)
Soient $\mathscr{S} \subset \mathbb{R}$ *un ensemble quelconque et* f *une fonction de deux variables définie sur* $\mathscr{S} \times \mathscr{S}$*. On dit que la fonction* f *est :*

(i) symétrique si pour tous $x, y \in \mathscr{S}$*,* $f(x,y) = f(y,x)$*,*

(ii) définie positive si pour tout entier naturel non nul $n \in \mathbb{N}$*, tous* $t_1, ..., t_n \in \mathscr{S}$ *et tous* $a_1, ..., a_n \in \mathbb{R}$*,* $\sum_{i=1}^n \sum_{j=1}^n a_i a_j f(t_i, t_j) \geq 0$*.*

Un théorème permet de caractériser les fonctions de covariance :

Théorème A.1 (Loève, 1955)
K *est une fonction de covariance si et seulement si* K *est une fonction symétrique et définie positive.*

Un cas particulier très important est le cas des processus gaussiens, dont la définition a été rappelée dans le chapitre 2. Une mesure gaussienne est une mesure de probabilité associée à un processus gaussien.

A.4 Intégrales stochastiques

Un processus stochastique $\mathrm{X} = (\mathrm{X}_t)_{t\in[0,\mathrm{T}]}$ est une famille de variables aléatoires indexée par $[0,\mathrm{T}]$. C'est donc une fonction de deux variables qui sont le temps $t \in [0,\mathrm{T}]$ et l'état de l'univers ω.

Définition A.7 (Intégrale d'un processus)
L'intégrale stochastique du processus X_t *sur l'intervalle* $a < t \leq b$ *est décrite par l'intégrale :*

$$\int_a^b \mathrm{X}_t \, dt.$$

Elle est définie comme la limite en moyenne quadratique des sommes de Riemann $\lim_{n\to+\infty} \sum_{k=1}^{n} \mathrm{X}(t_k)\,(t_k - t_{k-1})$*, sur l'ensemble des subdivisions de l'intervalle* $a < t \leq b$ *aux temps* $a = t_0 < t_1 < \cdots < t_n = b$ *et où* $max_{k=1,\ldots,n}\,(t_k - t_{k-1})$ *tend vers 0.*

Théorème A.2 (Loève, 1960)
$\int_a^b \mathrm{X}_t \, dt$ *existe si, et seulement si, la fonction* $(s,t) \mapsto \mathbb{E}\mathrm{X}_s\mathrm{X}_t$ *est intégrable au sens de Riemann sur l'intervalle* $a < t \leq b$.

A.5 Cas particulier du chapitre 3

Dans cette section, prenons le cas particulier où $\Sigma = \Sigma_0$. Afin d'implémenter l'algorithme défini par l'équation 3.4.1, il est nécessaire de calculer les densités $p(\mathrm{Y}_i)$ et $p(\mathrm{Y}_i|c, \mathrm{Y}^{-i})$, pour tout i et pour tout c, relativement à une même mesure de référence.

Proposition A.3
Dans le modèle (3.2.2), les lois prédictive a priori et prédictive a posteriori correspondent respectivement à :

$$\begin{aligned} \mathrm{Y}_i &\sim \mathrm{P}_{\mu, 2\Sigma}, \\ \mathrm{Y}_i|c, \mathrm{Y}^{-i} &\sim \mathrm{P}_{\frac{\sum_{j\neq i: c_j = c} \mathrm{Y}_j + \mu}{n_c+1}, \frac{n_c+2}{n_c+1}\Sigma}, \end{aligned}$$

où $n_c = \#\{j \neq i : c_j = c\}$.

PREUVE. La loi prédictive *a priori* provient de la généralisation du résultat pour des gaussiennes multivariées. Pour la loi prédictive *a posteriori*, nous pouvons montrer d'après les résultats du chapitre 3 que la loi de $\phi|c, \mathrm{Y}^{-i}$, loi *a posteriori* de ϕ en se basant sur le *prior* G_0 et toutes les observations Y_k pour lesquelles $k \neq i$ et $c_k = c$, est le processus gaussien $\mathrm{P}_{\frac{\sum_{j\neq i: c_j = c} \mathrm{Y}_j + \mu}{n_c+1}, \frac{1}{n_c+1}\Sigma}$.

