Détermination de classes de modalités de dégradation significatives pour le pronostic et la maintenance

Xuanzhou Wang

Thèse
de doctorat
de l'UTT

Xuanzhou WANG

Détermination de classes de modalités de dégradation significatives pour le pronostic et la maintenance

Spécialité :
Optimisation et Sûreté des Systèmes

2013TROY0022 **Année 2013**

utt
université de technologie
Troyes

THESE

pour l'obtention du grade de

DOCTEUR de l'UNIVERSITE DE TECHNOLOGIE DE TROYES

Spécialité : OPTIMISATION ET SURETE DES SYSTEMES

présentée et soutenue par

Xuanzhou WANG

le 15 novembre 2013

Détermination de classes de modalités de dégradation significatives pour le pronostic et la maintenance

JURY

M. D. BRIE	PROFESSEUR DES UNIVERSITES	Président
M. C. AMBROISE	PROFESSEUR DES UNIVERSITES	Rapporteur
M. P. BEAUSEROY	PROFESSEUR DES UNIVERSITES	Directeur de thèse
M. C. BIERNACKI	PROFESSEUR DES UNIVERSITES	Rapporteur
M. A. GRALL	PROFESSEUR DES UNIVERSITES	Examinateur
Mme É. GRALL	MAITRE DE CONFERENCES	Directrice de thèse

Remerciements

Je tiens dans un premier temps à remercier Edith Grall-Maës et Pierre Beauseroy. En tant que directeurs de thèse, ils m'ont guidé dans mon travail et m'ont aidé à trouver des solutions pour avancer. Edith et Pierre, merci de m'avoir fait confiance, conseillé et pour toutes les heures que vous avez consacrées à diriger ce travail de recherche au cours de ces trois dernières années.

J'adresse mes remerciements à monsieur Christophe Ambroise et monsieur Christophe Biernacki pour avoir accepté d'être rapporteurs de mes travaux et membres du jury de soutenance. Je les remercie aussi pour leurs nombreuses remarques et suggestions qui ont permis d'améliorer la qualité de ce mémoire. De même, je remercie monsieur David Brie et monsieur Antoine Grall qui ont accepté d'évaluer ce travail de thèse en tant que membres du jury.

J'aimerais remercier sincèrement tous mes collègues et amis grâce auxquels j'ai pu travailler dans un cadre très agréable. Je remercie en particulier Elias Khoury et Tuan Huynh de m'avoir aidé et changé les idées quand j'en avais besoin.

Mes derniers remerciements vont vers ma famille, et surtout mes parents, de m'avoir toujours fait confiance et encouragé dans mes choix tout au long de ces années.

Xuan Zhou WANG

Résumé

Les travaux présentés dans ce manuscrit traitent de la détermination de classes de systèmes selon leur mode de vieillissement dans l'objectif de prévenir une défaillance et de prendre une décision de maintenance. L'évolution du niveau de dégradation observée sur un système peut être modélisée par un processus stochastique paramétré. Un modèle usuellement utilisé est le processus Gamma. On s'intéresse au cas où tous les systèmes ne vieillissent pas identiquement et le mode de vieillissement est dépendant du contexte d'utilisation des systèmes ou des propriétés des systèmes, appelé ensemble de covariables. Il s'agit alors de regrouper les systèmes vieillissant de façon analogue en tenant compte de la covariable et d'identifier les paramètres du modèle associé à chacune des classes.

Dans un premier temps la problématique est explicitée avec notamment la définition des contraintes : incréments d'instants d'observation irréguliers, nombre quelconque d'observations par chemin décrivant une évolution, prise en compte de la covariable. Ensuite des méthodes sont proposées. Elles combinent un critère de vraisemblance dans l'espace des incréments de mesure du niveau de dégradation, et un critère de cohérence dans l'espace de la covariable. Une technique de normalisation est introduite afin de contrôler l'importance de chacun de ces critères. Des études expérimentales sont effectuées pour illustrer l'efficacité des méthodes proposées.

Abstract

The work presented in this thesis deals with the problem of determination of classes of systems according to their aging mode in the aim of preventing a failure and making a decision of maintenance. The evolution of the observed deterioration levels of a system can be modeled by a parameterized stochastic process. A commonly used model is the Gamma process. We are interested in the case where all the systems do not age identically and the aging mode depends on the condition of usage of systems or system properties, called the set of covariates. Then, we aims to group the systems that age similarly by taking into account the covariate and to identify the parameters of the model associated with each class.

At first, the problem is presented with the definition of several constraints : time increments of irregular observations, any number of observations per path which describes an evolution, consideration of the covariate. Then the methods are proposed. They combine a likelihood criterion in the space of the increments of deterioration levels, and a coherence criterion in the space of the covariate. A normalization technique is introduced to control the importance of each of these two criteria. Experimental studies are performed to illustrate the effectiveness of the proposed methods.

Table des matières

Principales notations

Nous listons les notations principales utilisées dans ce mémoire. En général, une valeur scalaire est représentée par un caractère maigre minuscule (e.g. $x, y, z, \ldots$), tandis qu'un vecteur est indiqué par un caractère minuscule en gras (e.g. $\mathbf{x}, \mathbf{y}, \mathbf{z}, \ldots$). Remarquons que tous les vecteurs sont considérés comme vecteurs lignes, donc leurs transpositions (e.g. $\mathbf{x}^T, \mathbf{y}^T, \mathbf{z}^T, \ldots$) sont les vecteurs colonnes. De plus, nous dénotons un ensemble par E_{ind} où l'indice *ind* caractérise les éléments dans cet ensemble. Par exemple, $E_{\mathbf{x}}$ représente un ensemble de vecteurs $\mathbf{x}$.

l_i : $i^{ème}$ individu, $i = 1, \ldots, N$

$\mathcal{X}$: *espace de l'attribut*, $\mathcal{X} \subset \mathbb{R}^p$

$\mathcal{Y}$: *espace de la covariable*, $\mathcal{Y} \subset \mathbb{R}^q$

X : *attribut* des individus

Y : *covariable* des individus

$\mathbf{x}_i$: *observation* sous la forme de vecteur du $i^{ème}$ individu dans l'espace de représentation, $i = 1, \ldots, N$

$\mathbf{y}_i$: *valeur de covariable* sous la forme de vecteur du $i^{ème}$ individu dans l'espace de covariable, $i = 1, \ldots, N$

K : nombre total de classes

$\mathcal{P}_K$: une partition des individus en K classes

C_k : $k^{ème}$ classe de la partition $\mathcal{P}_K$, $k = 1, \ldots, K$

z_i : classe d'appartenance du $i^{ème}$ individu, $i = 1, \ldots, N$

$P(.)$: fonction de probabilité

$f(.)$: fonction de densité de probabilité (p.d.f)

Chapitre 1

Introduction

1.1 Contexte et problématique générale

Depuis quelques décennies, les normes de sécurité des systèmes et les exigences en termes de fonctionnalités et de coût sont de plus en plus drastiques. Elles nécessitent des études approfondies de sûreté. L'objectif de ces dernières est de prévenir, éviter ou corriger les dysfonctionnements de systèmes en réduisant les coûts associés. Dans le contexte du pronostic et de la maintenance, une des problématiques importantes est de prendre des décisions sur des modèles de défaillance en utilisant des mesures qui permettent de caractériser l'état de vieillissement du système. On s'intéresse alors à des systèmes ou composants pour lesquels on applique des modèles de défaillance paramétrés et dont les valeurs de paramètres dépendent de caractéristiques du système (par exemple la composition physico-chimique de composants ou les conditions d'utilisation du système).

L'évolution de la mesure de vieillissement au cours du temps forme une trajectoire. Ces trajectoires sont considérées comme des réalisations d'un processus stochastique de dégradation. En pratique, un modèle paramétrable est choisi pour modéliser le processus. Le même processus, défini par un modèle et des valeurs de paramètres, produit des trajectoires différentes. Dans un contexte d'apprentissage, les paramètres des processus peuvent être estimés en utilisant un ensemble de réalisations. Cependant, la loi de vieillissement de plusieurs exemplaires d'un même système peut être modélisée par un seul processus ou peut nécessiter plusieurs processus selon que tous les exemplaires vieillissent identiquement ou pas. On peut, par exemple, imaginer que des moteurs identiques utilisés dans des contextes différents ne subissent pas les même dégradations alors que les moteurs employés dans les conditions analogues vieillissent de la même façon. Dans le cas où tous les systèmes surveillés ne vieillissent pas selon les même modalités, un même modèle avec des paramètres adaptés aux caractéristiques d'utilisation peut être utilisé. Il convient alors d'identifier les paramètres en fonction de ces caractéristiques. En faisant l'hypothèse où il existe plusieurs modalités de fonc-

tionnement, il est alors nécessaire de former des groupes de système ayant les même propriétés de vieillissement pour ensuite estimer les lois de dégradation. Dans la pratique la formation des groupes n'a rien d'un problème simple.

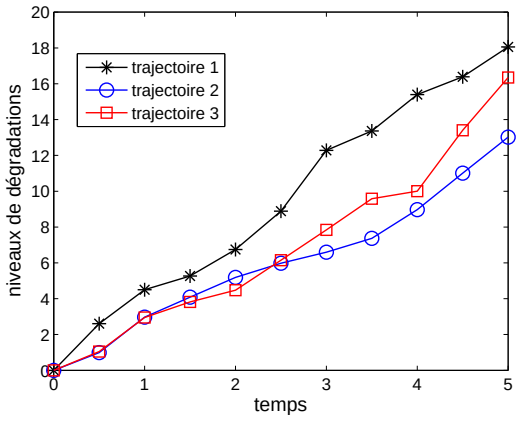
D'un point de vue de la reconnaissance des formes, le regroupement est aussi appelé la classification non-supervisée (ou *clustering* en anglais). Au contraire de la classification supervisée dont l'objectif est d'utiliser les individus déjà classés pour apprendre un modèle qui permet ensuite de classer un nouvel individu, le clustering vise à attribuer à chaque individu un label de classe, de sorte que les individus similaires sont regroupés dans la même classe. Dans notre cas, partant de trajectoires de vieillissement, on souhaite construire des groupes de trajectoires de vieillissement similaires et identifier les lois de vieillissement associées à chacun de ces groupes. La figure 1.1a illustre un exemple de 3 trajectoires issues de la même loi de vieillissement modélisée par le processus Gamma homogène avec les paramètres $m = 4$, $\sigma^2 = 2$. La figure 1.1b montre 2 lois bien distinctes ($m_1 = 4, m_2 = 8, \sigma_1^2 = \sigma_2^2 = 2$) dont chacune contient 3 trajectoires, tandis que la figure 1.1c montre 2 lois distinctes mais proches ($m_1 = 4, m_2 = 5, \sigma_1^2 = \sigma_2^2 = 2$) où le regroupement des trajectoires n'est pas évident.

D'un point de vue interdisciplinaire, la problématique générale de cette thèse porte sur le clustering de trajectoires de dégradation dans des classes caractérisant leur mode d'évolution. Dès que les classes sont formées, la loi de vieillissement de chaque classe peut être déterminée. Ces modèles permettront alors de prendre des décisions sur le pronostic ou la maintenance des systèmes.

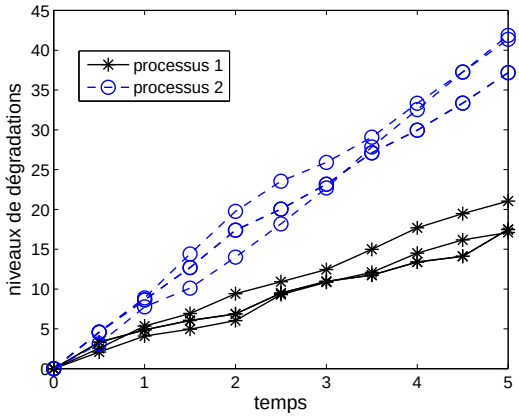
1.2 Contributions de la thèse

En se basant sur la problématique générale, cette thèse consiste à développer des méthodes de clustering qui s'appuient sur des connaissances issues du domaine de la reconnaissance des formes d'une part et du domaine de sûreté de fonctionnement d'autre part. Les contributions principales de la thèse peuvent être résumées comme suit :

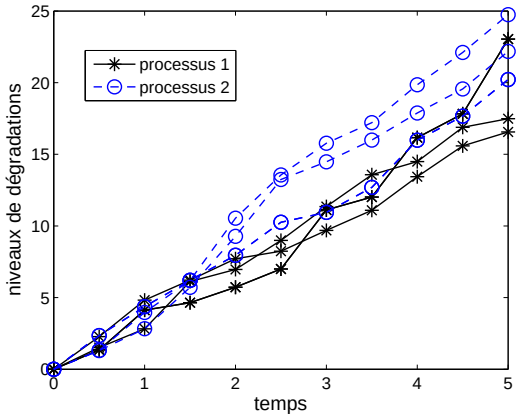
- Étude des spécificités du problème. Dans cette thèse, les trajectoires de vieillissement sont caractérisées par d'une part, les mesures caractérisant l'état de dégradation du système au cours du temps, et d'autre part, les conditions sous lesquelles le système est employé. Généralement, les mesures sont compliquées ou coûteuses. Par exemple, mesurer la longueur d'une fissure sur une digue demande une opération d'envergure, extraire un indicateur d'une cuve de centrale nucléaire ne peut être qu'exceptionnel, car leur nombre est limité et fixé lors



(a) 3 trajectoires provenant d'un même processus



(b) 2 processus bien différents, chacun contient 3 trajectoires



(c) 2 processus peu différents, chacun contient 3 trajectoires

Figure 1.1 – problématique générale

de la construction de la centrale. Ainsi, le nombre de mesures est souvent limité et les instants de mesure sont non réguliers. Par conséquent, les incréments de dégradation d'une mesure à la suivante ne suivent pas le même loi, car ils

dépendent de l'incrément de temps. Dans ce travail, on considère en outre que l'évolution du système dépend d'une covariable. D'un point de vue naturel, les systèmes sous des conditions analogues ont plus de chance de vieillir de la même manière que ceux sous les conditions distinctes. Autrement dit, les trajectoires avec des valeurs de covariable similaires ont plus de chance d'être modélisé par le même processus de dégradation. Ceci permet d'ajouter une connaissance *a priori* au cours du regroupement des trajectoires. Il est nécessaire d'étudier ces spécificités d'un premier temps avant de traiter le problème.

- Proposition de méthodes de clustering pour le problème envisagé. Nous avons traité le problème de façon incrémentale, en partant d'hypothèses restrictives et en relâchant progressivement certaines hypothèses. Dans un premier temps, nous avons abordé les études en supposant qu'il existe une mesure par trajectoire avec une période d'inspection identique. De plus, le nombre de classes est fixé à 2 et la dimension de la covariable est limitée à 1. Une méthode MLC a été proposée pour traiter ce cas particulier. Dans une deuxième étape, nous avons considéré qu'il peut exister plusieurs classes et la dimension de la covariable peut être quelconque. En inspirant par la méthode MLC, deux autres méthodes ont été proposées : la méthode basée sur un critère local et celle basée sur un critère global. Après avoir souligné les avantages de cette dernière méthode, nous l'avons adapté pour traiter le cas général en relâchant toutes les hypothèses.

1.3 Organisation du document

Le chapitre 2 présente l'état de l'art sur les méthodes de clustering et les processus de dégradation. Les méthodes de clustering sont introduites en fonction de deux catégories selon le type de données à regrouper : les méthodes avec un seul type d'informations, appelées aussi l'attribut, et celles exploitant deux types d'informations correspondant à l'attribut et la covariable. La première catégorie regroupe les méthodes de clustering classiques parmi lesquelles nous présentons cinq types de méthodes : les méthodes hiérarchiques, les méthodes de partitionnement, les méthodes à noyau, les méthodes basées sur la densité et les méthodes basées sur les graphes. La deuxième catégorie, généralement appelé clustering spatial dans la littérature, contient les méthodes spatiales qui peuvent être traitées à l'aide d'approches probabilistes ou non probabilistes. Enfin, nous présentons les processus de dégradation et plus particulièrement le processus Gamma.

Le chapitre 3 est consacré à la présentation du problème général et à la proposition

d'une solution dans un cas simple. Nous présentons le problème général en décrivant les caractéristiques générales, la formulation mathématique et l'évaluation de la performance d'une solution. Ensuite, un cas simple est étudié dans le cadre où chaque trajectoire contient une observation avec une période d'inspection constante. Plus précisément, ce cas simple correspond à deux classes uni-modales et à une covariable uni-dimensionnelle. Une méthode nommée MLC est développée qui pour construire toutes les partitions compatibles avec les contraintes et hypothèses du problème, et choisir la partition avec la plus grande vraisemblance.

Dans le cadre d'une seule mesure par trajectoire, d'une période d'inspection constante, d'un nombre de classes et de la dimension de covariable quelconques, deux méthodes sont proposées dans le chapitre 4. La première méthode vise à optimiser pour chaque individu la vraisemblance locale qui tient compte de l'effet de ses voisins, tandis que la deuxième méthode est basée sur un critère global qui combine les informations dans l'espace de représentation et l'espace de covariable. Des études expérimentales sont commentées pour ces deux méthodes en utilisant une base de données simulées. De plus, la méthode basée sur un critère global est comparée avec deux autres méthodes de la littérature en utilisant deux bases de données réelles.

Le chapitre 5 présente une solution dans le cas général où le nombre de classes et la dimension de covariable sont quelconques, la période d'inspection n'est pas constante et le nombre de mesures par trajectoire peut être quelconque. La méthode basée sur le critère global présentée dans le chapitre précédent est développée afin de s'adapter au cas général.

Enfin, le chapitre 6 conclut ce mémoire en évoquant des perspectives de recherche.

Chapitre 2

État de l'art

2.1 Introduction

Ce chapitre commence par une vue générale du clustering en section 2.2. Nous donnons la définition du clustering et introduisons les deux catégories de méthodes de clustering : le clustering avec un type d'information et celui avec deux types d'information. La relation entre les données de dégradation et le clustering est aussi introduite. En section 2.3, nous rappelons la définition de mesures de distance et la notion de système de voisinage. En section 2.4 et 2.5, les méthodes les plus utilisées sont présentées en distinguant les méthodes classiques qui utilisent un type d'information, et les méthodes spatiales qui permettent de traiter deux types d'information. Enfin, en section 2.6, nous présentons ce qu'est le processus de dégradation et comment il est modélisé à l'aide d'un processus stochastique correspondant à la problématique décrite dans le chapitre précédent. Nous concluons en indiquant et en motivant les options choisies pour ce travail.

2.2 Généralités sur le clustering

Étant une des activités premières des êtres humains, le clustering joue un rôle important pour traiter les données rencontrées quotidiennement [4]. C'est une idée naturelle de catégoriser les données en groupe, puisqu'il est ainsi plus efficace et plus rapide de traiter les données par groupe. De plus, il est logique que les données associées à un même groupe soient *similaires* d'après certains critères. Autrement dit, si les données peuvent être représentées par des caractéristiques, les groupes sont formés par des données qui partagent des caractérisations analogues. Contrairement à la classification supervisée dont l'objectif est de prédire l'appartenance aux groupes en disposant de données déjà classées [32, 57], le clustering vise à regrouper des données similaires sans connaître leur appartenance. Pour la suite de la section, nous présentons tout d'abord

la définition du clustering, sa formulation mathématique et les étapes nécessaires pour réaliser une tâche de clustering. Ensuite, nous introduisons brièvement la définition des deux catégories de méthodes de clustering en fonction des types de données dont on dispose pour décrire le problème à traiter. Enfin, nous établissons le lien entre la problématique de ce travail et les catégories de méthodes introduites au préalable.

2.2.1 Définition du clustering

Généralement, le clustering désigne une méthode dont l'objectif est de permettre de déterminer à partir de données, des catégories formées "naturellement" par les individus décrits. Le problème de clustering peut être formalisé comme suit :

Définition 2.1. Soit $E_l = \{l_1, \ldots, l_N\}$ un ensemble de N individus définis dans un espace $\mathcal{X}$. Le K clustering de E_l est un processus de calcul qui permet de définir une partition $\mathcal{P}_K$ de cet ensemble dans K sous-ensembles $C_1, C_2, \ldots C_K$, qui vérifient les trois conditions suivantes :

1. $C_k \neq \emptyset$, $k = 1, \ldots, K$
2. $\cup_{k=1}^{K} C_k = E_l$
3. $C_k \cap C_s = \emptyset$, $k, s = 1, \ldots, K$

Cette définition est souvent utilisée dans la littérature, car elle présente de façon explicite l'idée de clustering [37, 74, 85]. Cependant, cette définition n'est pas universelle et correspond au *clustering dur* où chaque individu appartient à un seul groupe. Alternativement, le *clustering flou* représente le cas où chaque observation est liée avec un degré d'appartenance variable à chaque groupe : $u_{ik} \in [0, 1]$ en vérifiant les deux contraintes suivantes :

1. $\sum_{k=1}^{K} u_{ik} = 1$
2. $0 < \sum_{i=1}^{N} u_{ik} < N$, $\forall k$

qui sont introduites en théorie des ensembles flous [87].

Par ailleurs, nous soulignons que la détermination de la valeur de K qui apparaît dans la définition est un problème majeur du clustering. En général, cette valeur peut être déterminée selon une des trois stratégies suivantes :

- On suppose que le nombre de classes est une connaissance *a priori*. Cette stratégie est utilisée en se basant sur des retours d'experts [36].
- On estime la valeur de K selon un certain critère défini *a priori*. Un ensemble de résultats de partition peut être obtenu avec différents choix de K. Chaque

partition obtenue correspond à une valeur de critère, et donc le critère peut être évalué par rapport au choix de K [75].

- L'estimation de la valeur K fait partie du processus de clustering. On définit un certain critère selon lequel les individus sont regroupés ou dégroupés. Donc, la valeur de K fait partie du résultat obtenu à la fin du processus du clustering [88].

2.2.2 Étapes du clustering

Généralement, une tâche de clustering est décrite selon les étapes suivantes :

- *Sélection de caractéristiques.* Les caractéristiques doivent être sélectionnées proprement afin de représenter le plus d'informations intéressantes possibles relatives au problème. De plus, la redondance parmi les caractéristiques choisies doit être minimale, car en limitant la dimension, on réduit largement la complexité du processus de clustering [38].
- *Choix de la mesure de proximité.* Cette mesure définit la *similarité* ou la *dissimilarité* entre les caractéristiques d'observations sélectionnées. Sachant que le regroupement est effectué sur les données similaires au sens de cette mesure, elle va affecter significativement le résultat du clustering [71].
- *Choix du critère de clustering.* Le critère est défini en fonction du choix de mesure de proximité de l'étape précédente. Il représente souvent la 'pertinence' de rassembler certains individus plutôt que d'autres. Donc, l'optimisation de ce critère permet de grouper les individus similaires, et de dégrouper les individus différents. Les différents critères s'adaptent aux différents problèmes, et il n'existe pas de critère qui peut représenter universellement tous les problèmes [43]. Il est donc important d'analyser le problème envisagé et de choisir le critère le plus pertinent.
- *Choix de la méthode de clustering.* Étant donné un ensemble d'individus à regrouper, la meilleure façon de répondre au problème est de construire toutes les partitions possibles et de sélectionner celle qui correspond à la valeur optimale du critère choisi. Cependant, cette procédure n'est possible qu'avec un petit nombre d'individus N. Si $S(N, K)$ représente le nombre de toutes les partitions possibles de N individus en K groupes, il devient vite très grand avec la croissance de N. En effet, $S(N, K)$ vérifie la relation suivante [50] :

$$S(N,K) = \frac{1}{K!}\sum_{i=0}^{K}(-1)^{K-i}\begin{pmatrix} K \\ i \end{pmatrix} i^N \tag{2.1}$$

Les valeurs de $S(N, K)$ pour certains choix de N et K sont présentées comme

suit :
- $S(10, 3) = 9330$
- $S(15, 3) = 2375101$
- $S(20, 3) = 580606446$

Par conséquent, nous utilisons des méthodes de clustering qui visent à optimiser le critère choisi de l'étape précédente sans faire une recherche exhaustive. Différentes méthodes peuvent être développées pour optimiser le critère de manières différentes (e.g. localement ou globalement).

- *Validation du résultat.* Dès que le résultat du clustering est obtenu, il est nécessaire de l'évaluer pour vérifier sa pertinence. En général, on peut utiliser des critères qui permettent d'évaluer la qualité des clusters selon certaines caractéristiques qui donnent des indications sur la compacité ou la séparabilité des clusters [25, 74, 84, 52].
- *Interprétation du résultat.* Les experts analysent les résultats pour juger de sa pertinence et de sa cohérence par rapport aux connaissances qu'ils ont du problème. Ce retour est important pour résoudre le problème envisagé en pratique.

2.2.3 Deux catégories de méthodes de clustering

La catégorisation des méthodes de clustering dans la littérature est très variée, car la façon de décrire un problème n'est pas unique. Dans ce mémoire, nous classifions les méthodes en deux catégories selon les différents types de données traitées : des données avec un seul type d'information et celles avec deux types d'information.

Clustering avec un type d'information

Dans ce cadre, chaque individu l_i, $(i = 1, \ldots, N)$, N étant le nombre d'individus, est caractérisé par un seul type d'information appelée aussi *attribut* X qui est une variable aléatoire. Les valeurs d'attribut sont appelées *observations* dont l'ensemble est décrit par $E_{\mathbf{x}} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$. Chaque observation peut s'écrire sous la forme d'un vecteur : $\mathbf{x}_i = (x_{i1}, \ldots x_{ip}) \in \mathcal{X} \subset \mathbb{R}^p$ où $\mathcal{X}$ est *l'espace de représentation.* L'appartenance d'un individu est présentée par z_i qui est une valeur de variable aléatoire Z. Donc, ces méthodes de clustering visent à trouver une partition, ou à associer un ensemble de labels $E_z = \{z_i, \ldots, z_N\}$ aux observations. La figure 2.1 illustre un exemple caractérisé par un attribut de dimension 2.

Beaucoup de méthodes classiques s'adaptent à de tels problèmes. Ces méthodes peuvent encore être classifiées en plusieurs groupes selon les problématiques envisagées.

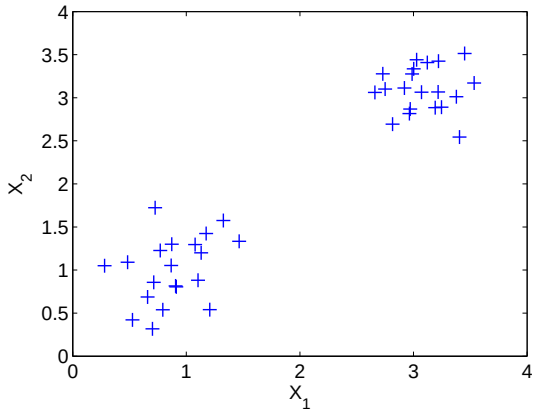


Figure 2.1 – exemple de données avec un attribut de dimension 2

Généralement, cinq groupes sont mentionnés dans la littérature :

- Méthodes hiérarchiques
- Méthodes de partitionnement
- Méthodes à noyau
- Méthodes basées sur la densité
- Méthodes basées sur les graphes

Toutes ces méthodes nécessitent de définir une mesure de proximité dans l'espace de représentation. En pratique, cette mesure est souvent définie par une distance dont les propriétés seront rappelées dans la section 2.3.1. De plus, un *système de voisinage* se basant sur la mesure de distance est aussi nécessaire pour les méthodes basées sur la densité et les méthodes basées sur les graphes. Pour ces deux types de méthodes, l'appartenance de chaque individu est influencée par ses *voisins*. La définition du système de voisinage est présentée dans la section 2.3.2.

Clustering avec deux types d'information

Au lieu de ne tenir compte que de l'attribut, les méthodes qui entrent dans ce cadre traitent deux types d'information : l'*attribut* X qui caractérise l'information discriminante, et la *covariable* Y qui apporte l'information supplémentaire. Chaque valeur de covariable associée à une observation est présentée par un vecteur : $\mathbf{y}_i = (y_{i1}, \ldots, y_{iq}) \in \mathcal{Y} \subset \mathbb{R}^q$. Du point de vue du problème de clustering, X et Y n'ont pas le même rôle, car on suppose que la distribution de X est conditionnée par la classe, et la classe dépend de la valeur de la covariable. On a notamment $f(E_{\mathbf{x}} \mid E_{\mathbf{y}}, E_z) = f(E_{\mathbf{x}} \mid E_z)$ et il y a un lien déterministe inconnu entre Y et Z. La difficulté pour ce genre de problème est d'intégrer proprement les deux types d'information pour que le résultat obtenu respecte la contrainte de proximité du point de vue de l'attribut et de

la covariable. Un exemple de cette problématique est illustré par la figure 2.2.

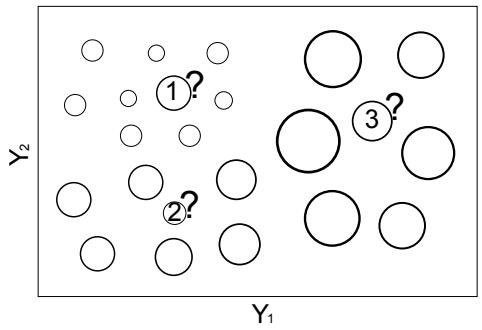


Figure 2.2 – exemple avec un attribut de dimension 1 et une covariable de dimension 2

Les valeurs de l'attribut (aussi appelées les observations dans ce mémoire) des données spatiales dans la figure 2.2 sont représentées par la taille des cercles, tandis que les valeurs de la covariable sont données par les localisations dans ce plan de dimension 2. Évidemment, les individus peuvent être classés en 3 groupes selon la valeur d'attribut petite, moyenne et grande. Si le clustering est basé seulement sur l'attribut, les individus 1 et 3 seront classés dans le groupe 'moyen' et l'individu 2 sera dans le groupe 'petit'. Toutefois, comme les 3 individus sont respectivement entourées dans l'espace de la covariable par des cercles petits, moyens et grands, il est plus pertinent de les classer dans les groupes 'petit', 'moyen' et 'grand' respectivement.

En pratique, les valeurs de covariable les plus utilisées sont les coordonnées spatiales, et ce type de problème est souvent appelé *clustering spatial*. Par exemple, dans le cadre du traitement d'image, le niveau de gris de chaque pixel définit l'attribut, tandis que la position de chaque pixel correspond à la valeur de covariable associée.

Afin de résoudre ce problème, nous pouvons chercher à optimiser un critère $c(E_z \mid E_{\mathbf{x}}, E_{\mathbf{y}})$ en fonction des labels E_z pour déterminer le lien entre la covariable et les labels . Dans un cas particulier où la covariable peut être interprétée de la même manière que l'attribut, le problème peut être résolu comme un problème de clustering de la première catégorie en assimilant les composants de la covariable à des composants supplémentaires de l'attribut. Dans ce cas, les méthodes de clustering de la première catégorie sont directement applicables, mais la contrainte d'implication entre la covariable Y le label Z est perdue.

2.2.4 Les données de dégradation et le clustering

D'après la présentation de la problématique dans le chapitre précédent, les données de surveillance dans ce mémoire sont représentées par le niveau de dégradation. Ce

dernier correspond à l'attribut dans le cadre du clustering. Par exemple, la propagation de fissures est un cas typique pour ce genre de problème. En général, nous mesurons la longueur de fissure pour suivre la propagation du phénomène [48]. Le système est en défaillance si la longueur de la fissure dépasse un certain seuil. Dans ce cas, on considère que la longueur mesurée caractérise la dégradation et elle est la grandeur d'intérêt du problème. Donc, la longueur est interprétée comme l'attribut X.

Dans de nombreux problèmes, la dégradation est souvent liée à certaines conditions. Prenons aussi la propagation de fissures comme un exemple, cette propagation dépend souvent de la température, des contraintes mécaniques, du matériau [54, 83]. Si deux systèmes se dégradent sous des conditions similaires, on peut s'attendre à ce que les propagations soient aussi similaires. Donc, ces conditions peuvent être interprétées comme la covariable Y.

Supposons qu'on dispose de mesures des propagations de fissures pour plusieurs systèmes qui sont exploitées dans des conditions différentes. L'objectif est de regrouper tous les systèmes qui se dégradent de façon similaire afin d'appliquer la même stratégie de maintenance. Du point de vue du clustering, c'est un problème de clustering avec deux types d'information. Il est nécessaire de concevoir une méthode pour regrouper les systèmes qui ont des comportements similaires concernant la propagation des fissures en sachant *a priori* que ceux avec des conditions similaires ont plus de chance d'appartenir au même groupe. Il s'agit donc bien de faire du clustering sur l'attribut (longueur de la fissure) mais en prenant en compte de la covariable (conditions environnementales).

2.3 Mesure de distance et système de voisinage

Selon les étapes de clustering décrites dans la section précédente, le choix de la mesure de proximité est un choix important. Nous commençons dans cette section par une introduction de la notion de distance qui représente très souvent la mesure de proximité. Ensuite, nous présentons la définition d'un système de voisinage. Ce dernier formalise la dépendance entre un individu et ses voisins. Par commodité, la mesure de distance et le système de voisinage sont présentés avec un type d'information où les individus sont caractérisés seulement par l'attribut.

2.3.1 Mesure de distance

Définition 2.2. Soit $E_{\mathbf{x}} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ un ensemble d'observations dont chacune est caractérisée par un vecteur $\mathbf{x}_i = (x_{i1}, \ldots x_{ip}) \in \mathcal{X} \subset \mathbb{R}^p$ avec $\mathcal{X}$ l'espace de

représentation. Une mesure de distance d entre deux observations est une fonction qui vérifie les propriétés suivantes :

1. Symétrie $d(\mathbf{x}_i, \mathbf{x}_j) = d(\mathbf{x}_j, \mathbf{x}_i), \quad \forall \mathbf{x}_i, \mathbf{x}_j \in \mathcal{X}$
2. Réflexivité $d(\mathbf{x}_i, \mathbf{x}_i) = 0, \quad \forall \mathbf{x}_i \in \mathcal{X}$
3. Positivité $d(\mathbf{x}_i, \mathbf{x}_j) \geqslant 0, \quad \forall \mathbf{x}_i, \mathbf{x}_j \in \mathcal{X}$
4. Inégalité triangulaire $d(\mathbf{x}_i, \mathbf{x}_j) \leqslant d(\mathbf{x}_i, \mathbf{x}_h) + d(\mathbf{x}_h, \mathbf{x}_j), \quad \forall \mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_h \in \mathcal{X}$

Plusieurs mesures de distances sont proposées par rapport aux différents problèmes envisagés. Nous listons ci-après quelques mesures fréquemment utilisées dans la littérature.

- **La distance euclidienne** :

$$d_{Eu}(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T} = \left(\sum_{r=1}^{p} \parallel x_{ir} - x_{jr} \parallel^2 \right)^{1/2} \tag{2.2}$$

  La distance euclidienne est la mesure la plus populaire utilisée pour les méthodes du clustering classique [37].

- **La distance de Manhattan** :

$$d_{Man}(\mathbf{x}_i, \mathbf{x}_j) = \sum_{r=1}^{p} \parallel x_{ir} - x_{jr} \parallel \tag{2.3}$$

  Cette distance est souvent appelée 'la distance du taxi', puisqu'elle mesure la distance entre deux points dans la ville parcourue par un taxi qui fait les déplacements horizontaux et verticaux. Cette mesure est utilisée pour un apprentissage flou dans [10]. La distance euclidienne mène à la définition de courbes iso-distances hyper-sphériques, alors que la distance de Manhattan conduit à des courbes iso-distances hyper-rectangulaires [85].

- **La distance de Tchebychev** :

$$d_{Tch}(\mathbf{x}_i, \mathbf{x}_j) = max_r \parallel x_{ir} - x_{jr} \parallel \tag{2.4}$$

  Cette mesure adopte la distance maximale parmi toutes les distances univariées.

