Modélisation et classification des données binaires en grande dimension : application à l'autopsie verbale

Seydou Nourou Sylla

UNIVERSITE GASTON BERGER DE SAINT-LOUIS

ECOLE DOCTORALE DES SCIENCES ET DES TECHNOLOGIES

U.F.R DE SCIENCES APPLIQUEES ET DE TECHNOLOGIES

Laboratoire d'**E**tudes et de **R**echerches en **S**tatistique et **D**éveloppement

THÈSE

pour obtenir le grade de

DOCTEUR EN MATHÉMATIQUES APPLIQUÉES
DE L'UNIVERSITÉ GASTON BERGER

Spécialité : Statistique Appliquée

présentée par

SEYDOU NOUROU SYLLA

Modélisation et classification des données binaires en grande dimension: application a l'autopsie verbale

soutenue publiquement 21 Décembre 2016, devant le jury ci-dessous :

PRESIDENT DU JURY	**Aboubakary Diakhaby**	Professeur à Université Gaston Berger
RAPPORTEURS	**Christophe Biernacki**	Professeur à l'Université Lille 1
	Ahmadou Alioum	Professeur à l'Université de Bordeaux
	Papa Ngom	Professeur à l'Université Cheikh Anta Diop de Dakar
EXAMINATEURS	**Aliou Diop**	Professeur à Université Gaston Berger
	Ali Souleymane Dabye	Professeur à Université Gaston Berger
	Stéphane Girard	Directeur de recherche à INRIA-Grenoble
DIRECTEUR DE THESE	**Abdou kâ Diongue**	Professeur à Université Gaston Berger

DÉDICACES

Je dédie ce travail à feue ma mère la Sainte Mariama WELLE, et à feue ma Grand- mère Khadidiatou SALL.

REMERCIEMENTS

Rendons grâce à DIEU et à son prophète Muhamed(PSL)

Ce travail a été réalisé grâce à la collaboration entre trois équipes de recherche:

- URMITE-IRD(Unité de Recherche sur les Maladies Infectieuses Tropicales et Emergents)
- MISTIS-INRIA(Modélisation et Inférence de Systèmes aléatoires complexes et structurés)
- LERSTAD-UGB(Laboratoire d'Études et de Recherche en Statistiques et Développement)

Cher professeur Stéphane Girard, j'ai été satisfait dés notre premier rencontre au sein de l'INRIA de Grenoble. J'ai beaucoup apprécié votre simplicité et vos connaissances. J'ai été particulièrement touché par votre sympathie et votre disponibilité à répondre à toutes mes préoccupations. Vos qualités scientifiques et humaines, suscitent en nous une admiration et un profond respect. Je suis heureux de l'honneur que vous m'avait fait en m'accompagnant dans ce travail. C'est pourquoi j'ai une entière reconnaissance pour la confiance qui vous m'avait accordé. Vous avez amplement mérité la bénédiction de mon grand-père(paroles de sénégalais).

Cher professeur Abdou Kâ Diongue, vous avez été au début et à la fin de ce travail malgré vos nombreuses occupations. Vos contributions dans la réalisation de ce travail ont été sans commune mesure. Merci pour le soutien sans faille.

Je tiens à remercier le Docteur Aldiouma Diallo et le Dr Cheikh Sokhna, Responsable du laboratoire URMITE-Dakar, pour avoir accepté de m'accueillir dans leur prestigieuse institution

en vue de la réalisation de ce travail. J'ai le sentiment d'avoir bénéficié d'un privilège à travers l'immense contribution matérielle, votre personnelle et vos différents services. Merci infiniment.

Je tiens à exprimer mes sincères remerciements à Monsieur Aboubakry Diakhaby, Professeur à Université Gaston Berger, pour l'honneur qu'il m'a fait en présidant ce jury. J'adresse ma vive reconnaissance à Monsieur Christophe Biernacki, Professeur à l'Université Lille 1 ,Monsieur Ahmadou Alioum, Professeur à l'Université de Bordeaux et Monsieur Papa Ngom, Professeur à l'Université Cheikh Anta Diop de Dakar pour l'intérêt qu'ils ont porté à mon travail en acceptant de rapporter ce manuscrit et d'avoir examiné minutieusement ce travail.

Je remercie Monsieur Souleymane Daby, Professeur à Université Gaston Berger et Monsieur Aliou Diop, Professeur à Université Gaston Berger, pour m'avoir fait l'honneur d'accepter de participer au jury en tant que examinateurs.

Un grand merci à Florence Forbes et à tous les membres de l'équipe MISTIS (Gildas, Alessandro, Brice, Clément, Thomas, Jean-Baptiste, Pablo, Aina, Alexis,...), à coté de vous j'ai beaucoup appris. Vos conseils, votre aide et votre soutien moral ne m'ont jamais manqué. J'espère que la théorie de mon grand-père vous serait bénéfique et mes meilleurs sentiments au baby-foot. Tous mes remerciements et ma profonde gratitude.

L'accès aux données qui ont été exploitées dans ce document a été facilité par la disponibilité remarquée l'Antenne IRD de Dakar. Je remercie toute l'équipe de l'IRD : Emilie Ndiaye, Paul Senghor, Prosper Ndiaye, Bruno Senghor, Djibril Dione, Ousmane Ndiaye, Pape Niokhor Diouf, Antoine Ndour, Mouhamadou Baba Sow et tous les autres.

Merci à tous mes professeurs de l'UFR SAT qui ont participé à ma formation, j'ai tant appris à vos coté.

Mention spécial à Oumy Niass, Papa Birahima Fall et Amadou Thiam, vos conseils et soutien ont été d'une importance capitale.

Mes remerciements vont à l'endroit de mon père et de ma tendre et très chère mère, pour la beauté de la vie, qu'ils nous ont fait découvrir jour après jour, à travers le langage de la spontanéité, le geste de la tendresse, la pureté de leurs regards, le merveilleux de leur présence.

Mention très spécial à ma femme qui est pour moi un rayon de soleil, une lumière dans ma vie.

Mention très spécial à ceux qui m'ont vu grandir, mes oncles Abdoulaye Omar Sall, Ahmed

Tidiane Sall et toute la famille Sall de Sam-Kaolack pour leur soutien sans faille à mon égard. Je remercie du fond du cœur mes familles d'accueil, la famille Sall de la Patte d'oies Bulders à Dakar et la famille Bâ de Diawling à Saint-Louis.

Je remercie du fond du cœur mes frères, ma sœur et toute la famille Sylla pour tout l'amour et le soutien qu'ils m'ont toujours apporté.

Merci à mes amis, Assane Seck, Moussa Seck, Mountaga Sall, Papa Ousmane Cissé, Pape Mbissane Faye, Ismaila Ndiaye,Madior Niang, Mansour Diop pour la route parcourue ensemble dans l'amour, la fidélité et l'espérance.

Merci à la famille de Mame Abdoul Aziz SY de Tivaouane, la famille Kanéne de Kaolack pour leurs prières et bénédictions.

A toutes ces personnes que j'ai omis imprudemment de mentionner, sachez que sans vous, ce modeste travail ne verrait pas le jour.

RÉSUMÉ

Résumé

Le manque de données fiables sur les niveaux et les causes de mortalité constitue encore un frein au développement dans les pays défavorisés. Dans ces pays, il n'est pas toujours facile d'obtenir des informations fiables sur la morbidité et la mortalité. L'autopsie verbale est devenue la principale source d'information sur les causes de décès dans ces localités. Cette méthode s'appuie sur des questionnaires structurés de manière à déterminer la symptomatologie et à obtenir des informations sur la cause probable du décès. Ces données collectées conduisent à l'élaboration de méthodes dites d'aide aux diagnostics qui reposent souvent sur les méthodes de classification.

La problématique porte sur l'élaboration d'une méthode de diagnostic automatique à partir des données d'enquête. L'objectif est d'obtenir des diagnostics en prenant en compte la présence ou l'absence de symptômes et des variables socio-démographiques. Il repose sur la construction de modèles de discrimination à partir de données multi-classes avec un nombre important de variables explicatives à caractère binaire.

Une partie de ce travail de thèse porte sur l'utilisation d'un modèle de mélange sous l'hypothèse d'indépendance conditionnelle et sur des techniques de réduction de la dimensionnalité. Le caractère binaire des réponses suppose des méthodes reposant sur les mesures de similarité. Ainsi, une généralisation de plusieurs mesures de similarités et de dissimilarités est exposée dans cette thèse. Nous avons également présenté une technique de construction de noyaux pour la classification à partir d'une mesure de similarité. La seconde partie de cette thèse présente une méthode de classification combinant à la fois les mesures de similarités et les modèles de mélange. La structure hiérarchique des questions posées lors de l'entretien et de leurs interactions nous a permis de définir une structure sur les données. Ainsi pour mieux prendre en compte cette structure, nous avons présenté lors de nos travaux un noyau hiérarchique avec effet d'interactions entre les variables. Ce noyau combine à la fois une structure hiérarchique des variables suivant un arborescence à deux niveaux et l'interaction de leurs sous variables jusqu'à un certain ordre fixé.

Mots clés: Autopsie verbale, données binaires, similarité, noyau, grande dimension, classification

Abstract

The lack of reliable data about the causes of mortality still constitutes an obstacle for the development of poor regions in the world. In these countries, it is not always easy to obtain reliable information about morbidity and mortality.

Verbal autopsy has become the main source of information about the causes of death in many places. This method is based on structured questionnaires to determine the symptoms and to get information about the possible cause of death. These data lead to the development of diagnosis assistance systems which are often based on classification methods.

The problem we tackle is the development of a method for automatic diagnosis using survey data. The final objective is to provide a diagnosis by taking into account the presence or absence of symptoms and sociodemographic variables. This approach is based on the construction of discrimination models from multi-class data with a large number of explanatory variables of binary nature.

The first part of this thesis uses a mixture model under the assumption of conditional independence together with dimensionality reduction techniques. The binary nature of the answers requires methods based on similarity measures. Thus, a generalization of several measures of similarity and dissimilarity is exposed in this thesis. Since kernels are of great importance in classification, we also present a kernel construction technique from a similarity measure. The second part of this thesis presents a classification method combining both similarity measures and mixture models. The hierarchical structure of the questions asked during the interview and their interactions allows us to define a structure over the data. To better take into account this structure, we present a hierarchical kernel that takes into account the interactions between variables. This kernel combines a hierarchical structure for the variables with a tree structure with two levels and interaction of variables up to a certain order.

Mots clés: Verbal autopsy, binary data, similarity, kernel, high dimension, classification.

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLES

INTRODUCTION

0.1 Contexte et Problématique

0.1.1 Contexte

Le manque de données fiables sur les niveaux et les causes de mortalité dans les régions défavorisées de la planète constitue encore un frein aux efforts déployés pour établir une base de données solide sur laquelle appuyer la politique, la planification, la surveillance et l'évaluation sanitaire. Les programmes de santé publique doivent s'appuyer sur des données fiables pour pouvoir identifier l'ampleur, la gravité des problèmes de santé afin d'y apporter des solutions et proposer des mécanismes adaptés pour leur évaluation. Cependant, dans les pays en voie de développement, il n'est pas toujours facile d'obtenir des informations fiables sur la morbidité et la mortalité.
Dans ces pays, particulièrement en Afrique, on ne dispose pas souvent de renseignements exacts sur la mortalité. De plus, les informations sur les causes de décès sont rares dans les pays à forte mortalité, particulièrement en zone rurale avec une couverture médicale faible. Une meilleure connaissance de l'importance relative des différentes maladies est primordiale pour définir et orienter les programmes de santé. La surveillance démographique est une solution adéquate et précieuse d'où la création des sites de suivi démographiques et l'application d'une méthode dite d'Autopsie Verbale (AV).

L'autopsie verbale est devenue la principale source d'information sur les causes de décès

dans des localités pour lesquelles il n'existe ni système d'état civil ni certificat médical de décès. Une autopsie verbale est une méthode qui consiste à interroger les membres de la famille ou les aidants de la personne décédée. Elle s'appuie sur des questionnaires structurés de manière à déterminer la symptomatologie et à obtenir d'autres informations permettant d'établir la cause probable du décès. Ces fiches d'enquête sont remises aux médecins. Ces médecins, à partir des symptômes et des variables socio-démographiques recueillis, donnent leurs diagnostics.

Cependant, le domaine médical nécessite le recueil de nombreuses données lors de l'examen du patient ou lors de l'entretien dans le cas d'une autopsie verbale. Ces données conduisent à l'élaboration de méthodes dites d'aide aux diagnostics. L'aide au diagnostic se développe de plus en plus et gagne en popularité. Il existe de nombreuses applications qui permettent d'assister les médecins et spécialistes dans leur démarche de prises de décision. Le système de l'aide au diagnostic repose souvent sur les méthodes de classification ou de reconnaissance de forme.

Ainsi, l'évaluation des causes de décès dans les milieux ruraux conduit à une classification des individus suivants les diagnostics définis a priori sur la population étudiée. Les données fournies permettent d'appliquer des méthodes décisionnelles pour discriminer les individus décédés en rapport avec les symptômes et variables socio-démographiques déclarés. De plus, l'objectif du diagnostic médical est entre autre d'obtenir un modèle regroupant les individus présentant des caractéristiques similaires. Ce regroupement s'obtient par la classification des données sur la présence et l'absence de symptômes. Les méthodes de classification, qui ont pour but de construire des partitions d'ensemble d'objets décrits par des variables de telle sorte que les classes soient les plus homogènes possible, semblent être une piste importante de recherche sur les données issues des enquêtes par autopsie verbale. A partir des données, l'apprentissage statistique permet une résolution automatique du problème sur la base d'une prise de décision faite à partir de ces observations. Après une phase d'apprentissage sur les diagnostics déjà établis par des médecins, les méthodes de classification permettent une automatisation de ces diagnostics.

0.1.2 Problématique

La problématique porte sur l'élaboration d'une méthode de diagnostic automatique à partir des données d'autopsie verbale via des méthodes statistiques. L'objectif des médecins est d'obtenir

des diagnostics en prenant en compte la présence ou l'absence de symptômes et des variables socio-démographiques. De ce fait, la problématique repose sur l'adaptation des méthodes statistiques à l'analyse des causes de décès. Cette adaptation se fait grâce à l'élaboration de nouvelles méthodes de classification basées sur les modèles de mélanges et sur l'introduction de mesures de similarité et de noyaux adaptés à des données binaires issues des autopsies verbales en milieu rural.

0.2 Contributions

La mise en œuvre de nos travaux s'est heurtée à plusieurs difficultés, dues en partie aux particularités du milieu médical et à la qualité des données. En effet, en menant plusieurs études, les enquêteurs ont pu recueillir un nombre important de variables sur un ensemble de patients qui sont décédés.

Le nombre d'observations (dans notre cas le nombre important de causes, de symptômes et d'individus) est essentiel car il influence considérablement la précision des performances de discrimination. Dès lors, l'un des enjeux de cette thèse est de construire des modèles de discrimination efficaces à partir de données multi-classes avec un nombre important de variables explicatives.

Les données dont on dispose sont multidimensionnelles et multi-classes et ne peuvent pas être modélisées par une distribution de probabilité classique. Ainsi, la structure des données permet de supposer l'existence de plusieurs sous-populations, chaque sous-population pouvant être modélisée de manière séparée. Le modèle de mélange est un outil pour modéliser ce type de données en s'appuyant sur des lois simples comme la loi multinomiale, normale,...
L'approche mélange permet de formaliser la notion de classes. L'existence de groupes de diagnostics hétérogènes dans la structure de nos données nous pousse ainsi à l'utilisation des modèles de mélange.

Le choix de la nature des variables est important en classification. Ce choix influe sur les modèles utilisés et sur les résultats de l'analyse. En général, les données s'expriment dans un tableau individus-variables. Les données sont souvent de deux types (qualitatives ou quantitatives). Dans certains cas, elles peuvent être mixtes ou hétérogènes. L'aspect présence ou absence des symptômes, nous pousse vers le choix de modèles à variables qualitatives. En

outre le caractère qualitatif des données et leur aspect multi-classes permettent d'envisager des méthodes à noyaux basés sur les mesures de similarités.

Les travaux effectués ont donné lieu à plusieurs contributions.

Méthode de classification supervisée par modèle de mélange

Une des contributions de cette thèse est l'utilisation d'un modèle de mélange sous l'hypothèse d'indépendance conditionnelle. Une étude sur les modèles de mélange avec des lois multinomiales est utilisée pour définir des classifieurs probabilistes. Le nombre important de paramètres à estimer dans les modèles multinomiaux nous impose l'utilisation de l'hypothèse d'indépendance conditionnelle sur les variables pour réduire au mieux le nombre de paramètres à estimer. La sélection des variables les plus significatives et la réduction du nombre de classes sont abordées afin de prédire l'appartenance des individus aux groupes a priori.

Une généralisation des mesures de similarités

Une multitude de mesures de similarité sont proposées dans la littérature. L'une des contributions de cette thèse est de proposer une mesure qui permet d'unifier plusieurs mesures de similarité. Cette unification permet de retrouver suivant les valeurs des paramètres proposés certaines mesures de similarité. Les grandes familles des mesures de Tversky et Baulieu s'avèrent être des cas particuliers de notre généralisation.

Une méthode de classification combinant mesure de similarité et modèles de mélanges

Les noyaux jouent un rôle essentiel en classification. De ce fait, dans nos travaux nous avons présenté une technique de construction de noyaux à partir d'une mesure de similarité. Ainsi, nous avons présenté une méthode de classification combinant à la fois les mesures de similarités et les modèles de mélange dans un cadre gaussien ou non.

Introduction d'un noyau hiérarchique

Afin de prendre en compte la spécificité des données, nous proposons un noyau hiérarchique. Ce noyau prend en compte la structure hiérarchique des variables. La structure hiérarchique des symptômes et de leurs interactions nous a permis dans un second temps de définir une structuration des symptômes et un nouveau noyau permettant de prendre en compte ces deux aspects.

0.3 Organisation de la thèse

Ce manuscrit est structuré en deux parties : la première partie concerne l'état de l'art et la seconde partie décrit nos contributions.

- Au chapitre 1, nous proposerons de contextualiser la recherche des causes de mortalité par autopsie verbale, de définir la méthodologie et les outils utilisés. Une étude rétrospective des méthodes statistiques utilisées dans d'autres sites sera proposée.
- Au chapitre 2, nous exposerons les méthodes classiques en analyse discriminante en passant par les modèles de mélange. Dans ce chapitre, on s'intéresse aussi à l'étude des méthodes à noyaux, des mesures de similarité et des méthodes hiérarchiques.
- Au chapitre 3, la structure des variables (liée au protocole expérimental: variables mesurées sur plusieurs classes) rend l'utilisation des modèles de mélange efficace. Ainsi, on fera l'étude de deux méthodes basées sur des modèles de mélange sous l'hypothèse d'indépendance conditionnelle: la sélection de variables pertinentes et une étude sur la réduction du nombre de classes.
- Au cours du chapitre 4, une étude combinant à la fois des modèles de mélange et des mesures de similarité sera abordée. Cette partie présentera des résultats sur des données autres que les autopsies verbales.
- Après avoir défini des noyaux sur des données binaires, nous nous focaliserons au chapitre 5 sur la structure hiérarchique des variables et sur leurs interactions. Nous évaluerons les méthodes hiérarchiques en définissant un nouveau noyau qui tient compte de la structure des variables et de leurs interactions jusqu'à un certain ordre.

- On conclura ces travaux et exposera ceux en cours ainsi que les perspectives de recherche dans le chapitre 6.

CHAPITRE 1

ETAT DE L'ART: AUTOPSIE VERBALE

Sommaire

1.1 Introduction

La mortalité reste très élevée dans nombre de pays africains au sud du Sahara, et particulièrement en milieu rural. Il n'est pas toujours facile d'obtenir des informations fiables sur la morbidité et la mortalité. Ce manque d'informations sur les causes de décès constitue un frein à l'élaboration d'une base de données solide sur laquelle peut s'appuyer la politique, la planification, la surveillance et l'évaluation sanitaire [2]. Ainsi, une meilleure connaissance des causes de décès permettrait, d'une part l'évaluation de l'impact des programmes dirigés qui visent à réduire la mortalité, et d'autre part l'allocation de ressources dans ces domaines.

L'application d'une méthode – dite d'autopsie verbale permet de disposer des causes probables de décès. L'autopsie verbale est devenue la principale source d'information sur les causes

de décès dans les populations pour lesquelles il n'existe ni système d'état civil, ni certificat médical de décès.

Dans ce chapitre, nous élaborerons un état de l'art des autopsies verbales. Dans la partie 1.2, nous définirons le concept de l'autopsie verbale. Nous énoncerons dans la partie 1.3 les caractéristiques et les techniques d'enquêtes associées. Dans la partie 1.4, nous exposerons la méthodologie des études par autopsie verbale et dans la partie 1.5 nous établirons une synthèse des études statistiques menées dans le cadre de cette problématique.

1.2 Autopsie Verbale

Autopsie médicale: Les mots "Autopsie "ou "examen post mortem "ou "nécropsie "désignent l'examen médical des cadavres. L'objectif de l'autopsie médicale est d'établir la cause de la mort (cause principale et causes indirectes s'il y a lieu), de déterminer l'état de santé du sujet avant son décès, et si les éventuels traitements reçus étaient adaptés. De nos jours, l'allongement de l'espérance de vie et les polypathologies font que, le plus souvent, un individu meurt de suite de plusieurs maladies. Dans ce cas, la cause de la mort n'est pas toujours évidente et l'autopsie intervient pour établir la réalité des faits, si elle est pratiquée dans les délais et conditions requis.

Autopsie verbale: L'autopsie verbale consiste à interroger les membres d'un ménage sur les circonstances du décès de l'un des leurs, survenu antérieurement, afin d'en déterminer les causes.

1.2.1 Historique de la méthode d'autopsie verbale

En l'absence de sources institutionnelles d'informations, l'amélioration des moyens d'investigation des problèmes de santé, dont, en premier lieu, la détermination des causes médicales de décès, constitue un enjeu important pour la recherche [3].
Dans [4], Biraud propose en 1954 d'utiliser les symptômes afin de faciliter l'enregistrement et l'estimation des causes possibles de décès.
Il suggère de former un personnel non médical pour enregistrer au moins :

- le sexe et l'âge de la personne,

- les circonstances du décès (accident, mort violente, maladie),
- les principaux symptômes, leurs sièges et leur durée,
- les affections épidémiques.

Le but est d'obtenir des diagnostics communautaires que pourraient recueillir les fonctionnaires en place ayant une connaissance préalable du pays, de ses coutumes et de ses pathologies. Il estime que d'importantes informations sur les causes de décès pourraient être collectées par cette voie.

En 1978, l'Organisation Mondiale pour la Sante (OMS) [5] reprend l'idée d'un système simplifié de collecte des données sur la mortalité, en proposant de classer les décès selon une cause unique, à partir des informations obtenues sur les symptômes de la maladie au cours d'un entretien non directif avec l'entourage de la personne décédée.

Une expérience d'enregistrement des causes de décès par des non médecins selon ce principe a été conduite à Matlab au Bengladesh en 1975 [6]. Il s'avère qu'au cours de cette expérience que le système d'entretien non directif favorise les biais d'enquête et les pertes d'informations. Le déroulement de l'entretien, les questions posées et les conclusions influençent énormément le bon sens de l'enquêteur et aussi sa délibération en faveur du diagnostic pour certaines maladies plutôt que d'autres [3]. A cela s'ajoute un risque de confusion entre concepts traditionnels et concepts médicaux.

Pour minimiser ces biais, il apparaît nécessaire de développer une procédure standardisée, surtout pour les chercheurs associés à des programmes de santé maternelle et infantile en zone rurale. Une nouvelle méthode de collecte est créée à partir de questionnaires standardisés, très structurés, portant sur la description des symptômes développés lors de la maladie et de leur succession dans le temps [6].

1.2.2 Mode d'utilisation

L'autopsie verbale permet d'analyser les facteurs médicaux et non médicaux dans la suite des événements ayant conduit au décès. Elle est donc une méthode visant à élucider les causes médicales du décès et mettre à jour les facteurs personnels, familiaux et communautaires susceptibles d'avoir contribué au décès, lorsque celui-ci est survenu en dehors d'un établissement de

santé. Elle permet aussi de recenser les facteurs liés à des décès survenus dans une communauté (source précieuse de renseignements sur les causes de mortalité). Son utilisation est encouragée, dans le cadre d'une base de données démographique, lorsque les décès sont nombreux en dehors des établissements de soins.

Cette technique permet d'assigner avec un minimum de risque d'erreur une cause de mortalité dans les régions dépourvues d'état civil, de registres de décès, ou lorsque les outils de diagnostic sont inexistants ou de fiabilité réduite.

Dans de nombreux pays, particulièrement dans les pays en voie de développement, l'autopsie verbale est indispensable pour effectuer des recherches démographiques ou épidémiologiques sur la mortalité [7].

L'OMS [2] dans son programme d'élaboration des normes en matière d'autopsie verbale préconise trois usages principaux:

- en premier lieu, on l'utilise essentiellement comme outil de recherche dans les études longitudinales de population et en recherche interventionnelle, c'est-à-dire en surveillance démographique et épidémiologique. La surveillance démographique vise entre autres à attribuer une cause de décès avec le plus d'objectivité possible pour favoriser la représentativité et la comparabilité des causes de décès à la fois dans le temps et dans l'espace. La surveillance épidémiologique vise à optimiser le diagnostic par rapport à l'objet de l'investigation, généralement chez les enfants ou encore pour déterminer la où les cause(s) de décès maternels.
- en second lieu, elle est devenue une source de données statistiques sur les causes de décès qui permet de répondre à la demande d'estimation de la charge de morbidité des populations, nécessaires pour l'élaboration des politiques, la planification, l'établissement des priorités et l'analyse comparative.
- enfin l'autopsie verbale fournit des données qui sont de mieux en mieux acceptées comme source de statistiques sur les causes de décès.

Mode opératoire L'autopsie verbale comporte cinq étapes essentielles:

1. Recensement des décès : par une enquête par sondage ou par une enquête exhaustive sur le terrain.

2. Interrogatoire de l'entourage de la personne décédée : il doit être effectué dans un délai raisonnable après le décès pour respecter le deuil de la famille tout en s'assurant d'un souvenir suffisamment récent. Il est mené par un enquêteur expérimenté à partir d'un questionnaire standardisé. Le figure 1.1 montre un entretien d'autopsie verbale dans la zone de Niakhar.

3. Le diagnostic primaire est effectué indépendamment par deux médecins qui rédigent chacun une fiche de synthèse standardisée.

4. La confirmation du diagnostic est apportée par un médecin tiers. Son rôle est de valider le diagnostic s'il est identique pour les deux diagnostics primaires ou de formuler un compromis lorsqu'ils sont voisins et (ou) compatibles.

5. En cas de divergence entre les deux diagnostics primaires, le dossier est examiné en réunion de consensus composé des trois précédents médecins et présidé par un quatrième médecin.

Cependant, le succès de cette méthode dépend:

- du type de questionnaire utilisé,
- de la classification de la mortalité utilisée,
- de la qualité des réponses, celle-ci étant elle même dépendante de plusieurs facteurs que sont:
  - l'enquêteur,
  - le répondant,
  - le délai entre le décès et l'enquête.

1.3 Les caractéristiques de la technique d'autopsie verbale

1.3.1 La classification de la mortalité

Selon le guide de l'OMS [2], les causes de décès sont définies comme toutes les maladies et les conditions morbides. Ces conditions morbides regroupent l'ensemble des circontances

qui aboutissent ou contribuent à la cause de décès. Cette définition inclut tous les processus complexes que traversent le patient avant le décès. Il n'est pas possible, en pratique, d'analyser tous ces processus à travers un interrogatoire rétrospectif de personnes proches du défunt et, seule une nécropsie minutieuse, complète pratiquée par des gens bien entraînés, pourrait permettre d'évaluer scientifiquement toutes les causes de décès.

Dans le but d'évaluer les causes de décès, l'approche à suivre au préalable pour élaborer un questionnaire est de lister une série de causes qui ont une probabilité élevée de survenir au sein de la population. Cela doit être effectué pour les divers groupes d'âges et de sexes et ensuite une série de symptômes est associée à chacune des causes. Les symptômes retenus doivent être assez spécifiques d'une maladie et exclusifs des autres. On obtient ainsi une classification de causes probables avec inévitablement une catégorie *Autres et non déterminées* destinée à rassembler les décès pour lesquels les symptômes ne permettent pas une affectation aux catégories précédentes.

Ainsi deux approches peuvent être utilisées pour développer et attribuer des diagnostics à partir de l'autopsie verbale:

- Approche restrictive : partant d'une classification de causes de mortalité donnée, l'outil d'autopsie verbale (comportant un questionnaire avec des procédures pour attribuer une cause) va classer les décès dans ces catégories pré- établies;
- Approche dite ouverte : la classification du décès est faite sur la base des diagnostics attribués à partir de l'autopsie verbale.

La caractéristique voulue dans la classification élargie de la mortalité ou approche dite ouverte est qu'elle peut être utilisée par d'autres études avec des modifications mineures. Une classification élargie de la mortalité devrait inclure toutes les causes de décès qui constituent des problèmes importants de santé publique. Les autres classifications sont reconnues comme répondant à des stratégies d'intervention. Ces catégories de maladies devraient autant que possible avoir des symptômes cliniques qui soient complexes, distincts et facilement reconnaissables. La connaissance de la structure des causes de décès de la population dans laquelle l'autopsie verbale va être appliquée pourrait faciliter une classification élargie de la mortalité selon des critères définis par [8].

1.3.2 Le questionnaire

Le questionnaire d'autopsie verbale peut avoir un certain nombre de formats:

1. Ouvert (une liste de symptômes),
2. Une liste avec des questions filtres,
3. La combinaison de 1 et 2.

Un questionnaire ouvert est une page blanche sur laquelle les enquêteurs rapportent les signes et symptômes ayant trait au décès. Pour la liste de symptômes, on établit la présence ou l'absence de symptômes. Pour la liste avec filtre, il y a une série de symptômes majeurs et de signes qui sont présents, suivis par un module de questions apparentées.

Pour exemple: dans le symptôme toux, une réponse positive à la question filtre sera alors suivie par une série de questions sur la durée, la sévérité de la toux et le type d'expectoration.

Un module peut être relié à un symptôme mais aussi à une catégorie spécifique de cause de décès et dans ce cas, il inclura des questions sur tous les symptômes exigés pour le diagnostic de la catégorie de la maladie en question. Pour exemple: l'absence ou non de toux peut être la question filtre pour entrer dans le module pneumonie qui inclura les questions sur la toux mais aussi sur les symptômes tels que la difficulté respiratoire, la polypnée et la fièvre, cela pour confirmer ou rejeter le diagnostic de pneumonie. La persistance ou non de la toux au delà de 4 semaines peut être un filtre pour entrer dans le module de tuberculose qui va alors inclure des questions telles que l'hémoptysie, l'amaigrissement, la fièvre et les difficultés respiratoires. La combinaison de format ouvert suivi d'une liste fermée avec ou sans filtre peut être utilisée. Les avantages et inconvénients de questionnaires ouverts ou structurés dans les études de santé par entrevue ont été discutés. Nous résumons les différentes aspects ci-dessous.

- Un format ouvert d'autopsie verbale requiert beaucoup plus d'expérience et probablement une connaissance médicale des enquêteurs ce qui augmente la variabilité des réponses entre différents enquêteurs.
- Une liste sans filtre ne nécessite pas d'enquêteurs médicalement formés et réduit les biais liès à l'enquêteur parce que ce dernier est forcé de considérer tous les symptômes pendant

l'enquête. Ce format ne retient pas tous les détails des symptômes rattachés au décès et peut accroître le nombre de symptômes faussement rapportés comme ayant été présents.

- Une liste avec filtre ne nécessite pas d'enquêteurs médicalement formés et peut être plus efficace pour la collecte de données. Là aussi, ce type de liste réduit le biais occasionné par les enquêteurs.

Par ailleurs les mérites relatifs aux formats variés d'autopsie verbale n'ont pas été formellement étudiés.
Le rôle important de la recherche qualitative des termes, des concepts locaux de maladies et terminologie, pour faciliter la procédure de traduction et de transcription des questionnaires a été souligné [3].

La présence de plusieurs langages ou dialectes dans les petites populations pose des problèmes pour le choix de la langue utilisée dans le questionnaire d'autopsie verbale. Dans ces situations, on peut adopter un questionnaire d'autopsie verbale pour chacune des langues locales de la population étudiée ou dans une langue principale avec une liste de symptômes, traduite dans les autres langues locales. Un modèle de questionnaire d'autopsie verbale devrait être adaptable aux différentes études par l'incorporation de concepts locaux de maladies, de symptômes, et de phraséologies.

1.3.3 Les enquêteurs

Les enquêteurs peuvent être des médecins, des infirmiers ou autres personnels de santé ou même de simples enquêteurs.
L'utilisation d'enquêteurs médicalement formés est préférable mais les capacités respectives des enquêteurs simples non formés et des enquêteurs médicalement formés pour la technique de l'autopsie verbale n'ont pas encore été étudiées.
Le personnel médical est coûteux, mais il est plus qualifié que les enquêteurs simples dans l'interprétation des réponses afin d'attribuer un diagnostic pendant l'entrevue. Aussi, le choix et la formation des enquêteurs sont déterminants pour assurer la qualité des réponses, ces enquêteurs doivent être du même groupe ethnique que les personnes enquêtées, originaires si possible de la même région ou village, du fait des problèmes posés par les concepts locaux de maladies et de terminologie c'est-à-dire la présence de plusieurs langages ou dialectes.

1.3.4 Les répondants

Le meilleur répondant est évidemment la personne qui connaît mieux la maladie finale et les circonstances du décès du patient.
Les mères sont ainsi les principales répondantes pour les décès d'enfants. Le meilleur répondant pour les décès d'adultes est moins facilement identifiable. Par exemple, un époux peut ne pas être le meilleur répondant pour le décès de sa femme, et il a été suggéré dans une étude que les sœurs sont meilleures répondantes en ce qui concerne le décès d'une femme [8].
Il est alors important de s'informer sur la personne qui a pris soin ou qui a vécu avec la personne décédée durant la maladie ayant conduit à la mort, mais aussi sur ses relations spécifiques interpersonnelles de façon à identifier et choisir le meilleur répondant.

1.3.5 Le délai d'enquête

Le temps écoulé entre le décès et l'interrogatoire est un élément important pour la qualité de l'information. Interrogés trop tôt, les parents pourraient être réticents à répondre du fait de certaines coutumes. Cependant quand l'interrogatoire est fait trop tard, ils peuvent avoir oublié certains détails de la séquence des événements nécessaires à l'établissement du diagnostic, entraînant ainsi des biais. Aussi, les décès d'adultes sont des événements assez rares et dans certaines sociétés les décès prématurés d'adultes sont considérés comme plus importants (graves) que les décès d'enfants. Il est alors possible d'utiliser des délais plus longs pour les décès d'adultes. Parler d'un décès très tardivement pourrait être la cause d'un stress et il serait alors plus judicieux de définir un minimum ou un maximum de délai d'enquête.

1.4 Méthodologie

Les informations sanitaires recueillies à partir des services sanitaires sont des sources utiles de statistiques régulières destinées à la planification des services et à l'allocation des ressources. Mais, elles ne donnent pas à elles seules une image complète du secteur de la santé, ni de la situation sanitaire des populations. Les sites de suivi démographique (Systemes de Surveillance Demographique (SSD)) permettent de satisfaire le besoin en information robuste, disponible régulièrement et de façon continue. Ces informations portent sur les événements démographiques

Figure 1.1: Un entretien d'autopsie verbale à Niakhar

et épidémiologiques. Le site de suivi démographique a pour rôle:

1. de suivre rigoureusement les changements dans la morbidité des maladies négligées ou rares, pour des interventions adaptées aux conditions épidémiologiques et socioculturelles,
2. de fournir des informations sur la santé qui reflètent précisément la charge actuelle de la maladie sur les populations,
3. de soutenir le contrôle et le suivi des nouvelles menaces qui pèsent sur la santé, et d'alerter les personnes ou autorités compétentes à prendre les mesures qui s'imposent,
4. de servir de plateforme d'essai et d'évaluation des interventions en matière de santé.

1.4.1 Présentation des sites d'études

Bandafassi

La zone d'étude de Bandafassi, au Sénégal, est située en zone de savane soudanienne, à 750 km au sud-est de Dakar, aux confins du Mali et de la Guinée. L'observatoire couvre une population s'élevant à 12 500 personnes en 2007 réparties dans 42 villages. Ceux-ci sont dispersés sur 600 km^2 dans une zone de plaines, de collines et de plateaux de faible altitude (300 à 400 mètres) mais particulièrement escarpés. Les villages sont de petite taille - 280 habitants en moyenne – et divisés en hameaux pour certains. La densité de population est de près de 20 habitants au km^2. La population totale est passée de 6 577 habitants à 12 500 habitants au début de 2007. La collecte d'informations démographiques s'effectue dans la zone de Bandafassi par enquête à passages répétés à intervalle annuel. Elle a commencé en 1970. Après un premier recensement, chaque village a été visité une fois par an, en général entre janvier et mars. Environ 5000 décès ont été enregistrés dans la période d'étude 1985 à 2010 dans la zone soit une moyenne de 165 décès par an. Tous les décès survenus n'ont pas fait l'objet d'une enquête par autopsie verbale.

Mlomp

Le site suivi de Mlomp, à environ 500 km de la capitale sénégalaise, se trouve dans la région de Ziguinchor, dans le département d'Oussouye, dans le sud-ouest du Sénégal, tout près de la

frontière longeant la Guinée-Bissau. Environ la moitié de l'arrondissement de Loudia-Ouolof y est contenue. La superficie du site de Mlomp est de 70 km^2. Les villages sont composés d'habitations regroupées en cercle pour former un diamètre d'à peu près 3 km. La zone de Mlomp est un ensemble de 11 villages ou quartiers. Le recensement initial a été la première phase de l'enquête démographique, puisqu'il a permis d'établir la première liste nominative des habitants de la zone d'étude. La population des 11 villages a été recensée fin 1984 début 1985. Les deux premiers passages démographiques ont eu lieu au mois d'octobre. Par la suite, ils ont eu lieu en début d'année, généralement entre les mois de janvier et de février.

Niakhar

Créée par Pierre Cantrelle en 1962, la zone d'étude de Niakhar, est un SSD situé à 135 km au sud-ouest de Dakar. Elle s'étend sur une latitude de 14° 30 Nord et une longitude de 16°30 Ouest. La zone d'étude de Niakhar, à cheval sur deux communautés rurales (Diarére et Ngayokhéme), comprend 30 villages répartis sur 230 km^2. En juin 2010, le SSD de Niakhar comptait 42 000 habitants et affichait une densité de population élevée de 131 habitants par km^2. Les 30 villages couverts par la surveillance démographique varient en importance : Darou (le plus petit) compte 60 habitants, alors que Toucar (le plus grand) en compte 3 150 ; trois autres villages ont plus de 2 000 habitants. Cette région faisait depuis 1962 l'objet d'un suivi démographique continu dans 8 villages de la zone et depuis 1983 dans les 30 villages actuels. Depuis lors, les habitants de ces villages sont visités à intervalles de temps réguliers. Depuis 1997, les enquêtes se déroulent chaque année en février, mai, août et novembre. A la suite de chaque passage, les informations, après vérification, sont saisies dans une base de données, régulièrement mise à jour.

