

Université de Montréal

**Analyse statistique de données fonctionnelles à
structures complexes**

par

ADJOGOU ADJOBÓ FOLLY DZIGBODI

Département de mathématiques et de statistique
Faculté des arts et des sciences

Thèse présentée à la Faculté des études supérieures
en vue de l'obtention du grade de
Philosophiæ Doctor (Ph.D.)
en Statistique

01 Mai 2017

© ADJOGOU ADJOBÓ FOLLY DZIGBODI, 2017

Université de Montréal

Faculté des études supérieures

Cette thèse intitulée

**Analyse statistique de données fonctionnelles à
structures complexes**

présentée par

ADJOGOU ADJOBO FOLLY DZIGBODI

a été évaluée par un jury composé des personnes suivantes :

Pierre Duchesne

(président-rapporteur)

Alejandro Murua

(directeur de recherche)

Mylène Bédard

(membre du jury)

Abbas Khalili (Université McGill)

(examineur externe)

Benoît Perron

(représentant du doyen de la FAS)

Thèse acceptée le

29 Septembre 2017

RÉSUMÉ

Les études longitudinales jouent un rôle prépondérant dans des domaines de recherche variés et leur importance ne cesse de prendre de l'ampleur. Les méthodes d'analyse qui leur sont associées sont devenues des outils privilégiés pour l'analyse de l'étude temporelle d'un phénomène donné. On parle de données longitudinales lorsqu'une ou plusieurs variables sont mesurées de manière répétée à plusieurs moments dans le temps sur un ensemble d'individus. Un élément central de ce type de données est que les observations prises sur un même individu ont tendance à être corrélées. Cette caractéristique fondamentale distingue les données longitudinales d'autres types de données en statistique et suscite des méthodologies d'analyse spécifiques. Ce domaine d'analyse a connu une expansion considérable dans les quarante dernières années. L'analyse classique des données longitudinales est basée sur les modèles paramétriques, non-paramétriques et semi-paramétriques. Mais une importante question abondamment traitée dans l'étude des données longitudinales est associée à l'analyse typologique (regroupement en classes) et concerne la détection de groupes (ou classes ou encore trajectoires) homogènes, suggérés par les données, non définis a priori de sorte que les individus dans une même classe tendent à être similaires les uns aux autres dans un certain sens et, ceux dans différentes classes tendent à être non similaires (dissemblables). Dans cette thèse, nous élaborons des modèles de clustering de données longitudinales et contribuons à la littérature de ce domaine statistique en plein essor. En effet, une méthodologie émergente non-paramétrique de traitement des données longitudinales est basée sur l'approche de l'analyse des données fonctionnelles selon laquelle les trajectoires longitudinales sont perçues comme étant un échantillon de fonctions (ou courbes) partiellement observées sur un intervalle de temps sur lequel elles sont souvent supposées lisses. Ainsi, nous proposons dans cette thèse, une revue de la littérature statistique sur l'analyse des données longitudinales et développons deux nouvelles méthodes de partitionnement fonctionnel basées sur des modèles spécifiques. En effet, nous exposons dans le premier volet de la présente thèse une revue succincte de la plupart des modèles typiques d'analyse des données longitudinales, des modèles paramétriques aux modèles non-paramétriques et semi-paramétriques. Nous présentons également les développements récents dans le domaine de l'analyse typologique de

ces données selon les deux plus importantes approches : l'approche non paramétrique et l'approche fondée sur un modèle. Le but ultime de cette revue est de fournir un aperçu concis, varié et très accessible de toutes les méthodes d'analyse des données longitudinales. Dans la première méthodologie proposée dans le cadre de cette thèse, nous utilisons l'approche de l'analyse des données fonctionnelles (ADF) pour développer un modèle très flexible pour l'analyse et le regroupement de tout type de données longitudinales (balancées ou non) qui combine adéquatement et simultanément l'analyse fonctionnelle en composantes principales et le regroupement en classes. La modélisation fonctionnelle repose sur l'espace des coefficients dans la base des splines et le modèle, conçu dans un cadre bayésien, est basé sur un mélange de distributions de Student. Nous proposons également un nouveau critère pour la sélection de modèle en développant une approximation de la log-vraisemblance marginale (MLL). Ce critère se compare favorablement aux critères usuels tels que AIC et BIC. La seconde méthode de regroupement développée dans la présente thèse est une nouvelle procédure d'analyse de données longitudinales qui combine l'approche du partitionnement fonctionnel basé sur un modèle et une double pénalisation de type Lasso pour identifier les classes homogènes ou les individus avec des tendances semblables. Les courbes individuelles sont approximées dans un espace dérivé par une base finie de splines et le nombre optimal de classes est déterminé en pénalisant un mélange de distributions de Student. Les paramètres de contrôle de la pénalité sont définis par la méthode d'échantillonnage par hypercube latin qui assure une exploration plus efficace de l'espace de ces paramètres. Pour l'estimation des paramètres dans les deux méthodes proposées, nous utilisons l'algorithme itératif espérance-maximisation.

Mots clés : Données longitudinales, partitionnement fonctionnel, classification non supervisée, modèles de mélange pour classification, analyse des données fonctionnelles, algorithme EM, statistique bayésienne.

ABSTRACT

Longitudinal studies play a salient role in many and various research areas and their relevance is still increasing. The related methods have become a privileged tool for analyzing the evolution of a given phenomenon across time. Longitudinal data arise when measurements for one or more variables are taken at different points of a temporal axis on individuals involved in the study. A key feature of such type of data is that observations within the same subject may be correlated. That fundamental characteristic makes longitudinal data different from other types of data in statistics and motivates specific methodologies. There has been remarkable developments in that field in the past forty years. Typical analysis of longitudinal data relies on parametric, non-parametric or semi-parametric models. However, an important question widely addressed in the analysis of longitudinal data is related to cluster analysis and concerns the existence of groups or clusters (or homogeneous trajectories), suggested by the data, not defined a priori, such that individuals in a given cluster tend to be similar to each other in some sense, and individuals in different clusters tend to be dissimilar. This thesis aims at contributing to that rapidly expanding field of clustering longitudinal data. Indeed, an emerging non-parametric methodology for modeling longitudinal data is based on the functional data analysis approach in which longitudinal trajectories are viewed as a sample of partially observed functions or curves on some interval where these functions are often assumed to be smooth. We then propose in the present thesis, a succinct review of the most commonly used methods to analyze and cluster longitudinal data and two new model-based functional clustering methods. Indeed, we review most of the typical longitudinal data analysis models ranging from the parametric models to the semi and non parametric ones, as well as the recent developments in longitudinal cluster analysis according to the two main approaches : non-parametric and model-based. The purpose of that review is to provide a concise, broad and readily accessible overview of longitudinal data analysis and clustering methods. In the first method developed in this thesis, we use the functional data analysis approach to propose a very flexible model which combines functional principal components analysis and clustering to deal with any type of longitudinal data, even if the

observations are sparse, irregularly spaced or occur at different time points for each individual. The functional modeling is based on splines and the main data groups are modeled as arising from clusters in the space of spline coefficients. The model, based on a mixture of Student's t-distributions, is embedded into a Bayesian framework in which maximum a posteriori estimators are found with the EM algorithm. We develop an approximation of the marginal log-likelihood (MLL) that allows us to perform an MLL based model selection and that compares favourably with other popular criteria such as AIC and BIC. In the second method, we propose a new time-course or longitudinal data analysis framework that aims at combining functional model-based clustering and the Lasso penalization to identify groups of individuals with similar patterns. An EM algorithm-based approach is used on a functional modeling where the individual curves are approximated into a space spanned by a finite basis of B-splines and the number of clusters is determined by penalizing a mixture of Student's t-distributions with unknown degrees of freedom. The Latin Hypercube Sampling is used to efficiently explore the space of penalization parameters. For both methodologies, the estimation of the parameters is based on the iterative expectation-maximization (EM) algorithm.

Keywords : Longitudinal data, functional clustering, model-based clustering, functional data analysis, EM algorithm, Bayesian framework.

TABLE DES MATIÈRES

Résumé.....	iii
Abstract.....	v
Liste des tableaux	x
Liste des figures	xi
Dédicace.....	xiii
Remerciements	xiv
Chapitre 1. Introduction	1
Bibliographie	6
Chapitre 2. A review of longitudinal data analysis and clustering methods	9
Abstract	9
2.1. Introduction	10
2.2. Analysis methods for longitudinal data	11
2.2.1. Parametric models for analysis of longitudinal data.....	12
2.2.1.1. Linear models for longitudinal data	12
2.2.1.2. Generalized linear models for longitudinal data.....	14
2.2.1.3. Non-linear models for longitudinal data	16
2.2.2. Non-parametric and semi-parametric models for longitudinal data analysis.....	16
2.2.2.1. Kernel-based non parametric methods.....	17
2.2.2.2. Splines-based non parametric methods	19
2.2.2.3. Semi-parametric methods	20
2.2.2.4. The estimation of the covariance in longitudinal data analysis	21
2.3. Clustering methods for longitudinal data	22

2.3.1. Non-parametric clustering methods.....	23
2.3.2. Model-based clustering methods.....	24
2.3.3. Recent developments in longitudinal cluster analysis.....	27
2.3.3.1. Longitudinal cluster analysis in gene expression data	27
2.3.3.2. Clustering methods using the functional data analysis approach	35
2.3.3.3. Software-implemented clustering methods.....	39
2.3.3.4. Some comparison studies among longitudinal data clustering methods	40
2.4. Conclusions and discussion.....	40
Bibliographie	43
Chapitre 3. Functional model-based clustering for longitudinal data.....	52
abstract.....	52
3.1. Introduction	53
3.2. Functional data analysis and clustering.....	55
3.2.1. The model.....	55
3.2.2. Extension to multiple dimensions.....	58
3.2.3. Parameter estimation	60
3.2.4. Model selection	63
3.3. Experiments with simulated and real data.....	64
3.3.1. Simulation study for the one-dimensional model	64
3.3.2. Simulation study for the two-dimensional model	67
3.3.3. Comparison study with real datasets	68
3.3.3.1. The Rats data	68
3.3.3.2. Growth data	69
3.3.3.3. ECG data.....	71
3.3.3.4. The yeast cycle data	71
3.4. Application to the PRRS viremia dataset.....	73
3.5. Conclusions and discussion.....	84
Bibliographie	86
Chapitre 4. Functional model-based clustering with Lasso-type penalization for longitudinal data	90

Abstract	90
4.1. Introduction	91
4.2. Functional model-based clustering with Lasso penalization	93
4.2.1. Fundamentals of the model	93
4.2.2. The penalized log-likelihood	94
4.2.3. Choosing the penalty parameters by cross-validation	95
4.2.4. The Bayesian Lasso functional clustering model	97
4.2.5. EM algorithm : Expectation and Maximization steps	98
4.3. Model selection	100
4.4. Simulation study	103
4.5. Chronic obstructive pulmonary disease	108
4.6. Conclusion	110
Bibliographie	111
Chapitre 5. Conclusion	114
Bibliographie	117
Annexe A. Some analytical details on partitioning and EM steps	A-i
A.1. Partitioning of an incomplete multivariate Gaussian data	A-i
A.2. Analytical developments for EM expectation step :	A-i
A.3. The updating EM equations for the mixed-effects model for PRRSV	A-v
Annexe B. An illustration of the code for the functional model-based clustering analysis	B-i
Bibliographie	B-i

LISTE DES TABLEAUX

3.1	Partition matrix for Growth data	70
3.2	Illustration of the partition matrices for the computation of the kappa coefficient	77
4.1	Values of postulated number of clusters according to G	104

LISTE DES FIGURES

3.1	Yeast cycle data. The observed (top row) and estimated (bottom row) variance-covariance matrices by cluster. The clusters are arranged from left to right, starting with Cluster 1.	58
3.2	Yeast cycle data. The overall observed (left) and estimated (right) variance-covariance matrices.	59
3.3	One-dimensional model. Boxplots of the ARI scores for the models selected by MLL (top), AIC (middle) and BIC (bottom). The light grey boxes correspond to $G = 3$, whereas the darker grey ones correspond to $G = 9$. The middle grey boxes correspond to $G = 6$. N stands for sample size. The boxplots are organized first by N and then by G	66
3.4	One-dimensional model. Proportion of times each criteria chose a particular number of clusters.	67
3.5	Two-dimensional model. Box plots of the ARI scores for the models selected by MLL (left), AIC (middle) and BIC (right).	68
3.6	Two-dimensional model. Example of a dataset with six clusters.	69
3.7	Original (left) and Predicted (right) curves for the Rats dataset.	70
3.8	Original (left) and Predicted (left) curves for the Growth dataset.	71
3.9	Model selection results for yeast cell cycle data.	72
3.10	Yeast cycle data. Observed and model-estimated mean curves for the four clusters yielded by the MLL and BIC criteria.	73
3.11	Yeast cycle data. Observed and model-estimated mean curves for the five clusters found by Cho et al. [7].	74
3.12	Yeast data. Overall mean curves (left) and distribution of the ν_i associated with the error term distribution of the data.	74
3.13	Illustration of the difference at days 40 and 42.	75
3.14	The M_0 $3G$ -partition mean curves.	78

3.15	The M_1 14G-partition mean curves by <i>Wur</i> category	79
3.16	The M_1 3G-partition mean curves by <i>Wur</i> category	79
3.17	The M_2 3G-partition mean curves by <i>Experiment</i> category	81
3.18	The M_{12} 3G-partition mean curves by <i>Wur</i> and <i>Experiment</i> for cluster 1	81
3.19	The M_{12} 3G-partition mean curves by <i>Wur</i> and <i>Experiment</i> for cluster 2	82
3.20	The M_{12} 3G-partition mean curves by <i>Wur</i> and <i>Experiment</i> for cluster 3	82
3.21	MLL Comparison for model selection	83
4.1	Comparison of the three criteria for model selection	106
4.2	Similarity ratio and model performance	107
4.3	Influence of postulated G	107
4.4	Comparison of True and estimated number of clusters	108
4.5	Cluster mean curves for the three partitions found by the Bayesian Lasso clustering model. The first row displays the clusters in the two-dimensional space of functional principal components (FPC) scores. The bottom row shows the cluster mean-curves	109

DÉDICACE

Je dédie ce travail à mon épouse **Ange Christelle**
et mes enfants **Ékoué Émidèl** et **Ayéle Asséna**.

REMERCIEMENTS

« Le temps met tout en lumière. » (Thalès)

En premier lieu et par dessus toute considération, je veux dire Merci à Dieu le Père Tout-puissant qui a permis la réalisation (et enfin, la fin) de cette thèse. À toi, Père Éternel : « Mes faibles mots te sanctifient et mon esprit te magnifie. Seigneur, je te dis Merci » (extrait du chant *Dans ton sanctuaire* de GAEL).

Plusieurs personnes ont vécu avec moi, ces longues années de doctorat.

Sur le plan familial et du haut de cette liste, je remercie ma charmante et douce épouse : « Chérie, Merci de tout coeur pour l'amour et le soutien. Merci également pour ces bonheurs immenses qui ont jalonné cette thèse, notamment ces deux magnifiques bébés ». Ensuite, je rends un vibrant hommage à mes parents, à ma très chère maman et à mon père pour leurs prières, conseils et rigueur. « C'est vous qui avez semé cette graine, l'avez arrosée et entretenue pendant toutes ces années ». Spécialement pour toi Tonton, j'ajoute ces vers de Alfred de Vigny : « A voir ce que l'on fut sur terre et ce qu'on laisse, seul le silence est grand, tout le reste est faiblesse ». Je veux également dire un gros gros Merci à ma mère Tanty et à mes soeurs chéries Amélé, Akwavi et Enyonam : « Oui, la distance nous joue des tours mais l'amour survit à tout ». Merci également à ma merveilleuse et bienveillante belle-famille au Canada et en Côte d'Ivoire.

Sur le plan académique et professionnel, j'ai beaucoup de reconnaissance et d'admiration à témoigner à mon directeur de recherche. Je n'en rajoute pas et peut-être même n'en dis-je pas assez en affirmant que je n'aurais pu rêver un meilleur directeur de thèse. « Disponible, tu as su valoriser ce qui est bien, et transformer ce qui l'est moins. Merci Alejandro pour tout ce que tu m'a appris, pour ton sens sourcilleux, ta générosité et ta sensibilité à mon contexte personnel. Ta compétence, ta rigueur scientifique et ta clairvoyance m'ont beaucoup édifié. J'aime à penser qu'au fil des années nous sommes devenus de bons amis ».

Je remercie également tous les professeurs et tout le valeureux personnel du Département de mathématique et statistique de l'Université de Montréal. Je ne saurais oublier tous mes collègues de la Direction de la modélisation des systèmes de transport du Ministère des Transports. Merci à tous pour vos encouragements notamment dans le sprint final. Et enfin, Merci à Amhos. Merci à tous mes amis. Merci à tous mes étudiants du DMS.

Chapitre 1

INTRODUCTION

Dans plusieurs domaines de recherche, notamment dans les sciences sociales, les études longitudinales sont devenues un outil essentiel pour *analyser l'évolution d'un phénomène donné dans le temps*, qui peut revêtir un caractère plus important que la simple connaissance du moment d'apparition d'un tel phénomène. Elles sont constituées de mesures répétées d'une ou de plusieurs variables, prises sur un ensemble d'individus, à différents points d'un axe temporel. Une caractéristique fondamentale de ce type de données est que les observations recueillies sur un même sujet tendent à être corrélées. Il s'agit de la *corrélation intra-sujet*. En effet, la connaissance de la valeur observée de la variable réponse à une date donnée fournit de l'information sur sa valeur probable à date future (voir Fitzmaurice and Ravichandran [9]). La corrélation entre les mesures répétées va à l'encontre de l'hypothèse fondamentale d'indépendance qui constitue la pierre angulaire de plusieurs techniques standard en statistique (test t, régression linéaire, anova). Cette particularité des données longitudinales est décrite dans plusieurs articles et ouvrages traitant du sujet et notamment dans Fitzmaurice et al. [8] qui présente également les différentes sources et nature de la corrélation dans ce type de données ainsi que les conséquences potentielles lorsque celle-ci n'est pas prise en compte dans l'analyse statistique. De plus, les données longitudinales diffèrent des autres types de données en statistique telles que les données multivariées, les études transversales, les séries temporelles, et leur analyse requiert par conséquent des méthodologies spécifiques.

La méthodologie statistique d'analyse des données longitudinales a connu au cours des trente dernières années un développement considérable, qui a été facilité par l'émergence de technologies nouvelles qui favorisent les applications numériques sur des ordinateurs de plus en plus puissants. Les méthodes les plus couramment utilisées dans l'analyse des données longitudinales sont basées sur des modèles paramétriques tels que les modèles linéaires à effets mixtes proposés par Laird and Ware [14] pour l'étude des variables réponses continues observées au fil du temps. Ces méthodes reposent sur une décomposition explicite de

la variation dans les données en variabilité inter et intra-sujet (Verbeke and Molenberghs [28]). La classe des modèles paramétriques comprend également les modèles marginaux et les modèles linéaires généralisés à effets mixtes (McCullagh and Nelder [16]) qui sont deux généralisations importantes des modèles linéaires aux cas où la variable réponse est discrète, ainsi que les modèles non linéaires à effets mixtes. Plusieurs exemples d'applications empiriques telles que Zeger and Diggle [32], Brumback and Rice [2], Lin and Ying [15] et Diggle et al. [5] montrent que les hypothèses paramétriques ne sont pas toujours appropriées pour modéliser la dynamique temporelle entre une variable d'intérêt et des variables explicatives dans une étude longitudinale. Ainsi, les méthodes non-paramétriques et semi-paramétriques ont émergé dans la littérature statistique afin de proposer des formes fonctionnelles plus flexibles dans l'analyse des données longitudinales. Il s'agit essentiellement d'adapter les méthodes par noyaux (Wand and Jones [30], Fan and Gijbels [6]) et les méthodes de lissage par splines (Wahba [29], Green and Silverman [11], Stone et al. [24]) qui ont été élaborées pour l'étude de données indépendantes, aux spécificités des données longitudinales notamment la corrélation intra-sujet entre les mesures répétées dans le temps.

Une autre méthodologie non paramétrique de la modélisation des données longitudinales est fournie par l'approche de l'analyse de données fonctionnelles (ADF) selon laquelle les mesures répétées recueillies auprès des individus sont considérées comme des portions de courbes (Ramsay and Silverman [21]). En effet, les trajectoires longitudinales sont perçues comme étant un échantillon de fonctions (ou courbes) partiellement observées sur un intervalle de temps sur lequel elles sont souvent supposées lisses. En d'autres termes, les données longitudinales sont considérées comme des données fonctionnelles irrégulièrement observées. L'objectif de cette approche est donc d'utiliser les outils de l'analyse de données fonctionnelles pour prédire chaque trajectoire individuelle à partir des mesures effectuées, en tenant compte des mesures provenant de tous les autres individus impliqués dans l'étude longitudinale.

Depuis quelques années, des méthodes d'extension des techniques de l'analyse de données fonctionnelles aux données longitudinales font l'objet d'un champ de recherche en pleine expansion. L'émergence de ce nouveau paradigme repose sur l'idée fondamentale que les méthodes de l'ADF peuvent constituer un outil remarquable pour optimiser l'analyse des données longitudinales (Zhao et al. [33], Rice [22]), notamment en ce qui concerne le regroupement en classes (partitionnement ou classification non supervisée ou encore clustering en anglais) tel qu'illustré dans Ullah and Finch [27] qui présente les récentes applications majeures de l'approche par l'ADF. Le regroupement en classes a toujours été une méthode

d'analyse de données privilégiée dans le cas des données longitudinales et porte sur l'identification de groupes, non définis a priori et suggérés par les données, de sorte que les individus dans un même groupe tendent à être similaires dans un certain sens (défini par le critère de regroupement) et ceux dans différents groupes tendent à être non similaires. Par exemple, le partitionnement des données d'expression génétique est un enjeu important en bioinformatique car la compréhension des gènes qui se comportent de façon similaire conduit à la découverte d'importantes informations biologiques (McNicholas and Subedi [18]). Les travaux qui s'inscrivent dans le contexte d'extension de l'ADF utilisent essentiellement l'analyse fonctionnelle en composantes principales (AFCP) qui a émergé comme un outil majeur en ADF et qui permet de réduire la dimensionalité d'un ensemble de données fonctionnelles en identifiant les modes de variation les plus significatifs. Ramsay and Silverman [21] propose une excellente présentation de la théorie de l'AFCP, des techniques de calcul des composantes principales telles que la discrétisation des fonctions observées et l'expansion dans une base de fonctions, ainsi qu'une étude comparative de ces différentes approches de l'AFCP.

L'objectif de cette thèse est de contribuer à l'élaboration de nouveaux modèles plus flexibles, basés sur l'approche de l'analyse des données fonctionnelles pour l'analyse et le regroupement en classes des données longitudinales ou des courbes d'évolution corrélées. La théorie classique de l'ADF s'intéresse aux données de dimension infinie telles que les courbes ou les images. Elle est donc essentiellement appropriée pour l'étude des données longitudinales *balançées* c'est à dire des mesures prises à intervalles réguliers (grille temporelle uniformément graduée), de sorte que tous les sujets sont évalués aux mêmes périodes de temps et admettent donc le même nombre de mesures. Mais dans la réalité, la plupart des études longitudinales résultent très souvent en des données *non balançées* (mesures irrégulières, prises à des périodes de temps assez différentes pour les individus, et de plus, tous les sujets peuvent ne pas avoir le même nombre de mesures).

Nous présentons dans le cadre de cette thèse, deux nouveaux modèles de classification non-supervisée de données longitudinales basés sur l'approche de l'analyse des données fonctionnelles, qui tiennent compte de la forme générale des variables-trajectoires et qui constituent des contributions notoires dans la panoplie des méthodes d'analyse disponibles dans ce domaine. Ces modèles font l'objet d'articles scientifiques et se démarquent non seulement par leur flexibilité, la pertinence et l'originalité de leurs hypothèses et lois a priori, mais aussi et surtout par le fait qu'ils sont conçus pour convenir à tous les types de données longitudinales, aussi bien balançées que non balançées. Par exemple, la première méthodologie proposée réalise le partitionnement de données longitudinales à partir d'une seule variable

(1D) ou de deux variables conjointement (2D) mais reste facilement extensible au regroupement en classes basé sur 3 ou plus de 3 variables avec prise en compte possible d'effets fixes. Par ailleurs, une contribution majeure dans la deuxième méthodologie proposée dans cette thèse est cette idée d'utiliser une pénalisation du type Lasso comme loi a priori, dans le sens qu'au lieu de chercher le paramètre optimal Lasso, il est incorporé directement comme paramètre du modèle et il faut impérativement estimer la constante de normalisation qui lui est associée. Nous présentons également dans cette thèse, une revue assez concise de la littérature existante sur les différents modèles d'analyse et de partitionnement des données longitudinales.

Le premier chapitre non introductif de cette thèse présente un aperçu général des différentes méthodes de traitement des données longitudinales selon une perspective historico-évolutive, en examinant les approches paramétrique, semi-paramétrique et non-paramétrique. Une emphase particulière est mise sur les méthodes et procédures de regroupement en classes (modèles de clustering) qui constituent une question importante et largement discutée dans la littérature sur les données longitudinales. En effet, depuis quelques décennies, plusieurs travaux méthodologiques ont porté sur l'extension aux données longitudinales, des différentes méthodes de partitionnement bien développées en matière de données indépendantes et qui ne sont pas adaptées au traitement des mesures répétées. La synthèse proposée dans ce chapitre se démarque d'autres études similaires existantes (Fitzmaurice et al. [7], Gibbons et al. [10], Jacques and Preda [12], Bouveyron and Brunet [1]) par sa diversité (différentes méthodologies selon différents angles de traitement des données longitudinales) et son intégration des méthodes et algorithmes de partitionnement plus récents.

Dans le deuxième chapitre de cette thèse, nous présentons le modèle flexible développé pour l'analyse et le partitionnement de données longitudinales (balancées ou non). Le modèle combine l'analyse fonctionnelle en composantes principales et le regroupement en classes, qui repose sur l'espace des coefficients dans la base des splines et un modèle de mélange de distributions de Student de degrés de liberté inconnus. Nous développons une approximation de la log-vraisemblance marginale (MLL) pour la sélection de modèles qui se compare favorablement aux critères usuels (AIC et BIC). Nous considérons également une extension du modèle aux courbes multidimensionnelles. Des études de simulations et de comparaison avec d'autres modèles du genre, ainsi que des applications sur des données réelles ont été menées pour évaluer la pertinence et la performance du modèle. En effet, la méthodologie a été appliquée sur plusieurs jeux de données assez connus tels que les données de rats publiées dans Crowder and Hand [4] et étudiées dans McNicholas and Murphy [17] dans un contexte longitudinal, les données de croissance provenant des études de croissance de

Berkeley (Tuddenham and Snyder [26]), les données d'activité électrique cardiaque étudiées dans Olszewski [19], l'ensemble de données génétiques sur le cycle cellulaire de Cho et al. [3] et les données sur le virus du syndrome respiratoire et de reproduction chez les porcs (voir Rowland et al. [23]). James and Sugar [13] avaient proposé un modèle similaire mais notre modèle présente l'avantage d'être plus précis et exact par rapport à la structure imposée sur les courbes individuelles et la procédure pour identifier les groupes suggérés par les données. L'estimation des paramètres est effectuée à partir de l'algorithme EM, de plus en plus utilisé et modifié en analyse de données longitudinales.