Ces trois distances sont des cas particuliers de la distance de Minkowski qui est définie comme suit :

$$d_{Min}(l_i, l_j) = \left(\sum_{r=1}^{p} \parallel x_{ir} - x_{jr} \parallel^n \right)^{1/n} \tag{2.5}$$

Elles correspondent respectivement à l'équation 2.5 avec $n = 2$, $n = 1$ et $n = \infty$. De plus, la distance de Manhattan peut être vue comme une surestimation de la distance euclidienne, alors que la distance de Tchebychev représente une sous estimation sachant la relation suivante :

$$d_{Tch}(l_i, l_j) \leqslant d_{Eu}(l_i, l_j) \leqslant d_{Man}(l_i, l_j) \tag{2.6}$$

2.3.2 Système de voisinage

Un système de voisinage représente un modèle de dépendance entre des observations proches. La proximité est définie au sens de la mesure de distance. Nous donnons tout d'abord la définition d'un système de voisinage. Ensuite, nous présentons deux types de systèmes de voisinage basés respectivement sur une fenêtre et sur un graphe.

Définition d'un système de voisinage

Définition 2.3. Un système de voisinage dans l'espace de représentation $\mathcal{X}$ représente l'ensemble des voisinages $\mathbf{V}$:

$$\mathbf{V} = \{V_i \mid \forall i = 1, \ldots, N\} \tag{2.7}$$

où V_i le voisinage de x_i vérifie les deux conditions suivantes :

1. Une observation n'est pas voisine d'elle-même : $\mathbf{x}_i \notin V_i, \quad i = 1, \ldots, N$
2. La relation de voisinage est symétrique : $\mathbf{x}_j \in V_i \Leftrightarrow \mathbf{x}_i \in V_j, \quad i, j = 1, \ldots, N$

D'après cette définition, deux méthodes présentées ci-dessous sont souvent utilisées pour construire le système de voisinage. Une méthode consiste à définir une fenêtre d'une certaine taille. L'autre méthode consiste à définir un graphe non orienté où chaque sommet représente une observation et chaque arête indique une relation d'adjacence. La construction du système de voisinage est indispensable pour les méthodes de clustering qui tiennent compte de la dépendance entre des observations proches.

Système de voisinage défini par une fenêtre

Selon une mesure de distance choisie, nous définissons une fenêtre dont la taille ϵ est spécifiée *a priori*. Cette fenêtre permet de définir les voisins de $\mathbf{x}_i$ comme l'ensemble

d'observations x_j qui vérifient :

$$d(\mathbf{x}_i, \mathbf{x}_j) < \epsilon,\ j \neq i \tag{2.8}$$

Autrement dit, la fenêtre permet de filtrer toutes les observations autour de $\mathbf{x}_i$. Un exemple avec la mesure de distance euclidienne est montré dans la figure 2.3. La taille de la fenêtre est caractérisée par le rayon ϵ. Le voisinage de chaque observation est : $V_1 = \{l_2\}, V_2 = \{l_1, l_3\}, V_3 = \{l_2, l_4\}, V_4 = \{l_3\}, V_5 = \emptyset$.

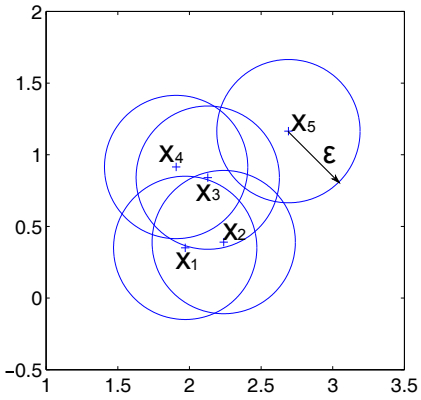


Figure 2.3 – fenêtre de voisinage en dimension 2 avec la distance euclidienne

Système de voisinage défini par un graphe

Une autre construction de système de voisinage est basée sur la théorie des graphes [80]. Un graphe peut être noté $G(E_{\mathbf{x}}, \mathbf{E})$ avec $E_{\mathbf{x}}$ l'ensemble des sommets et $\mathbf{E}$ l'ensemble des arêtes. Un graphe non orienté est capable de définir un système de voisinage, car chaque arête représente une adjacence binaire entre deux sommets [28].

Trois types de graphes sont couramment mentionnés dans la littérature : la triangulation de Delaunay [19, 15], le graphe de Gabriel [55] et l'arbre couvrant de poids minimal [27, 89].

- La triangulation est une méthode qui permet de faire une approximation naturelle d'une surface par décompositions en triangles [15]. Parmi toutes les triangulations, la triangulation de Delaunay vise à maximiser le plus petit angle de l'ensemble des angles des triangles qui composent le modèle de surface. Cette triangulation est telle qu'aucune observation ne soit à l'intérieur du cercle qui circonscrit un triangle sauf les sommets du triangle. Donc, elle évite tous les triangles 'allongés'. Un exemple de la triangulation de Delaunay est présenté par la

figure 2.4a. Il est possible d'étendre la définition dans le cas multi-dimensionnel en remplaçant les triangles et cercles respectivement par des simplexes et hyper-sphères.

- Le graphe de Gabriel est un sous-graphe de la triangulation de Delaunay. Deux sommets $\mathbf{x}_i$ et $\mathbf{x}_j$ sont connectés par une arête du graphe de Gabriel si et seulement si le cercle ayant la distance $d(\mathbf{x}_i, \mathbf{x}_j)$ comme diamètre ne contient aucune autre observation. En partant d'une triangulation de Delaunay, le graphe de Gabriel est obtenu en enlevant simplement toutes les arêtes e_{ij} entre les points x_i et x_j qui ne vérifient pas la condition suivante :

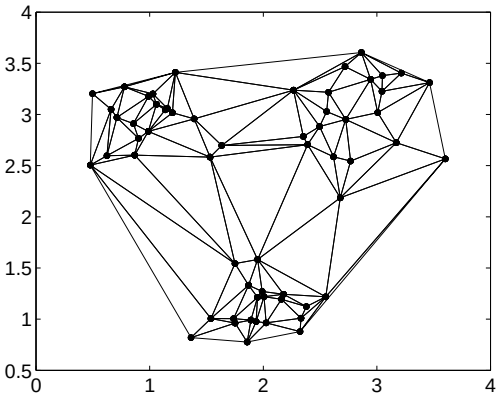
$$d(\mathbf{x}_i, \mathbf{x}_j)^2 < d(\mathbf{x}_i, \mathbf{x}_h)^2 + d(\mathbf{x}_j, \mathbf{x}_h)^2, \ \forall \mathbf{x}_h \in \mathcal{X} \tag{2.9}$$

  Nous illustrons le graphe de Gabriel dans la figure 2.4b par rapport à la triangulation de Delaunay de la figure 2.4a pour la mesure de distance euclidienne.
- Étant donné un graphe non orienté connexe dont les arêtes sont pondérées, un arbre couvrant de poids minimal (MST) est un arbre couvrant dont la somme des poids des arêtes, correspondant à la distance euclidienne dans notre cas, est minimale. L'arbre couvrant de poids minimal est unique pour un graphe dont chaque arête a un poids distinct. Deux méthodes sont souvent utilisées pour obtenir un MST : la méthode de Prim [62] et la méthode de Kruskal [45]. Nous appliquons le premier algorithme sur les sommets montrés dans la figure 2.4a, le résultat est illustré par la figure 2.4c.

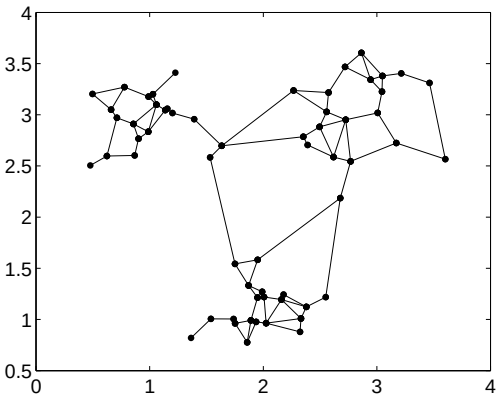
Les trois graphes sont capables de définir un système de voisinage. Toutefois, la triangulation de Delaunay connecte deux sommets qui sont sur l'enveloppe convexe, même s'ils sont relativement éloignés. Quant au MST, le nombre d'arêtes est assez limité. Dans ce cas, il est possible que deux observations ne soient pas liées même si elles sont relativement proches. Cela conduit au fait qu'il y a moins de relations de voisinages.

2.4 Clustering classique

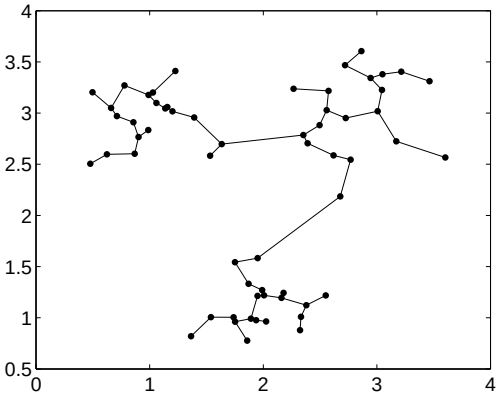
Dans cette section, nous rappelons quelques méthodes classiques de clustering de la première catégorie où les individus sont caractérisés seulement par l'attribut. Nous présentons cinq catégories de méthodes les plus mentionnées dans la littérature : les méthodes hiérarchiques, les méthodes de partitionnement, les méthodes à noyau, les méthodes basées sur la densité et les méthodes basées sur les graphes. Les deux dernières méthodes s'appuient sur un système de voisinage.



(a) la triangulation de Delaunay



(b) le graphe de Gabriel



(c) l'arbre couvrant de poids minimal

Figure 2.4 – des systèmes de voisinage basés sur les graphes

2.4.1 Méthodes hiérarchiques

Les méthodes hiérarchiques visent à chercher récursivement une hiérarchie de clusters en produisant un dendrogramme. Ce dernier permet d'illustrer l'arrangement des

groupes en une structure hiérarchique. La racine du dendrogramme représente l'ensemble des observations, tandis que les feuilles désignent individuellement les observations. En général, ce dendrogramme peut être construit par deux approches [39] :

1. L'approche ascendante. Elle démarre en considérant que chaque observation représente une classe, puis fusionne successivement les paires d'observations les plus similaires. Elle répète cette fusion jusqu'à ce que toutes les observations soient regroupées dans une seule classe.
2. L'approche descendante. Elle démarre avec une classe unique qui contient toutes les observations, puis divise successivement la classe en séparant les observations les plus éloignées. Elle répète la division jusqu'à ce qu'il existe autant de classes que d'observations.

En pratique, l'approche ascendante est souvent préférée, car elle correspond mieux à la conception de regroupement [23]. Étant donné un ensemble de N observations $E_{\mathbf{x}} = \{\mathbf{x}_i\}_{i=1}^{N}$, nous décrivons les étapes de l'approche ascendante comme suit :

1. Démarrer avec une partition initiale telle que chaque individu représente une classe : $\mathcal{P}^t = \{C_i^t = \{\mathbf{x}_i\}\}_{i=1}^{N}$. Nous définissons au départ $t = 0$ comme le niveau de hiérarchie.
2. Calculer la matrice de distance $D^t = \left[d(C_i^t, C_j^t)\right]_{(N-t)\times(N-t)}$ en se basant sur une mesure de distance. $d(C_i^t, C_j^t)$ est une mesure entre les classes d'observations, et nous présentons dans les paragraphes suivants les différentes définitions possibles de cette mesure.
3. Déterminer la paire de classes à regrouper, e.g. (C_i^t, C_j^t) parmi toutes les paires possibles tel que :

$$d(C_i^t, C_j^t) = \min_{m,n} d(C_m^t, C_n^t),\ \forall C_m^t, C_n^t \in \mathcal{P}^t,\ m \neq n \tag{2.10}$$

4. Au niveau suivant $t \leftarrow t + 1$, regrouper C_i^t et C_j^t et mettre à jour $\mathcal{P}^t$
5. Répéter les étapes 2,3,4 jusqu'à ce qu'il n'existe qu'une seule classe dans $\mathcal{P}^t$

Le problème essentiel de telles méthodes se trouve à l'étape 2 qui concerne la mesure de distance entre les classes. Les méthodes hiérarchiques se distinguent les unes des autres par la définition de cette mesure. Pour illustrer cela, nous commençons par introduire deux méthodes qui sont les plus connues dans ce cadre : *single-link* [68] et *complete-link* [70].

Pour la méthode single-link, la mesure de distance $d(C_i, C_j)$ est égale à la distance minimale entre toutes les paires d'observations issues de C_i et C_j. Au contraire, la

méthode complete-link utilise la distance maximale. Ces deux mesures sont respectivement définies mathématiquement comme suit :

$$\begin{aligned} d(C_i, C_j) &= \min d(\mathbf{x}_h, \mathbf{x}_s),\ \forall \mathbf{x}_h \in C_i,\ \forall \mathbf{x}_s \in C_j \\ d(C_i, C_j) &= \max d(\mathbf{x}_h, \mathbf{x}_s),\ \forall \mathbf{x}_h \in C_i,\ \forall \mathbf{x}_s \in C_j \end{aligned}$$

Ces deux méthodes sont illustrées avec un exemple de 5 observations montré dans la figure 2.5. Nous montrons les mises à jour des matrices D^t à chaque étape dans les deux cas. Ensuite, les dendrogrammes sont également présentés. La distance euclidienne a été utilisée comme mesure de distance entre observations pour cet exemple.

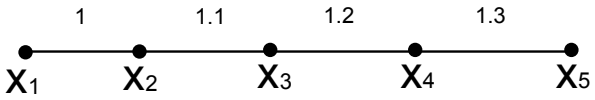


Figure 2.5 – exemple de 5 données

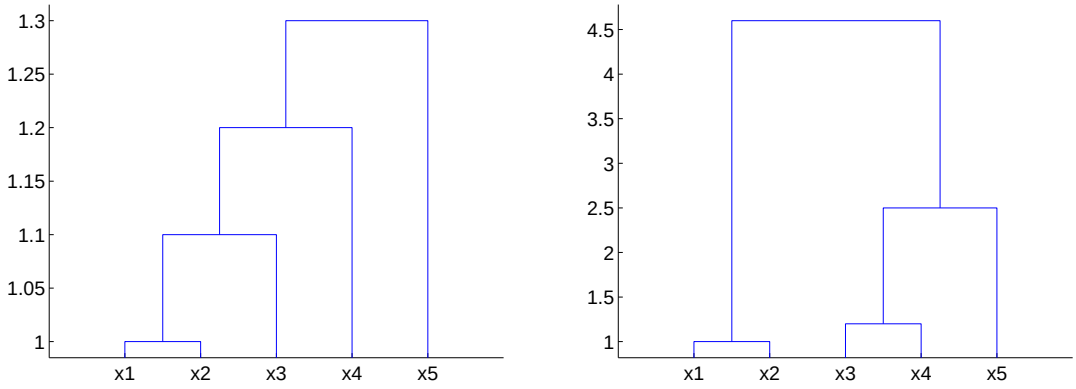
$$D^0_{\mathbf{X}} = \begin{bmatrix} 0 & 1 & 2.1 & 3.3 & 4.6 \\ & 0 & 1.1 & 2.3 & 3.6 \\ & & 0 & 1.2 & 2.5 \\ & & & 0 & 1.3 \\ & & & & 0 \end{bmatrix}$$

$$D^1_{\mathbf{X}} = \begin{bmatrix} 0 & 1.1 & 2.3 & 3.6 \\ & 0 & 1.2 & 2.5 \\ & & 0 & 1.3 \\ & & & 0 \end{bmatrix} \qquad D^1_{\mathbf{X}} = \begin{bmatrix} 0 & 2.1 & 3.3 & 4.6 \\ & 0 & 1.2 & 2.5 \\ & & 0 & 1.3 \\ & & & 0 \end{bmatrix}$$

$$D^2_{\mathbf{X}} = \begin{bmatrix} 0 & 1.2 & 2.5 \\ & 0 & 1.3 \\ & & 0 \end{bmatrix} \qquad D^2_{\mathbf{X}} = \begin{bmatrix} 0 & 3.3 & 4.6 \\ & 0 & 2.5 \\ & & 0 \end{bmatrix}$$

$$D^3_{\mathbf{X}} = \begin{bmatrix} 0 & 1.3 \\ & 0 \end{bmatrix} \qquad D^3_{\mathbf{X}} = \begin{bmatrix} 0 & 4.6 \\ & 0 \end{bmatrix}$$

Les clusters obtenus avec le single-link sont formés par des distances petites sur le dendrogramme, puisqu'on cherche à chaque étape les observations dont la distance est minimale. Elle a tendance à favoriser les groupes allongés (ceci est aussi appelé effet de chaîne). Au contraire, la méthode complete-link trouve les clusters avec la distance la plus importante et elle est mieux adaptée aux clusters compacts.



(a) Single-link (b) Complete-Link

Figure 2.6 – Single-link et Complete-link

La matrice de distance est mise à jour à chaque itération selon le procédé suivant : étant donné la fusion des classes C_i^t et C_j^t à la hiérarchie t pour former la classe $C_{i,j}^{t+1}$ à la hiérarchie $t+1$, la mise à jour des distances entre $C_{i,j}^{t+1}$ et les autres clusters existants C_s^t, $(s \neq i, j)$ peut être présentée par l'équation suivante [46] :

$$d(C_{i,j}^{t+1}, C_s^t) = a_i d(C_i^t, C_s^t) + a_j d(C_j^t, C_s^t) + b d(C_i^t, C_j^t) + c|d(C_i^t, C_s^t) - d(C_j^t, C_s^t)| \quad (2.11)$$

Les deux méthodes single-link et complete-link correspondent à des cas particuliers de cette équation. Nous obtenons le single-link en prenant $a_i = a_j = \frac{1}{2}, b = 0, c = -\frac{1}{2}$, et la méthode complete-link, en prenant $a_i = a_j = \frac{1}{2}, b = 0, c = \frac{1}{2}$ [74].

Il existe d'autres méthodes qui peuvent être considérées comme des compromis entre single-link et complete-link. Elles se distinguent par des valeurs différentes des coefficients a_i, a_j, b et c dans l'équation 2.11. Nous les résumons dans le tableau 2.1 en utilisant la même terminologie que [37]. L'acronyme 'PGM' représente 'la méthode du groupe en paire' (Pair Group Method). 'W' et 'U' représentent respectivement 'pondéré' (Weighted) et 'non pondéré' (Un-weighted). Une méthode pondérée donne le même poids à chaque classe, donc les observations d'une classe moins peuplée ont plus de poids que celles des classes les plus grandes. 'A' et 'C' représentent 'moyen' (Average) et 'centroïde'. n_i est le nombre d'observations dans le cluster C_i.

Une fois la hiérarchie construite en fonction de la méthode choisie, il suffit de déterminer le niveau de coupure du dendrogramme. Pour cela, il est nécessaire de préciser le nombre de clusters attendu, ou bien le seuil de dissimilarité entre classes.

Ces méthodes sont classiques, mais elles souffrent du problème de très grande complexité qui peut atteindre $O(N^3)$. Quand la base de données est grande, le temps de

Tableau 2.1 – Différents algorithmes du clustering hiérarchique

	a_i	a_j	b	c
Single-link	$\frac{1}{2}$	$\frac{1}{2}$	0	$-\frac{1}{2}$
Complete-link	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2}$
WPGMA	$\frac{1}{2}$	$\frac{1}{2}$	0	0
UPGMA	$\frac{n_i}{n_i+n_j}$	$\frac{n_j}{n_i+n_j}$	0	0
WPGMC	$\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{4}$	0
UPGMC	$\frac{n_i}{n_i+n_j}$	$\frac{n_j}{n_i+n_j}$	$\frac{-n_i n_j}{(n_i+n_j)^2}$	0
Ward	$\frac{n_i+n_s}{n_i+n_j+n_s}$	$\frac{n_j+n_s}{n_i+n_j+n_s}$	$\frac{-n_s}{n_i+n_j+n_s}$	0

calcul devient problématique. C'est pourquoi plusieurs autres méthodes ont été proposées [30, 53, 31, 41] dans l'objectif de réduire la complexité.

2.4.2 Méthodes de partitionnement

Contrairement aux méthodes hiérarchiques qui construisent une séquence ordonnée de partitions, les méthodes de partitionnement visent à trouver directement une partition de l'ensemble d'observations en K groupes. Le principe de ces méthodes consiste à optimiser un certain critère avec K défini *a priori*. Généralement, les méthodes peuvent être divisées en deux catégories en fonction des différents types de critères : les critères basés sur la dispersion et les critères qui utilisent les fonctions de densité.

Clustering basé sur le critère de la dispersion

Le critère de l'erreur quadratique est le plus utilisé dans le cadre du clustering de partitionnement. S'appuyant sur les notations définies précédemment, il peut être formalisé comme suit :

$$J(\mathbf{m}_1, \ldots, \mathbf{m}_K) = \sum_{k=1}^{K} \sum_{\mathbf{x}_i \in C_k} \| \mathbf{x}_i - \mathbf{m}_k \|^2 \tag{2.12}$$

avec $\mathbf{m}_k$ le vecteur moyen des observations dans la classe C_k. Nous pouvons aussi faire le lien entre ce critère et les matrices de dispersion [18] où la matrice de dispersion totale S_T est décrite par la somme de la matrice de dispersion *intra-classe* S_W et de la matrice de dispersion *inter-classes* S_B :

$$S_T = S_W + S_B \tag{2.13}$$

où S_W est définie par :

$$S_W = \sum_{k=1}^{K} \sum_{\mathbf{x}_i \in C_k} (\mathbf{x}_i - \mathbf{m}_k)^T (\mathbf{x}_i - \mathbf{m}_k) \tag{2.14}$$

et S_B est définie par :

$$S_B = \sum_{k=1}^{K} n_k (\mathbf{m}_k - \mathbf{m})^T (\mathbf{m}_k - \mathbf{m}) \tag{2.15}$$

avec n_k le nombre d'observations dans C_k et $\mathbf{m}$ le vecteur moyen de l'ensemble $E_{\mathbf{x}}$.

Le critère 2.12 est équivalent à la trace trS_W définie comme la somme d'éléments diagonaux de S_W. Par conséquent, minimiser l'erreur quadratique J revient à minimiser trS_W. De plus, comme la trace trS_T est constante pour un ensemble $E_{\mathbf{x}}$, elle est indépendante de la partition. Minimiser trS_W est aussi equivalent à maximiser trS_B. Beaucoup de méthodes basées sur S_W et S_B ont été proposées parmi lesquelles la plus connue est la méthode *K-means* dont le principe le suivant :

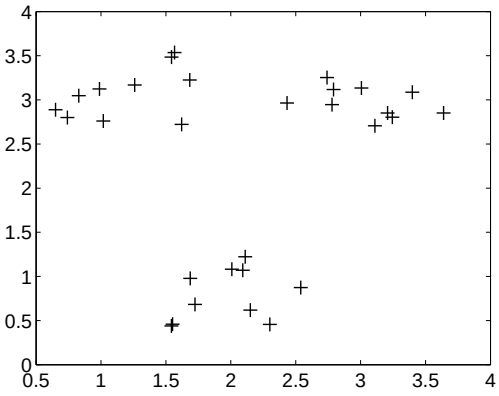
1. Sélectionner K centroïdes, puis répéter les étapes 2 et 3 jusqu'à ce que les appartenances de données se stabilisent.
2. Générer une nouvelle partition en distribuant chaque donnée à la classe dont le centroïde est le plus proche.
3. Mettre à jour les centroïdes.

La figure 2.7 montre un exemple de déroulement de K-means. En partant de l'étape initiale avec les observations à regrouper, les trois classes trouvées se stabilisent successivement en mettant à jour leurs centroïdes qui sont illustrés par les cercles dans la figure.

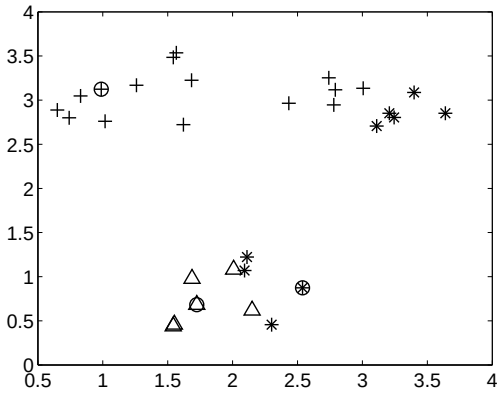
La méthode K-means est une des méthodes les plus classiques, car elle est simple à implémenter. Toutefois, il y a aussi quelques désavantages à noter.

Un problème important pour cette méthode concerne le choix du nombre de clusters. Certaines solutions heuristiques ont été proposées [75] pour déterminer ce paramètre. Typiquement, la méthode est mise en oeuvre pour plusieurs valeurs de K et il s'agit de sélectionner parmi toutes les partitions obtenues celle qui correspond le mieux au problème. Malheureusement, il n'existe pas encore de solution universelle pour ce choix, et la sélection du K dépend beaucoup de l'expérience d'expert.

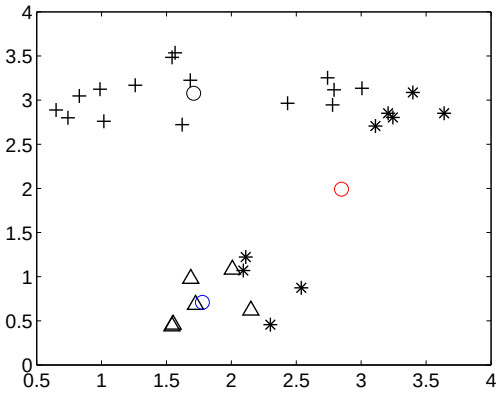
En outre, K-means est sensible à la partition initiale. Elle permet de converger vers une solution optimale locale et aucune méthode ne permet de savoir si cet optimum local correspond aussi à l'optimum global. Dans la plupart des cas, la solution de ce



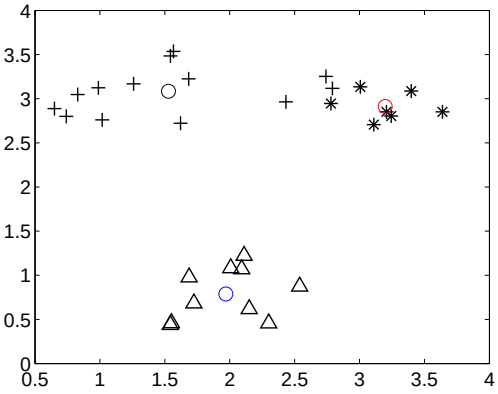
(a) exemple de données



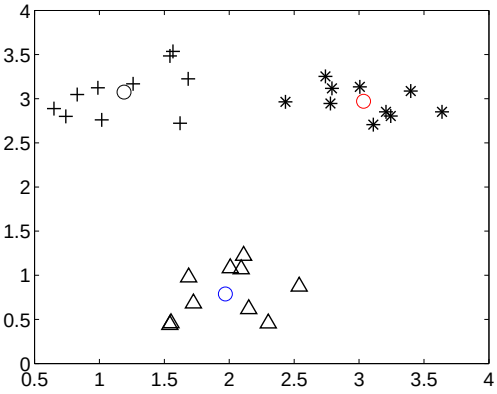
(b) itération 1



(c) itération 2



(d) itération 3



(e) itération 4

Figure 2.7 – exemple d'application de la procédure K-means

problème consiste à choisir plusieurs initialisations avec une valeur de K fixée, puis à sélectionner la partition qui correspond à la valeur minimale du critère 2.12 [60].

De plus, cette méthode est sensible aux observations atypiques. La valeur du centroïde de C_k dépend de toutes les observations dans cette classe, même s'il existe des

observations atypiques qui sont relativement éloignées du centroïde. C'est pourquoi une variante du critère de l'erreur quadratique appelée K-médoïdes a été proposée [42]. Par rapport à K-means, cette méthode a l'objectif de minimiser la somme de la distance entre les observations et les médoïdes. Le médoïde de la classe C_k est défini comme l'observation dont la distance en moyenne avec toutes les autres observations dans cette classe est minimale. Autrement dit, le médoïde dans une classe est l'observation située la plus au centre. Supposons que l'ensemble de médoïdes est $E_{\mathbf{x}^\lambda} = \{\mathbf{x}_1^\lambda, \ldots, \mathbf{x}_K^\lambda\}$, nous cherchons à minimiser le critère ci après :

$$J(\mathbf{x}_1^\lambda, \ldots, \mathbf{x}_K^\lambda) = \sum_{\mathbf{x}_k^\lambda \in E_{\mathbf{x}^\lambda}} \sum_{\mathbf{x}_i \in C_k \setminus \mathbf{x}_k^\lambda} d(\mathbf{x}_i, \mathbf{x}_k^\lambda) \tag{2.16}$$

En général, la méthode K-médoïdes a deux avantages par rapport à K-means. Tout d'abord, elle est moins sensible aux données atypiques, comme un médoïde est une observation et non un vecteur moyen. De plus, cette méthode peut s'adapter aux données dans le domaine discret, tandis que K-means est applicable uniquement dans le domaine continu. Cependant, la complexité algorithmique de K-médoïdes est plus grande.

Critère basé sur la fonction de densité de probabilité

D'un point de vue probabiliste, nous supposons que toutes les observations sont générées selon une certaine fonction de densité de probabilité. L'hypothèse est faite que les observations dans le même cluster sont issues de la même fonction. Les méthodes dans ce cadre consistent à faire l'hypothèse de la distribution et nécessitent d'estimer les paramètres inconnus. Pour la suite, nous présentons tout d'abord le modèle de mélange fini et l'algorithme EM. Ensuite, deux approches basées sur l'algorithme EM sont introduites.

Le modèle de mélange fini considère que les observations $E_{\mathbf{x}} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ proviennent des différentes classes avec les appartenances $E_z = \{z_1, \ldots, z_N\}$ inconnues [56]. L'appartenance de chaque observation z_i correspond à la numérotation de classe où $z_i = k$ signifie que $\mathbf{x}_i$ appartient à la classe k. Le modèle de mélange contenant K composantes peut être écrit sous la forme :

$$\begin{aligned} f(\mathbf{x}_i) &= \sum_{k=1}^{K} P(z_i = k) f(\mathbf{x}_i \mid z_i = k) \\ &= \sum_{k=1}^{K} \pi_k f_k(\mathbf{x}_i) \end{aligned} \tag{2.17}$$

où $\pi_k = P(z_i = k)$ représente la proportion du mélange :

$$\pi_k \in [0,1], \quad \forall k \qquad \text{et} \qquad \sum_{k=1}^{K} \pi_k = 1 \tag{2.18}$$

Le terme f_k représente la fonction de densité de la composante k. Donc, $f(\mathbf{x}_i)$ peut être vue comme une distribution mélangée dont chaque composante $f_k(\mathbf{x}_i)$ est associée à une classe.

En général, on suppose que toutes les fonctions appartiennent à une même famille mais chacune est caractérisée par des paramètres différents définis par $\Theta = \{\theta_k\}_{k=1}^{K}$. L'équation 2.17 peut alors s'écrire :

$$f(\mathbf{x}_i \mid \Phi) = \sum_{k=1}^{K} \pi_k f(\mathbf{x}_i \mid \theta_k) \tag{2.19}$$

où $\Phi = (\pi_1, \ldots, \pi_K, \theta_1, \ldots, \theta_k)$ représente l'ensemble des paramètres du modèle. Un exemple de modèle de mélange est proposé par la Figure 2.8.

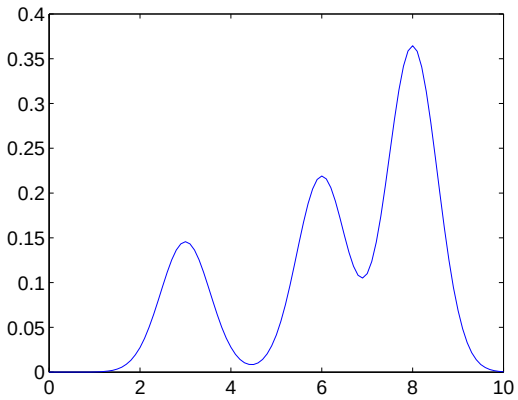


Figure 2.8 – un exemple de modèle de mélange de 3 gaussiennes en 1-D

Le clustering avec le modèle de mélange peut se faire selon deux approches [29] : l'approche mélange et l'approche classification. Les deux approches s'appuient sur l'algorithme EM (Expectation-Maximization) [16, 44] dont nous présentons brièvement le principe.

L'algorithme EM vise à estimer les paramètres Φ en maximisant la vraisemblance ou de façon équivalente la log-vraisemblance d'un modèle de mélange avec les observa-

tions :

$$L(E_{\mathbf{x}} \mid \Phi) = \log \left(\prod_{i=1}^{N} f(\mathbf{x}_i \mid \Phi) \right) = \sum_{i=1}^{N} \log \left(\sum_{k=1}^{K} \pi_k f(\mathbf{x}_i \mid \theta_k) \right) \tag{2.20}$$

La fonction 2.20 est non linéaire par rapport à Φ et aucune solution analytique n'est disponible. L'algorithme EM cherche une solution en utilisant une procédure itérative dans laquelle on introduit la notion de données complètes, fondamentale pour cet algorithme.

L'ensemble d'observations $E_{\mathbf{x}}$ défini précédemment peut être interprété comme une information partielle des données complètes, notées $E_{\mathbf{w}} = (E_{\mathbf{x}}, E_z)$ où E_z est appelé information manquante. Dans le contexte du clustering, $E_z = \{z_1, \ldots, z_N\}$ correspond à la classe d'appartenance qui indique les origines des observations. La vraisemblance calculée à partir de $E_{\mathbf{w}}$ est appelée la vraisemblance complétée. Selon le théorème de Bayes, on obtient la relation de densité de probabilité sous la forme :

$$f(E_{\mathbf{w}} \mid \Phi) = f(E_{\mathbf{w}} \mid E_{\mathbf{x}}, \Phi) f(E_{\mathbf{x}} \mid \Phi) \tag{2.21}$$

En prenant le logarithme de cette égalité, une relation entre la log-vraisemblance des observations définie par $L(E_{\mathbf{x}} \mid \Phi)$ et la log-vraisemblance complétée définie par $L_c(E_{\mathbf{w}} \mid \Phi)$ est donnée par :

$$L(E_{\mathbf{x}} \mid \Phi) = L_c(E_{\mathbf{w}} \mid \Phi) - \log f(E_{\mathbf{w}} \mid E_{\mathbf{x}}, \Phi) \tag{2.22}$$

En introduisant la notation Φ' qui représente les estimations des paramètres actuels, puis en multipliant chaque terme de l'équation 2.22 par $f(E_z \mid E_{\mathbf{x}}, \Phi')$, et en faisant la somme par rapport à E_z on obtient les espérances comme suit :

$$\begin{aligned} L(E_{\mathbf{x}} \mid \Phi) &= \sum_{E_z} f(E_z \mid E_{\mathbf{x}}, \Phi') L_c(E_{\mathbf{w}} \mid \Phi) - \sum_{E_z} f(E_z \mid E_{\mathbf{x}}, \Phi') \log f(E_{\mathbf{w}} \mid E_{\mathbf{x}}, \Phi) \\ &= \mathbf{E}\left[L_c(E_{\mathbf{w}} \mid \Phi) \mid E_{\mathbf{x}}, \Phi' \right] - \mathbf{E}\left[\log f(E_{\mathbf{w}} \mid E_{\mathbf{x}}, \Phi) \mid E_{\mathbf{x}}, \Phi' \right] \\ &= Q(\Phi, \Phi') - H(\Phi, \Phi') \end{aligned}$$

Le terme de gauche $L(E_{\mathbf{x}} \mid \Phi)$ ne change pas, car il est indépendant de E_z. On peut aussi montrer que la fonction $H(\Phi, \Phi')$ est maximum pour $\Phi = \Phi'$ selon l'inégalité de Jensen [16]. Donc le paramètre Φ qui maximise le terme $Q(\Phi, \Phi')$ conduit au fait que : $L(E_{\mathbf{x}} \mid \Phi) \geqslant L(E_{\mathbf{x}} \mid \Phi')$. Par conséquent, il suffit de construire une succession de $\Phi^{q+1} = \operatorname{argmax} Q(\Phi, \Phi^q)$ à partir d'une initialisation Φ^0 et on trouve une suite croissante de $L(E_{\mathbf{x}} \mid \Phi)$.