1.4.2 Source de données des autopsies verbales

Dans ces trois sites, lors de leur passage dans les concessions, les enquêteurs collectent les événements qui ont eu lieu depuis le dernier passage.

Les décès survenus sur cet intervalle sont enregistrés sur des cahiers de terrain. Il s'agit du pré-enregistrement du décès. Durant cette première phase, on procède à l'identification de la personne décédée (nom, date de naissance, date et lieu de décès) et éventuellement à l'enregistrement d'une cause supposée.

C'est dans la deuxième phase que l'on introduit le questionnaire d'autopsie verbale (voir en annexe). La méthode d'autopsie verbale est appliquée dans ces trois sites pour déterminer les causes de décès mais son application a pu varier au cours du temps suivant la zone d'étude ou la période.

Depuis 2003, un effort d'harmonisation de la méthode dans les trois sites est mené [9]. Cela s'est traduit par un changement important dans les sites de Mlomp et Bandafassi: alors qu'auparavant les diagnostics étaient établis dans ces deux sites par un seul médecin, une double lecture des autopsies par deux médecins a été mise en place [9]. Ainsi sur ces trois sites, lorsqu'un décès se produit, l'agent enquêteur rencontre des proches du défunt et remplit un questionnaire dans lequel il identifiera la personne décédée et précisera l'évolution et les symptômes de sa maladie. Ensuite deux médecins parcourent ce questionnaire pour donner leur diagnostic. En cas de désaccord entre eux, on rassemble d'autres médecins afin qu'ils s'entendent sur le diagnostic.

La Classification Internationale des Maladies (CIM-9) de l'OMS est utilisée pour coder les causes de décès sous-jacentes les plus probables.

La présente étude porte sur l'ensemble des autopsies verbales qui ont eu lieu entre 1985 et 2010 à Bandafassi, Mlomp et Niakhar.

L'enregistrement du décès est suivi de l'administration du questionnaire d'autopsie verbale. La source de données sera constituée de la lecture des fiches d'autopsies verbales des trois sites.

Une base de données [1] réunissant l'ensemble des signes et symptômes recueillis pour chaque individu recensé comme décédé durant la période d'étude et les causes de décès déclarées par la famille et les médecins en charge du diagnostic est enregistrée. D'autres variables comme le recours au service de soins avant le décès, le lien de parenté du répondant avec le décédé, le délai d'enquête et tous les événements socio-démographiques sont aussi enregistrées dans cette base.

Dans la suite, nous définissons par:

- X les variables globales (items) représentant les questions *principales* par exemple **Fièvre ou Corps chaud**. Ces variables sont qualitatives et sont notées par **1** s'il y'a présence du symptôme ou **0** sinon.

[1] voir documentation de la base de données en annexe

Les variables **Sexe** et **Âge** des individus sont aussi codées comme une variable globale X. Une particularité est à noter sur le codage des variables **Sexe** et **Âge**. La variable **sexe** est codée **1** pour le sexe masculin et **0** pour le sexe féminin. Pour la variable **Âge**, une discrétisation en groupes d'âges est utilisée. Les personnes sont groupées en **neonate**[2]**, infant**[3]**, under5**[4]**, child**[5]**, adult**[6]**, midage**[7] et **elder**[8] et pour chaque groupe on code par **1** si la personne est dans ce groupe et **0** sinon.
Toute autre variable indépendante à ces rubriques (voir questionnaire en annexe) est considérée comme une variable globale X.

- Les variables Z notées sous items représentent toutes les sous-réponses relatives à la présence d'une variable globale X.
  Si X vaut **0** toutes ses sous-réponses Z sont nulles.

  Par exemple, dans la rubrique **Fièvre ou Corps chaud** (voir Figure 1.2):

  - La variable X représente la réponse à la question absence ou présence de **Fièvre ou Corps chaud**.
  - Les variables Z représentent les réponses aux questions suivantes: **très forte, moyenne, intermittente, continue, présence de sueurs et de frissons** relatives à la rubrique **Fièvre ou Corps chaud**.

  Si la personne répond par **0** à la présence de **Fièvre ou Corps chaud**, toutes les questions relatives à cette rubrique sont omises.
  Ces variables sont souvent quantitatives mais un processus de codage binaire est utilisé. Par exemple, la durée de la **Fièvre ou Corps chaud** est codée par **1** si elle est inférieure à deux semaines et **0** sinon.

  Le nombre de sous items pour un item donné dépend de la structure du questionnaire utilisé (voir l'exemple de questionnaire en annexe). Ce même procédé est appliqué pour toutes les parties du questionnaire.

[2] moins de 28 jours
[3] entre 1 et 11 mois
[4] âgé de 1 à 4 ans
[5] âgé de 5 à 14 ans
[6] âgé de 15 à 49 ans
[7] âgé de 50 à 64 ans
[8] âgé de plus de 65 ans

- ▶ A la lecture de la fiche d'autopsie verbale, deux médecins en charge des diagnostics proposent chacun une cause probable. Pour diminuer le biais relatif à l'attribution des causes probables par les deux médecins, on note par Y les causes communes aux deux médecins. Plusieurs causes sont ainsi notées et sont regroupées, sous la direction des médecins, en des groupes de causes, notées $G_1, \ldots, G_K$.

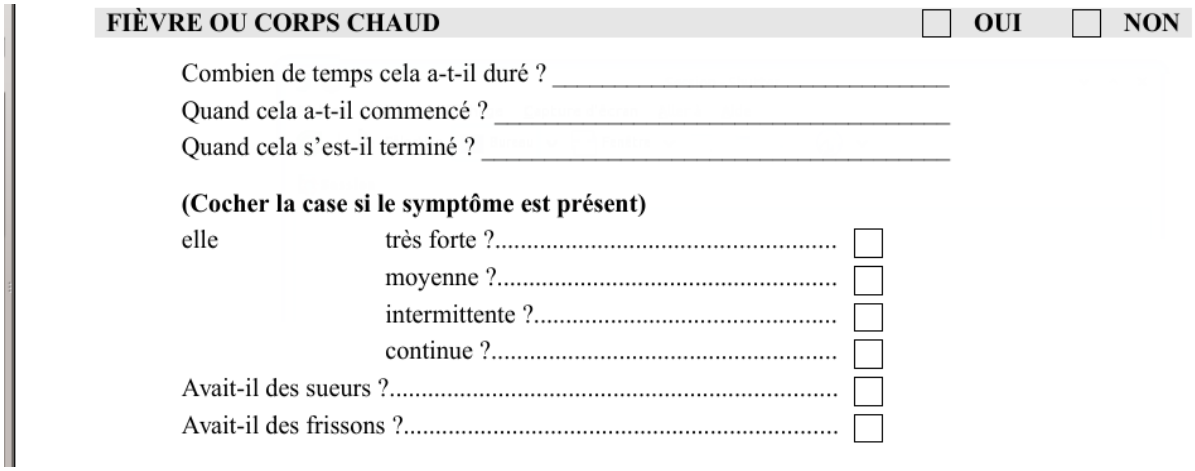
Toutes ces informations sont consignées dans une base de données. L'architecture et les outils utilisés lors de la conception de la base de données sont décrites en annexe. Pour le traitement et l'épuration de la base de données, nous avons utilisé en même temps les logiciels *Stata*, *R* et *Matlab*. La conception et l'implémentation des méthodes proposées dans ce document sont réalisées avec *R* et *Matlab*.

Dans la suite du document, on dispose d'un échantillon de n individus décrits par les variables explicatives X et Z ainsi que leur appartenance à l'un des K groupes (variable Y). Cet échantillon dit échantillon d'apprentissage va être utilisé pour construire une règle de décision.

Nous formalisons les réponses du questionnaire comme suit:
Ces variables binaires représentent la présence (1) ou l'absence (0) des symptômes et des variables non symptomatiques pour un individu donné. Pour chaque symptôme donné, un ensemble de sous items peut être aussi collecté. Au final, on a:

- les variables aléatoires binaires $X = (X_j, j = 1, \ldots, p)$ qui définissent les symptômes et variables socio-démographiques.
- les variables aléatoires $Z = (Z_j^\ell, j = 1, \ldots, p, \ell = 1, \ldots, p_j)$ qui représentent les p_j sous-items pour chaque variable X_j.
- la variable aléatoire $Y \in \{1, \ldots, K\}$ qui est la variable à expliquer représentant le groupe (diagnostics des médecins).

Dans le cas précis de nos travaux, après épuration de la base de données, nous avons $n = 2500$ individus déclarés décédés durant la période de 1985 à 2010. Un ensemble de près de $\sum_{j=1}^{p}(1 + p_j) = 100$ variables (incluant les items et les sous items) est recensé pour chaque individu. Un ensemble de $K = 22$ causes de décès déclarées par les médecins est ainsi recensé.



FIÈVRE OU CORPS CHAUD ☐ OUI ☐ NON

Combien de temps cela a-t-il duré ? _____

Quand cela a-t-il commencé ? _____

Quand cela s'est-il terminé ? _____

(Cocher la case si le symptôme est présent)

elle très forte ?..... ☐

moyenne ?..... ☐

intermittente ?..... ☐

continue ?..... ☐

Avait-il des sueurs ?..... ☐

Avait-il des frissons ?..... ☐

Figure 1.2: Extrait du questionnaire

1.5 État de l'art des méthodes mathématiques

La fiabilité de la méthode de l'autopsie verbale dépend de nombreux facteurs: la cause du décès, les informations recueillies, les médecins,... La méthode de l'autopsie verbale peut parfois entraîner des erreurs de diagnostic [10]. L'assignation des causes de décès par la méthode des autopsies verbales est le plus souvent faite par des médecins. Cette assignation est souvent assujettie à des biais dûs à la spécialité du médecin et à son expérience. Pour réduire au mieux les biais dûs aux diagnostics des médecins et améliorer les performances du diagnostic, des méthodes automatisées sont présentées dans la littérature. Des algorithmes experts sont ainsi développés, la validation des résultats de ces méthodes est souvent faite par une comparaison des diagnostics des médecins notés Physician-Certified Verbal Autopsy (PCVA) avec les diagnostics de la méthode proposée.

Des approches probabilistes basées sur la méthode de Bayes, sont utilisées dans d'autres sites [11] [12][13]. Dans ces approches, on cherche à classer les causes de décès en rapport avec des indicateurs (symptômes, âge, sexe, ...). Plus particulièrement, on détermine la probabilité d'une cause C étant donné la présence d'un indicateur particulier I. Pour chaque indicateur et pour chaque cause une probabilité d'apparition est calculée parmi les cas de décès. Ainsi pour une cause donnée, sa probabilité est d'abord calculée parmi les décès dans la population, ce qui représente la fraction de la mortalité par cause (CSMF). Pour chaque individu, 3 causes probables sont énumérées avec leurs probabilités respectives. A partir de cette sortie, il est

également possible d'estimer un facteur de risque pour chaque cause, défini comme la somme des probabilités des 3 cas divisée par 3. Cette méthode est à la base d'un logiciel dénommé InterVA.

D'autres approches combinent les profils symptomatiques [14] des individus avec la méthode de Bayes [15] pour déterminer les causes probables de décès décrites par autopsie verbale. Cette méthode décrite dans [14] combine les avantages de la méthode proposée par [15] et est notée par SP (Symptôme Simplifié). La méthode SP utilise deux échantillons. Dans le premier échantillon, la cause de décès est connue avec certitude mais l'échantillon n'a pas besoin d'être représentatif dans la population d'étude. Ces données sont issues des hôpitaux, dispensaires ou cases de santé. Dans le second échantillon, les données sont issues des autopsies verbales et il est plus représentatif. Au niveau des données hospitalières, on calcule les probabilités de réponses pour chaque symptôme au niveau du premier échantillon. Ces probabilités des symptômes permettent en premier lieu de calculer les Cause Specific Mortality Fraction (CSMF) au niveau de la population et en second lieu d'utiliser les CSMF pour l'assignation de la cause de décès en rapport avec chaque réponse au niveau des autopsies verbales. Cette CSMF est ainsi utilisée pour affiner la cause de la mort de manière individuelle sur la population. La méthode SP fournit des évaluations précises des CSMF spécifiques à chaque cause.

La méthode des forêts aléatoires [16] est aussi utilisée pour prédire la cause de décès. Pour prédire la cause de décès, cette méthode créé des arbres pour chaque paire de causes. À chaque nœud, l'algorithme choisit sur un sous-ensemble aléatoire de symptômes pour discriminer les individus suivant les deux causes. Pour chaque paire de causes issue de la liste des causes, 100 arbres de décision sont ainsi créés pour discriminer au mieux la paire de causes. Pour l'agrégation de tous ces arbres, un système de vote est effectué en comptant le nombre d'arbres ayant prédit une cause spécifique donnée. Ensuite un système de normalisation (cf [16]) est proposé pour déterminer le score de chaque cause.

La méthode dite "Le tarif " proposée par [17] calcule un score pour chaque cause et pour chaque symptôme, à travers une base de données d'autopsie verbales validée. Ce score est calculé sur les réponses données lors de l'enquête et permet de prévoir la cause probable de décès. Cette méthode permet d'identifier des signes et symptômes les plus fortement significatifs pour une cause donnée. L'approche est la suivante. Un tarif est calculé pour chaque symptôme. Ce tarif permet d'attribuer sur la base des réponses un pouvoir informatif à ce symptôme pour une cause donnée. Pour chaque cause donnée, les tarifs relatifs aux symptômes sont additionnés

pour constituer le tarif de cette cause. Pour un individu donné, la cause ayant obtenue le plus grand score par rapport aux tarifs des symptômes déclarés lui est attribuée.

Dans [18], une étude comparative de ces 5 méthodes automatisées de diagnostics des autopsies verbales est présentée. La comparaison entre les méthodes de InterVA-4 [19], de forêt aléatoire [16], du Modèle SP [14], de la méthode "Le tarif " [17] et celle utilisant la méthode de Bayes [15] est ainsi effectuée.

Certaines de ces méthodes utilisent des retours d'expérience pour évaluer la performance des résultats proposés. La plupart de ces approches sont appliquées dans des zones spécifiques où le VIH sida est prédominant [20]. Dans ces méthodes la détermination des causes de décès et le contexte épidémiologique sont fortement corrélés. Dans nos sites d'études, le paludisme est endémique contrairement aux autres sites étudiés. La plupart des études menées dans les zones rurales du Sénégal, sur la détermination des causes probables de décès portent sur ces deux critères [10, 21, 22, 23, 24]:

- spécificité: tous les décès diagnostiqués pour une cause sont réellement dûs à cette cause,
- sensibilité: tous les décès réellement dûs à une cause sont bien diagnostiqués comme tels.

Cependant on sait que les principaux facteurs de mortalité interagissent de façon complexe et sont tributaires, à leur tour, d'un ensemble de variables, de facteurs sociaux tout aussi complexes, du fait que rares sont les causes de décès qui sont pathognomoniques.

Aucune méthode mathématique n'est encore utilisée pour la détermination et/ou l'analyse des causes probables de décès décrites par autopsie verbale dans les trois sites d'étude du Sénégal (Bandafassi, Niakhar et Mlomp). De plus, aucune des méthodes citées plus haut ne prend en compte les individus diagnostiqués pour une cause donnée comme une sous-population de la population d'étude.

Ainsi, les modèles de mélanges permettent de prendre en compte cette hétérogénéité des causes de décès et la population totale comme un mélange de sous populations homogènes. Diagnostiquer, c'est détecter les similarités entre individus, c'est aussi prendre en compte l'absence et la présence de symptômes dans le processus de diagnostic. Ainsi dans la méthode

proposée au chapitre 4, on effectue une combinaison entre les modèles de mélanges et les mesures de similarité.

Dans le chapitre suivant, nous proposons tout d'abord une revue des méthodes de mélange, en passant par les méthodes à noyaux et les mesures de similarité.

CHAPITRE 2

ÉTAT DE L'ART: MÉTHODES STATISTIQUES

Sommaire

2.1 Introduction

Notre travail s'inscrit dans le cadre des méthodes d'apprentissage statistique. Ce chapitre a pour but d'introduire les notions et concepts clés en analyse discriminante, sur les méthodes à noyaux et sur les mesures de similarité.

Un modèle mathématique est une représentation d'un phénomène étudié, dont la qualité dépend essentiellement de la connaissance de ce phénomène et des moyens dont on dispose pour construire le modèle. La connaissance étant souvent imparfaite (limitation de la compréhension du phénomène, du nombre de données et des expériences) et les moyens limités (scientifiques et numériques). De plus, expliquer un phénomène, revient à comprendre son fonctionnement, étudier sa complexité et prendre une décision au vu de prédire son comportement. Avec le nombre croissant de données dans le monde moderne, le besoin en analyse augmente et devient de plus en plus complexe. Cette modélisation peut se faire de différentes manières.

Une approche pertinente pour aborder les problèmes complexes est la fouille de données. La fouille de données est l'ensemble des méthodes permettant d'explorer et d'analyser des données de façon automatique ou semi automatique, dans le but de détecter des structures particulières ou cachées en établissant des règles de décision. Une des parties importantes des méthodes de fouille de données est la classification.

La classification est une méthode d'analyse des données qui vise à regrouper en classes homogènes un ensemble d'observations. Elle a pour but la résolution de manière automatique des problèmes par la prise de décision sur la base des observations induites aux problèmes. Son objectif est principalement de définir des règles permettant de classer des objets à partir de variables qualitatives ou quantitatives caractérisant ces objets. Elle joue un rôle de plus en plus important dans de nombreux domaines scientifiques et techniques.

La classification revêt de plus en plus une place importante en analyse de données exploratoire et décisionnelle. La classification se subdivise en deux approches:

Approche explicative Les méthodes explicatives cherchent à découvrir les variables ou les associations de variables. Elles cherchent ainsi à découvrir les variables les plus pertinentes pour décrire les différences entre groupes a priori et à déterminer si ces différences sont significatives.

Approche décisionnelle L'approche décisionnelle cherche à affecter tout nouvel individu ou objet à des groupes définis a priori. Pour cela la méthode proposée doit s'adapter à la diversité des données comme l'abondance, la rareté des individus ou objets, le nombre de descripteurs et le nombre de classes.

La classification se subdivise souvent en trois types de méthodes:

Méthode non supervisée La classification non supervisée dite aussi clustering permet de modéliser le système qui a généré les données observées et à partir de cette modélisation construit la règle de décision. Elle cherche une typologie ou segmentation c'est-à-dire une partition ou une répartition des individus en classes ou groupes. Le classement se fait en optimisant un critère permettant de regrouper les individus dans des classes les plus homogènes et les plus distinctes possible. Le nombre de partitions n'est pas souvent connu a priori.

Méthode semi supervisée L'approche semi supervisée permet de prendre en compte les données non classées pour améliorer la règle de classification. Elle modélise la distribution conjointe des prédicteurs et de la classe. L'auto-apprentissage constitue une des premières méthodes utilisées en classification semi-supervisée. Elle consiste à subdiviser les données en deux parties. Dans une des parties, on construit une règle de classement à partir des données classées. Ainsi à partir de ce classifieur, on classe les autres parties des données sans leurs étiquettes [25].

Méthode supervisée La classification supervisée construit directement la règle de décision à partir des données observées. Elle s'applique à des populations décrites par des variables et munies d'une partition définie a priori et d'un intérêt particulier. Elle vise à séparer au mieux les classes de la partition à l'aide des variables explicatives.

De plus la performance de certaines méthodes de classification repose souvent sur le choix d'une bonne mesure de similarité. De ce fait, depuis plusieurs années, des mesures de similarité sont proposées dans des domaines divers et variés. Leur terminologie varie suivant les domaines considérés (similarité, ressemblance, proximité, ...) et suivant les types de données. La classification peut aussi s'appuyer sur l'utilisation de noyau. Les méthodes à noyaux constituent une classe de modèles qui étendent astucieusement les méthodes linéaires au cas non linéaire.

L'utilisation des noyaux ne modifie pas fondamentalement la nature des données et du problème posé. Ce chapitre se propose d'énoncer les principaux concepts mathématiques utilisés dans nos travaux. Nous élaborons une étude sur les méthodes de classification particulièrement en classification supervisée. Au niveau du paragraphe 2.2 nous nous focaliserons sur les méthodes paramétriques avec l'introduction des modèles de mélanges sur des données quantitatives et qualitatives. Au paragraphe 2.3 nous décrirons les méthodes non paramétriques. Nous aborderons les noyaux et les stratégies de construction des fonctions noyaux sur des données structurées et non structurées. Nous nous focaliserons par la suite sur trois méthodes de classification utilisant un noyau. Les mesures de similarités étant importantes pour la construction des noyaux, de ce fait dans le paragraphe 2.4 nous aborderons une étude élargie des mesures de similarités.

2.1.1 Données et notations

Dans la suite du document, on s'intéresse à la classification supervisée. Nous adopterons les notations suivantes.

On dispose d'un échantillon de n individus décrits par p variables à expliquer $X_1, \ldots, X_p$. L'appartenance de chaque individu à l'un des K groupes a priori $C_1, \ldots, C_K$ est connue. On suppose que les observations $\{x_1, \ldots, x_n\}$ sont des réalisations d'un vecteur aléatoire $X = (X_1, \ldots, X_p) \in \mathbb{R}^p$. On suppose en outre que $\{y_1, \ldots, y_n\}$ décrivent l'appartenance des observations $x = \{x_1, \ldots, x_n\}$ et sont des réalisations de la variable aléatoire $Y \in \{1, \ldots, K\}$.

Soit $x = (x_1, \ldots, x_p)$ les caractéristiques d'un individu donné, on note:

- $\mathbb{P}(X)$ la loi de probabilité du vecteur X,
- $\mathbb{P}(C_k)$ la probabilité de la classe C_k,
- $\mathbb{P}(X|C_k)$ la loi de probabilité du vecteur X sachant que X appartient à la classe C_k,
- $\mathbb{P}(C_k|X)$ la probabilité que X appartienne à la classe C_k.

La distribution du couple des vecteurs aléatoires (X, Y) est construite en supposant que:

- La variable Y est munie de la loi de probabilité $\pi_1, \ldots, \pi_k$. π_k représente la probabilité a priori d'un groupe i.e
  $\mathbb{P}(Y = k) = \mathbb{P}(C_k) = \pi_k$, , $k = 1, \ldots, K$.

- Sachant $Y = k$, les vecteurs de description suivent la loi $\mathbb{P}(X|C_k)$.

2.1.2 Règles de décision

La classification supervisée construit une règle de décision permettant d'affecter un individu décrit par p variables à l'un des K groupes $C_1, \ldots, C_K$ d'une partition définie a priori sur la population étudiée. Pour construire une telle règle, il faut disposer d'un échantillon d'apprentissage de la population pour lequel l'affectation de chaque individu à l'un des groupes est connue.

Une règle de décision est une application ξ qui associe au vecteur $x \in \mathbb{R}^p$ un vecteur $\xi(x) \in \{1, \ldots, K\}$

$$\begin{array}{rccl} \xi & : & \mathbb{R}^p & \rightarrow & \{1, \ldots, K\} \\ & & x & \mapsto & \xi(x) = k. \end{array}$$

La qualité d'une règle de décision se mesure souvent par une matrice des coûts de mauvais classement $C(\ell|k)$ qui représente le coût de classement d'un individu dans C_ℓ d'un individu du groupe C_k. De plus si $\ell = k$ alors $C(k|k) = 0$. La règle de décision optimale[1] pour les modèles statistiques est la règle de Bayes.

Elle minimise l'espérance du coût de mauvaise classification $R(\xi)$. $R(\xi)$ est décrit dans [27] comme :

$$R(\xi) = \sum_{k=1}^{K} \pi_k \sum_{\ell=1}^{K} C(\ell|k) \int_{R_{\xi(l)}} \mathbb{P}(x|C_k)dx$$

avec $\int_{R_{\xi(l)}} \mathbb{P}(x|C_k)dx$ représentant la probabilité que x soit affecté à C_ℓ sachant qu'il appartient à C_k et $R_{\xi(l)} = \{x \in \mathcal{X}/\xi(x) = \ell\}$.

La règle de Bayes consiste à affecter l'observation x à la classe la plus probable *a posteriori*:

$$\xi^*(x) = \underset{\ell=1,\ldots,K}{\operatorname{argmax}} \mathbb{P}(x|C_\ell)\pi_\ell.$$

Cette règle dépend des probabilités a priori π_k et des probabilités a posteriori par groupe. Ces deux quantités sont le plus souvent estimées comme suit.

[1] Cf ([26], Annexe technique) pour la démonstration

- Pour tout $k = 1, \ldots, K$, les π_k sont interprétées comme des probabilités a priori des observations x appartenant à la classe C_k. Elles sont souvent estimées par $\frac{n_k}{n}$ où n_k est le cardinal de la classe k et n est le nombre total d'observations.

- Les probabilités a posteriori sont les valeurs des probabilités $\mathbb{P}(X = x|Y = k)$. L'estimation des probabilités a posteriori dépend souvent du schéma d'échantillonnage[2]. Des méthodes d'estimation des probabilités a posteriori dans le cadre quantitatif et qualitatif sont décrites dans les paragraphes 2.2.2.2 et 2.2.3.2.

2.1.3 Estimation des performances d'une méthode de classification

Après avoir construit une règle de classification, il est important de mesurer sa performance. Ainsi, pour évaluer la qualité d'une méthode de classification, on sépare les données disponibles entre une base d'apprentissage et une base de test. La règle de classification ξ est appliquée d'abord à la base d'apprentissage puis on utilise ξ pour retrouver la classe d'appartenance des éléments de la base de test.

L'estimation du taux d'erreur e de classification est faite le plus souvent sur l'échantillon test par :

$$e(\xi) = \frac{\sum_{i=1}^{n_T} 1(\hat{y}_i \neq y_i)}{n_T}$$

avec n_T la cardinalité de l'échantillon test et $\hat{y}_i$ la classe estimée de l'individu i.

Souvent la performance est calculée en terme de Taux Correct de Classification (TCC) qui mesure le nombre d'accords communs entre le modèle proposé et le système qui a généré les données:

$$TCC(\xi) = (1 - e(\xi)) \times 100.$$

Il existe plusieurs méthodes d'estimation de l'erreur de classification. Nous allons énumérer les plus populaires.

Estimation par resubtitution Cette estimation utilise l'échantillon d'apprentissage pour construire et valider en même temps la règle de décision. Le taux issu de cette estimation est

[2]Cf [27], pour les différentes schémas d'échantillonnage

appelé *taux apparent d'erreur*. Ce taux est optimiste du fait que le même échantillon est utilisé comme apprentissage et test. Il ne conduit pas à une estimation correcte du taux d'erreur exact.

Estimation par échantillon test L'échantillon initial est scindé en deux échantillons: apprentissage et test. La règle de décision est construite sur l'échantillon d'apprentissage et sa validation est faite sur l'échantillon test. Sa précision dépend de la taille de l'échantillon test et exige un nombre d'observations important dans les deux échantillons.

Estimation par validation croisée Les données sont rarement de taille suffisante pour que l'estimation par échantillon test donne une estimation performante de l'erreur de prédiction du modèle. La validation croisée est l'une des méthodes les plus utilisées dans l'étude de performance des classifieurs.

La validation croisée est une méthode d'estimation de fiabilité d'un modèle fondée sur une technique d'échantillonnage.
Son principe est le suivant:

- Diviser les données d'apprentissage en r sous-échantillons de tailles égales.
- Retenir l'un de ces échantillons, supposons le $i$ème échantillon, pour le test et apprendre sur les $r-1$ autres.
- Calculer l'erreur empirique $\hat{e}^i(\xi)$ sur l'échantillon i avec ξ la règle de classification utilisée.
- Recommencer r fois en faisant varier l'échantillon i de 1 à r.

L'erreur dite de validation croisée est la moyenne des erreurs mesurées:

$$\hat{e}(\xi) = \frac{1}{r}\sum_{i=1}^{r} \hat{e}^i(\xi).$$

Cette procédure fournit une estimation non biaisée du taux d'erreur réel. Il est souvent courant de prendre des valeurs de r comprises entre 5 et 10. Ainsi on utilise une grande partie des données pour l'apprentissage tout en obtenant une mesure précise du taux d'erreur réel.

2.2 Méthodes paramétriques

L'approche paramétrique de la classification consiste à supposer que les individus à classer sont des réalisations indépendantes d'une variable aléatoire de distribution de probabilité connue. Les observations $\{x_1, \ldots, x_n\}$ à classer sont des réalisations d'un vecteur aléatoire X à valeurs dans $\mathbb{R}^p$ de la loi de probabilité $f(\bullet, \alpha)$ avec α le paramètre à estimer.

Les hypothèses sur la loi de probabilité considérée permettent de définir différents modèles probabilistes de classification : le modèle de multi-modalité, le modèle de partitions fixes et le modèle de mélange. Dans ce document, nous nous focaliserons sur le cas du modèle de mélange.

2.2.1 Modèles de mélange

L'approche de classification basée sur les modèles de mélange est apparue en 1894 avec une étude de Pearson [28]. L'approche mélange permet de formaliser l'idée de données hétérogènes. Elle permet de modéliser des lois de probabilité par groupes. On suppose que les données proviennent d'une source contenant plusieurs sous-populations. Chaque sous-population est modélisée de manière séparée, la population totale est un mélange de ces sous-populations. Le modèle résultant est un modèle de mélange fini.

Un mélange fini est une combinaison convexe de lois de probabilités $f_k(\bullet, \theta_k)(k = 1, \ldots, K)$:

$$f(\bullet; \alpha) = \sum_{k=1}^{K} \pi_k f_k(\bullet; \theta_k).$$

Les coefficients π_k sont les proportions du mélange *ie* $\forall k : \pi_k > 0$ et $\sum_{k=1}^{K} \pi_k = 1$.
Les paramètres des composantes du mélanges sont notés $\theta_k, k = 1, \ldots, K$ et on note $\alpha = (\pi_1, \ldots, \pi_K, \theta_1, \ldots, \theta_K)$. Les composantes du mélange $f_k, k = 1, \ldots, K$ sont souvent de la même famille de lois et leur choix dépend de la nature des données. La nature des données conduit souvent à une modélisation gaussienne pour des données quantitatives et à une modélisation multinomiale pour des données catégorielles.

Plusieurs auteurs se sont penchés sur l'étude des modèles de mélanges surtout dans les modèles de mélanges de gaussiennes [29, 30, 31, 32, 33, 34] et sur des modèles de mélanges de lois multinomiales [27, 35, 36, 37, 38, 39, 40].

2.2.2 Modèle de mélange sur variables quantitatives

Dans ce paragraphe, nous nous focaliserons particulièrement sur le modèle de mélange gaussien. Nous définissons tout d'abord la densité gaussienne, puis le modèle de mélange gaussien.

2.2.2.1 Modèle de mélange gaussien

La loi normale ou distribution gaussienne est la distribution la plus populaire parmi les distributions probabilistes. Elle est le plus souvent utilisée pour modéliser des variables aléatoires continues. Dans le cas multidimensionnel, sa fonction de densité multivariée s'exprime de la manière suivante:

$$\mathcal{N}(x;\theta_k) = \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma_k|^{\frac{1}{2}}} \exp(-\frac{1}{2}(x-\mu_k)^t {\Sigma_k}^{-1}(x-\mu_k))$$

avec $\theta_k = (\mu_k, \Sigma_k)$, μ_k le vecteur des moyennes de dimension p, Σ_k la matrice de covariance et $|\Sigma_k|$ représente le déterminant de la matrice Σ_k. Cette densité gaussienne modélise une classe ellipsoïdale de centre μ_k et ses caractéristiques géométriques sont associées à la décomposition spectrale de la matrice de variance Σ_k.

Le modèle de mélange le plus populaire est le modèle de mélange de type gaussien. Dans ce cas, les densités de probabilité des variables explicatives conditionnellement aux classes sont supposées être celles de loi normales de moyennes μ_k et de matrice de variances Σ_k:

$$f(x;\alpha) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x;\theta_k).$$

Pour prédire la variable Y à partir des variables explicatives, il faut s'appuyer sur les probabilités a posteriori comme décrit dans le paragraphe 2.1.2.

L'hypothèse d'égalité ou non des matrices de variance-covariance Σ_k entre les classes détermine la manière d'estimer les paramètres du mélanges.

2.2.2.2 Estimation paramétrique: cas quantitatif

Les estimateurs paramétriques sont utilisés afin d'estimer le paramètre inconnu θ. Dans ce paragraphe, nous passerons en revue la méthode du maximum de vraisemblance et ses adaptations pour l'estimation des paramètres en grande dimension.

Maximum de vraisemblance L'estimation par maximum de vraisemblance est une méthode courante utilisée pour inférer les paramètres de la distribution de probabilité d'un échantillon donné. Pour appliquer l'estimation par maximum par vraisemblance, on se donne un échantillon fini $\{x_1, \ldots, x_n\}$ généré par un n-échantillon $\{X_1, \ldots, X_n\}$ de variables aléatoires indépendantes de loi f_θ, que l'on souhaite modéliser. On souhaite ainsi estimer le paramètre θ.

L'idée est alors de choisir θ^*, sous l'hypothèse d'indépendance des variables aléatoires X_i, qui maximise la probabilité d'observer $\{x_1, \ldots, x_n\}$:

$$\theta^* = \underset{\theta}{\operatorname{argmax}}\{L(x_1, \ldots, x_n; \theta)\}$$

où $L(x_1, \ldots, x_n; \theta)$ est la vraisemblance du modèle avec

$$L(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} f_\theta(x_i).$$

La vraisemblance s'exprime ainsi par le produit de toutes les densités de probabilités marginales. L'estimateur fourni par la méthode du maximum de vraisemblance possède des propriétés de consistance, de convergence en probabilité et est asymptotiquement sans biais [41]. L'expression de la vraisemblance dépend des caractéristiques des labels *ie* du cas supervisé ou non supervisé. L'affectation peut aussi se faire en trouvant le minimum sur k de la fonction du score discriminant:

$$s_k(x) = \frac{1}{2}(x - \hat{\mu}_k)^t \hat{\Sigma}_k^{-1}(x - \hat{\mu}_k) + \frac{1}{2}\log|\hat{\Sigma}_k| - \log\hat{\pi}_k$$

Dans le cas où les distributions dans chaque classe ont même matrice de covariance Σ, il suffit de prendre

$$s_k(x) = \frac{1}{2}(x - \hat{\mu}_k)^t \hat{\Sigma}^{-1}(x - \hat{\mu}_k) - \log\hat{\pi}_k$$

Si de plus les probabilités a priori des classes sont supposées égales, le score discriminant coïncide avec la distance de Mahalanobis:

$$s_k(x) = \frac{1}{2}(x - \hat{\mu}_k)^t \hat{\Sigma}^{-1}(x - \hat{\mu}_k).$$

La règle d'affectation bayésienne devient ainsi la recherche du centre le plus proche selon cette distance.

En grande dimension, le problème de l'inversibilité des matrices peut se poser. Dans le paragraphe suivant, nous énoncerons certaines techniques développées pour la grande dimension.

Estimation en grande dimension Dans ce paragraphe, nous allons énoncer les problèmes d'estimation en grande dimension. La classification de données en grande dimension est un problème délicat qui requiert souvent d'autres techniques d'estimation. Ce problème apparaît dans des domaines comme l'analyse de gènes, d'images, ...

Pour la classification des données en grande dimension, le modèle de mélange gaussien est le plus utilisé. Mais ce dernier s'avère limité lorsque la taille de l'échantillon est faible devant le nombre de variables. Ce problème est noté "curse of dimensionality", voir [33, 42, 43]. Pour le modèle de mélange gaussien, le nombre de paramètres de la matrice de covariance à estimer croit avec le carré de la dimension. Comme souligné dans [43], il existe des méthodes pour pallier ce problème comme les méthodes de réduction de dimension [44], les modèles parcimonieux [45] et les méthodes de régularisation [46].

Nous focalisons notre attention sur les modèles parcimonieux. Nous allons tout d'abord énoncer les propriétés géométriques des modèles gaussiens.

Dans le cas gaussien, les caractéristiques géométriques sont souvent décrites par la décomposition spectrale de la matrice de variance Σ_k [47]. Cette décomposition spectrale s'écrit pour chaque matrice de variance:

$$\Sigma_k = \lambda_k D_k A_k {D_k}^t$$

avec $\lambda_k = |\Sigma_k|^{1/d}$ représentant le volume de la classe, D_k étant la matrice des vecteurs propres de Σ_k représentant l'orientation de la classe et A_k étant la matrice diagonale, telle que $\det(A_k) = 1$ dont la diagonale est constituée des valeurs propres normalisées de Σ_k rangées en ordre croissant.

Une re-paramétrisation est proposée dans [33]. Cette re-paramétrisation combine à la fois

réduction de la dimension, modèles parcimonieux et régularisation.

L'hypothèse fondamentale de cette méthode est que les données de chacune des classes vivent dans des sous-espaces différents dont les dimensions intrinsèques peuvent être différentes. Pour ce faire, on se donne une matrice orthogonale Q_k composée des vecteurs propres de Σ_k et on définit la matrice de covariance Δ_k dans l'espace propre de Σ_k par:

$$\Delta_k = Q_k^t \Sigma_k Q_k.$$

La matrice Δ_k est une matrice diagonale contenant les valeurs propres de Σ_k. Dans ce modèle, on suppose Δ_k n'a que $d_k + 1$ valeurs propres différentes notées $a_{k1}, \dots a_{kd_k}$ et $p - d_k$ valeurs propres communes notée b_k. L'hypothèse fondamentale de cette méthode est $a_{kj} > b_k,\ j = 1, \dots, d_k$ avec $d_k \in \{1, \dots, p-1\}$ la dimension intrinsèque du sous espace des données, supposée inconnue.

Cette modélisation suppose que les variables latentes X conditionnellement aux classes $Z = k$ vivent dans un sous-espace E_k engendré par les d_k vecteurs propres associées aux valeurs propres a_{kj} et tel que $\mu_k \in E_k$. Le paramètre b_k modélise la variance d'un bruit gaussien ε sur l'orthogonal de E_k. La relation linéaire des variables aléatoires X et Y conditionnellement à $Z = k$ est:

$$Y_{|Z=k} = Q_k X_{|Z=k} + \varepsilon_{|Z=k}.$$

Ce modèle gaussien parcimonieux est noté $[a_{kj} b_k Q_k d_k]$. Dans le cas de la classification supervisée, à partir de cette re-paramétrisation, les auteurs définissent une méthode dite "High-dimensional discriminant analysis (HDDA) " [48].

Dans ce modèle, les probabilités a priori sont estimées par $\hat{\pi}_k = \frac{n_k}{n}$ et les moyennes par $\hat{\mu}_k = \frac{1}{n_k} \sum_{i \in C_k} y_i$ où n_k est le nombre d'individus dans la classe C_k.

Les estimations des paramètres spécifiques du modèle requièrent la connaissance de la dimension intrinsèque de chaque classe. L'approche proposée dans [48] se base sur les valeurs propres de la matrice de covariance empirique W_k de chaque classe. L'estimation des paramètres spécifiques par maximum de vraisemblance est réalisée comme suit:

- les d_k premières colonnes de Q_k sont estimées par les vecteurs propres associés aux d_k

plus grandes valeurs propres λ_{kj} de la matrice de covariance empirique

$$W_k = \frac{1}{n_k} \sum_{j/z_j=k} (x_j - \hat{\mu}_i)(x_j - \hat{\mu}_i)^t$$

et z_j indique le numéro de la classe de l'observation x_j.