Dans le troisième chapitre, nous présentons une nouvelle procédure de partitionnement fonctionnel (*functional clustering*) pour l'analyse des données longitudinales. L'idée originale du modèle proposé est inspirée des récents travaux dans le domaine de la sélection de variables pour le partitionnement de données à très grande dimension dont une revue est présentée par Bouveyron and Brunet [1], dans laquelle une emphase particulière est mise sur le critère de pénalisation dans la classification non supervisée. Il s'agit en effet d'introduire, à l'instar de Pan and Shen [20] et Wang and Zhou [31], des termes de pénalité du type L_1 ou L_∞ dans la fonction de log-vraisemblance. La méthode que nous proposons diffère de celles existantes par son approche. Au lieu de simultanément partitionner les données et en réduire la dimensionalité en déterminant les variables les plus pertinentes pour le processus de regroupement, notre procédure de partitionnement basée sur un modèle utilise l'approche fonctionnelle et une pénalisation double du type Lasso (Tibshirani [25]) pour simultanément déterminer la dimension appropriée de la base finie de fonctions (réduction de la dimension) et le nombre approprié de groupes homogènes (partitionnement). La performance et l'utilité de la procédure sont démontrées par la simulation et l'application sur des données réelles, notamment les données d'exposition des rats à la fumée de tabac dans le cadre d'études cliniques.

Bibliographie

- [1] Bouveyron, C. and C. Brunet (2014). Model-based clustering of high-dimensional data : A review. *Computational Statistics and Data Analysis* 71, 52–78.
- [2] Brumback, B. and J. Rice (1998). Smoothing spline models for the analysis of nested and crossed samples of curves. *Journal of the American Statistical Association* 93, 961–976.
- [3] Cho, R. J., M. J. Campbell, E. A. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T. G. Wolfsberg, A. E. Gabrielian, D. Landsman, D. J. Lockhart, and R. W. Davis (1998). A genome-wide transcriptional analysis of the mitotic cell cycle. *Molecular Cell* 2, 65–73.
- [4] Crowder, M. J. and D. J. Hand (1990). *Analysis of Repeated Measures*. London : Chapman and Hall.
- [5] Diggle, P. J., P. J. Heagerty, K.-Y. Liang, and S. L. Zeger (2002). *Analysis of Longitudinal Data, 2nd Ed.* Oxford : Oxford University Press.
- [6] Fan, J. and I. Gijbels (1996). *Local Polynomial Modelling and Its Applications*. London, Chapman Hall.
- [7] Fitzmaurice, G., M. Davidian, G. Verbeke, and G. Molenberghs (2008). *Longitudinal Data Analysis*. Chapman & Hall CRC Handbooks of Modern Statistical Methods.
- [8] Fitzmaurice, G. M., N. M. Laird, and J. H. Ware (2012). *Applied Longitudinal Analysis*. John Wiley and Sons, Second edition.
- [9] Fitzmaurice, G. M. and C. Ravichandran (2008). A primer in longitudinal data analysis. *Circulation* 118, 2005–2010.
- [10] Gibbons, R., D. Hedeker, and S. DuToit (2010). Advances in analysis of longitudinal data. *Annual Review of Clinical Psychology* 6, 79–107.
- [11] Green, P. J. and B. Silverman (1994). *Nonparametric Regression and Generalized Linear Models A Roughness Penalty Approach*. London, Chapman Hall.
- [12] Jacques, J. and C. Preda (2014). Functional data clustering : a survey. *Advances in Data Analysis and Classification, Springer Verlag* 8(3).
- [13] James, G. and C. A. Sugar (2003). Clustering for sparsely sampled functional data. *Journal of the American Statistical Association* 98, 397–408.

- [14] Laird, N. M. and J. H. Ware (1982). Random-effects models for longitudinal data. *Biometrics* 38, 963–974.
- [15] Lin, D. and Z. Ying (2001). Semiparametric and nonparametric regression analysis of longitudinal data. *Journal of the American Statistical Association* 96, 103–126.
- [16] McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models, 2nd Edition*. London : Chapman and Hall.
- [17] McNicholas, P. D. and T. B. Murphy (2010). Model-based clustering of longitudinal data. *Canadian Journal of Statistics* 38, 153–168.
- [18] McNicholas, P. D. and S. Subedi (2012). Clustering gene expression time course data using mixtures of multivariate t-distributions. *Journal of Statistical Planning and Inference* 142, 1114–1127.
- [19] Olszewski, R. (2001). Generalized feature extraction for structural pattern recognition in time-series data. Ph.D. thesis, Carnegie Mellon University.
- [20] Pan, W. and X. Shen (2007). Penalized model-based clustering with application to variable selection. *Journal of Machine Learning Research* 8, 1145–1164.
- [21] Ramsay, J. and B. Silverman (2005). *Functional Data Analysis*. New York : Springer.
- [22] Rice, J. (2004). Functional and longitudinal data analysis : Perspectives on smoothing. *Statistica Sinica*.
- [23] Rowland, R., J. Lunney, and J. Dekkers (2012). Control of porcine reproductive and respiratory syndrome (prrs) through genetic improvements in disease resistance and tolerance. *Front. in Gen.* 3 :260.
- [24] Stone, C., M. Hansen, C. Kooperberg, and Y. K. Truong (1997). Polynomial splines and their tensor products in extended linear modeling (with discussion). *Annals of Statistics* 25, 1371–1470.
- [25] Tibshirani, R. (2011). Regression shrinkage and selection via the lasso : a retrospective. *Journal of the Royal Statistical Society Series B* 73, 273–282.
- [26] Tuddenham, R. and M. Snyder (1954). Physical growth of california boys and girls from birth to eighteen years. *Universities of California Public Child Development* 1, 188–364.
- [27] Ullah, S. and C. F. Finch (2013). Applications of functional data analysis : A systematic review. *BMC Medical Research Methodology* 1471-2288, 13–43.
- [28] Verbeke, G. and G. Molenberghs (2000). *Linear Mixed Models for Longitudinal Data*. Springer series in statistics. New York.
- [29] Wahba, G. (1990). *Spline Models for Observational Data*. Philadelphia : CBMS-NSF Regional Conference Series, SIAM.
- [30] Wand, M. P. and M. C. Jones (1995). *Kernel Smoothing*. London : Monographs on Statistics and Applied Probability, Chapman & Hall.

- [31] Wang, S. and J. Zhou (2008). Variable selection for model-based high dimensional clustering and its application to microarray data. *Biometrics* 64, 440–448.
- [32] Zeger, S. L. and P. J. Diggle (1994). Semiparametric models for longitudinal data with application to cd4 cell numbers in hiv seroconverters. *Biometrics* 50, 689–699.
- [33] Zhao, X., J. Marron, and M. Wells (2004). The functional data analysis view of longitudinal data. *Statistica Sinica* 14, 789–808.

Chapitre 2

A REVIEW OF LONGITUDINAL DATA ANALYSIS AND CLUSTERING METHODS

ABSTRACT

We review the recent developments in longitudinal cluster analysis according to the two main approaches, non-parametric and model-based. We deemed it useful and relevant to present at first, most of the typical longitudinal data analysis methods ranging from the parametric models to the semi and non parametric ones. Then, the clustering methods are discussed. The main purpose of this review is to provide a concise, broad and readily accessible overview of the most important and available methods for analyzing and clustering longitudinal data.

Key words : Longitudinal data, model-based clustering, sparse longitudinal data, functional data analysis, gene expression, mixture student.

2.1. INTRODUCTION

Longitudinal studies have become an essential tool for studying the evolution in time of a given phenomena as they play a salient role in many various research areas. Longitudinal data arises when one or more outcomes are measured at a sequence of observation times on multiple subjects (Diggle et al. [23]). For each subject, the measurements are taken at different times during a certain period leading to a sequence of observations. That fundamental characteristic of possible correlation between observations coming from the same subject (*within-subject correlation*) makes longitudinal data different from other types of data in statistics (such as multivariate data, cross-sectional data, time series data) and calls for specific methodologies. For example, longitudinal data differ from classical time series data because they consist of a large number of independent trajectories that are sparsely and potentially irregularly sampled over time, rather than a few random processes that are uniformly sampled over time (Heggseth [51]). Similarly, historical development of ideas related to longitudinal studies and their advantages over cross-sectional studies are presented in many books or papers such as Rajulton [100]; Hedeker and Gibbons [50] and Rindfleisch et al. [105].

The statistical methodology for the analysis of longitudinal data is essentially based around three axes : parametric models, non-parametric models and semi-parametric models. In the past forty years, that statistical methodology has evolved remarkably, due to increasingly sophisticated technologies which allow numerical applications on high performance machines. Therefore, several new avenues have been explored. For example, the works of Zhao et al. [134] and Rice [103] have shown that longitudinal data can be viewed as a type of functional data. As a consequence, a non-parametric methodology for modeling longitudinal data and based on the functional data analysis (FDA) approach has emerged. In the FDA approach pioneered by Ramsay and Silverman [101], longitudinal trajectories are viewed as a sample of partially observed functions or curves on some interval where these functions are often assumed to be smooth.

In this review, we are particularly interested in an important question widely addressed in the analysis of longitudinal data : the issue of cluster analysis (Hennig et al. [52]; Bruckers [8]). It concerns the existence of groups or clusters (or homogeneous trajectories), suggested by the data, not defined a priori, such that individuals in a given cluster tend to be similar to each other in some sense, and individuals in different clusters tend to be dissimilar. The extension of cluster analysis to longitudinal data has been the focus of a lot of methodological work. These methods range from heuristic approaches such as k-means (Hartigan and Wong [47]; Tarpey [117]; Genolini and Falissard [41]), distances or dissimilarities-based

algorithms (Komarek and Komarkova [65]; Hennig et al. [52]) to model-based procedures (Fraley and Raftery [33]; Fraley and Raftery [35]; James and Sugar [60]). Most methods can be categorized into one of two approaches : nonparametric and model-based methods.

The first section of the this paper, along the same lines as reviews such as Fitzmaurice et al. [31] and Gibbons et al. [44], highlights the longitudinal data analysis methods ranging from the parametric models to the semi and non parametric models. We present herein the most notable developments with an emphasis on the main characteristics of each model. The second section is mainly dedicated to the review of the two clustering categories including methods based on the functional data analysis approach.

2.2. ANALYSIS METHODS FOR LONGITUDINAL DATA

The most popular and extensively used methods in the analysis of longitudinal data are based on the parametric models which often contain random effects, such as the *linear mixed-effects model* proposed by Laird and Ware [67] for continuous responses. The parametric models also include marginal models, generalized linear mixed models for discrete responses and non-linear mixed-effects models. Many examples of empirical applications such as Brumback and Rice [9], Zeger and Diggle [131], Lin and Ying [72] and Diggle et al. [23] demonstrate that parametric assumptions are not always appropriate to modelize the temporal dynamic between a response variable and covariates in longitudinal studies. Hence, non-parametric and semi-parametric methods have emerged in the statistical literature in order to propose more flexible functional forms in the analysis of longitudinal data. This mainly relates to extending and adapting kernel methods (Wand and Jones [124]; Fan and Gijbels [29]) and smoothing splines methods (Green and Silverman [45]; Wahba [122]; Stone et al. [114]), originally developed for independent data, to the particularities of longitudinal data, especially the within-subject correlation among repeated measures over time. Another non parametric methodology in modeling longitudinal data is provided by the functional data analysis (FDA) approach in which the sequence of measurements collected for each individual are considered as portions of curves. The observed measurements correspond to values of a random trajectory corrupted by measurement error. The objective is to use FDA tools to predict individual trajectories from the measurements made for a subject, borrowing strength from the entire sample of subjects (Fitzmaurice et al. [31, chap. 10]). The research associated with the extension of FDA methodology to the analysis of longitudinal data, especially the use of functional principal components analysis (FPCA), is an area in great expansion.

2.2.1. Parametric models for analysis of longitudinal data

Parametric models, which often contain random effects, represent an analysis approach commonly used by statisticians. There is a large variety of such models, depending on the type of the response variable or the objective of the statistical analysis. Amongst others, there are linear models, generalized linear models and non linear models.

2.2.1.1. Linear models for longitudinal data

Linear parametric models are essentially based on normality assumptions and are useful for continuous longitudinal data. Introduced by Laird and Ware [67], the linear mixed-effects model is probably the most popular method in this class of models. In general, a linear mixed-effects model is specified as (Verbeke and Molenberghs [121]) :

$$\begin{cases} \mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i \\ \mathbf{b}_i \sim \mathcal{N}_q(\mathbf{0}, \mathbf{D}) \quad ; \quad \boldsymbol{\epsilon}_i \sim \mathcal{N}_{n_i}(\mathbf{0}, \boldsymbol{\Sigma}_i) \\ \mathbf{b}_1, \dots, \mathbf{b}_N, \boldsymbol{\epsilon}_1, \dots, \boldsymbol{\epsilon}_N \quad \text{are independent} \end{cases} \quad (2.2.1)$$

where \mathbf{Y}_i is the n_i -dimensional vector of measurements of the subject i with $1 \leq i \leq N$; and N being the total number of subjects in the longitudinal study. \mathbf{X}_i and \mathbf{Z}_i are respectively, the covariates matrices of dimensions (n_i, p) and (n_i, q) . $\boldsymbol{\beta}$ is a p -dimensional vector representing the fixed effects; \mathbf{b}_i is a q -dimensional vector representing the random effects; $\boldsymbol{\epsilon}_i$ is the n_i -dimensional vector corresponding to the errors. The matrices \mathbf{D} and $\boldsymbol{\Sigma}_i$ represent the variance-covariance matrices of random effects and errors, respectively.

Conditionally to the random effects \mathbf{b}_i , \mathbf{Y}_i has a multivariate normal distribution of mean $\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i$ and variance-covariance matrix $\boldsymbol{\Sigma}_i$. Unless the model is estimated using a Bayesian approach (Gelman et al. [39]), the parameter estimation and the inference are based on the marginal distribution of \mathbf{Y}_i which is provided by :

$$p(\mathbf{y}_i) = \int p(\mathbf{y}_i|\mathbf{b}_i)p(\mathbf{b}_i)d\mathbf{b}_i.$$

Given the Gaussian distribution of the random effects \mathbf{b}_i , it is demonstrated that the response vector \mathbf{Y}_i of each individual follow a multivariate normal distribution with mean $\mathbf{X}_i\boldsymbol{\beta}$ and variance-covariance matrix $\mathbf{V}_i = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^\top + \boldsymbol{\Sigma}_i$.

Many models are proposed in the statistical literature depending on the choice of the covariance structure of the errors $\boldsymbol{\Sigma}_i$. Most models are special cases of the general model proposed by Diggle et al. [23]. These authors propose the general linear model with the assumption that the covariance structure of the sequence of measures collected on each individual can be specified by a certain number of unknown parameters represented by the vector $\boldsymbol{\alpha}$. Indeed, if $\mathbf{y}_i = (y_{i1}, \dots, y_{in_i})$ is the n_i -dimensional vector of measurements of the individual i and the

vector $\mathbf{t}_i = (t_{i1}, \dots, t_{in_i})$ represents the measurement time points, then the \mathbf{y}_i are assumed to be realizations of independent Gaussian vectors $\mathbf{Y}_i \sim \mathcal{N}_{n_i}(\mathbf{X}_i\boldsymbol{\beta}, V_i(\mathbf{t}_i, \boldsymbol{\alpha}))$. The originality of the presentation in Diggle et al. [23] relies on two important points. First, they make an explicit distinction between the mean and covariance structures by stipulating $\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i$ with $\boldsymbol{\varepsilon}_i \sim \mathcal{N}_{n_i}(\mathbf{0}, V_i(\mathbf{t}_i, \boldsymbol{\alpha}))$. Second, they propose an additive formulation of the different sources of random variation in longitudinal data which are : *random effects*, *within-subject variability* and *measurement errors*. Formally, the vector $\boldsymbol{\varepsilon}_i$ and the term ε_{ij} associated to the j^{th} measurement of individual i are defined as follows :

$$\begin{cases} \boldsymbol{\varepsilon}_i = \mathbf{Z}_i\mathbf{b}_i + W_i(\mathbf{t}_i) + \boldsymbol{\varepsilon}_i \\ \varepsilon_{ij} = \mathbf{Z}_{i(j,\cdot)}^T\mathbf{b}_i + W_i(t_{ij}) + \boldsymbol{\varepsilon}_{ij}; \quad i = 1, \dots, N; \quad j = 1, \dots, n_i. \end{cases} \quad (2.2.2)$$

In this decomposition, the \mathbf{b}_i represent the random effects and are a set of N q -dimensional independent Gaussian vectors with mean vector zero and covariance matrix \mathbf{D} . The $\mathbf{Z}_{i(j,\cdot)}$ are q -dimensional vectors of explanatory variables attached to individual measurements and represent the j^{th} row of the (n_i, q) -dimensional matrix \mathbf{Z}_i . The terms $\{W_i(t_{ij})\}$ represent the within-subject serial correlation and are N independent realizations of a Gaussian stationary process of mean zero and variance σ^2 with correlation function $\rho(u)$. The $\boldsymbol{\varepsilon}_{ij}$ represent measurement errors and are a set of M mutually independent normal random variables with mean zero and variance τ^2 where $M = [\sum_{i=1}^N n_i]$. Let \mathbf{R}_i be the (n_i, n_i) -dimensional matrix with the $(j, k)^{\text{th}}$ element being the correlation h_{ijk} between $W_i(t_{ij})$ and $W_i(t_{ik})$ defined as $h_{ijk} = \rho(|t_{ij} - t_{ik}|)$. Let \mathbf{I}_{n_i} be the (n_i, n_i) -identity matrix. The covariance matrix of $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \dots, \varepsilon_{in_i})$ is defined as :

$$\text{Var}(\boldsymbol{\varepsilon}_i) = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^T + \sigma^2\mathbf{R}_i + \tau^2\mathbf{I}_{n_i}. \quad (2.2.3)$$

With this specification of the covariance structure for longitudinal data, it is up to the analyst to introduce one or several sources of variability in the linear model according to the context of the experience or study. One can see that the linear mixed-effects model in Equation (2.2.1) corresponds to the particular case of Equation (2.2.2) where the within-subject serial correlation is omitted and a diagonal covariance structure is imposed on the measurement errors.

The inference and parameter estimation of these models (Verbeke and Molenberghs [121]; Diggle et al. [23]; Fitzmaurice et al. [32]; Molenberghs and Verbeke [92]; Fitzmaurice et al. [31]) are based on the well-known principle of maximum likelihood (ML) or the restricted maximum likelihood (REML) estimations. The REML is used to adjust the bias introduced by the maximum likelihood estimation of the covariance components.

There are in the literature several algorithms for the computation of maximum likelihood

or REML estimators. Laird and Ware [67] showed how the expectation-maximization (EM) algorithm proposed by Dempster et al. [22] can not only be applied to obtain ML estimators, but can be useful to compute REML estimators through an empirical Bayesian approach.

Other alternative procedures such as the Newton-Raphson algorithm, the quasi-Newton algorithm or the simplex algorithm of Nelder and Mead (1965) are used very often. However, a Bayesian approach through determination of the posterior probability distribution is sometimes preferred to estimate the random effects b_i in the linear mixed-effects models (Molenberghs and Verbeke [92, chap. 10]).

2.2.1.2. *Generalized linear models for longitudinal data*

Generalized linear models (McCullagh and Nelder [82]) are a class of regression models on the independent observations of a discrete or continuous variable. Statisticians have then developed extensions of generalized linear models for longitudinal data in order to take into account the context of correlated observations. We present three main extensions : *marginal models*, *generalized linear mixed-effects models* and *transition models*.

Marginal models : In a marginal model, the regression of a response variable on some explanatory variables and the within-subject correlation are analyzed separately according to the following assumptions :

- (1) The mean of each response $E(Y_{ij}|\mathbf{X}_{ij}) = \mu_{ij}$ depends on the explanatory variables \mathbf{X}_{ij} through $\gamma(\mu_{ij}) = \mathbf{X}_{ij}^T \boldsymbol{\beta}$ where $\gamma(\cdot)$ is a known *link function* such as the logit for binary responses.
- (2) The variance of each Y_{ij} given the covariates, is assumed to depend on the mean through $Var(Y_{ij}|\mathbf{X}_{ij}) = \varphi v(\mu_{ij})$ where φ is a scale parameter and v is a known variance function.
- (3) The correlation between Y_{ij} and Y_{ik} is a function of the corresponding means and additional parameters α through $Cor(Y_{ij}, Y_{ik}) = \rho(\mu_{ij}, \mu_{ik}; \alpha)$ where ρ is a known function.

It is the third component of the marginal model specification, the within-subject correlation between measurements collected for the same individual, that represents the main extension of generalized linear models to longitudinal data.

Generalized linear mixed-effects models : In a certain sense, marginal models take into account the within-subject correlation but they provide no explanation about the potential source of that correlation. An alternative approach that takes into account that within-subject correlation and provides its source consists in the introduction of random

effects in the model. This is the case of generalized linear mixed-effects models defined by the following assumptions :

- (1) Conditionally to the random effects vector \mathbf{b}_i , the Y_{ij} are independent and have distributions from an exponential family with a conditional mean that depends on fixed effects as well as random effects according to $\gamma\{E(Y_{ij}|\mathbf{b}_i)\} = \mathbf{X}_{ij}^T\boldsymbol{\beta} + \mathbf{Z}_{ij}^T\mathbf{b}_i$ where $\gamma(\cdot)$ is a known link function.
- (2) The conditional variance is a function of the conditional mean through : $Var(Y_{ij}|\mathbf{X}_{ij}) = \varphi v\{E(Y_{ij}|\mathbf{b}_i)\}$ where v is a known variance function and φ is a scale parameter known or to be estimated.
- (3) The random effects \mathbf{b}_i are independent from explanatory variables \mathbf{X}_{ij} and follow a multivariate normal distribution with mean zero and covariance matrix \mathbf{G} of dimension (q, q) .

The interpretation of the regression parameters is distinct in the marginal models and in the generalized linear mixed-effects models due to the difference in the target of inference. Generalized linear mixed-effects models are more appropriate when the study objective is to make inference on individuals rather than on whole population (Fitzmaurice et al. [31, chap. 2]).

Transition models : In a transition model, each realization Y_{ij} of a longitudinal sequence \mathbf{Y}_i is defined as a function of past responses and covariates. The dependence between repeated measures is modeled as resulting from the influence of past values on the present observation. Thus, in transition models, one assumes that $\gamma(E\{Y_{ij}|\mathbf{X}_{ij}, \mathbf{H}_{ij}\}) = \mathbf{X}_{ij}^T\boldsymbol{\beta} + \sum_{r=1}^s \alpha_r \eta_r(\mathbf{H}_{ij})$, where $\mathbf{H}_{ij} = (Y_{i1}, \dots, Y_{ij-1})$ represents the history of measures collected before the j^{th} occasion, and the η_r are known functions (often linear but not necessarily). With generalized linear mixed-effects models and transition models, it is possible to estimate unknown parameters using traditional maximum likelihood methods (Diggle et al. [23]). The most commonly used methods for estimation in generalized linear mixed-effects models are : maximum likelihood (ML), penalized quasi-likelihood (PQL) and Monte Carlo Markov chain methods (MCMC). For example, to obtain ML estimates, the numerical integration or the Monte Carlo integration is combined to optimization algorithms such as Newton-Raphson, Fisher Scoring or EM algorithm (Fitzmaurice et al. [31]). For parameter estimation in marginal models, the most commonly used approach is the generalized estimating equations (GEE) method developed by Zeger and Liang [132]. Molenberghs and Verbeke [92] present a standard iterative procedure for estimation using GEE.

2.2.1.3. Non-linear models for longitudinal data

Non-linear models are an important class of longitudinal data analysis models. Unlike generalized linear models where a certain restricted form of non-linearity can be introduced through the link function, non-linear models are fundamentally non-linear. Indeed, in marginal and generalized mixed-effects models, a non-linear link function $\gamma(\cdot)$ determines an appropriate scale on which the transform of the mean of a measurement Y_{ij} is linear in the regression parameters (and random effects according to the underlying model). However, in a non-linear model, the concept of linear predictor is abandoned and one obtains, in the case of marginal model for example :

$$E(Y_{ij}|\mathbf{X}_{ij}) = \eta(\mathbf{X}_{ij}, \boldsymbol{\beta}), \quad (2.2.4)$$

where $\eta(\cdot)$ is an arbitrary function of covariates and parameters. All three types of generalized linear models can be extended to non-linear models. Mixed-effects non-linear models are the most commonly used. An excellent description of those models is presented in Molenberghs and Verbeke [92] and Fitzmaurice et al. [31].

2.2.2. Non-parametric and semi-parametric models for longitudinal data analysis

In longitudinal data analysis, one is usually interested in the estimation of the underlying curve that generate the observed measures. For that purpose, parametric models are very often proposed. However, the main problem with parametric modeling is the quest of a suitable model with a limited number of parameters and that is the best fit to the data. Furthermore, these models suffer from an inflexibility to analyze structures sometimes very complicated and challenging in longitudinal data. It would then be more appropriate that the relationship between the mean of a response variable and covariates does not rely completely on parametric assumptions. Non-parametric and semi-parametric models then represent an alternative to parametric models and have encountered very significative developments in the past years. Non-parametric and semi-parametric models for independent data have then been extended to longitudinal data where the presence of within-subject correlation is a major challenge to take up.