L'algorithme EM peut être mis en oeuvre en deux étapes : l'étape d'espérance et l'étape de maximisation. Ces deux étapes sont décrites comme suit :

- Étape d'espérance : cette étape consiste à calculer le terme $Q(\Phi, \Phi')$ connaissant l'ensemble des observations $E_{\mathbf{x}}$ et les paramètres à la $q^{ème}$ itération Φ^q. Nous définissons la valeur binaire de ν telle que $\nu_{ik} = 1$ si $z_i = k$ et $\nu_{ik} = 0$ sinon.

$$\begin{aligned}
Q(\Phi, \Phi^q) &= E\left[L_c(E_{\mathbf{w}} \mid \Phi) \mid E_{\mathbf{x}}, \Phi^q\right] \\
&= E\left[\log\left(\prod_{i=1}^{N} f(\mathbf{x}_i, z_i \mid \Phi)\right) \mid E_{\mathbf{x}}, \Phi^q\right] \\
&= E\left[\sum_{i=1}^{N}\sum_{k=1}^{K} \nu_{ik} \log\left(P(z_i = k) f(\mathbf{x}_i | z_i = k, \theta_k)\right) \mid E_{\mathbf{x}}, \Phi^q\right] \\
&= E\left[\sum_{i=1}^{N}\sum_{k=1}^{K} \nu_{ik} \log\left(\pi_k f(\mathbf{x}_i \mid \theta_k)\right) \mid E_{\mathbf{x}}, \Phi^q\right] \\
&= \sum_{i=1}^{N}\sum_{k=1}^{K} E\left[\nu_{ik} \mid E_{\mathbf{x}}, \Phi^q\right] \log\left(\pi_k f(\mathbf{x}_i \mid \theta_k)\right) \\
&= \sum_{i=1}^{N}\sum_{k=1}^{K} p(\nu_{ik} = 1 \mid E_{\mathbf{x}}, \Phi^q) \log\left(\pi_k f(\mathbf{x}_i \mid \theta_k)\right)
\end{aligned}$$

  Afin d'obtenir la valeur de $Q(\Phi, \Phi^q)$, il est nécessaire de calculer le terme $P(\nu_{ik} = 1 \mid E_{\mathbf{x}}, \Phi^q)$, i.e. la probabilité *a posteriori* que $\mathbf{x}_i$ appartienne à la composante k. Cette probabilité est souvent décrite comme c_{ik}^q sous la forme :

$$c_{ik}^q = P(\nu_{ik} = 1 \mid E_{\mathbf{x}}, \Phi^q) = \frac{\pi_k f(\mathbf{x}_i \mid \theta_k^q)}{\sum_{l=1}^{K} \pi_l f(\mathbf{x}_i \mid \theta_l^q)} \tag{2.23}$$

  Par conséquent, $Q(\Phi, \Phi^q)$ peut s'écrire comme suit :

$$Q(\Phi, \Phi^q) = \sum_{i=1}^{N}\sum_{k=1}^{K} c_{ik}^q \log \pi_k + \sum_{i=1}^{N}\sum_{k=1}^{K} c_{ik}^q \log f(\mathbf{x}_i \mid \theta_k) \tag{2.24}$$

- Étape de maximisation : cette étape consiste à mettre à jour les paramètres Φ en maximisant l'équation 2.24. Cette maximisation peut être réalisée respectivement par rapport aux proportions de mélange π_k en maximisant le premier terme de cette équation et aux paramètres θ_k en maximisant le deuxième terme. Pour le premier terme, la valeur de proportion de chaque itération π_k^{q+1} est donnée par la relation ci après :

$$\pi_k^{q+1} = \frac{\sum_{i=1}^{N} c_{ik}^q}{N}$$

La maximisation du deuxième terme dépend du modèle de mélange et la solution peut être obtenue en résolvant l'équation de vraisemblance :

$$\sum_{i=1}^{N}\sum_{k=1}^{K} c_{ik}^{q}\frac{\partial}{\partial\theta_k}\log f(\mathbf{x}_i \mid \theta_k) = 0 \tag{2.25}$$

L'approche mélange et l'approche classification sont deux approches de clustering basées sur l'algorithme EM. Elles sont décrites ci-après :

1. L'approche mélange peut être interprétée comme une étape supplémentaire de l'algorithme EM à la fin duquel on détermine les paramètres du modèle de mélange et les valeurs de c_{ik}. L'approche mélange consiste à allouer chaque observation à la classe qui maximise la valeur c_{ik}. Autrement dit, les appartenances d'individus sont déterminées selon le principe de MAP.
2. L'approche classification vise à maximiser directement la vraisemblance complétée en s'appuyant sur les données complétées qui contiennent les observations et les appartenances [73] :

$$L_c(E_{\mathbf{w}} \mid \Phi) = \sum_{i=1}^{N}\sum_{k=1}^{K} \nu_{ik}\log\left(\pi_k f(\mathbf{x}_i \mid \theta_k)\right) \tag{2.26}$$

Une méthode CEM (Classification Espérance-Maximisation) est proposé dans [11] pour trouver une solution à la maximisation du critère 2.26. Cette méthode peut être vue comme une variante de l'algorithme EM en ajoutant une étape de classification entre l'étape d'espérance et celle de maximisation. À chaque itération, on attribue à chaque itération les observations aux classes selon le principe de MAP. La méthode CEM est décrite comme ci-après :

- Étape initiale : choisir arbitrairement les paramètres initiaux $\Phi^0 = (\pi_1^0, \ldots, \pi_K^0, \theta_1^0, \ldots, \theta_K^0)$
- Étape E : estimer les probabilités *a posteriori* de chaque observation selon l'équation 2.23.
- Étape C : créer une partition en mettant chaque observation dans la classe qui lui donne la probabilité *a posteriori* maximale. Plus formellement, l'appartenance de la $i^{ème}$ observation vérifie la relation suivante :

$$z_i^{q+1} = \operatorname{argmax} c_{ik}^{q} \quad (i = 1 \ldots N;\ k = 1 \ldots K)$$

Cette étape consiste à remplacer les probabilités c_{ik} par les valeurs 0 et 1.

- Étape M : maximiser la vraisemblance conformément aux $z_i^{q+1}\,(i = 1 \ldots N)$. Les

proportions à l'itération $q+1$ sont données par

$$\pi_k^{q+1} = \frac{n_k^q}{N}$$

où n_k^q représente le nombre d'observations attribuées à la classe k. Comme dans l'algorithme EM, le calcul des θ_k^{q+1} dépend du modèle de mélange choisi.

2.4.3 Méthodes à noyau

Les méthodes à noyau sont basées sur le théorème de Cover [13] : soit un ensemble de données qui ne peut pas être séparé de manière linéaire, il a plus de chance d'être séparable linéairement si on le projette dans un espace de dimension supérieure. L'idée principale de ces méthodes est de produire des clusters dans un espace de redescription $\mathcal{F}$ de grande dimension [78, 65]. Par exemple, si les individus dans l'espace initial sont complexes et non linéairement séparables, nous pouvons les transformer par une fonction de redescription ϕ dans un espace où nous avons la possibilité de les séparer de façon linéaire [34].

Le problème des méthodes à noyau concerne d'une part, le choix adéquat de la fonction de redescription ϕ, et d'autre part, le calcul dans l'espace $\mathcal{F}$ dont la dimension est souvent grande. Cependant, il est possible de ne pas avoir à connaître explicitement la fonction ϕ et d'effectuer les calculs grâce à l'utilisation d'une *fonction noyau*. La définition de cette dernière est présentée comme suit :

Définition 2.4. Soit $E_x = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ un ensemble d'observations où $\mathbf{x}_i \in \mathcal{X}$. Une fonction noyau est une fonction $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ satisfaisant :

$$\kappa(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j) \tag{2.27}$$

avec ϕ une fonction dotée d'un produit scalaire de $\mathcal{X}$ vers $\mathcal{F}$: $\phi : \mathcal{X} \to \mathcal{F}$

Cette définition implique que le choix de la fonction noyau κ induit la fonction ϕ. Quelques exemples de fonction noyau sont présentés dans le tableau 2.2. La propriété la plus utilisée en pratique de cette fonction est de calculer la distance euclidienne dans $\mathcal{F}$ sans connaître explicitement ϕ. C'est ce que l'on appelle l'astuce de noyau (*kernel trick*) [58] :

$$\begin{aligned} \| \phi(\mathbf{x}_i) - \phi(\mathbf{x}_j) \|^2 &= \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_i) + \phi(\mathbf{x}_j) \cdot \phi(\mathbf{x}_j) - 2\phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j) \\ &= \kappa(\mathbf{x}_i, \mathbf{x}_i) + \kappa(\mathbf{x}_j, \mathbf{x}_j) - 2\kappa(\mathbf{x}_i, \mathbf{x}_j) \end{aligned}$$

Tableau 2.2 – Exemples de fonction noyau

Linéaire	$\kappa(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i \cdot \mathbf{x}_j$
Polynomial du degré p	$\kappa(\mathbf{x}_i, \mathbf{x}_j) = (1 + \mathbf{x}_i \cdot \mathbf{x}_j)^p,\ p \in \mathbb{N}$
Gaussien	$\kappa(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right),\ \sigma \in \mathbb{R}$

En se basant sur cette astuce, toutes les méthodes de clustering basées sur la mesure de distance euclidienne peuvent être appliquées dans un espace de redescription $\mathcal{F}$ [24, 1]. Une des méthodes les plus mentionnées dans la littérature est le *K-means à noyau* [64]. Cette méthode vise à minimiser le critère de l'erreur quadratique dans l'espace $\mathcal{F}$:

$$J(\mathbf{m}_1, \ldots, \mathbf{m}_K) = \sum_{i=1}^{N} \sum_{k=1}^{K} I_{C_k}(\mathbf{x}_i) \parallel \phi(\mathbf{x}_i) - \mathbf{m}_k \parallel^2 \tag{2.28}$$

avec

$$I_{C_k}(\mathbf{x}_i) = \begin{cases} 1 & \text{si } \mathbf{x}_i \in C_k \\ 0 & \text{sinon} \end{cases} \quad \text{et} \quad \mathbf{m}_k = \frac{\sum_{i=1}^{N} I_{C_k}(\mathbf{x}_i)\phi(\mathbf{x}_i)}{\sum_{i=1}^{N} I_{C_k}(\mathbf{x}_i)} \tag{2.29}$$

Les $\mathbf{m}_k$ sont les centroïdes dans $\mathcal{F}$ et ils ne sont pas directement disponibles, car la fonction de redescription ϕ est souvent inconnue. Cependant, selon l'astuce de noyau, le carré de la distance euclidienne peut être reformulé comme ci-après :

$$\begin{aligned} \parallel \phi(\mathbf{x}_i) - \mathbf{m}_k \parallel^2 &= \kappa(\mathbf{x}_i, \mathbf{x}_i) - \frac{2\sum_{j=1}^{N} I_{C_k}(\mathbf{x}_j)\kappa(\mathbf{x}_i, \mathbf{x}_j)}{\sum_{j=1}^{N} I_{C_k}(\mathbf{x}_j)} \\ &+ \frac{\sum_{j=1}^{N}\sum_{r=1}^{N} I_{C_k}(\mathbf{x}_j) I_{C_k}(\mathbf{x}_r)\kappa(\mathbf{x}_j, \mathbf{x}_r)}{\sum_{j=1}^{N}\sum_{r=1}^{N} I_{C_k}(\mathbf{x}_j) I_{C_k}(\mathbf{x}_r)} \end{aligned} \tag{2.30}$$

Le principe de la méthode K-means à noyau peut être décrit selon les quatres étapes suivantes :

1. Sélectionner arbitrairement K centroïdes, puis répéter les étapes 2 à 4 jusqu'à ce que les appartenances d'individus se stabilisent.
2. Pour chaque observation $\mathbf{x}_i$ et chaque classe C_k, calculer $\parallel \phi(\mathbf{x}_i) - \mathbf{m}_k \parallel^2$ en utilisant l'équation 2.30.
3. Générer une nouvelle partition en distribuant chaque observation à la classe dont le centroïde est le plus proche. Précisément, l'appartenance de chaque observation est mise à jour en vérifiant :

$$I_{C_k}(\mathbf{x}_i) = \begin{cases} 1 & \text{si } \parallel \phi(\mathbf{x}_i) - \mathbf{m}_k \parallel^2 < \parallel \phi(\mathbf{x}_i) - \mathbf{m}_t \parallel^2 \\ 0 & \text{sinon} \end{cases} \quad \forall t \neq k \tag{2.31}$$

L'avantage principal des méthodes à noyau est d'améliorer la séparabilité d'un ensemble de données en le transformant dans un espace $\mathcal{F}$ de dimension supérieure. De plus, l'astuce du noyau permet d'écrire le carré de la distance des observations dans l'espace $\mathcal{F}$ par la fonction noyau. Cela veut dire que toutes les méthodes de clustering classiques basées sur la mesure de distance peuvent être reformulées et appliquées dans l'espace $\mathcal{F}$.

2.4.4 Méthodes basées sur la densité

Ces méthodes considèrent qu'un cluster est un ensemble d'observations dans les zones les plus denses selon un système de voisinage. Généralement, ces méthodes sont capables de trouver des clusters de forme quelconque. De plus, les observations atypiques peuvent être détectées et traitées de façon très efficace. Afin d'illustrer l'idée principale de ces méthodes, nous présentons dans cette section la méthode DBSCAN [21] qui est une des méthodes les plus mentionnées dans la littérature.

La méthode DBSCAN définit un système de voisinage $\mathbf{V} = \{V_i \mid \forall i = 1, \ldots, N\}$ à l'aide d'une fenêtre paramétrée par ϵ selon la définition introduite en section 2.3.2. En utilisant la distance euclidienne, la fenêtre définit chaque voisinage V_i comme une hyper-sphère centrée en $\mathbf{x}_i$ avec le rayon ϵ spécifié *a priori*. On dénote aussi $n(\mathbf{x}_i)$ le nombre d'observations qui appartiennent au voisinage V_i. De plus, une observation $\mathbf{x}_i$ est appelé *noyau* s'il y a au moins n_s voisins dans son voisinage où n_s est un autre coefficient fixé *a priori*. En partant de ces notations, on a deux définitions importantes :

1. L'observation $\mathbf{x}_j$ est *densité-accessible directe* de $\mathbf{x}_i$ si :
   - $\mathbf{x}_j \in V_i$
   - $n(\mathbf{x}_i) \geqslant n_s$
2. L'observation $\mathbf{x}_j$ est *densité-accessible* de $\mathbf{x}_i$ s'il existe une suite d'observations $\mathbf{x}_i, \ldots, \mathbf{x}_m, \ldots, \mathbf{x}_j$ telles que chaque observation est *densité-accessible directe* de celle qui la précède dans la suite.

Reprenons l'exemple de la figure 2.3 avec le coefficient $n_s = 2$. Selon les deux définitions, aucune observation n'est densité-accessible directe de $\mathbf{x}_1$, puisque $n(\mathbf{x}_1) < n_s$. Cependant, $\mathbf{x}_1$ et $\mathbf{x}_3$ sont densité-accessibles directes de $\mathbf{x}_2$. De plus, $\mathbf{x}_4$ est densité-accessible de $\mathbf{x}_2$, car elle est densité-accessible directe de $\mathbf{x}_3$, elle même densité-accessible directe de $\mathbf{x}_2$.

Une observation qui n'est pas 'noyau' peut soit être une observation *limite* qui tombe sur le bord d'un cluster et qui est *densité-accessible* depuis un 'noyau', soit être un *bruit* qui n'est *densité-accessible* depuis aucune observation. La méthode DBSCAN

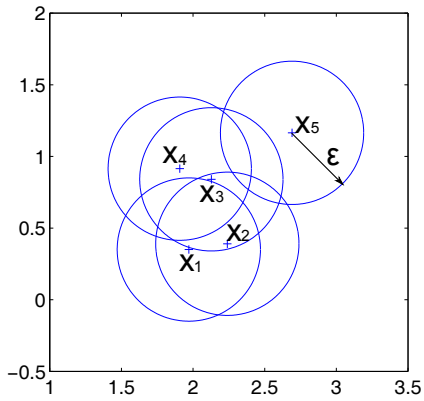


Figure 2.9 – observations et leurs voisinages

comporte les deux étapes suivantes :

1. Choisir aléatoirement une observation $\mathbf{x}_i \in E_{\mathbf{x}}$.
2. Si $\mathbf{x}_i$ n'est pas un noyau, alors la marquer comme un bruit. Sinon, rassembler toutes les observations qui sont densité-accessibles de $\mathbf{x}_i$ dans une classe, même celles déjà marquées comme des bruits.
3. Sélectionner une autre observation et reprendre l'étape 2 jusqu'à ce que toutes les observations dans $E_{\mathbf{x}}$ soient considérées.

Nous n'avons pas besoin de spécifier le nombre de clusters en appliquant cette méthode, car il permet de le trouver lui-même. Cependant, le résultat obtenu est très sensible au choix des coefficients ϵ et n_s, et dépend de la bonne séparabilité des groupes.

2.4.5 Méthodes basées sur les graphes

Les notions de graphe et certains exemples sont présentés dans la section 2.3.2. En général, les méthodes de clustering basées sur les graphes visent à couper les arêtes d'un graphe connexe non orienté où il existe une suite d'arêtes permettant d'atteindre un sommet en partant d'un autre. En enlevant les arêtes *inconsistantes*, le graphe devient non connexe avec des composantes qui sont considérées comme des clusters. Cette suppression est basée sur un critère qui exprime l'influence des arêtes locales d'un système de voisinage. L'avantage de cette approche est de pouvoir trouver des clusters avec des formes variées [74]. Il existe plusieurs méthodes dans ce cadre parmi lesquelles la plus citée est la méthode de clustering basée sur le MST (Minimum Spanning Tree) [88]. La triangulation de Delaunay est aussi un outil important du clustering sur laquelle plusieurs méthodes ont été proposées [19, 22]. Nous décrivons brièvement une méthode basée sur le MST et une méthode basée sur la triangulation de Delaunay.

Méthode basée sur le MST

Étant donné un MST (présenté à la section 2.3.2), la méthode consiste à rechercher les arêtes inconsistantes parmi les arêtes du graphe. Cette méthode est paramétrée par un entier r définissant une profondeur d'exploration du graphe et un réel positif c définissant un seuil de sélection d'arêtes. Pour une arête e_{ij} entre 2 sommets $\mathbf{x}_i$ et $\mathbf{x}_j$, les arêtes e_{kl} de profondeur r depuis le sommet $\mathbf{x}_i$ sont recherchées et de même depuis $\mathbf{x}_j$. Les poids w_{kl} associés à ces arêtes sont utilisés pour calculer la moyenne $m_{e_{ij}}$ et l'écart-type $\sigma_{e_{ij}}$. La décision d'inconsistance de l'arête e_{ij} est prise selon la règle :

$$w_{ij} > m_{e_{ij}} + c\sigma_{e_{ij}} \tag{2.32}$$

Prenons l'exemple illustré sur la figure 2.10 avec $r = 2$ et $c = 6$. On considère l'arête e_{ij} qui est notée e_0 sur la figure. Les arêtes de profondeur $r = 2$ de e_{ij} sont $e_1, e_2, e_3, e_4, e_5, e_7, e_8, e_9, e_{11}$. Pour cette arête, on a $m_{e_{ij}} = 2.33$ et $\sigma_{e_{ij}} = 1.56$. Donc, l'écart entre $w_{ij} = 15$ et $m_{e_{ij}}$ est égale à 8 fois $\sigma_{e_{ij}}$. comme $c < 8$, l'arête e_{ij} est déclarée inconsistante.

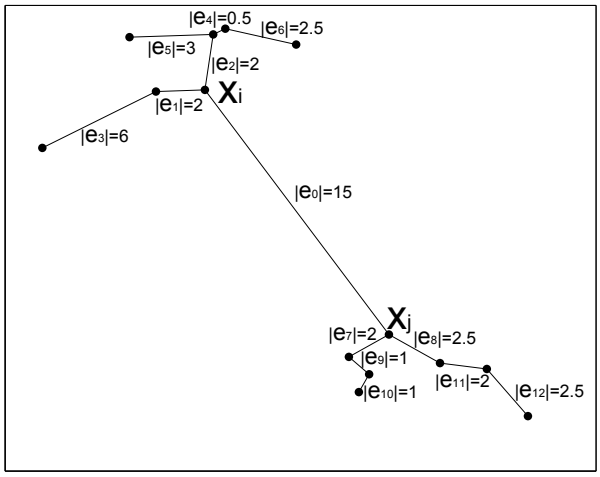


Figure 2.10 – exemple de MST avec les poids d'arêtes

La méthode basée sur cette idée peut être décrite simplement comme suit :

- Construire l'arbre couvrant de poids minimal avec les sommets $E_{\mathbf{x}}$. Calculer le poids w_{ij} pour chaque arête.
- Enlever les arêtes inconsistantes.

Méthode basée sur la triangulation Delaunay

La triangulation Delaunay est beaucoup utilisée dans le cadre du clustering. AUTOCLUST [22] est une méthode souvent mentionnée fondée sur cette triangulation. Le principe général est le suivant : les arêtes sont classées en trois catégories selon les critères locaux et globaux : les arêtes courtes, longues et les autres. On commence par enlever toutes les arêtes courtes et longues. Ensuite, on rétablit quelques arêtes courtes dont les sommets appartiennent à la même classe. Enfin, certaines arêtes de la catégorie 'les autres' sont enlevées si elles créent des chemins entre des sommets à la frontière.

En général, la méthode AUTOCLUST a deux avantages principaux par rapport aux autres méthodes basées sur les graphes. le premier avantage est qu'elle permet de trouver des clusters dispersés à coté de clusters compacts. De plus, elle est capable de résoudre le problème où deux clusters sont liés par plusieurs chaînes.

2.5 Clustering spatial

La section précédente a présenté les méthodes de clustering qui s'adaptent aux individus caractérisés uniquement par l'attribut. En revanche, cette section donne un aperçu des méthodes spécifiques utilisées pour le clustering de données spatiales. Le problème de clustering spatial a été brièvement rappelé dans le section 2.2.3. Il se caractérise par le fait que l'espace des mesures peut être divisé en 2 sous-espaces : l'espace de l'attribut $\mathcal{X}$ et l'espace de covariable $\mathcal{Y}$. Donc, l'ensemble des données est divisé en deux selon chaque sous-espace : $E_{\mathbf{x}}$ et $E_{\mathbf{y}}$.

Chaque individu est dénoté $l_i = (\mathbf{x}_i, \mathbf{y}_i)$, $(i = 1, \ldots, N)$ avec $\mathbf{x}_i \in \mathcal{X}$ et $\mathbf{y}_i \in \mathcal{Y}$. Souvent, la covariable correspond aux coordonnées géographiques. Donc, les méthodes de clustering spatial peuvent être considérées comme traitement des problèmes avec des observations géo-localisées. Le principe de ce genre de méthodes consiste à combiner les effets de la proximité dans l'espace $\mathcal{X}$ et dans l'espace $\mathcal{Y}$. Nous décrivons respectivement dans les sections 2.5.1 et 2.5.2 les approches non probabilistes et les approches probabilistes.

2.5.1 Méthodes non probabilistes

Transformation des attributs

Afin de prendre en compte l'adjacence géographique pour le clustering, une idée naturelle est de définir un système de voisinage qui représente une proximité des individus

dans l'espace géographique. Ce voisinage est ensuite utilisé pour corriger la valeur de chaque observation en fonction des valeurs de ses voisins.

Par exemple, le filtre médian est une méthode basée sur cette approche [6]. Il est souvent utilisé pour le traitement d'image où les niveaux de gris sont considérés comme les observations et les coordonnées de pixels donnent les informations géographiques. Cette technique consiste à remplacer le niveau de gris de chaque pixel par la valeur médiane des niveaux de gris de tous les pixels qui appartiennent à un voisinage donné.

Utilisation de la matrice de distance

Cette méthode, discutée dans [59], calcule tout d'abord la matrice de distance $D = [d_{ij}]_{N\times N}$ de toutes les réalisations dans l'espace $\mathcal{X}$. Cette matrice est ensuite modifiée en intégrant les informations géographiques selon $D^* = \left[d^*_{ij}\right]_{N\times N}$ où $d^*_{ij} = d_{ij} \times g(w_{ij})$ et $g(w_{ij})$ est une fonction des distances w_{ij} dans l'espace de covariable. La nouvelle matrice D^* mélange les informations d'attribut et de covariable. Ensuite, cette matrice est transformée en un tableau individus/variables par une analyse factorielle. Puis, l'algorithme K-means est appliqué pour partitionner les individus décrits dans les sous-espaces des vecteurs propres associés aux plus grandes valeurs propres.

Utilisation de la contrainte des graphes

Selon la section 2.3.2, un graphe non orienté est capable de représenter naturellement la relation de voisinage entre les observations. Cette méthode vise à construire un système de voisinage dans l'espace géographique, et regrouper les observations dans $\mathcal{X}$ en tenant compte de la contrainte du voisinage dans $\mathcal{Y}$. Une méthode, appelée DBSC (density-based spatial clustering) [51], se base sur la triangulation de Delaunay. Cette méthode s'appuie sur des définitions de l'accessibilité analogues aux définitions introduites dans la section 2.4.4. Elles sont formulées comme suit :

- L'individu l_j est *densité accessible spatial* de l_i si :
  1. $\mathbf{y}_j \in V_i$ où V_i, $(i = 1, \ldots, N)$ représente une relation de voisinage définie par la triangulation Delaunay dans l'espace $\mathcal{Y}$.
  2. $d(\mathbf{x}_i, \mathbf{x}_j) \leqslant S$ avec S un seuil spécifié *a priori*.
- L'individu l_j est *accessible spatial* d'une classe C s'il vérifie :
  1. $d(\mathbf{x}_j, \mathbf{m}_C) \leqslant S$ où $\mathbf{m}_C$ est le vecteur moyen de la classe C dans l'espace $\mathcal{X}$.
  2. $\mathbf{y}_j \in V_i$ et $l_i \in C$

- Un *indicateur de densité* $DI(l_i)$ pour l'individu l_i est défini sous la forme :

$$DI(l_i) = n_{sdr}(l_i) + n_{sdr}(l_i)/n_v(l_i) \tag{2.33}$$

  où n_{sdr} est le nombre d'individus qui sont densités accessibles spatiales de l_i, et $n_v(l_i)$ représente le nombre total d'individus qui appartiennent au voisinage V_i
- Un *noyau spatial* est défini comme l'individu qui porte l'indicateur de densité maximale parmi tous les individus non classifiés.
- Un *noyau d'expansion* est un individu avec au moins un voisin qui est densité accessible spatial avec lui.

En se basant sur ces définitions, la méthode se déroule de façon itérative comme suit :

1. Construire la triangulation de Delaunay dans l'espace géographique $\mathcal{Y}$. Enlever les arêtes globalement et localement longues.
2. D'après le graphe de Delaunay modifié, calculer les *indicateurs de densités* pour chaque individu. Sélectionner ensuite un noyau spatial l_i et les noyaux d'expansion qui appartiennent au voisinage V_i.
3. Ajouter successivement les noyaux d'expansion qui vérifient la condition de la *densité accessible spatiale* de l_i. Un cluster peut être trouvé à la fin de cette étape.
4. Ajouter successivement tous les individus qui vérifient la condition *accessible spatiale* au cluster précédent.
5. Répéter les étapes 1 à 4 jusqu'à ce que chaque individu soit affecté à une classe.

Cette méthode est capable de trouver des clusters avec des formes arbitraires sans avoir besoin de définir *a priori* le nombre de classes K. Elle est robuste à la présence de bruit. Par ailleurs, la complexité en temps est $O(N\log N)$.

Bilan

Les méthodes basées sur l'approche non probabiliste présentées brièvement ci-dessus ont pour objectif de regrouper les données spatiales sans utiliser de modèle probabiliste. Pour les différentes méthodes dans ce cadre, on combine la proximité des observations dans l'espace $\mathcal{X}$ et la proximité des informations dans l'espace $\mathcal{Y}$ selon diverses techniques. Aucune loi probabiliste *a priori* n'est associée à l'attribut, ni à la covariable. La section suivante présente des méthodes basées sur l'approche probabiliste où la proximité dans chaque espace peut avoir une interprétation probabiliste.

2.5.2 Méthodes probabilistes

Principe de l'approche probabiliste

Les méthodes probabilistes visent à modéliser la proximité dans l'espace $\mathcal{X}$ et la proximité dans l'espace $\mathcal{Y}$ par des lois de probabilité, et ensuite à exprimer les probabilités *a posteriori* des labels ou la vraisemblance de ces labels pour définir la partition à retenir.

La distribution *a posteriori* de l'ensemble des appartenances $E_z = \{z_1, \ldots, z_N\}$ sachant l'ensemble d'observations $E_{\mathbf{x}} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ peut être exprimée selon le théorème Bayes :

$$P(E_z \mid E_{\mathbf{x}}) = \frac{f(E_{\mathbf{x}} \mid E_z)P(E_z)}{f(E_{\mathbf{x}})} \tag{2.34}$$

où $P(E_z)$ décrit la probabilité *a priori*, et $f(E_{\mathbf{x}} \mid E_z)$ représente la densité de probabilité de l'ensemble d'observations conditionnellement à l'ensemble des classes. La stratégie bayésienne consiste à minimiser le coût *a posteriori* :

$$\hat{E}_z = arg\min_{E_z} g(E_z \mid E_{\mathbf{x}}) \tag{2.35}$$

avec :

$$g(E_z \mid E_{\mathbf{x}}) = \mathbf{E}\left[c(E_z, E_Z^*) \mid E_{\mathbf{x}}\right] = \sum_{E_Z^*} c(E_z, E_Z^*)P(E_Z^* \mid E_{\mathbf{x}}) \tag{2.36}$$

où $c(E_z, E_Z^*)$ est la fonction qui définit le coût de la partition actuelle quand la partition réelle est donnée par E_Z^*. Une des fonctions les plus utilisées est celle du coût 0-1 : $c(E_z, E_Z^*) = I(E_z \neq E_Z^*)$ qui vaut 0 pour la bonne décision et 1 sinon. Donc, l'objectif est de minimiser le critère 2.36 qui peut être développé comme suit :

$$g(E_z \mid E_{\mathbf{x}}) = \sum_{E_Z^* \neq E_z} P(E_Z^* \mid E_{\mathbf{x}}) = 1 - P(E_z \mid E_{\mathbf{x}}) \tag{2.37}$$

sachant $\sum_{E_Z^*} P(E_Z^* \mid E_{\mathbf{x}}) = 1$. Donc, cette approche bayésienne revient à déterminer l'estimateur *a posteriori* $\hat{E}_z$ qui vérifie :

$$\hat{E}_z = arg\max_{E_z} P(E_z \mid E_{\mathbf{x}}) \tag{2.38}$$

Selon le principe de l'approche bayésienne, le problème qui reste est de choisir les modèles pour $f(E_{\mathbf{x}} \mid E_z)$ et $P(E_z)$. Dans la plupart des cas, on suppose que la distribution $f(E_{\mathbf{x}} \mid E_z)$ est issue d'une famille connue avec l'ensemble des paramètres $\Theta = \{\theta_1, \ldots, \theta_K\}$ inconnu. Ces paramètres peuvent être estimés si les appartenances

d'individus sont déterminées. Quant à la distribution $P(E_z)$, on utilise souvent les champs de Markov qui permettent de décrire la partition d'un point de vue probabiliste en se basant sur la proximité d'individus dans l'espace géographique. Nous introduisons ensuite les champs de Markov et les méthodes qui permettent d'estimer les paramètres.

Champs de Markov

Le modèle du MRF (Markov Random Field) est très utilisé pour modéliser la probabilité $P(E_z)$. La définition du MRF peut s'écrire comme ci-après :

Définition 2.5. Soit $\mathbf{V} = \{V_i\}_{i=1}^N$ un système de voisinage pour un ensemble de N données et $E_Z = \{Z_i\}_{i=1}^N$ un ensemble de variables aléatoires dont les valeurs sont dans l'espace $\mathcal{Z}$. E_Z est un champ de Markov par rapport à $\mathbf{V}$ si :

1. $P(E_Z = E_z) > 0, \forall E_z \in \mathcal{Z}$
2. $P(Z_i = z_i \mid \{Z_j = z_j\}_{j \neq i}) = P(Z_i = z_i \mid \{Z_j = z_j\}_{\mathbf{y}_j \in V_i})$

Dans le contexte du clustering spatial, les données dans la définition peuvent être interprétées par les coordonnées géographiques qui correspond à l'ensemble $E_{\mathbf{y}} = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$ dans ce mémoire. La définition du MRF n'est pas directement applicable en pratique sans le théorème de Hammersley-Clifford qui montre que la distribution du MRF est équivalente à la distribution de Gibbs :

$$G(E_z) = P(E_Z = E_z) = \frac{1}{W} e^{-H(E_z)} \tag{2.39}$$

où $W = \sum_{E_z} e^{-H(E_z)}$ est une constante de normalisation souvent appelée *fonction de partition*, et $H(E_z)$ indique la fonction d'énergie décrite sous la forme :

$$H(E_z) = \sum_{c \in \mathcal{C}} \varphi_c(E_z)$$

où $\mathcal{C}$ représente l'ensemble des cliques. Une clique contient un ensemble de données où deux données quelconques sont adjacentes. $\varphi_c(E_z)$ est la fonction potentielle qui a une valeur réelle non-négative pour chaque clique c. Un exemple de la fonction d'énergie bien connue dans la littérature est le modèle de Strauss [72] défini ci-après :

$$H(\beta, E_z) = -\frac{1}{2}\beta \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{k=1}^{K} c_{ik} c_{jk} v_{ij} \tag{2.40}$$

avec $c_{ik} = 1$ si la $i^{ème}$ donnée relève de la classe k et 0 sinon, et v_{ij} est défini selon

l'équation suivante :

$$v_{ij} = \begin{cases} 1 & \text{si } \mathbf{y}_j \in V_i \\ 0 & \text{sinon} \end{cases} \quad i, j = 1 \dots N \tag{2.41}$$

$\beta > 0$ est *paramètre de lissage* qui permet de contrôler l'importance de la fonction d'énergie. Si β tend vers 0, $H(\beta, E_z)$ tend vers 0 et $G(E_z)$ tend vers $\frac{1}{W}$.

Méthodes basées sur l'approche probabiliste

Selon l'équation 2.34 et 2.38, les méthodes basées sur l'approche probabiliste permettent de déterminer l'estimateur $\hat{E}_z$ qui maximise la distribution *a posteriori* à partir d'une part, de la probabilité *a priori* $P(E_z)$ qui peut être interprétée comme un modèle de MRF selon l'équation 2.39, et d'autre part, de la distribution conditionnelle $f(E_{\mathbf{x}} \mid E_z)$ dont la famille est supposée connue.