- l'estimateur de a_{kj} est $\hat{a_{kj}} = \lambda_{kj}, j = 1, \ldots, d_k$.
- l'estimateur de b_k est $\hat{b_k} = \frac{1}{p-d_k}(trace(W_k) - \sum_{j=1}^{d_k} \lambda_{kj})$.

L'affectation d'un nouvel individu se fait grâce à la règle du maximum a posteriori.

2.2.3 Modèle de mélange sur variables qualitatives

La loi de probabilité pour modèles de mélanges sur des variables qualitatives est le plus souvent soit une loi multinomiale [27, 35, 37, 38, 39, 40] ou de Dirichlet [49]. Certains auteurs proposent l'utilisation des modèles de mélange gaussiens dans le cadre des variables qualitatives [31].

2.2.3.1 Modèle de mélange multinomial

Nous allons expliciter les lois multinomiales qui sont le plus souvent utilisées pour les variables qualitatives.

Loi binomiale: La loi binomiale modélise le nombre de succès obtenus lors de la répétition indépendante de plusieurs expériences aléatoires identiques. Elle est décrite par deux paramètres: n le nombre d'expériences réalisées et P_1 la probabilité de succès. Elle représente une somme d'épreuves de Bernoulli. La variable aléatoire prend la valeur 1 lors d'un succès et 0 sinon. La probabilité de k succès dans une répétition de n expériences est donnée par:

$$\mathbb{P}(X = k) = \binom{n}{k} P_1^k (1 - P_1)^{n-k}.$$

Loi multinomiale La loi multinomiale est une généralisation d'une loi binomiale à m résultats possibles au lieu de 2 avec $p_1, \ldots, p_m$ les probabilités correspondantes:

$$\mathbb{P}(N_1 = n_1, \ldots, N_m = n_m) = \frac{n!}{n_1! \ldots n_m!} p_1^{n_1} \ldots p_m^{n_m},$$

avec $\sum_{i=1}^{m} N_i = n$ et $\sum_{i=1}^{m} p_i = 1$.

Dans le cadre des variables qualitatives, les données sont codées suivant un tableau disjonctif complet $n \times p$. Lorsque les variables explicatives sont qualitatives, la loi naturelle proposée est la loi multinomiale. La variable aléatoire X sachant $Y = k$ suit une loi multinomiale à 2^p états possibles.

Dans les modèles de mélanges multinomiaux on fait souvent l'hypothèse d'indépendance conditionnelle des paramètres suivant les classes.

L'expression de la loi de chaque composante s'écrit par :

$$f_k(x; \theta_k) = \prod_{j=1}^{p} \prod_{h=1}^{m_j} (\theta_{kjh})^{x_{jh}}$$

avec $\theta_k = (\theta_{kjh}; j = 1, \ldots, p; h = 1, \ldots, m_j)$ tel que $\theta_{kjh} = \mathbb{P}(X = x_{jh} | Y = k)$: probabilité que la variable j présente la modalité h dans la classe k.

Les paramètres du mélange sont $\alpha = (\pi_1, \ldots, \pi_K, \theta_1, \ldots, \theta_K)$.

2.2.3.2 Estimation paramétrique: cas qualitatif

Les modèles de mélange multinomiaux exigent une estimation de $2^p - 1$ paramètres pour chaque classe. Ce nombre devient très vite énorme si le nombre de paramètres augmente. Pour pallier ce problème, des méthodes d'estimation parcimonieuses sont utilisées ou des hypothèses d'indépendance conditionnelle des paramètres sont appliquées. Nous aborderons dans les paragraphes suivants ces deux aspects.

Modèles parcimonieux multinomiaux Pour estimer les paramètres dans le cadre d'un modèle de mélange qualitatif, des modèles parcimonieux sont aussi proposés [36, 50]. Ces modèles reparamétrisent les lois conditionnelles des classes θ_{kj} en se focalisant sur une modalité majoritaire γ_{kj} et une redistribution de la masse de probabilités restante entre les autre modalités de manière équiprobable β_{kj}. Ainsi, θ_{kj} est réécrite de la manière suivante $(\beta_{kj}, \ldots, \beta_{kj}, \gamma_{kj}, \beta_{kj}, \ldots, \beta_{kj})$ avec $\gamma_{kj} \geq \beta_{kj}$ et $\beta_{kj} = \frac{1-\gamma_{kj}}{m_j - 1}$.

De plus, pour rendre majoritaire la modalité γ_{kj}, on impose que $\gamma_{kj} \geq \frac{1}{m_j}$. En se basant sur la position de la modalité majoritaire et la distribution de la masse de probabilité restante aux modalités minoritaire, les auteurs décrivent des modèles parcimonieux notés $[\varepsilon_k^j]$.

Les modèles $[\varepsilon_k^j]$ permettent ainsi une réduction de l'estimation de $m_j - 1$ paramètres par variable et par classe en une estimation d'un paramètre par variable et par classe. Pour l'estimation de ce paramètre, plusieurs hypothèses sont faites sur sa distribution par rapport aux classes et aux variables [50, 51].

Indépendance conditionnelle Pour réduire le nombre de paramètres à estimer, on suppose que les variables sont indépendantes à l'intérieur de chaque groupe. Cette hypothèse signifie que les variables sont dépendantes, mais cette dépendance entre variables est entièrement expliquée par la connaissance des groupes a priori. Cette hypothèse réduit considérablement le nombre de paramètres à estimer pour chaque groupe a priori de p variables au lieu de $2^p - 1$. Cette méthode sera utilisée dans le chapitre suivant pour estimer les paramètres.

2.3 Méthodes non paramétriques

En l'absence de toute hypothèse sur la loi des observations, certains auteurs ont proposé plusieurs types de méthodes d'estimation de la densité: les estimateurs par bases d'ondelette, l'estimation par des splines, les estimateurs à noyaux,.... Nous exposerons dans ce paragraphe les noyaux avant de se focaliser sur les méthodes utilisant ces noyaux.

2.3.1 Noyaux

Les noyaux jouent un rôle important dans les méthodes de classification et de recherche de motifs. L'élaboration et l'utilisation des noyaux s'est faite suivant des motivations différentes.

1. La première des motivations s'est faite dans le cadre de la méthode des plus proches voisins (cf paragraphe 2.3.3.2).

   On se donne un exemple d'apprentissage $\{(x_i, y_i)_{i=1,\ldots,n}\}$. Les pondérations de chaque sortie dépendent de la position relative de x_i dans $\mathcal{X}$ et du point considéré. La fonction

définissant ces pondérations notée κ est appelée fonction noyau. Une fonction de noyau entre deux points x et x' est écrite sous ce formalisme par :

$$\kappa(x, x^{'}) = f(\frac{d(x, x^{'})}{\sigma})$$

où d est une distance définie sur $\mathcal{X}$, σ est un facteur d'échelle et f est une fonction décroissante.

2. Une autre approche introduit la notion d'espace de redescription. On utilise une fonction dite de redescription pour résoudre un problème non linéaire sur $\mathcal{X}$ de telle sorte que le problème devienne linéaire dans ce nouvel espace. Cette nouvelle formulation suppose l'utilisation d'une fonction de redescription ϕ et le calcul de produits scalaires dans le nouvel espace [52].

Définition 2.3.1 *Une fonction noyau est une fonction* $\kappa : x, x^{'} \in \mathcal{X}^2 \rightarrow \mathbb{R}$ *vérifiant*

$$\kappa(x, x^{'}) = \langle \phi(x), \phi(x^{'}) \rangle$$

où ϕ *est une fonction de* $\mathcal{X}$ *vers un espace de redescription* $\mathcal{F}$ *doté d'un produit scalaire:*

$$\phi : x \mapsto \phi(x) \in \mathcal{F}.$$

Cette fonction de redescription permet de reproduire les données dans un autre espace de dimension supérieure. Les classifieurs utilisent les noyaux afin de produire un classifieur linéaire dans l'espace de redescription. Les noyaux permettent de calculer un produit scalaire dans un espace de dimension plus grande. De ce fait avec un ensemble d'exemples on peut calculer ces produits scalaires. Ce calcul conduit souvent à l'élaboration d'une matrice de ressemblances dite *matrice de Gram* G.

Définition 2.3.2 *Pour toute fonction noyau* $\kappa : \mathcal{X}^2 \rightarrow \mathbb{R}$ *et un ensemble d'exemples* $\{x_1, \ldots, x_n\} \in \mathcal{X}$*, on définit la matrice de Gram* G *comme la matrice* $n \times n$ *des produits scalaires des exemples entre eux* $G_{ij} = \kappa(x_i, x_j)$.

Cette matrice caractérise le noyau dans l'ensemble étudié. Elle vérifie des propriétés intéressantes:

Propriété 2.3.1 *Toute matrice de gram G vérifie les propriétés suivantes:*

- *G est symétrique,* $\forall(i,j) \in \{1,\ldots,n\}^2, G_{ij} = G_{ji}$,
- *G est semi définie positive,* $\forall\lambda \in Spectre(G), \lambda \geq 0$.

Plusieurs théorèmes permettent de caractériser les noyaux sans pour autant utiliser la fonction de redescription ϕ. Ainsi dans [53], on définit une fonction noyau comme suit:

Définition 2.3.3 *Une fonction $\kappa : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ est une fonction noyau si et seulement si elle est symétrique et semi-définie positive.*

En d'autres termes, la définition suivante permet de caractériser un noyau semi-défini positif.

Définition 2.3.4 *Une fonction $\kappa : \mathcal{X}^2 \rightarrow \mathbb{R}$ est un noyau semi-défini positif si*

$$\sum_{i,j=1}^{n} c_i c_j \kappa(x_i, x_j) \geq 0 \quad \forall x_1,\ldots,x_n \in \mathcal{X} \quad \forall c_1,\ldots,c_n \in \mathbb{R}.$$

De manière équivalente, une fonction noyau κ définit une matrice de Gram semi-définie positive pour tout tout ensemble de points de $\mathcal{X}$.

3. D'autres utilisent le théorème de Mercer pour caractériser les noyaux. Ce théorème fournit les conditions pour qu'une fonction symétrique soit un noyau. Elle se base sur la décomposition spectrale de la matrice G. Elle permet de l'exprimer en fonction des paramètres spectraux (vecteurs et valeurs propres).

Théorème 2.3.1 *Si $\kappa(,)$ est une fonction continue symétrique d'un opérateur intégral*

$$g(x') = Af(x') = \int_b^a \kappa(x,x')f(x)dx + h(x')$$

vérifiant:

$$\int_{\mathcal{X}\times\mathcal{X}} \kappa(x,x^{'})f(x)f(x^{'})dxdx^{'} \geq 0$$

pour toute fonction $f \in L_2(\mathcal{X})$,$\mathcal{X}$ étant un sous ensemble compact de $\mathbb{R}^d$, et h une fonction linéaire dans $\mathcal{X} = \mathbb{R}^d$ alors la fonction κ peut être développée en une série

entière uniformément convergente en fonction des valeurs propres positives λ_i et des fonctions propres ψ_i:

$$\kappa(x, x^{'}) = \sum_{j=1}^{N} \lambda_j \psi_j(x) \psi_j(x^{'})$$

où N est le nombre de valeurs propres positives.

De plus étant donné un domaine $\mathcal{X}$, on appelle noyau de Mercer une application $\kappa : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ qui produit une matrice de Gram semi-définie positive quel que soit l'ensemble de vecteurs $E = \{x_1, \ldots, x_n\}$ sur lequel elle est calculée [26, 54, 55].

Le théorème de Mercer affirme que tout noyau positif peut s'exprimer comme un produit scalaire sur l'espace d'image de la fonction de transformation Φ :

$$\kappa(x, x') = \langle \Phi(x), \Phi(x') \rangle.$$

En d'autres termes, le théorème de Mercer fournit les conditions suffisantes pour qu'une fonction symétrique κ soit une fonction noyau. De plus, il stipule l'existence d'une fonction Φ mais n'apporte pas de précisions sur sa construction de manière analytique car il n'existe pas de correspondance bijective entre le noyau κ et la fonction de transformation Φ.

4. Un autre approche utilise les approximations de fonctions dans un espace de Hilbert. Cette approche utilise des noyaux reproduisants; ce sont des noyaux dans des espaces de Hilbert (*RKHS (Reproducing Kernel Hilbert Space)*). Étant donné un ensemble $\mathcal{X}$, $\mathbb{R}^{\mathcal{X}}$ est l'ensemble des fonctions de $\mathcal{X}$ dans $\mathbb{R}$.

Définition 2.3.5 *$\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$ est un espace de Hilbert à noyau reproduisant si et seulement si*

(a) $\mathcal{F}$ est un sous-espace vectoriel de $\mathbb{R}^{\mathcal{X}}$.

(b) $\mathcal{F}$ est un espace de Hilbert , c'est-à-dire muni d'un produit scalaire $< ., . >$ et toute suite de Cauchy sur $\mathcal{F}$ est convergente et sa limite est dans $\mathcal{F}$.

(c) Pour tout $z \in \mathcal{X}$, la fonction d'évaluation linéaire $\delta_z : \mathcal{F} \rightarrow \mathbb{R}$, définie pour tout $h \in \mathcal{F}$ par $\delta_z(f) = f(z)$ est bornée.

Si $\mathcal{F}$ est un Reproducing Kernel Hilbert Space (RKHS), alors pour tout $z \in \mathcal{X}$, δ_z est une forme linéaire bornée sur $\mathcal{F}$ donc continue.

En d'autres termes étant donnée une fonction noyau, il est possible de lui associer un espace de Hilbert de fonctions régulières. Ainsi de manière explicite on énonce la définition suivante:

Définition 2.3.6 *Soit $\mathcal{F}$ un espace de Hilbert de fonctions réelles définies sur un ensemble $\mathcal{X}$:*

$$\mathcal{F} = \{\sum_{i=1}^{n} \alpha_i \kappa(x_i, \bullet) : n \in \mathbb{N}, x_i \in \mathcal{X}, \alpha_i \in \mathbb{R}, i = 1, \ldots, n\}$$

$\mathcal{F}$ est appelé espace de Hilbert à noyau reproduisant doté d'un produit scalaire noté $\langle \cdot, \cdot \rangle_{\mathcal{F}}$ s'il existe une fonction $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ ayant les propriétés suivantes:

(a) pour tout élément $x \in \mathcal{X}, \kappa(x, \bullet)$ appartient à $\mathcal{F}$, et

(b) la fonction κ est une fonction noyau reproduisant : pour toute fonction $f \in \mathcal{F}$, on a : $\langle f, \kappa(x, \bullet) \rangle_{\mathcal{F}} = \sum_{i=1}^{n} \alpha_i \kappa(x_i, x) = f(x)$.

Cette définition permet de montrer que pour un ensemble d'éléments $x_1, \ldots, x_n \in \mathcal{X}$, l'ensemble $\mathcal{F}$ de fonctions a une structure d'espace vectoriel.

La théorie des RKHS nous permet ainsi de trouver une fonction de redescription adéquate et de calculer les produits scalaires dans un espace dont la dimension peut être finie ou pas [26, 52, 54]. Sa particularité repose sur le fait qu'il ne dépend que de la fonction noyau contrairement à celui de la définition de Mercer qui dépend de la mesure définie sur $\mathcal{X}$.

Cependant avec les fonctions à noyaux, on peut effectuer les calculs explicites des produits scalaires sans pour autant utiliser la fonction de redescription. Les méthodes à noyaux permettent l'utilisation d'algorithmes linéaires pour des problèmes non linéairement séparables. L'utilisation des noyaux ne modifie pas fondamentalement la nature des données et du problème posé. L'usage de la fonction noyau prend généralement le nom de kernel trick (littéralement « l'astuce du noyau »). Son usage est décrit au niveau de la Figure 2.1. Sa popularité est due à l'apparition des Séparateurs à Vaste Marge (SVM) (paragraphe 2.3.3.1).

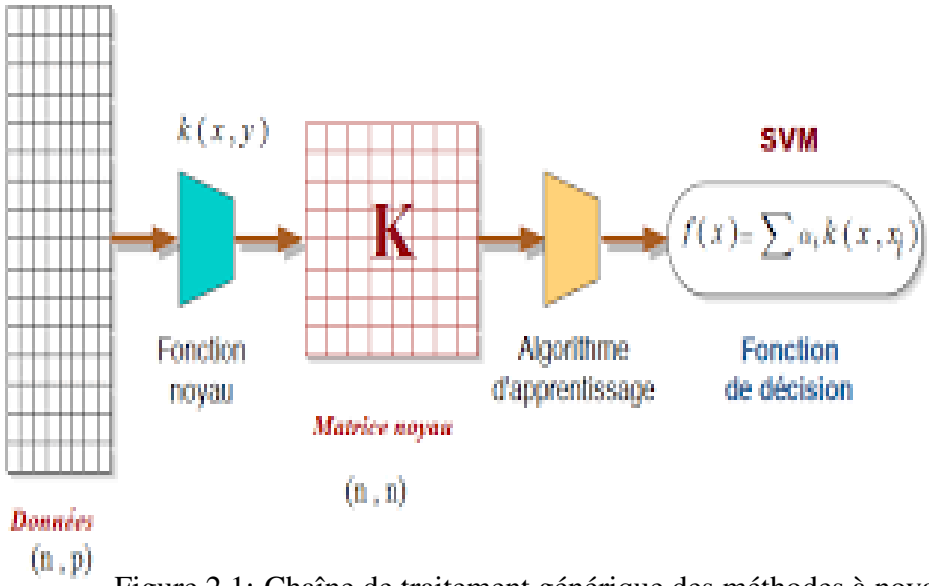


Figure 2.1: Chaîne de traitement générique des méthodes à noyaux.

La construction d'une fonction noyau appropriée est une étape importante suivant le domaine et suivant les objectifs de l'étude. Il existe plusieurs noyaux dans la littérature, nous allons en énumérer quelques uns:

- Linéaire:

$$\kappa(x, x') = \langle x, x' \rangle$$

  Il correspond au produit scalaire sans aucune transformation.

- Gaussien RBF:

$$\kappa(x, x') = \exp(-\frac{\|x - x'\|^2}{2\sigma^2})$$

  Ce noyau a la particularité d'appliquer une gaussienne sur la distance entre les exemples.

- Polynomial:

$$\kappa(x, x') = (\langle x, x' \rangle + c)^u$$

  Il permet d'appliquer le principe de maximisation de la marge aux classifieurs polynomiaux.

σ, c et u sont les paramètres du noyau.

La construction d'un nouveau noyau se fait souvent à partir des noyaux déjà définis.

Propriété 2.3.2 *Soient κ_1 et κ_2 deux fonctions noyaux :*

- *La somme de fonctions noyaux est un noyau:* $\kappa_1 + \kappa_2$.
- *Le produit de deux fonctions noyaux est un noyau:* $\kappa_1\kappa_2$.
- *La multiplication d'une fonction noyau par un réel α positif non nul est une fonction noyau:* $\alpha\kappa_1$.

Les propriétés de construction d'un noyau sont explicitées dans [26, 52, 54, 55].

2.3.2 Noyaux sur des données structurées

Ces types de noyaux sont le plus souvent utilisés dans le traitement de données textuelles. Ces noyaux permettent de prendre en compte la structure des données. Nous passerons en revue les principaux types de noyaux proposés pour des données structurées (arbres, graphes, ...).

Noyaux de convolution Le noyau de convolution est la référence des noyaux appliqués sur des données structurées telles que les arbres et les graphes. Les données structurées sont des objets décomposables en sous-objets. On se donne un ensemble d'objets composites, c'est-à-dire constitués de sous-parties. Soit $x \in \mathcal{X}$ un élément décomposable[3], on note $x_p \in \mathcal{X}_p, p = 1, \ldots, P$ l'ensemble des sous-parties de x.

Dans les noyaux de convolution, on considère l'ensemble des décompositions comme une relation $R(x_1, x_2, \ldots, x_P, x)$ c'est-à-dire que $x_1, x_2, \ldots, x_P$ constituent l'objet composite de x.

Son principe repose sur la définition d'une fonction noyau entre objets composites à partir de mesures de similarité entre leurs sous-parties. L'idée est de définir des fonctions noyau locales κ_p définies sur $\mathcal{X}_p \times \mathcal{X}_p$ puis en remontant récursivement , niveau par niveau et d'effectuer la somme des différentes contributions à la ressemblance globale [26, 54, 56].

Définition 2.3.7 *(Fonction noyau de convolution)*
Le noyau de convolution de $\kappa_1, \kappa_2, \ldots, \kappa_P$ selon la relation de décomposition R est défini par:

$$\kappa_1 * \ldots * \kappa_P(x, x') = \sum_{R} \prod_{p=1}^{P} \kappa_p(x_p, x_p^{'})$$

[3]Remarque: x n'est plus un vecteur mais représente les objets de $\mathcal{X}$

La somme est calculée selon toutes les décompositions permises par R pour décomposer l'objet x en $x_1, \ldots, x_P$ et l'objet x' en $x_1^{'}, \ldots, x_P^{'}$. La principale difficulté de ces noyaux réside dans le choix d'une relation R consistante en rapport avec le problème considéré.

Noyau p-Spectrum Ce noyau permet d'évaluer le nombre de sous ensembles de taille p que deux ensembles ont en communs. Ainsi, plus le nombre de sous-ensembles en commun est important et plus la similarité entre ces deux ensembles sera importante. Le spectre d'ordre p d'un ensemble (séquence) s peut être aussi interprété comme l'histogramme des fréquences de tous les sous-ensembles de longueur p. Le produit interne entre les spectres définit un noyau [57, 58] appelé *spectrum kernel function*. Une application pour la classification de séquences de protéines a été effectuée dans [57] avec SVM.

Noyaux sur des structures hiérarchisées Les arbres ont des structures hiérarchisées, ils représentent souvent des objets composites et structurés. La plupart des fonctions noyau sont évaluées sur le nombre de sous arbres commun entre deux arbres. On définit un arbre T comme ayant une racine et au moins un nœud fils. Le sous-arbre S de T est caractérisé comme un arbre dont les sommets et les arcs s'apparient avec les étiquettes des sommets et arcs correspondant dans T. L'un des premiers noyau utilisé sur les données structurées de type arbre est le noyau dit *Tree kernel* [26, 59].

Définition 2.3.8 *La fonction de noyau est définie comme:*

$$\kappa(T_1, T_2) = \sum_{\substack{S_1 \in \mathcal{S}(T_1) \\ S_2 \in \mathcal{S}(T_2)}} \kappa_{\mathcal{S}}(S_1, S_2)$$

avec T_1 et T_2 des arbres, S_1 et S_2 des sous-arbres respectifs de T_1 et T_2 , $\mathcal{S}(T_1)$ et $\mathcal{S}(T_1)$ l'ensemble des sous-arbres respectifs de T_1 et T_2 et $\kappa_{\mathcal{S}}(S_1, S_2)$ un noyau de base sur les sous arbres.

La fonction de noyau *Tree kernel* dépend de la nature du noyau de base $\kappa_{\mathcal{S}}(S_1, S_2)$.

De ce fait on a les spécificités suivantes:

1. Si $\kappa_{\mathcal{S}}(S_1, S_2) = \delta(S_1 = S_2)$ alors $\kappa(T_1, T_2)$ compte uniquement le nombre de sous-arbres communs.

2. Si $\kappa_{\mathcal{S}}(S_1, S_2) = \delta(S_1, S_2)\delta(n(S_1) = N)\delta(n(S_2) = N)$ alors $\kappa(T_1, T_2)$ compte le nombre de sous-arbres communs de longueur N mesurée par le nombre de nœuds $n(S_i)$ de S_i.

3. Si $\kappa_{\mathcal{S}}(S_1, S_2) = \delta(S_1, S_2)\max(D + 1 - |d(S_1) - d(S_2)|, 0)$ alors $\kappa(T_1, T_2)$ compte le nombre de sous-arbres ayant la même profondeur avec $d(S_i)$ la profondeur de la racine S_i dans l'arbre $T_i, i = 1, 2$ et D la profondeur maximale fixée,

avec $\delta()$ la fonction de Kroenecker.

Plusieurs méthodes de classification utilisent les fonctions noyaux, nous allons en énumérer quelques-unes utilisées dans ce manuscrit.

2.3.3 Méthodes à noyaux

2.3.3.1 Séparateurs à Vaste Marge (SVM)

Les Séparateurs à Vaste Marge [54, 60, 61] sont une famille de classifieurs binaires en apprentissage supervisé destinés à résoudre des problèmes de discrimination et de régression.
Les SVM sont le plus souvent appliqués dans le cas binaire $Y \in \{+1, -1\}$ mais il existe des méthodes dédiées au cas multi-classes. Les SVM permettent de trouver l'hyperplan de marge maximale, optimal du point de vue de l'apprentissage.

Cas de deux classes: Dans un premier temps, nous nous restreignons au cas binaire : $Y \in \{+1, -1\}$. Si les observations sont linéairement séparables, l'hyperplan de marge maximale $h : \mathcal{X} \to \mathcal{Y}$ prend la forme d'une équation linéaire du type $h(x) = \omega^t x + b = \langle \omega, x \rangle + b = 0$ où $\omega \in \mathbb{R}^p$ et $b \in \mathbb{R}$.
La règle de décision est donnée par : $f(x) = signe(\omega^t x + b)$.
Ainsi, un individu x_i est bien classé si $y_i(\omega^t x_i + b) > 0$ et mal classé sinon.

Le problème revient à définir un hyperplan séparateur optimal c'est-à-dire le couple de paramètres (ω, b).

Le risque empirique décrit dans le paragraphe 2.1.3 ne permet pas souvent de déterminer un couple unique de paramètres à partir de l'échantillon d'apprentissage.

L'hyperplan séparateur optimal est celui qui créé la plus grande marge (écart) entre les observations de la classe positive et celles de la classe négative.

L'algorithme des SVM consiste alors à choisir les paramètres qui maximisent un critère d'ordre géométrique. Ce critère géométrique repose sur la minimisation de $\|\omega\|$ sous les contraintes $y_i f(x_i) > 0$ avec $\|\omega\|$ représentant la distance d'un point à l'hyperplan défini par l'équation $h(x) = 0$, voir [26, 53, 54, 55].

Les multiplicateurs de Lagrange permettent de résoudre ce problème d'optimisation et conservent les points les plus proches de l'hyperplan appelés vecteurs supports.

Dans le cas où les observations ne sont pas séparables, on peut conserver l'objectif de maximiser la marge mais en utilisant des méthodes à noyaux. Dans ce cas on introduit ainsi une fonction de transformation non linéaire ϕ de l'espace des entrées $\mathcal{X}$ et un espace de redescription $\phi(\mathcal{X})$. En utilisant la formulation duale du problème d'optimisation dans [26] et les motivations du noyau décrites dans le paragraphe 2.3.1, la fonction de décision de l'hyperplan séparateur prend la forme suivante:

$$h(x) = \sum_{i=1}^{m} \alpha_i y_i \kappa(x, x_i)$$

où α_i sont les multiplicateurs de Lagrange [26, 52, 53, 54, 54, 55, 61].

Stratégies multi classes: Dans ce paragraphe, nous nous focalisons à des problèmes de discrimination à K catégories, avec $3 \leq K < \infty$.

L'adaptation des Support a Vecteurs Machines (SVM) bi classes au cas multi classes se fait généralement suivant deux stratégies:

1. *Un contre un*: Pour K classes, $\frac{K(K-1)}{2}$ classifieurs sont entraînés, chacun opposant une classe à une autre. Elle consiste à entraîner des SVM sur chacun des couples de classes, puis de décider la classe gagnante soit par un vote majoritaire soit par l'estimation des probabilités a posteriori.

2. *Un contre tous* consiste à entraîner un SVM biclasse en utilisant les éléments d'une classe contre toutes les autres classes. K classifieurs sont définis. Les nouvelles données sont classées selon la prédiction la plus forte donnée parmi tous les classifieurs. Dans la suite

du document, nous utiliserons cette approche.

2.3.3.2 Méthode des plus proches voisins (KNN)

Les (k-Nearest Neighbors en anglais) KNN sont une méthode à base de voisinage, non paramétrique.
Dans un contexte de classification d'une nouvelle observation x, l'idée fondatrice est de faire voter les plus proches voisins de cette observation. La classe de x est déterminée en fonction de la classe majoritaire parmi les k plus proches voisins de l'observation x.

Une nouvelle observation est classée au regard de ses plus proches voisins dans l'échantillon d'apprentissage. La détermination de leur similarité se base sur des mesures de distance ou des fonctions noyaux. Plusieurs stratégies sont considérées suivant le nombre de voisins:

1-NN: Pour classer un nouveau individu, on détermine le plus proche voisin dans l'échantillon d'apprentissage par $d(x, x_{(i)}) = \min_i d(x, x_i)$ avec $x_{(i)}$ le plus proche voisin de x et d la distance considérée. Les distances euclidienne et de Minkowski sont les plus utilisées.

k-NN: Pour éviter que la classe prédite ne soit déterminée que par une seule observation, on utilise les k plus proches voisins. La règle de décision s'applique aux k plus proches voisins et la classe prédite sera alors la classe majoritaire. L'estimation de la classe d'appartenance d'une nouvelle observation se fait en prenant en compte (de façon identique) les k échantillons d'apprentissage suivant une distance entre la nouvelle observation et ses voisins: la décision est en faveur de la classe majoritaire.

Choix de k: Le degré de voisinage est déterminé par le paramètre k et est choisi par l'utilisateur. Le meilleur choix de k dépend souvent du jeu de données. En général, les grandes valeurs de k réduisent l'effet du bruit sur la classification et donc le risque de sur-apprentissage, mais rendent les frontières entre classes moins distinctes. Il convient donc de faire un choix de compromis avec la variabilité associée à une faible valeur de k [62, 63, 64]. Un bon choix de k est souvent fait par validation croisée.

L'introduction des fonctions noyau permet la pondération des observations par rapport à un point de référence de sorte que plus une observation est proche de la référence, plus son poids est important. Le noyau ainsi défini doit être maximal pour des distances nulles et décroissant pour des distances assez grandes. Ce noyau est généralement symétrique.

Algorithme: On se donne un échantillon d'apprentissage $\{(x_i, y_i), i = 1, \ldots, n\}$ et une nouvelle observation x et y sa classe à prédire.

1. Sélection des $k+1$ plus proches voisins de x selon une fonction de noyau $\kappa(x, x_{(i)})$
2. La classe de x est choisie d'après le vote pondéré des k plus proches voisins.

La performance de cette méthode dépend du choix du noyau κ et du nombre de voisins k.

2.3.3.3 Processus gaussien parcimonieux en analyse discriminante (pgpDA)

Des algorithmes de classification conventionnels peuvent être modifiés grâce aux noyaux dès lors que la méthode proposée dépend des données via un produit scalaire. Le produit scalaire est simplement remplacé par une évaluation du noyau, conduisant à une transformation des algorithmes linéaires en non-linéaires. En outre, un belle propriété des algorithmes d'apprentissage à noyau est la possibilité de traiter tout type de données. La seule condition est d'être en mesure de définir une fonction définie positive sur les paires d'éléments à classer [65].

Dans [31], les auteurs proposent une méthode utilisant à la fois les avantages des modèles de mélange et des méthodes à noyaux. Cette méthode de classification vise à résoudre le problème de classification sur des données non quantitatives. C'est une méthode qui permet de prendre en compte des données à la fois qualitatives, quantitatives, fonctionnelles, en somme des données hétérogènes en général.

C'est une méthode à base probabiliste qui permet une interprétation des résultats de la classification et une méthode pouvant contenir toutes les types de données grâce aux spécificités et à la performance des noyaux. Cette méthode utilise des familles de processus gaussiens.

Considérons un échantillon d'apprentissage $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ constitué de réalisations indépendantes d'un vecteur aléatoire X binaire et où les étiquettes $\{y_1, \ldots, y_n\}$ sont des réali-

sations indépendantes d'une variable aléatoire $Y \in \{1, \ldots, K\}$ indiquant l'appartenance des observations aux K classes, ie $y_i = k$ signifie que x_i appartient à la $k$ème classe C_k.

Soit κ une fonction noyau définie par $\kappa : \{0,1\}^2 \to \mathbb{R}^+$ satisfaisant les conditions de Mercer. Pour tout $k = 1, \ldots, K$, on introduit également la fonction $\rho_k : \{0,1\}^2 \to \mathbb{R}$ définie par

$$\rho_k(x,y) = \kappa(x,y) - \frac{1}{n_k} \sum_{x_\ell \in C_k} (\kappa(x_\ell, y) + \kappa(x, x_\ell)) + \frac{1}{n_k^2} \sum_{x_\ell, x_{\ell'} \in C_k} \kappa(x_\ell, x_{\ell'}),$$

où n_k est le cardinal de la classe C_k.

On définit alors la matrice M_k par $(M_k)_{\ell,\ell'} = \rho_k(x_\ell, x_{\ell'})/n_k$ pour tout $(\ell, \ell') \in \{1, \ldots, n_k\}^2$.

On définit la règle de classification suivante : $x \to C_i$ si et seulement si $i = \arg\min_{k=1,\ldots,K} D_k(x)$ où D_k est la fonction de classification proposée par [31] :

$$\begin{aligned} D_k(x) = \frac{1}{n_k} \sum_{j=1}^{d_k} \frac{1}{\lambda_{kj}} \left(\frac{1}{\lambda_{kj}} - \frac{1}{\lambda} \right) \left(\sum_{x_\ell \in C_k} \beta_{kj\ell} \rho_k(x, x_\ell) \right)^2 + \frac{1}{\lambda} \rho_k(x,x) \\ + \sum_{j=1}^{d_k} \log(\lambda_{kj}) + (d_{\max} - d_k) \log(\lambda) - 2 \log(\pi_k). \end{aligned} \tag{2.1}$$

On a noté $\pi_k = n_k/n$, λ_{kj} la $j$ème plus grande valeur propre de la matrice M_k , β_{kj} le vecteur propre associé ($\beta_{kj\ell}$ représente sa ℓème coordonnée), $j = 1, \ldots, d_k$ et $d_{max} = \max(d_1, \ldots, d_K)$. Enfin, on a

$$\lambda = \sum_{k=1}^{K} \pi_k (\text{trace}(M_k) - \sum_{j=1}^{d_k} \lambda_{kj}) \Big/ \sum_{k=1}^{K} \pi_k (r_k - d_k)$$

et les paramètres d_k et r_k représentent respectivement la dimension intrinsèque de la classe C_k (à estimer) et une caractéristique (connue, cf [31], Table 2) du noyau κ calculé sur C_k, $k = 1, \ldots, K$. r_k est la dimension de la classe C_k une fois projetée dans un espace non linéaire avec le noyau κ. Dans la pratique, on a $r_k = \min(n_k, p)$ pour un noyau linéaire et $r_k = n_k$ pour les noyaux non linéaires.

Seuls les vecteurs propres associés aux d_k plus grandes valeurs propres de M_k doivent être estimés. Cette propriété est une conséquence de l'hypothèse cruciale de cette méthode: Les données de chaque classe C_k vivent dans un sous-espace spécifique (de dimension d_k) de l'espace (de dimension r_k) défini par le noyau κ. Cette hypothèse permet ainsi de contourner

l'inversion instable des matrices M_k, $k = 1, \ldots, K$ qui est généralement nécessaire dans les versions kernelisées des modèles de mélange gaussien, voir [66, 67, 68, 69, 70]. Dans la pratique, d_k est estimée grâce à l'éboulis des valeurs propres et au scree-test de Cattell [71]. L'estimation des autres paramètres est explicitée dans le paragraphe 2.2.2.2. La dimension ainsi sélectionnée est celle pour laquelle les différences des valeurs propres sélectionnées sont plus petites qu'un seuil t. Le seuil t peut être fourni par l'utilisateur ou sélectionné par validation croisée.

Nous renvoyons le lecteur à [31] pour plus de détails sur les fondements statistiques de cette règle de classification. Formellement, elle s'apparente à la règle de classification quadratique obtenue à partir d'un modèle de mélange gaussien dans laquelle les produits scalaires sont remplacés par une fonction de similarité non-linéaire. La fonction de classement associée à ce noyau est explicitée dans [31, 72]. La fonction de classification ainsi définie peut être appliquée sans connaissance aucune de la fonction ϕ. Il suffit de définir une fonction noyau et sa matrice de Gram associée. Des extensions de ce modèle pour des données fonctionnelles sont explicitées dans [31].

La mise en œuvre de cette méthode nécessite le choix d'une fonction noyau κ mesurant la similarité entre deux vecteurs binaires. Dans le paragraphe suivant, nous allons décrire les mesures de similarités.

2.4 Mesure de similarité

La performance de certaines méthodes de classification repose sur le choix d'une bonne mesure de similarité. Comme explicité dans [26], une fonction noyau correspond à une mesure de similarité entre deux entrées x et x'. Cette mesure de similarité est naturelle du fait que deux objets présentant des caractéristiques similaires présentent des variables cibles tout autant similaires.

Ainsi, depuis plusieurs années, des mesures de similarité sont proposées dans des domaines divers et variés. Chacune des mesures proposées présente des propriétés synthétiques et souvent spécifiques au domaine d'application. En 1901, Jaccard [73] définit une mesure de similarité qui est fortement utilisée en écologie et en biologie. Les mesures de similarités et de dissimilarités jouent un rôle essentiel dans les problèmes d'analyse de motif tels que la classification, le clustering,...

2.4.1 Définitions

Une mesure de similarité est un concept qui dépend fortement des types de données utilisées. Leur terminologie varie suivant les domaines considérés (similarité, ressemblance, proximité, ...) et suivant les types de données. La similarité et la dissimilarité permettent de mesurer un lien entre les individus d'un même ensemble. Ces indices mesurent le degré de ressemblance entre deux individus à comparer. Elles s'opposent à la distance qui mesure le degré de différence entre deux individus. Dans les distances, on cherche les individus les plus proches (distance minimale) contrairement aux mesures de similarité où on cherche plutôt les individus les plus similaires (similarité maximale).

En reprenant les concepts ensemblistes de Lerman [74, 75], on note par:

- A l'ensemble des caractéristiques présentes dans l'objet x,
- B l'ensemble des caractéristiques présentes dans l'objet $x^{'}$.
  Une mesure de similarité se définit en se basant sur quatre concepts essentiels:
  1. Nombre de caractéristiques communes aux objets:
  $$|A \cap B| = a,$$
  2. Nombre de caractéristiques possédées par x et non par $x^{'}$:
  $$|A - B| = b,$$
  3. Nombre de caractéristiques possédées par $x^{'}$ et non par x:
  $$|B - A| = c,$$
  4. Nombre de caractéristiques possédées ni par l'un ni par l'autre:
  $$|\overline{A} \cap \overline{B}| = d.$$

Définition 2.4.1 *D est un indice de dissimilarité s'il vérifie les propriétés suivantes:*

- $\forall x, x^{'} \in \mathcal{X}; D(x, x^{'}) = D(x^{'}, x)$ *(propriété de symétrie),*

- $\forall x \in \mathcal{X}; D(x, x^{'}) \geq D(x, x) = 0$ *(propriété de positivité).*

Pour un indice de dissimilarité D, le lien entre deux individus est d'autant plus fort que sa valeur est petite.