Il suffit alors d'appliquer à nouveau le résultat pour la loi prédictive *a priori* afin d'obtenir le résultat escompté. ■

Nous observons que les fonctions de covariance des processus ci-dessus sont des multiples l'une de l'autre. Comme les fonctions de covariance sont différentes, nous n'avons pas trouvé dans la littérature de résultat direct pour exprimer ces densités par rapport à une même mesure de référence. L'article de Striebel [123] répond cependant à ce problème. L'auteur propose de calculer des dérivées de Radon-Nikodym entre deux processus gaussiens par le biais des lois finidimensionnelles.

Pour simplifier, définissons les deux fonctions de covariance suivantes :

$$\begin{aligned} \mathrm{C}_0(u, v) &= \frac{\sigma^2}{2\beta} e^{-\beta|u-v|}, \\ \mathrm{C}_1(u, v) &= \frac{k\sigma^2}{2\beta} e^{-\beta|u-v|}, \end{aligned}$$

où σ et β sont deux réels strictement positifs. C_0 et C_1 sont donc deux fonctions de covariance multiples l'une de l'autre et correspondent à des processus d'Ornstein-Uhlenbeck. Supposons que l'on veuille calculer la dérivée de Radon-Nikodym entre les deux processus $\mathrm{P}_{0,\mathrm{C}_0}$ et $\mathrm{P}_{0,\mathrm{C}_1}$. En notant alors C_0^n et C_1^n les matrices de variance-covariance de dimension n, associées aux temps d'observation $t_1, t_2, \ldots, t_n$, et f_0 et f_1 respectivement les densités par rapport à la mesure de Lebesgue des lois normales multivariées $\mathcal{N}(0, \mathrm{C}_0^n)$ et $\mathcal{N}(0, \mathrm{C}_1^n)$, l'objectif d'après Striebel est de trouver un réel $r > 1$ tel que la quantité :

$$\lim_{n\to\infty} \int \left(\frac{f_1(x)}{f_0(x)}\right)^r f_0(x)dx,$$

soit finie.

D'après Striebel [123], il suffit de considérer les temps d'observation $t_i = i\tau$ avec $\tau = \frac{\mathrm{T}}{n}$, qui

forment une suite de points dense dans $[0,T]$. Ainsi, nous pouvons noter plus précisément :

$$C_0^n = \frac{\sigma^2}{2\beta}\begin{pmatrix} 1 & e^{-\beta\tau} & \cdots & e^{-(n-1)\beta\tau} \\ e^{-\beta\tau} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & e^{-\beta\tau} \\ e^{-(n-1)\beta\tau} & \cdots & e^{-\beta\tau} & 1 \end{pmatrix}.$$

L'inverse de cette matrice est alors :

$$(C_0^n)^{-1} = \frac{2\beta}{\sigma^2(1-e^{-2\beta\tau})}\begin{pmatrix} 1 & -e^{-\beta\tau} & 0 & \cdots & 0 \\ -e^{-\beta\tau} & 1+e^{-2\beta\tau} & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & 1+e^{-2\beta\tau} & -e^{-\beta\tau} \\ 0 & \cdots & 0 & -e^{-\beta\tau} & 1 \end{pmatrix}.$$

De la même façon, nous pouvons écrire :

$$C_1^n = \frac{k\sigma^2}{2\beta}\begin{pmatrix} 1 & e^{-\beta\tau} & \cdots & e^{-(n-1)\beta\tau} \\ e^{-\beta\tau} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & e^{-\beta\tau} \\ e^{-(n-1)\beta\tau} & \cdots & e^{-\beta\tau} & 1 \end{pmatrix},$$

et :

$$(C_1^n)^{-1} = \frac{2\beta}{k\sigma^2(1-e^{-2\beta\tau})}\begin{pmatrix} 1 & -e^{-\beta\tau} & 0 & \cdots & 0 \\ -e^{-\beta\tau} & 1+e^{-2\beta\tau} & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & 1+e^{-2\beta\tau} & -e^{-\beta\tau} \\ 0 & \cdots & 0 & -e^{-\beta\tau} & 1 \end{pmatrix}.$$