Let's consider a longitudinal study with a single explanatory variable \mathbf{X} . Let (Y_{ij}, X_{ij}) be the dependent variable and the covariate for individual i ($i = 1, \dots, N$) measured at time point t_{ij} ($j = 1, \dots, n_i$). The dependent variable can be continuous or discrete. The marginal mean and the marginal variance of Y_{ij} are given by : $\mu_{ij} = E(Y_{ij}|X_{ij})$ and $Var(Y_{ij}|X_{ij}) = \varphi^{-1}v(\mu_{ij})$ where $v(\cdot)$ is a variance function and φ is a scale parameter. It is assumed that the marginal

mean depends on X_{ij} according to :

$$\gamma(\mu_{ij}) = \theta(X_{ij}), \quad (2.2.5)$$

where $\theta(\cdot)$ is an unknown smooth function and $\gamma(\cdot)$ is a known link function. The commonly used link functions include the *identity link* [$\gamma(\mu) = \mu$] for Gaussian realizations; the *logit link* [$\gamma(\mu) = \log\{\mu/(1 - \mu)\}$] or the *probit link* [$\gamma(\mu) = \Phi^{-1}(\mu)$] for binary realizations (Φ is the cumulative distribution function of a Gaussian distribution); and the *log link* [$\gamma(\mu) = \log(\mu)$] for Poisson type realizations. Among non-parametric models, which are distinct according to the method of estimation of the θ function, there are models based on kernel methods and models based on spline methods.

2.2.2.1. Kernel-based non parametric methods

There are two main methods : the local polynomial kernel generalized estimating equations (LPK-GEE) estimator and the kernel-seemingly unrelated estimator (kSUR). The LPK-GEE estimator (Lin and Carroll [73]) is an extension of the conventional local polynomial kernel (Fan and Gijbels [29]) estimator to longitudinal data through the introduction of a covariance matrix in a way similar to the method of generalized estimating equations for generalized linear models. Specifically, at a target point x , $\theta(X_{ij})$ is locally approximated by a d^{th} -order polynomial as :

$$\theta(X_{ij}) \approx \alpha_0 + \dots + \alpha_d(X_{ij} - x)^d = \mathbf{X}_{i(j,\cdot)}^\top \boldsymbol{\alpha} \quad (2.2.6)$$

where $\mathbf{X}_{i(j,\cdot)} = \{1, \dots, (X_{ij} - x)^d\}$ and $\boldsymbol{\alpha} = (\alpha_0, \dots, \alpha_d)^\top$. The estimation equation of the symmetric LPK-GEE estimator is :

$$\sum_{i=1}^N \mathbf{X}_i^\top \boldsymbol{\Delta}_i \mathbf{K}_{ih}^{1/2} \mathbf{V}_i^{-1} \mathbf{K}_{ih}^{1/2} \{\mathbf{Y}_i - \boldsymbol{\mu}_i\} = 0 \quad (2.2.7)$$

where \mathbf{X}_i is the matrix whose j^{th} row is $\mathbf{X}_{i(j,\cdot)}$; $\boldsymbol{\Delta}_i = \text{diag}\{\delta_{ij}\}$ with $\delta_{ij} = 1/\gamma'(\mu_{ij})$; $\mathbf{K}_{ih} = \text{diag}\{K_h(X_{ij} - x)\}$. Note that $K_h(s) = h^{-1}K(s/h)$ where h is a bandwidth and $K(\cdot)$ is a kernel function that is often chosen as a symmetric probability density function of mean zero. The matrix \mathbf{V}_i is a ‘‘working’’ covariance matrix defined as $\mathbf{V}_i = [\mathbf{S}_i^{1/2} \mathbf{R}_i(\xi) \mathbf{S}_i^{1/2}]$ with $\mathbf{S}_i = \text{diag}\{\varphi^{-1}v(\mu_{ij})\}$ and \mathbf{R}_i is an invertible correlation matrix, which possibly depends on a vector of parameters ξ that can be estimated using the method of moments. Also, $\boldsymbol{\mu}_i = \{\mu_{i1}, \dots, \mu_{in_i}\}^\top$ with $\mu_{ij} = \gamma^{-1}(\mathbf{X}_{i(j,\cdot)}^\top \boldsymbol{\alpha})$. Most commonly used kernel functions $K(\cdot)$ include the Gaussian kernel, the uniform kernel and the Epanechnikov kernel.

Note that \mathbf{R}_i is a user-specified working correlation matrix. It is used to account for the within-subject correlations of responses and to estimate the true correlation structure which is unknown. The three most common used working correlation structures are *Independent*

($Cor(Y_{ij}, Y_{ik}) = 0, j \neq k$), *Exchangeable* or *Compound symmetry* ($Cor(Y_{ij}, Y_{ik}) = \rho, j \neq k$) and *AR(1)* or *First-order autoregressive* ($Cor(Y_{ij}, Y_{ik}) = \rho^{|j-k|}, j \neq k$). Hu et al. [54], among others, discussed the choice of a good working correlation structure and develop a nonparametric and data-adaptive method for selecting the correlation structure.

Equation (2.2.7) can be solved using the Fisher Scoring algorithm via iteratively re-weighted least squares. Let $\hat{\boldsymbol{\alpha}}$ be the solution of that equation. Then, the estimator LPK-GEE of $\theta(x)$ is $\hat{\theta}_K(x) = \hat{\alpha}_0$. Lin and Carroll [73] have also considered a non-symmetric LPK-GEE estimator by replacing $[\mathbf{K}_{ih}^{1/2} \mathbf{V}_i^{-1} \mathbf{K}_{ih}^{1/2}]$ by $[\mathbf{V}_i^{-1} \mathbf{K}_{ih}]$ and the asymptotic performance of the estimator is similar to the symmetric case. In Chen and Jin [15], the authors have proposed an improved version of the LPK-GEE estimator which uses a degenerate local working covariance matrix. Unlike the original LPK-GEE estimator which is more efficient while the within-subject correlation is ignored (Wu and Zhang [127]), the efficiency of the kernel estimator of Chen and Jin [15] is the same, no matter whether one ignores correlation or accounts for correlation.

A heuristic explanation concerning the failure of the LPK-GEE estimator to efficiently account for the within-subject correlation is that it is based on the principle of local likelihood (Fitzmaurice et al. [31, chap. 9]). The design of a kernel-based estimator that accounts for the correlation in longitudinal data should then not rely on the traditional local likelihood. Thus, Wang [125] proposed the kernel-based seemingly unrelated (kSUR) estimator. Specifically, let's consider a d^{th} -order polynomial kernel estimator using the approximation in Equation (2.2.7). If $\hat{\theta}_K^{[l]}(x)$ represents the kSUR estimator of $\theta(x)$ at the l^{th} iteration, then at the $(l+1)^{th}$ iteration, one has $\hat{\theta}_K^{[l+1]}(x) = \hat{\alpha}_0$ where $\hat{\boldsymbol{\alpha}} = (\hat{\alpha}_0, \hat{\alpha}_1, \dots, \hat{\alpha}_d)^T$ is solution to the following equation :

$$\sum_{i=1}^N \sum_{j=1}^{n_i} K_h(X_{ij} - x) \mathbf{X}_i^T \mathbf{V}_i^{-1} \{\mathbf{Y}_i - \boldsymbol{\mu}_{i(j)}\} = 0. \quad (2.2.8)$$

In Equation (2.2.8), \mathbf{V}_i is the working covariance matrix; \mathbf{X}_i is a $(n_i, d+1)$ -dimensional matrix of zeros except at the j^{th} row which is $\{1, (X_{ij} - x), \dots, (X_{ij} - x)^d\}^T$; K_h is a kernel function and

$$\boldsymbol{\mu}_{i(j)} = \left\{ \hat{\theta}_K^{[l]}(X_{i1}), \dots, \hat{\theta}_K^{[l]}(X_{i,j-1}), \sum_{k=0}^d (X_{ij} - x)^k \alpha_k, \hat{\theta}_K^{[l]}(X_{i,j+1}), \dots, \hat{\theta}_K^{[l]}(X_{in_i}) \right\}^T.$$

The kSUR estimator $\hat{\theta}_K^*(x) = \hat{\alpha}_0$ is obtained at the convergence. The Fisher Scoring algorithm can be used to iteratively solve Equation (2.2.8). The kSUR estimator is convergent and effectively accounts for the within-subject correlation. Simulation results showed that it is more efficient than the kernel-based GEE estimator in term of the quadratic mean

error (Wang [125]). Lin and Carroll [75] have extended the kSUR method to the likelihood principle.

2.2.2.2. Splines-based non parametric methods

An alternative method to non-parametrically estimate the function $\theta(x)$ in Equation (2.2.5) consists in the use of smoothing splines. A smoothing spline estimates the non-parametric regression function $\theta(x)$ using a piecewise polynomial function with all the observed covariate values X_i used as knots, where smoothness constraints are assumed at those knots (Wahba [122], Green and Silverman [45]). For longitudinal data, there are essentially, the generalized smoothing splines estimator, the P-splines estimator and the regression splines estimator. To illustrate key features of the generalized smoothing splines estimator, let's consider the Gaussian realizations (*identity link*), thus the model :

$$Y_{ij} = \theta(X_{ij}) + \epsilon_{ij}, \quad (2.2.9)$$

where the $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \dots, \epsilon_{in_i})^\top$ are independent with zero mean and covariance matrix $\boldsymbol{\Sigma}$. By assuming a working covariance matrix \mathbf{V}_i , the r^{th} -order smoothing splines estimator minimizes

$$-\frac{1}{2N} \sum_{i=1}^N \{\mathbf{Y}_i - \boldsymbol{\theta}_i\}^\top \mathbf{V}_i^{-1} \{\mathbf{Y}_i - \boldsymbol{\theta}_i\} - \frac{1}{2} \lambda \int \{\theta^{(r)}(x)\}^2 dx \quad (2.2.10)$$

$$= -\frac{1}{2N} \sum_{i=1}^N \{\mathbf{Y}_i - \boldsymbol{\theta}_i\}^\top \mathbf{V}_i^{-1} \{\mathbf{Y}_i - \boldsymbol{\theta}_i\} - \frac{1}{2} \lambda \boldsymbol{\theta}^\top \boldsymbol{\Omega} \boldsymbol{\theta} \quad (2.2.11)$$

where $\boldsymbol{\theta}_i = \{\theta(X_{i1}), \dots, \theta(X_{in_i})\}$; λ is a tuning parameter controlling the trade-off between the fitting quality and smoothing level of the curve $\theta(\cdot)$; $\boldsymbol{\Omega}$ is the smoothing matrix (Green and Silverman [45]). Consequently, the r^{th} -order smoothing splines estimator of $\theta(x)$ is :

$$\hat{\boldsymbol{\theta}}_S = (\tilde{\mathbf{V}}^{-1} + N\lambda\boldsymbol{\Omega})^{-1} \tilde{\mathbf{V}}^{-1} \mathbf{Y}, \quad (2.2.12)$$

where $\tilde{\mathbf{V}} = \text{diag}(\mathbf{V}_1, \dots, \mathbf{V}_N)$ and $\mathbf{Y} = (\mathbf{Y}_1^\top, \dots, \mathbf{Y}_N^\top)^\top$. Lin et al. [76] have studied the theoretical properties of the smoothing splines estimator and demonstrated that it is asymptotically equivalent to the kernel SUR estimator. Their results indicated that the generalized smoothing splines estimator is a higher order kernel SUR estimator (for example, the cubic smoothing splines estimator ($r = 2$) is a fourth-order kernel SUR estimator). In addition, the generalized smoothing splines estimator is convergent and the most efficient estimator $\hat{\boldsymbol{\theta}}_S$ is obtained when the within-subject correlation is accounted for.

Smoothing splines use all measurement points as knots. Hence, for very large datasets their computations can be long and complex. Regression splines (Stone et al. [114]) that use a small number of knots have been proposed for non-parametric regression on longitudinal

data under the model in Equation (2.2.9) (Rice and Wu [104]; Huang et al. [56]). A regression spline approximates $\theta(\cdot)$ by $\theta(x) = \sum_{l=0}^L B_l(x)\alpha_l$ where the number of knots is small and $\{B_l(\cdot)\}_{l=0}^L$ is a set of L basis functions such as *B-splines*. The coefficients α_l are estimated by weighted least squares.

Another alternative consists in using the P-splines (Eilers and Marx [26]; Ruppert et al. [107]) that require a moderate number of knots (usually smaller than the sample size but greater than the number of knots in the regression splines case). If $\{(x-x_1)_+^3, \dots, (x-x_M)_+^3\}$ refers to a *plus functions basis* built from the M knots x_1, \dots, x_M ($a_+ = a$ if $a > 0$ and 0 otherwise), then $\theta(\cdot)$ can be approximated by

$$\theta(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \alpha_3 x^3 + \sum_{m=1}^M (x-x_m)_+^3 \alpha_{m+3} = \mathbf{B}(x)^\top \boldsymbol{\alpha} \quad (2.2.13)$$

where $\mathbf{B}(x) = \{1, x, x^2, x^3, (x-x_1)_+^3, \dots, (x-x_M)_+^3\}^\top$ and $\boldsymbol{\alpha} = (\alpha_0, \dots, \alpha_M)^\top$. Note that in this case, $L = M + 3 + 1$. Under the model in Equation (2.2.9), the coefficients $\boldsymbol{\alpha}$ are estimated using the penalized log-likelihood.

2.2.2.3. Semi-parametric methods

Semi-parametric models have parametric as well as non-parametric components. Let's assume that Y_{ij} is the j^{th} measurement ($j = 1, \dots, n_i$) of individual i ($i = 1, \dots, N$). Equation (2.2.14) represents a marginal semi-parametric model where \mathbf{X}_{ij} is a p -dimensional vector of covariates whose effects are modeled parametrically and W_{ij} is another scalar covariate whose effects are modeled non-parametrically. Specifically,

$$\gamma(\mu_{ij}) = Y_{ij} = \mathbf{X}_{ij}^\top \boldsymbol{\beta} + \theta(W_{ij}) \quad (2.2.14)$$

where $\gamma(\cdot)$ is a known link function and $\theta(\cdot)$ is an unknown smooth function as in Equation (2.2.5). Following Zeger and Diggle [131], Lin and Carroll [74] have studied the model of Equation (2.2.14) by first estimating $\theta(\cdot)$, given $\boldsymbol{\beta}$, using the non-parametric method LPK-GEE. Then, given the resulting local polynomial kernel GEE estimator of $\theta(\cdot)$ denoted by $\hat{\theta}(w; \boldsymbol{\beta})$ and equal to $\hat{\alpha}_0$ (defined as in Section 2.2.2.1), the regression coefficients $\boldsymbol{\beta}$ are estimated through the *profile method* which consists in solving the profile estimating equation :

$$\sum_{i=1}^N \frac{\partial \boldsymbol{\mu}_i^\top}{\partial \boldsymbol{\beta}} \mathbf{V}_i^{-1} [\mathbf{Y}_i - \boldsymbol{\mu}_i] = 0, \quad (2.2.15)$$

where $\hat{\boldsymbol{\theta}}_i = \{\hat{\theta}(W_{i1}; \boldsymbol{\beta}), \dots, \hat{\theta}(W_{in_i}; \boldsymbol{\beta})\}^\top$; \mathbf{V}_i is a working covariance matrix and $\boldsymbol{\mu}_i$ is the vector whose j^{th} component is $\mu_{ij} = \gamma^{-1}\{\mathbf{X}_{ij}^\top \boldsymbol{\beta} + \hat{\theta}(W_{ij}; \boldsymbol{\beta})\}$.

The results of Lin and Carroll [74] indicate that these estimations fail to analyze correctly longitudinal data when the within-subject correlation is accounted for in the model. Therefore, Wang et al. [126] have proposed to estimate $\theta(w)$ using the kSUR estimator (Section 2.2.2.1) and β still by the profile method in Equation (2.2.15). They have then demonstrated that the *profile/kSUR* estimator of β is convergent for any given working covariance matrix and is the most efficient when the working matrix \mathbf{V}_i is equal to the true covariance of \mathbf{Y}_i .

A useful extension of the marginal semi-parametric model in Equation (2.2.14) to the likelihood paradigm is the generalized mixed-effects semi-parametric model in which one assumes that, conditionally to the random effects \mathbf{b}_i , Y_{ij} follows a distribution from an exponential family with mean μ_{ij} according to :

$$\gamma(\mu_{ij}) = \mathbf{X}_{ij}^T \beta + \theta(W_{ij}) + \mathbf{Z}_{ij}^T \mathbf{b}_i, \quad (2.2.16)$$

where \mathbf{X} , \mathbf{W} , β , and $\theta(\cdot)$ are defined as in Equation (2.2.14). The \mathbf{Z}_{ij} is a q -dimensional vector of covariates associated with random effects \mathbf{b}_i which are normally distributed $\mathcal{N}_q(\mathbf{0}, \mathbf{D}(\xi))$. The estimation of $\theta(\cdot)$ and (β, ξ) are done through the *profile/kSUR* method where $\theta(\cdot)$ is estimated using the non-parametric kSUR method and the estimation of (β, γ) is based on log-likelihood maximization of :

$$\sum_{i=1}^N l\{\mathbf{Y}_i; \beta, \xi, \hat{\theta}(W_{i1}; \beta, \xi), \dots, \hat{\theta}(W_{in_i}; \beta, \xi)\}. \quad (2.2.17)$$

An alternative estimation method in semi-parametric models is based on the use of splines (smoothing splines or P-splines) to estimate the non-parametric function $\theta(\cdot)$.

2.2.2.4. *The estimation of the covariance in longitudinal data analysis*

All those non-parametric and semi-parametric models require the specification of a covariance matrix in order to adequately estimate the mean, but they do not provide any systematic procedure to estimate the covariance structure.

Some authors have worked on the non-parametric modeling of the covariance structure. Capra and Müller [10] as well as Staniswalis and Lee [113] have developed implementations with kernel-based smoothing methods. Briefly, the mean function is estimated by smoothing the aggregated data $(t_{ij}, Y_{ij}) \quad i = 1, \dots, N; \quad j = 1, \dots, n_i$ where t_{ij} and Y_{ij} are respectively time point and value of the j^{th} measurement from the individual i . Once the mean function $\mu(t)$ is obtained, raw covariances are computed from all observed pairs $(t_{ij}, Y_{ij}), (t_{il}, Y_{il})$

according to :

$$V_{ijl} = (Y_{ij} - \hat{\mu}(t_{ij}))(Y_{il} - \hat{\mu}(t_{il}))$$

Then, a second surface smoothing step is applied on the cloud $((t_{ij}, t_{il}), V_{ijl})$ in order to obtain the estimated covariance surface. Several types of smoothing can be applied and the smooth covariance function is then discretized in a uniform temporal grid, where it is represented as the covariance matrix.

Diggle and Verbyla [24] have introduced a non-parametric estimator of the covariance matrix for longitudinal data through the smoothing of the sample variogram. More recent works on the issue include Wu and Pourahmadi [128] and Huang et al. [55]. Also, Fan et al. [30] have studied sparse longitudinal data and proposed a semi-parametric method based on maximization of quasi-likelihood to estimate the covariance function. In that method, the variance function is modeled non-parametrically as a function of time while the correlation function is assumed parametric. The natural extension of that method is to estimate the covariance function fully non-parametrically, which is more and more implemented in functional data analysis (FDA). Li [70] proposed a method that combines the semi-parametric estimator through the profile method (described in Section 2.2.2.3) and the non-parametric estimation for the covariance.

2.3. CLUSTERING METHODS FOR LONGITUDINAL DATA

Sometimes, the analysis of longitudinal data, in addition to the typical analyses presented in Section 2.2, consists in building groups or clusters suggested by the data, not defined a priori. Indeed, clustering the subjects from a longitudinal study and analyzing them by cluster turns out to be an approach increasingly adopted by data analysts from various fields of study. Moreover, the obtained clusters can then be used in regression models to predict outcomes. Hence, methods have been developed to extend multivariate cluster analysis to longitudinal data with the objective of clustering subject trajectories. As discussed in He [49], that extension is necessary and motivated by the difference in the longitudinal data structure (generally unequal number of measurements for individuals, observations not necessarily obtained at the same time points, and presence of missing data).

Clustering refers to unsupervised classification and consists in performing a cluster analysis on a dataset with no cluster information other than the observed values. Most available clustering methods to be adapted to longitudinal data can be categorized into two approaches : **a non-parametric approach** and **a model-based approach**. Those two approaches aggregate the five clustering algorithm categories presented in Elavarasi et al. [27] : Hierarchical, Partition, Spectral, Grid-based and Density-based. They are also consistent with

the three types of clustering algorithms proposed in Ren [102] : connectivity models (based on the distance and linkage criteria), centroid models (such as the k-means algorithm), and distribution models (assuming the existence of a mixture of distributions and using conventionally the EM-algorithm for estimation). The classes of clustering models presented in He [49] (hierarchical models, centroid models and mixture models) can also be merged in those two approaches.

The non-parametric approach contains usual partition-based clustering techniques such as k-means (and its many extensions), hierarchical clustering, methods using specific distances or dissimilarities and methods using new heuristics or geometric criteria to cluster. The model-based clustering approach considers a probability distribution and mostly involves fitting a finite mixture of distributions.

2.3.1. Non-parametric clustering methods

This approach relates to classical algorithmic methods and is essentially based on identifying similar groups through the quantification of the similarity between two objects. The similarity measure is metric-based rather than based on a probability distribution. As indicated in Heggeseth [51], the three key ingredients to these methods are *the dissimilarity measure, the clustering algorithm, and the number of clusters*. Metric, distance, dissimilarity and similarity are all related concepts. Among the most popular metrics are the Euclidean distance, Manhattan distance, Pearson's correlation for continuous features, Spearman's rank correlation, Kendall's Tau for ordinal features, simple matching coefficient and Jaccard coefficient for binary features. The choice of the similarity measure should consider the features type and scale, the desired interpretation of similarity (e.g., proximity or association), sensitivity to outliers, and underlying distributional assumptions for the features (see Bruckers [8]).

Regarding the clustering algorithm, there is a distinction to be made between hierarchical and partitional clustering methods. In hierarchical clustering, a hierarchy of clusters is created, which can be represented by a tree structure called dendrogram. One can either start with the leaves of the tree (each individual as a separate cluster) and merge the clusters together to the root (agglomerative), or alternatively start at the root of the tree and split the clusters into leaves (divisive) (see Dufour [25]). The split or combination of clusters is done using the similarity measure in multiple ways : single linkage, complete linkage and average linkage. Partitional techniques produce a single partition of the objects into $k \geq 2$ disjoint clusters, by optimizing a criterion function. Two popular methods in this class are k-means and partitioning around medoids (PAM). The k-means algorithm attempts to minimize the sum of the squared distances between the objects and their cluster centers, by

iteratively reallocating objects to the clusters until convergence. The PAM method aims at partitioning N objects into G clusters in which each object is assigned to the cluster with the closest medoid (see Lin [71]). For these methods, the number of clusters needs to be specified unlike hierarchical methods.

Liu and Luo [77] mentioned the sensitiveness to data, the difficulty to determine the number of clusters of the well-known partition-based clustering algorithm k-means and the failure of the Euclidean distance (typically used in k-means as similarity measure) to get a good performance in pattern clustering on longitudinal data. They propose then, the *Max-Difference iterative clustering algorithm* (combining a new distance called Max-Difference distance and an iterative optimization) which, according to their results, has advantages on performance, computational complexity and non-sensitiveness to data compared with k-means. Similarly, Usami [119] discussed the importance of constraints in clustering methods. They proposed *a new constrained k-means method* with lower bound constraints on cluster proportions and distances among clusters at focused variables and time points to fulfill various needs in clustering longitudinal data. The new method, deemed to be satisfactory based on simulated and real data, assumes a large number of clusters at the onset and iteratively deletes and combines clusters according to these constraints and directly estimate the unknown number of clusters.

Lin [71] develops a new clustering method to deal with a new type of longitudinal data, category-ordered data which has features of categorical and ordinal scales. For these data, the implemented model-based clustering methods are hard to use, due to missing values of type missing at random, and the existing dissimilarity functions are not suitable. Thus, the authors proposed a method that involves a new dissimilarity function, the so-called p-dissimilarity. In addition, the method lets the data select the appropriate clustering algorithm (hierarchical linkage or PAM) based on cluster stability and coherence, respectively measured via the Prediction Strength (PS) index and the Average Silhouette Width (ASW).

2.3.2. Model-based clustering methods

Model-based clustering is the increasingly popular area of cluster analysis that relies on a probabilistic description of data via finite mixture models. In this approach, each cluster is mathematically represented by a parametric distribution. The data are assumed to be generated by a mixture of underlying distributions described by a set of parameters. In the most general form of a mixture, the density of a random variable Y takes the form :

$$p(\mathbf{y}|\boldsymbol{\pi}) = \sum_{g=1}^G \pi_g p_g(\mathbf{y}) \quad (2.3.1)$$

where π_g represents the g^{th} mixing proportion or the probability that an observation belongs to the g^{th} group (or cluster) with corresponding density p_g called the g^{th} mixing or group density. The number G represents the total number of groups and is to be additionally estimated when it is unknown. Note that $\boldsymbol{\pi} = (\pi_1, \pi_2, \dots, \pi_G)^\top$ with $\pi_g \in (0, 1)$ and $\sum_{g=1}^G \pi_g = 1$. Usually, the component densities p_g are assumed to be of parametric form $p_g(\mathbf{y}) = p_g(\mathbf{y}|\boldsymbol{\Pi}_g)$ with a completely known functional form. Equation (2.3.1) can then be rewritten and takes the form

$$p(\mathbf{y}|\boldsymbol{\Pi}) = \sum_{g=1}^G \pi_g p_g(\mathbf{y}|\boldsymbol{\Pi}_g) \quad (2.3.2)$$

of Equation (2.3.2), which is referred to as a finite mixture model with parameter vector $\boldsymbol{\Pi} = (\boldsymbol{\pi}^\top, \boldsymbol{\Pi}_1^\top, \boldsymbol{\Pi}_2^\top, \dots, \boldsymbol{\Pi}_G^\top)^\top$. Further review of finite mixture models can be found in McLachlan and Peel [85] and Melnykov and Maitra [90].

The most common type of mixture considered in the literature is unquestionably the Gaussian mixture model, but other models investigated are the Poisson mixture, the skew-Normal and t-distribution mixtures (as heavy-tailed alternative to Gaussian mixtures, see McLachlan and Peel [84]). The Gaussian mixture model assumes a multivariate Gaussian distribution for each group and is expressed as :

$$p(\mathbf{y}) = \sum_{g=1}^G \pi_g \phi(\mathbf{y}|\boldsymbol{\mu}_g, \boldsymbol{\Gamma}_g) \quad (2.3.3)$$

where π_g is as previously defined and $\phi(y|\boldsymbol{\mu}_g, \boldsymbol{\Gamma}_g)$ is the density of a multivariate Gaussian distribution with mean $\boldsymbol{\mu}_g$ and covariance matrix $\boldsymbol{\Gamma}_g$. Many works such as Banfield and Raftery [3], Celeux and Govaert [12], Fraley and Raftery [34, 36, 37] parameterized the group covariance structure $\boldsymbol{\Gamma}_g$ through an eigenvalue decomposition in the form $\boldsymbol{\Gamma}_g = \lambda_g \boldsymbol{\Lambda}_g \boldsymbol{\Delta}_g \boldsymbol{\Lambda}_g^\top$ where $\boldsymbol{\Lambda}_g$ is the orthogonal matrix of eigenvectors, $\boldsymbol{\Delta}_g$ is a diagonal matrix whose elements are proportional to the eigenvalues and λ_g is an associated constant of proportionality. The idea underlying this decomposition is to treat λ_g , $\boldsymbol{\Delta}_g$ and $\boldsymbol{\Lambda}_g$ as independent sets of parameters and either constrain them to be the same for each cluster or allow them to vary among clusters in order to give a wide range of parsimonious covariance structures. The diverse exploitations of that decomposition resulted in a well-known family of mixture models for model-based clustering : the Mclust family (Fraley and Raftery [37]) which consists of ten mixture models that arise from the imposition of constraints upon the cluster covariance (Fraley and Raftery [38]). The clustering algorithm attempts to find the best estimates of the parameters by maximizing the log-likelihood function via the EM algorithm (see Melnykov and Maitra [90]; McLachlan and Peel [85]).