Définition 2.4.2 *Un opérateur S est dit indice de similarité s'il vérifie la propriété de symétrie et les deux propriétés suivantes:*

- $\forall \quad x, x^{'} \in \mathcal{X}; S(x, x^{'}) \geq 0$ *(propriété de positivité),*
- $\forall \quad x, x^{'} \in \mathcal{X} \quad ; S(x, x) \geq S(x, x^{'})$ *(propriété de maximisation).*

Si l'indice est celui de la similarité, le lien est d'autant plus fort que sa valeur est grande. Le passage d'un indice S à un indice D peut s'exprimer de la manière suivante:

$$\forall \quad x, x^{'} \in \mathcal{X}; D(x, x^{'}) = 1 - S(x, x^{'}).$$

2.4.2 Mesure de similarité pour des données binaires

Les données binaires correspondent à des données d'absence/présence où la présence d'une caractéristique est indiquée par le chiffre 1 et l'absence par le chiffre 0. La comparaison revient à évaluer le nombre de 1 communs ou/et le nombre de 0 communs ou/et le nombre de leurs différences.

Lorsqu'on compare deux individus, il y a toujours une caractéristique qui est absente au niveau des deux individus. Une question importante pour choisir une mesure de similarité adéquate est de savoir si le fait qu'une caractéristique soit absente pour les deux individus contribue - ou ne contribue pas - à augmenter leur similarité.

Si la présence est un fait démontrable, il est en effet plus difficile de le faire pour le caractéristique absence pour certains cas de figures et donc, de lui donner le même poids que celle d'une présence. De nombreuses mesures permettent de ne pas tenir compte des doubles-absences et d'autres permettent d'en tenir compte.

De plus le rôle des modalités d'une variable binaire occupe une place importante dans l'étude des mesures de similarité.

Mesures asymétriques Les mesures symétriques ne prennent en compte que les caractéristiques présentes dans A et/ou dans B mais ne considèrent pas les doubles absences. Tversky [76] propose une approche qui unifie plusieurs mesures de similarité de ce type. Il considère la similarité entre deux individus comme le nombre de propriétés en commun pondéré par le nombre de propriétés spécifiques à chaque individus.

Définition 2.4.3 *Similarités de Tversky:*

$$S_{Tversky}(x, x') = \frac{a}{a + \theta'(b + c)}$$

avec θ' positif.

Mesures symétriques Ces mesures prennent en compte les quatre quantités associées à un couple d'individus. Les mesures qui considèrent le double-zéro *ie* d comme une ressemblance sont dits symétriques. Ces mesures ont la particularité d'être croissantes en a et d et décroissantes en b et c. Mais, certaines mesures ne sont pas croissantes en d comme celle de Russel-Rao. Les mesures les plus générales dans cette catégorie peuvent être regroupées dans les dissimilarités de Baulieu [75, 77]:

Définition 2.4.4 *Similarités de Baulieu*

$$S_{Baulieu}(x, x') = \frac{-(b + c)}{\alpha' a + (b + c) + \beta' d}$$

avec α' et β' des réels positifs.

Les similarités sont calculées en prenant le complément de ces dernières.

Les études comparatives Avec la multitude de mesures de similarité proposées dans la littérature, certains auteurs ont tenté de faire des études comparatives sur certaines mesures proposées dans leurs domaines respectifs:

- Hubalek [78] a recueilli 43 mesures de similarité, dont 20 d'entre elles ont été utilisées pour grouper et analyser des données sur les champignons.

- Jackson et col. [79] a fait une étude comparative de 8 mesures de similarité sur des espèces de poissons.
- Zhang et col. [80] a comparé 7 mesures pour l'identification des écritures manuscrites.
- Willett [81] a évalué 13 mesures de similarité pour des données d'empreintes digitales.
- Seung-Seok et col [1]: 76 mesures de similarités binaires définies au cours du siècle dernier ont été collectées et analysées fournissant ainsi la plus vaste enquête sur ces mesures.

De plus, certaines mesures proposées dans des contextes différents se révèlent équivalentes pour des données binaires [72].

CHAPITRE 3

CLASSIFICATION SUPERVISÉE PAR MODÈLE DE MÉLANGE MULTINOMIAL POUR LES AUTOPSIES VERBALES

Sommaire

3.1 Introduction

Les données analysées dans ce chapitre représentent les caractéristiques des personnes décédées durant la période de 1985 à 2010 dans les trois sites de l'IRD (Niakhar, Bandafassi et Mlomp) du Sénégal.

Les variables explicatives notées $X = (X_j, j = 1, \ldots, p)$ représentent la présence (1) ou l'absence (0) des symptômes et des caractéristiques non symptomatiques sur un individu donné. Les diagnostics des médecins représentent la variable cible notée Y. Le jeu de données considéré ici comporte $n = 2500$ individus répartis dans $K = 18$ classes et caractérisés par $p = 100$ variables.

Le nombre élevé de variables peut être source de perturbations lorsqu'elles sont introduites dans une méthode de classification. Un problème qu'on peut rencontrer lors de la mise en œuvre des modèles de mélange concerne la taille du vecteur de paramètres à estimer. La quantité de données nécessaire à une estimation fiable peut alors être trop importante. Une solution couramment employée est la réduction de la dimension. On peut aussi choisir parmi toutes les variables disponibles celles qui apporteront le plus d'information à l'analyse et éliminer les

variables ne présentant que peu d'intérêt (sélection de variables). Ainsi, on évalue la stabilité des résultats de la classification par rapport aux différentes perturbations que les données peuvent subir.

Ici, l'aspect qualitatif de nos variables nous pousse à choisir un modèle de mélange multinomial. Pour réduire la complexité, nous avons utilisé trois pistes de réflexions.

La première est de réduire le nombre de paramètres à estimer. Le défaut des modèles multinomiaux est le nombre important de paramètres à estimer de l'ordre de $2^p - 1$ pour chaque groupe a priori. Dans notre cas, le nombre de paramètres à estimer est de l'ordre de 2.10^{31}. Pour résoudre ce problème de complexité, nous avons supposé l'indépendance conditionnelle de nos variables dans les classes.

Considérant le nombre important de variables de l'ordre de 100 et le nombre élevé de classes 18, les deux autres pistes de réflexion sont une réduction du nombre de classes et une sélection de variables afin d'en choisir les plus pertinentes.

Pour évaluer la performance de la classification sur les données d'apprentissage, nous avons calculé les taux de biens classés c'est-à-dire le taux d'accord entre le modèle proposé et les diagnostics établis par les médecins à la lecture de la fiche d'autopsie verbale. Ce taux est noté Taux Correct de Classification (TCC).

Ce chapitre s'organise comme suit. Nous présenterons tout d'abord au paragraphe 3.2, un modèle de mélange multinomial. Nous y exposerons les motivations de l'hypothèse d'indépendance conditionnelle. Dans le paragraphe 3.3, nous exposerons une méthode de réduction du nombre de classes par une méthode de $k - medoids$ sur la matrice des probabilités a posteriori des classes. Une méthode séquentielle de sélection de variables est présentée dans le paragraphe 3.4. Nous présenterons indépendamment les résultats des méthodes dans le paragraphe 3.5. Nous exposerons nos conclusions et nos perspectives dans le paragraphe 3.6.

3.2 Modèle de mélange sous hypothèse d'indépendance conditionnelle.

Pour la modélisation de données binaires, le modèle multinomial est le plus naturellement utilisé dans les modèles de mélanges. Les lois multinomiales considérées ont ainsi 2^p états possibles.

Ainsi, pour le modèle multinomial complet, il revient à estimer un nombre de $2^p - 1$ paramètres pour chaque classe qui est de l'ordre de 2.10^{31} dans notre cas. Le nombre de paramètres à estimer est ainsi considérable et présente une grande complexité. Pour réduire la complexité, nous supposons l'indépendance conditionnelle entre nos variables. L'hypothèse d'indépendance conditionnelle permet de supposer ainsi que les p variables binaires sont indépendantes à l'intérieur de chaque classe. La dépendance entre variables est expliquée par la connaissance des classes a priori. Le modèle d'indépendance conditionnelle permet d'avoir une méthode optimale en parfait accord avec la règle de décision construite [27].

Ainsi, les variables explicatives sont supposées indépendantes à l'intérieur de chaque groupe $X_i \perp X_j|Y$ pour tout $i \neq j$:

$$\mathbb{P}(X = x|Y = k) = \prod_{j=1}^{p} \mathbb{P}(X_j = x_j|Y = k).$$

De plus, cette hypothèse et le théorème des probabilités totales permettent ainsi d'obtenir la loi marginale de X :

$$\mathbb{P}(X = x) = \sum_{k=1}^{K} \mathbb{P}(Y = k) \prod_{j=1}^{p} \mathbb{P}(X_j = x_j|Y = k),$$

et selon le théorème de Bayes, les probabilités a posteriori d'appartenance aux classes s'écrivent:

$$\mathbb{P}(Y = k|X = x) \propto \mathbb{P}(Y = k) \prod_{j=1}^{p} \mathbb{P}(X_j = x_j|Y = k).$$

La variable Y est modélisée par une loi multinomiale à K niveaux de probabilités $\pi_1, \ldots, \pi_K$, c'est-à-dire $\mathbb{P}(Y = k) = \pi_k$ pour $k = 1, \ldots, K$.

Conditionnellement à $Y = k$, X_j est modélisée par une loi de Bernoulli de paramètre $\theta_{j,k}$, pour tout $j = 1, \ldots, p$ et $k = 1, \ldots, K$.

L'affection d'un individu à l'une des classes est faite par la règle du maximum a posteriori : x est affecté au groupe ℓ si et seulement si

$$\ell = \arg \max_{k=1,\ldots,K} \pi_k \prod_{j=1}^{p} \theta_{j,k}^{x_j} (1 - \theta_{j,k})^{1-x_j}.$$

L'estimation des paramètres du modèle s'effectue par la méthode du maximum de vraisemblance. Ainsi, $\theta_{j,k}$ est estimé par la proportion d'individus présentant le symptôme j parmi les individus qui ont été diagnostiqués porteurs de la maladie k.

L'estimation a priori des classes p_k a été réalisée dans deux cas différents:

- probabilités égales des classes: π_k est estimée par $\frac{1}{K}$
- probabilités différentes des classes: $\frac{n_k}{n}$ où n_k est le nombre de personnes diagnostiquées pour une cause k et n est le nombre total d'individus $n = \sum_{k=1}^{K} n_k$.

3.3 Réduction du nombre de classes

Il est à noter que la qualité d'une méthode de classification est souvent liée au choix du nombre de classes. La plupart des méthodes de classification supervisées considèrent souvent 2 classes. Dans notre cas, le nombre de classes est de 18 qui représentent les causes de décès établis par les médecins. Ce nombre de classes étant très élevé, nous avons chercher à le réduire.

Etant donné le caractère supervisé de nos données, nous avons juger utile de ne pas proposer un nouveau classement des individus. Ainsi, nous avons étudié dans quelle mesure il était possible de regrouper les classes $C_1, \ldots, C_K$ en des groupes de classes $G_1, \ldots, G_r$ avec $r << K$.

Plusieurs méthodes sont proposées pour réduire la complexité d'un classification en regroupant les classes. Parmi ces méthodes, il existe les méthodes $k - means$ et $k - medoids$. Ces méthodes utilisent un algorithme de minimisation de la somme des distances entre le point référence (point central) et les points de la classe. La stratégie de ces méthodes est de trouver r groupes homogènes parmi n objets (individus) de manière aléatoire.

Cependant l'algorithme $k - means$ utilise la moyenne des observations dans une classe comme point de référence. La moyenne étant sensible aux valeurs atypiques, si une série de données présente des valeurs aberrantes alors ces dernières peuvent considérablement modifier sa distribution.

Pour contourner ce problème, certains auteurs [82, 83, 84] préconisent l'utilisation des centres comme point de référence d'où la méthode des $k - medoids$. L'algorithme souvent

proposé est le PAM (Partitioning Around Medoids). Cet algorithme est implanté dans le package *cluster* du logiciel *R*. La stratégie proposée repose sur le classement par $k - medoids$ des probabilités a posteriori d'appartenance des individus aux classes. Chaque classe k est caractérisée par le vecteur $(\theta_{1,k}, \ldots, \theta_{p,k})^t \in \mathbb{R}^p$.

La particularité de notre méthode repose sur l'utilisation de la matrice des probabilités a posteriori:

$$\theta_{j,p} = M(x_j, y_\ell) = [\mathbb{P}(X = x_j | Y = y_\ell)]_{j,\ell}, j = 1, \ldots, p, \ell = 1, \ldots, K.$$

En utilisant cette matrice, nous effectuons un classement des prédicteurs en fonction des classes a posteriori. Nous avons effectuer une partition de causes de décès en fonction des réponses positives sur les symptômes des individus décédés. Cette partition est validée par le collège des médecins en charge des diagnostics des autopsies verbales.

L'algorithme de cette méthode est ainsi résumé comme suit:

1. Initialisation

   • Calcul de:

   $$M(x_j, y_\ell) = [\mathbb{P}(X = x_j | Y = y_\ell)]_{j,\ell}, j = 1, \ldots, p, \ell = 1, \ldots, K$$

   • Calcul de la matrice de similarité: $M' = M^t M$

   • Choisir aléatoirement r médoids qui représentent les positions centrales $\{g_1^{(1)}, \ldots, g_r^{(1)}\}$,

2. A la $t^{\text{ème}}$ itération, assigner chaque vecteur colonne $m_\ell^{'}$ de M' au cluster plus proche:

   $$G_i^{(t)} = \left\{ \mathbf{m}_\ell^{'} : \left\| \mathbf{m}_\ell^{'} - \mathbf{g}_i^{(t)} \right\| \leq \left\| \mathbf{m}_\ell^{'} - \mathbf{g}_{i^*}^{(t)} \right\| \forall\, i^* = 1, \ldots, r \right\}, i = 1, \ldots, r,$$

   avec $G_i^{(t)}$ l'ensemble des éléments du cluster i à la $t^{\text{ème}}$ itération.

3. Mettre à jour les medoids de chaque cluster :

   $$\mathbf{g}_i^{(t+1)} = \frac{1}{|G_i^{(t)}|} \sum_{\mathbf{m}_\ell^{'} \in G_i^{(t)}} \mathbf{m}_\ell^{'}, i = 1, \ldots, r$$

4. Retour en 2. jusqu'à convergence.

Le problème de choix du nombre optimal de classes et des points initiaux se pose naturellement. Pour le choix des points initiaux, nous avons adopté la méthode aléatoire tandis que sur le choix du nombre de classes optimal nous avons utilisé les largeurs des silhouettes présentées ci-dessous. La méthode dite silhouette est souvent utilisée pour le choix du nombre optimal de classes [83, 84] dans les méthodes $k - medoids$. En utilisant la fonction pam du library *cluster*, cette fonction nous informe sur les informations particulières à chaque classe, comme le nombre d'objets, le diamètre ou encore les distances maximum et minimum entre deux objets de la même classe. La silhouette renseigne sur le nombre de classes représentatives à choisir.

Elle se calcule par:

$$s(i) = \frac{D_{intra}(i) - D_{inter}(i)}{\max(D_{intra}(i), D_{inter}(i))}, i = 1, \ldots, r \text{ avec } \quad -1 < s(i) < 1$$

où $D_{intra}(i)$ est la distance moyenne entre tous les objets d'une classe, $D_{inter}(i)$ la distance moyenne entre un objet i et sa classe voisine.

A partir de la silhouette, on calcule les largeurs moyennes des silhouettes notées $S_\ell, \ell = 1, \ldots, r$ qui représentent la somme des silhouettes en chaque point divisé par le nombre d'objets dans chaque groupe. Plus la largeur moyenne des silhouettes est proche de 1 plus le nombre de classes correspondant est optimal [83] voir Tableau 3.1 pour l'interprétation de la significativité du nombre de classes en fonction de la largeur des silhouettes.

3.4 Sélection de variables

Les données issues de la méthode d'autopsie verbale sont obtenues à partir d'un protocole complexe où plusieurs étapes peuvent introduire du bruit sur les données. Etant donné que les données analysées présentent de nombreuses variables, il est nécessaire de proposer une méthode de sélection de variables.

Le problème de la sélection de variables en discrimination se rencontre généralement lorsque le nombre de variables pouvant être utilisées pour expliquer la classe d'un individu est très élevé. Le rôle de la procédure de sélection de variables consiste à sélectionner un sous ensemble de

variables permettant d'expliquer la classe. Une procédure de sélection de variables est composée de deux parties fondamentales [85]:

- Une mesure de pertinence mesurant l'influence d'un sous ensemble de variables sur la variable à expliquer,
- Un algorithme de recherche permettant de parcourir l'ensemble des sous ensembles de variables à la recherche d'un sous ensemble optimal au sens de la mesure de pertinence.

On classe les procédures de sélection de variables en deux groupes [85]: les procédures filtres et les procédures modèle-dépendantes.
Dans les procédures filtres, la sélection de variables est totalement indépendante du modèle de discrimination choisi et s'effectue en tant que traitement préalable à la phase d'estimation.
Dans le cas des procédures modèle-dépendantes, la mesure de pertinence est définie à l'aide du modèle de discrimination choisi, généralement en fonction de l'erreur du modèle sur un ensemble test.

Notre stratégie se repose sur les procédures modèle-dépendantes. Nous proposons de mesurer l'apport de chaque variable sur le modèle considéré. Les variables pertinentes sont celles dont les valeurs influencent systématiquement le taux des biens classés. Autrement dit, une variable X_i est pertinente si la connaissance de sa valeur change les probabilités sur les valeurs de la classe Y. On mesure la pertinence par rapport à une augmentation ou non du taux d'accord entre le modèle et les diagnostics établis.
Ainsi, les symptômes sont choisis de manière à prédire le groupe de diagnostic auquel pourrait appartenir un individu décédé. On pourra ainsi penser que plus l'on dispose de variables pertinentes, meilleur sera le diagnostic.

La méthodologie appliquée est la suivante: les variables sont ordonnées selon leur performance de classification lorsqu'elles sont utilisées seules. On parcourt ensuite cet ensemble ordonné et l'on retient les variables qui donnent le meilleur taux d'erreur par validation croisée. L'algorithme est le suivant:

- Entrée:

  $\mathcal{X}$: L'ensemble des variables

  $TCC(\mathcal{X}')$: Taux de Classification Correct de $\mathcal{X}'$ pour tout $\mathcal{X}' \subset \mathcal{X}$

- Sortie:

  $\mathcal{X}^*$: L'ensemble des variables sélectionnées

Étape I :

- Chercher la meilleure variable $x^* \in \mathcal{X}$ telle que $x^* = \underset{x \in \mathcal{X}}{\operatorname{argmax}} TCC(\{x\})$
- $\mathcal{X}^* = \{x^*\}$ et $\mathcal{X}_0 = \mathcal{X} - \{x^*\}$

Étape II:

- Tant que $\mathcal{X}_0 \neq \emptyset$ et $\exists x \in \mathcal{X}_0$ tel que $TCC(\mathcal{X}^* \bigcup \{x\}) > TCC(\mathcal{X}^*)$

  $x^* = \underset{x \in \mathcal{X}_0}{\operatorname{argmax}} TCC(\mathcal{X}^* \bigcup \{x\})$

  $\mathcal{X}^* = \mathcal{X}^* \bigcup \{x^*\}$ et $\mathcal{X}_0 = \mathcal{X}_0 - \{x^*\}$

- Fin tant que

3.5 Résultats

L'application du modèle multinomial sous l'hypothèse d'indépendance conditionnelle sur 18 causes de décès et 100 variables donne un taux d'accord entre le modèle et les diagnostics de l'ordre de 55.16% voir Tableau 3.5.

3.5.1 Réduction du nombre de classes

Nous allons tout d'abord mettre l'accent sur l'estimation des probabilités des classes a priori. La Figure 3.1 présente les taux de classification selon que les probabilités a priori des classes soient

supposées égales (courbe rouge) ou différentes (courbe bleue).
Il apparaît tout d'abord qu'il est préférable de ne pas se restreindre à des probabilités a priori de classes égales si le nombre de classes est élevé. L'estimation des probabilités a priori des classes différentes ou égales n'apporte pas une amélioration du taux de classification pour un petit nombre de classes. Sans surprise, plus le nombre de classes considéré est faible plus le taux de bien classés est important. L'estimation des probabilités de classes différentes ou égales n'a de sens que lorsque le nombre de classes est élevé. Pour un nombre de classes de l'ordre de 2, l'hypothèse d'estimation de probabilités différentes ou égales des classes a priori n'apporte pas de changement dans la prédiction. Pour un nombre de classes inférieur à 6, le taux de classification est supérieur à 75%. La figure 3.2 représente le nombre de classes en fonction de la largeur des silhouettes. Elle permet de déterminer le nombre de classes optimal. Les groupements en $4, 5$ et 6 classes présentent des largeurs de silhouettes supérieures à 0.3, ce qui représente les groupements de classes les plus raisonnables, voir Tableau 3.1. Selon ce critère, le nombre de classes optimal est égal à 4, voir Figure 3.2.

Largeur de la silhouette	Interprétation
$0.5 - 1$	Structure forte
$0.3 - 0.5$	Structure raisonnable
$0 - 0.30$	Structure faible
< 0	Pas de structures intéressantes

Tableau 3.1: Interprétation de la largeur des silhouettes

Nous ne commentons que les regroupements en 4, 5 et 6 causes de décès, voir Tableau 3.2.

- Dans tous les regroupements de causes, on note un groupe comprenant les maladies gastriques et un autre groupe représentant les maladies dites fébriles. Ces deux groupes sont quasiment invariants quel que soit le groupement envisagé voir en gras dans le Tableau 3.2.

- Les autres groupes varient en fonction du nombre de groupes considéré. En variant le nombre de classes dans chaque groupes de 4 à 6, on découpe de plus en plus le premier et le dernier groupe.

[1]bpp= Broncho-pneumopathie

	Groupe 1	*Groupe* 2	*Groupe* 3	*Groupe* 4	*Groupe* 5	*Groupe* 6
4 *Groupes*	cardiopathie néphropathie hépatopathie bpp[1] prostate cause-inconnue maladie-sang diabète tumeur	**digestive** **parasitaire** **malnutrition**	**fébrile** **méningite** **fièvre** **abcès**	nouveau-né épilepsie		
5 *Groupes*	cardiopathie néphropathie bpp prostate cause-inconnue	hépatopathie maladie-sang diabète tumeur	**digestive** **parasitaire** **malnutrition**	**fébrile** **méningite** **fièvre** **abcès**	nouveau-né épilepsie	
6 *Groupes*	cardiopathie néphropathie bpp prostate cause-inconnue	hépatopathie maladie-sang diabète tumeur	**digestive** **parasitaire** **malnutrition**	**fébrile** **méningite** **fièvre** **abcès**	nouveau-né	épilepsie

Tableau 3.2: Répartition des diagnostics, en gras : groupes invariants

3.5.2 Performances de la méthode

Les mesures de performance de la méthode proposée sont la précision, le rappel et le F-score, définies comme suit:

$$\text{Rappel}_k = \text{R}_k = \frac{\upsilon_{p,k}}{\upsilon_{p,k} + f_{n,k}},$$

$$\text{Précision}_k = \text{P}_k = \frac{\upsilon_{p,k}}{\upsilon_{p,k} + f_{p,k}},$$

$$\text{F-score}_k = \text{F}_k = \frac{2 \times \text{R}_k \times \text{P}_k}{\text{R}_k + \text{P}_k},$$

avec

- $\upsilon_{p,k}$ les vrais positifs: le modèle proposé diagnostique à raison des individus appartenant à la classe k,
- $f_{p,k}$ les faux positifs: le modèle proposé diagnostique à tort des individus appartenant à la classe k,
- $f_{n,k}$ les faux négatifs: le modèle proposé diagnostique à tort des individus n'appartenant pas à la classe k.

Si la précision d'une méthode est élevée, cela signifie que les causes non pertinentes sont détectées par le modèle et ce dernier est considéré comme étant précis. Une méthode idéale fournira une précision égale à 1. Si la concordance entre les diagnostics établis par le modèle et ceux de médecins est importante alors le taux de rappel est élevé. Une méthode de diagnostic optimale fournit des réponses dont la précision et le rappel sont égaux à 1.

Dans le Tableau 3.3, on résume les mesures de performances de toutes les causes de décès. Le modèle de mélange multinomial permet de capter avec une précision supérieure à 50% les causes comme la cardiopathie, l'hépatopathie, les broncho-pneumoniques, les maladies digestives, les maladies fébriles et les maladies touchant spécifiquement les nouveaux-nés. Les maladies fébriles présentent la précision la plus élevée de l'ordre de 90% mais présentent un taux de concordance de 50%. Les maladies touchant les nouveaux-nés présentent le taux de concordance le plus élevé de l'ordre de 86% c'est à dire que le modèle permet de bien détecter ces types de maladies avec une précision de 73%. A l'inverse les maladies comme abcès et tumeur présentent des taux de rappel respectivement de 44% et de 20% avec une précision respectivement de 18% et de 32%. De ce fait le modèle ne présente pas une bonne détection de ces types de maladies. Pour affiner les résultats, nous avons présenté dans le Tableau 3.4 les performances des groupes de causes obtenus avec la méthode de réduction du nombre de classes proposée. On note que les groupes invariants présentent des précisions de plus de 75% et des rappels élevés de l'ordre de 75%. En regroupant les classes, on affine la précision avec un taux de concordance élevé.

3.5.3 Sélection de variables

Au niveau du Tableau 3.5, on présente les résultats de la méthode de sélection de variables proposée. Les taux de classification de la méthode de sélection de variables sont très légèrement meilleurs que ceux de la méthode sans sélection. Ces accroissements sont faibles et varient en fonction des groupes de classes considérés. Le nombre de variables retenues ne dépend pas des groupes de classes. Ce nombre de variables varie en fonction des groupes, aucune spécificité ne se dégage sur le nombre de variables à retenir.

	Précision	Rappel	F score
cardiopathie	0.72	0.46	0.56
néphropathie	0.41	0.42	0.42
hepatopathie	0.62	0.72	0.67
bpp	0.62	0.77	0.69
prostate	0.31	0.81	0.45
maladie_digestive	0.77	0.49	0.60
maladie_fébrile	0.91	0.58	0.70
maladie_sang	0.41	0.33	0.37
maladie_nouveau_ne	0.73	0.86	0.79
maladie_parasitaire	0.26	0.58	0.36
cause_inconnu	0.23	0.64	0.34
épilepsie	0.59	0.83	0.69
diabète	0.40	0.63	0.49
méningite	0.25	0.77	0.38
fièvre	0.14	0.53	0.23
malnutrition	0.30	0.66	0.42
tumeur	0.32	0.20	0.25
abcès	0.18	0.44	0.26

Tableau 3.3: Mesures de performances des 18 causes de décès

	Sans sélection	Sélection	
Nombre de groupes (G_ℓ)	TCC	TCC	Nombre de variables retenues
18 groupes	55.16%	55.61%	43
17 groupes	55.66%	55.95%	42
16 groupes	56.05%	57.17%	47
15 groupes	57.81%	58.33%	47
14 groupes	60.01%	60.72%	41
13 groupes	59.58%	60.61%	45
12 groupes	62.33%	63.12%	53
11 groupes	65.85%	65.92%	39
10 groupes	68.25%	68.43%	49
9 groupes	69.17%	69.88%	49
8 groupes	70.70%	71.44%	54
7 groupes	75.07%	75.17%	59
6 groupes	76.93%	77.19%	64
5 groupes	73.21%	73.59%	54
4 groupes	80.08%	80.33%	56
3 groupes	79.71%	80.15%	63

Tableau 3.5: Taux de bien classés sans ou avec sélection de variables et nombre de variables retenues.

		Précision	Rappel	F score
4 *Groupes*	Groupe 1	0.84	0.76	0.80
	Groupe 2	0.75	0.75	0.75
	Groupe 3	0.85	0.77	0.81
	Groupe 4	0.36	0.92	0.52
5 *Groupes*	Groupe 1	0.81	0.74	0.77
	Groupe 2	0.48	0.54	0.51
	Groupe 3	0.78	0.71	0.74
	Groupe 4	0.86	0.76	0.81
	Groupe 5	0.37	0.92	0.53
6 *Groupes*	Groupe 1	0.81	0.75	0.78
	Groupe 2	0.49	0.56	0.52
	Groupe 3	0.78	0.72	0.75
	Groupe 4	0.86	0.77	0.81
	Groupe 5	0.59	0.92	0.72
	Groupe 6	0.26	0.93	0.40

Tableau 3.4: Performances des groupes de causes

3.6 Conclusions et perspectives

Nous avons, dans ce chapitre, présenté un modèle de mélange multinomial sur des données d'autopsies verbales. La première contribution développée est de réduire le nombre de groupes de diagnostics afin d'améliorer les résultats du modèle multinomial. Une sélection des symptômes et variables socio démographiques est établie pour augmenter la pertinence des diagnostics et diminuer les bruits dûs au protocole d'enquête. Les premiers résultats sur les données d'enquête montrent une adéquation entre les diagnostics par autopsie verbale et un modèle de mélange avec un taux de classification de 80% pour 4 groupes de causes. L'intégralité de cette procédure de réduction de classes est validée par les experts du domaine. La réduction du nombre de classes permet d'augmenter le taux de classification de manière considérable. A l'inverse la méthode de sélection de variables n'améliore pas de manière significative les résultats de la classification.

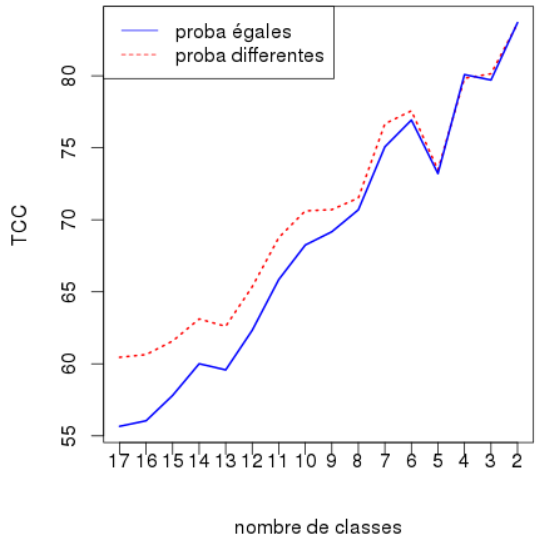


Figure 3.1: Taux de bien classés en fonction du nombre de classes

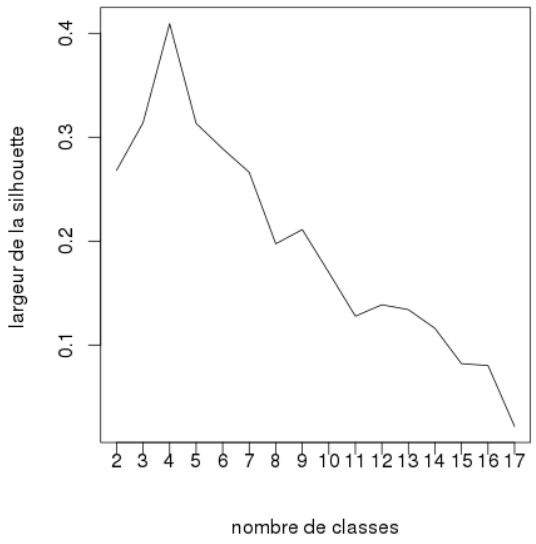


Figure 3.2: Choix du nombre optimal de classes

CHAPITRE 4

UNE MÉTHODE DE CLASSIFICATION COMBINANT MESURES DE SIMILARITÉ ET MODÈLES DE MÉLANGES

Sommaire

4.1 Introduction

La classification supervisée vise à construire une règle de décision qui attribue une observation x vivant dans une espace $\mathcal{X}$ dont la classe est inconnue à l'une des K classes connues $C_1, \ldots, C_K$. Pour la construction de ce classificateur, un ensemble de données d'apprentissage $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ est utilisé, où une observation est notée $x_i \in \mathcal{X}$ et $y_i \in \{1, \ldots, K\}$ indique la classe d'appartenance de x_i, $i = 1, \ldots, n$.

Certaines méthodes de classification supposent que les prédicteurs $\{x_1, \ldots, x_n\}$ sont des réalisations indépendantes d'un vecteur aléatoire X sur $\mathcal{X}$ et que la distribution conditionnelle des classes de X est paramétrique.

Lorsque $\mathcal{X} = \mathbb{R}^p$, il existe plusieurs distributions paramétriques possibles, mais souvent le modèle gaussien est utilisé et dans ce cas, la distribution marginale de X est donc un mélange de gaussiennes.

L'estimation des paramètres du modèle peut être obtenue par maximum de vraisemblance, voir [86]. Certaines extensions dédiées aux données de grandes dimensions sont faites voir [29, 30, 87, 88, 89, 90, 91].

Même si les méthodes de classification sont généralement appréciées pour leurs multiples avantages, elles sont souvent limitées à des données quantitatives.
Seuls quelques travaux existent pour les données catégorielles en utilisant une loi multinomiale [40] ou une loi de Dirichlet [49].

Récemment, une nouvelle méthode de classification, dénommée «processus gaussien parcimonieux en analyse discriminante» (pgpDA), a été proposée [31] pour aborder le cas des données de natures différentes. Voir, par exemple [92] pour une application à la classification des données hyperspectrales.

L'idée de base est d'introduire une fonction de noyau dans la règle de classification gaussienne. Les données hétérogènes sont prises en compte grâce à ce noyau.

Les fonctions noyau permettent une transformation des données d'origine vers un ensemble de description par l'utilisation d'un produit scalaire. Le produit scalaire est ainsi remplacé par une évaluation du noyau, ce qui conduit à une transformation des algorithmes linéaires vers des algorithmes non linéaires.
En outre, une belle propriété des algorithmes d'apprentissage par noyau est la possibilité de traiter tout type de données.

En un sens profond, une fonction de noyau $\kappa(x_\ell, x'_\ell)$ correspond à une mesure de similarité entre x_ℓ et x'_ℓ, mesure qu'il est naturel utiliser en induction puisqu'un a priori évident est de supposer que deux entrées similaires doivent être associées à des sorties similaires [26]. Pour un ensemble de points $\{x_\ell, \ell = 1, \ldots, n\}$, on peut calculer la matrice de Gram (matrice noyau) dont les éléments sont $M_{\ell,\ell'} = \kappa(x_\ell, x'_\ell)$. La seule condition pour que κ soit un noyau est que la matrice de Gram soit semi définie positive sur les paires d'éléments à classer [65].

Nous proposons l'utilisation d'une fonction noyau généralisée sur des données binaires. Ainsi, ce chapitre expose l'introduction d'une famille de noyaux exponentiels dans les méthodes

de classification supervisée sur des prédicteurs binaires. Son originalité est de combiner une règle de classification basée sur un modèle de mélange avec des mesures de similarité grâce à une nouvelle famille de noyaux exponentiels. À cette fin, nous montrons comment les nouveaux noyaux peuvent être construits en se basant sur des mesures de similarité ou de dissimilarité. En particulier, 76 mesures sont considérées. Certains liens sont établis entre ces mesures quand elles sont appliquées sur les prédicteurs binaires. Une nouvelle famille de mesures est également introduite pour unifier les mesures existant dans la littérature.

En conséquence, nous nous retrouvons avec une nouvelle méthode de classification supervisée dédiée à des prédicteurs binaires combinant les mesures de similarité et les modèles de mélange.

La performance de la nouvelle méthode de classification est illustrée sur deux ensembles de données réelles (données d'autopsie verbale et les données de chiffres manuscrits) en utilisant 76 mesures de similarité différentes.

Ce chapitre est organisé de la manière suivante. Le principe de pgpDA est expliqué dans le paragraphe 4.2. Un examen des mesures de similarité et de dissimilarité est proposé dans le paragraphe 4.3. La construction de nouveaux noyaux à partir de mesures de similarité est présentée dans le paragraphe 4.4. La méthode est illustrée sur des données réelles dans le paragraphe 4.5 et quelques observations finales sont fournies dans le paragraphe 4.6. Les preuves sont reportées au chapitre 8.

4.2 Classification grâce à une fonction noyau

Ce paragraphe présente une méthode de classification utilisant des noyaux, la méthode pgpDA.

Considérons un échantillon d'apprentissage $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ constitué de réalisations indépendantes d'un vecteur aléatoire X binaire et où les étiquettes $\{y_1, \ldots, y_n\}$ sont des réalisations indépendantes d'une variable aléatoire $Y \in \{1, \ldots, K\}$ indiquant l'appartenance des observations aux K classes, ie $y_i = k$ signifie que x_i appartient à la $k$ème classe C_k.

Soit κ une fonction noyau définie par $\kappa : \{0,1\}^2 \to \mathbb{R}^+$ satisfaisant les conditions de Mercer voir paragraphe 2.3.1. Pour tout $k = 1, \ldots, K$, on introduit également la fonction

$\rho_k : \{0,1\}^2 \to \mathbb{R}$ définie par

$$\rho_k(x,y) = \kappa(x,y) - \frac{1}{n_k}\sum_{x_\ell \in C_k}(\kappa(x_\ell, y) + \kappa(x, x_\ell)) + \frac{1}{n_k^2}\sum_{x_\ell, x_{\ell'} \in C_k}\kappa(x_\ell, x_{\ell'}),$$

où n_k est le cardinal de la classe C_k. On définit alors la matrice M_k par $(M_k)_{\ell,\ell'} = \rho_k(x_\ell, x_{\ell'})/n_k$ pour tout $(\ell, \ell') \in \{1, \ldots, n_k\}^2$. On définit la règle de classification suivante : $x \to C_i$ si et seulement si $i = \arg\min_{k=1,\ldots,K} D_k(x)$ où D_k est la fonction de classification proposée par [31] :

$$\begin{aligned} D_k(x) = &\frac{1}{n_k}\sum_{j=1}^{d_k}\frac{1}{\lambda_{kj}}\left(\frac{1}{\lambda_{kj}} - \frac{1}{\lambda}\right)\left(\sum_{x_\ell \in C_k}\beta_{kj\ell}\rho_k(x, x_\ell)\right)^2 + \frac{1}{\lambda}\rho_k(x,x) \\ &+ \sum_{j=1}^{d_k}\log(\lambda_{kj}) + (d_{\max} - d_k)\log(\lambda) - 2\log(\pi_k). \end{aligned} \tag{4.1}$$

On a noté $\pi_k = n_k/n$, λ_{kj} la $j$ème plus grande valeur propre de la matrice M_k , β_{kj} le vecteur propre associé ($\beta_{kj\ell}$ représente sa ℓème coordonnée), $j = 1, \ldots, d_k$ et $d_{max} = \max(d_1, \ldots, d_K)$. Enfin, on a

$$\lambda = \sum_{k=1}^{K}\pi_k(\text{trace}(M_k) - \sum_{j=1}^{d_k}\lambda_{kj}) \Big/ \sum_{k=1}^{K}\pi_k(r_k - d_k)$$

et les paramètres d_k et r_k représentent respectivement la dimension intrinsèque de la classe C_k (à estimer) et une caractéristique (connue, cf [31], Table 2) du noyau κ calculé sur C_k, $k = 1, \ldots, K$, pour plus de détails se référer au paragraphe 2.3.3.3 .

En pratique, on a $r_k = \min(n_k, p)$ pour des noyaux linéaires et $r_k = n_k$ pour des noyaux non linéaires, voir [31], Table 2 pour plus d'exemples.