Ainsi, pour tout réel r nous avons :

$$r(C_1^n)^{-1} - (r-1)(C_0^n)^{-1} = \frac{2(r-(r-1)k)\beta}{k\sigma^2(1-e^{-2\beta\tau})}\begin{pmatrix} 1 & -e^{-\beta\tau} & 0 & \cdots & 0 \\ -e^{-\beta\tau} & 1+e^{-2\beta\tau} & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & 1+e^{-2\beta\tau} & -e^{-\beta\tau} \\ 0 & \cdots & 0 & -e^{-\beta\tau} & 1 \end{pmatrix}.$$

Il reste à calculer alors le déterminant de la matrice suivante :

$$\begin{pmatrix} 1 & -e^{-\beta\tau} & 0 & \cdots & 0 \\ -e^{-\beta\tau} & 1+e^{-2\beta\tau} & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & 1+e^{-2\beta\tau} & -e^{-\beta\tau} \\ 0 & \cdots & 0 & -e^{-\beta\tau} & 1 \end{pmatrix}.$$

En écrivant :

$$\begin{array}{rcl} i & = & n, \\ \alpha & = & 1, \\ \gamma & = & 1, \\ \eta & = & 1+e^{-2\beta\tau}, \\ \lambda & = & -e^{-\beta\tau}, \end{array}$$

il vient le déterminant suivant :

$$\frac{\frac{1}{2}\left(r_1^{n-2}-r_2^{n-2}\right)\left(\frac{1+e^{-2\beta\tau}}{2}-e^{-2\beta\tau}-e^{-2\beta\tau}+\frac{1+e^{-2\beta\tau}}{2}\right)}{\left(\left(\frac{1+e^{-2\beta\tau}}{2}\right)^2-e^{-2\beta\tau}\right)^{1/2}}+\frac{1}{2}\left(r_1^{n-2}+r_2^{n-2}\right)\left(1-e^{-2\beta\tau}\right).$$

En se préoccupant plus tard des expressions respectives de r_1 et r_2, on peut d'ores et déjà écrire ce terme de la façon suivante :

$$\frac{\frac{1}{4}\left(r_1^{n-2}-r_2^{n-2}\right)\left(1-e^{-2\beta\tau}\right)^2}{\frac{1}{2}\left(1-e^{-2\beta\tau}\right)}+\frac{1}{2}\left(r_1^{n-2}+r_2^{n-2}\right)\left(1-e^{-2\beta\tau}\right).$$

En simplifiant alors tous les termes, il reste :

$$r_1^{n-2}\left(1-e^{-2\beta\tau}\right).$$

Comme par le calcul $r_1=1$ le déterminant vaut finalement $1-e^{-2\beta\tau}$. En conclusion, nous avons la formule suivante :

$$|r(\mathrm{C}_1^n)^{-1}-(r-1)(\mathrm{C}_0^n)^{-1}|=\left(\frac{2(r-(r-1)k)\beta}{k\sigma^2(1-e^{-2\beta\tau})}\right)^n\times\left(1-e^{-2\beta\tau}\right).$$

Revenons maintenant à la formule principale à calculer, qui est :

$$\int\left(\frac{f_1(x)}{f_0(x)}\right)^r f_0(x)dx=\frac{|\mathrm{C}_0^n|^{(r-1)/2}}{|\mathrm{C}_1^n|^{r/2}}|r(\mathrm{C}_1^n)^{-1}-(r-1)(\mathrm{C}_0^n)^{-1}|^{-1/2}.$$

Il vient d'une part la formule précédente :

$$|r(\mathrm{C}_1^n)^{-1}-(r-1)(\mathrm{C}_0^n)^{-1}|=\left(\frac{2(r-(r-1)k)\beta}{k\sigma^2(1-e^{-2\beta\tau})}\right)^n\times\left(1-e^{-2\beta\tau}\right),$$

et d'autre part la formule suivante, clairement :

$$\frac{|\mathrm{C}_0^n|^{(r-1)/2}}{|\mathrm{C}_1^n|^{r/2}}=\frac{1}{k^{nr/2}|\mathrm{C}_0^n|^{1/2}}.$$