Another family of eight Gaussian mixture models with parsimonious covariance structure has been introduced in McNicholas and Murphy [86] by extending the mixture of factor analyzers model (Ghahramani and Hinton [43]). Under the general mixture of factor analyzers model, the density of an observation in group g is multivariate Gaussian with mean $\boldsymbol{\mu}_g$ and the covariance structure is assumed of the form $\boldsymbol{\Gamma}_g = \boldsymbol{\Lambda}_g \boldsymbol{\Lambda}_g^\top + \boldsymbol{\Delta}_g$, where the loading matrix $\boldsymbol{\Lambda}_g$ is a $p \times q$ matrix of parameters typically with $q \ll p$ and the noise matrix $\boldsymbol{\Delta}_g$ is a diagonal matrix. The maximum likelihood estimates for the parameters in these models are found using the Alternating Expectation-Conditional Maximization (AECM) algorithm (Meng and van Dyk [91]). That algorithm is an extension of the EM algorithm that uses different specifications of missing data at each stage.

In model-based clustering, clusters are defined as observations coming most likely from the same distribution. For parameter estimation in mixture models, possible approaches include the method of moments and distance-based procedures but the maximum likelihood estimation carried out by means of the EM algorithm and the Bayesian approach (via Markov chain Monte Carlo procedures) are by far the most popular methods. More developments on the estimation issues (possibility of unbounded likelihood function, singular covariance matrices, spurious solutions, the EM initialization difficulties etc.) of finite mixture models are presented in Celebi [11, chap. 1] and Melnykov [89]. The uncertainty for cluster-membership assignment of each observation is naturally quantified via the posterior probabilities. Indeed, each individual is classified into the group to which it has the highest estimated posterior probability.

For the optimal model selection in model-based clustering, an information criterion is usually chosen and the two most popular are the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). Overall, the two criteria minimize the negative log likelihood function augmented by some penalty or adjustment imposed to reflect model complexity. If $\mathcal{L}(\hat{\boldsymbol{\Pi}}|\mathbf{y})$ denotes the likelihood function evaluated at the ML estimate of $\boldsymbol{\Pi}$, the criteria are calculated for each fitted candidate model as : AIC = $-2 \log \mathcal{L}(\hat{\boldsymbol{\Pi}}|\mathbf{y}) + (2 \times d)$ and BIC = $-2 \log \mathcal{L}(\hat{\boldsymbol{\Pi}}|\mathbf{y}) + (\log(N) \times d)$ where d is the dimension of $\boldsymbol{\Pi}$ and N is the size of the data. As discussed in studies such as Everitt et al. [28] and Melnykov and Maitra [90], AIC and BIC are easily implemented criteria with good performance in the selection of the adequate number of clusters but they have some inconsistencies. For example, AIC tends to overestimate G while BIC tends to underestimate G when the sample size is small.

2.3.3. Recent developments in longitudinal cluster analysis

For longitudinal data specifically, the objective of cluster analysis leads to the exercise of finding groups of subjects with similar trajectories or patterns in one or more variables. Longitudinal data consist of measurements taken at different times on each individual, with the typical feature, especially in clinical settings or behavioral sciences, that both the number of measurements and the time points may differ across the individuals. The correlation structure associated with that type of data also presents significant modeling challenges. The usefulness of clustering methods dedicated specifically to longitudinal data analysis has only recently become well recognized among researchers (Usami [119]). In the last two decades, there has been extensive methodological work to extend the methods from multivariate cluster analysis to the area of longitudinal data. Most of the clustering methods proposed to deal with longitudinal data are purposely designed to address a specific inherent characteristic of such data by modifying or adjusting a technical aspect of existing clustering methods for multivariate data from the two approaches (non-parametric and model-based).

2.3.3.1. Longitudinal cluster analysis in gene expression data

A rapid expansion of algorithms for longitudinal data originates, amongst others, from their need in gene-expression data analysis where the clustering of co-regulated genes is an important task as it is critical for reliable inference of the underlying biological processes. Statistically, the problem of clustering time course data is a special case of the more general problem of clustering longitudinal data (McNicholas and Subedi [88]). As discussed in Chan et al. [14], the k-means algorithm, although efficient and regularly used in that area, is prone to produce only locally optimal solutions that are, in addition, sensitive to the initial conditions. In order to alleviate these problems, Chan et al. [14] proposed a novel global clustering method called the greedy elimination method (GEM). They showed that the GEM is effective in enhancing the global optimality and consistency of the clustering solutions, based on real gene expression data.

Another approach widely adopted in that area of bioinformatics applications is the model-based clustering with a challenge associated with the correlation structure. The application of the model-based approach to clustering gene expression data was first discussed in Yeung et al. [130], where a Gaussian mixture model was used. Following the mixture modeling framework, De la Cruz-Mesia et al. [20] showed some merits of model-based clustering over non-probabilistic clustering techniques and introduced, as an extension of the work of Pauler and Laird [96], a mixture of nonlinear hierarchical models in which each component density is subject-specific. Indeed, a vector $\mathbf{y}_i = (y_{i1}, y_{i2}, \dots, y_{in_i})$ of measurements taken at different

times for an individual i is assumed to follow a mixture model :

$$\mathbf{y}_i \sim \sum_{g=1}^G \pi_g p_g(h_g(\mathbf{\Pi}_{ig}, x_{ig}); \mathbf{\Sigma}_{ig}) \quad (2.3.4)$$

where the densities $p_g(h_g(\mathbf{\Pi}_{ig}, x_{ig}); \mathbf{\Sigma}_{ig})$ are indexed by a mean $h_g(\cdot)$ and a $(n_i \times n_i)$ covariance matrix $\mathbf{\Sigma}_{ig}$ which only depends on i for its dimension. They assume $h_g(\cdot)$ to be a nonlinear function of unknown individual-specific parameters $\mathbf{\Pi}_{ig}$ and known covariates x_{ig} . For each g , the parameters vector $\mathbf{\Pi}_{ig}$ of dimension p follows a multivariate Gaussian distribution $\mathcal{N}_p(\boldsymbol{\mu}_g, \mathbf{\Gamma}_g)$. For parameter estimation, the authors studied both frequentist (maximum likelihood estimation via an EM-type algorithm) and Bayesian (a sequence of Gibbs and Metropolis-Hastings steps) approaches. Concerning model selection, the BIC criterion was chosen and from a Bayesian viewpoint, they used the Bayes factor as selection tool. However, the only modeling of the covariance structure in their model was the imposition of the *isotropic constraint* $\mathbf{\Sigma}_{ig} = \sigma_g^2 \mathbf{I}_{n_i}$, suggesting that the variability is the same at all time points.

McNicholas and Murphy [87] introduced a family of mixture models with a covariance structure considered to be specifically designed for the model-based clustering of longitudinal data. They assume a Gaussian mixture model with a modified Cholesky decomposition for each group covariance structure $\mathbf{\Gamma}_g$ of the form :

$$\mathbf{\Gamma}_g^{-1} = \mathbf{T}_g^\top \mathbf{\Delta}_g^{-1} \mathbf{T}_g \quad (2.3.5)$$

where \mathbf{T}_g is a unique $(p \times p)$ lower triangular matrix with diagonal elements 1 and $\mathbf{\Delta}_g$ is a unique $(p \times p)$ diagonal matrix with strictly positive diagonal entries. Equation (2.3.5) is the equivalent expression of the known modified Cholesky decomposition of $\mathbf{\Gamma}_g$ which is originally expressed as $\mathbf{\Delta}_g = \mathbf{T}_g \mathbf{\Gamma}_g \mathbf{T}_g^\top$ and has been used in Krzanowski et al. [66] and Pourahmadi [98]. Based on the decomposition in Equation (2.3.5), the density of an observation \mathbf{y}_i in group g is given by :

$$\phi(\mathbf{y}_i | \boldsymbol{\mu}_g, \mathbf{T}_g, \mathbf{\Delta}_g) = \frac{1}{\sqrt{(2\pi)^p |\mathbf{\Delta}_g|}} \exp\left\{-\frac{1}{2}(\mathbf{y}_i - \boldsymbol{\mu}_g)^\top \mathbf{T}_g^\top \mathbf{\Delta}_g^{-1} \mathbf{T}_g (\mathbf{y}_i - \boldsymbol{\mu}_g)\right\} \quad (2.3.6)$$

Let \mathbf{z} denote the group membership indicators ($z_{ig} = 1$ if individual i belongs to group g and 0 otherwise). Let $\mathcal{L}(\mathbf{\Pi} | \mathbf{y}, \mathbf{z})$ denote the likelihood of the complete-data (\mathbf{y}, \mathbf{z}) and let $\mathcal{Q}(\mathbf{\Pi})$ denote the conditional expectation of the complete-data log likelihood with respect to the missing data \mathbf{z} , given the observed data \mathbf{y} and the set of the parameters $\mathbf{\Pi}$. The

expressions of $\mathcal{L}(\mathbf{\Pi}|\mathbf{y}, \mathbf{z})$ and $\mathcal{Q}(\mathbf{\Pi})$ (the expected value) are given by :

$$\begin{cases} \mathcal{L}(\mathbf{\Pi}|\mathbf{y}, \mathbf{z}) &= \prod_{i=1}^N \prod_{g=1}^G \left[\pi_g \phi(\mathbf{y}_i | \boldsymbol{\mu}_g, (\mathbf{T}_g^\top \boldsymbol{\Delta}_g^{-1} \mathbf{T}_g)^{-1}) \right]^{z_{ig}} \\ \mathcal{Q}(\mathbf{\Pi}) &= \sum_{g=1}^G \sum_{i=1}^N \hat{z}_{ig} \log \left[\pi_g \phi(\mathbf{y}_i | \boldsymbol{\mu}_g, (\mathbf{T}_g^\top \boldsymbol{\Delta}_g^{-1} \mathbf{T}_g)^{-1}) \right] \\ &= \left(\sum_{g=1}^G N_g \log \pi_g \right) - \left(\frac{np}{2} \log(2\pi) \right) - \left(\sum_{g=1}^G \frac{N_g}{2} \log |\boldsymbol{\Delta}_g| \right) \\ &\quad - \left(\sum_{g=1}^G \frac{N_g}{2} \text{tr} \{ \mathbf{T}_g \mathbf{S}_g \mathbf{T}_g^\top \boldsymbol{\Delta}_g^{-1} \} \right) \end{cases} \quad (2.3.7)$$

where $\mathbf{S}_g = \frac{1}{N_g} \sum_{i=1}^N z_{ig} (\mathbf{y}_i - \boldsymbol{\mu}_g)(\mathbf{y}_i - \boldsymbol{\mu}_g)^\top$ and $N_g = \sum_{i=1}^N z_{ig}$. As stated in Shaikh [110], the group-covariance decomposition of Equation (2.3.5) is particularly useful for longitudinal data as the diagonal entries of matrix $\boldsymbol{\Delta}_g$ reflect the variations within each time point of group g while the sub-diagonal elements of \mathbf{T}_g represent relationships between time points of group g . The various constraints that can be placed upon \mathbf{T}_g and $\boldsymbol{\Delta}_g$ result in a family of eight Gaussian mixture models that are fitted using an EM algorithm. The Bayesian information criterion (BIC) is used to select the best member of this family.

Shaikh et al. [111] extended that modeling framework (with modified Cholesky-decomposed covariance structure) to accommodate *incomplete longitudinal data*, meaning that there is missing data in addition to the missing group memberships (Shaikh [110]). Their main contribution consists in developing a modified EM algorithm in which the missing data are imputed at each iteration and taken into account in the next iteration. The imputed values are then combined with the observed values to produce an approximation to the expected value of the complete-data log-likelihood given by :

$$\mathcal{Q}(\mathbf{\Pi}) = \sum_{g=1}^G \sum_{i=1}^N z_{ig} \log \left[\pi_g \phi(\tilde{\mathbf{y}}_i | \boldsymbol{\mu}_g, (\mathbf{T}_g^\top \boldsymbol{\Delta}_g^{-1} \mathbf{T}_g)^{-1}) \right]. \quad (2.3.8)$$

The vector $\tilde{\mathbf{y}}_i$ is the i^{th} individual observations where \tilde{y}_{ij} is the imputed value of y_{ij} if it is missing or the observed value if it is available. This expression of $\mathcal{Q}(\mathbf{\Pi})$ is identical to the one of McNicholas and Murphy [87] in Equation (2.3.7) except that $\tilde{\mathbf{y}}_i$ is used instead of \mathbf{y}_i . Indeed, the vector \mathbf{y}_i of dimension p is partitioned so that $\mathbf{y}_i^1 \in \mathbb{R}^r$ represents the r missing values and $\mathbf{y}_i^2 \in \mathbb{R}^{p-r}$ represents the observed values. As in the usual EM algorithm (used in McNicholas and Murphy [87]), the group membership indicators \hat{z}_{ig} are estimated at each iteration. In addition, the imputed values in $\hat{\mathbf{y}}_i^1$ used to replace the missing values in the expectation step are calculated via the expression :

$$\hat{\mathbf{y}}_i^1 = \sum_{g=1}^G z_{ig} \left[\boldsymbol{\mu}_g^1 + \boldsymbol{\Gamma}_{12g} \boldsymbol{\Gamma}_{22g}^{-1} (\mathbf{y}_i^2 - \boldsymbol{\mu}_g^2) \right] \quad (2.3.9)$$

where $\boldsymbol{\mu}_g^1 \in \mathbb{R}^r$ is the mean of the g^{th} group at the r time points corresponding to the missing \mathbf{y}^1 ; $\boldsymbol{\mu}_g^2 \in \mathbb{R}^{p-r}$ is the mean of the g^{th} group at the $p-r$ time points corresponding to the observed \mathbf{y}^2 ; $\boldsymbol{\Gamma}_{12g}$ and $\boldsymbol{\Gamma}_{22g}$ are block matrices for group g defined as in Appendix

A.1. The algorithm used in Shaikh et al. [111] is not a proper EM algorithm and is termed *pseudo-EM algorithm* due to that approximation to the expected value of the complete-data log-likelihood.

Furthermore, McNicholas and Subedi [88] utilized the more robust mixtures of multivariate t-distributions (motivated by McLachlan and Peel [84]) as a heavier-tailed alternative to the Gaussian mixture model for model-based clustering. Equation (2.3.10) represents the form of the density for a mixture of multivariate t-distributions model with G components :

$$p(\mathbf{y}|\mathbf{\Pi}) = \sum_{g=1}^G \pi_g p_g(\mathbf{y}|\boldsymbol{\mu}_g, \boldsymbol{\Gamma}_g, \nu_g) \quad (2.3.10)$$

where $p_g(\mathbf{y}|\boldsymbol{\mu}_g, \boldsymbol{\Gamma}_g, \nu_g)$ is the density of a multivariate t-distribution with mean $\boldsymbol{\mu}_g$, scale matrix $\boldsymbol{\Gamma}_g$ and ν_g degrees of freedom. For data $\mathbf{y} = (\mathbf{y}_1, \dots, \mathbf{y}_N)^\top$ where each \mathbf{y}_i is measured at n time points t_1, \dots, t_n , McNicholas and Subedi [88] use the modified Cholesky decomposition of Equation (2.3.5) on each group scale matrix $\boldsymbol{\Gamma}_g$ and consider a linear model for the group mean of the form $\boldsymbol{\mu}_g = \mathbf{Q}\boldsymbol{\beta}_g$ where

$$\mathbf{Q} = \begin{pmatrix} 1 & 1 & \dots & 1 \\ t_1 & t_2 & \dots & t_n \end{pmatrix}^\top, \quad \boldsymbol{\beta}_g = \begin{pmatrix} a_g \\ b_g \end{pmatrix}. \quad (2.3.11)$$

Like McNicholas and Murphy [87], constraints can be placed upon the group scale matrices, leading to a novel family of eight mixture models. Parameters, including the group degrees of freedom ν_g , are estimated using an EM algorithm. For model selection, McNicholas and Subedi [88] considered two approaches : the BIC criterion and the integrated completed likelihood (ICL) as an alternative to the BIC. The ICL has been proposed by Biernacki et al. [6] and essentially penalizes the BIC for estimated mean entropy, thereby punishing mixture components that are more spread out. An approximate ICL is used in practice and is given by :

$$\text{ICL} \approx \text{BIC} + \sum_{i=\xi+1}^N \sum_{g=1}^G \text{MAP}\{\hat{z}_{ig}\} \log \hat{z}_{ig} \quad (2.3.12)$$

where ξ is the number of free parameters in the model; $\text{MAP}\{\hat{z}_{ig}\}$ is the maximum a posteriori classification given \hat{z}_{ig} , that is $\text{MAP}\{\hat{z}_{ig}\} = 1$ if $\max\{\hat{z}_{ig}\}$ occurs in group g and $\text{MAP}\{\hat{z}_{ig}\} = 0$ otherwise.

According to Ciampi et al. [17], the spectral decomposition of the matrices (parameterization of the scale matrices or the covariance matrices, up to a multiplicative constant) as considered in the papers cited above, is of limited help when analyzing longitudinal data with non negligible correlations, since it does not address the special form that the

variance-covariance matrices may take. Along the same lines with the modeling of the correlation structure, Ciampi et al. [17] studied a model that is a mixture of regressions, with variance-covariance matrices that are allowed to vary within the extended linear mixed model (ELMM, see Pinheiro and Bates [97]) family. Longitudinal data are usually unbalanced : both the number of measurements and the time points may differ across individual units. Let $Y_i(t_{ij})$ be the observation of the i^{th} individual at time t_{ij} for $i = 1, \dots, N$, $j = 1, \dots, n_i$, where N is the total number of individuals and n_i is the number of time points at which the i^{th} individual has been observed. The ELMM can be written as :

$$\begin{cases} \mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i \\ \mathbf{X}_i = \begin{pmatrix} k_1(t_{i1}) & \dots & k_p(t_{i1}) \\ \dots & \dots & \dots \\ k_1(t_{in_i}) & \dots & k_p(t_{in_i}) \end{pmatrix}, \quad \mathbf{Z}_i = \begin{pmatrix} h_1(t_{i1}) & \dots & h_q(t_{i1}) \\ \dots & \dots & \dots \\ h_1(t_{in_i}) & \dots & h_q(t_{in_i}) \end{pmatrix} \\ \boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^\top, \quad \mathbf{b}_i = (b_{i1}, \dots, b_{iq})^\top, \quad \boldsymbol{\epsilon}_i \sim \mathcal{N}_{n_i}(0, \sigma^2\boldsymbol{\Sigma}_i) \end{cases} \quad (2.3.13)$$

where \mathbf{X}_i and \mathbf{Z}_i are design matrices ; \mathbf{b}_i and $\boldsymbol{\epsilon}_i$ are assumed independent ; \mathbf{Y}_i is independent of \mathbf{Y}_j for $i \neq j$ and $\boldsymbol{\Sigma}_i$ is a $n_i \times n_i$ matrix that may depend on i through the time intervals t_{ij} but not otherwise. Typically, $\boldsymbol{\Sigma}_i$ is parameterized in terms of a relatively small number of variance parameters. Furthermore, the distribution of the random effects \mathbf{b}_i is assumed to be $\mathcal{N}_q(0, \mathbf{D})$ where \mathbf{D} is a symmetric positive definite matrix which may depend on parameters to be estimated. The k_i 's and h_i 's denote the elements of a basis in function space and in practice, the columns of \mathbf{Z}_i are often chosen as subset of the columns of \mathbf{X}_i . These descriptions lead to the following distributions for $\mathbf{Y}_i|\mathbf{b}_i$ and \mathbf{Y}_i :

$$\mathbf{Y}_i|\mathbf{b}_i \sim \mathcal{N}_{n_i}(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i, \sigma^2\boldsymbol{\Sigma}_i) \quad \text{and} \quad \mathbf{Y}_i \sim \mathcal{N}_{n_i}(\mathbf{X}_i\boldsymbol{\beta}, \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^\top + \sigma^2\boldsymbol{\Sigma}_i) \quad (2.3.14)$$

The random effects \mathbf{b}_i may be considered as missing data and maximum likelihood estimation is done by the EM algorithm. Under the assumption that N individuals are sampled from G distinct component or group distributions, Ciampi et al. [17] write :

$$p(\mathbf{Y}_i|\mathbf{b}_i) = \sum_{g=1}^G \pi_g \phi(\mathbf{Y}_i | (\mathbf{X}_i\boldsymbol{\beta}_{(g)} + \mathbf{Z}_i\mathbf{b}_{i(g)}), \sigma_{(g)}^2\boldsymbol{\Sigma}_{(g)}) \quad (2.3.15)$$

where the π_g 's are still the mixing coefficients. In the model formulation of Equation (2.3.15), each component g is distinct and uniquely defined by the parameters $\boldsymbol{\beta}_{(g)}$, $\sigma_{(g)}^2$, $\mathbf{D}_{(g)}$ and $\boldsymbol{\Sigma}_{(g)}$. In addition, the distributions in Equation (2.3.14) apply for each couple $(\mathbf{Y}_{i(g)}, \mathbf{b}_{i(g)})$ with the corresponding group parameters and the joint log-likelihood of $(\mathbf{Y}_{i(g)}, \mathbf{b}_{i(g)})$ denoted $\log L_{(g)}(\boldsymbol{\beta}_{(g)}, \sigma_{(g)}^2, \mathbf{D}_{(g)}, \boldsymbol{\Sigma}_{(g)} | \mathbf{y}_i, \mathbf{b}_{i(g)})$ is equal to the log-likelihood of $((\mathbf{Y}_{i(g)}|\mathbf{b}_{i(g)}), \mathbf{b}_{i(g)})$ and is obtained as :

$$\log L_{(g)}(\cdot|\cdot) = -\frac{1}{2} \sum_{i=1}^N \left(\log(|2\pi\mathbf{D}_{(g)}|) + \log(|2\pi\sigma_{(g)}^2\boldsymbol{\Sigma}_{(g)}|) + \mathbf{b}_{i(g)}^\top \mathbf{D}_{(g)}^{-1} \mathbf{b}_{i(g)} \right) \quad (2.3.16)$$

$$+ \frac{(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta}_{(g)} - \mathbf{Z}_i\mathbf{b}_{i(g)})^\top \boldsymbol{\Sigma}_{(g)}^{-1} (\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta}_{(g)} - \mathbf{Z}_i\mathbf{b}_{i(g)})}{\sigma_{(g)}^2}$$

Note that $\mathbf{Y}_{i(g)} = \mathbf{X}_i\boldsymbol{\beta}_{(g)} + \mathbf{Z}_i\mathbf{b}_{i(g)} + \boldsymbol{\epsilon}_{i(g)}$. The log-likelihood of the mixture model is defined as $\log L = \sum_{i=1}^N \log(\sum_{g=1}^G \pi_g \cdot e^{\log L_{(g)}(\cdot|\cdot)})$. As direct maximization of that log-likelihood can be quite difficult due to the sum of terms inside the logarithm, the data can be "completed" by considering the unobserved latent indicator variables z_{ig} which are equal to 1 if the observation i belongs to cluster g and 0 otherwise. Then, the complete data log-likelihood is rewritten as :

$$\log L = \sum_{i=1}^N \sum_{g=1}^G \left\{ z_{ig} \log(\pi_g) + z_{ig} \log L_{(g)}(\boldsymbol{\beta}_{(g)}, \sigma_{(g)}^2, \mathbf{D}_{(g)}, \boldsymbol{\Sigma}_{(g)} | \mathbf{y}_i, \mathbf{b}_{i(g)}) \right\} \quad (2.3.17)$$

The maximum likelihood estimates of the parameter vector $\boldsymbol{\Pi} = (\boldsymbol{\pi}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2, \mathbf{D}, \boldsymbol{\Sigma})$ are obtained using the EM algorithm.

Recently, Coffey et al. [19] have addressed the issue of clustering longitudinal profiles in time-course gene expression data. In the introduction to their approach, they recalled the irrelevance of multivariate clustering methods for time-course gene expression data, due to the fact that such studies result generally in extremely high-dimensional data and exhibit problems such as missing values, unequal sampling times and/or large measurement errors. According to them, the techniques developed to cope with these difficulties such as Bayesian mixture models (Wakefield et al. [123]), mixtures of linear mixed effects models (Celeux et al. [13], Ng et al. [94], Qin and Self [99], Nueda et al. [95]), clustering based on shape similarity (Hestilow and Huang [53]) or clustering of time-course data using self-organizing maps (Chen [16]), do not facilitate the removal of noise from the measured data thus ignoring any smoothness that may be evident in the gene expression profiles. And, that has led to the emergence of curve-based clustering methods which assume that gene expression over time is a continuous process to be represented by a continuous smooth curve or function. Some of the earliest papers describing the curve-based methods include Bar-Joseph et al. [4], Luan and Li [78], James and Sugar [60], Leng and Müller [69], Song et al. [112], Kim et al. [63], Kim and Kim [64]¹.

Coffey et al. [19] emphasized that the estimation of the cluster mean curves in the curve-based methods requires choosing an optimal number of basis functions (or the joint points for these functions, called knots) and that is a complex problem essentially due to the difficulty to control the degree of smoothing applied to the data. One solution, as implemented in papers such as Ma et al. [79], Déjean et al. [21], Ma et al. [81], Ma and Zhong [80], is

1. These methods are reviewed in the next section.

to use smoothing splines regression, where a knot is placed at each unique time point and the resulting over-fitting is controlled by adding a penalty term to the optimization criterion. However, a major drawback of clustering using smoothing spline regression is the high computational overhead associated with these methods (numerical evaluation of the integral associated with the penalty term, choice of an optimal value for the smoothing parameter λ for each cluster).