En outre, nous soulignons que seuls les vecteurs propres associés aux d_k plus grandes valeurs propres de M_k sont estimés.

Cette propriété est une conséquence de l'hypothèse cruciale de cette méthode: Les données de chaque classe C_k vivent dans un sous-espace spécifique (de dimension d_k) de l'espace (de dimension r_k) défini par le noyau κ. Cette hypothèse permet de contourner l'inversion instable des matrices M_k $k = 1, \ldots, K$ qui est généralement nécessaire dans les versions "kernelized" de modèles de mélange gaussien, voir par exemple [66, 67, 68, 69, 70].

Dans la pratique, d_k est estimée grâce au test de Cattell [71] qui identifie une cassure sur

l'éboulis des valeurs propres. La dimension sélectionnée est celle pour laquelle les différences de valeurs propres sont plus petites qu'un seuil t.
Le seuil t peut être fourni par l'utilisateur ou sélectionné par validation croisée.

La mise en œuvre de cette méthode nécessite la sélection d'une fonction noyau κ qui mesure la similitude entre deux vecteurs binaires. La remarque d'invariance suivante peut être faite:

Lemme 4.2.1 *Soit κ une fonction symétrique définie positive $\kappa : \{0,1\}^p \times \{0,1\}^p \to \mathbb{R}^+$. Alors, pour tout $\eta > 0$ et $\mu \in \mathbb{R}$, les règles de classification associées à κ et $\tilde{\kappa} := \eta\kappa + \mu$ par (4.1) sont identiques.*

En conséquence, pour définir une méthode de noyau appropriée [65], il suffit de trouver une version décalée de κ qui est une fonction définie positive *ie*

$$\exists \mu \in \mathbb{R}, \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j [\kappa(x_i, x_j) + \mu] \geq 0 \ , \ \forall n \in \mathbb{N}, \ (c_i, c_j) \in \mathbb{R}^2, \ (x_i, x_j) \in \{0,1\}^p \times \{0,1\}^p. \quad (4.2)$$

Formellement, la méthode pgpDA s'apparente à la règle de classification quadratique obtenue à partir d'un modèle de mélange gaussien dans laquelle les produits scalaires sont remplacés par une fonction de similarité non-linéaire. La mise en œuvre de cette méthode nécessite le choix d'une fonction noyau κ mesurant la similarité entre deux vecteurs binaires.

4.3 Mesures de similarité et dissimilarité

La performance de certaines méthodes de classification repose sur le choix d'une bonne mesure de similarité [1, 93]. Historiquement, toutes les mesures binaires observées ont une performance significative dans leurs domaines respectifs. Ainsi depuis longtemps, plusieurs mesures de similarité sont proposées dans des domaines divers. De plus dans [1], 76 mesures de similarités binaires ont été collectées et analysées fournissant ainsi la plus vaste enquête sur ces mesures.

Dans la suite du document, nous ne considérons que des mesures de similarité appliquées sur des données binaires. Pour ces type de données, la comparaison s'appuie essentiellement sur le nombre de caractéristiques communes entre les individus et sur le nombre de caractéristiques qui les distinguent.

Notations Soient x_ℓ et $x_{\ell'}$ deux individus dans $\{0,1\}^p$ que l'on souhaite comparer.

Soit $a =< x_\ell, x_{\ell'} >$ et $d =< \mathbf{1} - x_\ell, \mathbf{1} - x_{\ell'} >$ respectivement le nombre de 1 et 0 communs entre les deux vecteurs ($\mathbf{1}$ désigne le vecteur de $\mathbb{R}^p$ dont toutes les composantes sont égales à 1). De même, on introduit $b =< \mathbf{1} - x_\ell, x_{\ell'} >$ et $c =< x_\ell, \mathbf{1} - x_{\ell'} >$ avec $a + b + c + d = p$.

L'existence de plusieurs mesures de similarités peut être un désavantage, tant dans le cadre de leur utilisation que de leur utilité. C'est dans ce cadre que certains auteurs ont cherché à faire des études de comparaisons des mesures de similarités dans des domaines variés [78, 79, 80, 81]. Il apparaît aussi que pour des variables binaires, plusieurs mesures de similarités sont équivalentes. Par exemple, la mesure de Hamming [1] eq. (15) est équivalente à celle de [1] eq. (17)–(23).

C'est dans cette optique, que nous avons présenté dans [72], une généralisation de plusieurs mesures de similarités.

4.3.1 Généralisation des mesures de similarités

Ici, nous proposons d'unifier la plupart des mesures proposées dans la littérature par l'introduction de la similarité suivante :

$$S(x, x') = \frac{\alpha a - \theta(b + c) + \beta d}{\alpha' a + \theta'(b + c) + \beta' d} \tag{4.3}$$

où $\alpha \geq 0$, $\beta \geq 0$, $\theta \geq 0$, $(\alpha', \beta') \in \mathbb{R}^2$ et $\theta' \neq 0$.

Cette généralisation permet de prendre en compte suivant les valeurs des paramètres de (4.3), les 4 quantités associées à un couple d'individus, à savoir, leur intersection a, leur différence $b + c$ et l'intersection de leurs complémentaires d.

La Table 4.1 résume 28 mesures de similarités de [1] réécrites dans notre formalisme (4.3).

Ainsi, le contrast model de Tversky [76] et la similarité de Baulieu [77] s'avèrent être des cas particuliers de (4.3).

Le contrast model de Tversky [76] peut être réécrit:

$$S_{\text{Tversky}}(x, x') = \frac{a}{a + \theta'(b + c)}$$

et qui est un cas particulier de (4.3) avec $\alpha = \alpha' = 1$ et $\theta = \beta = \beta' = 0$.

La similarité de Baulieu [77] définie par:

$$S_{\text{Baulieu}}(x, x') = \frac{-(b + c)}{\alpha' a + (b + c) + \beta' d}$$

peut être obtenue avec (4.3) pour $\alpha = \beta = 0$ et $\theta = \theta' = 1$.

Certaines mesures de [1] n'entrent pas dans notre formalisme (4.3) mais s'avèrent équivalentes: La mesure de Forbesi [1] eq. (34) est équivalente à celle de Cosine [1] eq. (31) , Kulczynski-II [1] eq. (41), Driver & Kroeber [1] eq. (42) et Johnson [1] eq. (43) sont équivalentes, Ochia1 [1] eq. (33) est équivalent à Otsuka [1] eq. (38), Hellinger [1] eq. (29) est équivalent à Chord [1] eq. (30) et Tarantula [1] eq. (75) est équivalent à Ample [1] eq. (76). Lors des conceptions des mesures, l'introduction du nombre de zéros communs (d) dans les mesures de similarités est discuté dans [94, 95]. Ainsi pour certains, l'introduction de d s'avère une nécessité pour d'autres non. Dans notre cas, elle permet de comparer des absences de symptômes qui est une information utile en matière de diagnostic. La valeur de d peut être aussi utile par exemple lorsque la règle de classification dépend du codage des données, comme nous le montrons dans le Lemme 4.4.1 ci-dessous. C'est dans cette optique, que nous avons défini dans le paragraphe 4.3.2 une nouvelle mesure de similarité comme étant une combinaison de l'absence et de présence d'attributs sur des individus à comparer.

4.3.2 Similarité de Sylla & Girard

Nous considérons la similarité suivante qui est aussi un cas particulier de (4.3):

$$S_{\text{Sylla \& Girard}}(x, x') = \alpha a + (1 - \alpha)d, \tag{4.4}$$

pour $\theta = 0$, $\beta = 1 - \alpha$ et $\alpha' = \beta' = \theta' = 1/p$. Cette nouvelle mesure peut être interprétée comme une extension de la mesure Intersection [1] eq. (12) et de celle de Russell & Rao [1] eq. (14) correspondant au cas où $\alpha = 1$. La nouvelle mesure $S_{\text{Sylla \& Girard}}$ peut aussi être vue comme une

Nom	α	θ	β	α'	θ'	β'	**équation**
Jaccard	1	0	0	1	1	0	(1)
Tanimoto	-	-	-	-	-	-	(65)
Dice	2	0	0	2	1	0	(2)
Czekanowski	-	-	-	-	-	-	(3)
Nei & li	-	-	-	-	-	-	(5)
3w-Jaccard	3	0	0	3	1	0	(4)
Sokal & Sneath-I	1	0	0	1	2	0	(6)
Sylla & Girard	α	0	$1-\alpha$	1	1	1	
Sokal & Michener	1	0	1	1	1	1	(7)
Innerproduct	-	-	-	-	-	-	(13)
Sokal & Sneath-II	2	0	2	2	1	2	(8)
Gower & Legendre	-	-	-	-	-	-	(11)
Roger & Tanimoto	1	0	1	1	2	1	(9)
Faith	1	0	0.5	1	1	1	(10)
Intersection	1	0	0	1	1	1	(12)
Russell & Rao	-	-	-	-	-	-	(14)
Hamming*	0	1	0	1	1	1	(15)
Squared-Euclid*	-	-	-	-	-	-	(17)
Canberra*	-	-	-	-	-	-	(18)
Manhattan*	-	-	-	-	-	-	(19)
Mean-Manhattan*	-	-	-	-	-	-	(20)
Cityblock*	-	-	-	-	-	-	(21)
Minkowski*	-	-	-	-	-	-	(22)
Vari*	-	-	-	-	-	-	(23)
Lance & Williams*	0	1	0	2	1	0	(27)
Bray & Curtis*	-	-	-	-	-	-	(28)
Sokal & Sneath-III	-1	0	-1	0	1	0	(56)
Kulczynski-I	-1	0	0	0	1	0	(64)
Hamann	1	1	1	1	1	1	(67)

Tableau 4.1: Mesures de similarité réécrites dans le formalisme (4.3). Les mesures marquées (*) sont obtenues en prenant l'opposé de la mesure de dissimilarité associée. La dernière colonne fait référence au numéro de l'équation dans [1].

extension des mesures de Sokal & Michener [1] eq. (7) et Innerproduct [1] eq. (13) correspondent au cas où $\alpha = 1/2$.

Le paramètre α dans $S_{\text{Sylla \& Girard}}$ permet d'équilibrer les poids associés aux nombres de 1 communs et de 0 communs. Les performances de cette mesure sur les deux jeux de données sont exposées dans le paragraphe 4.5.4.

4.4 Construction de noyaux associés à des observations binaires

Le but de ce paragraphe est de construire des noyaux adaptés aux données binaires à partir des mesures de similarité et de dissimilarité présentées dans le paragraphe 4.3. Les noyaux peuvent

être utilisés dans les règles de classification décrites dans le paragraphe 2.3.1 pour construire de nouvelles méthodes de classification conçues pour les données binaires. Dans un premier temps, nous considérons le cas du noyau linéaire et du noyau Radial Basic Function (RBF). Nous montrons ensuite dans un second temps comment le noyau RBF peut être étendu à une catégorie plus large de noyaux exponentiels.

Noyaux linéaires. Le noyau linéaire est la plus simple des fonctions noyaux. Elle est donnée par $\kappa_{\text{linéaire}}(x_\ell, x_{\ell'}) =< x_\ell, x_{\ell'} >= a$, ce qui, dans le cas binaire, revient à comptabiliser le nombre de 1 communs entre x_ℓ et $x_{\ell'}$.

Il est démontré (voir [31], Proposition 3) que la règle de classification associée (4.1) est quadratique et peut donc être interprétée comme un cas particulier de la méthode HDDA [29]. Le Lemme 4.4.1 suivant montre que la règle de classification associée avec un noyau linéaire est indépendante du codage des données.

Lemme 4.4.1 *Soit x_ℓ, $x_{\ell'} \in \{0,1\}^p$ et soit $\tilde{\kappa}_{\text{linéaire}}(x_\ell, x_{\ell'}) =< \mathbf{1} - x_\ell, \mathbf{1} - x_{\ell'} >= d$ (Ce noyau compte le nombre de 0 communs entre x et x'). Les règles de classification (4.1) associées à $\kappa_{\text{linéaire}}$ et $\tilde{\kappa}_{\text{linéaire}}$ sont équivalentes.*

Noyaux exponentiels. Le noyau de type exponentiel le plus connu est le noyau RBF :

$$\kappa_{\text{RBF}}(x_\ell, x_{\ell'}) = \exp\left(-\frac{\|x_\ell - x_{\ell'}\|^2}{2\sigma^2}\right),$$

où σ est un paramètre positif.
Pour des observations binaires, on peut réécrire ce noyau avec une mesure de Hamming (cf Table 4.1 ou [1] eq.(15)):

Lemme 4.4.2 *Soit $x_\ell, x_{\ell'} \in \{0,1\}^p$ alors*

$$\kappa_{RBF}(x_\ell, x_{\ell'}) = \exp\left(\frac{S_{Hamming}(x_\ell, x_{\ell'})}{2\sigma^2}\right).$$

Il apparaît ainsi que le noyau RBF peut être retrouvé en utilisant une mesure de similarité pour des données binaires.

Nous proposons d'étendre cette construction à toutes les mesures de similarité S du Tableau 4.1 et plus généralement aux 76 mesures recensées dans [1] en posant :

$$\kappa(x_\ell, x_{\ell'}) = \exp\left(\frac{S(x_\ell, x_{\ell'})}{2\sigma^2}\right). \tag{4.5}$$

Dans la pratique, S peut être choisie comme (4.3), (4.4), ou plus généralement dans l'ensemble des 76 mesures S décrit dans [1].

Le résultat suivant est l'analogue du Lemme 4.2.1 pour les mesures de similarité.

Lemme 4.4.3 *Soit S une mesure de similarité $S : \{0,1\}^p \times \{0,1\}^p \to \mathbb{R}^+$. Alors, pour tout $\eta > 0$ et $\mu \in \mathbb{R}$, les règles de classification associées à S et $\tilde{S} := \eta S + \mu$ suivant (4.1) et (4.5) sont équivalentes.*

Clairement, deux mesures de similarité ne différant que par des constantes multiplicatives ou additives donneront des règles de classification équivalentes.

Le résultat suivant démontre que tout noyau défini à partir de (4.5) et (4.3) vérifie les conditions de (4.2).

Proposition 4.4.1 *Pour tout $\alpha \geq 0$, $\beta \geq 0$, $\theta \geq 0$, $(\alpha', \beta') \in \mathbb{R}^2$ et $\theta' \neq 0$, la famille de noyaux exponentiels:*

$$\kappa(x, x') = \exp\left(\frac{1}{2\sigma^2}\frac{\alpha a - \theta(b+c) + \beta d}{\alpha' a + \theta'(b+c) + \beta' d}\right)$$

définit un noyau positif.

Par la suite, nous considérons le cas particulier suivant:

$$\kappa_{\text{Sylla \& Girard}}(x, x') = \exp\left(\frac{\alpha a + (1-\alpha)d}{2\sigma^2}\right), \tag{4.6}$$

où $\theta = 0$, $\beta = 1 - \alpha$ et $\alpha' = \beta' = \theta' = 1/p$.

4.5 Applications

4.5.1 Données

Autopsie Verbale Une liste de p symptômes possibles est établie et les données recueillies $X = (x_1, \ldots, x_p)$ sont composées de l'absence ou de la présence (codées 0 ou 1) de chaque symptôme sur la personne décédée. La cause probable est donnée par un médecin et est codée comme une variable aléatoire qualitative Y.

Nous nous référons à [96] pour un examen des méthodes automatiques pour l'attribution des causes de décès Y suivant les données d'autopsie verbale X. En particulier, des méthodes de classification basées sur la règle de Bayes ont été proposées, voir [11].

Ici, nous nous concentrons sur des données mesurées sur des personnes décédées durant la période de 1985 à 2010 dans les trois sites de l'IRD (Institut de recherche pour le développement) (Niakhar, Bandafassi et Mlomp) du Sénégal. L'ensemble de données comprend $n = 2500$ individus (personnes décédées) répartis dans $K = 18$ classes (causes de décès) et caractérisés par $p = 100$ variables. Ces variables représentent les symptômes et les variables socio-démographiques (âge, sexe, saison pluies, saison sèche, ...).

Chiffres manuscrits binarisés La reconnaissance de chiffres manuscrits est un des problèmes les plus populaires pour l'évaluation et la comparaison de classifieurs, avec par exemple, une application évidente dans les services postaux.
Ici, nous nous concentrons sur la base de données US Postal Service (USPS) des chiffres manuscrits qui se compose de $n = 9298$ images 16×16 en niveaux de gris [97]. L'ensemble des données est disponible en ligne à `http://yann.lecun.com/exdb/mnist`.
Le vecteur aléatoire X est une image numérisée et est représentée comme un vecteur de dimension p avec $p = 256$. Chaque pixel d'une image 16×16 est représenté en blanc ou en noir. La classe à prédire Y est le chiffre avec $K = 10$. Un échantillon extrait de l'ensemble des données est représenté sur la Figure 4.1.

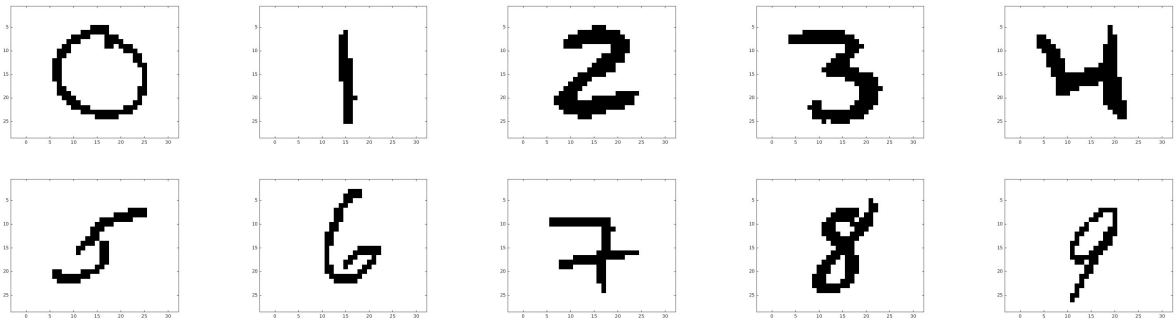


Figure 4.1: Un échantillon binarisé de chiffres manuscrits.

4.5.2 Méthodologie

La mise en œuvre de la méthode de classification nécessite la sélection des hyper paramètres $\omega = (t, \sigma)$ où t est le seuil (voir paragraphe 4.2) et σ est le paramètre du noyau voir équation (4.5). A cette fin, une technique de double validation croisée est utilisée.

L'échantillon total de taille n est subdivisé aléatoirement $M = 50$ fois en un échantillon apprentissage $\mathcal{L}_m$ de taille τn et en un échantillon test $\mathcal{T}_m$ de taille $(1 - \tau)n$ avec $\tau \in (0, 1)$ représentant une proportion et $m = 1, \ldots, M$.

Les paramètres α, σ du noyau et d_i sont choisis par validation croisée sur le jeu d'apprentissage (la dimension d_i est déterminée par l'intermédiaire d'un `seuil` sur la variance cumulée, voir [31] pour plus de détails) : 5 fois consécutivement, 100 individus sont retirés aléatoirement de l'échantillon d'apprentissage, et les paramètres (α, σ, `seuil`) sont estimés par maximisation du taux de bien classés sur les 100 individus retirés. Le taux de bien classés global est estimé sur l'échantillon test en répétant l'ensemble du procédé 50 fois.

De ce fait, sur chaque échantillon d'apprentissage $\mathcal{L}_m$, l'hyperparamètre optimal $\hat{\omega}_m$ est estimé par validation croisée pour $m = 1, \ldots, M$.

De plus, l'hyperparamètre optimal $\hat{\omega}$ est calculé comme le mode empirique de l'ensemble $\{\hat{\omega}_1, \ldots, \hat{\omega}_M\}$. Le taux moyen des biens classés est calculé sur l'échantillon l'apprentissage $\mathcal{L}_m, m = 1, \ldots, M$ et sur l'échantillon test $\mathcal{T}_m, m = 1, \ldots, M$.

4.5.3 Résultats obtenus avec 76 noyaux de [1]

Le but de ce paragraphe est de comparer les performances des méthodes de classification obtenues en combinant les 76 mesures de similarité et de dissimilarité présentées dans [1] avec le noyau exponentiel (4.5). Par souci d'exhaustivité, les résultats obtenus avec le noyau Sylla & Girard sont également présentés. Les résultats de classification sont résumés dans le Tableau 4.2 pour $\tau = 63\%$ sur l'ensemble des données d'apprentissage et de test.

Seuls les résultats associés aux 18 meilleurs noyaux (en termes de Taux Correct de Classification TCC calculés sur l'ensemble de test) sont rapportés. Il semble que ces noyaux ont de bons résultats de classification sur les deux ensembles de données avec TCC $\in [83\%, 88\%]$. Il est également intéressant de noter que 9 noyaux sur les 76 de [1] apparaissent parmi les 18 meilleurs sur les deux jeux de données, c'est le cas de: Euclid, Pearson, Hellinger, Dice, 3w-Jaccard, Orchia1, Gower & Legendre, Roger & Tanimoto et RBF.
Relevons également que le noyau Sylla & Girard devrait également être inclu, sur la liste des 10 noyaux présentant de bons résultats sur les deux ensembles de données.

Les résultats sont consignés dans le Tableau 4.2 et sont ordonnés suivant les valeurs de TCC calculé sur l'échantillon test. L'échantillon test inclut $\tau = 63\%$ des individus de l'ensemble des données initial. La dernière colonne du Tableau 4.2 se réfère au numéro de l'équation dans [1].

4.5.4 Résultats obtenus avec le noyau Sylla & Girard

Dans cette partie, nous mettons en lumière l'application de la mesure Sylla & Girard (4.4). Cette première expérience a pour but de souligner l'influence de la valeur de α sur le noyau développé. Les TCC sont calculés pour $\alpha \in \{0, 0.1, \ldots, 1\}$ et pour τ par double validation croisée comme décrit dans le paragraphe précédent. Il apparaît en premier lieu sur la Figure 4.2 que les courbes de taux de classification ne sont pas symétriques par rapport à $\alpha = 0,5$. Cela signifie que le codage des observations a un effet sur le classement. Ceci est différent du cas du noyau linéaire, voir le Lemme 4.4.1.

Cependant, dans les deux cas considérés, $\alpha = 0.1$ permet d'avoir des performances supérieures au noyau RBF associé à $\alpha = 0.5$. Ainsi, la sélection d'une valeur optimale de α est d'intérêt. Elle peut être facilement réalisée en introduisant α comme un hyper paramètre supplémentaire dans ω et donc en le sélectionnant par double validation croisée. Enfin, on peut

noter que certaines valeurs de α donnent lieu à des TCC assez élevés sur l'échantillon test. En particulier, un TCC de 87% est atteint sur des données complexes de type autopsie verbale pour un $\tau = 78\%$.
A titre de comparaison, une classification basée sur un modèle de mélange multinomial sous l'indépendance conditionnelle présente des taux de classification de l'ordre de 55.16% voir Tableau 3.5.
La figure 4.2 est une comparaison des taux de classification moyens entre RBF (Hamming) et $S_{\text{Sylla \& Girard}}$ appliqués sur des données d'autopsies verbales et de données de chiffres manuscrits. Elle montre les TCC classés selon les proportions d'échantillons de test et d'apprentissage. On note que la taille de l'échantillon influe sur les taux de classification.
De ce fait pour des données de types autopsie verbale, un bon échantillonnage est nécessaire pour disposer des taux de classification élevés. Mais avec le nombre élevé de classes, il est souvent difficile d'avoir un échantillon d'une grande taille pour chaque classe considérée. Nous notons en moyenne un pourcentage important de bien classés de l'ordre de 80% pour l'échantillon apprentissage et de 70% pour l'échantillon test. Les résultats obtenus avec le noyau RBF ($\alpha = 0.5$) sont en rouge. A gauche, nous avons les TCC sur l'ensemble d'apprentissage et à droite celui du test. En haut, on a les données d'autopsie verbale et en bas les données de chiffres manuscrits.

4.5.5 Comparaison entre méthodes de classification

Enfin, la méthode de classification proposée est comparée avec la méthode des Forêts Aléatoires (`package RandomForest`,Version 4.6.10 du logiciel `R`) et avec la méthode de SVM (`bibliothèque libsvm`, version 3.2 de `Matlab`). La méthode `un contre tous` de SVM est utilisée. Nous nous limitons à la mise en œuvre du noyau Sylla & Girard dans les méthodes pgpDA et SVM.

Il apparaît ainsi dans le Tableau 4.5 que sur l'ensemble de données d'autopsie verbale, la méthode pgpDA présente des résultats significativement meilleurs que les méthodes SVM et Forêts Aléatoires sur l'ensemble de test. De plus le TCC obtenu avec Forêts Aléatoires est plus élevé pour l'échantillon d'apprentissage, on peut ainsi soupçonner un sur-apprentissage des données. On peut aussi observer que le TCC associé à ppgDA est légèrement dépendant de α (TCC $\in [82.08\%, 86.89\%]$) tandis que le TCC associé à SVM est très sensible à α (TCC $\in [62.67\%, 80.00\%]$).

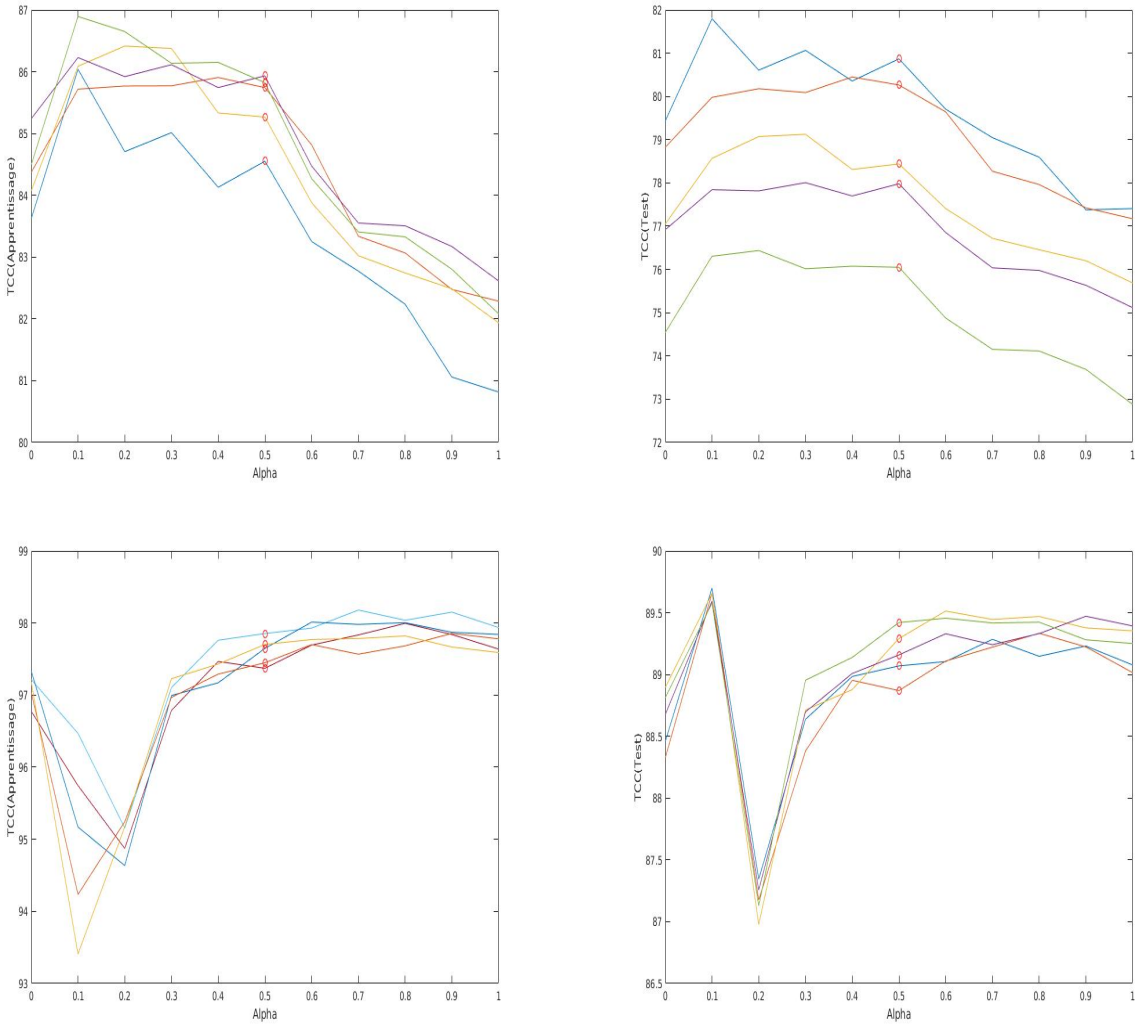


Figure 4.2: TCC de $S_{\text{Sylla \& Girard}}$ en fonction de α avec en bleu $\tau = 0.78$, en rouge $\tau = 0.74$, en jaune $\tau = 0.7$, en violet $\tau = 0.66$ et en vert $\tau = 0.63$. A gauche, nous avons les TCC sur l'ensemble d'apprentissage et à droite celui du test. En haut, on a les données d'autopsie verbale et en bas les données de chiffres manuscrits.

A l'opposé, SVM et Forêts Aléatoires donnent de meilleurs résultats que pgpDA sur les données de chiffres manuscrits. Ceci peut être dû au petit nombre de classes ($K = 10$ pour les données des chiffres manuscrits contrairement à $K = 18$ pour les données d'autopsie verbale). Le TCC associé à ppgDA est cependant satisfaisant, il est plus grand que 87% quelle que soit la valeur de α. Les TCC associés au paramètre optimal $\alpha^\star$ sélectionné par validation croisée est noté en gras dans le Tableau 4.5. Un taux $\tau = 63\%$ des individus de l'ensemble sur les données initial est pris pour l'échantillon apprentissage. Les TCC sont donnés pour des données autopsie verbale (en haut) et pour les données des chiffres manuscrits (en bas). Le noyau Sylla & Girard est appliqué dans pgpDA, SVM et k NN pour $\alpha \in \{0.1, 0.2, \ldots, 1\}$.

Les TCC sur les données d'autopsies verbales et sur les données des chiffres manuscrits obtenus avec les paramètres par défaut `nodesize = 1` et `ntree=500` sont mis en évidence (en gras), et rapportés dans le Tableau 4.6. Les TCC obtenus avec Forêts Aléatoires pour plusieurs valeurs de `nodesize` et `ntree` sur l'ensemble de données d'autopsies verbales sont reportés dans le Tableau 4.7. Les (TCC) obtenus avec Forêts Aléatoires pour plusieurs valeurs de `nodesize` et `ntree` sur l'ensemble des chiffres manuscrits sont aussi reportés dans le Tableau 4.7.

4.5.6 Performances de la méthode proposée

Nous utilisons les mesures de performances décrites dans le paragraphe 3.5.2.

Le Tableau 4.3 résume la performance de la méthode proposée pour chaque cause de décès. Cette performance est évaluée par validation croisée avec $\tau = 63\%$. Pour $\alpha = 0.3$, nous notons pour les maladies diabète, Néphropathie et Méningite une précision égale 1. Cela signifie que la méthode proposée peut diagnostiquer sans aucune erreur possible les personnes atteintes de ces maladies. Pour la maladie Méningite, il y a 100% de précision mais avec un rappel de 61%. Pour ce cas, la méthode proposée est très précise mais moyennement inefficace, car elle n'a su trouver que 61% des réponses possibles. Une méthode dont le rappel est fort, mais avec une faible précision fournira des diagnostics erronés, en plus de ceux qui sont pertinentes. Il est possible de citer le cas des maladies comme Abcès et Prostate avec une précision respectivement de 75% et 71% avec un rappel de 100%. Dans notre cas, nous avons de bons scores quelle que soit la maladie considérée avec des scores de plus de 50%.

En appliquant les mêmes groupements de classes que dans le paragraphe 3.3, nous avons

déterminé les performances de notre méthode sur ces groupements dans le Tableau 4.4. En regroupant les maladies, la méthode perd en précision, on ne note jamais une précision de 100%. Mais quel que soit le groupement proposé, la précision est de plus de 65% avec des rappels assez élevés de plus de 80%. Le modèle de noyau proposé permet de détecter avec une bonne précision les diagnostics sur les groupements proposés. Le compromis entre le rappel et la précision mesuré par le $F - score$ est aussi important quel que soit le groupement proposé.

En comparant les $F - score$ des Tableaux 3.3 et 4.3 ou des Tableaux 3.4 et 4.4 , on note que la méthode à noyau présente des résultats sensiblement meilleurs que ceux de la méthode basée sur un modèle multinomial.

4.6 Conclusion

Ce travail a été motivé par deux constations: tout d'abord, de nombreuses mesures de similarité binaires ont été utilisées dans divers domaines scientifiques. Deuxièmement, les modèles de mélange fournissent une réponse cohérente au problème de classification et à l'interprétation probabiliste sur des cas multi-classes. En se basant sur ces remarques, notre principale contribution est la proposition d'une nouvelle méthode de classification combinant à la fois les avantages des modèles de mélange et des mesures de similarité binaires. La méthode proposée présente de bonnes performances de classification sur des ensembles de données complexes (nombre élevé des prédicteurs et de classes).

Nous croyons que cette méthode peut se révéler utile sur des problèmes de classification binaires dans des domaines divers et variés. En tant que sous-produit de ce travail, de nouvelles mesures de similarité sont proposées afin d'unifier celles existantes dans la littérature. Ce travail pourrait être étendu à la classification des données quantitatives et binaires mixtes. Comme suggéré dans [31], pour faire face à ces données, on peut construire un noyau combiné en mélangeant un noyau basé sur une mesure de similarité (tel que proposé ici) sur des prédicteurs binaires et un noyau RBF pour les prédicteurs quantitatifs. Le noyau combiné pourrait être par exemple la somme pondérée ou le produit des deux noyaux, voir [98] pour plus de détails sur l'apprentissage par noyaux multiples.

Noyau	α	σ	`seuil` t	**TCC (Apprentissage)**	**TCC (Test)**	**équation**
Euclid		4	0.60	87.99	83.82	(16)
Pearson		10	0.95	87.72	83.25	(51)
Hellinger		6	0.60	87.68	83.21	(29,30)
Dice		2	0.60	87.32	83.00	(2,3,5)
3w-Jaccard		2	0.75	87.21	82.87	(4)
Ochia1		2	0.60	87.15	82.77	(33,38)
Gower & Legendre		4	0.80	86.61	82.64	(8,11)
Roger & Tanimoto		2	0.65	85.89	82.39	(9)
Sylla & Girard	0.1	1.9	0.90	85.81	81.50	
Sylla & Girard	0.3	2.2	0.85	85.47	81.46	
Sylla & Girard = RBF	0.5	1.4	0.80	85.05	81.35	(15,17,...,23)
Sylla & Girard	0.2	1.8	0.80	85.58	81.11	
Godman & Kruskal		4	0.95	84.28	80.77	(69)
Sylla & Girard	0.4	2.5	0.80	84.67	80.64	
Sokal & Sneath 5		4	0.95	84.72	80.49	(57)
Sylla & Girard	0.6	3.09	0.80	83.19	79.61	
Sylla & Girard	0.7	3.34	0.95	82.96	79.47	
Sokal & Sneath1		2	0.05	83.37	78.65	(6)

Noyau	α	σ	`seuil` t	**TCC (Apprentissage)**	**TCC (Test)**	**équation**
Hellinger		8	0.5	97.57	89.70	(29,30)
Euclid		8	0.5	97.54	89.67	(16)
Sylla & Girard	0.1	3.16	1	92.29	89.58	
Sylla & Girard	0.6	6.19	0.5	97.52	89.46	
Sylla & Girard	0.7	6.69	0.5	97.47	89.41	
Dice		2	0.5	97.40	89.41	(2,3,5)
Ochia1		2	0.5	97.36	89.38	(33,38)
Sylla & Girard = RBF	0.5	5.65	0.5	97.49	89.38	(15,17,. . . ,23)
Roger & Tanimoto		2	0.4	97.29	89.35	(9)
Sylla & Girard	0.8	8	0.5	97.38	89.31	
Sylla & Girard	1	8	8	92.29	89.26	(9)
3w-Jaccard		4	0.5	97.28	89.25	(4)
Sylla & Girard	0.9	7.15	0.5	97.27	89.23	
Jaccard		4	0.4	97.23	89.21	(1)
Gower & Legendre		10	0.8	97.36	89.14	(8,11)
Sylla & Girard	0.4	6.3	0.5	97.21	89.10	
Sylla & Girard	0.3	5.4	0.5	96.86	88.67	
Sylla & Girard	0.2	4.4	0.45	97.94	86.82	

Tableau 4.2: TCC pour les données des autopsies verbales (en haut) et les données des chiffres manuscrits (en bas)

Maladies	Précision	Rappel	F score
Cardiopathie	0.89	0.86	0.87
Néphropathie	1.00	0.77	0.87
hépatopathie	0.85	0.85	0.85
bpp	0.86	0.96	0.91
Prostate	0.71	1.00	0.83
Digestive	0.94	0.90	0.92
Fébrile	0.85	0.81	0.83
Maladie-sang	0.69	0.75	0.72
Nouveau-né	0.78	0.92	0.84
Parasitaire	0.50	1.00	0.66
Cause-inconnue	0.70	0.68	0.69
Épilepsie	0.72	0.80	0.76
Diabète	1.00	1.00	1.00
Méningite	1.00	0.61	0.76
Fièvre	0.50	0.75	0.60
Malnutrition	0.75	0.60	0.66
Tumeur	0.83	0.50	0.62
Abcès	0.75	1.00	0.85

Tableau 4.3: Les mesures de performance sur l'ensemble de données autopsie verbale pour chaque cause de décès pour $\alpha = 0.3$ (α optimal)

		Précision	Rappel	F score
4 $Groupes$	Groupe 1	0.89	0.93	0.91
	Groupe 2	0.88	0.91	0.90
	Groupe 3	0.95	0.86	0.91
	Groupe 4	0.82	0.8	0.81
5 $Groupes$	Groupe 1	0.90	0.92	0.91
	Groupe 2	0.75	0.89	0.81
	Groupe 3	0.92	0.91	0.91
	Groupe 4	0.95	0.87	0.91
	Groupe 5	0.82	0.80	0.81
6 $Groupes$	Groupe 1	0.91	0.92	0.91
	Groupe 2	0.75	0.89	0.81
	Groupe 3	0.92	0.91	0.91
	Groupe 4	0.95	0.87	0.91
	Groupe 5	0.85	0.92	0.88
	Groupe 6	0.66	0.60	0.63

Tableau 4.4: Performances des groupes de causes pour $\alpha = 0.3$

	pgpDA		**SVM**		**kNN**	
α	**TCC (Apprentissage)**	**TCC (Test)**	**TCC (Apprentissage)**	**TCC (Test)**	**TCC (Apprentissage)**	**TCC (Test)**
0.1	86.9	76.3	85.3	74.6	64.5	53.1
0.2	86.6	76.4	**79.9**	**70.8**	67.4	57.6
0.3	**86.1**	**76.0**	79.5	70.4	**68.3**	**59.5**
0.4	86.1	76.1	76.0	67.9	69.1	60.9
0.5	85.8	76.1	72.7	65.3	69.0	61.0
0.6	84.3	74.9	70.3	63.5	69.2	61.8
0.7	83.4	74.2	69.2	62.6	68.3	60.9
0.8	83.3	74.1	68.7	62.2	68.5	60.9
0.9	82.8	73.7	68.2	61.7	67.7	59.8
1	82.1	72.0	67.6	61.2	64.6	56.4
	pgpDA		**SVM**		**kNN**	
α	**TCC (Apprentissage)**	**TCC (Test)**	**TCC (Apprentissage)**	**TCC (Test)**	**TCC (Apprentissage)**	**TCC (Test)**
0.1	93.4	89.6	**100.0**	**93.1**	91.5	91.4
0.2	95.2	87.1	99.9	97.5	94.3	93.8
0.3	97.2	88.9	99.9	97.8	**95.5**	**94.3**
0.4	**97.4**	**89.1**	99.7	97.7	95.3	93.4
0.5	97.7	89.4	99.4	97.4	94.7	92.0
0.6	97.8	89.4	99.3	97.2	92.5	88.7
0.7	97.8	89.4	99.1	97.0	89.3	83.5
0.8	97.8	89.4	98.3	96.2	82.5	74.7
0.9	97.7	89.3	98.0	96.0	72.5	62.2
1	97.6	89.3	97.7	95.7	56.1	45.2

Tableau 4.5: TCC en fonction des méthodes et de α.