De plus, les matrices étant inversibles, il vient :

$$\begin{array}{rcl} |\mathrm{C}_0^n| & = & \frac{1}{|(\mathrm{C}_0^n)^{-1}|}, \\ & = & \frac{1}{\left(\frac{2\beta}{\sigma^2(1-e^{-2\beta\tau})}\right)^n(1-e^{-2\beta\tau})}, \\ & = & \left(\frac{\sigma^2}{2\beta}\right)^n(1-e^{-2\beta\tau})^{n-1}. \end{array}$$

Au final, nous avons le résultat qui suit :

$$\begin{aligned} \int \left(\frac{f_1(x)}{f_0(x)}\right)^r f_0(x)dx &= \left(\frac{1}{k^r \frac{\sigma^2}{2\beta} \frac{2(r-(r-1)k)\beta}{k\sigma^2}}\right)^{n/2}, \\ &= \left(\frac{1}{k^{r-1}(r-(r-1)k)}\right)^{n/2}. \end{aligned}$$

D'après l'article de Striebel [123], si l'on trouve un réel $r > 1$ tel que cette limite existe lorsque n tend vers l'infini, alors les deux processus admettent une dérivée de Radon-Nikodym. Or, après maintes recherches numériques, il se trouve que nous n'avons pas pu trouver de réel $r > 1$ qui satisfasse cette condition. Soit $r \leq 1$, ce qui n'est pas la condition demandée. Soit $r > 1$ et le seul réel k valable dans ce cas est $k = 1$, ce qui revient à dire qu'un processus admet une densité par rapport à lui-même. Plus généralement, après avoir essayé numériquement différentes fonctions de covariance, nous ne sommes jamais arrivés à appliquer le résultat présent dans l'article de Striebel.

A.6 Preuve de résultats du chapitre 4

Lemme A.1
Soient $(X_n)_{n\geq 0}$ *une suite de variables aléatoires gaussiennes et* X *une variable aléatoire, telles que :*

$$X_n \xrightarrow{L^2} X,$$

où L^2 *signifie que la convergence est en moyenne quadratique. Alors,* X *est une variable aléatoire gaussienne. Par ailleurs, si la convergence se fait seulement en loi,* X *est toujours une variable aléatoire gaussienne, mais éventuellement dégénérée.*

PREUVE. Fixons un entier $n \in \mathbb{N}$ et écrivons la fonction caractéristique de X_n :

$$\psi_n(t) = \exp\left(i\mu_n t - \frac{1}{2}\sigma_n^2 t^2\right),$$

où μ_n et σ_n^2 désignent respectivement la moyenne et la variance de la variable aléatoire X_n. En particulier, d'après l'hypothèse de convergence en loi, $\lim_{n\longrightarrow+\infty}\psi_n(1)$ existe et nous avons également l'égalité suivante :

$$\left|\psi_n(1)\right| = \exp\left(-\frac{1}{2}\sigma_n^2\right).$$

Ceci prouve que $\lim_{n\longrightarrow+\infty}\sigma_n^2$ existe ; notons σ^2 cette limite. Comme les variables X_n convergent en loi, nous avons de plus $\sigma^2 < \infty$. En conséquence, la partie exponentielle complexe de la fonction caractéristique converge également et nous écrivons :

$$\lim_{n\longrightarrow+\infty}\exp\left(i\mu_n t\right) = \phi(t).$$

Ainsi, nous avons également la convergence suivante :

$$\lim_{n\longrightarrow+\infty}\psi_n(t) = \phi(t)\exp\left(-\frac{1}{2}\sigma^2 t^2\right).$$

Montrons maintenant que $\phi(t)$ peut s'écrire sous la forme $\exp\left(i\mu t\right)$ avec $\mu \in \mathbb{R}$. Pour cela, notons :

$$\phi_n(t) = \exp\left(i\mu_n t\right),$$

de sorte que la suite de fonctions $\left(\phi_n(t)\right)_{n\geq 0}$ admette pour limite la fonction $\phi(t)$. De plus, cette suite de fonctions $\left(\phi_n(t)\right)_{n\geq 0}$ est une suite de fonctions caractéristiques, donc continues en 0, et la fonction limite ϕ est également continue en 0. D'après le théorème de Bochner (voir théorème 1.6.2 de Bisgaard et al. [7]), la fonction ϕ est donc aussi une fonction caractéristique.