Coffey et al. [19] used, in contrast, penalized splines (P-splines) smoothing (Eilers and Marx [26], Ruppert et al. [106]) to model the gene expression profiles in each cluster. Indeed, Coffey et al. [19] proposed an alternative method that exploits the connection between the linear mixed effects model and P-spline (low-rank smoothing, moderate number of basis functions, easy to compute since the penalty is discrete) to simultaneously smooth the gene expression data to remove any measurement error/noise and cluster the expression profiles using finite G -mixtures of mixed effects models. The observed gene expression data for a single gene measured at a discrete number n of time points t_j ($j = 1, \dots, n$) is modeled as

$$y_j = \theta(t_j) + \varepsilon_j, \quad (2.3.18)$$

where $\theta(t_j)$ is the value at time point t_j of the smooth expression profile $\theta(\cdot)$ and ε_j is measurement error. While other basis functions such as B-splines are possible, Coffey et al. [19] uses for demonstration purpose, the expression of $\theta(\cdot)$ in a p^{th} degree truncated power basis with M knots $\kappa_1, \dots, \kappa_M$ as : $[\theta(t_j) = \beta_0 + \beta_1 t_j + \dots + \beta_p t_j^p + \sum_{m=1}^M \beta_{1m} (t_j - \kappa_m)_+^p]$ and re-writes Equation (2.3.18) as a linear regression model to estimate the coefficients β_l .

$$\begin{cases} \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}; & \boldsymbol{\varepsilon} \sim \mathcal{N}_n(0, \sigma_\varepsilon^2 \mathbf{I}); & \boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p, \beta_{11}, \dots, \beta_{1M})^\top; \\ \mathbf{X} = \begin{pmatrix} 1 & t_1 & \dots & t_1^p & (t_1 - \kappa_1)_+^p & \dots & (t_1 - \kappa_M)_+^p \\ 1 & t_2 & \dots & t_2^p & (t_2 - \kappa_1)_+^p & \dots & (t_2 - \kappa_M)_+^p \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & t_n & \dots & t_n^p & (t_n - \kappa_1)_+^p & \dots & (t_n - \kappa_M)_+^p \end{pmatrix} \end{cases} \quad (2.3.19)$$

For the estimation of the coefficients, the proposed method uses the P-splines smoothing approach to choose a relatively large number of basis functions by choosing the number of knots $[M = \max(5, \min(\frac{n}{4}, 35))]$ placed at the quantiles of the data, and introduces a ridge penalty term in the fitting criterion to account for over-fitting. Note that the number of basis functions depends on the number of knots. The formula for choosing M is a rule-of-thumb provided by Ruppert [108]. The fitting criterion is then to minimize $\text{PRSS} = [||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2 + \lambda \boldsymbol{\beta}^\top \boldsymbol{\Omega} \boldsymbol{\beta}]$ where $\boldsymbol{\Omega}$ is a penalty matrix penalizing the basis function coefficients and λ is a tuning parameter. In the case of the p^{th} degree truncated power basis, only the coefficients of the truncated line basis functions are penalized and the matrix $\boldsymbol{\Omega}$ is

chosen so that $[\boldsymbol{\beta}^\top \boldsymbol{\Omega} \boldsymbol{\beta} = \sum_{m=1}^M \beta_{1m}^2]$.

Further in the proposed method, Coffey et al. [19] relates the fitting criterion to a linear mixed effects model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\varepsilon}$ by letting the fixed effects design matrix \mathbf{X} be consisted of the first $(p+1)$ basis functions and the random effects design matrix \mathbf{Z} be consisted of the remaining M basis functions. The vector of fixed effects is $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)^\top$ and the vector of random effects is $\mathbf{b} = (b_1, b_2, \dots, b_M)^\top$. The minimization criterion for the penalized smoothing problem can then be expressed as $\text{PRSS} = [||\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{Z}\hat{\mathbf{b}}||^2 + \lambda ||\hat{\mathbf{b}}||^2]$ where $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{b}}$ are the best linear unbiased predictors (BLUP) of the mixed model :

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\varepsilon}; \quad \mathbf{b} \sim \mathcal{N}_M(0, \sigma_b^2 \mathbf{I}); \quad \boldsymbol{\varepsilon} \sim \mathcal{N}_n(0, \sigma_\varepsilon^2 \mathbf{I}). \quad (2.3.20)$$

Note that $[\sigma_b^2 = \sigma_\varepsilon^2 / \lambda]$; $[Var(\mathbf{y}) = (\sigma_b^2 \mathbf{Z}\mathbf{Z}^\top + \sigma_\varepsilon^2 \mathbf{I})]$; \mathbf{b} and $\boldsymbol{\varepsilon}$ are assumed to be mutually independent. Further, that smoothing in a mixed model framework for a single gene has been generalized to model gene expression clusters with gene-specific shifts around the cluster mean. Assuming that a gene i is known to be in cluster g , its expression profile is written as :

$$y_{ij} = \mu_g(t_{ij}) + \tilde{b}_i + \varepsilon_{ij}, \quad i = 1, \dots, N_g; \quad j = 1, \dots, n_i \quad (2.3.21)$$

where $\mu_g(t)$ is the mean expression curve in cluster g ; $\tilde{b}_i \sim \mathcal{N}(0, \sigma_{bg}^2)$ is an additional random effect to allow for gene-specific shifts from that mean curve; $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma_{\varepsilon g}^2)$; N_g is the total number of genes in cluster g ; n_i is the number of measures for gene i . The curve $\mu_g(t)$ and the \tilde{b}_i for all N_g genes in that cluster are estimated by stacking the data from the N_g genes and using the linear mixed effects model representation of P-splines such that :

$$\left\{ \begin{array}{l} \mathbf{Y}_g = \underbrace{\mathbf{X}_{g,s} \boldsymbol{\beta}_{g,s} + \mathbf{Z}_{g,s} \mathbf{b}_{g,s}}_{\mu_g(t)} + \tilde{\mathbf{Z}}_{g,b} \tilde{\mathbf{b}}_g + \boldsymbol{\varepsilon}_g; \\ \mathbf{Y}_g = (\mathbf{y}_1^\top, \mathbf{y}_2^\top, \dots, \mathbf{y}_{N_g}^\top)^\top; \quad \mathbf{y}_i = (y_{i1}, y_{i2}, \dots, y_{in_i})^\top \\ \mathbf{X}_{g,s} = (\mathbf{X}_{1,s}, \mathbf{X}_{2,s}, \dots, \mathbf{X}_{N_g,s})^\top; \quad \mathbf{Z}_{g,s} = (\mathbf{Z}_{1,s}, \mathbf{Z}_{2,s}, \dots, \mathbf{Z}_{N_g,s})^\top \\ \mathbf{X}_{i,s} = \begin{pmatrix} 1 & t_{i1} & \cdots & t_{i1}^p \\ 1 & t_{i2} & \cdots & t_{i2}^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & t_{in_i} & \cdots & t_{in_i}^p \end{pmatrix}; \quad \mathbf{Z}_{i,s} = \begin{pmatrix} (t_{i1} - \kappa_1)_+^p & \cdots & (t_{i1} - \kappa_M)_+^p \\ (t_{i2} - \kappa_1)_+^p & \cdots & (t_{i2} - \kappa_M)_+^p \\ \vdots & \ddots & \vdots \\ (t_{in_i} - \kappa_1)_+^p & \cdots & (t_{in_i} - \kappa_M)_+^p \end{pmatrix} \\ \tilde{\mathbf{Z}}_{g,b} = \text{diag}(\tilde{\mathbf{Z}}_{1,b}, \tilde{\mathbf{Z}}_{2,b}, \dots, \tilde{\mathbf{Z}}_{N_g,b}); \quad \tilde{\mathbf{Z}}_{i,b} = (1, 1, \dots, 1)^\top \\ \mathbf{b}_{g,s} \sim \mathcal{N}_M(0, \sigma_{bg}^2 \mathbf{I}); \quad \tilde{\mathbf{b}}_g \sim \mathcal{N}_{N_g}(0, \tilde{\sigma}_{bg}^2 \mathbf{I}); \quad \boldsymbol{\varepsilon}_g \sim \mathcal{N}_{(\sum_{i=1}^{N_g} n_i)}(0, \sigma_{\varepsilon g}^2 \mathbf{I}). \end{array} \right. \quad (2.3.22)$$

The vectors $\mathbf{b}_{g,s}$, $\tilde{\mathbf{b}}_g$ and $\boldsymbol{\varepsilon}_g$ are assumed to be independent. The model for gene i in cluster g can be written as

$$\mathbf{y}_i = \underbrace{\mathbf{X}_{i,s}\boldsymbol{\beta}_{g,s} + \mathbf{Z}_{i,s}\mathbf{b}_{g,s}}_{\mu_g(\mathbf{t}_i)} + \tilde{\mathbf{Z}}_{i,b}\tilde{\mathbf{b}}_i + \boldsymbol{\varepsilon}_i; \quad (2.3.23)$$

where $\mathbf{X}_{i,s}$ of dimension $(n_i, p + 1)$, $\mathbf{Z}_{i,s}$ of dimension (n_i, M) and $\tilde{\mathbf{Z}}_{i,b}$ of dimension $(n_i, 1)$ are the design submatrices for the i^{th} gene and \mathbf{t}_i is the vector of measurement time points for gene i .

In practice, the cluster membership is unknown and it is assumed that \mathbf{y}_i comes from a mixture of G components such that $\{p(\mathbf{y}_i|\boldsymbol{\Pi}) = \sum_{g=1}^G \pi_g p_g(\mathbf{y}_i|\boldsymbol{\Pi}_g)\}$ where $p_g(\cdot)$ are the component densities that depend on the vector of unknown parameters $\boldsymbol{\Pi}_g = (\boldsymbol{\beta}_{g,s}, \sigma_{bg}^2, \tilde{\sigma}_{bg}^2, \sigma_{\varepsilon_g}^2)$. Also $\boldsymbol{\Pi} = (\boldsymbol{\Pi}_1^T, \dots, \boldsymbol{\Pi}_G^T, \pi_1, \dots, \pi_G)$. Since the EM algorithm can be quite unstable in high-dimensional settings, a modification of the standard EM algorithm called rejection-controlled EM (RCEM) is employed for parameter estimation. RCEM is described by Liu et al. (1998) and Ma et al. (2006), who use the algorithm to speed up and stabilize the standard EM algorithm. For model selection, the model-fitting process is repeated for varying values of G (number of components for the mixture) and the solution with minimum BIC is chosen. As mentioned in Coffey et al. [19], the proposed methodology was presented in the context of time-course gene expression data, but it can be applied to any longitudinal dataset where cluster analysis is required.

2.3.3.2. Clustering methods using the functional data analysis approach

In longitudinal studies, measurements collected at different time points for a single subject can be seen as trajectories (Genolini and Falissard [41]). Indeed, longitudinal data are usually lined up to trajectories based on time and Functional data analysis (FDA) is a form of longitudinal analysis that is used to model such trajectory/trend patterns in time. Clustering functional data has received particular attention in the last decade, notably since the idea that FDA methods can be a very useful complement to the tools for the analysis of longitudinal data has been exposed in Zhao et al. [134]. As an illustration, one of the major applications of the FDA approach as highlighted in Ullah and Finch [118] is an apparent increasing interest in clustering. Consideration of clustering problems using the FDA setting provides ways to take time dependency into account by using tools such as basis function expansion or functional principal component analysis (FPCA) to describe the partially observed curves.

Let's consider a longitudinal study involving N subjects, and assume that n_i measurements were collected for subject i , $i = 1, \dots, N$. In the functional data analysis approach (Ramsay and Silverman [101]), for each subject $i \in \{1, \dots, N\}$, the observed measurements

$\mathbf{y}_i = (y_{i1}, \dots, y_{in_i})$ are assumed to be realizations with measurement errors of a random function \mathcal{X}_i at times $\mathbf{t}_i = (t_{i1}, \dots, t_{in_i})$ according to the model :

$$y_{ij} = \mathcal{X}_i(t_{ij}) + \epsilon_{ij} \quad j = 1, \dots, n_i \quad i = 1, \dots, N; \quad (2.3.24)$$

where the ϵ_{ij} 's are independent and identically distributed, with moments $E(\epsilon_{ij}) = 0$ and $E(\epsilon_{ij}^2) = \sigma^2$. The functions $\{\mathcal{X}_1, \dots, \mathcal{X}_N\}$ are assumed to be independent realizations of a second order stochastic process $\mathcal{X}(t)$ defined on a compact domain \mathcal{I} (the temporal axis), with mean $\mu(t) = E(\mathcal{X}(t))$ and covariance function $V(s, t) = \text{Var}(\mathcal{X}(s), \mathcal{X}(t))$ ($t, s \in \mathcal{I}$).

As stated in Jacques and Preda [59], the main source of difficulty when dealing with functional data consists in the fact that the observations are supposed to belong to an infinite dimensional space, whereas in practice one only has sampled curves observed into a finite set of time points. Due to that fact, the first step in FDA is often the reconstruction of the functional form of the data from discrete observations, through the expansion of the sampled curves in a finite dimensional space spanned by some basis of functions. Indeed, the dimension reduction consists generally in approximating the curves into a finite basis of functions (such as B-splines, wavelet basis, Fourier basis), or using FPCA. From a computational point of view, one generally needs to use also a basis approximation of the curves in the case of FPCA.

In the case of the basis function approach, the dimension reduction results in the expression

$$\mathcal{X}_i(t_{ij}) = \sum_{l=1}^L \alpha_{il} B_l(t_{ij}) = \mathbf{B}_i \boldsymbol{\alpha}_i \quad (2.3.25)$$

where $\{B_l\}_{l=1}^L$ is a set of basis functions and $\{\alpha_{il}\}_{l=1}^L$ is a set of the corresponding coefficients for the i^{th} curve, with $\boldsymbol{\alpha}_i = (\alpha_{i1}, \dots, \alpha_{iL})^\top$ and $\mathbf{B}_i = \{B_l(t_{ij})\}_{1 \leq j \leq n_i; 1 \leq l \leq L}$. As presented in Song et al. [112], in the basis function approach, three types of computational issues need to be addressed : (a) choosing an appropriate type of basis function, (b) determine the number of basis functions, and (c) computing the best linear combination. The least squares approach is a standard method to determine the approximating basis expansion by minimizing the sum of squares $\|\mathbf{y}_i - \mathbf{B}_i \boldsymbol{\alpha}_i\|^2$ which leads to the estimates $[\hat{\boldsymbol{\alpha}}_i = (\mathbf{B}_i^\top \mathbf{B}_i)^{-1} \mathbf{B}_i^\top \mathbf{y}_i]$. Regarding the FPCA, it has become a major tool in FDA to achieve dimension reduction, by reducing random trajectories to a set of k FPC scores. Each individual curve can be expressed as in Equation (2.3.26) which results from the Karhunen-Loeve expansion of a second-order L^2 -continuous stochastic process \mathcal{X} in Equation (2.3.27).

$$\mathcal{X}_i(t) = \hat{\mu}(t) + \sum_{j=1}^k \alpha_{ij} \hat{f}_j(t) \quad (2.3.26)$$

$$\mathcal{X}(t) = \hat{\mu}(t) + \sum_{j \geq 1} \alpha_j \hat{f}_j(t), \quad t \in \mathcal{T}, \quad (2.3.27)$$

where $\hat{\mu}(t)$ is the mean function, $\hat{f}_j(\cdot)$ are the functional principal components (FPC) and the α_{ij} are the FPC scores. See Ramsay and Silverman [101], Leng and Müller [69], Jacques and Preda [59] among others, for computational methods related to FPCA and the Karhunen-Loeve expansion. However, as discussed in Yao et al. [129], this method encounters difficulties when applied to sparse longitudinal data, with measurements collected at only few different times per subject. They develop a version of FPCA referred as principal components analysis through conditional expectation (PACE) for longitudinal data. In that method, the FPC scores are framed as conditional expectations and the authors demonstrated that it extends the applicability of FPCA to situations in longitudinal data analysis, where only few and sufficiently irregularly spaced measurements are available per subject.

Jacques and Preda [59] reviewed the main contributions to functional data clustering, and particularly the **three methodologies** on which are based most approaches used for clustering functional data : (1) dimension reduction before clustering, (2) nonparametric methods and (3) model-based clustering methods. In the first methodology also named two-stage methods, functional data are summarized either by their coefficients in a basis of functions or by their first principal component scores and then, usual clustering algorithms are used to estimate the clusters. The second methodology uses specific distances or dissimilarities between curves and the third methodology assumes a probabilistic distribution on either the principal components (modeling the FPC scores) or the coefficients of functional data expansion into a finite dimensional basis of functions. In the latter, contrary to the two-stage methods in which the estimation of these coefficients is done before clustering, these two tasks are performed simultaneously with model-based techniques.

Besides the functional clustering methods surveyed in Jacques and Preda [59], few other clustering methods have been proposed specifically for functional data, and some of them have found successful applications to time course microarray data. In gene expression data, Luan and Li [78] used linear combinations of basis functions to model the mean expression profile in each cluster and cluster the estimated basis function coefficients. They proposed the following mixed-effects model for the observed expression level at time t_{ij} for the gene i in cluster $g = 1, \dots, G$:

$$\mathcal{X}_i(t_{ij}) = \left(\sum_{l=1}^{L_1} \bar{\alpha}_l^{(g)} \bar{B}_l(t_{ij}) \right) + \left(\sum_{l=1}^{L_2} \alpha_{il} B_l(t_{ij}) \right) + \epsilon_{ij} \quad (2.3.28)$$

where the first term is used to model the mean average expression profile of the g^{th} cluster with a basis of B-splines $\bar{B} = \{\bar{B}_l(), l = 1, \dots, L_1\}$ for all the G clusters, and the second term is used to model the random effect curve for the i^{th} curve with a basis of B-splines $B = \{B_l(), l = 1, \dots, L_2\}$ possibly different from \bar{B} . The α_{il} are normal random coefficients with mean 0 and covariance matrix $Var(\gamma_i) = \mathbf{\Gamma}$ varying across genes. Finally, the last term is used to model the uncorrelated normal measurement errors ϵ_{ij} with $E(\epsilon_{ij}) = 0$, $Var(\epsilon_{ij}) = \sigma^2$. The clustering was based on a mixture model using an EM algorithm.

Bar-Joseph et al. [4] independently developed a similar model using the same cubic spline basis for both mean and random effects. They also present a discussion on the differences between cubic and B-splines.

A curve-based clustering method called functional clustering model (FCM) has been introduced in James and Sugar [60] to cluster sparsely sampled time course data. In their specification, gene curve i , given cluster membership g , is modeled as

$$\mathbf{y}_i = \mathbf{B}(\mathbf{t})(\boldsymbol{\alpha}_0 + \mathbf{\Lambda}\boldsymbol{\alpha}_g + \mathbf{b}_i) + \boldsymbol{\epsilon}_i \quad (2.3.29)$$

where \mathbf{t} is a uniform time grid $\mathbf{t} = (t_1, \dots, t_j, \dots, t_n)^\top$; n is the number of sampling points; $\mathbf{y}_i = (y_{i1}, \dots, y_{ij}, \dots, y_{in})^\top$; $\mathbf{B}(\mathbf{t})$ is the $(n \times L)$ spline basis matrix; $\boldsymbol{\alpha}_0$ is the basis coefficient vector for the overall shape function; $\boldsymbol{\alpha}_g$ is the q -dimensional basis coefficient vector for the g^{th} cluster shape function and $\mathbf{\Lambda}$ is the $(L \times q)$ transition matrix to reduce the parameter dimension from L to q with $q \leq \min(L, G - 1)$. The model admits random individual specific coefficients $\mathbf{b}_i \sim \mathcal{N}_L(0, \mathbf{D})$ and errors $\boldsymbol{\epsilon}_i \sim \mathcal{N}_n(0, \sigma^2 \mathbf{I}_n)$. With \mathbf{b}_i integrated out, the marginal sampling model for \mathbf{y}_i is $\{\mathbf{y}_i \sim \sum_{g=1}^G \pi_g \mathcal{N}_n(\mathbf{B}_i(\boldsymbol{\alpha}_0 + \mathbf{\Lambda}\boldsymbol{\alpha}_g), \boldsymbol{\Sigma}_i)\}$ where π_g is the cluster g membership probability, \mathbf{B}_i is the spline basis matrix evaluated on the time grid and $[\boldsymbol{\Sigma}_i = \sigma^2 \mathbf{I}_n + \mathbf{B}_i \mathbf{D} \mathbf{B}_i^\top]$. Under several identifiability conditions, the model is fitted via the EM algorithm and the number of clusters G can be determined through model selection using an alternative approach based on a "distorsion function" and suggested by Sugar and James [116].

In the same vein, Leng and Müller [69] has represented the expression profiles using a linear combination of functional principal components and performed functional logistic regression of the scores to classify the expression profiles into clusters. Song et al. [112] determined the FPC using basis functions expansion and clustered based on the FPC scores. Kim et al. [63] used a linear combination of Fourier basis functions to represent the expression profiles for clustering. Kim and Kim [64] clustered based on the derivative coefficients of a Fourier series.

Ma et al. [79] developed the smoothing splines clustering (named SSClust method) in which

gene expression curves are modeled to include random intercepts within a finite location mixture. Specifically, for gene i at time t_j , given cluster membership g , it is assumed that

$$y_{ij} = \mu_g(t_j) + b_i + \epsilon_{ij} \quad (2.3.30)$$

where $\mu_g(\cdot)$ is the g^{th} cluster specific shape function of time, $b_i \sim \mathcal{N}(0, \sigma_{bg}^2)$ is the gene specific random intercept and $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ is the error term. With b_i integrated out, the sampling model for the observed vector $\mathbf{y}_i = (y_{i1}, \dots, y_{in})^\top$ can be written as a finite mixture $\{\mathbf{y}_i \sim \sum_{g=1}^G \pi_g \mathcal{N}_n(\boldsymbol{\mu}_g, \boldsymbol{\Gamma}_g)\}$, where $\boldsymbol{\Gamma}_g = \sigma_{bg}^2 \mathbf{1}_n \mathbf{1}_n^\top + \sigma^2 \mathbf{I}_n$ and $\boldsymbol{\mu}_g = (\mu_g(t_1), \dots, \mu_g(t_n))^\top$. The estimates of $\boldsymbol{\mu}_g$ are obtained by maximizing a penalized likelihood function and the number of clusters G is determined via BIC. Ma et al. [81] extended that model to a Bayesian setting and Ma and Zhong [80] included additional covariates in the clustering algorithm. Déjean et al. [21] used smoothing splines regression to estimate the derivatives of the gene expression profiles before clustering based on the principal component scores of the discretized derivative functions. More recent papers in the field of clustering functional data include Suarez and Ghosal [115], Hasenstab et al. [48] and Ciollaro et al. [18].

The recent works of Adjogou et al. [1] result in the development of a flexible and Bayesian-embedded model based on B-splines in which the clusters are modeled by a mixture of Student t-distributions. The proposed model combines functional principal components analysis and clustering to deal with any type of longitudinal data even if the observations are sparse, irregularly spaced or occur at different time points for each individual.

2.3.3.3. Software-implemented clustering methods

Due to computational advances, many clustering methods for longitudinal data have been implemented in specific packages. The most popular ones associated with the model-based approach are : the longitudinal mixture modeling analysis procedure in SAS named *Proc Traj* (see Jones et al. [62]; Jones and Nagin [61]); *FlexMix* in the software R which implements a general framework for finite mixtures of regression models using the EM algorithm and allowing the modeling of longitudinal trajectories (see Leisch [68]; Gruen and Leisch [46]); *Funclust* from the Funclustering package in R (see Jacques and Preda [57, 58]) and *fclust* in R (available directly from James's webpage, James and Sugar [60]). The procedure *SSSclust* (Ma et al. [79]) is also implemented in the R package Model Based Functional Data Analysis. *Mplus* (Muthen and Muthen [93]) is also a statistical software that provides a general framework that can deal with mixture modeling on longitudinal data. Regarding the non-parametric approach, *KmL* (see Genolini and Falissard [40]; Genolini and Falissard

[41]) is an implementation of k-means method designed in R to work specifically on longitudinal data. It deals with missing values and runs the algorithm several times, varying the starting conditions and/or the number of clusters sought.

2.3.3.4. *Some comparison studies among longitudinal data clustering methods*

Many studies related to clustering longitudinal data have focused on comparing one method to another. Amongst them, Genolini and Falissard [40] compared the performances of *KmL* to *Proc Traj* based on artificial and real data. According to their report, the two techniques give very close clustering results when trajectories follow polynomial curves and *KmL* gives much better results on non-polynomial trajectories. Similarly, Dufour [25] compared the k-means method (with their own implementation) to a model-based method (implemented using *FlexMix* in R) according to the ability to correctly classify subject trajectories into groups (*Correct Classification Rate, CCR*). The results based on a simulation study revealed that both methods are found to perform well under most circumstances, but in 64% of the scenarios examined, the model-based method outperforms the k-means approach. For Schramm et al. [109], the main issues with the current methods in the context of longitudinal cluster analyses are sample size and variability of the times of measurements. After recalling the increasing use of the current parametric and non-parametric methods in medical research, they discussed some limitations of those methods. These limitations suggest the need for a new method to take into account the treatment effect when there is both a small sample and variability in the times of measurement. For that purpose, they propose a clustering of longitudinal data with an extended baseline (CLEB method) comprising two steps : first, building a linear mixed model with an extended baseline and second clustering the random predictions through a model-based or a non-parametric algorithm.

2.4. CONCLUSIONS AND DISCUSSION

In this paper, we present an overview of the main methodologies of clustering longitudinal data. We also make a brief and concise presentation of methods to analyze such data. As discussed throughout the paper, the field of longitudinal data analysis in general and particularly the one of longitudinal cluster analysis is in constant evolution and requires insight from various other fields such as high-dimensional and functional data analysis. Although most of the recent developments in the field of longitudinal cluster analysis have been presented, the current review does not purport to be exhaustive. The topic of longitudinal data analysis is related to so many issues that an exhaustive review needs to cover too many fields of statistics.

Broadly, apart from the category (non-parametric or model-based) in which each method

can be classified, the significant differences among the clustering methods are related to various issues such as the structure of the data targeted (balanced, unbalanced or both), the type of the outcome (continuous, discrete), the assumptions underlying the model (for example normal or Student distributions for the mixtures; constraints on the covariance matrices) or the degree of smoothing applied to the data. The current review has the advantage to enumerate a variety of clustering methods for longitudinal data but for practical reasons, data analysts are more likely to choose the software-implemented methods. Thus, it would be a wise move for researchers and specialists in the domain to facilitate as much as possible the implementation of the clustering methods.