Une méthode, appelée *recuit simulé* est proposée pour construire une succession d'estimateurs $\hat{E}_z$ tendant vers l'estimateur du maximum *a posteriori* [28]. L'idée principale de cette méthode consiste à introduire un paramètre de température T dans le modèle de MRF :

$$\pi(E_z, T) = \frac{1}{W(T)} e^{-H(E_z)/T} \tag{2.42}$$

Quand T est grand, cela revient à lisser la fonction et donc à la rendre plus convexe. A chaque étape de recuit simulé, la décroissance de T assez lente vers 0 conduit à la convergence de la probabilité $P(E_z \mid E_{\mathbf{x}})$ vers le maximum global. Cependant, il demande un temps de calcul énorme. C'est pourquoi une autre méthode, appelée ICM (Iterated Conditional Modes) est proposée [8] pour trouver une solution plus rapidement. Le principe d'ICM peut se décrire comme suit : soit $\hat{E}_z$ l'estimateur actuel de la partition théorique E_z^*, nous cherchons à mettre à jour l'étiquette $\hat{z}_i$ du $i^{ème}$ individu avec toutes les informations disponibles. Donc, il est pertinent de choisir la nouvelle étiquette qui maximise la probabilité de $\mathbf{x}_i$ conditionnellement à l'observation $\mathbf{x}_i$ et aux étiquettes actuelles $\{\hat{z}_j\}_{j \neq i}$. Dès que $\hat{z}_i$ est mis à jour, nous obtenons une nouvelle partition selon laquelle les paramètres θ_k, $(k = 1, \dots, K)$ peuvent être estimés. D'après le théorème de Bayes et la définition des MRF, la nouvelle $\hat{z}_i$ maximise le terme suivant :

$$P(z_i \mid \mathbf{x}_i, \{\hat{z}_j\}_{j \neq i}) \propto f(\mathbf{x}_i \mid z_i; \theta_{z_i}) P(z_i \mid \{\hat{z}_j\}_{\mathbf{y}_j \in V_i}) \tag{2.43}$$

Dès que ce principe est appliqué à chaque individu, un cycle de cette procédure est fini. Il suffit de répéter les cycles jusqu'à la convergence. La méthode ICM a l'avantage

de converger très rapidement. Par contre, elle conduit à un maximum local qui dépend des conditions initiales.

La méthode NEM (Neighborhood EM), proposée dans [2], cherche aussi une solution dans le cadre probabiliste. Cette méthode peut être interprétée comme une méthode MAP et vise à optimiser un critère qui peut être décrit comme la vraisemblance pénalisée. La vraisemblance, présentée dans l'équation 2.20, peut être reformulée de manière équivalente à une fonction $D(\mathbf{c}, \Phi)$ selon [33] :

$$D(\mathbf{c}, \Phi) = \sum_{i=1}^{N}\sum_{i=1}^{K} c_{ik} \log \pi_k f(\mathbf{x}_i \mid \theta_k) - \sum_{i=1}^{N}\sum_{i=1}^{K} c_{ik} \log c_{ik} \tag{2.44}$$

où $\mathbf{c}$ est la matrice des probabilités *a posteriori* présentée comme suit :

$$\mathbf{c} = \{c_{ik}\},\ (0 \leqslant c_{ik} \leqslant 1,\ \sum_{k=1}^{K} c_{ik} = 1) \tag{2.45}$$

avec chaque élément c_{ik} représentant la probabilité *a posteriori* décrite dans l'équation 2.23.

En outre, le terme de la pénalisation tient compte de la proximité spatiale et favorise les classes homogènes dans l'espace $\mathcal{Y}$. Il est donné comme suit :

$$G(\mathbf{c}) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{K} c_{ik} c_{jk} v_{ij} \tag{2.46}$$

avec v_{ij} défini dans l'équation 2.41.

Le critère à optimiser $U(\mathbf{c}, \Phi)$ combine le terme $D(\mathbf{c}, \Phi)$ et $G(\mathbf{c})$ par un coefficient β :

$$U(\mathbf{c}, \Phi) = D(\mathbf{c}, \Phi) + \beta G(\mathbf{c}) \tag{2.47}$$

La méthode EM présentée dans la section 2.4.2 est utilisée pour optimiser le critère 2.47. La convergence de cette méthode est prouvée dans [3] par rapport à la valeur de β. En effet, la méthode converge rapidement vers un optimum local si $\beta < \frac{1}{n_V^{max}}$ où $n_V^{max} = \max_i \sum_j v_{ij}$ représente le nombre maximal des voisins d'un individu dans l'espace $\mathcal{Y}$. Le résultat de la partition peut être obtenue à la convergence de la méthode en choisissant la classe qui maximise c_{ik} pour chaque individu.

Bilan

L'approche probabiliste du clustering spatial est basé sur le principe bayésien qui vise à déterminer l'estimateur MAP des labels des individus. Cela est équivalent à minimiser la fonction de coût 0-1 qui vaut 0 pour une bonne décision et 1 pour une mauvaise décision. Afin d'effectuer l'estimation MAP, il est nécessaire de connaître la probabilité *a priori* des labels $P(E_z)$ qui peut être modélisée par des champs de Markov. Ce dernier montre la relation entre les coordonnées géographiques et la partition. Trois méthodes probabilistes ont été brièvement présentées. La méthode de recuit simulé permet d'obtenir un estimateur MAP globalement optimal si le paramètre T suit une loi de décroissance pertinente. Le problème de cette méthode est que le temps de calcul est souvent très long. En pratique, deux autres méthodes ICM et NEM sont souvent utilisées. Ces deux méthodes permettent de trouver une solution d'estimateur MAP rapidement, mais cette solution correspond à un optimum local.

2.6 Processus de dégradation

Cette section introduit le processus de dégradation correspondant à la problématique présentée dans le chapitre précédent. Nous introduisons brièvement les processus stochastiques qui sont généralement utilisés pour modéliser une dégradation. Puis, nous présentons en détail le processus Gamma qui est fréquemment utilisé en sûreté de fonctionnement [77].

2.6.1 Processus stochastiques comme modèles de dégradation

Considérons qu'un système possède plusieurs états : l'état de marche, l'état de panne et des états intermédiaires. Il s'agit d'une description de l'évolution du système de l'état neuf à l'état défaillant. On représente les états par *niveaux de dégradation* dont l'évolution est décrite par un processus stochastique appelé *processus de dégradation*.

L'intérêt du processus de dégradation est qu'il peut être lié à des données de surveillance, e.g. le longueur des fissures d'un bâtiment, ou l'épaisseur du métal en cas de corrosion. Surtout, le développement des techniques de mesure permet de surveiller, de plus en plus précisément des caractéristiques du système qui peuvent être associées au degré de dégradation.

La première difficulté est de choisir le processus de dégradation pertinent qui peut s'adapter au problème étudié. Il s'agit donc de trouver un processus stochastique qui

décrit au mieux l'évolution de la dégradation. De nombreux modèles mathématiques sont utilisés parmi lesquels on note principalement dans la littérature le processus de Lévy [7, 69, 86]. Ce dernier est le processus ayant la propriété markovienne (indépendance de passages entre les états, absence de mémoire, etc.). Précisément, un processus $\{X(t),\ t \geqslant 0\}$ est un process de Lévy s'il vérifie :

1. $X(0) = 0$.
2. Indépendance des incréments : pour tout i, tel que $0 < i \leqslant N$, les $X(t_i) - X(t_{i-1})$ sont indépendants les uns des autres.
3. Continuité stochastique :
$$\lim_{h\to 0} P(\parallel X(t+h) - X(t) \parallel \geq \epsilon) = 0 \quad \forall \epsilon > 0.$$

On parle de processus de Lévy 'homogène' si les incréments du processus sont stationnaires, et 'non-homogène' dans le cas contraire [5].

La famille des processus de Lévy comporte principalement deux types de processus : les processus de (diffusion de) Wiener (*Wiener diffusion process*) et les processus Gamma.

Un processus de Wiener ou un mouvement Brownien avec dérive (*Brownian motion with drift*) est un processus de Lévy à trajectoire continue dont les incréments suivent des lois normales. Il est utilisé pour la modélisation de la dégradation dans plusieurs travaux [14, 81, 82]. Avec ce modèle, l'accroissement de la dégradation a une probabilité non nulle d'être négatif. Il ne permet donc pas de modéliser facilement des dégradations monotones de type propagation de fissures ou perte de matière [9].

Contrairement au processus de Wiener, le processus Gamma est un processus à accroissements positifs, il est donc adapté à la modélisation de dégradation continue monotone et il est largement utilisé dans divers cas comme le fluage du béton [12], la propagation des fissures [48] et la corrosion de matériaux [26]. Nous nous intéressons au processus Gamma qui est présenté dans la section suivante.

2.6.2 Définition du processus Gamma

Le processus Gamma est défini mathématiquement comme ci-après [77] :

Définition 2.6. Soit $A(t)$ une fonction non décroissante avec $t \geqslant 0$ et $A(0) \equiv 0$, le processus Gamma avec la fonction de forme $A(t)$ et le paramètre d'échelle $b > 0$ est un processus stochastique en temps continu $\{X(t),\ t \geqslant 0\}$ qui vérifie :

1. $X(0) = 0$.
2. $\{X(t),\ t \geqslant 0\}$ est un processus stochastique à incréments indépendants.

3. L'incrément $X(t) - X(s)$ suit une loi Gamma non décroissante $\mathrm{Ga}(A(t) - A(s),\, b)$ pour $0 \leqslant s < t$.

   Une variable aléatoire X qui suit une distribution Gamma $\mathrm{Ga}(a, b)$ a une densité de probabilité de la forme :

$$f_{a,b}(x) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx} \mathbf{I}_{(0,\,\infty)}(x) \tag{2.48}$$

   avec le paramètre de forme $a > 0$ et le paramètre d'échelle $b > 0$. $\mathbf{I}_{(0,\,\infty)}(x)$ représente la fonction indicatrice et $\Gamma(.)$ est la fonction Gamma.

4. $\forall t \geqslant 0$, l'espérance et la variance de $X(t)$ correspondent aux équations ci-après :

$$E(X(t)) = \frac{A(t)}{b} \quad \mathrm{Var}(X(t)) = \frac{A(t)}{b^2} \tag{2.49}$$

Quand la fonction de forme est linéaire : $A(t) = at$, $(a > 0)$, on obtient un processus Gamma homogène. Dans ce cas, $X(t) - X(s)$ suit la distribution $\mathrm{Ga}(a(t-s),\, b)$ et les incréments du processus ne dépendent que des incréments de temps $t - s$. Un exemple de trois trajectoires d'un processus Gamma homogène est illustré par la figure 2.11 où $a = 8$ et $b = 2$.

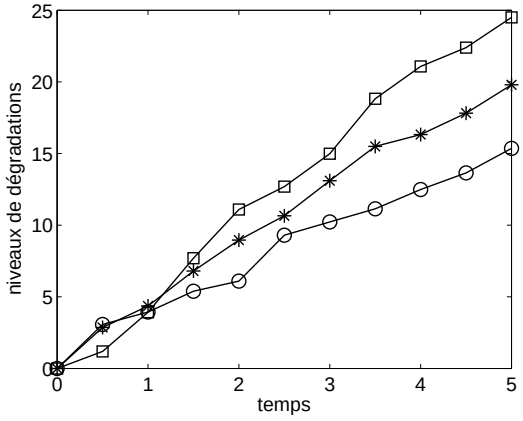


Figure 2.11 – exemple de trois trajectoires d'un processus Gamma homogène avec $a = 8$, $b = 2$

Pour un process non homogène, nous utilisons souvent une approximation de la fonction de forme : $A(t) = at^u$, $(a > 0, u > 0)$. Il est indiqué dans [20] que les différentes valeurs de u correspondent aux différents cas de dégradation (e.g. $u = 1$ pour la corrosion du béton et $u = 2$ pour l'attaque du sulfate).

2.6.3 Estimation des paramètres d'un processus Gamma

Il existe plusieurs méthodes pour estimer les paramètres d'un processus Gamma parmi lesquelles la méthode de maximum de vraisemblance (ML) et la méthode de moments sont les plus utilisées [63].

Méthode du maximum de vraisemblance

Étant donné les valeurs stochastiques x_i d'un processus Gamma aux instants t_i, les incréments du processus peuvent être décrits comme : $\Delta x_i = x_i - x_{i-1}$ $(i = 1, \ldots, N)$. Nous pouvons donner la fonction de vraisemblance par rapport à Δx_i grâce à la propriété d'indépendance :

$$l = \prod_{i=1}^{N} f(\Delta x_i \mid A(t_i) - A(t_{i-1}),\, b) \tag{2.50}$$

La fonction de log-vraisemblance de $L_{a,b}$ prend la forme :

$$\begin{aligned} L_{a,b} &= \log l(a, b \mid \Delta x_1, \ldots, \Delta x_N, t_0, \ldots, t_N, u) \\ &= \sum_{i=1}^{N} \{a(t_i^u - t_{i-1}^u)\log(b) - \log\Gamma[a(t_i^u - t_{i-1}^u)] \\ &\quad + [a(t_i^u - t_{i-1}^u) - 1]\log(\Delta x_i) - b\Delta x_i\} \end{aligned} \tag{2.51}$$

Pour le cas d'un processus homogène, l'équation 2.51 devient :

$$L_{a,b} = \sum_{i=1}^{N} (a\triangle t_i \log(b) - \log\Gamma(a\triangle t_i) + [(a\triangle t_i) - 1]\log(\Delta x_i) - b\Delta x_i) \tag{2.52}$$

Donc, la dérivée partielle de la vraisemblance $L_{a,b}$ par rapport à a et b est donnée par :

$$\begin{cases} \frac{\partial}{\partial a} L_{a,b} = \sum_{i=1}^{N} \Delta t_i \log(b) - \Delta t_i \psi(a\Delta t_i) + \Delta t_i \log(\Delta x_i) \\ \\ \frac{\partial}{\partial b} L_{a,b} = \sum_{i=1}^{N} (\frac{a\Delta t_i}{b} - \Delta x_i) \end{cases} \tag{2.53}$$

où $\psi(x) = \frac{\Gamma'(x)}{\Gamma(x)}$. Les paramètres optimaux s'obtiennent lorsque ces dérivées partielles s'annulent. On obtient alors :

$$\hat{a} = \hat{b}\frac{\sum_{i=1}^{N} \Delta x_i}{\sum_{i=1}^{N} \Delta t_i} \tag{2.54}$$

et

$$\sum_{i=1}^{N} \Delta t_i \log\left(\hat{a}\frac{\sum_{j=1}^{N}\Delta t_j}{\sum_{j=1}^{N}\Delta x_j}\right) + \sum_{i=1}^{N}\Delta t_i(\log(\Delta x_i) - \psi(\hat{a}\Delta t_i)) = 0 \tag{2.55}$$

Il est montré dans [12] que les estimateurs de maximum de vraisemblance $\hat{a}$ et $\hat{b}$ peuvent être trouvés en résolvant les équations 2.54 et 2.55.

Méthode des moments

Cette méthode nécessite une transformation de variable selon l'approximation de $A(t) : w_i = t_i^u - t_{i-1}^u$. Par conséquent, l'incrément Δx_i suit la distribution Gamma avec le paramètre de forme aw_i et le paramètre d'échelle b. Il est précisé dans [12] que les estimateurs de moments $\hat{a}$ et $\hat{b}$ pour le cas non homogène sont calculés en résolvant les équations suivantes :

$$\frac{\hat{a}}{\hat{b}} = \frac{\sum_{i=1}^{N}\Delta x_i}{\sum_{i=1}^{N} w_i} = \eta \tag{2.56}$$

$$\frac{\sum_{i=1}^{N}\Delta x_i}{\hat{b}}\left(1 - \frac{\sum_{i=1}^{N} w_i^2}{[\sum_{i=1}^{N} w_i]^2}\right) = \sum_{i=1}^{N}(\Delta x_i - \eta w_i)^2 \tag{2.57}$$

L'équation (2.57) est souvent reformulée pour obtenir un estimateur de la variance :

$$\frac{\hat{a}}{\hat{b^2}} = \frac{\sum_{i=1}^{N}(\Delta x_i - w_i\frac{\sum_{j=1}^{N}\Delta x_j}{\sum_{j=1}^{N} w_j})^2}{\sum_{i=1}^{N} w_i - \frac{1}{\sum_{i=1}^{N} w_i}\sum_{i=1}^{N}(w_i)^2} \tag{2.58}$$

Pour le cas homogène, les estimateurs peuvent être directement calculés selon les équations 2.56 et 2.58 en mettant $w_i = \Delta t_i$.

2.7 Conclusion

Suite au chapitre précédent qui a indiqué que le problème considéré est un problème de clustering des processus de dégradation, ce chapitre est principalement consacré à la présentation des méthodes de clustering et à l'introduction de processus de dégradation. Les méthodes de clustering sont classifiées en deux catégories selon le type de données à regrouper. La première catégorie contient les méthodes avec un seul type d'information, appelé aussi attribut dans ce mémoire, alors que la deuxième catégorie contient des méthodes avec deux types d'information : attribut et covariable.

Nous avons présenté les méthodes de clustering en fonction de ces deux catégories. Pour la première catégorie, nous avons présenté 5 types de méthodes parmi lesquelles les méthodes hiérarchiques, les méthodes de partitionnement et les méthodes à noyau qui se basent directement sur la mesure de distance. Les méthodes hiérarchiques sont souvent utilisées lorsque le nombre d'individus à regrouper est petit, car ces méthodes construisent une hiérarchie de la partition et le temps de calcul croit rapidement avec le nombre d'observations. En revanche, on utilise souvent les méthodes de partitionnement si on connaît *a priori* le nombre de clusters. Les méthodes à noyau notamment permettent d'améliorer la séparabilité des individus en les transformant dans un espace de redescription qui est souvent de très grande dimension. En outre, les méthodes basées sur la densité et les méthodes basées sur les graphes s'appuient sur un système de voisinage qui définit une adjacence entre les observations. En général, la connaissance *a priori* du nombre de clusters n'est pas nécessaire avec ces deux types de méthodes, mais les clusters trouvés dépendent du choix d'autres paramètres.

Les méthodes de clustering de la seconde catégorie peuvent être traitées à l'aide d'approches probabilistes ou non probabilistes. Les premières méthodes s'adaptent aux cas où on a aucune information *a priori* sur les distributions ou des grands volumes de données. Dans ce cas, on tient compte conjointement de la proximité dans l'espace $\mathcal{X}$ et dans l'espace $\mathcal{Y}$ dans le cadre non probabiliste. En revanche, les méthodes basées sur l'approche probabiliste traitent les individus dont les observations et les coordonnées suivent respectivement certaines lois probabilistes. Donc, le problème peut se reformulé comme un problème d'estimation des paramètres des lois.

Notre problématique concerne des données de dégradation qui sont caractérisées par d'une part, les observations qui suivent certains processus stochastiques de dégradation, et d'autre part, les valeurs de covariable qui influent le processus de dégradation. Il est considéré que les classes sont parfaitement séparables selon les valeurs de covariable, mais que celles ci ne permettent pas de déterminer la frontière. En revanche, les observations de dégradation permettent de discriminer les classes mais dans cet espace les classes ne sont pas parfaitement séparables. La problématique correspond au clustering de la seconde catégorie. Le clustering spatial correspond alors partiellement à notre problématique, car les coordonnées géographiques sont généralement bi-dimensionnelles, tandis que la dimension de la covariable peut être quelconque. Les méthodes proposées dans les chapitres suivants sont inspirées des méthodes de clustering spatial.

Chapitre 3

Présentation du problème général et d'une solution dans un cas simple

3.1 Introduction

Ce chapitre vise à donner une formulation détaillée du problème et proposer une première solution qui correspond à un cas particulier. Dans la section 3.2, nous décrivons la formulation du problème avec les notations indiquées dans les chapitres précédents. Dans la section 3.3, le problème est traité par une étude dans un cas simple. Nous présentons cette étude en deux parties : la proposition de la méthode MLC, et les analyses de sa performance. Nous concluons ce chapitre dans la section 3.4.

3.2 Présentation du problème

La compréhension et le positionnement du problème ont constitué une étape fondamentale dans ce travail de thèse. Nous présentons les caractéristiques générales dans la section 3.2.1. La formulation du problème traité est donné dans la section 3.2.2 et nous précisons l'évaluation de la performance d'une solution dans la section 3.2.3.

3.2.1 Caractéristiques générales

La section 2.2.4 montre brièvement que notre problème entre dans le cadre du clustering avec covariable, car chaque individu est caractérisé par une observation et une valeur de covariable. L'étude de l'état de l'art dans le chapitre 2 montre que les méthodes de clustering classiques ne sont pas adaptées à ce problème. Les méthodes de clustering spatial, pour leur part, correspondent partiellement au problème. Par conséquent, il est nécessaire de développer des nouvelles méthodes qui s'adaptent mieux au problème en tenant compte des trois particularités suivantes :

- La première particularité est que les observations dont nous disposons sont celles qui caractérisent un processus de dégradation. Une observation est considérée comme un incrément de niveau de dégradation qui dépend d'un incrément du temps. Quand les incréments ne sont pas constants, toutes les observations ne suivent pas la même loi, et la similarité d'attribut ne peut pas être mesurée directement par leur valeur. C'est pourquoi les méthodes classiques comme K-means ne sont pas applicables dans ce cas.
- Chaque trajectoire est présentée par une ou plusieurs observations qui partagent la même valeur de covariable. S'il existe plusieurs observations par trajectoire, nous avons alors une connaissance *a priori* du fait que ces observations appartiennent au même processus. Cette connaissance permet d'ajouter une information complémentaire au clustering.
- Dans cette étude, la covariable est supposée être un facteur influant sur le processus de dégradation, e.g. la température, l'humidité, la pression. En général, la covariable est décrite par un vecteur avec la dimension $q \geqslant 1$. Un vecteur de covariable correspond à un système dont la trajectoire de dégradation peut contenir plusieurs observations. De plus, les trajectoires avec les vecteurs de covariables similaires ont plus de chance d'avoir le même vecteur de paramètres et donc d'appartenir au même processus de dégradation. Remarquons que dans le contexte du clustering spatial rappelé dans la section 2.5, nous avons la même contrainte qu'avec des coordonnées géographiques. Cependant, cette contrainte du clustering spatial ne correspond pas complètement au problème dans notre cas, car les coordonnées géographiques sont généralement bi-dimensionnelles, tandis que la dimension de la covariable peut être quelconque.

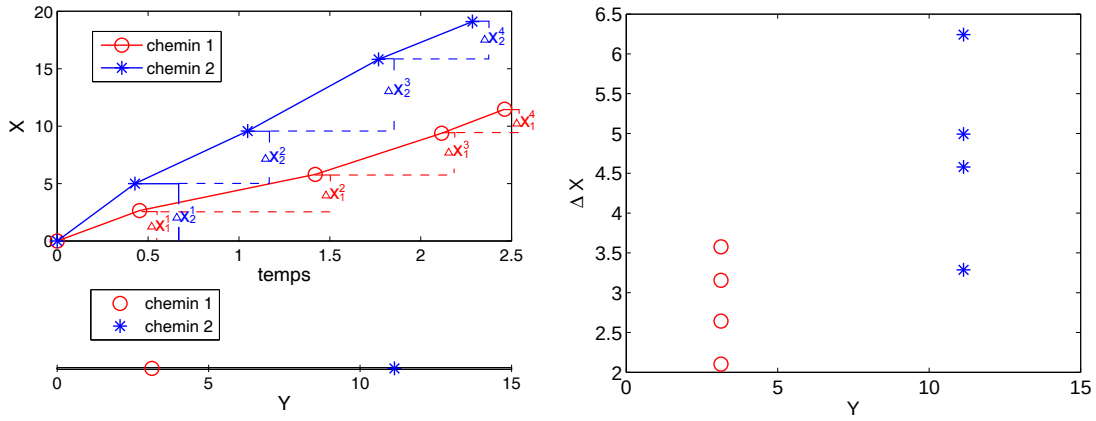
Nous listons des cas particuliers du problème traité afin de montrer les trois particularités ainsi leur relation avec les méthodes de clustering.

- Δt **constant, 1 observation par processus, pas de covariable**. Le problème revient au problème de clustering classique dans l'espace d'attribut $\mathcal{X}$. Les méthodes classiques comme K-means peuvent être utilisées dans ce cas.
- Δt **constant, 1 observation par processus, avec une covariable bi-dimensionnelle**. Le problème est équivalent au problème de clustering spatial. Les méthodes de clustering spatial comme NEM et ICM peuvent être appliquées dans ce cas.
- Δt **constant, plusieurs observations par processus, pas de covariable**. Dans ce cas, le problème est un problème de clustering avec contraintes qui imposent que certaines observations appartiennent au même groupe. Les méthodes décrites dans [67, 47] peuvent être utilisées.

3.2.2 Formulation

Afin de faciliter les études, nous donnons une formulation du problème dans cette section qui vise à présenter le problème sous forme analytique.

L'ensemble d'individus $\mathbf{L}$ est supposé provenir de N trajectoires. Chaque trajectoire p_i, $(i = 1, \ldots, N)$ est associée à un vecteur de covariable $\mathbf{y}_i = (y_{i1}, \ldots, y_{iq}) \in \Omega_y \subset \mathcal{R}^q$ et est composé d'une ou plusieurs observations. Le $j^{ème}$ individu de la trajectoire p_i est caractérisé par l'instant t_i^j et le niveau de dégradation $x_i^j = X(t_i^j) \in \Omega_x \subset \mathcal{R}$. Dans le cas du processus Gamma, nous utilisons les incréments des niveaux de dégradations, car ils sont indépendants les uns des autres selon la définition du processus Gamma dans la section 2.6.2. En outre, pour un processus Gamma homogène ils sont indépendants de la position temporelle. Donc chaque individu de la réalisation p_i d'un processus est représenté par $l_i^j = (\Delta t_i^j, \Delta x_i^j)$, $(j = 1, \ldots, \mid p_i \mid)$ où $\Delta t_i^j = t_i^j - t_i^{j-1}$, $\Delta x_i^j = x_i^j - x_i^{j-1}$, $\mid p_i \mid$ est le nombre d'observations de la trajectoire p_i. Par convention on choisit $t_i^0 = 0$, $x_i^0 = 0$. La figure 3.1a donne un exemple d'évolution d'une dégradation en fonction du temps et séparément la valeur de la covariable. La figure 3.1b illustre les observations dans les espaces conjoints de la covariable et de l'incrément de dégradation.



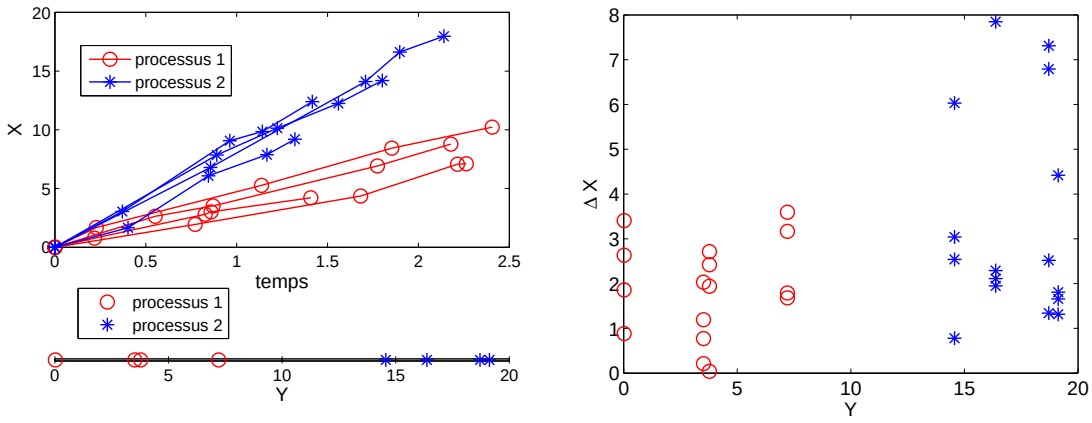
(a) exemple de 2 trajectoires avec 4 observations par trajectoire et les valeurs de covariable correspondantes

(b) les observations dans les 2 espaces conjoints

Figure 3.1 – description de l'attribut et de la covariable

Nous supposons que les individus peuvent être divisés en K classes de processus $C_1, \ldots, C_K$. Dans ce travail il est supposé que K est une connaissance *a priori*. Chaque classe est caractérisée par un vecteur de paramètre θ_k, $(k = 1, \ldots, K)$ qui dépend de la covariable. Le vecteur θ_k est défini par 2 paramètres du processus Gamma, qui sont le paramètre de forme a_k et le paramètre d'échelle b_k. Le label inconnu de chaque trajectoire est notée z_i où $z_i = k$ signifie que la réalisation du processus p_i provient de la classe k, et tous les individus de p_i appartiennent à la classe k. L'ensemble

d'appartenance $E_z = \{z_1, \ldots, z_N\}$ définit une partition des réalisations de processus. Les partitions E_z^* et $\hat{E}_z$ sont utilisées respectivement pour décrire la partition théorique et estimée. Par ailleurs, afin de décrire la dépendance des vecteurs de paramètres par rapport aux appartenances des trajectoires, nous écrivons la densité de probabilité de chaque incrément comme $f(\Delta x_i^j \mid z_i; \theta_{z_i})$. Les figure 3.2a et 3.2b correspondent à un exemple de 2 processus avec 4 trajectoires par processus. Donc, l'objectif est de déterminer les appartenances inconnues pour chaque trajectoire et les vecteurs de paramètres qui caractérisent les processus.



(a) exemple de 2 processus avec 4 trajectoires par processus.

(b) les observations dans les 2 espaces conjoints

Figure 3.2 – exemple de problème avec 2 processus et la covariable en dimension 1

3.2.3 Évaluation de la performance d'une solution

La performance d'une solution dépend bien sûr de l'importance donnée à l'attribut et à la covariable.

Lorsque le résultat théorique est inconnu, un critère combinant l'adéquation du modèle Gamma à l'attribut et la cohérence dans l'espace de covariable est proposé à la section 4.4. Remarquons qu'il n'existe pas de critère dans la littérature dans ce cas.

Lorsque le résultat théorique est connu, le taux d'erreur de classification peut être calculé. Toutefois il est important de remarquer que plus les classes ont des paramètres similaires, moins les erreurs ont d'importance. En effet, dans le cas de classes similaires avec des erreurs de classification, les erreurs d'estimation de paramètres sont faibles car les observations mal classées ressemblent aux observations bien classées. En outre, dans le cas de processus Gamma assez similaires, les prédictions d'évolution sont peu différentes. En revanche, dans le cas de processus Gamma distincts, des erreurs d'esti-

mation de paramètres ont un impact beaucoup plus important sur une utilisation des modèles pour faire de la prédiction. En conséquence, lorsque les processus sont peu distincts, il y a un plus grand risque d'erreur mais les erreurs sont moins critiques.

Les méthodes proposées ont donc été évaluées en calculant le taux d'erreur de classification mais aussi les valeurs des paramètres estimées, car ce sont eux qui caractérisent le modèle utilisé pour la prédiction.

Il est à noter que les paramètres a et b d'un processus Gamma sont difficilement interprétables. En revanche, la moyenne et la variance sur un temps unitaire sont plus explicites. Elles sont définies comme suit :

$$\theta_k = \begin{cases} m_k = \frac{a_k}{b_k} \\ \sigma_k^2 = \frac{a_k}{b_k^2} \end{cases} \qquad k = 1, \ldots, K \tag{3.1}$$

Il peut exister une erreur sur a et b mais la moyenne peut être correcte. Dans la prédiction, c'est la moyenne, et en moindre mesure, qui ont de l'importance.

3.3 Étude avec deux classes uni-modales et une covariable uni-dimensionnelle

Dans un premier temps, nous abordons l'étude par un cas particulier où $K = 2$, $q = 1$ et $\mid p_1 \mid = \cdots = \mid p_N \mid = 1$. De plus, les deux classes sont supposées uni-modales et séparées par un seuil inconnu s^*, tel que :

$$z_i = \begin{cases} 1 & \text{si } y_i \leqslant s^* \\ 2 & \text{si } y_i > s^* \end{cases} \quad i = 1, \ldots, N \tag{3.2}$$

Le vecteur de paramètre qui correspond à chaque classe de processus Gamma homogène est présenté par : $\theta_1 = (m_1, \sigma_1^2), \theta_2 = (m_2, \sigma_2^2)$. En partant de ce cas particulier, nous proposons une méthode que nous décrivons dans la section 3.3.1. Les études expérimentales sont présentées dans la section 3.3.2 avec des données simulées.

3.3.1 Méthode de maximum de vraisemblance pour clustering (MLC)

La méthode MLC a été proposée pour traiter ce cas particulier. L'idée principale de cette méthode est de partitionner les individus selon la valeur de la covariable en

choisissant la partition qui maximise la vraisemblance. Tout d'abord, nous trions par ordre croissant tous les individus en fonction de leur valeur de covariable. Les individus ordonnés sont dénotés comme $l_{(1)}, l_{(2)}, \ldots, l_{(N)}$ avec $y_{(1)} < y_{(2)} \cdots < y_{(N)}$. Ensuite, une succession de partitions est définie selon l'indice $2 \leqslant r \leqslant N-1$, de telle sorte que $\{l_{(1)}, \ldots, l_{(r)}\}$ forment la classe 1 et $\{l_{(r+1)}, \ldots, l_{(N)}\}$ forment la classe 2. En se basant sur chaque partition, on calcule les estimations des paramètres $\hat{\theta}_1^r, \hat{\theta}_2^r$ selon lesquels les log-vraisemblances $\hat{L}_1^r, \hat{L}_2^r$ peuvent être obtenues. Nous déterminons la vraisemblance totale $\hat{L}^r = \hat{L}_1^r + \hat{L}_2^r$ et la valeur r^* qui maximise $\hat{L}^r$ correspondant à la partition optimale. Nous définissons $2 \leqslant r \leqslant N-1$, car il est nécessaire d'avoir au moins deux observations pour l'estimation des paramètres θ_{z_i}.

La figure 3.3 illustre l'idée principale de cette méthode qui est développée en pseudo code dans l'Algorithme 1.