Or, l'expression de ϕ_n correspond à la fonction caractéristique d'une variable aléatoire δ_{μ_n}. De plus, la suite de fonctions $\left(\phi_n(t)\right)_{n\geq 0}$ converge. D'après le théorème de convergence de Lévy (voir théorème 1.5.7 de Bisgaard et al. [7]), la suite de variables aléatoires δ_{μ_n} converge en loi vers δ_μ avec $\mu \in \mathbb{R}$. Ainsi nous pouvons écrire $\phi(t) = \exp\left(i\mu t\right)$ donc :

$$\lim_{n\longrightarrow+\infty}\psi_n(t) = \exp\left(i\mu t - \frac{1}{2}\sigma^2 t^2\right),$$

ce qui prouve que la variable aléatoire X est une normale multivariée $\mathcal{N}\left(\mu, \sigma^2\right)$. ■

Le lemme suivant est une généralisation du théorème 3A de Parzen [95]. Dans son livre, Parzen énonce le résultat mais ne donne aucune preuve. Nous donnons ici une proposition de preuve.

Lemme A.2
Soit $\mathrm{X} \sim \mathrm{P}_{m,\mathrm{W}}$ *un processus à trajectoires continues p.s., où* m *et* W *sont deux fonctions continues, respectivement sur* $[0,\mathrm{T}]$ *et* $[0,\mathrm{T}]\times[0,\mathrm{T}]$. *Soit* Z *une fonction déterministe continue sur* $[0,\mathrm{T}]$. *Nous pouvons alors définir le processus* $\mathscr{X}$ *de la manière suivante :*

$$\mathscr{X}(t) = \int_0^t \mathrm{X}(u)\mathrm{Z}(u)du.$$

De plus, $\mathscr{X}$ *est un processus gaussien* $\mathrm{P}_{\widetilde{m},\widetilde{\mathrm{W}}}$, *donné par :*

$$\begin{array}{rcl} \widetilde{m}(t) & = & \int_0^t m(u)\mathrm{Z}(u)du, \\ \widetilde{\mathrm{W}}(s,t) & = & \int_0^s\int_0^t \mathrm{W}(u,v)\mathrm{Z}(u)\mathrm{Z}(v)dudv. \end{array}$$

PREUVE. D'une part, nous pouvons écrire pour chaque réel $t\in[0,\mathrm{T}]$ fixé, $\mathscr{X}(t)$ comme la somme de Riemann $\mathscr{X}(t) = \lim_{n\to+\infty}\frac{t}{n}\sum_{k=1}^n \mathrm{X}\left(\frac{tk}{n}\right)\mathrm{Z}\left(\frac{tk}{n}\right)$. Cette somme peut être interprétée en moyenne quadratique au sens de Parzen [95], et la convergence est ainsi en moyenne quadratique. Pour cela, il suffit d'après le résultat de Loève [67] page 472, de montrer que :

$$\int_0^t\int_0^t \mathbb{E}\mathrm{X}(u)\mathrm{Z}(u)\mathrm{X}(v)\mathrm{Z}(v)dudv < \infty.$$

Or, par définition de la fonction de covariance :

$$\mathbb{E}\mathrm{X}(u)\mathrm{X}(v) = m(u)m(v) + \mathrm{W}(u,v).$$

De plus, comme Z est une fonction non aléatoire par hypothèse :

$$\begin{array}{rcl} \int_0^t\int_0^t \mathbb{E}\mathrm{X}(u)\mathrm{Z}(u)\mathrm{X}(v)\mathrm{Z}(v)dudv & = & \int_0^t\int_0^t \mathrm{Z}(u)\mathrm{Z}(v)\mathbb{E}\mathrm{X}(u)\mathrm{X}(v)dudv, \\ & = & \int_0^t\int_0^t \mathrm{Z}(u)\mathrm{Z}(v)\left(m(u)m(v)+\mathrm{W}(u,v)\right)dudv. \end{array}$$

Les fonctions m et Z étant continues sur le compact $[0,t]$ et la fonction W étant continue sur le compact $[0,t]\times[0,t]$, la fonction $(u,v)\mapsto \mathrm{Z}(u)\mathrm{Z}(v)\left(m(u)m(v)+\mathrm{W}(u,v)\right)$ est nécessairement continue sur le compact $[0,t]\times[0,t]$. Toute fonction continue sur un compact étant bornée, nous en déduisons que la double intégrale ci-dessus est finie. En conséquence, nous avons bien la relation :

$$\int_0^t\int_0^t \mathbb{E}\mathrm{X}(u)\mathrm{Z}(u)\mathrm{X}(v)\mathrm{Z}(v)dudv < \infty.$$