Longitudinal data involve measuring one or many outcome variables on a relatively big set of individuals repeatedly through time and usually result in high-dimensional data. Assent [2] provided an overview of the effects of high-dimensional spaces and their implications for different clustering paradigms. They also reviewed models and algorithms that address clustering in high dimensions, with pointers to the literature, and sketched open research issues. According to Bouveyron and Brunet-Saumard [7], classical model-based clustering methods show a disappointing behavior in high-dimensional spaces and that is mainly due to the fact that they are dramatically over-parametrized in that case (the well-known *curse of dimensionality* introduced by Bellman [5]). Focusing on model-based clustering of high-dimensional data, they reviewed recent works in dimension reduction approaches, regularization-based techniques, parsimonious modeling, subspace clustering methods and clustering methods based on variable selection.

On another issue, Ren [102] discussed the fact that in longitudinal studies one is often interested in simultaneously clustering observations at both subject and time-levels. The goal is to cluster subjects with similar profiles, and within each subject-level cluster, one wants also to cluster the consecutive time points such that the profiles during those periods are relatively stable. For that specific purpose, they presented a *non-parametric Bayesian method (Dirichlet process mixture model)* to hierarchically cluster both subjects and consecutive time points for a longitudinal data by defining a specific base measure. The Gibbs sampler, a well-known MCMC algorithm, is implemented for the Bayesian posterior distributions and estimates.

It is worth noting that some aspects of clustering longitudinal data such as the clustering of joint trajectories (as in Genolini et al. [42]) or multivariate longitudinal data analysis (as in Verbeke et al. [120]) are beyond the scope of this paper. Indeed, it is increasingly likely (due to budget constraints for example) that a longitudinal study involves more than

one outcome (of the same or different types) collected repeatedly on a set of individuals. And it might be a challenge to analyze those variables jointly or cluster the so-called joint-trajectories. Furthermore, the current review does not thoroughly address the increasingly important challenge of clustering high-dimensional longitudinal data especially the case of brain images.

Bibliographie

- [1] Adjogou, F., K. Dorman, and A. Murua (2017). Functional model-based clustering for longitudinal data. Article to be submitted.
- [2] Assent, I. (2012). Clustering high dimensional data. *WIREs Data Mining and Knowledge Discovery, Issue 4 2*, 340–350.
- [3] Banfield, J. D. and A. E. Raftery (1993). Model-based gaussian and non-gaussian clustering. *Biometrics 49*, 803–821.
- [4] Bar-Joseph, Z., G. Gerber, D. Gifford, T. Jaakkola, and I. Simon (2003). Continuous representations of time-series gene expression data. *Journal of Bioinformatics and Computational Biology 10*, 341–356.
- [5] Bellman, R. (1957). *Dynamic Programming*. Princeton University Press.
- [6] Biernacki, C., G. Celeux, and G. Govaert (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence 22(7)*, 719–725.
- [7] Bouveyron, C. and C. Brunet-Saumard (2014). Model-based clustering of high-dimensional data : A review. *Computational Statistics and Data Analysis 71*, 52–78.
- [8] Bruckers, L. (2014). *Challenges in Cluster Analyses for Longitudinal Data*. Ph. D. thesis, Interuniversity Institute for Biostatistics and statistical Bioinformatics.
- [9] Brumback, B. and J. Rice (1998). Smoothing spline models for the analysis of nested and crossed samples of curves. *Journal of the American Statistical Association 93*, 961–976.
- [10] Capra, W. B. and H. G. Müller (1997). An accelerated-time model for response curves. *Journal of the American Statistical Association 92*, 72–83.
- [11] Celebi, E. (2014). *Partitional Clustering Algorithms*. Springer, 2014.
- [12] Celeux, G. and G. Govaert (1995). Gaussian parsimonious clustering models. *Pattern Recognition 28*, 781–793.
- [13] Celeux, G., O. Martin, and C. Lavergne (2005). Mixture of linear mixed models for clustering gene expression profiles from repeated microarray experiments. *Statistical Modelling 5*, 243–267.

- [14] Chan, Z. S., L. Collins, and N. Kasabov (2006). An efficient greedy k-means algorithm for global gene trajectory clustering. *Expert Systems with Applications* 30, 137–141.
- [15] Chen, K. and Z. Jin (2005). Local polynomial regression analysis of clustered data. *Biometrika* 92, 59–74.
- [16] Chen, X. (2009). Curve-based clustering of time course gene expression data using self-organizing maps. *Journal of Bioinformatics and Computational Biology* 7, 645–661.
- [17] Ciampi, A., H. Campbell, A. Dyachenko, B. Rich, J. McCusker, and M. G. Cole (2012). Model-based clustering of longitudinal data : Application to modeling disease course and gene expression trajectories. *Communications in Statistics-Simulation and Computation* 41, 992–1005.
- [18] Ciollaro, M., C. R. Genovese, and D. Wang (2016). Nonparametric clustering of functional data using pseudo-densities. eprint arXiv :1601.07872.
- [19] Coffey, N., J. Hinde, and E. Holian (2014). Clustering longitudinal profiles using p-splines and mixed effects models applied to time-course gene expression data. *Computational Statistics and Data Analysis* 71, 14–29.
- [20] De la Cruz-Mesia, R., F. A. Quintana, and G. Marshall (2008). Model-based clustering for longitudinal data. *Computational Statistics and Data Analysis* 52, 1441–1457.
- [21] Déjean, S., G. Martin, A. Baccini, and P. Besse (2007). Clustering time-series gene expression data using smoothing spline derivatives. EURASIP Journal on Bioinformatics and Systems Biology. Article ID 70561.
- [22] Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the em algorithm (with discussion). *Journal of the Royal Statistical Society* 39, 1–38.
- [23] Diggle, P. J., P. J. Heagerty, K.-Y. Liang, and S. L. Zeger (2002). *Analysis of Longitudinal Data, 2nd Ed.* Oxford : Oxford University Press.
- [24] Diggle, P. J. and A. P. Verbyla (1998). Nonparametric estimation of covariance structure in longitudinal data. *Biometrics* 54, 401–415.
- [25] Dufour, A. B. (2013). *Cluster analysis of longitudinal trajectories*. Ph. D. thesis, Boston University, Graduate School of Arts and Sciences.
- [26] Eilers, P. H. and B. D. Marx (1996). Flexible smoothing with b-splines and penalties (with discussion). *Statistical Science* 11, 89–121.
- [27] Elavarasi, S., J. Akilandeswari, and B. Sathiyabhama (2011). A survey on partitioning clustering algorithms. *International Journal of Enterprise Computing and Business Systems, Issue1. 1.*
- [28] Everitt, B., S. Landau, M. Leese, and D. Stahl (2011). *Cluster Analysis*. Wiley Series in Probability and Statistics. Wiley.

- [29] Fan, J. and I. Gijbels (1996). *Local Polynomial Modelling and Its Applications*. London, Chapman Hall.
- [30] Fan, J., T. Huang, and R. Li (2007). Analysis of longitudinal data with semiparametric estimation of covariance function. *Journal of the American Statistical Association* 102, 632–641.
- [31] Fitzmaurice, G., M. Davidian, G. Verbeke, and G. Molenberghs (2008). *Longitudinal Data Analysis*. Chapman & Hall CRC Handbooks of Modern Statistical Methods.
- [32] Fitzmaurice, G. M., N. M. Laird, and J. H. Ware (2004). *Applied Longitudinal Analysis*. John Wiley and Sons, New York.
- [33] Fraley, C. and A. E. Raftery (1993). Model based gaussian and non gaussian clustering. *Biometrics* 49, 803–821.
- [34] Fraley, C. and A. E. Raftery (1998). How many clusters? which clustering methods? answers via model-based cluster analysis. *The Computer Journal* 41, 578–588.
- [35] Fraley, C. and A. E. Raftery (1999). Mclust : Software for model-based cluster analysis. *Journal of Classification* 16(2), 297–306.
- [36] Fraley, C. and A. E. Raftery (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association* 97(458), 611–632.
- [37] Fraley, C. and A. E. Raftery (2003). Enhanced software for model-based clustering, density estimation, and discriminant analysis : Mclust. *Journal of Classification* 20, 263–286.
- [38] Fraley, C. and A. E. Raftery (2006). Mclust version 3 for r : Normal mixture modeling and model-based clustering. Technical Report 504. Department of Statistics, University of Washington.
- [39] Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (1995). *Bayesian Data Analysis*. London, Chapman Hall.
- [40] Genolini, C. and B. Falissard (2010). Kml : K-means for longitudinal data. *Computational Statistics* 25, 317–332.
- [41] Genolini, C. and B. Falissard (2011). Kml : A package to cluster longitudinal data. *Computer Methods and Programs in Biomedicine* 104, 34.
- [42] Genolini, C., J. B. Pingault, T. Driss, S. Côté, R. E. Tremblay, F. Vitaro, C. Arnaud, and B. Falissard (2013). Kml3d : A non-parametric algorithm for clustering joint trajectories. *Computer Methods and Programs in Biomedicine* 109, 104–111.
- [43] Ghahramani, Z. and G. Hinton (1997). The EM algorithm for factor analyzers. Technical report. Technical Report CRG-TR-96-1, University of Toronto, Toronto.
- [44] Gibbons, R., D. Hedeker, and S. DuToit (2010). Advances in analysis of longitudinal data. *Annual Review of Clinical Psychology* 6, 79–107.

- [45] Green, P. J. and B. Silverman (1994). *Nonparametric Regression and Generalized Linear Models A Roughness Penalty Approach*. London, Chapman Hall.
- [46] Gruen, B. and F. Leisch (2008). Flexmix version 2 : Finite mixtures with concomitant variables and varying and constant parameters. *Journal of Statistical Software* 28, 1–35.
- [47] Hartigan, J. and M. Wong (1979). A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 28, 100–108.
- [48] Hasenstab, K., C. Sugar, D. Telesca, S. Jeste, and D. Senturk (2016). Robust functional clustering of erp data with application to a study of implicit learning in autism. *Biostatistics Advance access*, 1–15.
- [49] He, Y. (2014). *Bayesian Cluster Analysis with Longitudinal Data*. Ph. D. thesis, University of California, Irvine.
- [50] Hedeker, D. and R. Gibbons (2006). *Longitudinal Data Analysis*. Wiley Series in Probability and Statistics.
- [51] Heggseth, B. C. (2013). *Longitudinal Cluster Analysis with Applications to Growth Trajectories*. Ph. D. thesis, University of California, Berkeley.
- [52] Hennig, C., M. Meila, F. Murtagh, and R. R. (2015). *Handbook of Cluster Analysis*. Chapman & Hall CRC Handbooks of Modern Statistical Methods.
- [53] Hestilow, T. and Y. Huang (2009). Clustering of gene expression data based on shape similarity. *EURASIP Journal on Bioinformatics and Systems Biology*. Article ID 195712.
- [54] Hu, J., P. Wang, and A. Qu (2015). Estimating and identifying unspecified correlation structure for longitudinal data. *Journal of Computational and Graphical Statistics* 24(2), 455–476.
- [55] Huang, J. Z., L. Liu, and N. Liu (2007). Estimation of large covariance matrices of longitudinal data with basis function approximations. *Journal of Computational and Graphical Statistics* 16, 189–209.
- [56] Huang, J. Z., C. O. Wu, and L. Zhou (2002). Varying-coefficient models and basis function approximations for the analysis of repeated measurements. *Biometrika* 89, 111–128.
- [57] Jacques, J. and C. Preda (2013a). Funclust : A curves clustering method using functional random variable density approximation. *Neurocomputing* 112, 164–171.
- [58] Jacques, J. and C. Preda (2013b). Model-based clustering for multivariate functional data. *Computational Statistics and Data Analysis*. In press.
- [59] Jacques, J. and C. Preda (2014). Functional data clustering : a survey. *Advances in Data Analysis and Classification, Springer Verlag* 8(3).
- [60] James, G. and C. A. Sugar (2003). Clustering for sparsely sampled functional data. *Journal of the American Statistical Association* 98, 397–408.

- [61] Jones, B. L. and D. S. Nagin (2007). Advances in group-based trajectory modeling and an sas procedure for estimating them. *Sociological Methods & Research* 35(4), 542–571.
- [62] Jones, B. L., D. S. Nagin, and K. Roeder (2001). A sas procedure based on mixture models for estimating developmental trajectories. *Sociological Methods & Research* 29(3), 374–393.
- [63] Kim, B. R., L. Zhang, A. Berg, J. Fan, and R. Wu (2008). A computational approach to the functional clustering of periodic gene-expression profiles. *Genetics* 180, 821–834.
- [64] Kim, J. and H. Kim (2008). Clustering of change patterns using fourier coefficients. *Bioinformatics* 24, 184–191.
- [65] Komarek, A. and L. Komarkova (2013). Clustering for multivariate continuous and discrete longitudinal data. *The Annals of Applied Statistics* 7 (1), 177–200.
- [66] Krzanowski, W. J., P. Jonathan, W. V. McCarthy, and M. R. Thomas (1995). Discriminant analysis with singular covariance matrices : Methods and applications to spectroscopic data. *Journal of the Royal Statistical Society Series C*, 4, 101–115.
- [67] Laird, N. M. and J. H. Ware (1982). Random-effects models for longitudinal data. *Biometrics* 38, 963–974.
- [68] Leisch, F. (2004). Flexmix : A general framework for finite mixture models and latent class regression in r. *Journal of Statistical Software* 11(8), 1–18.
- [69] Leng, X. and H. Müller (2006). Classification using function data analysis for temporal gene expression data. *Bioinformatics* 22, 68–76.
- [70] Li, Y. (2011). Efficient semiparametric regression for longitudinal data with nonparametric covariance estimation. *Biometrika* 98(2), 355–370.
- [71] Lin, C.-J. (2014). *A pattern-clustering method for longitudinal data - heroin users receiving methadone*. Ph. D. thesis, University College London.
- [72] Lin, D. and Z. Ying (2001). Semiparametric and nonparametric regression analysis of longitudinal data. *Journal of the American Statistical Association* 96, 103–126.
- [73] Lin, X. and R. J. Carroll (2000). Nonparametric function estimation for clustered data when the predictor is measured without with error. *Journal of the American Statistical Association* 95, 520–534.
- [74] Lin, X. and R. J. Carroll (2001). Semiparametric regression for clustered data using generalized estimating equations. *Journal of the American Statistical Association* 96, 1045–1056.
- [75] Lin, X. and R. J. Carroll (2006). Semi-parametric estimation in general repeated measures problems. *Journal of the Royal Statistical Society Series B* 68, 69–88.
- [76] Lin, X., N. Wang, A. H. Welsh, and R. J. Carroll (2004). Equivalent kernels of smoothing splines in nonparametric regression for clustered longitudinal data. *Biometrika* 91, 177–193.

- [77] Liu, Y. and N. Luo (2014). A new approach in pattern clustering on longitudinal data. *Journal of Computational Information Systems* 14, 6209–6222.
- [78] Luan, Y. and H. Li (2003). Clustering of time-course gene expression data using a mixed-effects model with b-splines. *Bioinformatics* 19, 474–482.
- [79] Ma, P., C. Castillo-Davis, W. Zhong, and J. Liu (2006). A data-driven clustering method for time course gene expression data. *Nucleic Acids Research* 34, 1261–1269.
- [80] Ma, P. and W. Zhong (2008). Penalized clustering of large-scale functional data with multiple covariates. *Journal of the American Statistical Association* 103, 625–636.
- [81] Ma, P., W. Zhong, Y. Feng, and J. Liu (2008). Bayesian functional data clustering for temporal microarray data. *International Journal of Plant Genomics*. Article ID 231897.
- [82] McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models, 2nd Edition*. London : Chapman and Hall.
- [83] McLachlan, G., R. W. Bean, and D. Peel (2002). A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics*. 18, 1–10.
- [84] McLachlan, G. and D. Peel (1998). Robust cluster analysis via mixtures of multivariate t-distributions. *Lecture Notes in Computer Science : Springer-Verlag, Berlin*. 1451, 658–666.
- [85] McLachlan, G. and D. Peel (2000). *Finite Mixture Models*. New York : Wiley, 2000.
- [86] McNicholas, P. D. and T. B. Murphy (2008). Parsimonious gaussian mixture models. *Statistics and Computing* 18(3), 285–296.
- [87] McNicholas, P. D. and T. B. Murphy (2010). Model-based clustering of longitudinal data. *Canadian Journal of Statistics* 38, 153–168.
- [88] McNicholas, P. D. and S. Subedi (2012). Clustering gene expression time course data using mixtures of multivariate t-distributions. *Journal of Statistical Planning and Inference* 142, 1114–1127.
- [89] Melnykov, V. (2013). Challenges in model-based clustering. *WIREs Comp Stat* 5, 135–148.
- [90] Melnykov, V. and R. Maitra (2010). Finite mixture models and model-based clustering. *Statistics Surveys* 4, 80–116.
- [91] Meng, X. and D. van Dyk (1997). The em algorithm ? an old folk song sung to the fast tune (with discussion). *J. Roy. Stat. Soc. Serie B* 59, 511–567.
- [92] Molenberghs, G. and G. Verbeke (2005). *Models for Discrete Longitudinal Data*. Springer. New York.
- [93] Muthen, L. K. and B. O. Muthen (1998-2010). *Mplus User’s Guide*. Los Angeles, CA.
- [94] Ng, S. K., G. McLachlan, K. Wang, L. Ben-Tovim Jones, and S.-W. Ng (2006). A mixture model with random-effects components for clustering correlated gene-expression profiles. *Bioinformatics* 22, 1745–1752.

- [95] Nueda, M., A. Conesa, J. Westerhuis, H. Hoefsloot, A. Smilde, M. Talo, and A. Ferrer (2007). Discovering gene expression patterns in time course microarray experiments by anova-sca. *Bioinformatics* 23, 1792–1800.
- [96] Pauler, D. K. and N. M. Laird (2000). A mixture model for longitudinal data with application to assessment of noncompliance. *Biometrics* 56, 464–472.
- [97] Pinheiro, J. C. and D. Bates (2000). *Mixed-Effects Models in S and S-PLUS*. New York : Springer.
- [98] Pourahmadi, M. (2000). Maximum likelihood estimation of generalised linear models for multivariate normal covariance matrix. *Biometrika* 87, 425–435.
- [99] Qin, L. X. and S. Self (2006). The clustering of regression models method with applications in gene expression data. *Biometrics* 62, 526–533.
- [100] Rajulton, F. (2001). The fundamentals of longitudinal research : An overview. *Special Issue on Longitudinal Methodology, Canadian Studies in Population* 28(2), 169–185.
- [101] Ramsay, J. and B. Silverman (2005). *Functional Data Analysis*. New York : Springer.
- [102] Ren, Y. (2012). *A Non-parametric Bayesian Method for Hierarchical Clustering of Longitudinal Data*. Ph. D. thesis, Department of Mathematical Sciences of the McMicken College of Arts and Sciences, University of Cincinnati.
- [103] Rice, J. (2004). Functional and longitudinal data analysis : Perspectives on smoothing. *Statistica Sinica*.
- [104] Rice, J. and C. Wu (2001). Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics* 57, 253–259.
- [105] Rindfleisch, A., A. J. Malter, S. Ganesan, and C. Moorman (2008). Cross-sectional versus longitudinal survey research : concepts, findings and guidelines. *Journal of Marketing Research* 45(3), 261–279.
- [106] Ruppert, D., M. Wand, and R. Carroll (2003a). *Semiparametric Regression*. Cambridge University Press, Cambridge, UK.
- [107] Ruppert, D., M. P. Wand, and R. J. Carroll (2003b). *Semiparametric Regression*. Cambridge : Cambridge University Press.
- [108] Ruppert, M. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics* 11, 735–757.
- [109] Schramm, C., C. Vial, A. Bachoud-Levi, and S. Katsahian (2015). Clustering of longitudinal data by using an extended baseline : A new method for treatment efficacy clustering in longitudinal data. *Statistical Methods in Medical Research*, 1–21.
- [110] Shaikh, M. (2009). *Clustering incomplete data*. Ph. D. thesis, University of Guelph.
- [111] Shaikh, M., P. D. McNicholas, and A. F. Desmond (2010). A pseudo-em algorithm for clustering incomplete longitudinal data. *The International Journal of Biostatistics, Issue 1, Article 8. 6*.

- [112] Song, J. J., H. Lee, J. S. Morris, and S. Kang (2007). Clustering of time-course gene expression data using functional data analysis. *Computational Biology and Chemistry* 31, 265–274.
- [113] Staniswalis, J. G. and J. J. Lee (1998). Nonparametric regression analysis of longitudinal data. *Journal of the American Statistical Association* 93, 1403–1418.
- [114] Stone, C., M. Hansen, C. Kooperberg, and Y. K. Truong (1997). Polynomial splines and their tensor products in extended linear modeling (with discussion). *Annals of Statistics* 25, 1371–1470.
- [115] Suarez, A. J. and S. Ghosal (2016). Bayesian clustering of functional data using local features. *Bayesian Analysis*, Number 1 11, 71–98.
- [116] Sugar, C. A. and G. M. James (2003). Finding the number of clusters in a data set : An information-theoretic approach. *Journal of the American Statistical Association* 98, 750–778.
- [117] Tarpey, T. (2007). Linear transformations and the k-means clustering algorithm : Applications to clustering curves. *The American Statistician* 61(1), 34–40.
- [118] Ullah, S. and C. F. Finch (2013). Applications of functional data analysis : A systematic review. *BMC Medical Research Methodology* 1471-2288, 13–43.
- [119] Usami, S. (2014). Constrained k-means on cluster proportion and distances among clusters for longitudinal data analysis. *Japanese Psychological Research*, No. 4 56, 361–372.
- [120] Verbeke, G., S. Fieuws, G. Molenberghs, and M. Davidian (2014). The analysis of multivariate longitudinal data : A review. *Statistical Methods in Medical Research* 23(1), 42–59.
- [121] Verbeke, G. and G. Molenberghs (2000). *Linear Mixed Models for Longitudinal Data*. Springer series in statistics. New York.
- [122] Wahba, G. (1990). *Spline Models for Observational Data*. Philadelphia : CBMS-NSF Regional Conference Series, SIAM.
- [123] Wakefield, J., C. Zhou, and S. Self (2003). Modelling gene expression data over time : curve clustering with informative prior distributions. In Bayesian Statistics 7 (eds. J. M. Bernardo, M. J. Bayarri, B. J. O, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West). Oxford : Clarendon Press.
- [124] Wand, M. P. and M. C. Jones (1995). *Kernel Smoothing*. London : Monographs on Statistics and Applied Probability, Chapman & Hall.
- [125] Wang, N. (2003). Marginal nonparametric kernel regression accounting for within-subject correlation. *Biometrika* 90, 43–52.
- [126] Wang, N., R. J. Carroll, and X. Lin (2005). Efficient semi-parametric marginal estimation for longitudinal/clustered data. *Journal of the American Statistical Association* 100, 147– 157.

- [127] Wu, H. and J. T. Zhang (2006). *Nonparametric Regression Methods for Longitudinal Data Analysis, Mixed-Effects Modeling Approaches*. Wiley Series in Probability and Statistics, John Wiley & Sons, 2006.
- [128] Wu, W. B. and M. Pourahmadi (2003). Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika* 90, 831–844.
- [129] Yao, F., H. G. Müller, and J. L. Wang (2005). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association* 100, 577–590.
- [130] Yeung, K. Y., C. Fraley, A. Murua, A. E. Raftery, and W. L. Ruzzo (2001). Model-based clustering and data transformations for gene expression data. *Bioinformatics* 17, 977–987.
- [131] Zeger, S. L. and P. J. Diggle (1994). Semiparametric models for longitudinal data with application to cd4 cell numbers in hiv seroconverters. *Biometrics* 50, 689–699.
- [132] Zeger, S. L. and K.-Y. Liang (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* 42, 121–130.
- [133] Zhang, Y., S. Horvath, R. Ophoff, and D. Telesca (2014). Comparison of clustering methods for time course genomic data : Applications to aging effects. UCLA : 346168. Retrieved from : <http://escholarship.org/uc/item/6pc0068s>.
- [134] Zhao, X., J. Marron, and M. Wells (2004). The functional data analysis view of longitudinal data. *Statistica Sinica* 14, 789–808.

Chapitre 3

FUNCTIONAL MODEL-BASED CLUSTERING FOR LONGITUDINAL DATA

ABSTRACT

The porcine reproductive and respiratory syndrome virus (PRRSV) causes respiratory symptoms in growing pigs and spontaneous abortions in pregnant sows. The annual economic losses due to PRRSV are estimated to be 670 million in the United States alone. A particular gene seems to be a major genetic determinant for the disease. Following previous studies, its effect appears to be different for different groups of pigs. We develop a flexible Bayesian model for the analysis of a multivariate longitudinal study carried out by the PRRSV Host Genetics Consortium. The model applies to general longitudinal or time-course data that presents unknown clusters of individuals. It combines functional principal component analysis and model-based clustering in order to simultaneously model and cluster the individual longitudinal trajectories. Model selection and inference are carried out using a Laplace approximation to the Bayes factors.

Key words : Longitudinal data, model-based clustering, sparse longitudinal data, functional data analysis, mixture Student, PRRSV.

3.1. INTRODUCTION

In many fields, longitudinal studies have become an essential tool for studying the evolution in time of a given phenomena. They are composed of measurements taken at different points of a temporal axis on individuals involved in the study. A fundamental characteristic of this type of data is that the observations on the same individual tend to be correlated. The statistical methodology for the analysis of longitudinal data has evolved remarkably in the past thirty years, due to increasingly sophisticated technologies that nowadays can be implemented on high performance machines. Commonly used methods for longitudinal data analysis are based on parametric models such as the *linear mixed-effects model* proposed by Laird and Ware [25].

Many empirical examples such as Brumback and Rice [6] in reproductive health, Zeger and Diggle [44], Lin and Ying [27] in longitudinal trajectories in AIDS research or Diggle et al. [12] in age effects on childhood respiratory diseases, have shown that fully parametric assumptions are not always appropriate to analyze the temporal dynamic between response variable and covariates in longitudinal studies. Non-parametric and semi-parametric models have been developed to propose more flexible functional forms to handle longitudinal data. These models are essentially based on kernel and spline smoothing methods. An emerging non-parametric methodology for modeling longitudinal data is based on the functional data analysis (FDA) approach in which longitudinal trajectories are viewed as samples of partially observed smooth functions or curves on some interval.

In this paper, we design an appropriate functional model for the analysis of the porcine reproductive and respiratory syndrome virus (PRRSV). This is a 15kb positive-stranded RNA virus in the family *Arteriviridae*. It emerged nearly simultaneously in the United States and Europe in the late 1980s (Wensvoort et al. [41]; Loula [28]), but has now spread to Asia (Tian et al. [36]). Upon infection of the pig, PRRSV replicates rapidly, remaining detectable in the blood for about 28 days, but persisting elsewhere, and rendering the pig infectious, for as many as 200 days (Rowland et al. [35]). The virus causes respiratory symptoms in growing pigs and spontaneous abortions in pregnant sows (Collins et al. [9]). Its tendency to cause prolonged subclinical infection is associated with a variety of debilitating syndromes involving co-infecting pathogens (Rowland et al. [35]). The annual economic losses due to PRRSV are estimated to be 670 million in the United States alone (Holtkamp et al. [17]).