	Forêts Aléatoires	
	TCC (Apprentissage)	**TCC (Test)**
Autopsie Verbale	88.7	67.4
Chiffres manuscrits	100.0	94.0

Tableau 4.6: TCC obtenus avec les Forêts Aléatoires pour (`nodesize=1` et `ntree=500`).

	TCC (Apprentissage)									
`nodesize` `ntree`	1	2	3	4	5	6	7	8	9	10
250	88.8	87.9	86.9	85.5	83.9	82.7	81.4	79.9	78.1	76.6
500	**88.7**	88.0	86.9	85.6	84.4	82.9	81.6	79.9	78.3	76.7
750	88.8	88.2	87.0	85.7	84.4	82.9	81.5	79.6	78.1	76.6
1000	88.7	88.2	87.2	85.7	84.2	83.1	81.6	80.0	78.1	76.8
	TCC (Test)									
`nodesize` `ntree`	1	2	3	4	5	6	7	8	9	10
250	67.3	67.1	67.0	67.6	67.0	67.1	66.8	66.3	66.1	65.9
500	**67.4**	67.8	67.3	67.0	67.4	66.9	66.7	66.4	65.9	65.7
750	67.5	67.6	67.4	67.1	67.2	66.9	66.7	66.5	65.8	65.8
1000	67.9	67.7	67.3	67.3	67.2	67.0	66.7	66.6	66.2	65.5

Tableau 4.7: TCC des Forêts Aléatoires obtenus avec les données d'autopsies verbales en fonction de `nodesize` et `ntree`

	TCC (Apprentissage)									
`nodesize` `ntree`	1	2	3	4	5	6	7	8	9	10
250	100.0	100.0	100.0	99.9	99.9	99.8	99.6	99.4	99.2	98.9
500	**100.0**	100.0	100.0	100.0	99.9	99.8	99.7	99.5	99.2	99.0
750	100.0	100.0	100.0	100.0	99.9	99.8	99.7	99.5	99.3	99.0
1000	100.0	100.0	100.0	100.0	99.9	99.8	99.7	99.5	99.3	99.0
	TCC (Test)									
`nodesize` `ntree`	1	2	3	4	5	6	7	8	9	10
250	93.9	93.8	93.8	93.6	93.5	93.5	93.2	93.1	93.1	93.0
500	**94.0**	94.0	93.7	93.8	93.7	93.5	93.3	93.3	93.1	93.2
750	93.9	94.0	93.8	93.8	93.7	93.5	93.5	93.3	93.3	93.1
1000	94.0	94.0	93.9	93.8	93.7	93.6	93.4	93.4	93.3	93.2

Tableau 4.8: TCC des Forêts Aléatoires obtenus avec les données des chiffres manuscrits en fonction de `nodesize` et `ntree`

CHAPITRE 5

MODÈLE DE MÉLANGE DE NOYAUX HIÉRARCHIQUES POUR LA CLASSIFICATION DE PRÉDICTEURS BINAIRES.

Sommaire

5.1 Introduction

Les méthodes de diagnostics font souvent appel à des données structurées. Ces données présentent une structure hiérarchique au niveau des questions posées lors de l'entretien avec le médecin ou avec l'enquêteur en charge des autopsies verbales. L'aspect hiérarchique des questions posées lors de l'interrogatoire ou de l'entretien suppose une prise en compte de cet aspect lors de l'analyse des données médicales. Ainsi, il convient de choisir des mesures de similarité prenant en compte cet aspect, afin de mieux représenter la réalité.

Pour la prise en compte de cette structure, on utilise un noyau global sur des données de type hétérogènes (vidéo, son, texte,...) ou hiérarchisées (items, sous-items,...). Chaque ensemble de ces caractéristiques peut nécessiter un noyau différent. Ainsi, pour construire un noyau global,

il est possible de définir simplement un noyau pour chacune de ces caractéristiques et de les combiner linéairement ou multiplicativement.

La classification hiérarchique permet d'intégrer une structure existante a priori dans les données. Cet a priori est représenté par une structure sur les variables ou les caractéristiques du jeu de données.

Dans ce chapitre, nous proposons l'utilisation d'un noyau hiérarchique sur des données binaires. On y expose l'introduction d'un noyau prenant en compte la structure hiérarchique et les interactions entre les sous-items dans des méthodes de classification supervisée. Ce noyau permet d'intégrer une connaissance issue du domaine d'application. Cette connaissance est relative à la façon dont les caractéristiques du problème sont organisées. De manière générale, nous nous intéressons aux problèmes dont les caractéristiques peuvent être structurées de manière hiérarchique. Dans le cadre de ces travaux, ces hiérarchies sont représentées par des arborescences à deux niveaux.

Ce chapitre est organisé de la manière suivante. Le principe de construction et d'utilisation des noyaux multiples est expliqué dans le paragraphe 5.2. Dans le paragraphe 5.3, nous décrivons d'abord les différentes étapes de la démarche qui permet d'aboutir au choix de ce noyau. Nous montrons que la formulation obtenue est adaptée à des arborescences à deux niveaux et à l'interaction des sous-variables. Par le biais de ces deux formulations, nous étudions les caractéristiques du noyau défini, et caractérisons la relation qui lie les niveaux de cette arborescence. La performance de la nouvelle méthode de classification est illustrée sur les données d'autopsies verbales dans le paragraphe 5.4. Des observations finales sont fournies dans le paragraphe 5.5. Les preuves sont reportées en annexe.

5.2 Noyau multiple et hiérarchique

Les méthodes dites “Multiple Kernels Learning ”visent à construire un modèle de noyau, où le noyau résultant est une combinaison linéaire de noyaux.

L'approche multi-noyaux [99, 100] a été introduite pour généraliser l'approche mono-noyau. L'utilisation des noyaux multiples permet de générer des noyaux de bases pour chaque source de données.

Les développements récents des noyaux multiples ont prouvé que leur utilisation permet d'améliorer l'interprétabilité et la flexibilité des méthodes [101]. L'approche des noyaux multiples est de plus en plus utilisée dans des contextes différents et surtout en reconnaissance d'images [102, 103, 104].

L'apprentissage du noyau consiste à optimiser des coefficients de pondération pour chaque noyau de base, plutôt que d'optimiser les paramètres d'un seul noyau. Il permet d'atténuer l'influence du noyau et permet en même temps de tester plusieurs noyaux presque simultanément.

Définition 5.2.1 *Les noyaux multiples sont définis comme suit:*

$$\kappa(x, x') = \sum_{m=1}^{M} d_m \kappa_m(x, x')$$

avec $d_m \geq 0, \forall\, m \in \{1, 2, \ldots, M\}$, $\sum_{m=1}^{M} d_m = 1$ *où* κ_m *est un noyau de base.*
M *est le nombre de noyaux de base et* d_m *le poids du noyau* κ_m*.*

Chaque noyau peut être calculé sur différents sous-ensembles de la base d'apprentissage, par différentes méthodes de description des données ou selon différentes formulations.

Le but de l'approche multi-noyaux consiste à ne retenir parmi cet ensemble de noyaux que les noyaux les plus pertinents. L'idée revient à ne sélectionner et combiner que les noyaux les plus pertinents parmi cet ensemble grâce à la pondération des coefficients d_m. De ce fait si un noyau est pertinent son poids d_m associé aura une valeur grande et inversement un noyau peu pertinent aura un poids faible. Ce choix de noyau est souvent fait par le calcul de poids optimaux des différents noyaux [105, 106].

Certaines méthodes combinent à la fois les noyaux multiples et la structure hiérarchique des données. Cette structure se présente souvent comme un graphe orienté, qui induit une hiérarchie entre les nœuds et les feuilles. Ces méthodes plus souvent connues sous l'acronyme "Hierarchical Kernel Learning " fournissent un ensemble de noyaux sur des données hiérarchisées.

5.3 Construction de noyaux hiérarchiques associés à des observations binaires

5.3.1 Données et notations:

Dans un questionnaire, il existe souvent des questions dites principales. Pour chaque question principale, il existe des questions dites secondaires. Les questions secondaires ne sont posées que lorsque la réponse à la question principale est positive.

En formalisant ce concept, la variable X_j représente la réponse à la question principale j. Pour chaque X_j donné, on a q_j réponses aux questions secondaires notées par les sous- variables $Z_1^j, \ldots, Z_{q_j}^j$.

Ainsi, en se rapportant au cas des données d'autopsie verbale, on note:

- Les variables aléatoires $X = (X_j, j = 1, \ldots, p)$ définissent les réponses aux questions principales représentant les symptômes et les variables socio-démographiques.
- Les variables aléatoires $Z = (Z_\ell^j, \ell = 1, \ldots, q_j, j = 1, \ldots, p)$ définissent les réponses aux questions secondaires représentant les q_j sous-variables pour chaque variable X_j.
- Les variables aléatoires $Y = (Y_k, k = 1, \ldots, K)$ définissent les variables à expliquer représentant le groupe (diagnostics des médecins).

Ces hiérarchies sont représentées par un arborescence à deux niveaux, comme celle illustrée sur la Figure 5.1.

Le premier niveau représente les réponses aux questions dites principales. Le second niveau représente les sous-variables c'est-à-dire les réponses aux questions secondaires relatives à une question principale.

De plus, le lemme suivant énonce la relation entre les niveaux de l'arbre.

Lemme 5.3.1 *Soit Z_ℓ^j est une sous variable de X_j, alors*

- $\forall \ell \in \{1, \ldots, q_j\}, \mathbb{P}(Z_\ell^j = 1 | X_j = 0) = 0$,
- $\exists\, \ell \in \{1, \ldots, q_j\}$ *tel que* $\mathbb{P}(Z_\ell^j = 1 | X_j = 1) = 1$.

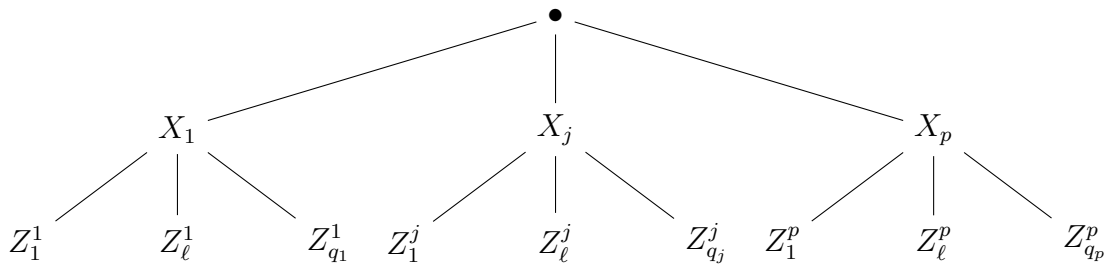


Figure 5.1: Exemple d'arborescence à deux niveaux associée aux variables explicatives

De plus

$$X_j = \max\{Z_1^j, \ldots, Z_\ell^j\} = 1 - \prod_{\ell=1}^{q_j}(1 - Z_\ell^j),$$

avec q_j le nombre de sous variables de X_j.

5.3.2 Noyaux hiérarchiques associés à des observations binaires

Le but de ce paragraphe est de construire des noyaux hiérarchiques adaptés aux données binaires à partir des mesures de similarité et de dissimilarité présentées dans le paragraphe 4.3. Les noyaux peuvent être utilisés dans les règles de classification décrites dans le paragraphe 2.3.1 pour construire de nouvelles méthodes de classification conçues pour les données binaires.

L'objectif est de construire un noyau associé à la structure hiérarchique des données et à l'interaction des sous- variables. Nous nous intéressons aux problèmes où les variables explicatives d'un jeu de données peuvent être structurées dans une arborescence. Dans cette structure, les caractéristiques dites sous-variables sont situées sur le niveau le plus bas de l'arbre. Le premier niveau permet d'identifier les variables dites principales auxquels appartiennent les sous-variables. Ces deux niveaux sont reliés par la relation décrite dans le Lemme 5.3.1.

Jusqu'à maintenant, nous avons considéré l'effet de chaque variable indépendant des autres variables. Ci-dessous nous proposons un nouveau noyau qui prend en compte les interactions entre variables. Une interaction doit avoir une traduction clinique significative, c'est-à-dire susceptible d'apporter des informations supplémentaires sur la méthode de diagnostic afin d'améliorer la précision des résultats.

Notre principe repose sur la transformation de la dissimilarité entre deux variables principales X_j et $X_{j'}$ en une combinaison de dissimilarités entre les variables principales X_j et $X_{j'}$ et leurs

sous-variables respectives $Z_\ell^j, \ell = 1, \ldots, q_j$ et $Z_{\ell'}^{j'}, \ell' = 1, \ldots, q_j'$.

D'après le Lemme 5.3.1, pour chaque variable X_j, on a q_j sous-variables $Z_1^j, \ldots, Z_{q_j}^j$. De plus on a:

$$X_j = \max\{Z_1^j, \ldots, Z_\ell^j\} \qquad = 1 - \prod_{\ell=1}^{q_j}(1 - Z_\ell^j) = \sum_{\ell=1}^{q_j}(-1)^{\ell-1} \sum_{k=1}^{\ell} \sum_{|i|=k} Z_{i_1} \ldots Z_{i_k}$$

où $|i| = k$ désigne la taille du multi-indice $i = (i_1, \ldots, i_k)$.

En calculant $\|x - x'\|^2$ on a :

$$\begin{aligned}
\|x - x'\|^2 &= \sum_{j=1}^{p} \left[\prod_{\ell=1}^{q_j}(1 - z_\ell^j) - \prod_{\ell=1}^{q_j}(1 - z_\ell^{'j}) \right]^2 \\
&= \sum_{j=1}^{p} \left[\sum_{\ell=1}^{q_j}(-1)^{\ell-1} \sum_{k=1}^{\ell} \left(\sum_{|i|=k} z_{i_1}^j \ldots z_{i_k}^j - \sum_{|i|=k} z_{i_1}^{'j} \ldots z_{i_k}^{'j} \right) \right]^2 \\
&= \sum_{j=1}^{p} \left[\sum_{\ell=1}^{q_j}(-1)^{\ell-1} \sum_{k=1}^{\ell} \sum_{|i|=k} s_{kji} \right]^2 \\
&= \sum_{j=1}^{p} \left[\sum_{\ell=1}^{q_j} \sum_{k=1}^{\ell} \sum_{|i|=k} (-1)^{\ell-1} s_{kji} \right]^2 \\
&= \sum_{j=1}^{p} \sum_{\ell=1}^{q_j} \sum_{k=1}^{\ell} \sum_{|i|=k} s_{kji}^2 + R
\end{aligned}$$

avec $s_{kji} = \left(z_{i_1}^j \ldots z_{i_k}^j - z_{i_1}^{'j} \ldots z_{i_k}^{'j} \right)$ et R la somme des doubles-prduits.

En posant

$$SC(z, z') = \sum_{j=1}^{p} \sum_{\ell=1}^{q_j} \sum_{k=1}^{\ell} \sum_{|i|=k} s_{kji}^2$$

on a donc la décomposition

$$\|x - x'\|^2 = SC(z, z') + R.$$

On définit une mesure de dissimilarité pour tout $\gamma \in [0,1]$ par:

$$D((x,z),(x',z')) = \gamma SC(z,z') + (1-\gamma)R = (1-\gamma)\|x-x'\|^2 + (2\gamma-1)SC(z,z').$$

En posant:

- $D_x(x,x') = \|x-x'\|^2$,
- $D_z(z,z') = SC(z,z')$,

la mesure de dissimilarité précédente se réécrit :

$$D((x,z),(x',z')) = (1-\gamma)D_x(x,x') + (2\gamma-1)D_z(z,z').$$

En utilisant la méthode de construction de noyau proposée dans l'équation (4.5), on introduit le noyau :

$$\kappa_{\text{SGH}}((x,z),(x',z')) = \kappa_x(x,x')^{1-\gamma}\kappa_z(z,z')^{2\gamma-1} \tag{5.1}$$

où,

- $\kappa_x(x,x') = \exp(-\|x-x'\|^2/2\sigma_x^2)$ est le noyau RBF,
- $\kappa_z(z,z') = \exp(-SC(z,z')/2\sigma_r^2)$.

Plus généralement dans le noyau (5.1):

1) Pour les variables principales X, on peut choisir un noyau de type:

$$\kappa_x(x,x') = \exp\left(\frac{S(x,x')}{2\sigma_x^2}\right). \tag{5.2}$$

avec S une mesure de similarité.

Cette mesure de similarité S peut être choisie par les mesures définies dans notre formalisme (4.3):

$$S(x,x') = \frac{\alpha a - \theta(b+c) + \beta d}{\alpha' a + \theta'(b+c) + \beta' d}$$

où $\alpha \geq 0$, $\beta \geq 0$, $\theta \geq 0$, $(\alpha',\beta') \in \mathbb{R}^2$ et $\theta' \neq 0$.

2) Les interactions des sous-variables Z sont prises en compte jusqu'à l'ordre r grâce au noyau suivant:

$$\kappa_z(z, z^{'}) = \exp\left(\frac{SC_{(r)}(z, z')}{2\sigma_r^2}\right) \tag{5.3}$$

avec

(a) r le nombre d'interactions,

(b) $SC_{(r)}$ la version tronquée au niveau r de SC:

$$\begin{aligned} SC_{(r)}(z, z') &= \sum_{j=1}^{p}\sum_{\ell=1}^{q_j}\sum_{k=1}^{r}\sum_{|i|=k} s_{kji}^2 \\ &= \sum_{j=1}^{p}\sum_{k=1}^{r}\sum_{\ell=1}^{q_j}\sum_{|i|=k} s_{kji}^2 \\ &= \sum_{j=1}^{p} q_j \sum_{k=1}^{r}\sum_{|i|=k} s_{kji}^2 \\ &= \sum_{j=1}^{p} q_j sc_{(r,j)} \end{aligned}$$

(c) $sc_{(r,j)}$ les interactions d'ordre r entre les sous-variables de la variable j définies par

$$\begin{aligned} sc_{(r,j)} &= \sum_{k=1}^{r}\sum_{|i|=k} s_{kji}^2 \\ &= \sum_{k=1}^{r}\sum_{|i|=k}\left(z_{i_1}^j \dots z_{i_k}^j - z_{i_1}^{'j} \dots z_{i_k}^{'j}\right)^2 \\ &= \sum_{i=1}^{p}(z_i^j - z_i^{'j})^2 + 2!\sum_{i<k}(z_i^j z_k^j - z_i^{'j} z_k^{'j})^2 \\ &+ 3!\sum_{i<k<t}(z_i^j z_k^j z_t^j - z_i^{'j} z_k^{'j} z_t^{'j})^2 + \dots + r!\sum_{i_1<i_2<\dots<i_r}(z_{i_1}^j \dots z_{i_r}^j - z_{i_1}^{'j} \dots z_{i_r}^{'j})^2. \end{aligned}$$

En combinant 1) et 2), on définit le noyau hiérarchique avec effet d'interactions d'ordre r suivant :

$$\kappa_{\text{SGH}}((x, z), (x', z')) = \kappa_x(x, x^{'})^{(1-\gamma)}\kappa_z(z, z^{'})^{(2\gamma-1)}$$

où κ_x est donné par (5.2) et κ_z par (5.3).

Pour certaines valeurs de γ, il apparaît ainsi que le noyau RBF peut être retrouvé pour des données binaires dans certains cas.
Si $\kappa_x = \kappa_{\text{RBF}}$ alors

- $\gamma = \frac{1}{2} \Rightarrow \kappa_{\text{SGH}}\left((x,z),(x',z')\right) = \kappa_{\text{RBF}}(x,x')$,
- $\gamma = 1$ et $r = 1 \Rightarrow \kappa_{\text{SGH}}\left((x,z),(x',z')\right) = \kappa_{\text{RBF}}(z,z^{'})$,
- $\gamma = \frac{2}{3}$ et $r = 1$
  $\Rightarrow \kappa_{\text{SGH}}\left((x,z),(x',z')\right) = \kappa_{\text{RBF}}((x \cup z),(x^{'} \cup z^{'}))$.

5.4 Application à des données d'autopsie verbale

5.4.1 Méthodologie

Les modèles que nous avons présentés dépendent du choix des hyperparamètres. La mise en œuvre de la méthode de classification nécessite la sélection des hyper paramètres $\omega = (\gamma, \alpha, \sigma_x, \sigma_r)$:

- γ le paramètre de pondération entre les variables principales et les variables secondaires désigné ici comme le paramètre hiérarchique,
- α le paramètre de pondération des présences et absences des variables principales du noyau $\kappa_x(x,x')$ comme décrit dans 4.6,
- σ_x le paramètre de lissage sur le noyau des variables principales,
- σ_r le paramètre de lissage du noyau d'interaction sur les variables secondaires.

A cette fin, une technique de double validation croisée est utilisée. L'échantillon total de taille n est subdivisé aléatoirement en $M = 50$ fois en échantillon apprentissage $\mathcal{L}_m$ de taille τn et en échantillon test $\mathcal{T}_m$ de taille $(1-\tau)n$ avec $\tau \in (0,1)$ représentant une proportion et $m = 1, \ldots, M$.

Le paramètre α définit par l'équation 4.6 est fixé à 0.5 pour retrouver le noyau RBF sur les variables principales et σ_x est fixé à 1.5 et le seuil à 0.8 comme valeur optimale calculée dans [72].

Le paramètre σ_r est choisi par validation croisée sur le jeu d'apprentissage : 5 fois consécutivement, 100 individus sont retirés aléatoirement de l'échantillon d'apprentissage, et est estimé par maximisation du taux de bien classés sur les 100 individus retirés. Le taux de bien classés global est estimé sur l'échantillon test en répétant l'ensemble du procédé 50 fois. De ce fait, sur chaque échantillon d'apprentissage $\mathcal{L}_m$, l'hyperparamètre optimal $\hat{\sigma}_{r_m}$ est estimé en 5 fois par validation croisée pour $m = 1, \ldots, M$.

De plus, l'hyperparamètre optimal $\hat{\sigma}_r$ est calculé comme le mode empirique de l'ensemble $\{\hat{\sigma}_{r_1}, \ldots, \hat{\sigma}_{r_M}\}$. Le taux moyen des biens classés est calculé sur l'échantillon l'apprentissage $\mathcal{L}_m, m = 1, \ldots, M$ et sur l'échantillon test $\mathcal{T}_m, m = 1, \ldots, M$.
En outre, pour un niveau d'interactions on choisit σ_r par validation croisée et on le maintient pour le niveau d'interactions suivant. Pour des problèmes de temps de calcul, nous fixons le nombre d'interactions maximum à 3.

5.4.2 Résultats obtenus avec un niveau d'interaction d'ordre 1

Au Tableau 5.1, nous présentons les résultats obtenus en nous limitant aux interactions d'ordre 1 ($r = 1$).

Il est à noter que respectivement pour $\gamma = 0.5$ et $\gamma = 1$, les taux de classification trouvés sont ceux associés respectivement aux variables principales et aux variables secondaires prises séparément.

Après évaluation des résultats, nous constatons que pour un $\gamma = 0.67$, nous retrouvons le même taux de classification qu'un noyau RBF, voir Tableau 4.2. Il est intéressant de constater qu'en considérant une structure sur les données, on améliore légèrement les taux de classification pour $\gamma = 0.7$ et 0.8.

γ	α	σ_x	σ_1	**TCC (Apprentissage)**	**TCC (Test)**
0.5	0.5	1.5	4.25	76.21	67.44
0.6	0.5	1.5	1.75	83.50	74.33
0.67	0.5	1.5	1.75	84.20	74.92
0.7	0.5	1.5	1.75	84.53	75.25
0.8	0.5	1.5	3.75	84.32	74.95
0.9	0.5	1.5	3.25	83.15	73.72
1	0.5	1.5	1.5	71.36	61.54

Tableau 5.1: Valeurs des paramètres ω et des taux de classification corrects pour $\gamma \in [0.5, 1]$ et $r = 1$

5.4.3 Résultats obtenus avec un niveau d'interaction d'ordre 2

Dans le Tableau 5.2, on résume les valeurs des paramètres pour un nombre d'interactions $r = 2$. Les valeurs optimales de σ_2 sont choisies par validation croisée. Les valeurs de σ_1 sont celles calculées dans le Tableau 5.1.

En considérant un niveau d'interactions d'ordre 2 ($r = 2$), le noyau ainsi défini présente une amélioration des résultats. A l'ordre d'interactions $r = 2$, on note que pour certaines valeurs de γ, les taux de classification obtenus dépassent les 75%. Pour $\gamma = 0.7$, le taux de classification obtenu est de 77%, ainsi les résultats sont sensiblement améliorés.

γ	α	σ_x	σ_1	σ_2	**TCC (Apprentissage)**	**TCC (Test)**
0.5	0.5	1.5	4.25	19	76.21	67.44
0.6	0.5	1.5	1.75	13	85.77	76.48
0.67	0.5	1.5	1.75	19	86.50	76.95
0.7	0.5	1.5	1.75	19	86.63	77.00
0.8	0.5	1.5	3.75	19	84.94	75.76
0.9	0.5	1.5	3.25	18	83.97	74.50
1	0.5	1.5	1.5	18	75.09	64.91

Tableau 5.2: Valeurs des paramètres ω et des taux de classification corrects pour $\gamma \in [0.5, 1]$ et $r = 2$

5.4.4 Résultats obtenus avec un niveau d'interaction d'ordre 3

Dans le Tableau 5.3, on résume les valeurs des paramètres pour un nombre d'interactions $r = 3$. Les valeurs optimales de σ_3 sont choisies par validation croisée. Les valeurs de σ_1 et de σ_2 sont celles calculées dans les Tableau 5.1 et 5.2. Pour $r = 3$ et $\gamma \in \{0.6, 0.67, 0.7\}$, on note des taux de classification légèrement supérieurs à 77.00%.

γ	α	σ_x	σ_1	σ_2	σ_3	**TCC (Apprentissage)**	**TCC (Test)**
0.5	0.5	1.5	4.25	19	27	76.21	67.44
0.6	0.5	1.5	1.75	13	22	86.59	77.07
0.67	0.5	1.5	1.75	19	31	86.93	77.19
0.7	0.5	1.5	1.75	19	36	86.93	77.14
0.8	0.5	1.5	3.75	19	44	85.10	75.57
0.9	0.5	1.5	3.25	18	44	83.01	73.21
1	0.5	1.5	1.5	18	29	74.72	64.59

Tableau 5.3: Valeurs des paramètres ω et des taux de classification corrects pour $\gamma \in [0.5, 1]$ et $r = 3$

5.4.5 Comparaison des niveaux d'interactions

En terme de comparaison, on note des taux de classification associés aux ordres d'interaction $r = 3$ supérieurs à ceux des ordres interaction $r = 1$ et $r = 2$. Pour $\gamma = 0.5$, le taux de classification est invariant vis-à-vis de l'ordre d'interaction. Ceci s'explique par le fait que pour $\gamma = 0.5$, le noyau proposé ne prend pas en compte les interactions et n'est calculé que sur les variables principales. Le meilleur taux de classification est obtenu pour $\gamma = 0.67$ avec un nombre d'interactions égal à 3. Le tableau 5.4 résume les taux de classification suivant le niveau d'interactions et la valeur de γ.

5.4.6 Performances de la méthode proposée

Dans ce paragraphe, nous allons appliquer les mêmes groupements de classes que dans le paragraphe 3.3. Nous utilisons les mesures de performances décrites dans le paragraphe 3.5.2 pour calculer les performances selon les causes de décès. La performance sera calculée pour

interactions	**r= 1**		**r = 2**		**r = 3**	
γ	**TCC (Apprentissage)**	**TCC (Test)**	**TCC (Apprentissage)**	**TCC (Test)**	**TCC (Apprentissage)**	**TCC (Test)**
0.5	76.21	67.44	76.21	67.44	76.21	67.44
0.6	83.50	74.33	85.77	76.48	86.59	77.07
0.67	84.20	74.92	86.50	76.95	86.93	**77.19**
0.7	84.53	**75.25**	86.63	**77.00**	86.93	77.14
0.8	84.32	74.95	84.94	75.76	85.10	75.57
0.9	83.15	73.72	83.97	74.50	83.01	73.21
1	71.36	61.52	75.09	64.91	74.72	64.59

Tableau 5.4: Résumé des taux de classification corrects pour $\gamma \in [0.5, 1]$

chaque niveau d'interactions et suivant la valeur de γ ayant réalisé le maximum des taux de classification dans l'échantillon test pour chaque niveau.

En examinant le Tableau 5.5, on note que les meilleurs taux de classification sont obtenus pour $\gamma \in \{0.67, 0.7\}$ quel que soit le regroupement considéré. En réduisant le nombre de classes, le taux de classification s'améliore de manière considérable. Cette amélioration se note aussi suivant les niveaux d'interactions. Dans l'échantillon test, on obtient des taux de classification de l'ordre de 83% en réduisant le nombre de classes.

Le Tableau 5.6 résume la performance du noyau hiérarchique proposé pour chaque cause de décès. Cette performance est évaluée par validation croisée. Pour les maladies Néphropathie, Prostate, Parasitaire, d'Épilepsie, d'abcès et celles des nouveaux nés, les performances restent invariantes quel que soit le niveau d'interactions. Ainsi, on peux affirmer que pour ces types de maladies, les interactions de variables n'apportent pas plus de précision lors du diagnostic. Pour les autres maladies, nous notons une légère amélioration des précisions suivant le niveau d'interactions des variables considérées. La précision la plus faible est de l'ordre de 50% et la plus forte de 93%. Ainsi, les causes non pertinentes sont détectées par le modèle. Pour ce qui est du rappel, il est compris entre 45% et 88%. Le modèle proposé a un fort taux de concordance avec les diagnostics établis par les médecins en charge du diagnostic. Le noyau hiérarchique est quasi optimal avec des taux de F-score de plus de 58%. Ce noyau permet une amélioration de la précision de certaines causes en considérant les interactions entre variables.

Interactions	r = 1		r = 2		r = 3	
Classes	**TCC (Apprentissage)**	**TCC (Test)**	**TCC (Apprentissage)**	**TCC (Test)**	**TCC (Apprentissage)**	**TCC (Test)**
18 classes	84.53 ($\gamma = 0.7$)	75.25	86.63($\gamma = 0.7$)	77.00	86.59($\gamma = 0.67$)	77.07
6 classes	87.48 ($\gamma = 0.7$)	82.37	89.36($\gamma = 0.7$)	84.14	89.99($\gamma = 0.67$)	84.42
5 classes	87.39 ($\gamma = 0.67$)	82.11	89.22($\gamma = 0.7$)	83.82	89.89($\gamma = 0.67$)	84.20
4 classes	88.05 ($\gamma = 0.67$)	84.31	89.87($\gamma = 0.7$)	86.07	90.64($\gamma = 0.67$)	86.37

Tableau 5.5: Taux de classification corrects en fonction du nombre de classes et du degré d'interaction.

Interactions	r = 1			r = 2			r = 3		
	Précision	Rappel	F-score	Précision	Rappel	F-score	Précision	Rappel	F-score
Cardiopathie	0.79	0.81	0.80	0.78	0.81	0.80	0.76	0.83	0.79
Néphropathie	0.80	0.54	0.65	0.80	0.54	0.65	0.80	0.54	0.65
hépatopathie	0.93	0.63	0.75	0.93	0.68	0.78	0.93	0.63	0.75
bpp	0.74	0.85	0.79	0.76	0.86	0.81	0.81	0.83	0.82
Prostate	0.83	0.45	0.58	0.83	0.45	0.58	0.83	0.45	0.58
Digestive	0.88	0.85	0.87	0.88	0.86	0.87	0.86	0.86	0.86
Fébrile	0.76	0.75	0.75	0.77	0.78	0.78	0.78	0.79	0.78
Maladie-sang	0.80	0.48	0.60	0.76	0.52	0.61	0.75	0.48	0.58
Nouveau-né	0.73	0.86	0.79	0.73	0.86	0.79	0.73	0.86	0.79
Parasitaire	0.66	0.88	0.76	0.66	0.88	0.76	0.66	0.88	0.76
Cause-inconnue	0.51	0.78	0.62	0.53	0.80	0.64	0.53	0.78	0.63
Épilepsie	0.81	0.76	0.78	0.81	0.76	0.78	0.81	0.76	0.78
Diabète	0.87	0.87	0.87	0.85	0.75	0.80	0.84	0.68	0.75
Méningite	0.77	0.43	0.56	0.77	0.43	0.56	0.80	0.50	0.61
Fièvre	0.55	0.76	0.64	0.55	0.76	0.64	0.52	0.76	0.62
Malnutrition	0.71	0.66	0.68	0.75	0.60	0.66	0.75	0.6	0.66
Tumeur	0.77	0.48	0.59	0.77	0.48	0.59	0.73	0.48	0.58
Abcès	0.87	0.58	0.70	0.87	0.58	0.70	0.87	0.58	0.70

Tableau 5.6: Mesures de performance sur l'ensemble des données autopsie verbale pour chaque cause de décès.

5.5 Conclusion

Ce travail a été motivé par la prise en compte de l'aspect hiérarchique des questions lors de l'entretien avec le médecin. Nous avons proposé un noyau prenant en compte une structure d'arborescence des niveaux de réponses aux questions lors de l'entretien. Ce noyau implémenté dans la méthode ppgda présente des performances de classification cohérentes. Un diagnostic se précise souvent par la présence ou l'absence de certains symptômes mais surtout de leur interaction.

Notre principale contribution est la proposition d'un noyau prenant en compte simultanément l'aspect hiérarchique et l'interaction des variables. Le noyau proposé présente de bonnes performances de classification sur un ensemble de données de diagnostics complexes (nombre

élevé des prédicteurs et de classes).

Une adaptation de ce noyau à des données à structure d'arbre et de graphe pourrait se révéler utile sur de nombreux problèmes.

Ce travail pourrait être étendu à la classification des données quantitatives et binaires mixtes en spécifiant l'aspect interaction des variables.

CHAPITRE 6

CONCLUSION

Dans ce dernier chapitre, nous effectuerons une synthèse des travaux présentés dans ce mémoire et les perspectives de recherches envisagées.

L'objectif de ce travail de thèse est de contribuer à l'amélioration des méthodes de diagnostics des autopsies verbales par l'analyse de données, en particulier celles dédiées à la classification supervisée, pour une meilleure prise en compte des causes établis par les médecins.

Pour l'essentiel, les approches développées dans cette thèse ont été motivées par le souci d'apporter des réponses à des problématiques de diagnostic en grande dimension sur des données binaires.

Pour tenir compte de la grande dimension des données ($n = 2500$ individus, $p = 100$ variables et $K = 18$ classes) , les méthodes proposées s'appuient le plus souvent sur l'hypothèse d'indépendance conditionnelle des variables et sur des méthodes de réduction de la dimensionnalité des problèmes.

6.1 Synthèse des travaux

Le thème de ce mémoire est la modélisation et la classification des données binaires en grande dimension. Nous ferons la synthèse des contributions apportées dans ce mémoire.

Autopsie verbale: Concept et Utilisation

Dans le chapitre 1, nous avons élaboré un revue de la littérature sur les autopsies verbales. Nous avons explicité le concept de l'autopsie verbale, les caractéristiques et les techniques d'enquêtes associées. En dernier lieu, nous avons exposé la méthodologie des études par autopsie verbale et une synthèse des études statistiques menées dans le cadre des méthodes de diagnostics par autopsie verbales.

Panorama des méthodes statistiques

Au niveau du chapitre 2 nous avons énoncer les principaux concepts mathématiques utilisés dans ce mémoire. Nous avons focalisé notre étude sur les méthodes de classification particulièrement en classification supervisée. Une étude rétrospective des méthodes paramétriques avec l'introduction des modèles de mélanges sur des données quantitatives et qualitatives est développée.

En plus des méthodes paramétriques, nous avons passé en revue les méthodes non paramétriques. De ce fait, on a abordé les noyaux et les stratégies de construction des fonctions noyaux sur des données structurées et non structurées. Notre approche s'est focalisé sur les trois méthodes de classification (pgpDA, SVM et kNN) utilisant un noyau. Les mesures de similarités jouent un rôle important pour la construction des noyaux. Nous avons abordé dans ce mémoire une étude élargie des mesures de similarités .

Classification supervisée par modèle de mélange multinomial pour les autopsies verbales:

Au chapitre 3, nous avons exposé un modèle de mélange multinomial. Pour réduire la complexité du problèmes, nous y avons exposé les motivations de l'hypothèse d'indépendance conditionnelle et une méthode de réduction du nombre de classes par une méthode de $k - medoids$ sur la matrice des probabilités a posteriori des classes. Une méthode séquentielle de sélection de variables est aussi présentée.

La première contribution développée dans ce chapitre est la réduction du nombre de classes en des groupes de diagnostics plus homogènes. Ce regroupement a permis améliorer les résultats

du modèle multinomial. Une sélection des symptômes et variables socio démographiques est établie pour augmenter la pertinence des diagnostics et diminuer les bruits dûs au protocole d'enquête.

Une méthode de classification combinant mesures de similarité et modèles de mélanges

En établissant des liens entre des mesures de similarité ou de dissimilarité, nous avons introduit dans ce chapitre une nouvelle famille de mesures unifiant plusieurs mesures existant dans la littérature. En utilisant cette généralisation de certaines mesures de similarité ou de dissimilarité, nous avons présenté une famille de noyaux exponentiels en fonction de la mesure de similarité proposée. Cette fonction noyau généralisée sur des données binaires est introduit des méthodes de classification supervisée.

Ainsi dans le chapitre 4, nous avons présenté une nouvelle méthode de classification supervisée dédiée à des prédicteurs binaires combinant les mesures de similarité et les modèles de mélange grâce à une nouvelle famille de noyaux exponentiels.

La performance de la nouvelle méthode de classification est illustrée sur deux ensembles de données réelles (données d'autopsie verbale et les données de chiffres manuscrits) en utilisant 76 mesures de similarité différentes.

Modèle de mélange de noyaux hiérarchiques pour la classification de prédicteurs binaires.

Dans le chapitre 5, nous avons explicité le principe de construction et d'utilisation des noyaux multiples. Dans ce chapitre, deux aspects sur la méthodologie des enquêtes et de la présentation des données associées ont motivé notre travail. L'un des aspects est la hiérarchie des questions posées lors de l'entretien avec le médecin. L'autre aspect repose sur une amélioration de la performance des diagnostics par l'interaction des symptômes car un diagnostic se précise souvent par la présence ou l'absence de certains symptômes mais surtout de leur interaction.