Nous avons ainsi défini la somme de Riemann en moyenne quadratique. D'autre part, nous remarquons que la variable aléatoire $\mathscr{X}(t)$ est gaussienne. En effet, d'après le lemme A.1, il s'agit de la limite en moyenne quadratique de variables aléatoires gaussiennes. De même, pour tout n-uplet $(t_1,...,t_n)\in[0,\mathrm{T}]^n$, le vecteur $(\mathscr{X}(t_1),...,\mathscr{X}(t_n))$ est gaussien et il en résulte que $\mathscr{X}$ est un processus gaussien car un processus est gaussien si, et seulement si, toutes ses lois finidimensionnelles sont gaussiennes.

Ce processus étant entièrement défini par sa fonction moyenne et sa fonction covariance, nous calculons pour tous réels fixés $s,t\in[0,\mathrm{T}]$:

*1) La moyenne de $\mathscr{X}$:

$$\mathbb{E}\mathscr{X}(t) = \mathbb{E}\int_0^t \mathrm{X}(u)\mathrm{Z}(u)\,du.$$

Utilisons le théorème de Fubini afin de montrer que l'on peut intervertir l'espérance et l'intégrale. Pour cela, il suffit de montrer que :

$$\int_0^t \mathbb{E}|\mathrm{X}(u)\mathrm{Z}(u)|\,du < \infty.$$

En appliquant l'inégalité de Hölder avec la fonction Z qui est non aléatoire, écrivons pour tout réel $u \in [0,t]$:

$$\mathbb{E}\Big|\mathrm{X}(u)\mathrm{Z}(u)\Big| \leq \Big|\mathrm{Z}(u)\Big|\sqrt{\mathbb{E}\mathrm{X}(u)^2}.$$

A nouveau, faisons usage de la relation $\mathbb{E}\mathrm{X}(u)^2 = m^2(u) + \mathrm{W}(u,u)$ pour écrire :

$$\mathbb{E}\Big|\mathrm{X}(u)\mathrm{Z}(u)\Big| \leq \Big|\mathrm{Z}(u)\Big|\sqrt{m^2(u) + \mathrm{W}(u,u)}.$$

Les quantités ci-dessus étant positives et l'inégalité étant valable pour tout réel $u \in [0,t]$, nous pouvons écrire également :

$$\int_0^t \mathbb{E}\Big|\mathrm{X}(u)\mathrm{Z}(u)\Big|\,du \leq \int_0^t \Big|\mathrm{Z}(u)\Big|\sqrt{m^2(u) + \mathrm{W}(u,u)}\,du.$$

Dans cette inégalité, les fonctions intervenant dans le membre de droite étant par hypothèse continues sur $[0,t]$, leur intégrale sur ce domaine est finie. D'où :

$$\int_0^t \mathbb{E}\Big|\mathrm{X}(u)\mathrm{Z}(u)\Big|\,du < \infty.$$

En appliquant alors le théorème de Fubini, nous pouvons conclure quant à la fonction moyenne de $\mathscr{X}$:

$$\begin{array}{rcl} \mathbb{E}\mathscr{X}(t) & = & \mathbb{E}\int_0^t \mathrm{X}(u)\mathrm{Z}(u)\,du, \\ & = & \int_0^t \mathbb{E}\mathrm{X}(u)\mathrm{Z}(u)\,du, \\ & = & \int_0^t m(u)\mathrm{Z}(u)\,du, \\ & = & \widetilde{m}(t). \end{array}$$

*2) La covariance de $\mathscr{X}$:

$$\begin{array}{rcl} Cov(\mathscr{X}(s),\mathscr{X}(t)) & = & \mathbb{E}\mathscr{X}(s)\mathscr{X}(t) - \mathbb{E}\mathscr{X}(s)\mathbb{E}\mathscr{X}(t), \\ & = & \mathbb{E}\int_0^s \mathrm{X}(u)\mathrm{Z}(u)\,du \int_0^t \mathrm{X}(v)\mathrm{Z}(v)\,dv - \widetilde{m}(s)\widetilde{m}(t). \end{array}$$