The PRRS Host Genetics Consortium (PHGC) was established to identify pig genetic determinants of PRRS susceptibility and tolerance (Lunney et al. [29]). In order to have

sufficient power, the PHGC experimentally infected hundreds of young pigs with PRRSV isolate NVSL97-7985 (GenBank accession AY545985) in a few large experimental trials over a few years. Each trial started with 200 pigs at weaning (3-4 weeks old) supplied by a single pig breeding company from a farm testing negative for PRRSV, *Mycoplasma hyopneumoniae*, and swine influenza virus. Pigs were from at least 30 litters from at least 10 sires mated with 3-8 dams/sire (Rowland et al. [35]). The pigs were transported to the biosecure Kansas State University testing facility, divided among at least 12 pens, and treated with broad spectrum antibiotics for seven days, then infected and followed for 42 days. Blood samples were collected at 0, 4, 7, 10, 14, 21, 28, 35, and 42 days post infection (dpi), and pigs were weighed weekly. The amount of virus in the blood, the virus load, was quantified using quantitative real-time Polymerase Chain Reaction (qPCR) (Kubista et al. [24]). All pigs were confirmed to be infected with PRRSV. The virus load peaked around 7-14 dpi (Rowland et al. [35]), then decayed to undetectable in most pigs by 28 dpi. A subset (10-20%) of pigs suffered a rebound in virus around 28-35 dpi that cleared by 42 dpi. Pig growth curves, as measured by weight gain, were suppressed compared to controls from the same litter that were not infected (Lunney et al. [29]). There was extensive variation in both pig growth curves and virus load trajectories. Later analyses revealed an important genetic determinant of this variation (Boddicker et al. [3]), suggesting the presence of discrete subpopulations of pigs with distinct responses to infection.

We use the functional data analysis approach to propose a very suitable and flexible model to analyze this type of data. The model combines multi-dimensional functional principal components analysis and clustering to deal with any type of longitudinal data even if the observations are sparse, irregularly spaced or occur at different time points for each individual. This is specially suitable for the analysis of PRRSV evolution which must take into account both the virus load and weight gain trajectories simultaneously. In addition to the longitudinal aspects of each trajectory (that is, different measurement days for different pigs), these trajectories present the added complexity that both variables, virus load and weight gain, were not necessarily measured on the same days for the same pigs.

The study of the effect of the gene WUR (for locus WUR10000125) which following the study of Boddicker et al. [3] appears to be a major genetic determinant of PRRSV disease, is of particular interest in the analysis. As suggested by Boddicker et al. [3], we model through clustering the presence of subpopulations of pigs that respond differently to infection so that there are latent effects associated with each subpopulation. The effect of WUR is then adjusted according to the latent effects which in turn hint to specific characteristics common in certain groups of pigs. The clusters are unknown and must be estimated along

the latent fixed and random effects involved in the model.

The paper is organized as follows. In Section 2, we define the model underlying the analysis of longitudinal data and the estimation procedure based on the Expectation-Maximization (EM) algorithm (Dempster et al. [11]). We also present the tools for model selection. Section 3 deals with simulation experiments as well as performance comparisons with other functional models applied on real data sets. The analysis of the porcine reproductive and respiratory syndrome virus is presented in Section 4.

3.2. FUNCTIONAL DATA ANALYSIS AND CLUSTERING

Consider a longitudinal study involving N subjects. Assume that n_i measures were collected for subject i . For each subject $i \in \{1, \dots, N\}$, the observed measurements $\mathbf{Y}_i = (y_{i1}, \dots, y_{in_i})$ are assumed to be realizations with measurement errors of a random function \mathcal{X}_i at time points $\mathbf{t}_i = (t_{i1}, \dots, t_{in_i})$ according to the model :

$$y_{ij} = \mathcal{X}_i(t_{ij}) + \epsilon_{ij} \quad j = 1, \dots, n_i \quad i = 1, \dots, N; \quad (3.2.1)$$

where the ϵ_{ij} 's are independent and identically distributed, with moments $E(\epsilon_{ij}) = 0$ and $E(\epsilon_{ij}^2) = \sigma^2$. The functions $\{\mathcal{X}_1, \dots, \mathcal{X}_N\}$ are assumed to be independent realizations of a second order stochastic process $\mathcal{X}(t)$ defined on a compact domain \mathcal{I} (the temporal axis).

3.2.1. The model

The model proposed here for longitudinal data combines functional principal components analysis and model based clustering. The functional principal components analysis utilizes the approach of James et al. [22], which is based on the mixed effects model, while the clustering model, although developed independently, is strongly related to the functional clustering model suggested by James and Sugar [23]. Our model differs from that of James and Sugar [23] in that we use a Bayesian framework for the parameters, and a different parameterization for the clusters. Our methodology also differs from that of James and Sugar [23] in that we perform model selection via an approximation to the Bayes factors. While James and Sugar [23] suggest several ad hoc but clever ways to choose the dimensions of several parameters, we adopt a unified and principled way to choose all these dimensions through our approximation to the Bayes factors. Below, we explain these points in more details.

Let $Y_i(\cdot)$ be the function or curve underlying the observed measurements \mathbf{Y}_i for subject i and $Y_i(t)$ be the evaluation of $Y(\cdot)$ at time point t . We assume that there exists an overall mean function $\hat{\mu}(t)$ and a finite orthogonal basis of norm-one (square-integrable) functions

in L^2 , $\{\hat{f}_1(t), \dots, \hat{f}_k(t)\}$ such that

$$\begin{aligned} Y_i(t) &= \hat{\mu}(t) + \sum_{j=1}^k \alpha_{ij} \hat{f}_j(t) + \epsilon_i(t) \\ &= \hat{\mu}(t) + \hat{\mathbf{f}}(t)^\top \boldsymbol{\alpha}_i + \epsilon_i(t) \quad i = 1, \dots, N \end{aligned} \quad (3.2.2)$$

for some random k -dimensional vectors $\boldsymbol{\alpha}_i = (\alpha_{i1}, \dots, \alpha_{ik})$ (the component *scores*), and error terms $\epsilon_i(t)$, $i = 1, \dots, N$, where $\hat{\mathbf{f}}(t)^\top = (\hat{f}_1(t), \dots, \hat{f}_k(t))$ (here and throughout the manuscript, the superscript \top indicates transposition). The functions $\hat{f}_j(t)$ are sometimes referred to as the principal component functions. The clustering model assumes a mixed effects framework on the component scores $\boldsymbol{\alpha}_i$:

$$\boldsymbol{\alpha}_i = \boldsymbol{\mu}_{\mathbf{z}_i} + \boldsymbol{\gamma}_i^{\mathbf{z}_i} \quad (3.2.3)$$

where \mathbf{z}_i denotes the unknown cluster the individual i belongs to. The cluster indicator \mathbf{z}_i takes the possible values $1, 2, \dots, G$ where G is the number of clusters. The vector $\boldsymbol{\mu}_{\mathbf{z}_i}$ represents the individual i 's cluster mean and $\boldsymbol{\gamma}_i^{\mathbf{z}_i}$ indicates the specific effect of the individual or the deviation from it's cluster effect. With this formulation, the functional clustering model can be written as :

$$Y_i(t) = \hat{\mu}(t) + \hat{\mathbf{f}}(t)^\top \boldsymbol{\mu}_{\mathbf{z}_i} + \hat{\mathbf{f}}(t)^\top \boldsymbol{\gamma}_i^{\mathbf{z}_i} + \epsilon_i(t) \quad i = 1, \dots, N. \quad (3.2.4)$$

Conditionally on the cluster to which an individual belongs, its longitudinal trajectory is decomposed into the sum of three components plus an error term. The first component represents the overall mean, the second component stands for the cluster or group effect, and the third component indicates the subject-specific effect (or deviation from its group effect). As in James et al. [22] and James and Sugar [23], we use a specification of the model in a finite-dimensional basis $\mathbf{b}(t)^\top = (b_1(t), \dots, b_q(t))$ of B-splines. Under this specification, we can write $\hat{\mu}(t) = \mathbf{b}(t)^\top \boldsymbol{\theta}_\mu$, with $\boldsymbol{\theta}_\mu \in \mathbf{R}^q$. Similarly, we can write $\hat{\mathbf{f}}(t)^\top = \mathbf{b}(t)^\top \boldsymbol{\Theta}$ for a $q \times k$ matrix $\boldsymbol{\Theta}$. This new parameterization yields

$$Y_i(t) = \mathbf{b}(t)^\top \boldsymbol{\theta}_\mu + \mathbf{b}(t)^\top \boldsymbol{\Theta} \boldsymbol{\mu}_{\mathbf{z}_i} + \mathbf{b}(t)^\top \boldsymbol{\Theta} \boldsymbol{\gamma}_i^{\mathbf{z}_i} + \epsilon_i(t) \quad (3.2.5)$$

which is equivalent to :

$$\mathbf{Y}_i = \mathbf{B}_i \boldsymbol{\theta}_\mu + \mathbf{B}_i \boldsymbol{\Theta} \boldsymbol{\mu}_{\mathbf{z}_i} + \mathbf{B}_i \boldsymbol{\Theta} \boldsymbol{\gamma}_i^{\mathbf{z}_i} + \boldsymbol{\epsilon}_i \quad (3.2.6)$$

where \mathbf{Y}_i is the vector of the measurements at time points \mathbf{t}_i and $\mathbf{B}_i = [\mathbf{b}(t_{i1}), \dots, \mathbf{b}(t_{in_i})]^\top$ is the matrix of the spline basis evaluated at those time points. The q -dimensional vector $\boldsymbol{\theta}_\mu$ and the matrix $\boldsymbol{\Theta}$ represent, respectively, the coefficients in the basis of the overall mean function $\hat{\mu}(t)$ and the principal components functions $\hat{f}(t)$. To accommodate for certain departure from normality in the data, we assume that the measurement errors $\boldsymbol{\epsilon}_i$ follow a multivariate Student's t distribution with unknown degrees of freedom ν_0 .

Bayesian framework :

Let $\mathbf{z}_i = (z_{i1}, \dots, z_{iG}) \in \{0, 1\}^G$ be such that $z_{ig} = 1$ if the individual i belongs to the cluster g ($1 \leq g \leq G$) and 0 otherwise ($i = 1, \dots, N$). These are the cluster membership indicators. They are assumed independent and identically distributed multinomials with parameters $(1, \pi_1, \dots, \pi_G)$, where π_g is the prior probability that an individual belongs to the g^{th} cluster (mixing probabilities). With a slight abuse of notation, we will denote by $\{\mathbf{z}_i = g\}$ the expression $\{\mathbf{z}_i = \mathbf{e}_g\}$ where \mathbf{e}_g is a G -dimensional vector with a 1 in the g^{th} coordinate and 0's elsewhere. To complete the setup of the proposed model, we further assume that :

$$\begin{cases} \boldsymbol{\mu}_g \sim \mathcal{N}_k(\mathbf{0}, \boldsymbol{\Gamma}_\mu) & \text{and } \boldsymbol{\Gamma}_\mu \sim \text{InvWishart}(m, (m - k - 1)I_k) \\ \boldsymbol{\gamma}_i^g \sim \mathcal{N}_k(\mathbf{0}, \boldsymbol{\Gamma}_g) & \text{and } \boldsymbol{\Gamma}_g \sim \text{InvWishart}(m, (m - k - 1)\mathbf{D}) \\ \mathbf{D} = \text{diag}(d_{11}, d_{22}, \dots, d_{kk}) & \text{with } d_{jj} \sim \text{Inv}\chi^2(m) \text{ and i.i.d. } (j = 1, \dots, k) \\ (\pi_1, \dots, \pi_G) \sim \text{Dirichlet}(a_1, \dots, a_G) \end{cases} \quad (3.2.7)$$

Let $\mathbf{B} = [\mathbf{b}(\min t_{ij}), \dots, \mathbf{b}(\max t_{ij})]^\top$ be the matrix of the spline basis evaluated on a fine grid of the range of the temporal axis, so that every observed time point in the data is included in the grid. Let $\delta_{ij} = 1$, if $i = j$, and be zero, otherwise. To approximate the orthogonality constraint $\{\int \hat{f}_j \hat{f}_l = \delta_{jl}\}$, we choose the basis of functions $\mathbf{b}(\cdot)$ so that $\mathbf{B}^\top \mathbf{B} = I_q$ and $\boldsymbol{\Theta}^\top \boldsymbol{\Theta} = I_k$. There are two important identifiability constraints in the model. These ensure that $\mathbf{B}_i \boldsymbol{\theta}_\mu$ is the overall mean and that $\mathbf{B}_i \boldsymbol{\Theta} \boldsymbol{\mu}_g$ is the mean curve of cluster g :

$$\begin{cases} \sum_{g=1}^G \pi_g \boldsymbol{\mu}_g = \mathbf{0} \\ \sum_{i=1}^N P_{ig} \boldsymbol{\gamma}_i^g = \mathbf{0}, \quad \text{for all } g = 1, \dots, G \end{cases} \quad (3.2.8)$$

where P_{ig} denote the posterior probability that the individual i belongs to group g . In order to simplify the formulas, we augment our model by representing the error terms as mixture between a Normal and an inverse-Chi-squared variables :

$$\begin{cases} \boldsymbol{\epsilon}_i | \nu_i \sim \mathcal{N}_{n_i}(0, \sigma^2 \nu_i I_{n_i}) & \text{and } \nu_i \sim \text{Inv}\chi^2(\nu_0) \Rightarrow \boldsymbol{\epsilon}_i \sim t_{\nu_0}(0, \sigma^2 I_{n_i}) \\ \text{with } \sigma^2 \sim \text{InvGamma}(\alpha_\sigma, \beta_\sigma). \end{cases} \quad (3.2.9)$$

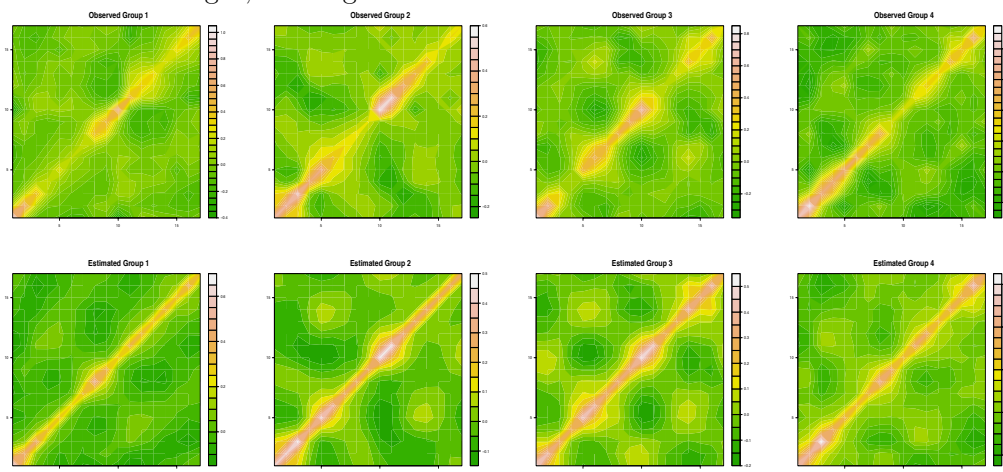
Note that m , α_σ , β_σ and the $a_g, 1 \leq g \leq G$ are hyper-parameters. The structure of this model is appealing since the clustering of the individuals relies essentially on a space of much reduced dimension than the original longitudinal trajectories. Note that choosing $k \leq 3$ would allow some form of visualization of the groups. The model of James and Sugar [23] adds another layer of parameterization in the clusters so that the clustering is forced to lie in a subspace of very small dimension. We prefer to let the data tell us which dimension better describes the clustering structure. Note that the model supposes that the variance-covariance matrix of the vector of measurements \mathbf{Y} , conditional on the hyper-parameters is

equal to :

$$\begin{aligned}
 \text{Var}(\mathbf{Y}) &= \text{Var}(E(\mathbf{Y}|g)) + E(\text{Var}(\mathbf{Y}|g)) & (3.2.10) \\
 &= \text{Var}(E(\mathbf{B}\boldsymbol{\theta}_\mu + \mathbf{B}\boldsymbol{\Theta}\boldsymbol{\mu}_g + \mathbf{B}\boldsymbol{\Theta}\boldsymbol{\gamma}^g + \boldsymbol{\epsilon})) + E(\text{Var}(\mathbf{B}\boldsymbol{\theta}_\mu + \mathbf{B}\boldsymbol{\Theta}\boldsymbol{\mu}_g + \mathbf{B}\boldsymbol{\Theta}\boldsymbol{\gamma}^g + \boldsymbol{\epsilon})) \\
 &= \text{Var}(\mathbf{B}\boldsymbol{\theta}_\mu) + E([\mathbf{B}\boldsymbol{\Theta}\boldsymbol{\Gamma}_\mu\boldsymbol{\Theta}^\top\mathbf{B}^\top] + [\mathbf{B}\boldsymbol{\Theta}\boldsymbol{\Gamma}_g\boldsymbol{\Theta}^\top\mathbf{B}^\top] + [(\sigma^2\nu_0/(\nu_0 - 2))I_n]) \\
 &= \mathbf{B}\boldsymbol{\Theta}(I_k + \mathbf{D})\boldsymbol{\Theta}^\top\mathbf{B}^\top + (\sigma^2\nu_0/(\nu_0 - 2))I_n.
 \end{aligned}$$

This can be estimated using the Maximum A Posteriori (MAP) estimators of the parameters, which are in turn obtained by the EM algorithm described next. As an illustration of this covariance estimate, consider the yeast cycle data described below in Section 3.3.3.4. Figure 3.1 shows the observed and estimated variance-covariance matrices associated with the four clusters found by our procedure. The overall (across clusters) yeast-cycle observed and estimated covariances are shown in Figure 3.2.

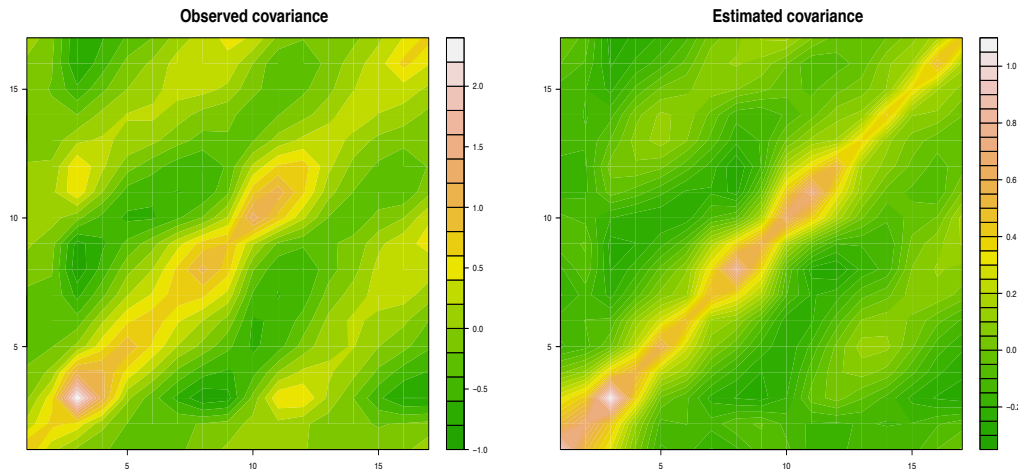
FIGURE 3.1. Yeast cycle data. The observed (top row) and estimated (bottom row) variance-covariance matrices by cluster. The clusters are arranged from left to right, starting with Cluster 1.



3.2.2. Extension to multiple dimensions

The model can be easily extended to a multidimensional model. The motivation for this extension relies on the belief that the explicit and simultaneous modeling of multiple curves from an individual carries more information than the modeling of a single curve that summarizes all the information from the individual. However, if the curves are highly correlated, then there is no loss of information in combining all the curves from an individual into a

FIGURE 3.2. Yeast cycle data. The overall observed (left) and estimated (right) variance-covariance matrices.



single representative curve. The approach is illustrated here for a bidimensional model because this is the dimension needed to analyze the PRRSV data, but it remains valid for multiple curves.

Suppose that we have N two-dimensional functional observations $\{Y_i^1(t), Y_i^2(t)\}$ with $t \in \mathcal{I}$ where $Y_i^l(t)$ is the value at time t of the l^{th} variable measured on the i -th individual, $l = 1, 2$; $i = 1, \dots, N$. These curves are assumed to come from the following model

$$\mathbb{Y}_i = \mathbb{B}_i \boldsymbol{\theta}_\mu + \mathbb{B}_i \boldsymbol{\Theta} \boldsymbol{\mu}_{\mathbf{z}_i} + \mathbb{B}_i \boldsymbol{\Theta} \boldsymbol{\gamma}_i^{\mathbf{z}_i} + \boldsymbol{\Upsilon}_i, \quad (3.2.11)$$

where $\mathbb{Y}_i = (\mathbf{Y}_i^1, \mathbf{Y}_i^2)^\top$ is the vector of dimension $(n_i^1 + n_i^2)$ arising from stacking into one single vector the observations from the curves $\{Y_i^1(t), Y_i^2(t)\}$; \mathbb{B}_i is the spline basis matrix derived from the one-dimensional basis matrix \mathbf{B}_i as

$$\mathbb{B}_i = \left(\begin{array}{c|c} \mathbf{B}_i & \mathbf{O} \\ \hline \mathbf{O} & \mathbf{B}_i \end{array} \right)$$

and the vector of residuals $\boldsymbol{\Upsilon}_i = (\boldsymbol{\epsilon}_i^1, \boldsymbol{\epsilon}_i^2)^\top$. Although here for simplicity in the exposition we have assumed that curves from the same individual are measured at the same time points, this assumption is not necessary. If the curves are measured at different time points, it suffices to change \mathbf{B}_i for \mathbf{B}_{il} , for each dimension l . As in the one-dimensional model, we assume that the clustering parameter priors are given as in (3.2.7) but where I_k is replaced by I_{2k} and τ_1 and τ_2 are set this time to $(m-2k-1)$ due to the increase in the dimension, and $\boldsymbol{\Upsilon}_i \sim t_{\nu_0}(0, \sigma^2 I_{2n_i})$. The parameter estimation and model selection for the multidimensional

model are performed in the same manner as in the one-dimensional case which is described next.

3.2.3. Parameter estimation

The estimation of the parameters is carried out by the maximization of the mixture model likelihood in the presence of the “latent data” \mathbf{W} given by the cluster labels \mathbf{z}_i and the random effects $\gamma_i^{\mathbf{z}_i}$. For that purpose, we use the EM algorithm. This maximizes the function Q defined as :

$$Q(\mathbf{\Pi}|\mathbf{\Pi}^{(t)}) = E_{\mathbf{W}|\mathbf{Y};\mathbf{\Pi}^{(t)}}[\log p(\mathbf{Y}, \mathbf{W}; \mathbf{\Pi})] \quad (3.2.12)$$

where \mathbf{Y} denotes the “observed” data, and $p(\cdot)$ denotes the corresponding probability function associated with the model (from now on we will write $p(\cdot)$ for all probability functions or densities which can be identified by the context in which they appear). The vector of unknown parameters is $\mathbf{\Pi}$ and $\mathbf{\Pi}^{(t)}$ denote the t^{th} updated value of $\mathbf{\Pi}$. In the expectation step (E-step), the cluster labels \mathbf{z}_i and random effects $\gamma_i^{\mathbf{z}_i}$ are estimated by conditional expectation of the complete log-likelihood given the observed data and the current value of the parameter vector. The function Q is maximized in the maximization step (M-step), and the $(t+1)^{\text{th}}$ updated value $\mathbf{\Pi}^{(t+1)}$ of $\mathbf{\Pi}$ is obtained through :

$$\mathbf{\Pi}^{(t+1)} = \underset{\mathbf{\Pi}}{\operatorname{argmax}} \quad Q(\mathbf{\Pi}|\mathbf{\Pi}^{(t)}). \quad (3.2.13)$$

The maximum likelihood estimator of $\mathbf{\Pi}$ can be obtained by repeating the E and M steps until convergence. Let $\vec{\nu} = \{\nu_1, \dots, \nu_N\}$, $\vec{\mu} = \{\mu_1, \dots, \mu_G\}$, $\vec{\Gamma} = \{\Gamma_1, \dots, \Gamma_G\}$ and $\mathbf{\Lambda} = \{\theta_\mu, \Theta, \mathbf{D}, \Gamma_\mu, \pi_1, \pi_2, \dots, \pi_G, \nu_0, \sigma^2\}$. The parameters of the model are given by $\mathbf{\Pi} = \{\vec{\nu}, \vec{\mu}, \vec{\Gamma}, \mathbf{\Lambda}\}$. The log-likelihood of the “complete data” (\mathbf{Y}, \mathbf{W}) is given by :

$$\begin{aligned} \left\{ \begin{aligned} \log[p(\mathbf{Y}, \mathbf{W}; \mathbf{\Pi})] &= \log \left\{ p(\mathbf{Y}, \mathbf{Z}, \gamma^{\mathbf{z}}; \vec{\nu}, \vec{\mu}, \vec{\Gamma}, \mathbf{\Lambda}) \right\} \\ &= \log \left\{ \left[\prod_{i=1}^N p(\mathbf{Y}_i, \mathbf{z}_i, \gamma_i^{\mathbf{z}_i} | \nu_i, \vec{\mu}, \vec{\Gamma}, \mathbf{\Lambda}) \times p(\nu_i | \vec{\mu}, \vec{\Gamma}, \mathbf{\Lambda}) \right] \left[\prod_{g=1}^G p(\mu_g, \Gamma_g) \right] [p(\mathbf{\Lambda})] \right\} \\ &= \sum_{i=1}^N \left\{ \log p(\mathbf{Y}_i, \mathbf{z}_i, \gamma_i^{\mathbf{z}_i} | \nu_i, \vec{\mu}, \vec{\Gamma}, \mathbf{\Lambda}) + \log p(\nu_i | \vec{\mu}, \vec{\Gamma}, \mathbf{\Lambda}) \right\} \\ &\quad + \sum_{g=1}^G \left\{ \log p(\mu_g, \Gamma_g) \right\} + \log p(\mathbf{\Lambda}) \end{aligned} \right. \quad (3.2.14) \end{aligned}$$

By introducing the probability density functions of the different distributions (the parameters and the hyper-parameters), the expression of the log-likelihood $\log[p(\mathbf{Y}, \mathbf{W}; \mathbf{\Pi})]$ becomes :

$$\begin{aligned}
\log[p(\mathbf{Y}, \mathbf{W}; \mathbf{\Pi})] &= \sum_{i=1}^N \left\{ \begin{aligned} &-\frac{n_i}{2} \log(\nu_i \sigma^2) - \frac{1}{2\nu_i \sigma^2} \left\| \mathbf{Y}_i - (\mathbf{B}_i \boldsymbol{\theta}_\mu + \mathbf{B}_i \boldsymbol{\Theta} \boldsymbol{\mu}_{\mathbf{z}_i} + \mathbf{B}_i \boldsymbol{\Theta} \boldsymbol{\gamma}_i^{\mathbf{z}_i}) \right\|^2 \\ &-\frac{1}{2} \log(|\boldsymbol{\Gamma}_{\mathbf{z}_i}|) - \frac{1}{2} \boldsymbol{\gamma}_{i, \mathbf{z}_i}^T \boldsymbol{\Gamma}_{\mathbf{z}_i}^{-1} \boldsymbol{\gamma}_i^{\mathbf{z}_i} + \sum_{g=1}^G Z_{ig} \log(\pi_g) \\ &+ \frac{\nu_o}{2} \log\left(\frac{\nu_o}{2}\right) - \log[\Gamma\left(\frac{\nu_o}{2}\right)] - \left(1 + \frac{\nu_o}{2}\right) \log(\nu_i) - \frac{\nu_o}{2\nu_i} \end{aligned} \right\} \\
&+ \sum_{g=1}^G \left\{ \begin{aligned} &-\frac{1}{2} \log(|\boldsymbol{\Gamma}_\mu|) - \frac{1}{2} \boldsymbol{\mu}_g^T \boldsymbol{\Gamma}_\mu^{-1} \boldsymbol{\mu}_g + \frac{m}{2} \log(|(m-k-1)\mathbf{D}|) \\ &-\frac{(m+k+1)}{2} \log(|\boldsymbol{\Gamma}_g|) - \frac{(m-k-1)}{2} \text{trace}[\mathbf{D}\boldsymbol{\Gamma}_g^{-1}] \end{aligned} \right\} \\
&+ \left\{ \frac{km}{2} \log(m-k-1) - \frac{(m+k+1)}{2} \log(|\boldsymbol{\Gamma}_\mu|) - \frac{(m-k-1)}{2} \text{trace}[\boldsymbol{\Gamma}_\mu^{-1}] \right\} \\
&+ \sum_{j=1}^k \left\{ +\frac{m}{2} \log\left(\frac{m}{2}\right) - \log[\Gamma\left(\frac{m}{2}\right)] - \left(1 + \frac{m}{2}\right) \log(d_{jj}) - \frac{m}{2d_{jj}} \right\} \\
&+ \left\{ \alpha_\sigma \log(\beta_\sigma) - \log[\Gamma(\alpha_\sigma)] - (\alpha_\sigma + 1) \log(\sigma^2) - \frac{\beta_\sigma}{\sigma^2} \right\} \\
&+ \left\{ -\log[B(a_1, \dots, a_G)] + \sum_{g=1}^G (a_g - 1) \log(\pi_g) \right\} \\
&+ \mathcal{C} \tag{3.2.15}
\end{aligned}$$

where \mathcal{C} is a normalizing constant and $B(a_1, \dots, a_G) = B(\mathbf{a})$ is the multivariate Beta function which can be expressed in terms of the Gamma function $\Gamma(\cdot)$ as $B(\mathbf{a}) = \frac{[\prod_{g=1}^G \Gamma(a_g)]}{\Gamma(\sum_{g=1}^G a_g)}$.