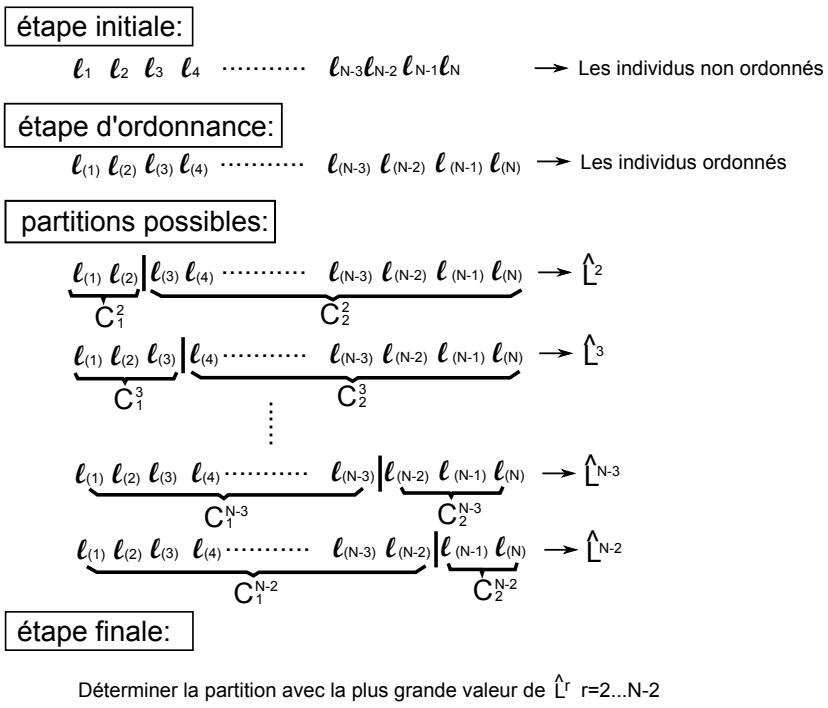


Figure 3.3 – explication de la méthode MLC

3.3.2 Études expérimentales

Application de la méthode MLC sur un ensemble d'individus

La figure 3.4 montre un exemple de la partition de référence par rapport à la covariable. Le nombre d'individus a été fixé à $N = 100$. De plus, nous avons défini que les valeurs de covariable sont distribuées uniformément dans l'intervalle $[0, 20]$. Les deux

Algorithme 1 Maximum Likelihood Clustering

1: Trier les individus selon les valeurs de covariable $\{y_i\}_{i=1}^{N}$ et déterminer $l_{(1)}, l_{(2)}, \ldots, l_{(N)}$.
2: **pour** $r = 2$ à $N - 2$ **faire**
3: Partitionner les individus en deux classes : $C_1^r = \{l_1, \ldots, l_r\}$ et $C_2^r = \{l_{(r+1)}, \ldots, l_{(N)}\}$.
4: Estimer les paramètres $\hat{\theta}_1^r, \hat{\theta}_2^r$ et calculer les log-vraisemblances $\hat{L}_1^r, \hat{L}_2^r$.
5: Calculer la log-vraisemblance totale $\hat{L}^r = \hat{L}_1^r + \hat{L}_2^r$.
6: **fin pour**
7: Determiner la partition optimale qui correspond à la valeur maximale des $\hat{L}^r$.

classes sont séparées par le seuil $s^* = 10$. Cette définition de la partition présentée dans l'équation 3.3 donne une égalité de la probabilité *a priori* d'appartenance de chaque individu.

$$z_i = \begin{cases} 1 & \text{si } 0 < y_i \leqslant s^* \\ 2 & \text{si } s^* < y_i < 20 \end{cases} \quad i = 1, \ldots, N \tag{3.3}$$

Par ailleurs, les vecteurs de paramètres avec la moyenne et la variance de chaque classe ont été définis comme suit :

$$\theta_1 = \begin{cases} m_1 = 4 \\ \sigma_1^2 = 2 \end{cases}, \theta_2 = \begin{cases} m_2 = 6 \\ \sigma_2^2 = 2 \end{cases} \tag{3.4}$$

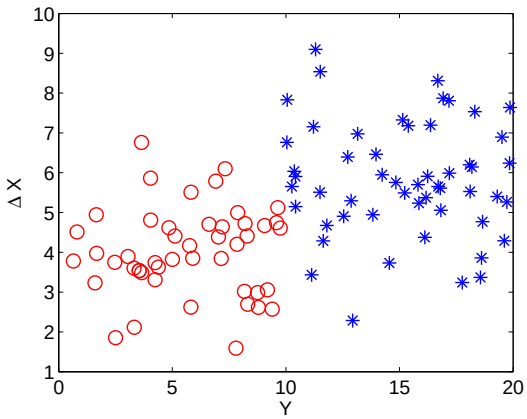
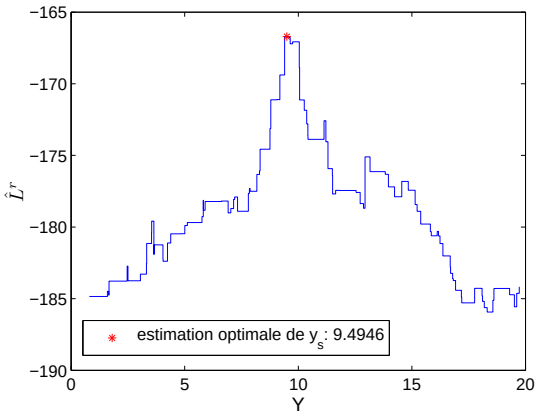
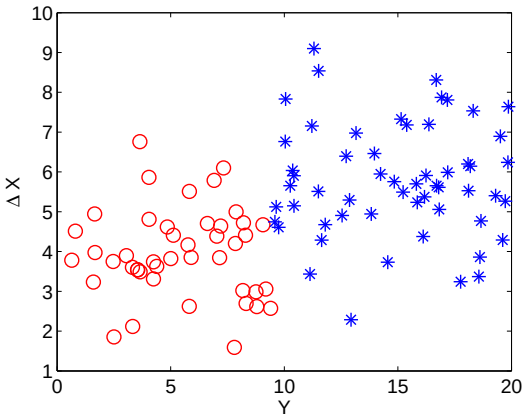


Figure 3.4 – ensemble d'individus et la partition théorique selon la covariable

La méthode MLC a été appliquée sur cet exemple simulé, et les valeurs du $\hat{L}^r$ ont été calculées pour chaque valeur de r qui correspond à un intervalle de seuil $s \in \left[y_{(r)}, y_{(r+1)}\right]$. Le résultat est illustré dans la figure 3.5 où la figure 3.5a montre l'évolution de la valeur $\hat{L}^r$, et la figure 3.5b montre le résultat de la partition obtenue.



(a) vraisemblance estimée en fonction du seuil s dépendant des valeurs de r



(b) partition trouvée avec l'algorithme MLC sur l'exemple

Figure 3.5 – application de la méthode MLC sur un ensemble d'individus

La valeur maximale de $\hat{L}^r$ se trouve à l'endroit où $y_{r^*} = 9.49$. Cela veut dire que parmi toutes les partitions possibles selon la covariable, la partition avec $y_{r^*} = 9.49$ correspond à la meilleure adéquation des données aux modèles de chaque classe. De plus, la partition théorique est retrouvée avec un taux d'individus mal classés égal à 0.03.

Analyses statistiques

Les analyses statistiques réalisées concernent deux aspects : l'influence du nombre d'individus et l'influence de la dissimilarité entre classes. Une première analyse a été effectuée en choisissant le nombre d'individus $N = 50, 100, 150, 200, 250, 300$. Pour chaque valeur de N, 200 expériences ont été réalisées et la méthode MLC a été appliquée sur chaque expérience. 200 partitions optimales avec le taux d'individus mal classés ont été estimées pour chaque valeur de N. Les résultats sont présentés par la figure 3.6.

Pour chaque valeur de N, une boîte à moustache présente la moyenne et les quartiles des taux d'erreurs. La marque centrale et l'astérisque représentent respectivement la valeur médianne et la valeur moyenne, les deux bords de la boîte correspondent aux quartiles, et la ligne en pointillé étend à la valeur extrême sans tenir compte des valeurs atypiques.

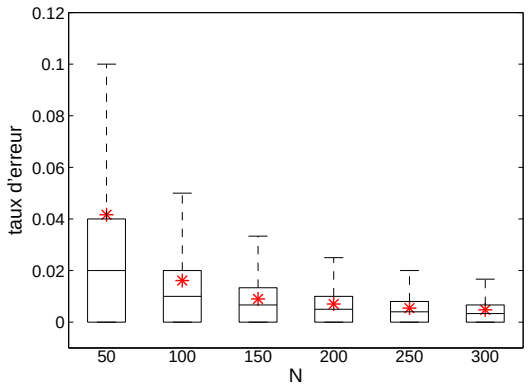


Figure 3.6 – taux d'erreur des partitions optimales estimés en fonction du nombre d'individus

Les paramètres des lois ont été estimés comparés avec les vrais paramètres pour chaque valeur de N. Les résultats sont rassemblés dans le tableau 3.1. $\hat{\mathrm{E}}\{\bar{m}\}$ et $\hat{\sigma}\{\bar{m}\}$ (resp. $\hat{\mathrm{E}}\{\bar{\sigma}^2\}$ et $\hat{\sigma}\{\bar{\sigma}^2\}$) représentent les estimations de l'espérance et de l'écart-type des moyennes (resp. variances) de 200 simulations avec les partitions théoriques. Aussi, $\hat{\mathrm{E}}\{\hat{m}\}$ et $\hat{\sigma}\{\hat{m}\}$ (resp. $\hat{\mathrm{E}}\{\hat{\sigma}^2\}$ et $\hat{\sigma}\{\hat{\sigma}^2\}$) représentent les estimations de l'espérance et de l'écart-type des moyennes (resp. variances) de 200 simulations avec les partitions obtenues par la méthode MLC.

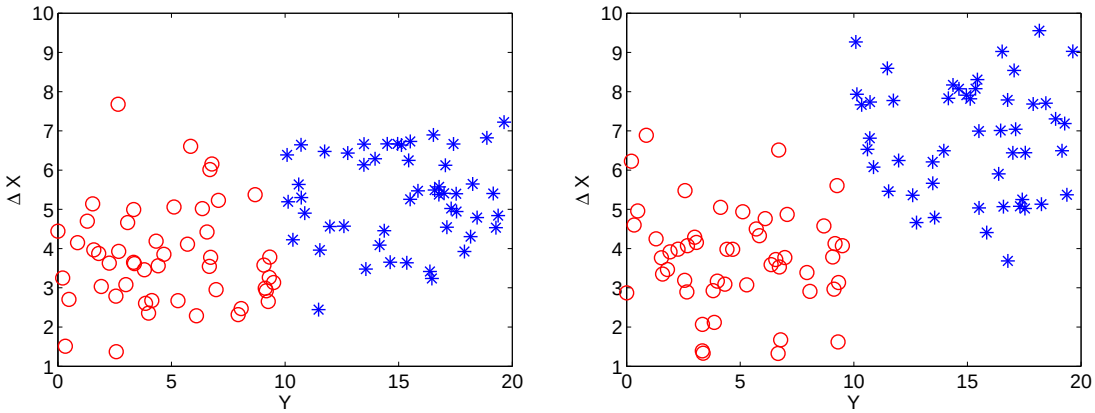
Selon les résultats du tableau 3.1, les paramètres sont mieux estimés quand N change de 50 à 150. Cependant, les résultats d'estimation sont quasiment constants à partir du $N = 150$.

L'effet de la dissimilarité entre classes théoriques a aussi été étudié. Une grande dissimilarité implique que les classes sont très distinctes l'une de l'autre, tandis que une petite signifie que les classes sont proches et moins séparables. La dissimilarité entre classes peut être caractérisée par la différence des vecteurs de paramètres $\theta_k = (m_k, \sigma_k^2)$, $(k = 1, \ldots, K)$, car m_k et σ_k^2 représentent respectivement la valeur moyenne et la variance des observations dans la classe k. Dans cette étude, nous avons dénoté Δm et $\Delta\sigma^2$ comme la différence des moyennes et des variances entre deux classes théoriques traitées. Il est évident que les deux différences interviennent pour définir la dissimilarité. Cependant, pour des classes de même variance, seul Δm peut définir la dissimilarité.

Tableau 3.1 – Analyse du nombre d'individus

	$N = 50$		$N = 100$		$N = 150$		$N = 200$		$N = 250$		$N = 300$	
	C_1	C_2	C_1	C_2	C_1	C_2	C_1	C_2	C_1	C_2	C_1	C_2
m	4	6	4	6	4	6	4	6	4	6	4	6
$\hat{\mathrm{E}}\{\bar{m}\}$	4.04	6.02	4.02	6.00	4.01	6.03	3.99	5.98	4.00	6.00	4.01	6.01
$\hat{\mathrm{E}}\{\hat{m}\}$	4.00	6.02	4.00	6.01	3.99	6.03	3.98	5.98	3.99	6.01	4.00	6.02
$\hat{\sigma}\{\bar{m}\}$	0.27	0.32	0.21	0.19	0.16	0.17	0.14	0.14	0.12	0.12	0.11	0.12
$\hat{\sigma}\{\hat{m}\}$	0.32	0.40	0.22	0.20	0.16	0.17	0.14	0.14	0.12	0.12	0.11	0.12
σ^2	2	2	2	2	2	2	2	2	2	2	2	2
$\hat{\mathrm{E}}\{\bar{\sigma}^2\}$	2.03	1.98	2.02	2.07	2.01	2.04	1.99	2.01	2.04	2.01	2.02	2.02
$\hat{\mathrm{E}}\{\hat{\sigma}^2\}$	1.95	1.91	1.99	2.04	1.99	2.02	1.98	1.99	2.02	2.00	2.01	2.01
$\hat{\sigma}\{\bar{\sigma}^2\}$	0.69	0.63	0.45	0.45	0.40	0.34	0.31	0.32	0.30	0.30	0.27	0.24
$\hat{\sigma}\{\hat{\sigma}^2\}$	0.76	0.71	0.46	0.46	0.40	0.34	0.31	0.32	0.30	0.30	0.27	0.24

Donc, afin de simplifier l'étude, nous avons défini les différentes dissimilarités selon les différents choix de Δm en fixant $\Delta\sigma^2 = 0$. La figure 3.7 montre deux exemples avec $\Delta m = 1$ et $\Delta m = 3$. Nous avons défini $\sigma_1^2 = \sigma_2^2 = 2$ pour ces deux exemples.



(a) $\Delta m = 1$ (b) $\Delta m = 3$

Figure 3.7 – deux exemples de dissimilarité par rapport aux valeurs de Δm

La performance de la méthode MLC a été analysée pour les différentes dissimilarités suivantes : $\Delta m = 1, 1.5, 2, 2.5, 3$. Précisément, nous avons défini $m_k = 4 + (k-1)\Delta m$ et $\sigma_k^2 = 2$. Pour chaque valeur de Δm, 200 simulations ont été générées. La figure 3.8 illustre le taux d'individus mal classés selon les différentes valeurs de Δm, et le tableau 3.2 montre les vecteurs de paramètres estimés à partir des partitions trouvées.

Selon la figure 3.8, le taux d'erreur diminue avec la croissance de la dissimilarité entre classes. Ce résultat est conforme à ce qui était attendu, car les classes théoriques sont de plus en plus séparables avec l'augmentation de la dissimilarité. De plus, le

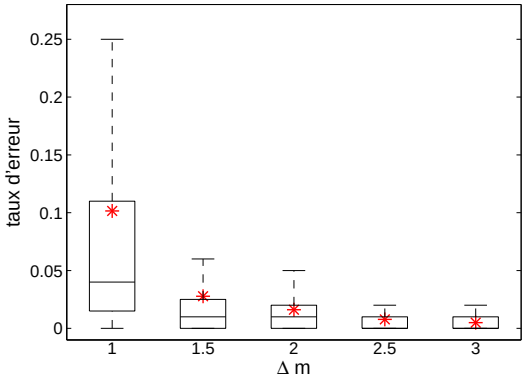


Figure 3.8 – taux d'erreur des partitions optimales estimées par rapport aux dissimilarités entre classes

Tableau 3.2 – Analyse de la dissimilarité entre classes

	$\Delta m = 1$		$\Delta m = 1.5$		$\Delta m = 2$		$\Delta m = 2.5$		$\Delta m = 3$	
	C_1	C_2	C_1	C_2	C_1	C_2	C_1	C_2	C_1	C_2
m	4	5	4	5.5	4	6	4	6.5	4	7
$\hat{\mathrm{E}}\{\bar{m}\}$	4.00	5.01	4.00	5.50	4.02	5.99	4.00	6.48	4.02	7.00
$\hat{\mathrm{E}}\{\hat{m}\}$	3.97	5.09	3.97	5.51	4.00	6.00	3.99	6.48	4.01	7.01
$\hat{\sigma}\{\bar{m}\}$	0.19	0.19	0.20	0.21	0.20	0.20	0.19	0.18	0.21	0.19
$\hat{\sigma}\{\hat{m}\}$	0.30	0.40	0.27	0.22	0.20	0.21	0.20	0.19	0.21	0.19
σ^2	2	2	2	2	2	2	2	2	2	2
$\hat{\mathrm{E}}\{\bar{\sigma}^2\}$	2.01	2.02	2.03	2.01	2.04	2.02	2.06	1.99	2.00	2.01
$\hat{\mathrm{E}}\{\hat{\sigma}^2\}$	1.96	1.90	1.99	1.98	1.99	2.01	2.04	1.98	1.98	2.00
$\hat{\sigma}\{\bar{\sigma}^2\}$	0.50	0.45	0.45	0.41	0.52	0.45	0.52	0.42	0.46	0.45
$\hat{\sigma}\{\hat{\sigma}^2\}$	0.64	0.64	0.48	0.42	0.50	0.47	0.53	0.43	0.45	0.46

tableau 3.2 montre que les paramètres sont mieux estimés quand Δm change de 1 à 1.5, et quasiment constant à partir du $\Delta m = 1.5$.

3.4 Conclusion

Ce chapitre présente le problème envisagé au cours de notre travail en décrivant trois particularités par rapport à l'étude de l'état de l'art dans le chapitre précédent. La première particularité est que les observations sont les incréments des niveaux de dégradation qui dépendent des incréments du temps. Cela veut dire que la similarité ne peut pas être mesurée directement par les valeurs d'attribut quand les incréments ne sont pas constants. La deuxième est que chaque trajectoire de dégradation peut

contenir plusieurs observations, qui impose une connaissance *a priori* du clustering. La troisième particularité est que chaque trajectoire est associée à une valeur de covariable, de sorte que les trajectoires avec les valeurs de covariable similaires ont plus de chance d'appartenir au même processus de dégradation. Selon cette description du problème, nous avons présenté la formulation du problème et l'évaluation de la performance d'une solution qui seront utilisées tout au long des travaux.

Le problème a tout d'abord été traité dans un cas particulier où il y a deux classes uni-modales et la covariable est uni-dimensionnelle. Nous avons proposé une méthode MLC dont l'idée principale est de construire toutes les partitions compatibles avec les contraintes et hypothèses du problème, et de choisir la partition avec la valeur maximale de la vraisemblance.

Cette méthode a été appliquée à un ensemble d'individus simulés. En moyenne, la valeur maximale de vraisemblance correspond bien à la partition la plus pertinente des deux classes. De plus, des analyses statistiques ont été effectuées par rapport à l'influence du nombre d'individus et à la dissimilarité entre classes. Ces résultats montrent que la méthode MLC proposée est capable de résoudre le problème pour cette première étude. Cependant, cette méthode est limitée au cas particulier considéré. Pour le cas général, la méthode n'est pas pertinente en raison du nombre de partitions possibles qui devient trop grand, ainsi que la difficulté à définir proprement toutes les partitions possibles. Donc, au lieu de construire *a priori* les partitions selon les valeurs de covariable, il est plus pertinent de construire un système de voisinage dans l'espace de covariable qui permet de définir la proximité locale des individus. Dans le chapitre 4 nous considérons le cas avec le nombre de classes $K \geqslant 2$, la dimension de la covariable quelconque, mais $\Delta t = 1$ et une observation par trajectoire. Dans le chapitre 5 nous considérons le cas général avec aucune contrainte particulière.

Chapitre 4

Solutions dans le cas avec une observation par trajectoire

4.1 Introduction

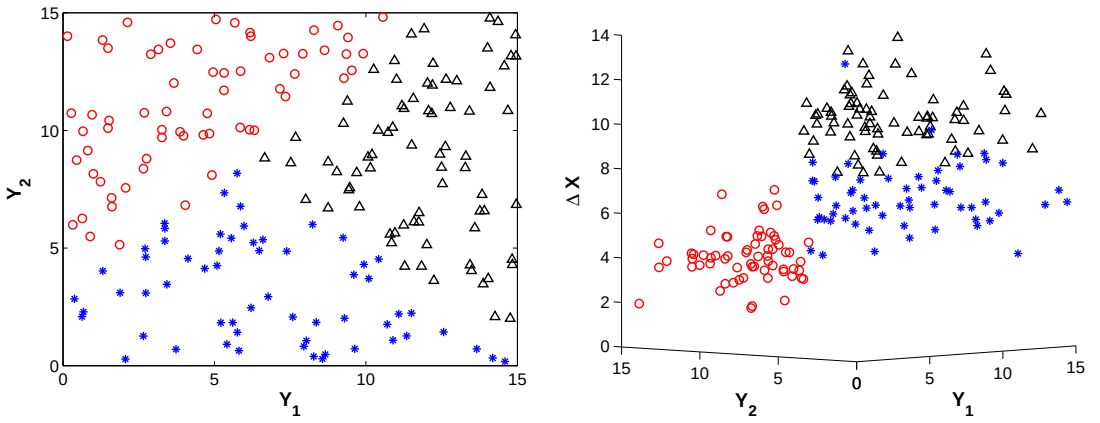
Ce chapitre propose des méthodes qui permettent de trouver des solutions dans le cas plus général que celui décrit dans le chapitre précédent. Nous commençons dans la section 4.2 par introduire brièvement le cas traité dans ce chapitre. Dans la section 4.3 on propose une méthode basée sur un critère local. Ce dernier permet de regrouper les individus proches d'une part, dans l'espace de représentation $\mathcal{X}$, et d'autre part, dans l'espace de covariable $\mathcal{Y}$. Une méthode basée sur un critère global est décrite dans la section 4.4. Contrairement à la première solution, celle-ci permet d'optimiser globalement un critère qui tient compte de la proximité des individus dans les espaces $\mathcal{X}$ et $\mathcal{Y}$. Des études expérimentales sont présentées pour chaque méthode. La conclusion est donnée dans la section 4.5.

4.2 Description du cas traité

Dans le chapitre précédent, nous avons présenté le problème en général et proposé une méthode MLC qui s'adapte à un cas particulier. Ce dernier correspond au cas de deux classes uni-modales et à une covariable de dimension 1. La méthode MLC est basée sur l'idée de maximisation de la vraisemblance totale des classes qui correspondent aux partitions définies selon les valeurs de covariable. Il est remarqué que le critère basé sur la vraisemblance est pertinent, puisque les observations dans notre problème sont supposées suivre des lois Gamma. Pourtant, la façon de partitionnement *a priori* selon les valeurs de covariable ne s'adapte pas pour le cas plus général, e.g. le nombre de classes $K > 2$,, ou la dimension de la covariable $q > 1$. Donc, les méthodes proposées dans ce chapitre retiennent le critère de la vraisemblance, mais utilisent d'autres stratégies

pour tenir compte de l'information dans l'espace de covariable.

Un exemple de ce cas est présenté dans la figure 4.1 où $K = 3$ et $q = 2$. Chaque individu est caractérisé par d'une part, une observation $\Delta x_i \in \Omega_x$ qui représente un incrément du processus Gamma et d'autre part, un vecteur de covariable $\mathbf{y}_i = (y_{i1}, y_{i2}) \in \Omega_y$. Nous considérons dans ce chapitre le cas d'une seule observation pour chaque trajectoire.



(a) dans l'espace de covariable (b) dans l'espace de représentation et de covariable

Figure 4.1 – exemple de partition d'un cas général avec 3 classes et la covariable de dimension 2

4.3 Solution avec un critère local

Dans cette section, nous proposons une méthode basée sur un critère local. Ce dernier peut être interprété comme une vraisemblances locale pondérée en utilisant les individus qui sont relativement proches dans l'espace de covariable et dans l'espace de représentation. Nous commençons par décrire ce critère sur lequel la méthode introduite est basée. Puis, nous présentons des études expérimentales.

4.3.1 Description du critère local

Nous proposons pour chaque individu l_r, $(r = 1, \ldots, N)$ un critère de vraisemblance locale L_{rk}, $(k = 1, \ldots, K)$ calculé avec les paramètres estimés $\hat{\theta}_k$. l_r est attribué à la classe k si L_{rk} est maximale parmi les K valeurs. Afin de tenir compte de la covariable, nous représentons la proximité dans l'espace $\mathcal{Y}$ par un système de voisinage défini dans la section 2.3.2. Il est à noter qu'il existe deux approches principales pour définir un

système de voisinage : l'approche de la fenêtre et l'approche du graphe. La première approche est utilisée dans ce cas. Cependant, au lieu d'utiliser une fenêtre uniforme qui effectue un filtrage simple, nous adoptons la fenêtre gaussienne qui permet d'attribuer un poids w_{ri}, $(i = 1, \ldots, N)$ à chaque individu l_i quand la fenêtre est centrée en individu l_r. La fonction du poids est décrite comme suit :

$$w_{ri} = e^{-\frac{1}{2}(\mathbf{y}_i - \mathbf{y}_r)\Sigma^{-1}(\mathbf{y}_i - \mathbf{y}_r)^T} \quad r, i = 1, \ldots, N \tag{4.1}$$

où $\Sigma = \sigma^2 I$ avec σ un paramètre choisi *a priori* qui représente la dispersion de la fenêtre. Un exemple de telle fenêtre est illustré dans la figure 4.2. Le poids w_{ri} peut être interprété par la distance euclidienne entre les individus l_r et l_i dans l'espace de covariable. Donc, nous pouvons construire une matrice de poids $W_{\mathbf{y}} = [w_{ri}]_{N \times N}$ dont chaque ligne r, $(r = 1, \ldots, N)$ représente le cas où la fenêtre gaussienne est centrée sur l'individu l_r. Cette matrice est symétrique avec tous les éléments diagonaux égaux à 1.

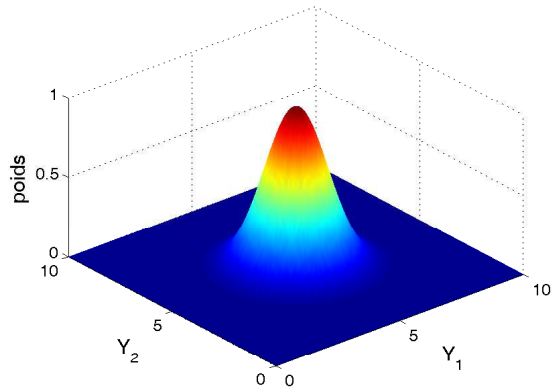


Figure 4.2 – exemple d'une fenêtre gaussienne dans l'espace de covariable

Muni de la matrice $W_{\mathbf{y}}$, nous calculons pour chaque ligne r les p.d.f. pondérées par les poids w_{ri}. Selon la formulation du problème dans la section 3.2.2, une observation est définie comme un incrément de processus Gamma Δx_i qui suit une loi Gamma caractérisée par un vecteur de paramètres inconnu θ_k, $(k = 1, \ldots, K)$. Donc, les p.d.f. pondérées avec la fenêtre centrée en l_r peuvent être décrites comme suit :

$$\bar{f}_{ri,k} = f(\Delta x_i \mid \theta_k) \cdot w_{ri} \quad (i = 1, \ldots, N;\ k = 1, \ldots, K) \tag{4.2}$$

La valeur de $\bar{f}_{ri,k}$ révèle d'une part, de la distance entre les individus l_i et l_r dans l'espace $\mathcal{Y}$, et d'autre part, de la cohérence de l'individu l_i d'être dans la classe k. Dans la figure 4.3 nous montrons la relation entre $\bar{f}_{ri,k}$ et les deux facteurs $f(\Delta x_i \mid \theta_k)$ et

w_{ri}. Dans la figure à gauche avec la fenêtre gaussienne, la cohérence d'individu d'être dans la classe k est plus grande si le carré est plus foncé.

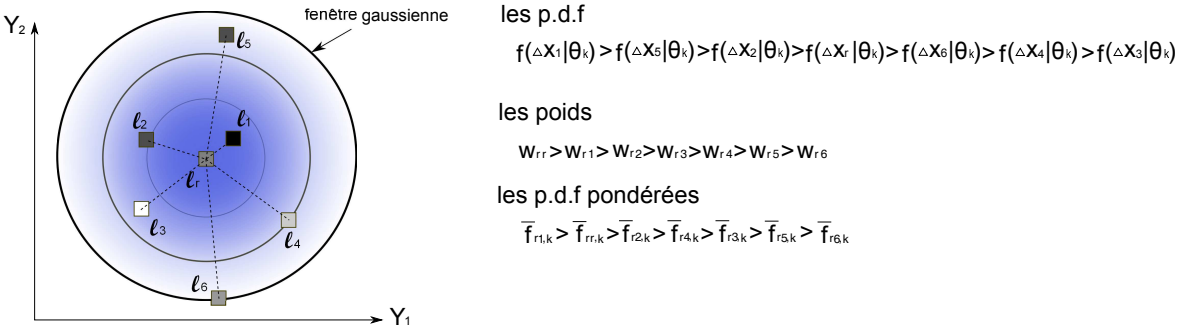


Figure 4.3 – p.d.f. pondérée avec l'individu central l_r

La vraisemblance locale pour la classe k où la fenêtre gaussienne est centrée en l_r est ensuite calculée par le produit des n_V plus grandes valeurs de $\bar{f}_{ri,k}$ où n_V est le nombre de voisins défini *a priori*. En pratique, il est équivalent et plus facile de calculer la log vraisemblance L_{rk} qui est décrite par la formule ci-après :

$$L_{rk} = \sum_{i=1}^{n_V} \log \bar{f}'_{ri,k} \tag{4.3}$$

où les valeurs de $\bar{f}'_{ri,k}$ correspondent aux valeurs de $\bar{f}_{ri,k}$ ordonnées pour vérifier la relation :

$$\bar{f}'_{ri,k} > \bar{f}'_{r(i+1),k}$$

Nous proposons ensuite une méthode de clustering basé sur ce critère local.

4.3.2 Méthode basée sur le critère local

La méthode basée sur le critère local peut se dérouler en trois étapes comme suit :

1. Définir le paramètre σ qui représente la dispersion de la fenêtre gaussienne et calculer la matrice du poids $W_{\mathbf{y}} = [w_{ri}]_{N\times N}$. Choisir le paramètre n_V qui définit le nombre de voisins d'un individu.
2. Générer une partition initiale en appliquant la méthode classique du K-means dans l'espace de représentation $\mathcal{X}$. Estimer les vecteurs de paramètres qui correspondent à la partition initiale.
3. Pour chaque individu l_r, calculer les p.d.f. pondérées de tous les individus selon l'équation 4.2. Trouver les n_V plus grandes valeurs de $\bar{f}'_{ri,k}$ et calculer la log vraisemblance locale selon l'équation 4.3 pour chaque valeur de k. Attribuer l_r à la

classe k qui correspond à la plus grande valeur de L_{rk}. Répéter cette étape jusqu'à ce que les appartenances des individus soient stables.

Cette méthode est développée en pseudo code dans l'Algorithme 2.

Algorithme 2 Méthode basée sur le critère local

1: Définir la valeur de σ et n_V.
2: Calculer $W_{\mathbf{y}} = [w_{ri}]_{N\times N}$
3: Mettre $s = 0$ qui représente le nombre initial d'itération.
4: Générer une partition initiale $\mathcal{P}_K^s$ et calculer les vecteurs de paramètres initiaux $\hat{\Theta}^s = \{\hat{\theta}_1^s, \ldots, \hat{\theta}_K^s\}$.
5: **répéter**
6: **pour** $r = 1$ à N **faire**
7: Calculer les p.d.f. pondérées avec l_r situé au centre de la fenêtre gaussienne :

$$\bar{f}_{ri,k}^s = f(\Delta x_i \mid \hat{\theta}_k^s) \cdot w_{ri} \quad \text{pour } i = 1, \ldots, N;\ k = 1, \ldots, K$$

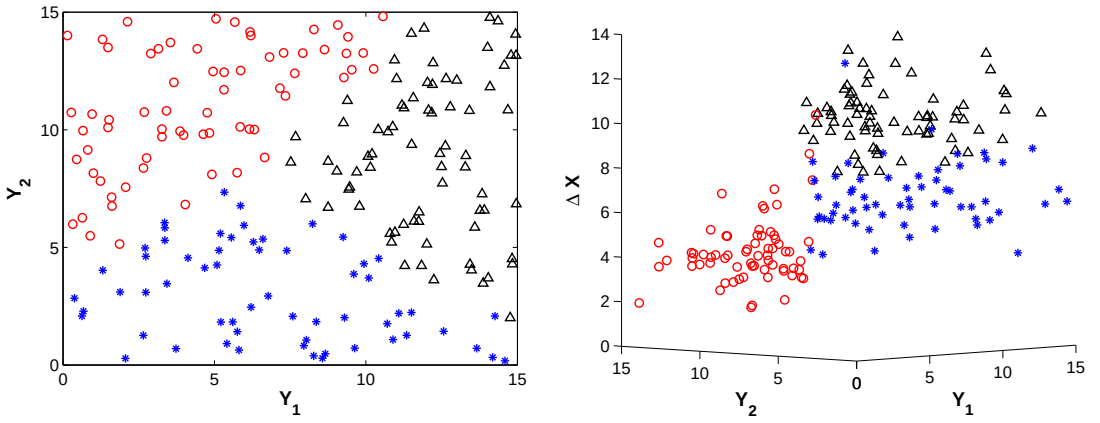
8: $\forall k$, prendre les n_V plus grandes valeurs de $\bar{f}_{ri,k}^s$ et les noter comme $\bar{f'}_{ri,k}^s$.
9: Calculer la log vraisemblance locale $L_{rk}^s = \sum_{i=1}^{n_V} \log \bar{f}'_{ri,k}$, pour $k = 1, \ldots, K$.
10: **fin pour**
11: $\forall r$, attribuer l_r à la classe C_k avec L_{rk}^s la plus grande valeur parmi les K valeurs.
12: Mettre à jours les vecteurs de paramètres $\hat{\Theta}^{s+1} = \{\hat{\theta}_1^{s+1}, \ldots, \hat{\theta}_K^{s+1}\}$ et calculer $\Delta\hat{\Theta} = \| \hat{\Theta}^{s+1} - \hat{\Theta}^s \|$.
13: $s = s + 1$.
14: $\hat{\Theta}^s = \hat{\Theta}^{s+1}$
15: **jusqu'à** $\Delta\hat{\Theta} = 0$

4.3.3 Études expérimentales

La méthode a été appliquée sur l'exemple présenté dans la figure 4.1. Nous précisons que $N = 200$, $K = 3$ et $q = 2$ dans ce cas. De plus, l'espace de covariable $\mathcal{Y}$ est supposé être un carré $d \times d$ avec $d = 15$. Les 3 classes sont séparées par deux frontières décrites respectivement par : $y_2 = y_1 + \frac{d(3-\sqrt{6})}{3}$ et $y_2 = d - y_1$. Cette définition de la partition donne une égalité de la probabilité *a priori* d'appartenance de chaque individu. Par ailleurs, les paramètres théoriques avec la moyenne et la variance sont définis comme ci-après :

$$\theta_1 = \begin{cases} m_1 = 4 \\ \sigma_1^2 = 2 \end{cases}, \ \theta_2 = \begin{cases} m_2 = 7 \\ \sigma_2^2 = 2 \end{cases}, \ \theta_3 = \begin{cases} m_3 = 10 \\ \sigma_3^2 = 2 \end{cases} \tag{4.4}$$

Différentes valeurs pour n_V et σ ont été testées, et nous avons choisi empiriquement $n_V = 7$ et $\sigma = 0.7$. La partition obtenue a été montrée dans la figure 4.4 :



(a) dans l'espace de covariable (b) dans l'espace de représentation et de covariable

Figure 4.4 – solution de la partition sur l'exemple avec $n_V = 8$ et $\sigma = 0.4$

La partition obtenue retrouve bien celle théorique montrée dans la figure 4.1. Nous avons calculé aussi le taux d'individus mal classés qui est égale à 0.02.

Une analyse de la dissimilarité entre classes théoriques a été effectuée. La dissimilarité peut être caractérisée par la valeur de Δm qui représente la différence des valeurs moyennes entre deux classes. Nous avons choisi alors $\Delta m = 1, 1.5, 2, 2.5, 3$. Précisément, nous avons défini $m_k = 4 + (k-1)\Delta m$ et $\sigma_k^2 = 2$. Pour chaque cas de dissimilarité, nous avons généré 200 expériences suivant la même partition théorique illustrée à la figure 4.1. Les valeurs moyennes des paramètres ont été estimées pour les 200 résultats obtenus. La figure 4.5 illustre le taux d'individus mal classés selon différentes valeurs de Δm, et le tableau 4.1 montre les paramètres estimés. $\hat{\mathrm{E}}\{\bar{m}\}$ et $\hat{\sigma}\{\bar{m}\}$ (resp. $\hat{\mathrm{E}}\{\bar{\sigma}^2\}$ et $\hat{\sigma}\{\bar{\sigma}^2\}$) représentent l'espérance et l'écart-type des valeurs moyennes (resp. variances) de 200 expériences théoriques. $\hat{\mathrm{E}}\{\hat{m}\}$ et $\hat{\sigma}\{\hat{m}\}$ (resp. $\hat{\mathrm{E}}\{\hat{\sigma}^2\}$ et $\hat{\sigma}\{\hat{\sigma}^2\}$) représentent l'espérance et l'écart-type des moyennes (resp. variances) de 200 partitions obtenues avec la méthode proposée.