De la même façon que précédemment, on démontre que :

$$\int_0^s\int_0^t \mathbb{E}\Big|\mathrm{X}(u)\mathrm{X}(v)\mathrm{Z}(u)\mathrm{Z}(v)\Big|\,du\,dv \leq$$

$$\int_0^s\int_0^t \Big|\mathrm{Z}(u)\mathrm{Z}(v)\Big|\sqrt{m^2(u) + \mathrm{W}(u,u)}\sqrt{m^2(v) + \mathrm{W}(v,v)} < \infty.$$

Les conditions d'application du théorème de Fubini étant vérifiées, nous pouvons écrire :

$$\begin{array}{rcl} Cov(\mathscr{X}(s),\mathscr{X}(t)) & = & \mathbb{E}\int_0^s\int_0^t \mathrm{X}(u)\mathrm{X}(v)\mathrm{Z}(u)\mathrm{Z}(v)\,du\,dv - \int_0^s m(u)\mathrm{Z}(u)\,du \int_0^t m(v)\mathrm{Z}(v)\,dv, \\ & = & \int_0^s\int_0^t \mathbb{E}\mathrm{X}(u)\mathrm{X}(v)\mathrm{Z}(u)\mathrm{Z}(v)\,du\,dv - \int_0^s m(u)\mathrm{Z}(u)\,du \int_0^t m(v)\mathrm{Z}(v)\,dv. \end{array}$$

Or, par définition, nous avons également :

$$\begin{aligned} Cov(X(u),X(v)) &= \mathbb{E}X(u)X(v) - \mathbb{E}X(u)\mathbb{E}X(v), \\ &= \mathbb{E}X(u)X(v) - m(u)m(v), \\ &= W(u,v), \end{aligned}$$

impliquant l'égalité $\mathbb{E}X(u)X(v) = m(u)m(v) + W(u,v)$. Ainsi nous pouvons conclure quant à la fonction covariance de $\mathscr{X}$:

$$\begin{aligned} Cov(\mathscr{X}(s),\mathscr{X}(t)) &= \int_0^s \int_0^t W(u,v)Z(u)Z(v)dudv \\ &\quad + \int_0^s \int_0^t m(u)m(v)Z(u)Z(v)dudv \\ &\quad - \int_0^s m(u)Z(u)du \int_0^t m(v)Z(v)dv, \\ &= \int_0^s \int_0^t W(u,v)Z(u)Z(v)dudv \\ &\quad + \int_0^s m(u)Z(u)du \int_0^t m(v)Z(v)dv \\ &\quad - \int_0^s m(u)Z(u)du \int_0^t m(v)Z(v)dv, \\ &= \int_0^s \int_0^t W(u,v)Z(u)Z(v)dudv, \\ &= \widetilde{W}(s,t). \end{aligned}$$

Ceci prouve que le processus $\mathscr{X}$ est gaussien et admet respectivement pour fonctions moyenne et covariance $\widetilde{m}$ et $\widetilde{W}$. La famille des lois fini-dimensionnelles étant consistante, nous avons bien défini un processus gaussien. ■

A.7 Modèle ajusté d'une covariable

Qu'il s'agisse des modèles de classification (DPMFc) ou (DPMFcs), la classification opère toujours directement sur les β_i. Or, il pourrait être envisagé un dernier modèle de classification basé sur les résidus. A cet effet, nous proposons ici le modèle de classification "résiduel" (DPMFcr) suivant :

$$(\text{DPMF}cr)\begin{cases} \mathrm{Y}_i(t) & = & \int_0^t \beta_i(u)\mathrm{Z}_i(u)du + \mathrm{U}_i(t) + \epsilon_i(t), \\ \epsilon_i & \overset{ind}{\sim} & \mathrm{P}_{0,\mathrm{C}}, \\ \beta_i & \overset{ind}{\sim} & \mathrm{P}_{\Lambda,\mathrm{K}}, \\ \mathrm{U}_i|\theta_i & \overset{ind}{\sim} & \mathrm{P}_{\theta_i,\Sigma}, \\ \theta_i|\mathrm{G} & \overset{ind}{\sim} & \mathrm{G}, \\ \mathrm{G} & \sim & \mathrm{DP}(\alpha_0, \mathrm{G}_0). \end{cases}$$