Nous avons ainsi proposé un nouveau noyau prenant en compte simultanément l'aspect hiérarchique des questions posées et l'interaction des symptômes. Nous avons montré que

la formulation obtenue est adaptée à des arborescences à deux niveaux et à l'interaction des sous-variables. Par le biais de ces deux formulations, nous avons étudié les caractéristiques du noyau défini.

Ce noyau est implémenté dans la méthode pgpDA et présente de bonnes performances de classification suivant l'ordre d'interactions des sous-variables sur un ensemble de données de diagnostics complexes (nombre élevé des prédicteurs et de classes).

6.2 Perspectives

Ce travail pourrait être étendu à la classification des données quantitatives, binaires et mixtes. Il serait intéressant de construire un noyau combiné en mélangeant un noyau basé sur une mesure de similarité (tel que proposé ici) sur des prédicteurs binaires et un noyau pour les prédicteurs quantitatifs. Une adaptation de ce noyau à des données à structure d'arbre ou de graphe pourrait se révéler utile sur de nombreux problèmes. Par ailleurs, la question de l'introduction d'un noyau hiérarchique sur des arbres avec plus de niveaux se serait une piste importante de recherche. La prise en compte simultanément de données binaires et non binaires dans la construction du noyau hiérarchique semble pertinente pour des problèmes en reconnaissance image, ... Il serait aussi intéressant de comparer le noyau proposé avec d'autre noyaux de type hiérarchiques.

CHAPITRE 7

LISTE DES TRAVAUX

7.1 Publications, Conférences, Séminaires et Workshops

Publications

1. S.N Sylla, S. Girard, A.K Diongue, A. Diallo and C. Sokhna,
   A classification method for binary predictors combining similarity measures and mixture models, Dependence Modeling, 3, 1090–1096, 2015,

2. B.Senghor,O. Diaw,S. Doucoure, S.N. Sylla, M. Seye, I. Talla, C. Bâ, A. Diallo and C. Sokhna,
   Efficacy of praziquantel against urinary schistosomiasis and reinfection in Senegalese school children where there is a single well-defined transmission period , Parasites & Vectors, vol. 8 (5), 2015,

3. B. Senghor, A. Diallo ,S.N. Sylla, S. Doucouré, M.O. Ndiath, L. Gaayeb, F. Djuikwo-Tendeng, C.T. Ba and C. Sokhna,
   Prevalence and intensity of urinary schistosomiasis among school children in the district of Niakhar, region of Fatick, Senegal, Parasites & Vectors, vol. 7 (5), 2014.

En préparation

4. S.N Sylla, S. Girard, A.K Diongue, A. Diallo and C. Sokhna,
   Hierarchical kernel mixture model for the classification of binary predictors.

5. S.N Sylla, S. Girard, A.K Diongue, A. Diallo and C. Sokhna,
   Supervised classification by multinomial mixture model for verbal autopsies

6. S.N Sylla, S. Girard, A.K Diongue, A. Diallo and C. Sokhna,
   Statistical Approach for the verbal autopsy diagnosis in rural areas in Senegal.

Communications orales

Conférences Internationales

1. S.N. Sylla, S. Girard, A.K. Diongue, A.Diallo and C. Sokhna.
   Hierarchical kernel mixture model for the classification of binary predictors..
   SADA'16, Cotonou, Benin (28 Novembre-3 Decembre 2016), à venir.

2. S.N. Sylla, S. Girard, A.K. Diongue, A.Diallo and C. Sokhna,
   Classification de données binaires via l'introduction de mesures de similarités dans les modèles de mélange.
   47èmes Journées de Statistique, Lille France, Juin 2015.

3. S.N. Sylla, S. Girard, A.K. Diongue, A.Diallo and C. Sokhna.
   Classification supervisée par modéle de mélange:Application aux diagnostics par autopsie verbale.
   46èmes Journées de Statistique,Rennes France, Juin 2014.

Conférences nationales

1. S.N. Sylla, S. Girard, A.K. Diongue, A.Diallo and C. Sokhna.
   Modèle de mélange de noyaux hiérarchiques pour la classification de prédicteurs binaires.

Application au diagnostic par autopsie verbale
cimpa,Université Gaston Berger ,Saint-louis, Avril 2016.

2. S.N. Sylla, S. Girard, A.K. Diongue, A.Diallo and C. Sokhna.
Classification supervisée par modéle de mélange:Application aux diagnostics par autopsie verbale
Dicdacs,Université Cheikh Anta Diop ,Dakar, Mars 2014.

Invitation à des séminaires

1. S.N. Sylla. *Modéle de mélange hiérarchique:Application aux diagnostics par autopsie verbale* Séminaire de Statistique de LERSTAD, Université Gaston Berger, Mars 2014.

2. S.N. Sylla. *Classification supervisée par modéle de mélange:Application aux diagnostics par autopsie verbale* Séminaire de Statistique de LERSTAD, Université Gaston Berger, Janvier 2014.

3. S.N. Sylla. *Classification supervisée par modéle de mélange:Application aux diagnostics par autopsie verbale*
journée de rentrée des doctorants en Statistique du Laboratoire Jean Kuntzmann, Grenoble Rhône-Alpes, France ,le 20 Novembre 2012.
`http://mistis.inrialpes.fr/doctorants-probastat2012`

BIBLIOGRAPHIE

[1] C. Seung-Seok, C. Sung-Hyuk, and C. Tappert. A survey of binary similarity and distance measures. *Systemics, Cybernetics and Informatics*, 8:43–48, 2010. (pages 11, 14, 74, 91, 95, 96, 97, 98, 99, 100 et 103)

[2] Organisation Mondiale pour la Santé. Normes d'autopsies verbales: Etablissements de la cause de décès. *OMS*, 2009. (pages 23, 26 et 27)

[3] M. Garenne and O. Fontaine. Enquêtes sur les causes probables de décès en milieu rural sénégalais. *ORSTOM. Dakar Sénégal*, 1988. (pages 24, 25 et 30)

[4] Y. Biraud. Méthode pour l'enregistrement par des non médecins des causes élémentaires de décès dans des zones sous-développées. *Genève,OMS*, 1956. (page 24)

[5] OMS. Notification d'informations sanitaires par un personnel non médical. *Genève*, 1978. (page 25)

[6] S. Zimicki. L'enregistrement des causes de décès par des non médecins: deux expériences au bangladesh. *INED*, 119:101–122, 1988. (page 25)

[7] J.P. Chippaux. Conception, utilisation et exploitation des autopsies verbales. *Médecine Tropicale*, 69(2):143–150, 2009. (page 26)

[8] D. Chandramohan, P. Setel, and M. Quigley. Effect of misclassification of causes of death in verbal autopsy: can it be adjusted? *Int J Epidemiol.*, 30(3):509–514, 2001. (pages 28 et 31)

[9] G. Duthé, S.H.D. Faye, E. Guyavarch, P. Arduin, M. Kante, A. Diallo, R. Laurent, A. Marra, and G. Pison. Changement de protocole dans la méthode d'autopsie verbale et mesure de la mortalité palustre en milieu rural sénégalais. *Bulletin de la Société de Pathologie Exotique.*, 103(5):317–332, 2010. (page 35)

[10] D. Chandramohan, P. Setel, and M. Quigley. Effect of misclassification of causes of death in verbal autopsy: can it be adjusted? *International Journal of Epidemiology*, 30(3):509–514, 2001. (pages 38 et 40)

[11] P. Byass, D.L Huong, and H. Van Minh. A probabilistic approach to interpreting verbal autopsies: methodology and preliminary validation in vietnam. *Scandinavian Journal of Public Health*, 31(62 suppl):32–37, 2003. (pages 38 et 101)

[12] P. Byass, E. Fottrell, D.L Huong, Y. Berhane, T. Corrah, K. Kahn, and L. Muhe. Refining a probabilistic model for interpreting verbal autopsy data. *Scandinavian journal of public health*, 34(1):26–31, 2006. (page 38)

[13] B. Weldearegawi, Y.A. Melaku, M. Spigt, and G.J. Dinant. Applying the interva-4 model to determine causes of death in rural ethiopia. *Global health action*, 7, 2014. (page 38)

[14] C.J. Murray, A.D. Lopez, D.M. Feehan, S.T Peter, and G. Yang. Validation of the symptom pattern method for analyzing verbal autopsy data. 4(11):e327. (pages 39 et 40)

[15] G. King, Y. Lu, et al. Verbal autopsy methods with multiple causes of death. *Statistical Science*, 23(1):78–91, 2008. (pages 39 et 40)

[16] A.D. Flaxman, A. Vahdatpour, S. Green, S.L. James, et al. Random forests for verbal autopsy analysis: multisite validation study using clinical diagnostic gold standards. *Population Health Metrics*, 9(1):1, 2011. (pages 39 et 40)

[17] S.L. James, A.D. Flaxman, C.J. Murray, et al. Performance of the tariff method: validation of a simple additive algorithm for analysis of verbal autopsies. *Population Health Metrics*, 9(1):31, 2011. (pages 39 et 40)

[18] C.J. Murray, R. Lozano, A.D. Flaxman, P. Serina, D. Phillips, A. Stewart, S.L. James, A. Vahdatpour, C. Atkinson, M.K. Freeman K, et al. Using verbal autopsy to measure causes of death: the comparative performance of existing methods. *BMC Medicine*, 12(1):5, 2014. (page 40)

[19] P. Byass, D. Chandramohan, S.J. Clark, et al. Strengthening standardised interpretation of verbal autopsy data: the new interva-4 tool. *Global Health Action*, 5, 2012. (page 40)

[20] B.A Lopman, R.V. Barnabas, J.T. Boerma, G. Chawira, K. Gaitskell, T. Harrop, P. Mason, et al. Creating and validating an algorithm to measure aids mortality in the adult population using verbal autopsy. *Plos Med*, 3(8):e312. (page 40)

[21] M. Anker. The effect of misclassification error on reported cause-specific mortality fractions from verbal autopsy. *International Journal of Epidemiology*, 26(5):1090–1096, 1997. (page 40)

[22] A. Desgrées du Loû, G. Pison, B. Samb, and J.F. Trape. L'évolution des causes de décès d'enfants en Afrique : une étude de cas au Sénégal avec la méthode d'autopsie verbale. *Population*, pages 845–882, 1996. (page 40)

[23] G. Duthé, S.H.D.Faye, E. Guyavarch, P. Arduin, A.M. Kanté, A. Diallo, R. Laurent, A. Marra, and G. Pison. Changement de protocole dans la méthode d'autopsie verbale et mesure de la mortalité palustre en milieu rural sénégalais. *Bulletin de la Société de Pathologie Exotique*, 103(5):327–332, 2010. (page 40)

[24] K.G. Shojania, E.C. Burton, K.M. McDonald, and L. Goldman. Changes in rates of autopsy-detected diagnostic errors over time: a systematic review. *Jama*, 289(21):2849–2856, 2003. (page 40)

[25] A.K. Agrawala. Learning with a probabilistic teacher. *Information Theory, IEEE Transactions on*, 16(4):373–379, 1970. (page 45)

[26] A. Cornuejols and L. Miclet. *Apprentissage artificiel: Concepts et algorithmes*. Eyrolles, 2010. (pages 47, 60, 61, 63, 64, 66, 70 et 92)

[27] G. Celeux and J.P. Nakache. *Analyse discriminante sur variables qualitatives*. Polytechnica, 1994. (pages 47, 48, 50, 55 et 78)

[28] K. Pearson. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, pages 71–110, 1894. (page 50)

[29] C. Bouveyron, S. Girard, and C. Cordelia. High-dimensional discriminant analysis. *Communications in Statistics—Theory and Methods*, 36(14):2607–2623, 2007. (pages 50, 92 et 99)

[30] C. Bouveyron and C. Brunet. Simultaneous model-based clustering and visualization in the Fisher discriminative subspace. *Statistics and Computing*, 22:301–324, 2012. (pages 50 et 92)

[31] C. Bouveyron, M. Fauvel, and S. Girard. Kernel discriminant analysis and clustering with parsimonious Gaussian process models. *Statistics and Computing*, pages 1–20, 2014. (pages 50, 55, 68, 69, 70, 92, 94, 99, 102 et 107)

[32] F. Forbes and D. Wraith. Location and scale mixtures of Gaussians with flexible tail behaviour: Properties, inference and application to multivariate clustering. *Computational Statistics & Data Analysis*, 90:61–73, 2015. (page 50)

[33] C. Bouveyron. *Modélisation et Classification des données de grandes dimension: Application à l'analyse d'images*. PhD thesis, Université Joseph Fourier, 2016. (pages 50 et 53)

[34] G. McLachlan and D. Peel. *Finite mixture models*. John Wiley & Sons, 2004. (page 50)

[35] G. Govaert and M. Nadif. Block clustering with Bernoulli mixture models: Comparaison of different approaches. *Computational Statistics & Data Analysies*, 52:3233–3245, 2008. (pages 50 et 55)

[36] G. Celeux and G. Govaert. Clustering criteria for discrete data and latent class models. *Journal of Classification*, 8(2):157–176, 1991. (pages 50 et 56)

[37] J. Jacques and C. Biernacki. Analyse discriminante sur données binaires lorsque les populations d'apprentissage et de test sont différentes. In *DMAS*, page 129, 2005. (pages 50 et 55)

[38] M. Marbac, C. Biernacki, and V. Vandewalle. Model-based clustering for conditionally correlated categorical data. *Journal of Classification*, 2(32):145–175, 2015. (pages 50 et 55)

[39] R. Lebret, S. Iovleff, F. Langrognet, C. Biernacki, G. Celeux, and G. Govaert. Rmixmod: the r package of the model-based unsupervised, supervised and semi-supervised classification mixmod library. *Journal of Statistical Software*, pages In–press, 2015. (pages 50 et 55)

[40] G. Celeux and G. Govaert. Clustering criteria for discrete data and latent class models. *Journal of Classification*, 8(2):157–176, 1991. (pages 50, 55 et 92)

[41] G. Saporta. *Probabilités, analyse de données et statistique*. Editions Technip, Paris, 1990. (page 52)

[42] T. Pavlenko and D. Von Rosen. Effect of dimensionality on discrimination. *Statistics*, 35(3):191–213, 2001. (page 53)

[43] C. Bouveyron. *Contributions à l'apprentissage statistique en grande dimension, adaptatif et sur données atypiques*. Habilitation à diriger des recherches en mathématiques appliquées, Université Panthéon-Sorbonne-Paris I, 2012. (page 53)

[44] M.M. Dundar and D.A. Landgrebe. Toward an optimal supervised classifier for the analysis of hyperspectral data. *Geoscience and Remote Sensing, IEEE Transactions On*, 42(1):271–277, 2004. (page 53)

[45] C. Fraley and A.E. Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458):611–631, 2002. (page 53)

[46] J.H. Friedman. Regularized discriminant analysis. *Journal of the American Statistical Association*, 84(405):165–175, 1989. (page 53)

[47] C. Biernacki. Pourquoi les modèles de mélange pour la classification. *Revue de MODULAD*, 40:1–22. (page 53)

[48] C. Bouveyron, S. Girard, and C. Schmid. High-dimensional discriminant analysis. *Communications in Statistics—Theory and Methods*, 36(14):2607–2623, 2007. (page 54)

[49] N. Bouguila, D. Ziou, and J. Vaillancourt. Novel mixtures based on the dirichlet distribution: application to data and image classification. *In Machine Learning and Data Mining in Pattern Recognition*, Springer:172–181, 2003. (pages 55 et 92)

[50] C. Biernacki, G. Celeux, G. Govaert, and F. Langrognet. Model-based cluster and discriminant analysis with the mixmod software. *Computational Statistics & Data Analysis*, 51(2):587–600, 2006. (pages 56 et 57)

[51] V. Vandewalle. *Estimation et sélection en classification semi-supervisée*. Theses, Université des Sciences et Technologie de Lille - Lille I, December 2009. (page 57)

[52] T. Hofmann, B. Schölkopf, and A.J. Smola. Kernel methods in machine learning. *The Annals of Atatistics*, pages 1171–1220, 2008. (pages 58, 61, 63 et 66)

[53] I. Steinwart and A. Christmann. *Support vector machines*. Springer Science & Business Media, 2008. (pages 59 et 66)

[54] B. Schölkopf and A.J. Smola. *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. The Mit Press, 2002. (pages 60, 61, 63, 65 et 66)

[55] J. Shawe-Taylor and N. Cristianini. *Kernel methods for pattern analysis*. Cambridge University Press, 2004. (pages 60, 63, 66 et 148)

[56] D. Haussler. Convolution kernels on discrete structures. Technical report, Citeseer, 1999. (page 63)

[57] C.S. Leslie, E. Eskin, and W.S. Noble. The spectrum kernel: A string kernel for svm protein classification. In *Pacific Symposium on Biocomputing*, volume 7, pages 566–575, 2002. (page 64)

[58] H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, and C. Watkins. Text classification using string kernels. *The Journal of Machine Learning Research*, 2:419–444, 2002. (page 64)

[59] M. Collins and N. Duffy. Convolution kernels for natural language. In *Advances in Neural Information Processing Systems*, pages 625–632, 2001. (page 64)

[60] V. Vapnik. *Statistical learning theory*. Wiley. (page 65)

[61] N. Cristianini and J. Shawe-Taylor. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge University Press, 2000. (pages 65 et 66)

[62] T.J. Hastie, R.J. Tibshirani, and J.H. Friedman. *The elements of statistical learning : data mining, inference, and prediction*. Springer, 2009. (page 67)

[63] K. Hechenbichler and K. Schliep. Weighted k-nearest-neighbor techniques and ordinal classification. Technical report, 2004. (page 67)

[64] N.S. Altman. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185, 1992. (page 67)

[65] T. Hofmann, B. Schölkopf, and A.J. Smola. Kernel methods in machine learning. *The Annals of Statistics*, pages 1171–1220, 2008. (pages 68, 92 et 95)

[66] M.M. Dundart and D.A. Landgrebe. Toward an optimal supervised classifier for the analysis of hyperspectral data. *Geoscience and Remote Sensing, IEEE Transactions on*, 42(1):271–277, 2004. (pages 70 et 94)

[67] B.Schölkopf and K.R. Müller. Fisher discriminant analysis with kernels. *Neural networks for Signal Processing*, 1(1), 1999. (pages 70 et 94)

[68] E. Pekalska and B. Haasdonk. Kernel discriminant analysis for positive definite and indefinite kernels. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(6):1017–1032, 2009. (pages 70 et 94)

[69] J. Wang, J. Lee, and C. Zhang. Kernel trick embedded Gaussian mixture model. In *International Conference Learning Theory*, pages 159–174. Springer, 2003. (pages 70 et 94)

[70] Z. Xu, K. Huang, J. Zhu, I. King, and M.R. Lyu. A novel kernel-based maximum a posteriori classification method. *Neural Networks*, 22(7):977–987, 2009. (pages 70 et 94)

[71] R.B. Cattell. The scree test for the number of factors. *Multivariate Behavioral Research*, 1(2):245–276, 1966. (pages 70 et 94)

[72] S.N. Sylla, S. Girard, A.K. Diongue, A. Diallo, and C. Sokhna. A classification method for binary predictors combining similarity measures and mixture models. *Dependence Modeling*, 3:1090–1096, 2015. (pages 70, 74, 96 et 122)

[73] P. Jaccard. Etude comparative de la distribution florale dans une portion des alpes et des jura. *Bull Soc Vandoise Sci Nat*, 37:547–579, 1901. (page 70)

[74] I.C. Lerman. Sur l'analyse des données préalable à une classification automatique (proposition d'une nouvelle mesure de similarité). *Mathématiques et Sciences Humaines*, 32:5–15, 1970. (page 71)

[75] M. Rifqi. *Similarité, raisonnement et modélisation de l'utilisateur*. Habilitation à diriger des recherches, Université Pierre et Marie Curie (UPMC), 2010. (pages 71 et 73)

[76] A. Tversky. Feature of similarity. *Psychological Review*, 84:327–352, 1977. (pages 73, 96 et 97)

[77] F.B. Baulieu. A classification of presence/absence based dissimilarity coefficients. *Journal of Classification*, 6:233–246, 1989. (pages 73, 96 et 97)

[78] Z. Hubalek. Coefficients of association and similarity, based on binary (presence-absence) data: An evaluation. *Biological Reviews*, 57(4):669–689, 1982. (pages 73 et 96)

[79] D.A. Jackson, K.M. Somers, and H.H. Harvey. Similarity coefficients: Measures of co-occurrence and association or simply measures of occurrence? *The American Naturalist*, 133(3):436–453, 1989. (pages 74 et 96)

[80] B. Zhang and S.N. Srihari. Binary vector dissimilarity measures for handwriting identification. In *Electronic Imaging*, pages 28–38. International Society for Optics and Photonics, 2003. (pages 74 et 96)

[81] P. Willett. Similarity-based approaches to virtual screening. *Biochemical Society Transactions*, 31(3):603–606, 2003. (pages 74 et 96)

[82] H. Park and C. Jun. A simple and fast algorithm for k-medoids clustering. *Expert Syst. Appl.*, 36(2):3336–3341, 2009. (page 79)

[83] P.J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987. (pages 79 et 81)

[84] T. Velmurugan and T. Santhanam. Computational complexity between k-means and k-medoids clustering algorithm for normal and uniform distributions of data points. *Journal of Computer Science*, 6:363–368, 2010. (pages 79 et 81)

[85] H. Liu and H. Motoda. *Feature Selection for Knowledge Discovery and Data Mining*. Kluwer Academic Publishers, 1998. (page 82)

[86] G. McLachlan. *Discriminant analysis and statistical pattern recognition*, volume 544. John Wiley & Sons, 2004. (page 92)

[87] C. Bouveyron, S. Girard, and C. Cordelia. High-dimensional data clustering. *Computational Statistics & Data Analysis*, 52(1):502–519, 2007. (page 92)

[88] G.J. McLachlan, D. Peel, and R.W. Bean. Modelling high-dimensional data by mixtures of factor analyzers. *Computational Statistics & Data Analysis*, 41(3):379–388, 2003. (page 92)

[89] P.D Mcnicholas and T.B. Murphy. Parsimonious Gaussian mixture models. *Statistics and Computing*, 18(3):285–296, 2008. (page 92)

[90] A. Montanari and C. Viroli. Heteroscedastic factor mixture analysis. *Statistical Modelling*, 10(4):441–460, 2010. (page 92)

[91] T.B. Murphy, N. Dean, and A.E. Raftery. Variable selection and updating in model-based discriminant analysis for high dimensional data with food authenticity applications. *The Annals of Applied Atatistics*, 4(1):396, 2010. (page 92)

[92] M. Fauvel, C. Bouveyron, and S. Girard. Parsimonious Gaussian process models for the classification of hyperspectral remote sensing images. *Geoscience and Remote Sensing Letters, IEEE*, 12(12):2423–2427, 2015. (page 92)

[93] V. Batagelj and M. Bren. Comparing resemblance measures. *Journal of Classification*, 12:73–90, 1995. (page 95)

[94] L.A Goodman & W.H. Kruskal. Measures of association for cross classifications. *Journal of the American Statistical Association*, 49:732–764, 1954. (page 97)

[95] P.H.A. Sneath, R. Sokal, et al. *Numerical taxonomy: The principles and practice of numerical classification.* W.H.Freeman, 1973. (page 97)

[96] B.C. Reeves and M. Quigley. A review of data-derived methods for assigning causes of death from verbal autopsy data. *International Journal of Epidemiology*, 26(5):1080–1089, 1997. (page 101)

[97] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. (page 101)

[98] M. Gönen and E. Alpaydın. Multiple kernel learning algorithms. *The Journal of Machine Learning Research*, 12:2211–2268, 2011. (page 107)

[99] G. Lanckriet, N. Cristianini, P. Bartlett, L. Ghaoui, and M.I. Jordan. Learning the kernel matrix with semidefinite programming. *The Journal of Machine Learning Research*, 5:27–72, 2004. (page 114)

[100] F. Bach, G. Lanckriet, and M.I. Jordan. Multiple kernel learning, conic duality, and the smo algorithm. In *Proceedings of the twenty-first international conference on Machine learning*, page 6. ACM, 2004. (page 114)

[101] G. Lanckriet, T. De Bie, N. Cristianini, M. Jordan, and W.S. Noble. A statistical framework for genomic data fusion. *Bioinformatics*, 20(16):2626–2635, 2004. (page 115)

[102] P. Gehler and S. Nowozin. On feature combination for multiclass object classification. In *Computer Vision, 2009 IEEE 12th International Conference*, pages 221–228. IEEE, 2009. (page 115)

[103] F. Suard, A. Rakotomamonjy, and A. Bensrhair. Model selection in pedestrian detection using multiple kernel learning. In *Intelligent Vehicle Symposium*, pages 13–14, 2007. (page 115)

[104] M. Varma and D. Ray. Learning the discriminative power-invariance trade-off. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference*, pages 1–8. IEEE, 2007. (page 115)

[105] A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet. More efficiency in multiple kernel learning. In *Proceedings of the 24th international conference on Machine learning*, pages 775–782. ACM, 2007. (page 115)

[106] R. Kachouri. *Classification multi-modeles des images dans les bases hétérogenes*. PhD thesis, Université d'Evry Val d'Essonne, 2010. (page 115)

CHAPITRE 8

ANNEXE

Preuve du Lemme 4.2.1. Pour tout $k = 1, \ldots, K$, soit $\tilde{\rho}_k$ une fonction définie par:

$$\begin{aligned}
\tilde{\rho}_k(x, x') &= \tilde{\kappa}(x, x') - \frac{1}{n_k} \sum_{x_\ell \in C_k} (\tilde{\kappa}(x_\ell, x') + \tilde{\kappa}(x, x_\ell)) + \frac{1}{n_k^2} \sum_{x_\ell, x_{\ell'} \in C_k} \tilde{\kappa}(x_\ell, x_{\ell'}) \\
&= \eta\kappa(x, x') - \frac{1}{n_k} \sum_{x_\ell \in C_k} (\eta\kappa(x_\ell, x') + \eta\kappa(x, x_\ell)) + \frac{1}{n_k^2} \sum_{x_\ell, x_{\ell'} \in C_k} \eta\kappa(x_\ell, x_{\ell'}), \\
&= \eta\rho_k(x, x').
\end{aligned}$$

Soit la matrice de Gram définie par, $(\tilde{M}_k)_{\ell,\ell'} := \tilde{\rho}_k(x_\ell, x_{\ell'})/n_k = \eta(M_k)_{\ell,\ell'}$ pour tout $(\ell, \ell') \in \{1, \ldots, n_k\}^2$.

Soient $\tilde{\lambda}_{k1} \geq \cdots \geq \tilde{\lambda}_{kn_k}$, les plus grandes valeurs propres de $\tilde{M}_k$ et $\tilde{\beta}_{k1}, \ldots, \tilde{\beta}_{kn_k}$ leurs vecteurs propres associés.

On note $\tilde{\lambda}_{kj} = \eta\lambda_{kj}$ et $\tilde{\beta}_{kj} = \pm\beta_{kj}$ pour tout $(j, k) \in \{1, \ldots, n_k\}^2$. Il en découle que:

$$\tilde{\lambda} := \sum_{k=1}^{K} n_k(\text{trace}(\tilde{M}_k) - \sum_{j=1}^{d_k} \tilde{\lambda}_{kj}) \Big/ \sum_{k=1}^{K} n_k(r_k - d_k) = \eta\lambda$$

et donc,

$$\begin{aligned}\tilde{D}_k(x) &:= \frac{1}{n_k}\sum_{j=1}^{d_k}\frac{1}{\tilde{\lambda}_{kj}}\left(\frac{1}{\tilde{\lambda}_{kj}}-\frac{1}{\tilde{\bar{\lambda}}}\right)\left(\sum_{x_\ell\in C_k}\tilde{\beta}_{kj\ell}\tilde{\rho}_k(x,x_\ell)\right)^2+\frac{1}{\tilde{\bar{\lambda}}}\tilde{\rho}_k(x,x)\\ &+ \sum_{j=1}^{d_k}\log(\tilde{\lambda}_{kj})+(d_{\max}-d_k)\log(\tilde{\lambda})-2\log(n_k)\\ &= D_k(x)+d_{\max}\log\eta.\end{aligned}$$

De plus $d_{\max}\log\eta$ ne dépend pas de k, et par conséquent les deux règles de classification sont équivalentes.

Preuve du Lemme 4.4.1. En simplifiant les notations, on a : $\kappa(x,x') =< x,x' >$ et

$$\begin{aligned}\tilde{\kappa}(x,x') &= <\mathbf{1}-x,\mathbf{1}-x'>\\ &= <\mathbf{1},\mathbf{1}>-<\mathbf{1},x>-<\mathbf{1},x'>+<x,x'>\\ &= \kappa(1,1)-\kappa(\mathbf{1},x)-\kappa(\mathbf{1},x')+\kappa(x,x').\end{aligned}$$

pour tout $k=1,\ldots,K$, en remplaçant $\tilde{\kappa}(x,x')$ dans le calcul de $\rho_k(x,x')$ on a :

$$\tilde{\rho}_k(x,x')=\tilde{\kappa}(x,x')-\frac{1}{n_k}\sum_{x_\ell\in C_k}(\tilde{\kappa}(x_\ell,x')+\tilde{\kappa}(x,x_\ell))+\frac{1}{n_k^2}\sum_{x_\ell,x_{\ell'}\in C_k}\tilde{\kappa}(x_\ell,x_{\ell'}),$$

ainsi $\tilde{\rho}_k(x,x')=\rho_k(x,x')$, de plus les deux fonctions de classification sont équivalentes.

Preuve du Lemme 4.4.3. En remarquant que

$$\tilde{\kappa}(x,x') = \exp\left(\frac{\tilde{S}(x,x')}{2\sigma^2}\right)=\exp\left(\frac{\eta S(x,x')+\mu}{2\sigma^2}\right)=\eta'\exp\left(\frac{S(x,x')}{2\sigma'^2}\right)$$

avec $\eta'=\exp(\mu/(2\sigma^2))$ et $\sigma'=\sigma/\sqrt{\eta}$.
On a: $\tilde{\kappa}(x,x')=\eta'\kappa(x,x')$ d'où l'équivalence d'aprés le Lemme 4.2.1

Preuve de la Proposition 4.4.1. En posant:

$$\begin{aligned} S_1(x,x') &= \alpha a - \theta(b+c) + \beta d, \\ S_2(x,x') &= \alpha' a + \theta'(b+c) + \beta' d, \end{aligned}$$

on a:

$$\kappa(x,x') = \exp\left(\frac{1}{2\sigma^2}\frac{S_1(x,x')}{S_2(x,x')}\right).$$

–Dans un premier temps, on montre que S_1 définit bien un noyau.
Notons que si $\theta = 0$, alors $S_1(x,x') = \alpha\kappa_{\text{linéaire}}(x,x') + \beta\tilde{\kappa}_{\text{linéaire}}(x,x')$ et on sait que la somme de deux noyaux est un noyau S_1 définit bien un noyau.

Dans le cas où $\theta > 0$, on peut écrire que

$$S_1(x,x') = \alpha a - \theta(p-a-d) + \beta d = \theta p(ua + vd - 1)$$

avec $u := (1+\alpha/\theta)/p > 0$ et $v := (1+\beta/\theta)/p > 0$.

ainsi S_1 vérifie les conditions (4.2).

– En second lieu, on montre que $1/S_2$ définit bien un noyau. On se focalise sur le cas où $0 \geq \alpha', \beta' < \theta'$. Notons que les autres cas sont similaires.
En posant $u' := (1-\alpha'/\theta')/p > 0$ et $v' := (1-\beta'/\theta')/p > 0$ on a:

$$S_2(x,x') = \alpha' a + \theta'(p-a-d) + \beta' d = \theta' p[1 - (u'a + v'd)]$$

où $u' \in [0,1)$ et $v' \in [0,1)$.

Pour $0 \leq u'a + v'd < 1$, en utilisant un développement limité on a:

$$\frac{1}{S_2(x,x')} = \frac{1}{\theta' p}\sum_{i=0}^{\infty}(u'a + v'd)^i.$$

Pour tout $N > 0$, on a:

$$S_{3,N}(x,x') = \frac{1}{\theta' p}\sum_{i=0}^{N}(u'a + v'd)^i.$$

Ainsi $S_{3,N}$ est une somme et produit de noyaux linéaires $\kappa_{\text{linéaire}}$ et $\tilde{\kappa}_{\text{linéaire}}$. D'après la Proposition

3.22 de [55] alors $S_{3,N}$ définit bien un noyau pour tout $N > 0$. Par conséquent, $S_{3,N}$ vérifie les conditions (4.2) pour tout $N > 0$. De même pour $N \to \infty$, on a aussi $1/S_2$ qui définit un noyau.

– Finalement, d'après les Proposition 3.22 et Proposition 3.24 de [55], alors κ définit bien un noyau.

OBSERVATOIRES DE POPULATION DE BANDAFASSI & MLOMP
SÉNÉGAL

Version de Janvier 2006

Collecte d'avril 2007

FICHE D'ENQUÊTE DÉCÈS

(Autopsie verbale)

Type de décès : ☐ Nouveau-né ☐ Enfant ☐☐
☐ Adulte femme ☐ Adulte homme

Enquêteur : _____ ☐☐

Date de visite : ☐☐ ☐☐ ☐☐☐☐

Village : ☐☐☐ _____

Concession : ☐☐☐ _____

Identité : ☐☐☐☐☐ _____

Mère : ☐☐☐☐☐ _____

Sexe : ☐ **Date de naissance** : ☐☐ ☐☐ ☐☐☐☐

Date de décès : ☐☐ ☐☐ ☐☐☐☐

Répondant :..... ☐☐

Âge déclaré au décès (jours ou semaines pour les bébés) :..... ☐☐☐

Lieu du décès :..... ☐☐☐

Cause déclarée :..... ☐☐

Quel est le nom local de la maladie ? _____
(en langue diola, peule, bedik ou malinké)

La personne a-t-elle été conduite au dispensaire ou à l'hôpital ?

☐ OUI où ?_____ Date : ☐☐ ☐☐ ☐☐☐☐
☐ NON

Diagnostic recopié du registre :_____

HISTOIRE DE LA MALADIE AYANT CONDUIT AU DÉCÈS

(Faire la chronologie des événements qui se sont déroulés au cours de la maladie ou après l'accident ayant conduit au décès. En cas de symptômes particuliers, faire décrire ou mimer par le répondant. Dans chaque cas, préciser les traitements reçus et l'ordre de succession des événements).

S'AGIT-IL D'UN ACCIDENT ? ☐ **OUI** ☐ **NON**

(si oui, ne remplir que l'histoire)

HISTOIRE DES SYMPTÔMES ET DES TRAITEMENTS :

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

Quel a été le 1er symptôme ? _____

Y a-t-il eu une maladie avant ce premier symptôme ? _____

Combien de temps a duré la maladie qui a conduit au décès ? _____

Quels ont été, dans l'ordre chronologique, les traitements reçus jusqu'au décès ?

Quels médicaments / soins ?	Où / par qui ?	A quelle date ?

FIÈVRE OU CORPS CHAUD ☐ **OUI** ☐ **NON**

Combien de temps cela a-t-il duré ? _____
Quand cela a-t-il commencé ? _____
Quand cela s'est-il terminé ? _____

(Cocher la case si le symptôme est présent)

elle très forte ?..... ☐
moyenne ?..... ☐
intermittente ?..... ☐
continue ?..... ☐
Avait-il des sueurs ?..... ☐
Avait-il des frissons ?..... ☐

Lui a-t-on donné de la nivaquine ou de la chloroquine au cours de la fièvre ?

☐ Oui ▭ combien de fois ?_____
combien de comprimés à chaque fois ?_____

☐ Non

Lui a-t-on donné un autre traitement en comprimés ou sirop ?

☐ Oui ▭ Lequel ?_____

☐ Non

A-t-il reçu une injection pour cette fièvre ?

☐ Oui ▭ Lieu :_____
Date : |__|__| |__|__| |__|__|__|__|

☐ Non

DIARRHÉE OU DYSENTERIE ☐ **OUI** ☐ **NON**

Combien de temps cela a-t-il duré ? _____
Quand cela a-t-il commencé ? _____
Quand cela s'est-il terminé ? _____
Combien avait-il de selles par jour ? _____

(Cocher la case si le symptôme est présent)

Les selles étaient-elles comme de l'eau (incolore) ?..... ☐
comme des crachats ?..... ☐
avec du sang ?..... ☐

SIGNES DE DÉSHYDRATATION ☐ OUI ☐ NON

	Oui	Non
Avait-il la bouche et la langue sèches ou était-il assoiffé ?	☐	☐
Avait-il les yeux enfoncés ?	☐	☐
Avait-il la fontanelle déprimée (enfant de moins de 2 ans) ?	☐	☐

VOMISSEMENTS ☐ OUI ☐ NON

Combien de temps cela a-t-il duré ?_____

Quand vomissait-il au cours de la maladie ? _____

De quelle couleur étaient ces vomissements ?_____

S'agissait-il de vomissements en jet (comme un robinet) ? ☐ Oui ☐ Non

CRISES CONVULSIVES ☐ OUI ☐ NON

Combien y a-t-il eu de crises (préciser sur quelle période) ?_____

Combien de temps a duré chaque crise ? _____

Quand ces crises sont-elles survenues au cours de la maladie ? _____

Description (signes pendant la crise : cocher la case si le symptôme est présent)

Avait-il des spasmes (mouvement brusque et incontrôlé) ? ☐

Criait-t-il ou pleurait-t-il ? ☐

Urinait-il ? ☐

Se mordait-il la langue ? ☐

Hypersalivait-il (bavait beaucoup) ? ☐

Respirait-il bruyamment ? ☐

La fontanelle était-elle gonflée (enfant de moins de 2 ans) ? ☐

Avait-il le cou tordu en arrière ? ☐

Avait-il le corps raidi en arrière ? ☐

Avait-il les jambes tendues ? ☐

Avait-il les jambes pliées ? ☐

Avait-il les bras tendus ? ☐

Avait-il les bras pliés ? ☐

Avait-il les poings fermés ? ☐

Avait-il la bouche fermée ou crispée (ne pouvait plus téter)? ☐

Perdait-il connaissance ? ☐

S'AGISSAIT-IL D'ÉPILEPSIE ☐ OUI ☐ NON

Depuis combien de temps faisait-il des crises ? _____

Etait-il soigné ?

☐ Oui ␣ Préciser où : _____

☐ Non

SIGNES NEUROLOGIQUES EN DEHORS D'UN CONTEXTE DE CRISES CONVULSIVES ☐ **OUI** ☐ **NON**

Y a-t-il eu perte de connaissance ou coma ?

☐ Oui ▭ Quand au cours de la maladie ? : _____

☐ Non

Y a-t-il eu paralysie du corps ou d'un membre ?