La figure 4.5 montre que le taux d'erreur diminue avec la croissance de la dissimilarité. Dans le tableau 4.1, les estimations des paramètres ($\hat{\mathrm{E}}\{\hat{m}\}$ et $\hat{\mathrm{E}}\{\hat{\sigma}^2\}$) se rapprochent des valeurs théoriques quand la dissimilarité passe de $\Delta m = 1$ à $\Delta m = 1.5$. Mais les estimations deviennent de moins en moins satisfaisantes avec $\Delta m > 1.5$. En effet, les estimations des paramètres dépendent non seulement du nombre d'individus mal classés, mais aussi de l'importance de ces individus. Concrètement, un individu mal classé dans un cas où les classes sont bien séparées est plus important que dans un cas où les classes sont moins séparables. Dans cette étude, le taux d'erreur atteint le minimum avec $\Delta m = 3$, mais les individus mal classés dans ce cas influencent beaucoup sur les estimations des paramètres.

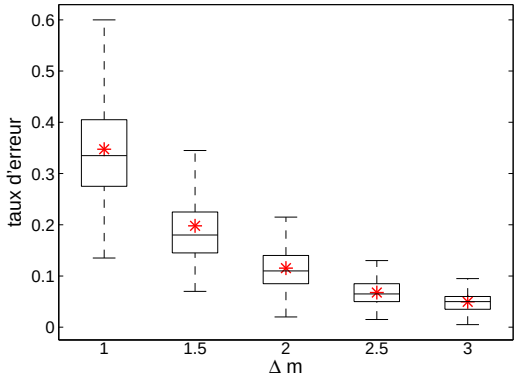


Figure 4.5 – taux d'erreur des partition optimales estimées par rapport aux dissimilarités entre classes

Tableau 4.1 – Résultat des paramètres estimés en fonction de la dissimilarité entre classes

	$\Delta m = 1$			$\Delta m = 1.5$			$\Delta m = 2$			$\Delta m = 2.5$		
	C_1	C_2	C_3	C_1	C_2	C_3	C_1	C_2	C_3	C_1	C_2	C_3
m	4	5	6	4	5.5	7	4	6	8	4	6.5	9
$\hat{\mathrm{E}}\{\bar{m}\}$	4.02	5.01	6.00	3.99	5.51	6.99	4.01	5.98	8.00	3.98	6.50	9.02
$\hat{\mathrm{E}}\{\hat{m}\}$	3.95	5.43	6.12	3.96	5.62	7.06	4.00	5.98	7.98	4.02	6.48	8.93
$\hat{\sigma}\{\bar{m}\}$	0.17	0.18	0.18	0.16	0.19	0.17	0.19	0.17	0.18	0.17	0.17	0.17
$\hat{\sigma}\{\hat{m}\}$	0.41	1.06	0.47	0.26	0.46	0.34	0.22	0.26	0.24	0.19	0.20	0.20
σ^2	2	2	2	2	2	2	2	2	2	2	2	2
$\hat{\mathrm{E}}\{\bar{\sigma}^2\}$	2.00	2.02	2.00	1.99	2.01	2.02	2.02	2.01	2.00	1.98	2.02	2.01
$\hat{\mathrm{E}}\{\hat{\sigma}^2\}$	1.90	2.07	1.86	1.97	2.08	2.04	2.12	2.17	2.15	2.22	2.24	2.33
$\hat{\sigma}\{\bar{\sigma}^2\}$	0.40	0.39	0.38	0.40	0.37	0.38	0.40	0.37	0.40	0.41	0.40	0.32
$\hat{\sigma}\{\hat{\sigma}^2\}$	0.66	1.33	0.76	0.51	0.54	0.66	0.52	0.41	0.48	0.56	0.48	0.50

4.4 Solution avec un critère global

Dans cette section, une solution avec un critère global est présentée. Ce critère, décrit dans la section 4.4.1, combine deux termes caractérisant respectivement la proximité dans l'espace de représentation et dans l'espace de covariable. L'utilisation du critère pour évaluer une partition $E_z = \{z_1, \ldots, z_N\}$ est présentée dans la section 4.4.2. Dans la section 4.4.3, une méthode basée sur ce critère global est développée dans l'objectif de déterminer une solution de la partition.

4.4.1 Description du critère global

Nous décrivons tout d'abord les deux termes du critère global : un terme de vraisemblance qui représente la proximité dans $\mathcal{X}$, et un terme du modèle de MRF (Markov Random Field) qui représente la proximité dans $\mathcal{Y}$ avec la présence du *paramètre de lissage* β. Ensuite, le critère global est présenté avec une technique de normalisation. Enfin, la validation de ce critère est montrée par une étude numérique.

Terme de proximité dans l'espace de représentation

Nous utilisons la vraisemblance pour mesurer la proximité dans l'espace de représentation. La log-vraisemblance de tous les incréments sachant l'ensemble des labels E_z peut être formulée comme suit :

$$L(E_z) = \sum_{i=1}^{N} \log f(\Delta x_i \mid z_i;\, \theta_{z_i}) \tag{4.5}$$

Afin d'obtenir la partition avec la plus grande proximité dans l'espace $\mathcal{X}$, les paramètres pour chaque classe $\theta_1, \ldots, \theta_K$ peuvent être estimés en maximisant la fonction 4.5. Cette idée est proposée dans [66] et une méthode itérative est proposée dans [76] pour réaliser cette optimisation :

- Générer une partition initiale en attribuant chaque observation dans une classe de manière arbitraire. Calculer les estimations des paramètres de chaque classe $\hat{\theta}_1, \ldots, \hat{\theta}_K$.
- Vérifier, pour chaque observation, si on peut améliorer la fonction 4.5 en basculant cette observation dans une autre classe. La mettre dans la classe qui correspond à la meilleure amélioration, et mettre à jour les estimations des paramètres en même temps.
- Arrêter cette itération si la fonction 4.5 ne s'améliore plus.

Pour un nombre de classes donné K, cette méthode permet d'obtenir une log-vraisemblance maximale locale. Nous dénotons cette valeur maximale en L_{ML-K}.

Terme de proximité dans l'espace de covariable

Le terme de proximité dans l'espace de covariable $\mathcal{Y}$ est représenté par le modèle du champ de Markov introduit dans la section 2.5.2. Sachant l'ensemble des labels E_z, le

terme de proximité dans ce cadre est dénoté par $G(\beta, E_z)$ qui vérifie la relation 2.39 :

$$G(\beta, E_z) = \frac{1}{W} e^{-H(\beta, E_z)} \tag{4.6}$$

avec $H(\beta, E_z)$ le modèle de Strauss [72] présenté dans l'équation 2.40.

Il est remarqué que $-H(\beta, E_z)$ est une fonction linéaire qui est proportionnelle au nombre de paires qui partagent le même label de classe. Pour un β fixé, $-H(\beta, E_z)$ atteint la valeur maximale si tous les individus sont attribués à la même classe. Dans ce cas, le terme $G(\beta, E_z)$ atteint aussi le maximum.

Critère global

Le critère global $U(\alpha, \beta, E_z)$ utilise un paramètre $\alpha \in [0, 1]$ qui sert à pondérer les deux espaces $\mathcal{X}$ et $\mathcal{Y}$ dont les proximités sont représentées respectivement par $L(E_z)$ et $G(\beta, E_z)$:

$$U(\alpha, \beta, E_z) = (1 - \alpha)L(E_z) + \alpha G(\beta, E_z) \tag{4.7}$$

Si $\alpha = 0$, l'influence de la covariable est ignorée et le critère global est donc équivalent au critère de la vraisemblance. Si α augmente, la proximité dans $\mathcal{Y}$ devient de plus en plus importante et l'influence de la proximité dans $\mathcal{X}$ diminue. Il est difficile de savoir quantitativement les influences de $L(E_z)$ et $G(\beta, E_z)$ sur $U(\alpha, \beta, E_z)$, car nous avons deux types d'information mesurée différemment. C'est pourquoi nous cherchons un choix pertinent du α avec la connaissance des valeurs de bornes de $L(E_z)$ et $G(\beta, E_z)$. Nous avons montré que la valeur maximale de la vraisemblance L_{ML-K} peut être obtenue avec le nombre de classes K donné. La valeur de maximum de vraisemblance atteint la valeur minimale L_{ML-1} si tous les individus sont dans une seule classe. Par commodité, ces deux valeurs de bornes pour $L(E_z)$ sont dénotées L_K and L_1. De plus, la maximisation de la vraisemblance seulement dans $\mathcal{X}$ avec K classes amène la partition qui correspond au cas 'meilleur attribut - pire covariable' où le terme $G(\beta, E_z)$ est dénoté $\bar{G}_K$. En revanche, la valeur de maximum de vraisemblance L_1 correspond au cas 'pire attribut - meilleur covariable' où le terme $G(\beta, E_z)$ est dénoté G_1. Il est remarqué que $\bar{G}_K$ n'est pas forcément la valeur minimale avec K classes, car une partition peut être moins homogène dans $\mathcal{Y}$ que la partition qui correspond à L_K. Selon les équations 2.39 et 2.40, une borne inférieure à $\bar{G}_K$ peut être obtenue comme $G_K = \frac{1}{W}$ s'il n'existe aucune paire qui partage le même label de classe. Aussi, G_K est une valeur théorique qui ne dépend pas de β et qui est seulement atteignable pour certains systèmes de voisinage.

La figure 4.6 montre les valeurs de bornes de $L(E_z)$ et $G(\beta, E_z)$ ainsi que la relation

entre $U(\alpha, \beta, E_z)$ et α. Deux segments $[L_K, G_K]$ et $[L_1, G_1]$ représentent respectivement le cas 'meilleur attribut - pire covariable' et 'pire attribut - meilleure covariable'. Le croisement des deux segments est noté α_+. Chaque partition E_z a une valeur de vraisemblance $L(E_z)$ entre L_1 et L_K, et une valeur du terme $G(\beta, E_z)$ entre G_K et G_1. Quand $\alpha = \alpha_+$, les deux termes sont dans la même échelle. Le terme $L(E_z)$ (resp. $G(\beta, E_z)$) est dominant si $\alpha < \alpha_+$ (resp. $\alpha > \alpha_+$). Donc, la valeur α_+ est pertinente pour la pondération si aucun terme n'est favorisé *a priori*.

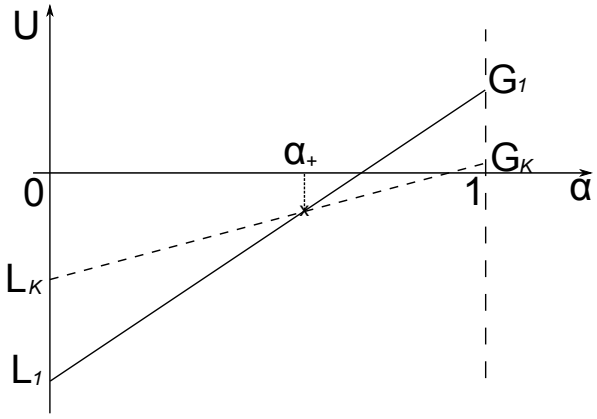


Figure 4.6 – valeurs de bornes de $L(E_z)$ et $G(\beta, E_z)$, et relation entre $U(\alpha, \beta, E_z)$ et α

Une étape de normalisation linéaire a été proposée afin de normaliser les deux termes pour n'importe quelle partition. Cette normalisation est réalisée comme suit :

$$U'(\alpha, \beta, E_z) = (1 - \alpha)L'(E_z) + \alpha G'(\beta, E_z) \tag{4.8}$$

avec

$$L'(E_z) = \frac{L(E_z) - L_1}{L_K - L_1}; \quad G'(\beta, E_z) = \frac{G(\beta, E_z) - G_K}{G_1 - G_K}$$

Le paramètre de lissage β est encore inclus dans le critère $U'(\alpha, \beta, E_z)$. Dans la suite, nous montrons analytiquement que β peut être considéré comme un paramètre redondant, puisqu'il joue le même rôle que α qui permet de contrôler l'importance du terme $G'(\beta, E_z)$ qui peut être écrit comme suit selon l'équation 2.39 :

$$G'(\beta, E_z) = \frac{\frac{1}{W}e^{-H(\beta,E_z)} - \frac{1}{W}}{\frac{1}{W}e^{-H(\beta,E_z^1)} - \frac{1}{W}} = \frac{e^{-H(\beta,E_z)} - 1}{e^{-H(\beta,E_z^1)} - 1} \tag{4.9}$$

où $H(\beta, E_z^1)$ représente la fonction d'énergie lorsque tous les individus appartiennent à une seule classe. L'équation 4.9 montre que $G'(\beta, E_z)$ ne dépend pas de W. Ce dernier est considéré comme une difficulté majeure pour calculer la valeur du modèle de champ de Markov [61, 79]. Par ailleurs, la fonction $-H(\beta, E_z)$ est proportionnelle au nombre

de paires de voisins qui partagent le même label de classe. Pour une partition donnée E_z, le nombre de paires est dénoté $A(E_z)$ et atteint la valeur maximale A_1 dans le cas 'pire attribut - meilleure covariable' et la valeur minimale $A_K = 0$ dans le cas 'meilleur attribut - pire covariable'. La valeur $A_K = 0$ est un minimum théorique qui n'est atteignable que pour certains systèmes de voisinage. Donc, nous avons la relation : $A_1 > A(E_z) \geqslant A_K = 0$, et l'équation 4.9 devient :

$$G'(\beta, E_z) = \frac{e^{A(E_z)\beta} - 1}{e^{A_1\beta} - 1}$$

Les deux valeurs limites avec $\beta \to 0$ et $\beta \to +\infty$ peuvent être présentées comme suit :

$$\lim_{\beta\to 0} G'(\beta, E_z) = \frac{A(E_z)}{A_1}; \quad \lim_{\beta\to +\infty} G'(\beta, E_z) = 0$$

De plus, la monotonie du terme $G'(\beta, E_z)$ par rapport au β peut être déterminée en calculant sa dérivée du premier ordre :

$$\begin{aligned}
\frac{dG'(\beta, E_z)}{d\beta} &= \frac{A(E_z)e^{A(E_z)\beta}(e^{A_1\beta} - 1) - (e^{A(E_z)\beta} - 1)A_1e^{A_1\beta}}{(e^{A_1\beta} - 1)^2} \\
&= \frac{(A_1e^{A_1\beta} - A(E_z)e^{A(E_z)\beta}) - (A_1 - A(E_z))e^{(A_1+A(E_z))\beta}}{(e^{A_1\beta} - 1)^2} \\
&= \frac{\sum_{n=0}^{\infty} \frac{\beta^n[A_1^{n+1} - A(E_z)^{n+1} - (A_1 - A(E_z))(A_1 + A(E_z))^n]}{n!}}{(e^{A_1\beta} - 1)^2}
\end{aligned}$$

avec $(e^{A_1\beta} - 1)^2 > 0$. Donc, il est plus simple de considérer seulement le numérateur pour déterminer le signe de la fonction ci-dessus. Précisément, nous nous sommes intéressés à la fonction ci-après, puisque β et n sont tous positifs.

$$r(n, E_z) = A_1^{n+1} - A(E_z)^{n+1} - (A_1 - A(E_z))(A_1 + A(E_z))^n$$

avec $n = 1, 2, 3 \ldots + \infty$. La récurrence peut être utilisée afin de montrer que $r(p, E_z) \leqslant 0$, $\forall p \geqslant 0$.

- $r(0, E_z) = r(1, E_z) = 0$
- $r(2, E_z) = A_1^3 - A(E_z)^3 - (A_1 - A(E_z))(A_1 + A(E_z))^2 < 0$
- On suppose que $r(p-1, E_z) = A_1^p - A(E_z)^p - (A_1 - A(E_z))(A_1 + A(E_z))^{p-1} < 0$ avec $n = p - 1 > 2$.

- Pour $n = p$, nous avons

$$\begin{aligned}
& (A_1 + A(E_z))r(p-1, E_z) < 0 \\
\Rightarrow\ & A_1^{p+1} - A(E_z)^{p+1} + A_1^p A(E_z) - A_1 A(E_z)^p - (A_1 - A(E_z))(A_1 + A(E_z))^p < 0 \\
\Rightarrow\ & r(p, E_z) + A_1^p A(E_z) - A_1 A(E_z)^p < 0 \\
\Rightarrow\ & r(p, E_z) < A_1 A(E_z)(A(E_z)^{p-1} - A_1^{p-1}) \\
\Rightarrow\ & \text{Comme } A_1 > A(E_z),\ 0 \text{ est un majorant de l'expression à droite, donc :} \\
\Rightarrow\ & r(p, E_z) < 0
\end{aligned}$$

- Donc, on montre que $r(n, E_z) < 0$ pour tout $n \geqslant 2$ en utilisant $A_1 > A(E_z) \geqslant 0$.

La dérivée du premier ordre $\frac{dG'(\beta, E_z)}{d\beta}$ est négative, donc $G'(\beta, E_z)$ décroît par rapport au β. Par conséquent, nous pouvons conclure que la réduction du β amène plus d'importance au terme $G'(\beta, E_z)$, qui peut être aussi obtenue en augmentant α. Le paramètre β peut être considéré redondant et il est plus pratique de choisir le cas limite avec $\beta \to 0$. Donc, un nouveau terme spatial $G(E_z)$ peut être défini dans l'équation 4.10 qui calcule simplement le nombre de paires de voisins qui partagent le même label de classe.

$$G(E_z) = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{k=1}^{K} c_{ik} c_{jk} v_{ij} \tag{4.10}$$

Par conséquent, le terme $\frac{A(E_z)}{A_1}$ est utilisé pour calculer le terme spatial normalisé qui fait partie du nouveau critère global normalisé. Ce dernier ne dépend pas du paramètre β et est reformulé comme suit :

$$U'(\alpha, E_z) = (1 - \alpha) L'(E_z) + \alpha G'(E_z) \tag{4.11}$$

avec

$$L'(E_z) = \frac{L(E_z) - L_1}{L_K - L_1}; \quad G'(E_z) = \frac{A(E_z)}{A_1}$$

4.4.2 Utilisation du critère global pour évaluer une partition

L'évaluation d'une partition a été effectuée en s'appuyant sur le critère global. En utilisant l'exemple illustré par la figure 4.1, nous calculons tout d'abord les valeurs de bornes du terme de la vraisemblance $L(E_z)$ dans l'espace $\mathcal{X}$ et du terme géographique $G(E_z)$ qui est basé sur le système de voisinage dans l'espace $\mathcal{Y}$ construit par le graphe de Gabriel illustré par la figure 4.7.

La figure 4.8a montre l'idée des bornes des termes $L(E_z)$ et $G(E_z)$ selon laquelle nous pouvons choisir empiriquement la valeur du paramètre α. Par exemple, si on

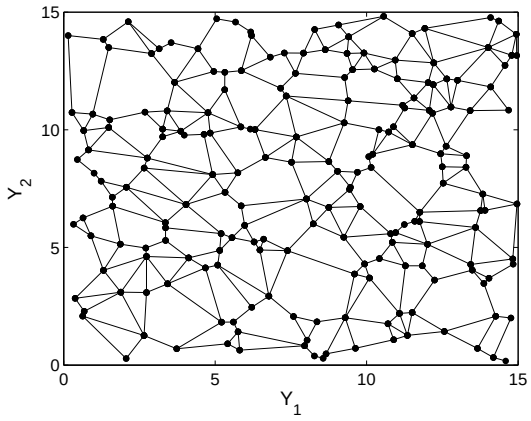
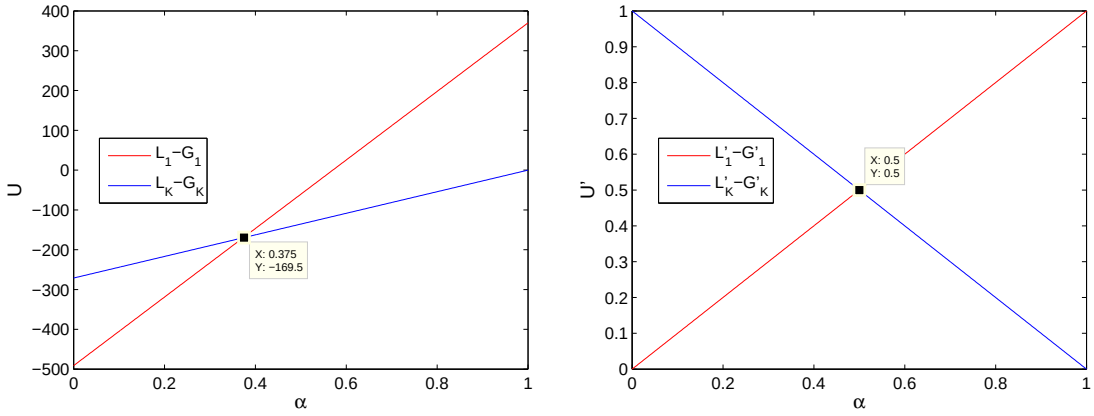


Figure 4.7 – exemple du graphe de Gabriel dans l'espace de covariable $\mathcal{Y}$



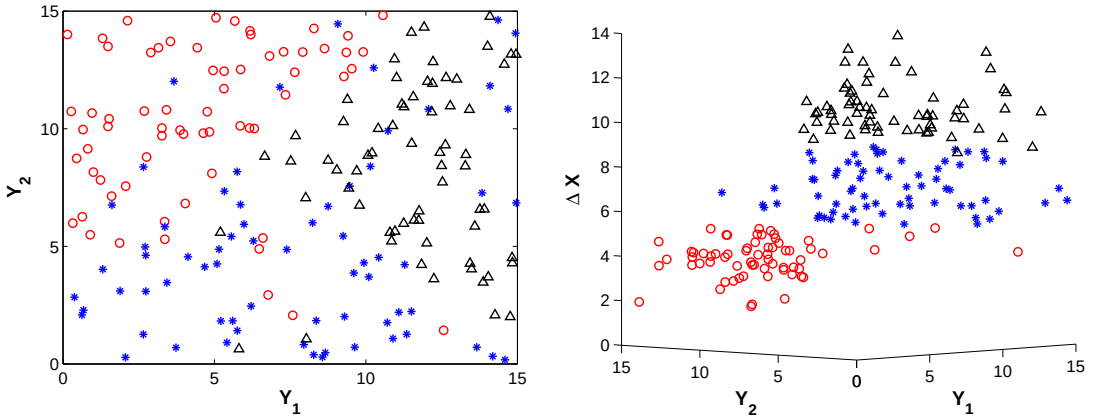
(a) critère global non normalisé (b) critère global normalisé

Figure 4.8 – critère global non normalisé et normalisé

souhaite donner la même importance à la proximité dans l'espace $\mathcal{X}$ et $\mathcal{Y}$, on choisit alors $\alpha = \alpha_+ = 0.375$ qui est équivalent à $\alpha = 0.5$ dans le cas normalisé montré dans la figure 4.8b.

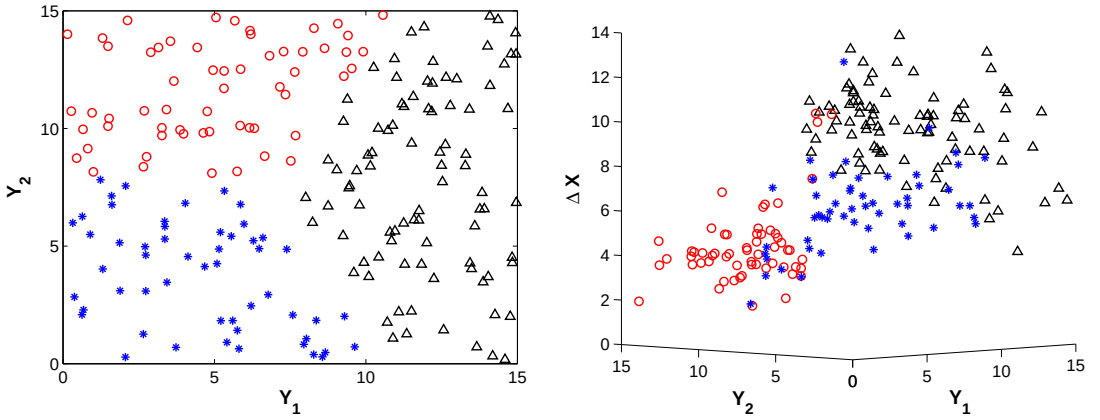
Des études du critère global normalisé ont été effectuées en se basant sur trois partitions différentes : celle théorique, celle avec une grande proximité dans l'espace de représentation et celle avec une grande proximité dans l'espace de covariable. Afin d'obtenir les deux dernières partitions, nous avons appliqué la méthode K-means respectivement dans $\mathcal{X}$ et $\mathcal{Y}$. Selon la section 2.4.2, la méthode K-means permet de minimiser la dispersion *intra-classe*. Les deux partitions obtenues avec K-means ont été montrées dans les figures 4.9 et 4.10. Nous avons tracé dans la figure 4.11 les valeurs du critère $U'(\alpha, E_z)$ par rapport à α. Comme prévu, $U'(\alpha, E_z)$ atteint la valeur maximale pour $\alpha = 0$ (resp. $\alpha = 1$) si la méthode K-means est appliquée seulement dans l'espace $\mathcal{X}$ (resp. $\mathcal{Y}$). Aussi, il est montré qu'avec une valeur de α adéquate (i.e. entre $[0.5, 0.9]$

dans ce cas), $U'(\alpha, E_z)$ a une plus grande valeur avec la partition théorique qu'avec les deux partitions trouvées par K-means.



(a) dans l'espace de covariable (b) dans l'espace de représentation et de covariable

Figure 4.9 – partition obtenue avec K-means dans l'espace $\mathcal{X}$



(a) dans l'espace de covariable (b) dans l'espace de représentation et de covariable

Figure 4.10 – partition obtenue avec K-means dans l'espace $\mathcal{Y}$

4.4.3 Utilisation du critère global pour rechercher une partition

Méthode proposée

Une méthode de clustering basée sur le critère global 4.11 a été développée. Elle permet de choisir une partition en déterminant la valeur du paramètre α.

Dans un premier temps, pour plusieurs valeurs de α, les partitions $\hat{E}_z^*(\alpha)$ sont déterminées. Pour une valeur de α donnée, $\hat{E}_z^*(\alpha)$ est déterminée en maximisant le

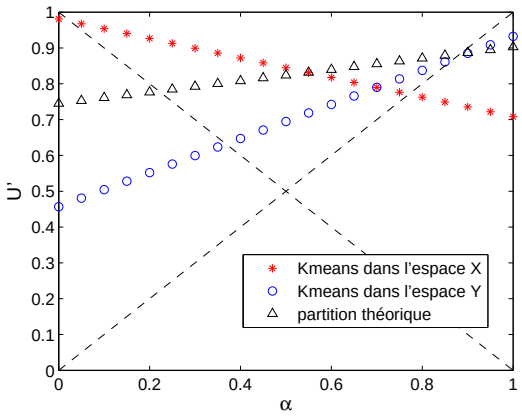


Figure 4.11 – valeurs du critère $U'(\alpha, E_z)$ en fonction de α

critère 4.11. Cette étape peut être formulée comme suit :

$$\hat{E}_z^*(\alpha) = arg\max_{E_z} U'(\alpha, E_z) \tag{4.12}$$

Cette maximisation peut être réalisée par une méthode itérative qui permet de construire une suite d'estimations $\hat{E}_z^s(\alpha)$, $(s = 0, 1, 2, \ldots)$ telles que la valeur du critère augmente à chaque itération. Cette méthode itérative, décrite dans Algorithme 3, est basée sur le principe de l'algorithme d'échange pour le clustering proposé dans [17]. Elle permet de trouver un optimum local, car l'algorithme d'échange est un algorithme glouton. Une solution souvent utilisée est d'appliquer la méthode avec plusieurs initialisations qui amènent différents maximums locaux, et de sélectionner finalement le maximum.

L'algorithme peut être répété pour un ensemble de valeurs *a priori* de α définies entre 0 et 1 avec un pas spécifié $\Delta\alpha$ pour trouver un ensemble des partitions $\hat{E}_z^*(\alpha)$, $(\alpha = 0 : \Delta\alpha : 1)$. Chaque partition $\hat{E}_z^*(\alpha)$ correspond à une valeur des termes $L'(\hat{E}_z^*(\alpha))$ et $G'(\hat{E}_z^*(\alpha))$ qui représentent respectivement la proximité dans l'espace $\mathcal{X}$ et dans l'espace $\mathcal{Y}$. Ces deux termes s'éloignent l'un de l'autre si l'un est favorisé particulièrement. Dans notre problème, on en tient compte de manière équilibrée, puisqu'aucune information *a priori* n'est disponible pour favoriser l'un des deux termes. Donc, la valeur de α peut être choisie en vérifiant le critère comme suit :

$$\hat{\alpha}^* = arg\min_{\alpha} \mid L'\left(\hat{E}_z^*(\alpha)\right) - G'\left(\hat{E}_z^*(\alpha)\right) \mid \tag{4.13}$$

La valeur de α est déterminée finalement comme $\hat{\alpha}^*$, et la partition correspondante est notée $\hat{E}_z^*(\alpha^*)$.

Algorithme 3 méthode itérative pour trouver $\hat{E}_z^*(\alpha)$ avec une valeur de α fixée

1: Mettre $s = 0$.
2: Initialiser $\hat{E}_z^s$ avec la méthode du ML (Maximum de Vraisemblance)
3: Calculer des paramètres initiaux $\hat{\Theta}(\hat{E}_z^s)$.
4: Calculer la valeur initiale du critère $U'(\alpha, \hat{E}_z^s)$.
5: **répéter**
6: **pour** $i = 1 : N$ **faire**
7: Créer des nouvelles partitions $\hat{E}^s_{z_{i,m}}$, $(m = 1, \ldots, K)$ en basculant $\hat{z}_i^s$ dans la classe m parmi K labels possibles.
8: Calculer les paramètres $\hat{\Theta}(\hat{E}^s_{z_{i,m}})$ selon les nouvelles partitions.
9: Calculer la valeur du critère $U'(\alpha, \hat{E}^s_{z_{i,m}})$ selon les nouvelles partitions.
10: **si** il existe m, tel que $U'(\alpha, \hat{E}^s_{z_{i,m}}) > U'(\alpha, \hat{E}_z^s)$ **alors**
11: Mettre $\hat{E}_z^{s+1} = \hat{E}^s_{z_{i,m}}$
12: Mettre $\hat{\Theta}(\hat{E}_z^{s+1}) = \hat{\Theta}(\hat{E}^s_{z_{i,m}})$
13: Mettre $U'(\alpha, \hat{E}_z^{s+1}) = U'(\alpha, \hat{E}^s_{z_{i,m}})$
14: $s \leftarrow s + 1$.
15: **fin si**
16: **fin pour**
17: **jusqu'à** $U'(\alpha, \hat{E}_z^s)$ ne peut plus être augmenté
18: Mettre $\hat{E}_z^*(\alpha) = \hat{E}_z^s$.

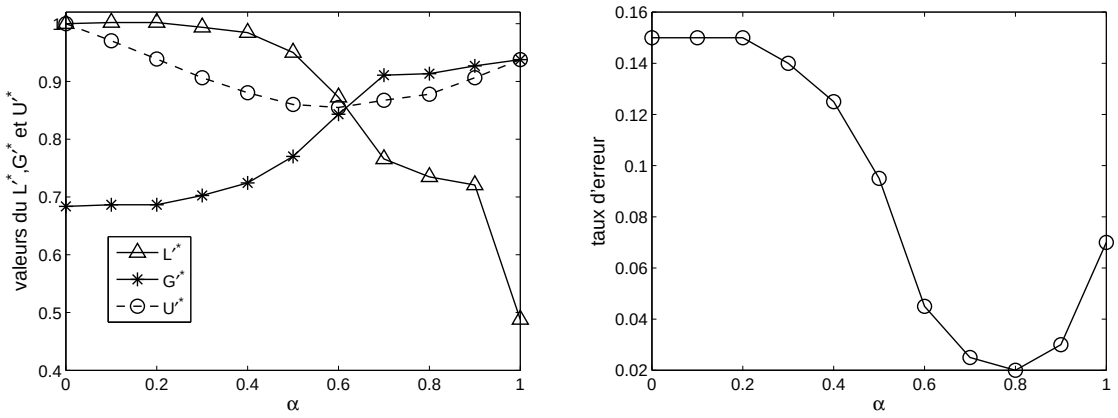
Études expérimentales

Dans cette section, nous évaluons tout d'abord la méthode basée sur le critère global en utilisant l'exemple donné dans la section 4.3.3. Puis, nous utilisons deux exemples réels et comparons les partitions obtenues avec la méthode proposée et deux méthodes de la littérature.

Un exemple simulé est illustré dans la figure 4.1. Il contient 200 individus appartenant à 3 classes. Chaque classe est associée à une distribution Gamma caractérisée par un vecteur de paramètres contenant la moyenne et la variance de chaque classe. Les paramètres sont définis comme $\theta_k = (m_k,\ \sigma_k^2)$ $(k = 1, 2, 3)$ avec $m_1 = 4$, $m_2 = 7$, $m_3 = 10$ et $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 2$. De plus, les valeurs de covariable sont distribuées de manière uniforme dans l'espace de covariable. Pour cet exemple, nous avons utilisé le graphe de Gabriel pour construire le système de voisinage dans $\mathcal{Y}$.

Nous avons illustré dans la figure 4.12a les valeurs des termes $L'(\hat{E}_z^*(\alpha))$, $G'(\hat{E}_z^*(\alpha))$ et du critère $U'(\alpha, \hat{E}_z^*(\alpha))$ par rapport aux différentes valeurs de α. Le taux d'individus mal classés est aussi présenté dans la figure 4.12b.

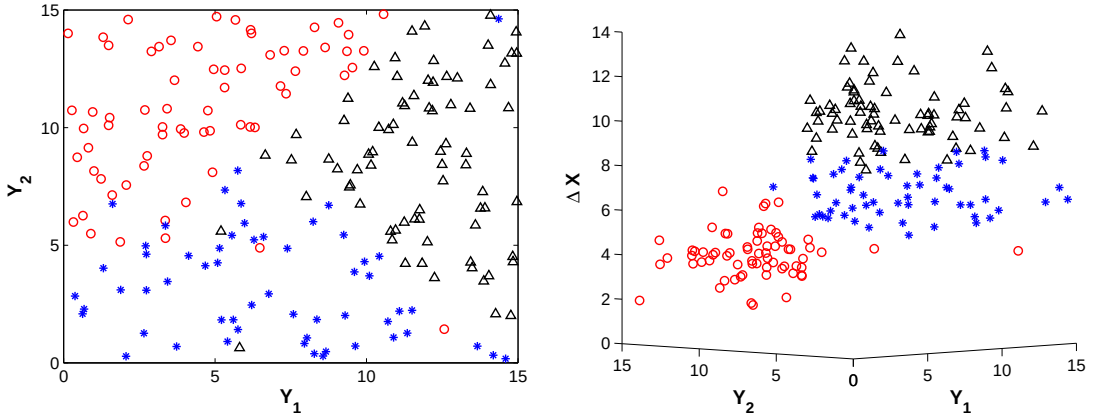
La valeur de $\hat{\alpha}^*$ déterminée selon la relation 4.13 vaut $\hat{\alpha}^* = 0.6$. Ce résultat a été



(a) $L'(\hat{E}_z^*(\alpha))$, $G'(\hat{E}_z^*(\alpha))$ et $U'(\alpha,(\hat{E}_z^*(\alpha)))$ (b) taux d'individus mal classés

Figure 4.12 – évaluation de la méthode proposée selon différentes valeurs de α

montré dans la figure 4.12a. La partition optimale correspondante $\hat{E}_z^*(\alpha^*)$ a été illustrée dans la figure 4.13. Le taux d'erreur maximal et minimal sont égaux respectivement à 0.15 et 0.02 avec $\alpha = 0$ et $\alpha = 0.8$. La partition obtenue par la méthode proposée correspond à un taux d'erreur de 0.04, qui n'est pas loin de l'erreur minimale.