Nous supposons de plus que pour tout entier i fixé, les processus β_i, U_i et ϵ_i sont indépendants entre eux. Ceci permet de ne pas introduire de surparamétrisation dans le modèle. D'après l'ensemble des résultats du chapitre 4, (DPMFcr) est alors équivalent au modèle suivant :

$$\begin{cases} \mathrm{Y}_i(t) & = & \mathscr{X}_i(t) + \mathrm{U}_i(t) + \epsilon_i(t), \\ \epsilon_i & \overset{ind}{\sim} & \mathrm{P}_{0,\mathrm{C}}, \\ \mathscr{X}_i & \overset{ind}{\sim} & \mathrm{P}_{m_i,\mathrm{W}_i}, \\ \mathrm{U}_i|\theta_i & \overset{ind}{\sim} & \mathrm{P}_{\theta_i,\Sigma}, \\ \theta_i|\mathrm{G} & \overset{ind}{\sim} & \mathrm{G}, \\ \mathrm{G} & \sim & \mathrm{DP}(\alpha_0, \mathrm{G}_0), \end{cases}$$

où pour tout entier $i \in \{1, \ldots, n\}$, les processus $\mathscr{X}_i$, U_i et ϵ_i sont à nouveau indépendants et où pour tous $s, t \in [0, \mathrm{T}]$, on a :

$$\begin{array}{rcl} m_i(t) & = & \int_0^t \Lambda(u)\mathrm{Z}_i(u)du, \\ \mathrm{W}_i(s,t) & = & \int_0^s \int_0^t \mathrm{K}(u,v)\mathrm{Z}_i(u)\mathrm{Z}_i(v)dudv. \end{array}$$

En conditionnant alors chaque hiérarchie par rapport à θ_i, on obtient le modèle équivalent final suivant :

$$\begin{cases} \mathrm{Y}_i|\theta_i & \overset{ind}{\sim} & \mathrm{P}_{m_i+\theta_i,\mathrm{C}+\mathrm{W}_i+\Sigma}, \\ \theta_i|\mathrm{G} & \overset{ind}{\sim} & \mathrm{G}, \\ \mathrm{G} & \sim & \mathrm{DP}(\alpha_0, \mathrm{G}_0). \end{cases}$$

Avec la paramétrisation en les variables c_i et ϕ_c, ce modèle devient équivalent à :

$$\begin{cases} \mathrm{Y}_i|c_i, \phi_c & \overset{ind}{\sim} & \mathrm{P}_{m_i+\theta_i,\mathrm{C}+\mathrm{W}_i+\Sigma}, \\ c_1, \ldots, c_n & \sim & \mathrm{CRP}(\alpha_0), \\ \phi_c|\mathrm{G}_0 & \overset{ind}{\sim} & \mathrm{G}_0. \end{cases}$$

On pourrait même envisager de modéliser les observations par une forme encore plus générale, c'est-à-dire du type :

$$(\text{DPMF}cg)\begin{cases} \mathrm{Y}_i(t) & = & (\mathrm{K}\beta_i)(t) + \epsilon_i(t), \\ \epsilon_i & \overset{ind}{\sim} & \mathrm{P}_{0,\mathrm{C}}, \\ \beta_i|\theta_i & \overset{ind}{\sim} & \mathrm{P}_{\theta_i,\Sigma}, \\ \theta_i|\mathrm{G} & \overset{ind}{\sim} & \mathrm{G}, \\ \mathrm{G} & \sim & \mathrm{DP}(\alpha_0, \mathrm{G}_0), \end{cases}$$

où K est un opérateur linéaire. Oya [92] donne la relation $(f, \mathbb{K}g) = (f, g)_{\mathrm{K}}$ sans preuve dans son article. Cette relation permet en particulier de faire le lien entre un produit scalaire dans un RKHS et un opérateur de covariance. Cependant, cela ne nous a pas permis d'écrire la loi *a posteriori* de θ_i dans le modèle :

$$\begin{cases} \beta_i|\theta_i & \overset{ind}{\sim} & \mathrm{K}\beta_i + \epsilon_i, \\ \theta_i & \sim & \mathrm{P}_{0,\Sigma_0}. \end{cases}$$