☐ Oui ▭ Préciser quelle(s) partie(s) : _____

☐ Non

DIFFICULTES À RESPIRER ☐ **OUI** ☐ **NON**

Combien de temps cela a-t-il duré ?_____

Quand cela a-t-il commencé ?_____

Quand cela s'est-il terminé ?_____

(Cocher la case si le symptôme est présent)

Respirait-il rapidement ?..... ☐

Respirait-il difficilement (s'étouffait) ?..... ☐

Respirait-il bruyamment ?..... ☐

Avait-il une respiration sifflante ?..... ☐

Avait-il les ailes du nez palpitantes ?..... ☐

La peau rentrait-elle dans les côtes ?..... ☐

TOUX ☐ **OUI** ☐ **NON**

Combien de temps cela a-t-il duré ?_____

Quand cela a-t-il commencé ?_____

Quand cela s'est-il terminé ?_____

(Cocher la case si le symptôme est présent)

Toussait-il la nuit ? ☐

Crachait-il après la toux ? ☐

▭ **si oui**, les crachats étaient-ils comme du pus ?..... ☐

comme de la mousse ?..... ☐

avec du sang ?..... ☐

nauséabonds ?..... ☐

Vomissait-il après la toux ? ☐

Perdait-il sa respiration en toussant ? ☐
Faisait-il des quintes de toux (groupes de toux)? ☐

BOUTONS ☐ **OUI** ☐ **NON**

Combien de temps cela a-t-il duré ?_____
Quand cela a-t-il commencé ?_____
Quand cela s'est-il terminé ?_____
A quel(s) endroit(s) du corps les boutons étaient situés ?_____
_____
_____
_____
_____

Sur quelle partie du corps sont-ils apparus en premier ?_____

(Cocher la case si le symptôme est présent)

Sont-ils apparus..... ensemble ?..... ☐
les uns après les autres ?..... ☐
Etaient-ils..... aplatis ?..... ☐
saillants ?..... ☐
grands ?..... ☐
petits ?..... ☐
Contenaient-ils..... de l'eau ?..... ☐
du pus ?..... ☐
Ont-ils cicatrisé avant le décès ?..... ☐
La peau a-t-elle desquamé ?..... ☐

PLAIES, BRÛLURES, ABCÈS ☐ **OUI** ☐ **NON**

	Oui	Non
Y avait-il des plaies ?.....	☐	☐
▭ si oui : étaient-elles infectées ?	☐	☐
Y avait-il des brûlures ?.....	☐	☐
Y avait-il des gonflements contenant du pus (abcès) ?.....	☐	☐

Pour les plaies, brûlures et abcès, indiquer la localisation :
_____
_____
_____
_____

SAIGNEMENTS ☐ **OUI** ☐ **NON**

Où étaient localisés ces saignements ?_____

Combien de fois a-t-il saigné ?_____

Quand au cours de la maladie a-t-il saigné ?_____

ŒDÈMES (CORPS GONFLÉ) ☐ **OUI** ☐ **NON**

Combien de temps cela a-t-il duré ?_____

Quand cela a-t-il commencé ?_____

Quand cela s'est-il terminé ?_____

Sur quelles parties du corps étaient-ils situés ?_____

VENTRE GONFLÉ ☐ **OUI** ☐ **NON**

Combien de temps cela a-t-il duré ?_____

Quand cela a-t-il commencé ?_____

Quand cela s'est-il terminé ? _____

Une ponction a-t-elle été pratiquée ?

☐ Oui ▭ Dans quelle formation sanitaire ? : _____

☐ Non

DIFFICULTÉS À URINER, PROBLÈMES URINAIRES ☐ **OUI** ☐ **NON**

Combien de temps cela a-t-il duré ?_____

Quand cela a-t-il commencé ?_____

Quand cela s'est-il terminé ?_____

Avait-il des douleurs en urinant ? ☐ Oui ☐ Non

COULEUR ANORMALE DES URINES ☐ **OUI** ☐ **NON**

De quelle couleur étaient ces urines ?_____

Quand au cours de la maladie ?_____

COULEUR ANORMALE DES SELLES ☐ **OUI** ☐ **NON**

De quelle couleur étaient ces selles ?_____

Quand au cours de la maladie ?____

MAL AUX YEUX, COULEUR ANORMALE DES YEUX ☐ **OUI** ☐ **NON**

Quand au cours de la maladie ?_____

		Oui	Non
Avait-il les yeux	rouges ?.....	☐	☐
	jaunes ?.....	☐	☐
	larmoyants ?.....	☐	☐

S'il s'agit d'un petit enfant (moins de 2 ans) ⊟ Aller à la page 9

MAUX DE POITRINE, MAUX DE CÔTES ☐ **OUI** ☐ **NON**

Combien de temps cela a-t-il duré ?_____
Quand cela a-t-il commencé ?_____
Quand cela s'est-il terminé ?_____

MAUX DE TÊTE ☐ **OUI** ☐ **NON**

Combien de temps cela a-t-il duré ?_____
Quand cela a-t-il commencé ?_____
Quand cela s'est-il terminé ?_____
Avait-il des bourdonnements d'oreilles ? ☐ Oui ☐ Non
Avait-il des troubles visuels ? ☐ Oui ☐ Non

MAUX DE VENTRE ☐ **OUI** ☐ **NON**

Combien de temps cela a-t-il duré ?_____
Quand cela a-t-il commencé ?_____
Quand cela s'est-il terminé ?_____

AUTRES SYMPTÔMES ☐ **OUI** ☐ **NON**

Préciser lesquels :_____
Combien de temps ont-t-ils duré ?_____
Quand cela a-t-il commencé ?_____
Quand cela s'est-il terminé ?_____

SIGNES GÉNÉRAUX

(Cocher la case si le symptôme est présent)

Avait-il des démangeaisons, prurit ?..... ☐
A-t-il maigri au cours de la maladie ?..... ☐
Etait-il déjà maigre au début de la maladie ?..... ☐
A-t-il arrêté de manger au cours de la maladie ?..... ☐
La couleur de la paume des mains a-t-elle changé ?..... ☐
Le corps a-t-il changé de couleur ?..... ☐
La langue était-elle pâle ?..... ☐
Mangeait-il de la terre ?..... ☐
Etait-il constipé ?..... ☐
Avait-il très soif durant la maladie ?..... ☐

D'autres personnes ou d'autres enfants ont-ils eu les mêmes symptômes à la même période ?

☐ Oui ▭ Dans quel village ? : _____
☐ Non

Remarques :

_____
_____
_____
_____
_____

LA PERSONNE SOUFFRAIT-ELLE D'UNE MALADIE CHRONIQUE ? ☐ **OUI** ☐ **NON**

☐ Oui ▭ Quelle maladie : _____
Depuis quand : _____
Quels traitements :_____
_____
_____

☐ Non

S'AGIT IL :

☐ D'un décès avant 5 ans ?
▭ L'enfant était-il sevré au moment du décès ? ☐ Oui ☐ Non
☐ D'un décès de nouveau né (survenu dans les 4 semaines suivant la naissance) ?
☐ D'un mort-né ?
☐ Du décès d'une femme enceinte ?
☐ D'un décès de femme âgée de 12 à 49 ans ?
▭ Quelle est la date de fin de la dernière grossesse |_|_| |_|_| |_|_|_|_|

LE RESTE DU QUESTIONNAIRE N'EST À REMPLIR QUE DANS LES CAS SUIVANTS :

- Mort-né
- Décès d'un nouveau-né (moins de 4 semaines)
- Décès d'une femme ayant entre 12 et 49 ans, et ayant eu une grossesse moins d'un an avant son décès

GROSSESSES PRÉCÉDENTES

Y a-t-il eu des problèmes pendant les grossesses et/ou accouchements précédents ?

☐ Oui ▭ lesquels ? : _____

☐ Non

Une césarienne a-t-elle été pratiquée lors d'une grossesse précédente ? ☐ Oui ☐ Non

HISTOIRE DE LA DERNIÈRE GROSSESSE

Combien de temps a duré la grossesse *?* |__|__| mois

La mère a-t-elle été malade durant la grossesse ?

(Cocher la case si le symptôme est présent)

☐ Oui ▭ Avait-elle les jambes enflées ?..... ☐
les mains enflées ?..... ☐
le visage enflé ?..... ☐
de l'hypertension artérielle ?..... ☐
des convulsions ?..... ☐
de la fièvre ?..... ☐
Saignait-elle ?..... ☐

☐ Non

A-t-elle été soignée au cours de la grossesse ?

☐ Oui ▭ Quels soins ? : _____

☐ Non

A-t-elle eu un régime particulier ?

☐ Oui ▭ Lequel ? : _____

☐ Non

Est-elle allée à la visite prénatale ?

☐ Oui ▭ Où ? : _____

☐ Non

A-t-elle reçu une injection contre le tétanos ?

☐ Oui ▭ Où ? : _____

☐ Non

ACCOUCHEMENT À LA SUITE DE LA DERNIÈRE GROSSESSE

L'accouchement a-t-il eu lieu à domicile ?..... ☐
pendant le transport ?..... ☐
dans un établissement de santé ?..... ☐

Si l'accouchement s'est déroulé dans un établissement de santé, préciser, quel(le) était :
la localité ?_____
l'établissement ?_____

L'accouchement a-t-il présenté des difficultés ou des complications ?
☐ Oui ▭ Lesquelles ? : _____
☐ Non

L'accouchement s'est-il déroulé par voie normale ?..... ☐.
par césarienne ?..... ☐.

Combien de temps a duré le travail ? ☐☐ Heures

	Oui	Non
S'agit-il d'une naissance multiple ?.....	☐	☐
La tête est-elle venue la première ?.....	☐	☐
La rupture de la poche des eaux s'est-elle faite plus de 12 heures avant l'accouchement ?.....	☐	☐
A-t-elle eu de la fièvre au-delà de 24 heures après l'accouchement ?..	☐	☐

Le placenta est-il venu normalement et en entier ?

☐ Oui ▭ Si oui à quel moment : Pendant le travail ☐
Après la délivrance ☐

☐ Non

La femme a-t-elle saigné longtemps ?
☐ Oui ▭ Combien de temps ? : _____
Quelle était la couleur du sang ? : _____
☐ Non

ÉTAT DE L'ENFANT

(Cocher la case correspondant à la situation)

L'enfant était-il né vivant ?..... ☐
mort-né ?..... ☐
avorté ?..... ☐
non né ?..... ☐

	Oui	Non
A-t-il crié rapidement après la naissance ?.....	☐	☐
A-t-il respiré normalement après la naissance ?.....	☐	☐
Etait-il trop gros ?.....	☐	☐
trop maigre ?.....	☐	☐
trop petit ?.....	☐	☐
trop grand ?.....	☐	☐
Avait-il la tête trop grosse ?.....	☐	☐
A-t-il uriné ?.....	☐	☐
A-t-il déféqué ?.....	☐	☐
En cas de mort-né, le fœtus était-il macéré ?.....	☐	☐

Avait-il une malformation visible à la naissance ?

☐ Oui ▭ Laquelle ? : _____

☐ Non

A-t-il tété ?

	Oui	Non
☐ Oui ▭ s'est-il arrêté de téter après quelques jours ?	☐	☐

☐ Non

Remarques : _____

_____

_____

_____

_____

_____

_____

_____

_____

Documentation de la base de données des autopsie verbales

Cette base de données des autopsies verbales est réalisé grâce à la collaboration de :

Mr Amadou Lamine NDONGO, informaticien

Mr Mouhamadou Baba SOW, responsable Base de données Bandafassi

Avec la participation de :

Mr Ekoué Kouévidjin, responsable Base de Données Niakhar

Mr Ousmane Ndiaye, responsable Base de Données Mlomp

INTRODUCTION

Le travail de la thèse est centré de manière générale sur la modélisation et l'analyse des données d'autopsies verbales.

Pour ce faire, nous devons fournir une base de données portant sur l'ensemble des fiches d'autopsies verbales collectées depuis 1985.

Cette base permet l'intégration de données multiples et hétérogènes. Elle enregistre tous les données sanitaires, médicales et sociodémographiques dans la suite du processus des autopsies verbales.

I. Architecture de la base

Cette base de données obéit à un schéma dite « schéma en étoile ».

Le schéma en étoile se résume en une classe principale (table fiche) et interconnecté à plusieurs tables (41 tables).

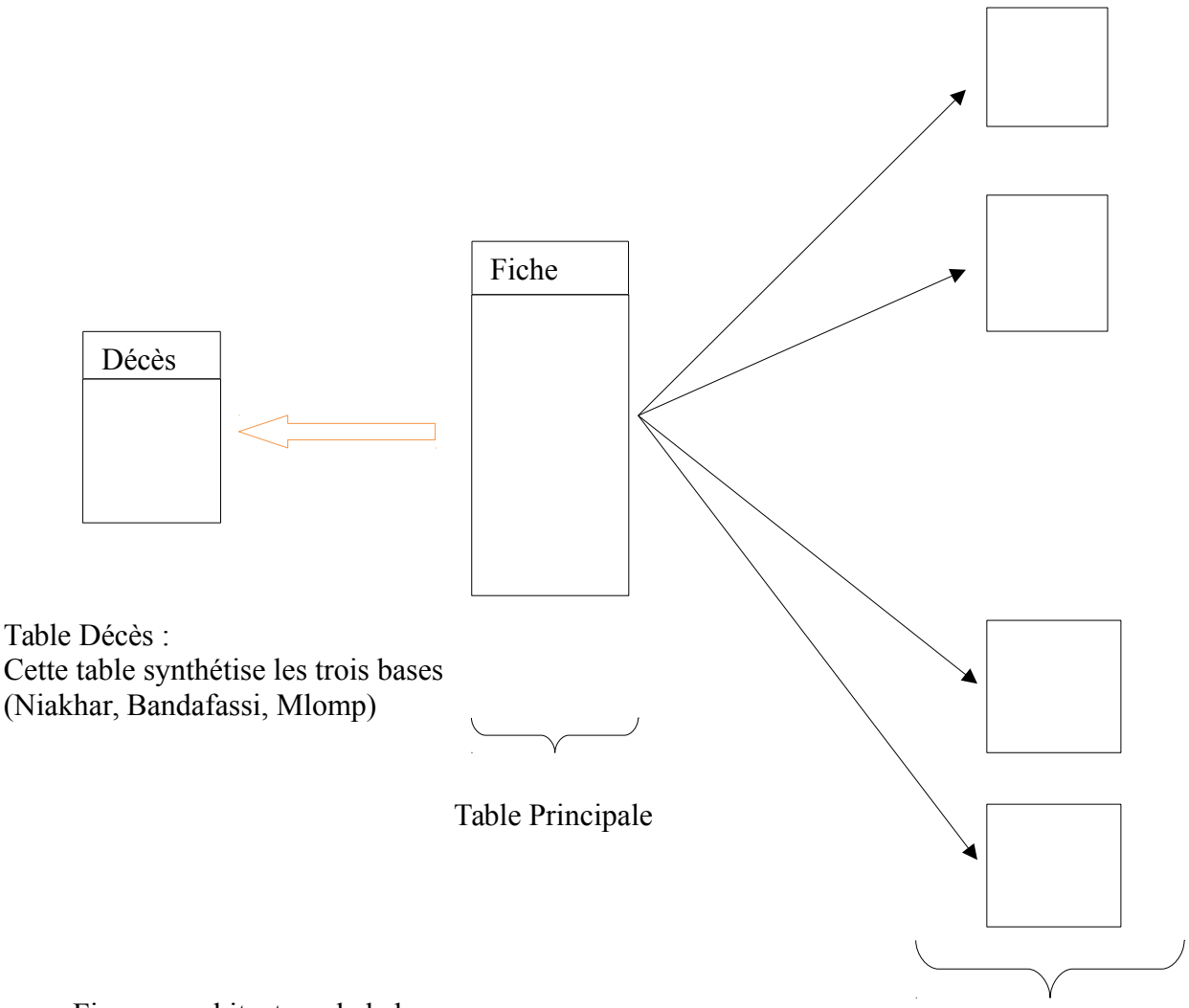


Figure : architecture de la base

II. Modèle Conceptuel de Données

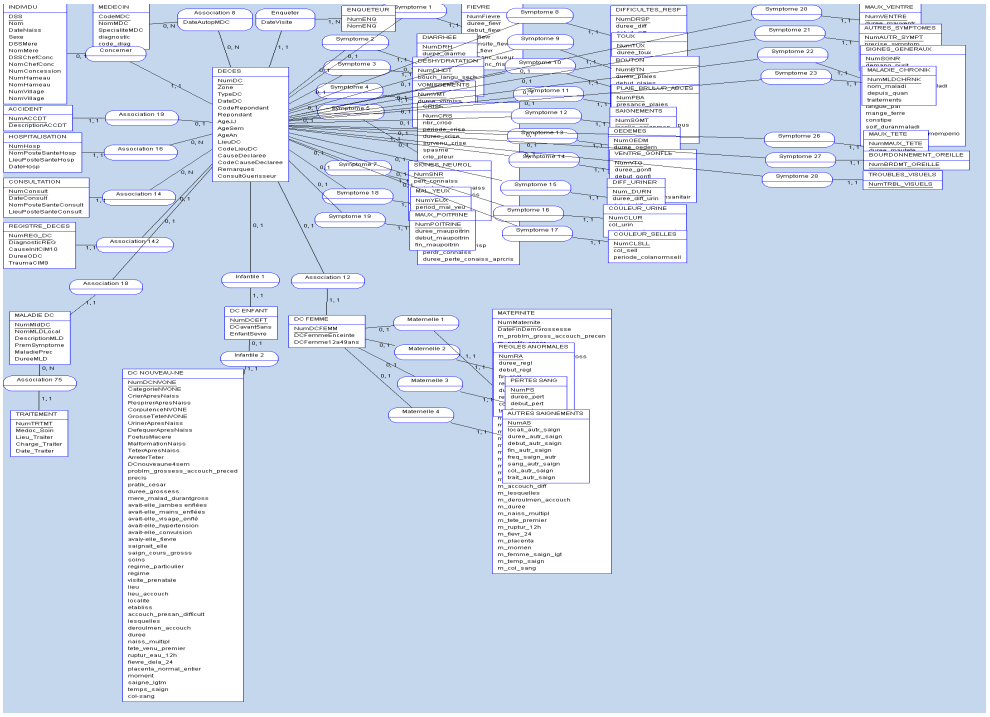


Figure : le MCD

III. Logiciels utilisés

- le Modèle Conceptuel de Données (MCD) et le dictionnaire de données est réalisé par le logiciel AnalyseSI-0.75
- L'architecture de la base de données est réalisée sous EasyPHP 5. 3.6.1
- Le formulaire de saisie est crée avec Adobe DreamWeaver CS5.5 et des applications en JavaScript pour assurer les contrôles.

Remarques : les variables mis en gras sont des clés étrangères.

IV. Composition des tables

Identification

Table décès

Variable	Type	Description
DSS	Texte	Identifiant de la personne décédée
Nom	Texte	Nom et prénom
DateNaiss	Texte	Date de naissance
Sexe	numérique	Le sexe du décédé
DSSMere	Texte	Identifiant de la mère
NomMere	Texte	Nom et prénom de la mère
DSSChefConc	Texte	Identifiant du chef de concession
NomChefConc	Texte	Nom et prénom du chef de concession
NumConcession	Numérique	Numéro de la concession
NumHameau	Numérique	Numéro du hameau
NomHameau	Texte	Nom du hameau
NumVillage	Numérique	Numéro du village
NomVillage	Texte	Nom du village

Table fiche

Variable	Type	Description
Numfiche		
TypeDC	Numérique	Type de décès
DateDC	Texte	Date du décès
CodeRepondant	Texte	Code lien de parenté répondant -personne décédée
Repondant	Texte	Nom du répondant
AgeJJ	numérique	Age en jours de la personne
AgeSem	numérique	Age en semaine de la personne
AgeAn	numérique	Age en année de la personne
LieuDC	Texte	Lieu de décès
CodeLieuDC	Numérique	Code du lieu de décès
CauseDeclaree	Texte	Cause de décès déclarée par le répondant
CodeCauseDeclaree	Numérique	Code cause déclarée(CIM)
Remarques	Texte	Remarques sur le décès
ConsultGuerisseur	Numérique	Consultation de guérisseur
Age	Numérique	Age de la personne
DateVisite	Texte	Date de visite de l'enquêteur
AutreLieuDC	Texte	Autre lieu de décès
DSS	Texte	

Table accident

Variable	Type	Description
NumAccident		
NumACCDT	Numérique	Il s'agit d'un accident
DescriptionACCDT	Texte	Histoire de l'accident ou l'histoire de la maladie
Numfiche		

Table enquêteur

Variable	Type	Description
NumENQ		
NomENQ	Texte	Nom de l'enquêteur

Table maladie chronique

Variable	Type	Description
NumMLDCHRNK		
nom_maladi	Texte	Nom de la maladie
depuis_quan	Texte	Depuis quand est survenu cette maladie
Traitements	Texte	Traitements reçus
Numfiche		

Table maladie au décès

Variable	Type	Description
NumMldDC		
NomMLDLocal	Texte	Nom local de la maladie
PremSymptome	Texte	Le premier symptôme
MaladiePrec	Texte	La maladie précédente
DureeMLD	Texte	Durée de la maladie
mem_symptom_period	Numérique	Ya-t-il les mêmes symptômes a la même période ?
precis_village	Texte	Nom du village
Numfiche		

Tables recours aux soins et traitements

Table consultation

Variable	Type	Description
NumConsult		
DateConsult1	Texte	Date de la première consultation
NomPosteSanteConsult1	Texte	Nom de la poste de santé
LieuPosteSanteConsult1	Texte	Lieu de consultation
DateConsult1	Texte	Date de la deuxième consultation
NomPosteSanteConsult1	Texte	Nom de la poste de santé
LieuPosteSanteConsult1	Texte	Lieu de consultation
Numfiche		

Table Hospitalisation

Variable	Type	Description
NumHosp		
NomPosteSanteHosp1	Texte	Nom du poste d'hospitalisation
LieuPosteSanteHosp1	Texte	Lieu d'hospitalisation
DateHosp1	Texte	Date de la première hospitalisation
NomPosteSanteHosp2	Texte	Nom du poste d'hospitalisation
LieuPosteSanteHosp2	Texte	Lieu d'hospitalisation
DateHosp2	Texte	Date de la deuxième hospitalisation
Numfiche		

Table traitement

Variable	Type	Description
NumTRTMT		
Medoc_Soin1	Texte	Médicaments reçu lors du premier traitement
Lieu_charge_Traiter1	Texte	Lieu de traitement
Date_Traiter1	Texte	Date du second traitement
Medoc_Soin2	Texte	Médicaments reçu lors du second traitement
Lieu_charge_Traiter2	Texte	Lieu de traitement
Date_Traiter2	Texte	Date du troisième traitement
Medoc_Soin3	Texte	Médicaments reçu lors du troisième traitement
Lieu_charge_Traiter3	Texte	Lieu de traitement
Date_Traiter3	Texte	Date du troisième traitement
NumMldDC		

Tables diagnostique :

Table diagnostiqué

Variable	Type	Description
NumFiche		
CodeMDC		
DateAutopMDC	Texte	Date de l'autopsie du médecin

Table médecin

Variable	Type	Description
CodeMDC	Numérique	Numéro du médecin chargé du diagnostic
Nom_medecin1	Texte	Nom du premier médecin en charge du diagnostic
SpecialiteMDC	Texte	Spécialiste du médecin
diagnostic1	Texte	Diagnostic du médecin
code_diag_gran_rubrik1	Texte	Code diagnostic par grand rubrique
code_diag_patho1	Texte	Code diagnostique par pathologie
signes_majeurs1	Texte	Présence signes majeurs
Lesquels1	Texte	Listing des signes majeurs

Table Registre fiche de décès

Variable	Type	Description
NumREG_DC	Numérique	Numéro du registre
DiagnosticREG	Texte	Diagnostic recopié du registre
CauseInitCIM10	Texte	Diagnostic recopié du CIM 10
DureeODC	Texte	Durée de la maladie jusqu'au décès
TraumaCIM9	Texte	Diagnostic recopié du CIM 9
Numfiche		

Tables profil symptomatologique

Table fièvre ou corps chaud

Variable	Type	Description
NumFievre		
duree_fievr	Texte	Durée de la fièvre
debut_fievr	Texte	Quand cela à t-il commencé
fin_fievr	Texte	Quant cela s'est-il terminé
intensite_fievre	Numérique	Intensité de la fièvre
eta_fievr	Numérique	Etat de la fièvre
presanc_sueur	Numérique	Présence de sueur
presanc_frisson	Numérique	Présence de frissons
Numfiche		

Table diarrhée

Variable	Type	Description
NumDRH		
duree_diarrhe	Texte	durée de la diarrhée
debut_diarrhe	Texte	Quand cela à t-il commencé
fin_diarrhe	Texte	Quant cela s'est-il terminé
nbr_selle_jr	Numérique	nombre de selles par jour
col_sel_eau	Numérique	selles incolores
col_sel_eau	Numérique	selles comme crachat
sel_avec_sang	Numérique	selles avec sang
col_sel_autr	Numérique	Un autre couleur pour les selles
precise_col_sel	Texte	Préciser couleur du sang
Numfiche		

Table autres symptômes

Variable	Type	Description
NumAUTR_SYMPT		
precise_symptom	Texte	Préciser le symptôme
duree_symptom	Texte	Durée du symptôme
debut_symptom	Texte	Quand cela à t-il commencé
fin_symptom	Texte	Quant cela s'est-il terminé
Numfiche		

Table bourdonnement oreille

Variable	Type	Description
NumBRDMT_OREILLE	Numérique	présence d'un bourdonnement d'oreille
Numfiche		

Table bouton

Variable	Type	Description
NumBTN		
duree_plaies	Texte	durée des boutons
debut_plaies	Texte	Quand cela à t-il commencé
fin_plaies	Texte	Quant cela s'est-il terminé
situation_plaies	Texte	Situation des plaies
localis_premier_plai	Texte	Localisation de la première plaie
mode_apparution_plaies	Numérique	Mode d'apparition des plaies
contenanc_plaies	Numérique	Contenance des plaies
cicatris_avan_dc	Numérique	La plaie s'est-elle cicatrisée avant le décès
peau_desquame	Numérique	La plaie s'est-elle desquamée

plaies_aplatis	Numérique	La plaie était-elle aplatie
plaies_sailants	Numérique	La plaie était-elle saillant
plaies_grand	Numérique	La plaie était-elle grand
plaies_petit	Numérique	La plaie était-elle petit
plaies_eau	Numérique	La plaie contenait-elle de l'eau
plaies_pus	Numérique	La plaie contenait-elle du pus
Numfiche		

Table couleur des selles

Variable	Type	Description
NumCLSLL		
col_sell	Texte	Couleur des selles
periode_colanormsell	Texte	Période de la couleur anormale des selles
Numfiche		

Table couleur urine

Variable	Type	Description
NumCLUR		
col_urin	Texte	Couleur des urines
periode_col_anormurin	Texte	période de la couleur anormal des urines
Numfiche		

Table crise

Variable	Type	Description
NumCRS		
nbr_crise	Numérique	Nombre de crise
periode_crise	Texte	Période de la crise
duree_crise	Texte	Durée de la crise
survenu_crise	Texte	Survenue de la crise
Spasme	Numérique	Présence de spasmes
crie_pleur	Numérique	La personne a –t- elle pleurait ou criait ?
Urine	Numérique	La personne a –t- elle urinait ?
mordr_langue	Numérique	La personne se mordait-elle la langue ?
Hypersaliv	Numérique	La personne hypersalivait elle ?
respir_bruyam	Numérique	La personne respirait-elle bruyamment ?
fontanell_gonfl	Numérique	La fontanelle était elle gonflée ?
cou_tordu_arrier	Numérique	Le cou est-il tordu en arrière ?
corp_raidi_arrier	Numérique	Le corps est-il raidi en arrière ?
jambes_tendu	Numérique	Jambes tendus ?
jambes_plies	Numérique	Jambes pliés ?
bras_tendu	Numérique	Bras tendu ?
bras_plies	Numérique	Bras plié ?
poings_fermes	Numérique	Poings fermés ?

bouche_ferme_crisp	Numérique	Bouche fermé et crispé ?
perdr_connaiss	Numérique	Perdre connaissance ?
duree_perte_conaiss_aprcris	Texte	Durée perte de connaissance
Numfiche		

Table déshydratation

Variable	Type	Description
NumDHDT		
bouch_langu_sech	Numérique	Bouche ou langue sèche ?
yeu_enfonc	Numérique	Yeux enfoncés ?
fontanel_deprim	Numérique	Fontanelle déprimée ?
Numfiche		

Table difficultés à respirer

Variable	Type	Description
NumDRSP		
duree_diff	Texte	Durée des difficultés respiratoires
debut_diff	Texte	Début des difficultés respiratoires
fin_diff	Texte	Fin des difficultés respiratoires
respir_rapid	Numérique	Respiration rapide ?
respir_diff	Numérique	Respiration difficile ?
respir_bruyam_diff	Numérique	Respiration difficile et bruyante ?
respir_sifflan	Numérique	Respiration sifflante ?
ailes_nez_palpitan	Numérique	Ailes ou nez palpitant ?
peau_rentre_cote	Numérique	La peau qui rentre dans les cotes ?
Numfiche		

Table difficultés à uriner

Variable	Type	Description
Num_DURN		
duree_diff_urin	Texte	Durée difficultés urinaires
debut_diff_urin	Texte	Durée difficultés urinaires
fin_diff_urin	Texte	Durée difficultés urinaires
douleur_urinan	Numérique	Douleur en urinant
Numfiche		

Table Mal aux yeux

Variable	Type	Description
NumYEUX		
period_mal_yeu	Texte	Période de la mal des yeux
col_yeu_rouge	Numérique	Yeux rouges ?
col_yeu_jaune	Numérique	Yeux jaunes ?
yeu_larmoyan	Numérique	Yeux larmoyants ?
Numfiche		

Table maux poitrine

Variable	Type	Description
NumPOITRINE		
duree_maupoitrin	Texte	Durée des maux de poitrine
debut_maupoitrin	Texte	Début des maux de poitrine
fin_maupoitrin	Texte	Fin des maux de poitrine
Numfiche		

Table maux tête

Variable	Type	Description
NumMAUX_TETE		
duree_mautete	Texte	Durée des maux de tête
debut_mautete	Texte	Début des maux de tête
fin_mautete	Texte	Fin des maux de tête
Numfiche		

Table maux ventre

Variable	Type	Description
NumVENTRE		
duree_mauventr	Texte	Durée des maux de ventre
debut_mauventr	Texte	Début des maux de ventre
fin_mauventr	Texte	Fin des maux de ventre
Numfiche		

Table Œdèmes

Variable	Type	Description
NumOEDM		
duree_oedem	Texte	Durée des œdèmes
debut_oedem	Texte	Début des œdèmes
fin_oedem	Texte	Fin des œdèmes
localisation_oedem	Texte	Localisation des œdèmes
Numfiche		

Table plaies brulures abces

Variable	Type	Description
NumPBA		
presance_plaies	Numérique	Présence de plaies
infect_pba	Numérique	Plaie infecté ?
presanc_brulur	Numérique	Présence brulures
gonflmen_contena_pus	Numérique	Plaie gonflé contenant du pus ?
Indication	Texte	Indiquer l'endroit
Numfiche		

Table signes neurologiques

Variable	Type	Description
NumSNR		
pert_connaiss	Numérique	Perte de connaissance ?
periode_pert_connais s	Texte	Période perte connaissance ?
duree_pert_connaiss	Texte	Durée perte connaissance ?
Paralysie	Numérique	Présence d'une paralysie ?
precis_paralysie	Texte	Préciser la paralysie
duree_paralysi	Texte	Durée paralysie
Numfiche		

Table toux

Variable	Type	Description
NumTUX		
duree_toux	Texte	Durée des toux
debut_toux	Texte	Début des toux
fin_toux	Texte	Fin des toux
touss_nuit	Numérique	Toussait-elle la nuit ?
cracha_apr_toux	Numérique	Crachait-elle après la toux ?
crach_com_pus	Numérique	Crachat comme du pus ?
crach_com_mouss	Numérique	Crachat comme de la mousse ?
crach_avec_sang	Numérique	Crachat avec du sang ?
crach_nauseabon	Numérique	Crachat nauséabond ?
vomi_apr_toux	Numérique	Vomissait-elle après la toux
perdr_respi_toussan	Numérique	Perdait-elle la respiration en toussant ?
quint_toux	Numérique	Avait-elle des quintes de toux ?
Numfiche		

Table troubles visuels

Variable	Type	Description
NumTRBL_VISUELS	Numérique	Présence de trouble visuels
Numfiche		

Table ventre gonflé

Variable	Type	Description
NumVTG		
duree_gonfl	Texte	Durée des gonflements
debut_gonfl	Texte	Début des gonflements
fin_gonfl	Texte	Fin des gonflements
pratik_ponction	Numérique	Pratique de ponction ?
specifi_formationsanitair	Texte	Spécifier la formation sanitaire
Numfiche		

Table vomissements

Variable	Type	Description
NumVMT		
duree_vomis	Texte	Durée des vomissements
period_vomis	Texte	Période des vomissements
col_vomis	Texte	Couleur des vomissements
vomis_jet	Numérique	Vomissements en jet ?
Numfiche		

Table saignements

Variable	Type	Description
NumSGMT		
localis_saignmen	Texte	Localisation saignements
nbr_saigne	Texte	Nombre de saignements
periode_saign	Texte	Période saignements
Numfiche		

Table signes généraux

Variable	Type	Description
NumSGNR		
demang_purit	Numérique	Démangeaisons et purit ?
maigri_courmaladi	Numérique	La personne a-t-elle maigri au cours maladie ?
maigr_debutmaladi	Numérique	La personne a-t-elle maigri au début maladie ?
arret_mange_coursmaladi	Numérique	La personne a-t-elle arrêté de manger au cours de la maladie ?

col_paum_chang	Numérique	La couleur de la paume a-t-elle changée ?
corps_chang_col	Numérique	La couleur du corps a-t-elle changée ?
langue_pal	Numérique	La personne a-t-elle la langue pale ?
mange_terre	Numérique	La personne mange-t-elle de la terre ?
Constipe	Numérique	La personne est-elle constipée ?
soif_duranmaladi	Numérique	La personne a-t-elle soif durant la maladie ?
mem_symp_prdotr_memperio	Numérique	D'autres personnes présentent-elles le même symptôme ?
village_SGNR	Texte	Préciser le village
remark_sign	Texte	Remarques
Numfiche		

Table décès maternelle

Table décès femme

Variable	Type	Description
NumficheFEMM		
DCFemmeEnceinte	Numérique	Décès d'une femme enceinte ?
DCFemme12a49ans	Numérique	Décès d'une femme âgée entre 12 et 49 ans ?
Numfiche		

Table maternité

Variable	Type	Description
NumMaternite		
DateFinDernGrossesse	Texte	Date de fin dernière grossesse
m_problm_gross_accouch_precen	Texte	Problème lors du dernier accouchement
m_pratik_cesar	Numérique	Pratique césarienne
m_duree_gross	Texte	Durée grossesse
m_mere_mala_duran_gross	Numérique	Mere malade durant grossesse
m_jambes_enflees	Numérique	Jambes enflées ?
m_mains_enflees	Numérique	Mains enflées?
m_visage_enfle	Numérique	Visage enflé ?
m_hypertension	Numérique	Présence d'une hypertension
m_convulsion	Numérique	Présence de convulsion
m_fievre	Numérique	Présence de fièvre
m_saign	Numérique	Présence de saignements
m_saign_cours_grosss	Numérique	Saignements au cours de la grossesse
m_soins	Texte	Soins recus
m_regime	Numérique	Pratique d'un régime alimentaire
m_lequel	Texte	Préciser le regime alimentaire
m_visit_prenatal	Numérique	Visite prénatale
m_lieu	Texte	Lieu de la visite
m_injection_tetanos	Numérique	Injection en tetanus
m_lieu_inject	Texte	Lieu d'injection
m_lieu_accouch	Texte	Lieu d'accouchement

m_localite	Texte	Préciser la localité
m_etabliss	Texte	Préciser l'établissement
m_accouch_diff	Numérique	Accouchements difficiles
m_lesquelles	Texte	Lesquelles
m_deroulmen_accouch	Texte	Déroulement de l'accouchement
m_duree	Texte	Durée de l'accouchement
m_naiss_multipl	Numérique	Naissance multiple
m_tete_premier	Numérique	Téte venue en premier
m_ruptur_12h	Numérique	Rupture en 12 h
m_fievr_24	Numérique	Fiévre durant les 24H
m_placenta	Numérique	Placenta venu normal et en entier
m_momen	Texte	Moment venu placenta
m_femme_saign_lgt	Numérique	La femme a-t-elle saigné longtemps ?
m_temp_saign	Texte	Durée des saignements
precise_prbl	Texte	Préciser les problems
m_col_sang	Texte	Couleur du sang
NumficheFEMM		

Table pertes de sang

Variable	Type	Description
NumPS		
duree_pert	Texte	Durée des pertes de sang
debut_pert	Texte	Début des pertes de sang
fin_pert	Texte	Fin des pertes de sang
freq_saign	Texte	Fréquence des pertes de sang
pert_tr_abondan	Numérique	Pertes de sang abondant ?
col_sang_pert	Texte	Couleur du sang
traitmen_perte	Texte	Traitement reçue
NumficheFEMM		

Table autres saignements

Variable	Type	Description
NumAS		
locali_autr_saign	Texte	Localisation des autres saignements
duree_autr_saign	Texte	Durée des saignements
debut_autr_saign	Texte	Début des saignements
fin_autr_saign	Texte	Fin des saignements
freq_saign_autr	Texte	Fréquence des saignements
sang_autr_saign	Texte	Présence de sang dans les saignements
col_autr_saign	Texte	Couleur des saignements
trait_autr_saign	Texte	Traitements reçues
NumficheFEMM		

Table Régles anormales

Variable	Type	Description
NumRA		
duree_regl	Texte	Durée des règles anormales
debut_regl	Texte	Début des règles anormales
fin_regl	Texte	Fin des règles anormales
regl_irregul_arret	Texte	Règle irrégulière ou arrêtée
duree_plu_longu	Texte	Durée plus longue
regl_abondant	Texte	Règle abondante
col_sang	Texte	Couleur du sang
traitmen_recu_pr_regl	Texte	Traitement reçue
NumficheFEMM		

Table décès enfant et nouveau-né

Table décès enfant

Variable	Type	Description
NumficheEFT		
DCavant5ans	Numérique	Décès d'un enfant avant 5 ans
EnfantSevre	Numérique	Enfant est-il sevré ?
DCnouveaune4sem	Numérique	Décès d'un nouveau-né
mortne	Numérique	Un mort-né
Numfiche		

Table décès d'un nouveau-né

Variable	Type	Description
NumficheNVONE		
CategorieNVONE	Texte	Catégorie de l'enfant
CrierApresNaiss	Texte	Enfant a-t-il crié après sa naissance ?
RespirerApresNaiss	Texte	Enfant a-t-il respiré après sa naissance ?
CorpulenceNVONE	Texte	Corpulence du nouveau-né
GrosseTeteNVONE	Texte	Enfant avait-il une grosse téte ?
UrinerApresNaiss	Texte	Enfant a-t-il urinait après sa naissance ?
DefequerApresNaiss	Texte	Enfant a-t-il déféqué après sa naissance ?
FoetusMacere	Texte	Fœtus macéré ?
MalformationNaiss	Texte	Une malformation a la naissance ?
TeterApresNaiss	Texte	Enfant a-t-il tété après sa naissance ?
ArreterTeter	Texte	Enfant a-t-il arrêté de tété ?
NumficheEFT		
NumMaternite		

Remarques : Si c'est un décès d'un nouveau-né, on pose les mêmes questions que sur la table maternité.