(a) dans l'espace de covariable (b) dans l'espace de représentation et de covariable

Figure 4.13 – résultat de la partition $\hat{E}_z^*(\alpha^*)$ avec $\hat{\alpha}^* = 0.6$

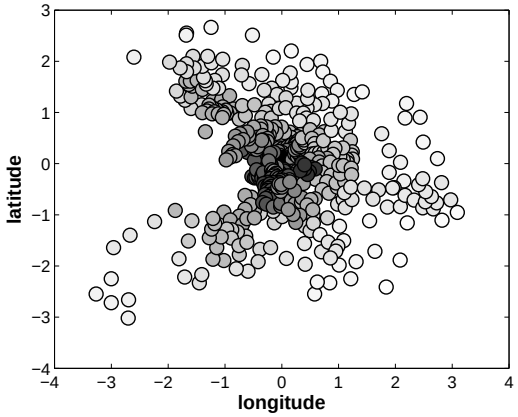
Deux exemples sur des données réelles ont été utilisés où les observations de chaque groupe suivent un modèle de distribution gaussienne. Cette dernière peut être formulée par $\mathcal{N}(m_k, \sigma_k^2)$ où m_k et σ_k^2 représentent respectivement la moyenne et la variance de la classe k. Les partitions obtenues avec la méthode proposée ont été comparées avec celles obtenues avec deux méthodes de la littérature : la méthode ICM (Iterated Conditional Mode) [8] et la méthode NEM (Neighborhood Expectation Maxi-

misation) [2]. Ces deux méthodes ont été brièvement rappelées dans la section 2.5.2. Leur poids, noté respectivement β_{ICM} et β_{NEM}, est utilisé dans chacune des méthodes pour pondérer la proximité dans l'espace géographique. Les partitions obtenues sont alors dépendantes de la valeur du poids.

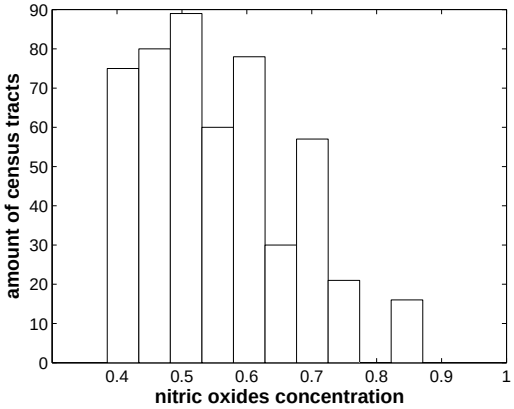
Pour la méthode ICM, la détermination du paramètre β_{ICM} est empirique. Dans le papier [8] β_{ICM} est choisi arbitrairement à 1.5 pour un problème de segmentation d'image. Dans la discussion du même papier, Greig, Porteous et Seheult mentionnent qu'une valeur plus petite de β_{ICM} est plus pertinente sans proposer une solution analytique. Pour la méthode NEM, la convergence de la méthode est prouvée avec $\beta_{NEM} < \frac{1}{n_V^{max}}$ où $n_V^{max} = \max_i \sum_j v_{ij}$ représente le nombre maximal des voisins d'un individu [3]. Il est précisé dans le même papier qu'une valeur plus grande de β_{NEM} peut être acceptée sans violer la convergence, puisque la démonstration de convergence utilise quelques approximations. Il est à noter que les partitions obtenues avec ces deux méthodes dépendent du paramètre géographique non fixé.

Pour la suite, nous avons comparé les méthodes ICM, NEM et celle proposée. La comparaison a été basée sur les partitions obtenues dénotées $\hat{E}_z^*(\gamma)$ où γ représente le coefficient pour chaque méthode, i.e. β_{ICM}, β_{NEM} ou $\hat{\alpha}^*$.

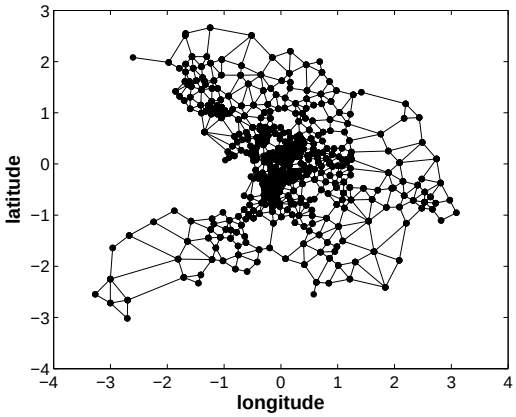
- La base de donnée du bâtiment, disponible dans [49], est utilisée dans le papier [35]. Cette base de donnée contient 14 attributs comme le taux de crime, le nombre d'enseignants et la valeur médiane des bâtiments des 506 quartiers de Boston. Parmi ces attributs, nous avons utilisé celui qui indique le pourcentage de la concentration des oxydes d'azote. Ce dernier est supposé dépendre de la localisation des bâtiments, de sorte que les bâtiments proches ont tendance à avoir une concentration similaire. Dans la figure 4.14a les valeurs de concentration ont été présentées en niveaux de gris. Les points clairs représentent une concentration basse, tandis que les points foncés représentent une concentration élevée. La figure 4.14b montre l'histogramme correspondant selon lequel on vise à regrouper les individus en deux classes comme dans [35]. De plus, nous avons construit le système de voisinage par le graphe de Gabriel illustré dans la figure 4.14c.
  La figure 4.15 montre les partitions obtenues avec chaque méthode mentionnée précédemment. Nous avons zoomé sur la zone centrale où les différences des résultats sont illustrées. La méthode ICM a été appliquée avec le coefficient géographique $\beta_{ICM} = 0.1$ et $\beta_{ICM} = 5$, tandis que la méthode NEM a été appliquée avec $\beta_{NEM} = 0.1$ et $\beta_{NEM} = 3$. Les résultats ont été montrés dans les figures 4.15a, 4.15b, 4.15c et 4.15d. De plus, le résultat avec la méthode proposée a été illustré par la figure 4.15e où $\hat{\alpha}^* = 0.7$ selon l'équation 4.13. L'évaluation des partitions trouvées a été montrée dans la figure 4.15f qui trace



(a) concentration en niveaux de gris



(b) histogramme des niveaux de gris

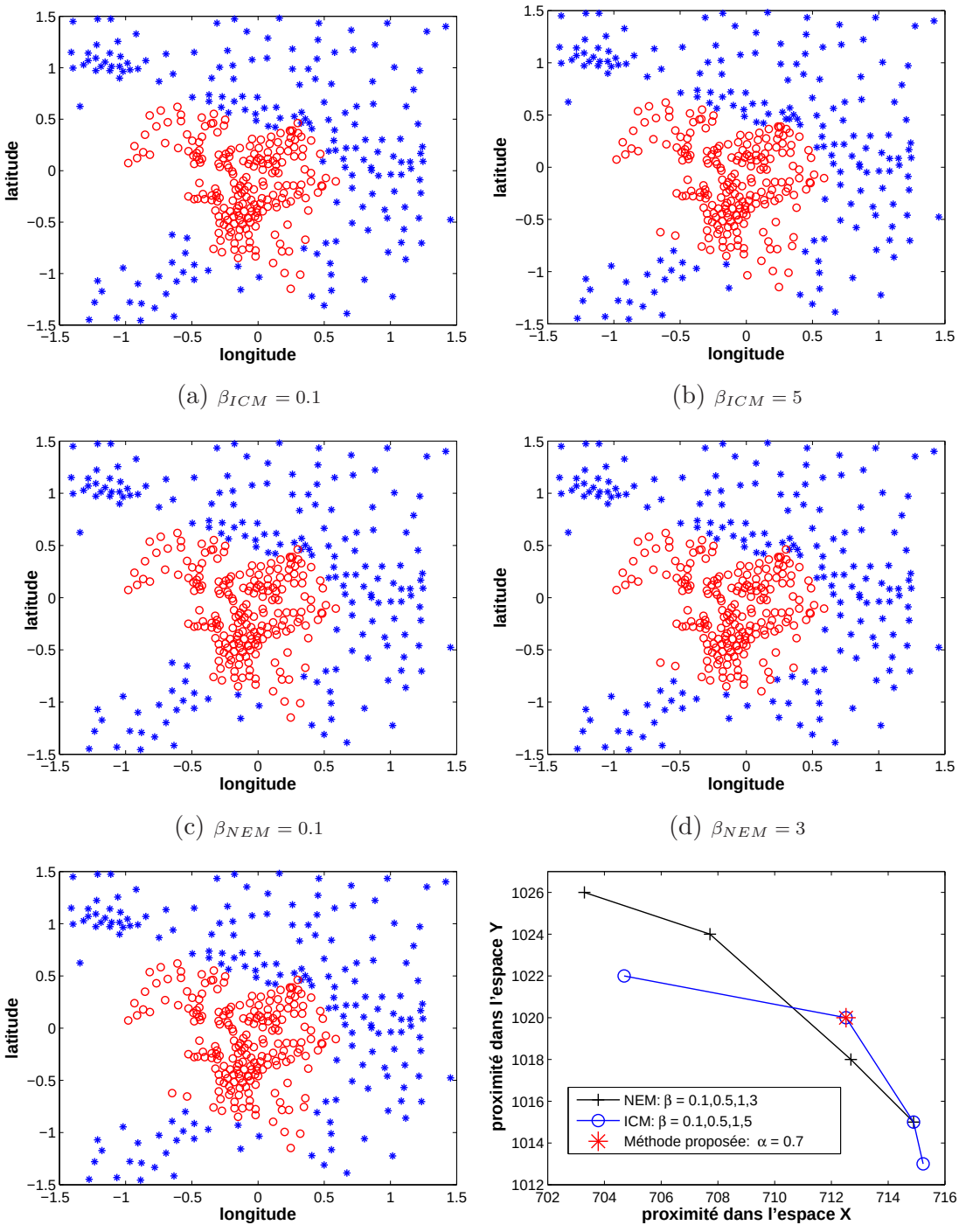


(c) système de voisinage

Figure 4.14 – la base de donnée du bâtiment

d'une part, la valuer de la vraisemblance calculée par l'équation 4.5, et d'autre part, la contiguïté géographique calculée par l'équation 4.10. Les résultats ont été tracés avec $\beta_{ICM} \in \{0.1, 0.5, 1, 5\}$ et $\beta_{NEM} \in \{0.1, 0.5, 1, 3\}$. La contiguïté géographique augmente et la vraisemblance diminue avec la croissance du coefficient géographique. Quant à la méthode proposée, la partition a été obtenue avec la même importance dans les espaces d'attribut et de la covariable. Cependant, une plus grande ou petite valeur de α pourrait être choisie dans l'objectif de favoriser l'un des deux espaces si une telle connaissance *a priori* était disponible.

- La base de donnée de l'image est une image de microscopie électronique en transmission (MET) illustrée dans la figure 4.16. Cette image, aussi utilisée dans [40], illustre la structure assemblée des nano-particules de Fe_3O_4. Il est indiqué dans [40] que les nano-particules de Fe_3O_4 peuvent s'assembler en une submicro-structure agrégée sphérique. La région la plus foncée montre une grande



(a) $\beta_{ICM} = 0.1$ (b) $\beta_{ICM} = 5$

(c) $\beta_{NEM} = 0.1$ (d) $\beta_{NEM} = 3$

(e) $\hat{\alpha}^* = 0.7$ (f) comparaison des partitions obtenues

Figure 4.15 – résultats de partition : (a)(b)(c)—méthode ICM ; (d)(e)(f)—méthode NEM ; (g)—méthode proposée ; (h)—proximité dans $\mathcal{X}$ et dans $\mathcal{Y}$

agrégation des nano-particles, tandis que la région moins foncée indique une agrégation plus faible. Aussi, la région la plus claire est le fond de MET. Donc, nous nous somme intéressés à partitionner cette image en trois clusters d'un point

de vue chimique. Le système de voisinage est régulier dans ce cas, de sorte que les voisins d'un pixel sont ceux qui se situent en haut, en bas, à gauche et à droite de l'individu considéré (sauf les pixels sur la borne de l'image). La méthode ICM a été appliquée avec le coefficient géographique $\beta_{ICM} = 0.1$ et $\beta_{ICM} = 5$, tandis que la méthode NEM a été appliquée avec $\beta_{NEM} = 0.1$ et $\beta_{NEM} = 1$. Les résultats ont été montrés dans les figures 4.15a, 4.15b, 4.15c et 4.15d. Par ailleurs, le résultat avec la méthode proposée a été donné dans la figure 4.17e avec $\hat{\alpha}^* = 0.9$. Les partitions obtenues ont été évaluées dans la figure 4.17f avec $\beta_{ICM} \in \{0.1, 0.5, 1, 5\}$ et $\beta_{NEM} \in \{0.01, 0.1, 0.5, 1\}$. La méthode ICM permet de trouver des solutions avec la plus grande contiguïté géographique, tandis que la méthode NEM ramène aux partitions avec la plus grande vraisemblance. La méthode proposée correspond à une solution intermédiaire.

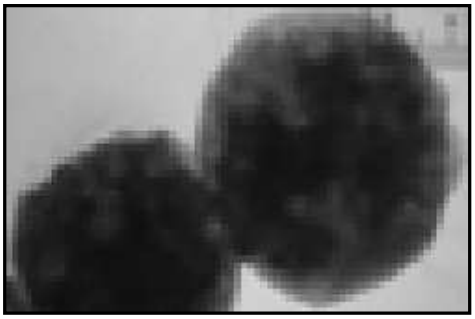
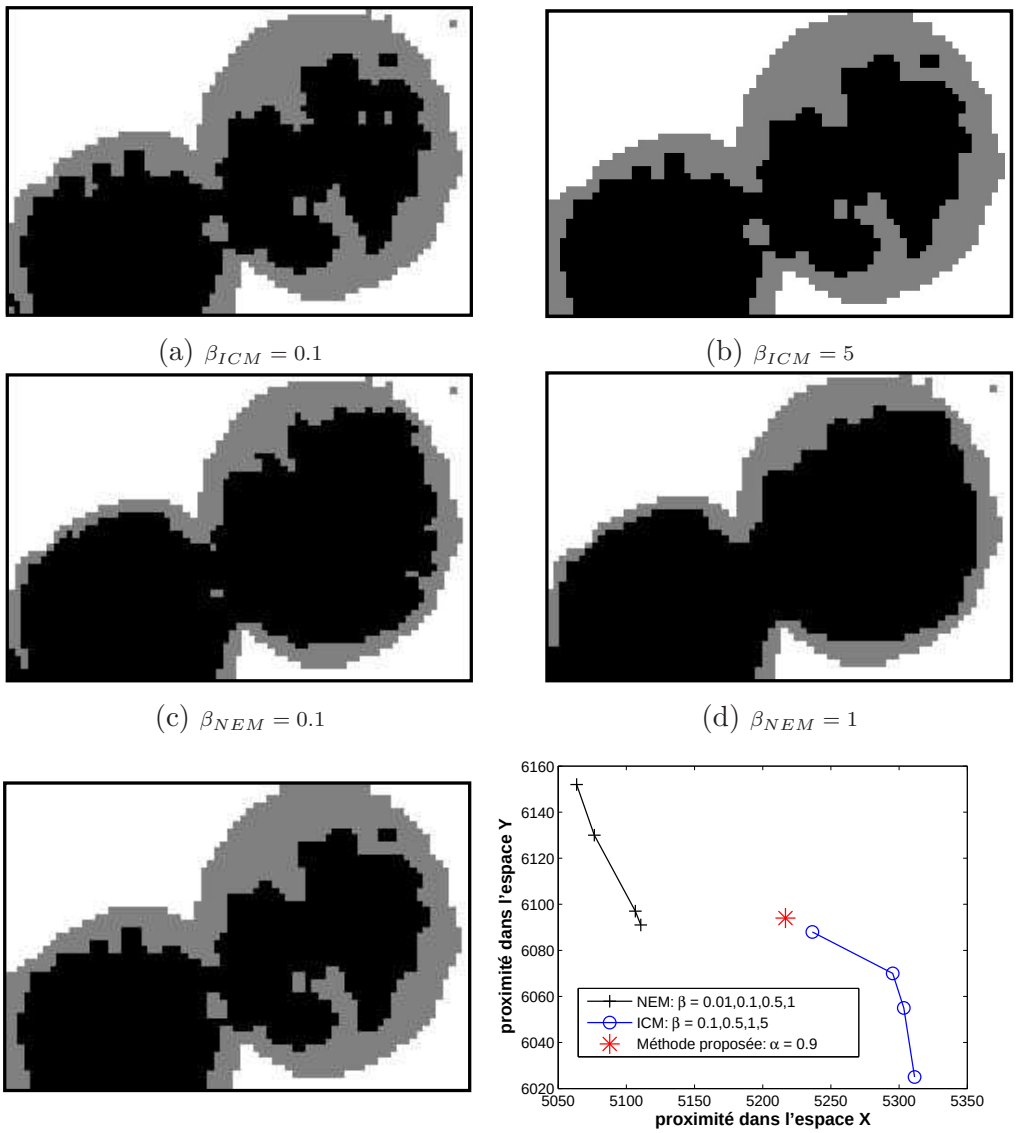


Figure 4.16 – exemple de l'image

4.5 Conclusion

Ce chapitre présente deux méthodes qui permettent de traiter le problème avec le nombre de classes $K > 2$ et la dimension de la covariable $q > 1$ dans le cadre d'une observation par trajectoire. La première méthode proposée est basée sur un critère local qui peut être interprété comme une log vraisemblance locale pondérée selon la proximité des individus dans l'espace de covariable $\mathcal{Y}$. Elle dépend du nombre de voisins sélectionnés n_V et de la dispersion σ de la fenêtre gaussienne. Les résultats numériques ont montré que cette méthode s'adapte bien au cas traité. Cependant, cette méthode ne converge pas dans certains cas particuliers où un individu peut basculer infiniment d'une classe à l'autre. Dans ce cas, la partition change à peine avec ce basculement d'un individu, et il suffit de préciser un nombre d'itérations maximale pour arrêter la méthode.

La deuxième méthode est basée sur un critère global qui combine deux termes



(a) $\beta_{ICM} = 0.1$ (b) $\beta_{ICM} = 5$

(c) $\beta_{NEM} = 0.1$ (d) $\beta_{NEM} = 1$

(e) $\hat{\alpha}^* = 0.9$ (f) comparaison des partitions obtenues

Figure 4.17 – résultats de partition : (a)(b)(c)—méthode ICM ; (d)(e)(f)—méthode NEM ; (g)—méthode proposée ; (h)—proximité dans $\mathcal{X}$ et dans $\mathcal{Y}$

de proximité dans l'espace $\mathcal{X}$ et l'espace $\mathcal{Y}$. Une technique de normalisation a été appliquée afin de mettre ces deux termes dans la même échelle. Le critère global permet d'évaluer toute partition en prenant en considération l'espace $\mathcal{X}$ et $\mathcal{Y}$. Par rapport à la première méthode, cette méthode a deux avantages principaux. Tout d'abord, elle permet de converger vers un optimal local dans tous les cas, puisque c'est une méthode itérative et le critère global augmente à chaque étape. En outre, il n'y a pas de paramètre supplémentaire à spécifier. Le paramètre de pondération α peut être déterminé en donnant la même importance à l'espace $\mathcal{X}$ et $\mathcal{Y}$. Des études numériques ont été effectuées en utilisant un exemple simulé qui correspond au problème de notre

travail, et deux bases de données réelles qui correspondent au problème classique du clustering spatial. Les résultats montrent que la performance de cette méthode est satisfaisante pour les deux cas.

Nous considérons dans ce chapitre qu'il y a seulement une observation pour chaque trajectoire du processus Gamma. Dans le chapitre suivant nous considérons le cas de classification de trajectoires, c'est à dire les observations de la même trajectoire partagent la même valeur de covariable et appartiennent au même processus.

Chapitre 5

Solution dans le cas général

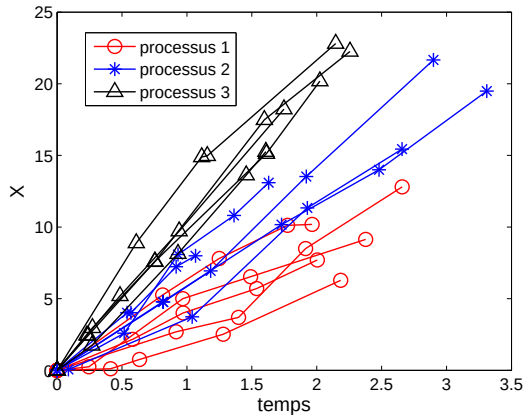
5.1 Introduction

En se basant sur le cas considéré dans le chapitre précédent, une solution a été proposée pour le cas avec plusieurs observations par trajectoire et est présentée dans ce chapitre. Nous rappelons brièvement dans la section 5.2 la notion de trajectoire et la contrainte introduite avec plusieurs observations par trajectoire. Dans la section 5.3, la méthode basée sur le critère global est développée pour qu'elle s'adapte au cas général. Des études expérimentales sont effectuées avec les exemples simulés dans la section 5.4. Nous concluons ce chapitre dans la section 5.5.

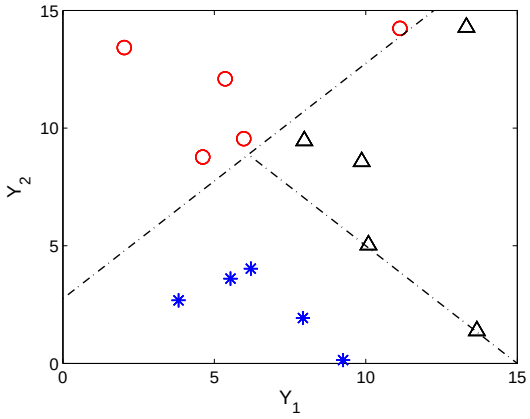
5.2 Description du cas général

Dans le chapitre précédent, le problème correspond au cas où il y a seulement une observation pour chaque trajectoire. Cette hypothèse correspond rarement au cas réel où on a plusieurs observations par trajectoire. Comme introduit dans la section 3.2.2, une trajectoire p_i, $(i = 1, \ldots, N)$ est associée à un vecteur de covariable $\mathbf{y}_i = (y_{i1}, \ldots, y_{iq})$. Toutes les observations x_i^j, $(j = 1, \ldots, \mid p_i \mid)$ provenant de la même trajectoire p_i partagent la même valeur de covariable où $\mid p_i \mid$ est le nombre d'observations de la trajectoire p_i. Un exemple dans ce cadre est donné dans la figure 5.1 avec $N = 15$, $K = 3$, $q = 2$ et $\mid p_1 \mid = \cdots = \mid p_N \mid = 4$. Dans cette section nous généralisons donc le problème étudié. En outre, nous ne nous restreignons plus au cas de l'incrément de temps unique $\Delta t = 1$.

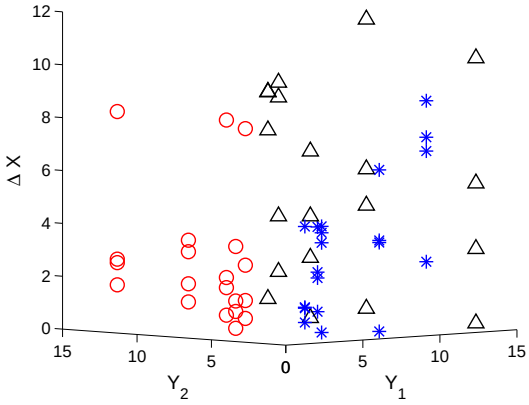
Deux méthodes ont été proposées dans le chapitre précédent : la méthode basée sur un critère local et celle basée sur un critère global. La deuxième méthode est avantageuse par rapport à la première méthode du fait qu'elle puisse converger dans tous les cas et qu'elle ne dépende pas de paramètres. Donc, nous l'avons développée pour qu'elle s'adapte au cas général.



(a) 15 trajectoires dont chacune contient 4 observations



(b) partition des trajectoires dans l'espace $\mathcal{Y}$



(c) partition des trajectoires dans $\mathcal{X}$ et $\mathcal{Y}$

Figure 5.1 – exemple d'un cas général du problème

5.3 Développement de la méthode basée sur le critère global

Le critère décrit dans l'équation 4.11 contient deux termes de proximité normalisés $L'(E_z)$ et $G'(E_z)$ respectivement dans l'espace $\mathcal{X}$ et $\mathcal{Y}$. Afin de développer ce critère pour qu'il s'adapte au cas général, nous devons redéfinir le terme $L(E_z)$ qui impose la contrainte d'appartenir à la même classe pour toutes les observations d'une même trajectoire. En revanche, le terme $G(E_z)$ est indépendant du nombre de réalisations par trajectoire puisqu'il ne dépend que de la covariable.

Basée sur l'équation 4.5, la vraisemblance dans ce cas peut être reformulée comme suit :

$$L(E_z) = \sum_{i=1}^{N} \sum_{j=1}^{|p_i|} \log f(\Delta x_i^j \mid z_i; \theta_{z_i}) \tag{5.1}$$

Cette équation est équivalente à l'équation 4.5 s'il y a seulement une observation par trajectoire, ou bien $\mid p_i \mid = 1$. La méthode ML proposée par [76] a été décrite dans la section 4.4.1. Elle permet de maximiser le critère de vraisemblance avec K classes sans covariable. Cette méthode itérative a été modifiée pour le cas où $\mid p_i \mid$ est quelconque :

- Générer une partition initiale : attribuer chaque trajectoire à une classe de façon arbitraire. Calculer les estimations des paramètres initiaux $\hat{\theta}_1, \ldots, \hat{\theta}_K$.
- Vérifier, pour chaque trajectoire, si on peut améliorer le critère 5.1 en basculant cette trajectoire dans une autre classe. Mettre la trajectoire dans la classe qui correspond à la meilleure amélioration, et mettre à jour les estimations des paramètres en même temps.
- Arrêter lorsque le critère 5.1 ne s'améliore plus.

La valeur ML obtenue par cette méthode itérative est dénotée L_K pour le nombre de classes K spécifié *a priori* et dénotée L_1 pour le cas où il existe seulement une classe.

La méthode pour déterminer la partition en K classes avec prise en compte de la covariable est alors similaire à la méthode avec une observation par trajectoire qui a été présentée dans la section 4.4.3. La seule différence porte sur la détermination de la vraisemblance $L(E_z)$ qui utilise la relation 5.1.

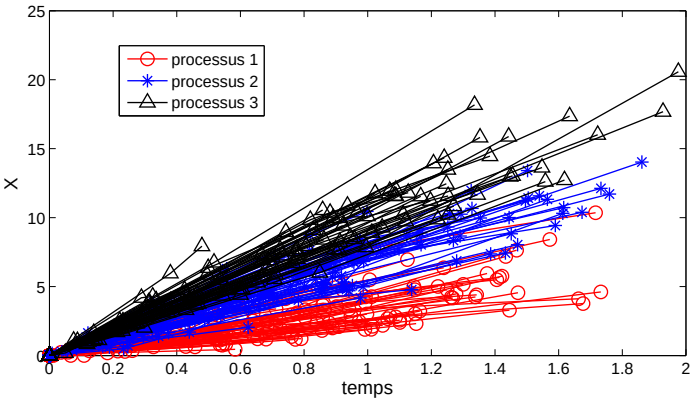
5.4 Études expérimentales

Dans cette section, des études expérimentales illustrent la solution avec plusieurs observations par trajectoire. Nous commençons par un exemple où on a le même nombre d'observations par trajectoire. C'est à dire $\mid p_1 \mid = \cdots = \mid p_N \mid$. Ensuite, nous traitons le problème avec différents nombres d'observations par trajectoire, ce qui est souvent le cas dans un problème réel.

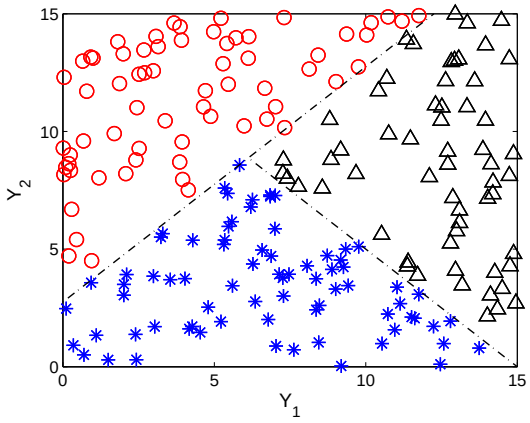
5.4.1 Même nombre d'observations par trajectoire

Une base de données a été simulée dans cet exemple avec $N = 200$, $K = 3$, $q = 2$ et $\mid p_1 \mid = \cdots = \mid p_N \mid = 2$. Δt suit une loi uniforme entre 0 et 1. La figure 5.2a montre 200 trajectoires provenant de 3 processus Gamma homogènes. Ces derniers sont caractérisés par les paramètres $\theta_k = (m_k,\ \sigma_k^2)$, $(k = 1, 2, 3)$ avec $m_1 = 4$, $m_2 = 7$, $m_3 = 10$, et $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 2$. Les trajectoires ne sont pas explicitement séparables et donc l'appartenance de chaque trajectoire n'est pas claire. Les figures 5.2b et 5.2c illustrent la partition théorique des trajectoires.

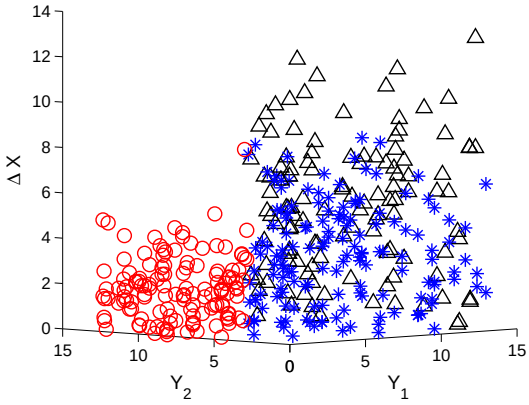
Le système de voisinage a été construit par le graphe de Gabriel et il est illustré



(a) exemple numérique avec 200 trajectoires dont chacune contient 2 observations



(b) partition des trajectoire dans l'espace $\mathcal{Y}$



(c) partition des trajectoire dans $\mathcal{X}$ et $\mathcal{Y}$

Figure 5.2 – exemple avec plusieurs observations par trajectoire

dans la figure 5.3. Les valeurs des termes $L'\left(\hat{E}_z^*(\alpha)\right)$, $G'\left(\hat{E}_z^*(\alpha)\right)$ et $U'\left(\alpha, \hat{E}_z^*(\alpha)\right)$ en fonction de α sont données dans la figure 5.4a et le taux de trajectoires mal classées est rapporté dans la figure 5.4b. Le taux d'erreur relatif aux trajectoires est égal à celui relatif aux individus, puisque le nombre d'individus par trajectoire est constant.

La valeur optimale de α est déterminée en minimisant la différence entre $L'(\hat{E}_z^*(\alpha))$ et $G'(\hat{E}_z^*(\alpha))$, ce qui donne $\hat{\alpha}^* = 0.6$. Avec cette valeur de α déterminée, la partition optimale estimée $\hat{E}_z^*(\alpha^*)$ est illustrée dans la figure 5.5. Les taux maximal et minimal de trajectoires mal classées sont égaux respectivement à 0.175 et 0.015 avec $\alpha = 0$ et $\alpha = 0.7$. La partition obtenue correspond à un taux d'erreur de 0.035, qui n'est pas loin du taux minimal.

Les paramètres du processus Gamma homogène ont été estimés selon les résultats de partition. Afin d'obtenir un résultat statistique, nous avons généré 200 expériences avec la même partition théorique montrée dans la figure 5.1. Ensuite, la méthode a été appliquée sur chaque expérience et les paramètres ont été estimés pour chaque partition

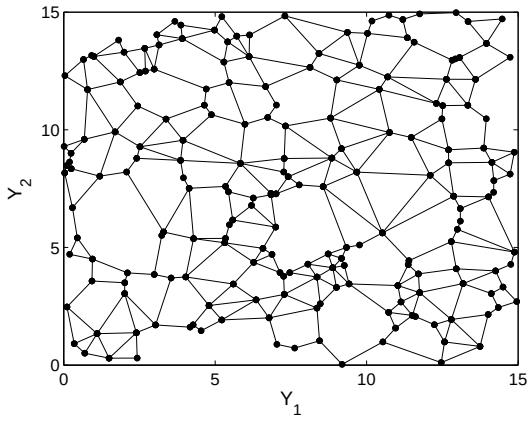
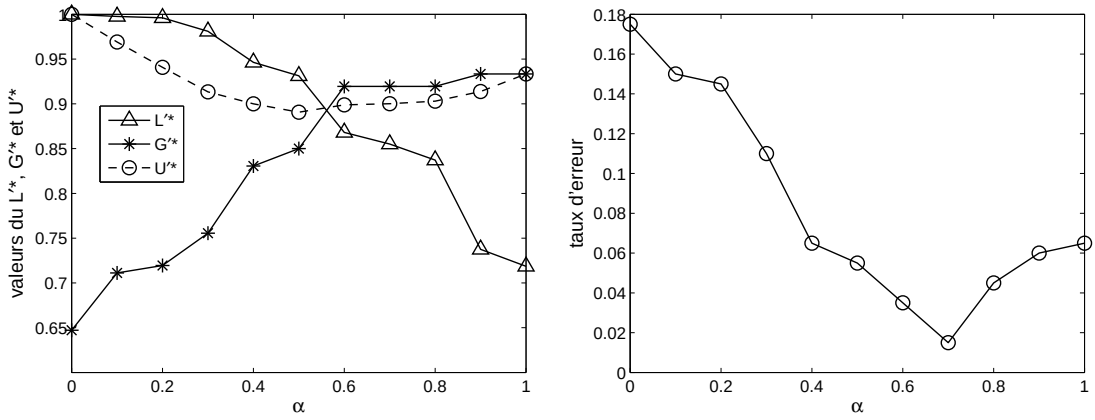
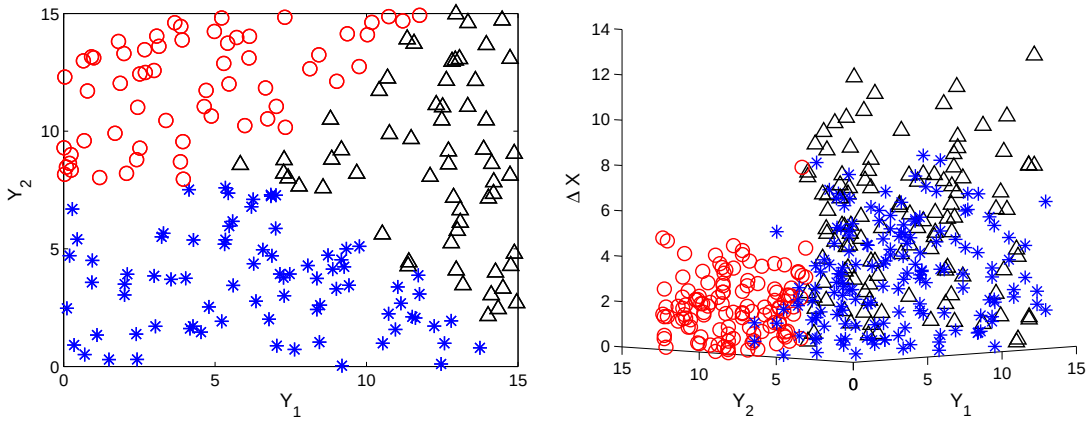


Figure 5.3 – graphe de Gabriel dans l'espace de covariable $\mathcal{Y}$



(a) $L'(\hat{E}_z^*(\alpha))$, $G'(\hat{E}_z^*(\alpha))$ et $U'(\alpha, \hat{E}_z^*(\alpha))$ (b) taux de trajectoires mal classées

Figure 5.4 – valeur des critères pour la méthode proposée selon différentes valeurs de α



(a) dans l'espace de covariable (b) dans l'espace de représentation et de covariable

Figure 5.5 – résultat de la partition $\hat{E}_z^*(\alpha^*)$ avec $\hat{\alpha}^* = 0.6$

Tableau 5.1 – Estimation des paramètres dans le cas où le nombre d'observations par trajectoire est identique

	C_1	C_2	C_3
m	4	7	10
$\hat{\mathrm{E}}\{\bar{m}\}$	4.01	7.01	9.99
$\hat{\mathrm{E}}\{\hat{m}\}$	3.93	6.99	10.06
$\hat{\sigma}\{\bar{m}\}$	0.17	0.19	0.17
$\hat{\sigma}\{\hat{m}\}$	0.20	0.26	0.19
σ^2	2	2	2
$\hat{\mathrm{E}}\{\bar{\sigma}^2\}$	1.99	2.02	1.97
$\hat{\mathrm{E}}\{\hat{\sigma}^2\}$	1.84	1.77	1.87
$\hat{\sigma}\{\bar{\sigma}^2\}$	0.34	0.37	0.29
$\hat{\sigma}\{\hat{\sigma}^2\}$	0.34	0.37	0.31

trouvée. Les résultats sont montrés dans le tableau 5.1. $\hat{\mathrm{E}}\{\bar{m}\}$ et $\hat{\sigma}\{\bar{m}\}$ (resp. $\hat{\mathrm{E}}\{\bar{\sigma}^2\}$ et $\hat{\sigma}\{\bar{\sigma}^2\}$) représentent l'espérance et l'écart-type des moyennes (resp. variances) de 200 simulations avec les partitions théoriques. Aussi, $\hat{\mathrm{E}}\{\hat{m}\}$ et $\hat{\sigma}\{\hat{m}\}$ (resp. $\hat{\mathrm{E}}\{\hat{\sigma}^2\}$ et $\hat{\sigma}\{\hat{\sigma}^2\}$) représentent l'espérance et l'écart-type des moyennes (resp. variances) de 200 simulations avec les partitions obtenues. Pour ce problème la méthode fournit de bons résultats.

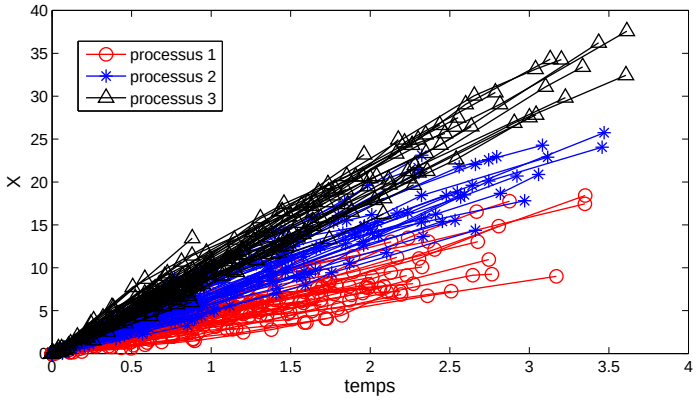
5.4.2 Nombre d'observations par trajectoire quelconque

Dans le cadre de la sûreté de fonctionnement des systèmes, les instants d'observations et le nombre d'observations ne sont généralement pas choisis et dépendent des conditions d'exploitation. Aussi, pour un problème réel, le nombre d'observations sur chaque trajectoire peut être quelconque.

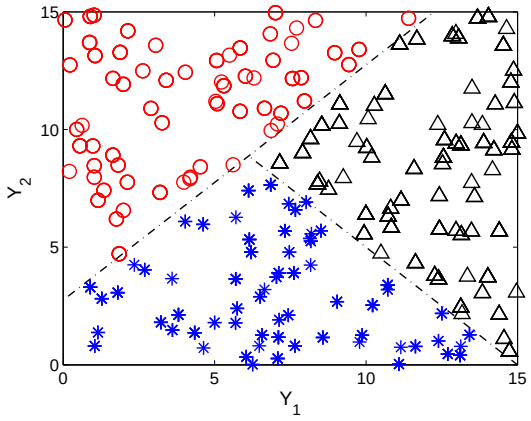
La figure 5.6a illustre un exemple avec 200 trajectoires appartenant à 3 processus Gamma homogènes caractérisés par les paramètres $\theta_k = (m_k,\ \sigma_k^2)$, $(k = 1, 2, 3)$ avec $m_1 = 4$, $m_2 = 7$, $m_3 = 10$, et $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 2$. Le nombre d'observations par trajectoire est une variable discrète uniformément distribuée entre 1 et 5. Les figures 5.6b et 5.6c montrent la partition théorique des trajectoires.

Les valeurs des termes $L'\left(\hat{E}_z^*(\alpha)\right)$, $G'\left(\hat{E}_z^*(\alpha)\right)$ et $U'\left(\alpha, \hat{E}_z^*(\alpha)\right)$ ont été tracées dans la figure 5.7a et le taux d'individus mal classés est montré dans la figure 5.7b. Aussi, la partition obtenue avec $\hat{\alpha}^* = 0.7$ est illustrée dans la figure 5.8. Ce résultat de partition correspond au taux d'erreur minimal qui est égal à 0.02.

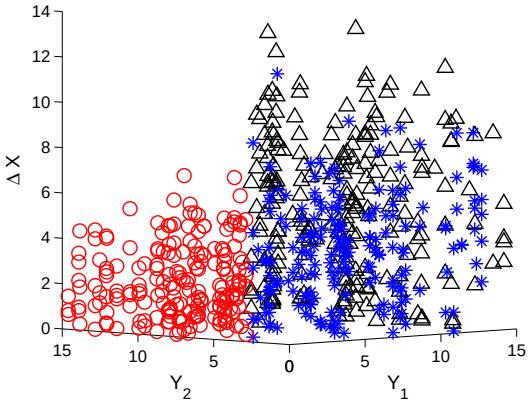
Les paramètres du processus Gamma homogène ont été estimés en se basant sur



(a) exemple numérique avec différents nombres d'observations par trajectoire

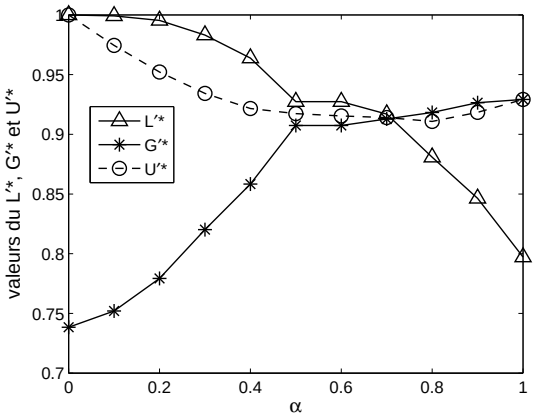


(b) partition des trajectoire dans l'espace $\mathcal{Y}$

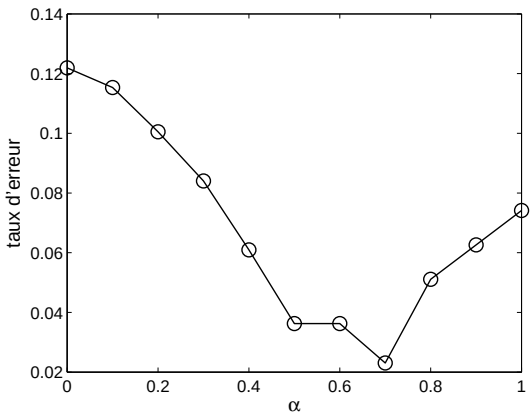


(c) partition des trajectoire dans $\mathcal{X}$ et $\mathcal{Y}$

Figure 5.6 – exemple d'un cas général avec différents nombres d'observations par trajectoire

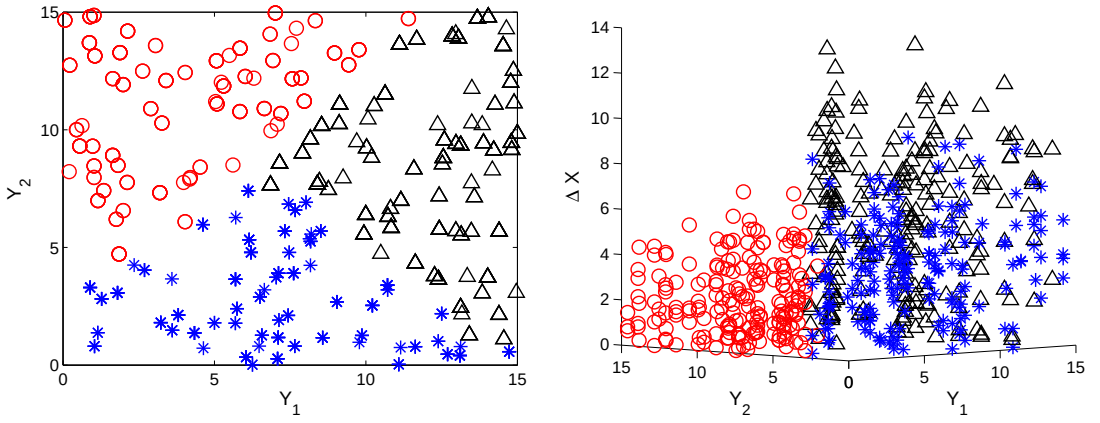


(a) $L'(\hat{E}_z^*(\alpha))$, $G'(\hat{E}_z^*(\alpha))$ et $U'(\alpha, (\hat{E}_z^*(\alpha)))$



(b) taux d'individus mal classées

Figure 5.7 – méthode proposée selon différentes valeurs de α



(a) dans l'espace de covariable (b) dans l'espace de représentation et de covariable

Figure 5.8 – résultat de la partition $\hat{E}_z^*(\alpha^*)$ avec $\hat{\alpha}^* = 0.7$

Tableau 5.2 – Estimation des paramètres dans le cas où le nombre d'observations par trajectoire est quelconque

	C_1	C_2	C_3
m	4	7	10
$\hat{\mathrm{E}}\{\bar{m}\}$	4	7.02	10
$\hat{\mathrm{E}}\{\hat{m}\}$	3.87	6.91	10.05
$\hat{\sigma}\{\bar{m}\}$	0.14	0.15	0.14
$\hat{\sigma}\{\hat{m}\}$	0.18	0.19	0.18
σ^2	2	2	2
$\hat{\mathrm{E}}\{\bar{\sigma}^2\}$	1.98	2.02	2
$\hat{\mathrm{E}}\{\hat{\sigma}^2\}$	1.81	1.74	1.89
$\hat{\sigma}\{\bar{\sigma}^2\}$	0.27	0.25	0.23
$\hat{\sigma}\{\hat{\sigma}^2\}$	0.32	0.33	0.28

les 200 expériences. Le résultat est montré dans le tableau 5.2. Il est à noter que les paramètres théoriques ont été retrouvés sans avoir une erreur importante.

5.5 Conclusion

Ce chapitre présente une solution dans le cas général où il existe plusieurs observations par trajectoire. Le critère global introduit dans le chapitre précédent a été adapté. Des études expérimentales ont été effectuées dans deux cas. Le premier concerne le même nombre d'observations par trajectoire, tandis que le deuxième cas consiste à traiter le problème où le nombre d'observations par trajectoire est quelconque. Les expériences ont montré la faisabilité de la méthode proposée.

Chapitre 6

Conclusion et perspectives

Dans ce dernier chapitre, la section 6.1 rapporte une synthèse des travaux qui ont été présentés dans les chapitres précédents. Ensuite, des pistes de travail restant à explorer sont présentées dans la section 6.2.

6.1 Synthèse des travaux

Ce mémoire présente les travaux de thèse intitulée *Détermination de classes de modalités de dégradation significatives pour le pronostic et la maintenance*. Les travaux principaux ont consisté à développer des méthodes de clustering adaptées à certaines catégories de données de surveillance qui caractérisent des modes d'évolution de vieillissement.

Le chapitre 1 a donné une vue générale du contexte des travaux et de la problématique générale. Cette dernière porte sur le clustering de systèmes dans des classes caractérisant un mode d'évolution de vieillissement dépendant de covariables caractéristiques du système ou de son utilisation.

Suite à la présentation dans le chapitre 1 qui montre que notre problème se situe à la charnière entre la thématique de la reconnaissance des formes et celle de la sûreté de fonctionnement, les méthodes de clustering ainsi que le processus de dégradation ont été rappelés dans le chapitre 2. Les méthodes de clustering dans la littérature se répartissent en deux catégories : le clustering classique et le clustering spatial. Les méthodes de clustering classique, qui traitent le problème caractérisé uniquement par l'attribut, ne sont pas directement utilisables car notre problème dépend aussi de la covariable. Par ailleurs, les méthodes de clustering spatial, qui considèrent souvent que la covariable est bi-dimensionnelle, correspond partiellement à notre problème. Nous avons aussi présenté le processus de dégradation en introduisant le modèle du processus Gamma qui est largement utilisé pour décrire une loi de vieillissement.

Nous avons décrit notre problème dans le chapitre 3. Par rapport aux problèmes de

clustering classiques, la spécificité principale de notre problème vient du fait que nous visons à regrouper des systèmes selon leur évolution de vieillissement. Cette dernière est caractérisée par d'une part, les mesures de vieillissement qui sont considérées comme les réalisations d'attribut, et d'autre part, la condition dans laquelle un système se dégrade et qui est interprétée comme la valeur de covariable. Une méthode MLC a été proposée pour trouver la solution dans un cas simple où il existe une seule mesure par évolution avec la période d'inspection constante, deux groupes de densités uni-modales et une covariable uni-dimensionnelle. L'idée principale de cette méthode est de construire toutes les partitions compatibles avec les contraintes considérées, et de choisir la partition avec la valeur maximale de la vraisemblance. Les études expérimentales ont été effectuées avec un ensemble d'individus simulés, et elles ont montré la performance de cette méthode selon le nombre de mesures et la dissimilarité entre classes. Toutefois, la méthode n'est plus pertinente pour les cas où le nombre de classes K est supérieur à 2 et la dimension de la covariable q est supérieur à 1. C'est parce que le nombre de partitions possibles devient trop grand, et donc il est impossible de définir toutes les partitions possibles.

Ensuite, deux méthodes ont été proposées dans le chapitre 4 pour le cas où le nombre de classes et la dimension de la covariable peuvent être quelconques. Afin de tenir compte de la connaissance introduite par la covariable, ces deux méthodes considèrent un système de voisinage dans l'espace de covariable. La première méthode est basée sur un critère local de chaque individu qui est décrit par une vraisemblance locale en sélectionnant les voisins de cet individu. Des études expérimentales ont montré que la partition trouvée est satisfaisante avec des paramètres bien choisis. Cependant, la convergence de la méthode n'est pas assurée dans certains cas particuliers où quelques individus basculent indéfiniment d'une classe à l'autre. La deuxième méthode est basée sur un critère global. Ce dernier contient d'une part, un terme de proximité dans l'espace de l'attribut décrit par la vraisemblance du modèle de dégradation, et d'autre part, un terme de proximité dans l'espace de covariable décrit par le modèle de MRF (champ de Markov). Ces deux termes sont combinés par un paramètre α qui contrôle l'importance de chacun. Cependant, comme ils représentent deux types d'information, il est difficile de savoir un terme est important ou pas par rapport à l'autre. C'est pourquoi nous avons proposé une normalisation linéaire qui permet de rendre ces deux termes dans la même échelle. De plus, nous avons montré que le terme du modèle de MRF est équivalent à un terme plus simple qui calcule le nombre de paires de voisins qui partagent le même label de classe. Cette méthode a été comparée avec deux autres méthodes de la littérature : la méthode ICM et la méthode NEM. L'avantage de la méthode proposée est sa capacité à déterminer le paramètre α en imposant la même

importance dans les deux espaces. Il est à noter que le résultat de la partition peut être amélioré en ajustant la valeur de α si on connaît *a priori* l'importance de chaque espace.

Nous avons proposé une solution avec plusieurs observations par trajectoire dans le chapitre 5. La méthode globale proposée dans la section 4.4.3 a été développée pour s'adapter à ce cas. Deux types d'exemples ont été simulés dans l'étude expérimentale : l'exemple avec le même nombre d'observations par trajectoire et celui avec différents nombres d'observations par trajectoire. Le premier est un exemple théorique, tandis que le deuxième exemple permet de se rapprocher du problème réel. Les partitions théoriques dans les deux exemples ont été déterminées avec peu d'erreurs, et l'estimation des paramètres du processus Gamma homogène est satisfaisante. Il est important de noter que plus de classes se ressemblent, plus la distinction des classes est difficile mais moins les erreurs sont importantes, car des modèles peu différents correspondent à des processus de dégradation peu différents et donc les conséquences sur le pronostic sont peu critiques.

6.2 Perspectives de recherche

Les travaux de recherche dans cette thèse ouvrent diverses perspectives de recherche.

En premier lieu, la méthode globale proposée dans le chapitre 4 pourrait être améliorée avec la connaissance *a priori* de l'importances des proximités dans l'espace de représentation et dans l'espace de covariable. Précisément, la valeur optimale du coefficient $\hat{\alpha}^*$ est déterminée en minimisant la différence entre $L'\left(\hat{E}_z^*(\alpha)\right)$ et $G'\left(\hat{E}_z^*(\alpha)\right)$, qui représentent respectivement la proximité dans $\mathcal{X}$ et $\mathcal{Y}$. Cela veut dire que le résultat de la partition est obtenu en donnant la même importance aux deux espaces. Cependant, il est peut être plus pertinent de favoriser une des deux proximités dans certains cas. Par exemple, dans le cas où les individus sont assez proches dans $\mathcal{X}$ et assez éloignées dans $\mathcal{Y}$, il est logique de favoriser l'espace $\mathcal{Y}$ et de négliger l'espace $\mathcal{X}$.

D'autre part, le nombre de classes K est supposé connu au cours de nos travaux. La détermination de ce paramètre est un problème majeur du clustering. Dans les travaux suivants, on pourrait tracer l'évolution de la valeur du critère global introduit dans le chapitre 4 par rapport aux différentes valeurs de K dont des deux termes du critère dépendent directement. Une première approche du problème pourrait être de déterminer les valeurs du critère global selon un ensemble des valeurs de K choisies. Il serait ensuite intéressant de voir si l'évolution de la valeur du critère présente des ruptures qui pourrait correspondre à des valeurs pertinentes pour K.

L'intérêt principale de ce travail réside dans le pronostic et la maintenance. Lorsque le résultat d'une partition est trouvé et les lois de vieillissement sont simultanément estimées, le mode de vieillissement et le modèle associé à un nouveau système peut être déterminés en s'appuyant sur les conditions dans lesquelles le système est employé. D'un point de vue applicatif, ces connaissances permettent de prédire l'évolution de vieillissement des systèmes et ainsi une politique de maintenance adaptée peut être proposée.

Bibliographie

[1] Alzate, C., and Suykens, J. A. Hierarchical kernel spectral clustering. *Neural Networks* (2012).

[2] Ambroise, C., Dang, M., and Govaert, G. Clustering of spatial data by the EM algorithm. *geoENV I-Geostatistics for Environmental Applications 9* (1997), 493–504.

[3] Ambroise, C., and Govaert, G. Convergence of an EM-type algorithm for spatial clustering. *Pattern recognition letters 19*, 10 (1998), 919–927.

[4] Anderberg, M. R. Cluster analysis for applications. Tech. rep., DTIC Document, 1973.

[5] Applebaum, D. *Lévy processes and stochastic calculus*, vol. 93. Cambridge university press, 2004.

[6] Arce, G. R. *Nonlinear signal processing : a statistical approach.* Wiley. com, 2005.

[7] Barker, C., and Newby, M. Optimal non-periodic inspection for a multivariate degradation model. *Reliability Engineering & System Safety 94*, 1 (2009), 33–43.

[8] Besag, J. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society. Series B (Methodological)* (1986), 259–302.

[9] Blain, C. *Modélisation des dégradations de composants passifs par processus stochastiques et prise en compte des incertitudes.* PhD thesis, 2008.

[10] Carpenter, G. A., Grossberg, S., and Rosen, D. B. Fuzzy art : Fast stable learning and categorization of analog patterns by an adaptive resonance system. *Neural networks 4*, 6 (1991), 759–771.

[11] Celeux, G., and Govaert, G. A classification em algorithm for clustering and two stochastic versions. *Computational statistics & Data analysis 14*, 3 (1992), 315–332.

[12] Cinlar, E., Osman, E., and Bazant, Z. P. Stochastic process for extrapolating concrete creep. *Journal of the Engineering Mechanics Division 103*, 6 (1977), 1069–1088.

[13] Cover, T. M. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *Electronic Computers, IEEE Transactions on*, 3 (1965), 326–334.

[14] Cox, D. Some remarks on failure-times, surrogate markers, degradation, wear, and the quality of life. *Lifetime Data Analysis 5*, 4 (1999), 307–314.

[15] De Berg, M., Cheong, O., Van Kreveld, M., and Overmars, M. *Computational geometry : algorithms and applications*. Springer, 2008.

[16] Dempster, A. P., Laird, N. M., and Rubin, D. B. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* (1977), 1–38.

[17] Duda, R., and Hart, P. *Pattern classification and scene analysis*. J. Wiley and Sons, New York, 1973.

[18] Duda, R. O., Hart, P. E., and Stork, D. G. Pattern classification and scene analysis 2nd ed.

[19] Eldershaw, C., and Hegland, M. Cluster analysis using triangulation. *Computational Techniques and Applications : CTAC97* (1997), 201–208.

[20] Ellingwood, B. R., and Mori, Y. Probabilistic methods for condition assessment and life prediction of concrete structures in nuclear power plants. *Nuclear Engineering and Design 142*, 2 (1993), 155–166.

[21] Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data mining* (1996), vol. 1996, AAAI Press, pp. 226–231.

[22] Estivill-Castro, V., and Lee, I. Autoclust : Automatic clustering via boundary extraction for mining massive point-data sets. In *In Proceedings of the 5th International Conference on Geocomputation* (2000), Citeseer.

[23] Everitt, B., and Hothorn, T. Cluster analysis. In *An Introduction to Applied Multivariate Analysis with R*. Springer, 2011, pp. 163–200.

[24] Filippone, M., Camastra, F., Masulli, F., and Rovetta, S. A survey of kernel and spectral methods for clustering. *Pattern recognition 41*, 1 (2008), 176–190.

[25] Fowlkes, E. B., and Mallows, C. L. A method for comparing two hierarchical clusterings. *Journal of the American statistical association 78*, 383 (1983), 553–569.

[26] Frangopol, D. M., Kallen, M.-J., and Noortwijk, J. M. v. Probabilistic models for life-cycle performance of deteriorating structures : review and future directions. *Progress in Structural Engineering and Materials 6*, 4 (2004), 197–212.

[27] Gallager, R. G., Humblet, P. A., and Spira, P. M. A distributed algorithm for minimum-weight spanning trees. *ACM Transactions on Programming Languages and systems (TOPLAS) 5*, 1 (1983), 66–77.

[28] Geman, S., and Geman, D. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 6 (1984), 721–741.

[29] Govaert, G. *Analyse des données*. Lavoisier, 2003.

[30] Guha, S., Rastogi, R., and Shim, K. Cure : an efficient clustering algorithm for large databases. In *ACM SIGMOD Record* (1998), vol. 27, ACM, pp. 73–84.

[31] Guha, S., Rastogi, R., and Shim, K. Rock : A robust clustering algorithm for categorical attributes. *Information systems 25*, 5 (2000), 345–366.

[32] Hand, D. J. Discrimination and classification. *Wiley Series in Probability and Mathematical Statistics, Chichester : Wiley, 1981 1* (1981).

[33] Hathaway, R. J. Another interpretation of the em algorithm for mixture distributions. *Statistics & Probability Letters 4*, 2 (1986), 53–56.

[34] Haykin, S. S. *Neural networks : a comprehensive foundation*. Prentice Hall Englewood Cliffs, NJ, 2007.

[35] Hu, T., and Sung, S. A hybrid EM approach to spatial clustering. *Computational statistics & data analysis 50*, 5 (2006), 1188–1205.

[36] Jain, A. K. Data clustering : 50 years beyond k-means. *Pattern Recognition Letters 31*, 8 (2010), 651–666.

[37] Jain, A. K., and Dubes, R. C. *Algorithms for clustering data*. Prentice-Hall, Inc., 1988.

[38] Jain, A. K., Duin, R. P. W., and Mao, J. Statistical pattern recognition : A review. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 22*, 1 (2000), 4–37.

[39] Jain, A. K., Murty, M. N., and Flynn, P. J. Data clustering : a review. *ACM computing surveys (CSUR) 31*, 3 (1999), 264–323.

[40] Jia, K., Zhao, R., Zhong, J., and Liu, X. Preparation and microwave absorption properties of loose nanoscale fe¡ sub¿ 3¡/sub¿ o¡ sub¿ 4¡/sub¿ spheres. *Journal of Magnetism and Magnetic Materials 322*, 15 (2010), 2167–2171.

[41] Karypis, G., Han, E.-H., and Kumar, V. Chameleon : Hierarchical clustering using dynamic modeling. *Computer 32*, 8 (1999), 68–75.

[42] Kaufman, L., and Rousseeuw, P. Clustering by means of medoids.

[43] Kleinberg, J. An impossibility theorem for clustering. *Advances in neural information processing systems* (2003), 463–470.

[44] Krishnan, T., and McLachlan, G. The em algorithm and extensions, 1997.

[45] Kruskal, J. B. On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical society 7*, 1 (1956), 48–50.

[46] Lance, G. N., and Williams, W. T. A general theory of classificatory sorting strategies 1. hierarchical systems. *The computer journal 9*, 4 (1967), 373–380.

[47] Law, M. H., Topchy, A. P., and Jain, A. K. Model-based clustering with probabilistic constraints. In *SDM* (2005).

[48] Lawless, J., and Crowder, M. Covariates and random effects in a gamma process model with application to degradation and failure. *Lifetime Data Analysis 10*, 3 (2004), 213–227.

[49] LeSage, J. Matlab toolbox for spatial econometrics, 1999.

[50] Liu, C. L. *Introduction to combinatorial mathematics*, vol. 181. McGraw-Hill New York, 1968.

[51] Liu, Q., Deng, M., Shi, Y., and Wang, J. A density-based spatial clustering algorithm considering both spatial proximity and attribute similarity. *Computers & Geosciences* (2012).

[52] Liu, Y., Li, Z., Xiong, H., Gao, X., and Wu, J. Understanding of internal clustering validation measures. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on* (2010), IEEE, pp. 911–916.

[53] Livny, T. Z. R. R. M. Birch : an efficient data clustering method for very large databases. In *ACM SIGMOD international Conference on Management of Data* (1996), vol. 1, pp. 103–114.

[54] Makhlouf, K., and Jones, J. Effects of temperature and frequency on fatigue crack growth in 18 % cr ferritic stainless steel. *International journal of fatigue 15*, 3 (1993), 163–171.

[55] Matula, D. W., and Sokal, R. R. Properties of gabriel graphs relevant to geographic variation research and the clustering of points in the plane. *Geographical analysis 12*, 3 (1980), 205–222.

[56] McLachlan, G., and Peel, D. *Finite mixture models*, vol. 299. Wiley-Interscience, 2000.

[57] McLachlan, G. J. *Discriminant analysis and statistical pattern recognition*, vol. 544. Wiley-Interscience, 2004.

[58] Muller, K.-R., Mika, S., Ratsch, G., Tsuda, K., and Scholkopf, B. An introduction to kernel-based learning algorithms. *Neural Networks, IEEE Transactions on 12*, 2 (2001), 181–201.

[59] Oliver, M., and Webster, R. A geostatistical basis for spatial weighting in multivariate classification. *Mathematical Geology 21*, 1 (1989), 15–35.

[60] Pena, J. M., Lozano, J. A., and Larranaga, P. An empirical comparison of four initialization methods for the¡ i¿ k¡/i¿-means algorithm. *Pattern recognition letters 20*, 10 (1999), 1027–1040.

[61] Potamianos, G., and Goutsias, J. A novel method for computing the partition function of markov random field images using monte carlo simulations. In *Acoustics, Speech, and Signal Processing, 1991. ICASSP-91., 1991 International Conference on* (1991), IEEE, pp. 2325–2328.

[62] Prim, R. C. Shortest connection networks and some generalizations. *Bell system technical journal 36*, 6 (1957), 1389–1401.

[63] Roussignol, M. Gamma stochastic process and application to maintenance.

[64] Schölkopf, B., Smola, A., and Müller, K.-R. Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation 10*, 5 (1998), 1299–1319.

[65] Schölkopf, B., and Smola, A. J. *Learning with kernels : support vector machines, regularization, optimization and beyond*. the MIT Press, 2002.

[66] Scott, A., and Symons, M. Clustering methods based on likelihood ratio criteria. *Biometrics* (1971), 387–397.

[67] Shental, N., Bar-Hillel, A., Hertz, T., and Weinshall, D. Computing gaussian mixture models with em using side-information. In *Proc. of workshop on the continuum from labeled to unlabeled data in machine learning and data mining* (2003).

[68] Sibson, R. Slink : an optimally efficient algorithm for the single-link cluster method. *The Computer Journal 16*, 1 (1973), 30–34.

[69] Singpurwalla, N. D. Survival in dynamic environments. *Statistical Science 10*, 1 (1995), 86–103.

[70] Sørensen, T. {A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons}. *Biol. skr. 5* (1948), 1–34.

[71] Spath, H. *Cluster analysis algorithms for data reduction and classification of objects*. Ellis Horwood, Ltd., 1980.

[72] Strauss, D. J. Clustering on coloured lattices. *Journal of Applied Probability* (1977), 135–143.

[73] Symons, M. J. Clustering criteria and multivariate normal mixtures. *Biometrics* (1981), 35–43.

[74] Theodoridis, S., and Koutroumbas, K. Pattern recognition, 1999.

[75] Tibshirani, R., Walther, G., and Hastie, T. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society : Series B (Statistical Methodology) 63*, 2 (2001), 411–423.

[76] Trauwaert, E., Kaufman, L., and Rousseeuw, P. Fuzzy clustering algorithms based on the maximum likelihood priciple. *Fuzzy Sets and Systems 42*, 2 (1991), 213–227.

[77] Van Noortwijk, J. A survey of the application of gamma processes in maintenance. *Reliability Engineering & System Safety 94*, 1 (2009), 2–21.

[78] Vapnik, V. *The nature of statistical learning theory*. springer, 1999.

[79] Wainwright, M., Jaakkola, T., and Willsky, A. A new class of upper bounds on the log partition function. *Information Theory, IEEE Transactions on 51*, 7 (2005), 2313–2335.

[80] West, D. B., et al. *Introduction to graph theory*, vol. 2. Prentice hall Upper Saddle River, NJ. :, 2001.

[81] Whitmore, G. Estimating degradation by a wiener diffusion process subject to measurement error. *Lifetime data analysis 1*, 3 (1995), 307–319.

[82] Whitmore, G., Crowder, M., and Lawless, J. Failure inference from a marker process based on a bivariate wiener model. *Lifetime Data Analysis 4*, 3 (1998), 229–251.

[83] Whitmore, G., and Schenkelberg, F. Modelling accelerated degradation data using wiener diffusion with a time scale transformation. *Lifetime Data Analysis 3*, 1 (1997), 27–45.

[84] Wu, J., Xiong, H., and Chen, J. Adapting the right measures for k-means clustering. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (2009), ACM, pp. 877–886.

[85] Xu, R., Wunsch, D., et al. Survey of clustering algorithms. *Neural Networks, IEEE Transactions on 16*, 3 (2005), 645–678.

[86] Yang, Y., and Klutke, G.-A. Lifetime-characteristics and inspection-schemes for levy degradation processes. *Reliability, IEEE Transactions on 49*, 4 (2000), 377–382.

[87] Zadeh, L. Fuzzy sets. *Information & Control 8* (1965), 338–353.

[88] Zahn, C. T. Graph-theoretical methods for detecting and describing gestalt clusters. *Computers, IEEE Transactions on 100*, 1 (1971), 68–86.

[89] Zhong, C., Miao, D., and Wang, R. A graph-theoretical clustering method based on two rounds of minimum spanning trees. *Pattern Recognition 43*, 3 (2010), 752–766.

Xuanzhou WANG

Doctorat : Optimisation et Sûreté des Systèmes

Année 2013

Détermination de classes de modalités de dégradation significatives pour le pronostic et la maintenance

Les travaux présentés dans ce manuscrit traitent de la détermination de classes de systèmes selon leur mode de vieillissement dans l'objectif de prévenir une défaillance et de prendre une décision de maintenance. L'évolution du niveau de dégradation observée sur un système peut être modélisée par un processus stochastique paramétré. Un modèle usuellement utilisé est le processus Gamma. On s'intéresse au cas où tous les systèmes ne vieillissent pas identiquement et le mode de vieillissement est dépendant du contexte d'utilisation des systèmes ou des propriétés des systèmes, appelé ensemble de covariables. Il s'agit alors de regrouper les systèmes vieillissant de façon analogue en tenant compte de la covariable et d'identifier les paramètres du modèle associé à chacune des classes.

Dans un premier temps la problématique est explicitée avec notamment la définition des contraintes: incréments d'instants d'observation irréguliers, nombre quelconque d'observations par chemin décrivant une évolution, prise en compte de la covariable. Ensuite des méthodes sont proposées. Elles combinent un critère de vraisemblance dans l'espace des incréments de mesure du niveau de dégradation, et un critère de cohérence dans l'espace de la covariable. Une technique de normalisation est introduite afin de contrôler l'importance de chacun de ces critères. Des études expérimentales sont effectuées pour illustrer l'efficacité des méthodes proposées.

Mots clés : classification - statistique mathématique - fiabilité, méthodes statistiques - processus stochastiques – apprentissage automatique.

Determination of Classes of Significant Deterioration Modalities for Prognosis and Maintenance

The work presented in this thesis deals with the problem of determination of classes of systems according to their aging mode in the aim of preventing a failure and making a decision of maintenance. The evolution of the observed deterioration levels of a system can be modeled by a parameterized stochastic process. A commonly used model is the Gamma process. We are interested in the case where all the systems do not age identically and the aging mode depends on the condition of usage of systems or system properties, called the set of covariates. Then, we aims to group the systems that age similarly by taking into account the covariate and to identify the parameters of the model associated with each class.

At first, the problem is presented especially with the definition of constraints: time increments of irregular observations, any number of observations per path which describes an evolution, consideration of the covariate. Then the methods are proposed. They combine a likelihood criterion in the space of the increments of deterioration levels, and a coherence criterion in the space of the covariate. A normalization technique is introduced to control the importance of each of these two criteria. Experimental studies are performed to illustrate the effectiveness of the proposed methods.

Keywords: classification - mathematical statistics - reliability (engineering), statistical methods - stochastic processes – machine learning.

Thèse réalisée en partenariat entre :

RÉGION CHAMPAGNE ARDENNE

utt université de technologie Troyes

Ecole Doctorale "Sciences et Technologies"