The Appendix A.2 presents all the analytical developments of the expectation step of the EM algorithm. It repeats, on purpose, some of the features of the model in order to be complete. At the end of the expectation step, one obtains the expression of the EM function Q introduced in Equation (3.2.12) as :

$$\begin{aligned}
Q(\mathbf{\Pi}|\mathbf{\Pi}^{(t)}) &= \sum_{i=1}^N \sum_{g=1}^G P_{ig} \times \left\{ \begin{aligned} &-\frac{n_i}{2} \log(\nu_i \sigma^2) - \frac{1}{2} \log(|\boldsymbol{\Gamma}_g|) + \log(\pi_g) \\ &-\frac{1}{2\nu_i \sigma^2} \left\{ \left\| \mathbf{Y}_i - \mathbf{B}_i \boldsymbol{\theta}_\mu - \mathbf{B}_i \boldsymbol{\Theta} \boldsymbol{\mu}_g - \mathbf{B}_i \boldsymbol{\Theta} \hat{\boldsymbol{\gamma}}_i^g \right\|^2 \right\} \\ &+\frac{1}{2\nu_i \sigma^2} \left\{ \text{trace} \left[\mathbf{B}_i \boldsymbol{\Theta} \hat{\mathbf{V}}_i^g \boldsymbol{\Theta}^T \mathbf{B}_i^T \right] \right\} \\ &-\frac{1}{2} \left\{ \hat{\boldsymbol{\gamma}}_{ig}^T \boldsymbol{\Gamma}_g^{-1} \hat{\boldsymbol{\gamma}}_{ig} + \text{trace} \left[\boldsymbol{\Gamma}_g^{-1} \hat{\mathbf{V}}_i^g \right] \right\} \\ &+ \frac{\nu_o}{2} \log\left(\frac{\nu_o}{2}\right) - \log[\Gamma\left(\frac{\nu_o}{2}\right)] - \left(1 + \frac{\nu_o}{2}\right) \log(\nu_i) - \frac{\nu_o}{2\nu_i} \end{aligned} \right\} \\
&+ \sum_{g=1}^G \left\{ \begin{aligned} &-\frac{1}{2} \log(|\boldsymbol{\Gamma}_\mu|) - \frac{1}{2} \boldsymbol{\mu}_g^T \boldsymbol{\Gamma}_\mu^{-1} \boldsymbol{\mu}_g + \frac{m}{2} \log(|(m-k-1)\mathbf{D}|) \\ &-\frac{(m+k+1)}{2} \log(|\boldsymbol{\Gamma}_g|) - \frac{(m-k-1)}{2} \text{trace}[\mathbf{D}\boldsymbol{\Gamma}_g^{-1}] \end{aligned} \right\} \\
&+ \left\{ \frac{km}{2} \log(m-k-1) - \frac{(m+k+1)}{2} \log(|\boldsymbol{\Gamma}_\mu|) - \frac{(m-k-1)}{2} \text{trace}[\boldsymbol{\Gamma}_\mu^{-1}] \right\} \\
&+ \sum_{j=1}^k \left\{ +\frac{m}{2} \log\left(\frac{m}{2}\right) - \log[\Gamma\left(\frac{m}{2}\right)] - \left(1 + \frac{m}{2}\right) \log(d_{jj}) - \frac{m}{2d_{jj}} \right\} \\
&+ \left\{ \alpha_\sigma \log(\beta_\sigma) - \log[\Gamma(\alpha_\sigma)] - (\alpha_\sigma + 1) \log(\sigma^2) - \frac{\beta_\sigma}{\sigma^2} \right\} \\
&+ \left\{ -\log[B(a_1, \dots, a_G)] + \sum_{g=1}^G (a_g - 1) \log(\pi_g) \right\} + \mathcal{C}
\end{aligned}$$

where the expressions of $\hat{\gamma}_i^{\mathbf{z}_i}$ and $\hat{V}_i^{\mathbf{z}_i}$ are :

$$\hat{\gamma}_i^{\mathbf{z}_i} = \left\{ \nu_i^{(t)} \sigma_{(t)}^2 \mathbf{\Gamma}_{\mathbf{z}_i}^{-1(t)} + \mathbf{\Theta}_{(t)}^\top \mathbf{B}_i^\top \mathbf{B}_i \mathbf{\Theta}_{(t)} \right\}^{-1} \mathbf{\Theta}_{(t)}^\top \mathbf{B}_i^\top \left\{ \mathbf{Y}_i - \mathbf{B}_i \boldsymbol{\theta}_\mu^{(t)} - \mathbf{B}_i \mathbf{\Theta}_{(t)} \boldsymbol{\mu}_{\mathbf{z}_i}^{(t)} \right\}$$

$$\hat{V}_i^{\mathbf{z}_i} = \left\{ \mathbf{\Gamma}_{\mathbf{z}_i}^{-1(t)} + \frac{\mathbf{\Theta}_{(t)}^\top \mathbf{B}_i^\top \mathbf{B}_i \mathbf{\Theta}_{(t)}}{\nu_i^{(t)} \sigma_{(t)}^2} \right\}^{-1}$$

The maximization step consists in maximizing the function $Q(\mathbf{\Pi}|\mathbf{\Pi}^{(t)})$ with respect to the vector of parameters $\mathbf{\Pi}$ which yields :

$$\left\{ \begin{array}{l} \boldsymbol{\mu}_g^{(t+1)} = \left[\left(\sum_{i=1}^N \frac{P_{ig}}{\nu_i^{(t)}} \mathbf{\Theta}_{(t)}^\top \mathbf{B}_i^\top \mathbf{B}_i \mathbf{\Theta}_{(t)} \right) + \sigma_{(t)}^2 \mathbf{\Gamma}_\mu^{-1(t)} \right]^{-1} \left[\sum_{i=1}^N \frac{P_{ig}}{\nu_i^{(t)}} \mathbf{\Theta}_{(t)}^\top \mathbf{B}_i^\top \left(\mathbf{Y}_i - \mathbf{B}_i \boldsymbol{\theta}_\mu^{(t)} - \mathbf{B}_i \mathbf{\Theta}_{(t)} \hat{\gamma}_i^g \right) \right] \\ \mathbf{\Gamma}_g^{(t+1)} = \left[\left(\sum_{i=1}^N P_{ig} \right) + (m+k+1) \right]^{-1} \left[\left(\sum_{i=1}^N P_{ig} (\hat{\gamma}_i^g \hat{\gamma}_i^{g\top} + \hat{V}_i^g) \right) + (m-k-1) \mathbf{D} \right] \\ \pi_g^{(t+1)} = \frac{\left(\sum_{i=1}^N P_{ig} \right) + (a_g - 1)}{N + \left(\sum_{g=1}^G a_g \right) - G} \quad (0 \leq \pi_g^{(t+1)} \leq 1) \\ \boldsymbol{\theta}_\mu^{(t+1)} = \left[\sum_{i=1}^N \left(\mathbf{B}_i^\top \mathbf{B}_i \right) \nu_i^{-1(t)} \right]^{-1} \left[\sum_{i=1}^N \sum_{g=1}^G \frac{P_{ig}}{\nu_i^{(t)}} \mathbf{B}_i^\top \left(\mathbf{Y}_i - \mathbf{B}_i \mathbf{\Theta}_{(t)} \boldsymbol{\mu}_g^{(t)} - \mathbf{B}_i \mathbf{\Theta}_{(t)} \hat{\gamma}_i^g \right) \right] \\ \boldsymbol{\Theta}_j^{(t+1)} = \left\{ \sum_{i=1}^N \sum_{g=1}^G \frac{P_{ig}}{\nu_i^{(t)}} \left[(\boldsymbol{\alpha}_{ig})_j^2 + (\hat{V}_i^g)_{jj} \right] \left(\mathbf{B}_i^\top \mathbf{B}_i \right) \right\}^{-1} \{ \Omega_1 - \Omega_2 \} \quad ; \quad \text{for } j = 1, \dots, k. \\ \Omega_1 = \sum_{i=1}^N \sum_{g=1}^G \frac{P_{ig}}{\nu_i^{(t)}} \left[(\boldsymbol{\alpha}_{ig})_j \mathbf{B}_i^\top \left(\mathbf{Y}_i - \mathbf{B}_i \boldsymbol{\theta}_\mu^{(t)} \right) \right]; \quad (\boldsymbol{\alpha}_{ig})_j = (\boldsymbol{\mu}_g + \hat{\gamma}_i^g)_j \\ \Omega_2 = \sum_{i=1}^N \sum_{g=1}^G \frac{P_{ig}}{\nu_i^{(t)}} \left[\sum_{h \neq j}^k \left((\boldsymbol{\alpha}_{ig})_j (\boldsymbol{\alpha}_{ig})_h + (\hat{V}_i^g)_{hj} \right) \left(\mathbf{B}_i^\top \mathbf{B}_i \right) \boldsymbol{\Theta}_h \right] \\ \sigma_{(t+1)}^2 = \frac{\frac{1}{2} \left\{ \sum_{i=1}^N \sum_{g=1}^G \frac{P_{ig}}{\nu_i^{(t)}} \left[\left\| \mathbf{Y}_i - \mathbf{B}_i \boldsymbol{\theta}_\mu^{(t)} - \mathbf{B}_i \mathbf{\Theta}_{(t)} \boldsymbol{\mu}_g^{(t)} - \mathbf{B}_i \mathbf{\Theta}_{(t)} \hat{\gamma}_i^g \right\|^2 + \text{trace} \left(\mathbf{B}_i \mathbf{\Theta}_{(t)} \hat{V}_{ig} \mathbf{\Theta}_{(t)}^\top \mathbf{B}_i^\top \right) \right] \right\} + \beta_\sigma}{\frac{1}{2} \left[\sum_{i=1}^N n_i \right] + (\alpha_\sigma + 1)} \\ \nu_0^{(t+1)} = \frac{b_{\nu_0} + 1 + \sqrt{2b_{\nu_0} + 1}}{b_{\nu_0}} \quad \text{with } b_{\nu_0} = \exp \left(\frac{1}{N} \sum_{i=1}^N (\log \nu_i^{(t)} + 1/\nu_i^{(t)}) - 1 \right) \\ \nu_i^{(t+1)} = \frac{\left\{ \nu_0^{(t)} + \sum_{g=1}^G \frac{P_{ig}}{\sigma_{(t)}^2} \left[\left\| \mathbf{Y}_i - \mathbf{B}_i \boldsymbol{\theta}_\mu^{(t)} - \mathbf{B}_i \mathbf{\Theta}_{(t)} \boldsymbol{\mu}_g^{(t)} - \mathbf{B}_i \mathbf{\Theta}_{(t)} \hat{\gamma}_i^g \right\|^2 + \text{trace} \left(\mathbf{B}_i \mathbf{\Theta}_{(t)} \hat{V}_{ig} \mathbf{\Theta}_{(t)}^\top \mathbf{B}_i^\top \right) \right] \right\}}{n_i + 2 + \nu_0^{(t)}} \\ d_{jj}^{(t+1)} = \frac{1}{2} \left[\frac{(mG - m - 2) + \sqrt{(mG - m - 2)^2 + 4m \times (m - k - 1) \times \sum_{g=1}^G \left\{ \mathbf{\Gamma}_g^{-1(t)} \right\}_{jj}}}{(m - k - 1) \times \sum_{g=1}^G \left\{ \mathbf{\Gamma}_g^{-1(t)} \right\}_{jj}} \right] \quad (j = 1, \dots, k) \\ \left\{ \mathbf{\Gamma}_\mu^{(t+1)} \right\}_{jj} = \frac{1}{m + k + 1 + G} \left[\left\{ \sum_{g=1}^G \boldsymbol{\mu}_g^{(t)} \boldsymbol{\mu}_g^{(t)\top} \right\}_{jj} + (m - k - 1) \right] \quad (j = 1, \dots, k). \end{array} \right.$$

The matrix \mathbf{D} is assumed to be diagonal for model identifiability reasons. Indeed, as the variance of \mathbf{Y}_i is obtained as $Var(\mathbf{Y}_i) = [\mathbf{B}\Theta(I_k + \mathbf{D})\Theta^\top\mathbf{B}^\top + (\sigma^2\nu_0/(\nu_0 - 2))I_{n_i}]$, one needs to impose \mathbf{D} diagonal in addition to $\Theta^\top\Theta = I_k$ so that not only the term $[\Theta\mathbf{D}\Theta^\top]$ will be identifiable but both Θ and \mathbf{D} . The non-imposition of that constraint can lead to the existence of two different values of the set (Θ, \mathbf{D}) giving the same value for $[\Theta\mathbf{D}\Theta^\top]$. The matrix Θ produced by the procedure will not necessarily be orthonormal. We transform it into an orthonormal matrix through the Gram-Schmidt algorithm (Golub and Van Loan [16]). An additional diagonality constraint is imposed on Γ_μ for simplicity.

Initialization of the algorithm :

The parameters θ_μ , Θ and σ^2 are initialized by assuming a model with a single cluster (i.e., $G = 1$). Then for a fixed $G > 1$, the cluster parameters such as μ_g , Γ_g , and π_g as well as the cluster membership indicators \mathbf{z}_i may be initialized by applying any clustering procedure to the scores α_i yielded by the single-cluster model. In our experiments, we used the Gaussian model-based clustering procedure implemented in the *mclust package* (see Fraley and Raftery [13], Fraley and Raftery [14], Fraley and Raftery [15]).

3.2.4. Model selection

An important feature of our model is the principled manner in which the values of the number of clusters G , the dimension k of the principal component function $\hat{f}(t)$, and the dimension q of the B-spline basis are chosen. We develop an approximation of the log marginal likelihood (or marginal log-likelihood, MLL) that allows us to perform an MLL-based model selection. The MLL is a key quantity used to choose between different models within a Bayesian model selection paradigm through Bayes factors. The marginal log-likelihood is the quantity that results from integrating out both the latent variables and the parameters. Using our model specification of observed data, latent variables and vector of parameters, we have :

$$\text{MLL} = \log \left\{ \int_{(\mathbf{W}, \mathbf{\Pi})} p(\mathbf{Y} | \mathbf{W}, \mathbf{\Pi}) p(\mathbf{W}, \mathbf{\Pi}) d(\mathbf{W}, \mathbf{\Pi}) \right\} \quad (3.2.16)$$

Traditionally, the marginal likelihood is approximated either using analytical methods or via sampling-based approaches such as Markov chain Monte Carlo. Here, we decide to use a multivariate Laplace approximation which yields

$$\int_{(\mathbf{W}, \mathbf{\Pi})} p(\mathbf{Y} | \mathbf{W}, \mathbf{\Pi}) p(\mathbf{W}, \mathbf{\Pi}) d(\mathbf{W}, \mathbf{\Pi}) \approx p(\mathbf{Y} | \hat{\mathbf{W}}, \hat{\mathbf{\Pi}}) \times p(\hat{\mathbf{W}}, \hat{\mathbf{\Pi}}) \times (2\pi)^{\frac{d}{2}} \times |-\mathbb{H}|^{-\frac{1}{2}}, \quad (3.2.17)$$

where $(\hat{\mathbf{W}}, \hat{\mathbf{\Pi}})$ is the the MAP estimator of the latent variables and parameters, d is the dimension of $(\mathbf{W}, \mathbf{\Pi})$ and \mathbb{H} is the Hessian of $\log \{p(\mathbf{Y} | (\mathbf{W}, \mathbf{\Pi})) p(\mathbf{W}, \mathbf{\Pi})\}$ evaluated at $(\hat{\mathbf{W}}, \hat{\mathbf{\Pi}})$. Taking the log of the expression in Equation (3.2.17), we obtain

$$\text{MLL} \approx \log \{p(\mathbf{Y} | \hat{\mathbf{W}}, \hat{\mathbf{\Pi}})\} + \log \{p(\hat{\mathbf{W}}, \hat{\mathbf{\Pi}})\} + \frac{d}{2} \log(2\pi) - \frac{1}{2} \log(|-\mathbb{H}|). \quad (3.2.18)$$

We further simplify this expression by using the approximation $|-\mathbb{H}| \simeq \left[\frac{\text{trace}(-\mathbb{H})}{d}\right]^d$, which arises from assuming that all the eigenvalues of $-\mathbb{H}$ are the same and necessarily equal to $[\text{trace}(-\mathbb{H})/d]$. Finally the expression of the MLL is given by :

$$\text{MLL} = \log p(\mathbf{Y} | \hat{\mathbf{W}}, \hat{\mathbf{\Pi}}) + \log p(\hat{\mathbf{W}}, \hat{\mathbf{\Pi}}) - \frac{d}{2} \log \left(\frac{\text{trace}(-\mathbb{H})}{2\pi d} \right). \quad (3.2.19)$$

In order to choose a model, we evaluate this latter MLL quantity for each potential triplet (q, k, G) . The model selected is the one that maximizes $\text{MLL}(q, k, G)$. In our experiments, we compare this MLL criterion with two classical criteria : the Akaike information criterion (AIC), and the Bayesian information criterion (BIC). We recall here that these two criteria are given, respectively by

$$\begin{cases} \text{AIC} &= \log p(\mathbf{Y} | \hat{\mathbf{W}}, \hat{\mathbf{\Pi}}) - d \\ \text{BIC} &= \log p(\mathbf{Y} | \hat{\mathbf{W}}, \hat{\mathbf{\Pi}}) - \frac{d}{2} \log(N), \end{cases} \quad (3.2.20)$$

where N is the number of individuals and d is the number of model parameters.

3.3. EXPERIMENTS WITH SIMULATED AND REAL DATA

In this section we start by showing the results of a simulation carried out to study the performance of our model. We test two main aspects of the model : its ability to reproduce the original clusters, and its ability to predict the original curves. The simulation study is based on four parameters : the sample size $N \in \{100, 500, 900\}$; the spline basis dimension $q \in \{10, 15, 20\}$; the score coefficients dimension $k \in \{2, 4, 6\}$ and the number of clusters $G \in \{3, 6, 9\}$.

3.3.1. Simulation study for the one-dimensional model

In order to simulate data for fixed values of (q, k, G) , we generate the model parameters in the list $\{\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_G, \boldsymbol{\Gamma}_1, \dots, \boldsymbol{\Gamma}_G, \boldsymbol{\theta}_\mu, \boldsymbol{\Theta}, \mathbf{D}, \boldsymbol{\Gamma}_\mu, \pi_1, \pi_2, \dots, \pi_G, \nu_0, \sigma^2\}$ at random based on the assumptions and constraints of our model. We choose a vector of measurement time points of length 21. The q -dimensional vector $\boldsymbol{\theta}_\mu$ is generated as a sample of q realizations of a standard normal distribution. The matrix $\boldsymbol{\Theta}$ is generated from a normal distribution and is then orthonormalized in order to satisfy the orthogonality constraint. Each d_{jj} ($j = 1, \dots, k$) is sampled from an Inverse- χ^2 distribution to make the diagonal matrix \mathbf{D} . The matrix $\boldsymbol{\Gamma}_\mu$ is

obtained via the diagonal values of a sampled matrix from an Inverse-Wishart distribution with parameters $m = (k + 2)$ and $(m - k - 1)I_k$. The number of individuals in each cluster n_g is chosen at random, and the probabilities π_g are set to the proportions $(\frac{n_g}{n})$. Given known \mathbf{D} and $\mathbf{\Gamma}_\mu$, the vectors $\boldsymbol{\mu}_g$ are generated from a multivariate normal distribution. The matrices $\mathbf{\Gamma}_g$ are generated from an Inverse-Wishart distribution and are then used to sample the vectors $\boldsymbol{\gamma}_i^g$ for each individual. The values of σ^2 and ν_0 are generated from an Inverse- χ^2 distribution. Finally, the error terms are generated from a Student t distribution.

Simulation results :

For each combination of (N, q, k, G) , we create several data sets as explained above. We try different values of (q, k, G) and choose the best model using our MLL criterion. For comparison purposes, we also apply the AIC and BIC criteria for model selection. The quality of the results is assessed by comparing the partitions (clustering) created by the model and the original (true) cluster memberships. The comparison is performed with a single measure of similarity, the Adjusted Rand Index (ARI) (Rand [33], Hubert and Arabie [18]). A perfect agreement between the two partitions yields an ARI score of 1. The closest the score is to 1, the more similar the partitions are.

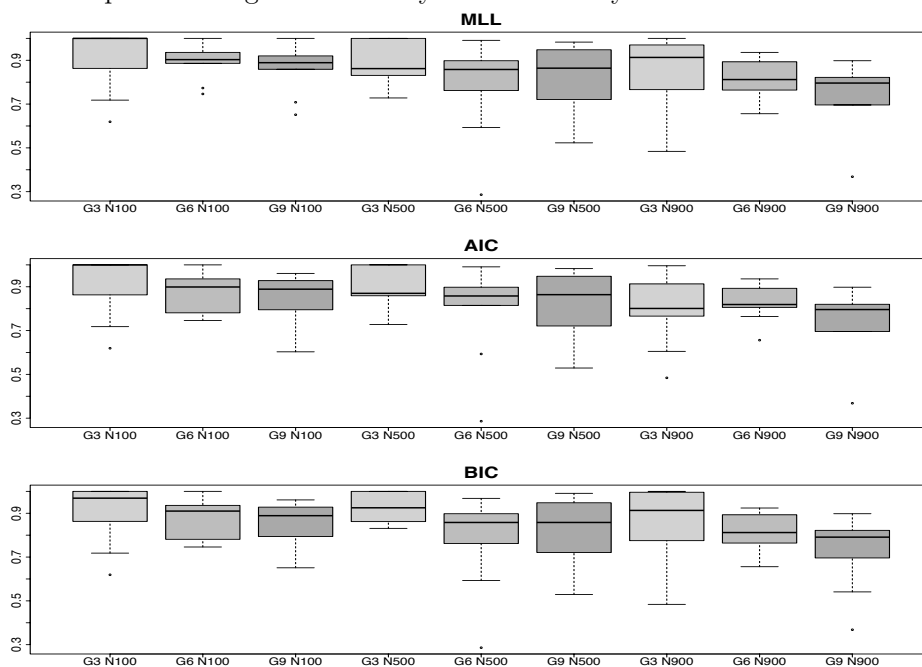
The examination of the simulation results is based on an analysis of variance of the ARI scores as a function of G, k, q and N . The ARI was first transformed to the inverse sinus function, $\arcsin(\text{ARI})$. The values of the two dimensional parameters k and q did not affect the ARI scores. Only the sample size N and the number of clusters were statistically significant. Figure 3.3 shows the results. Note that all three model selection criteria, MLL, AIC and BIC, performed very similarly. Clearly, the ARI score decreases with the number of clusters, as the data structure becomes more complex. There is no such clear pattern with the sample size. It appears that a large sample size makes the discovery of the cluster structure more difficult. However, this effect might be attributable to more curves filling in the space or gap between the clusters. So we decided to study the effect of the separation between clusters. We recall here two measures of separability between clusters : the inter-gap and the intra-gap. The inter-gap measure is defined as the mean distance between the cluster mean curves. The intra-gap measure as presented in Tibshirani et al. [37] is the pooled within-cluster sum of squares of the Euclidean distances to the cluster means. More explicitly, for N curves clustered in G clusters C_1, C_2, \dots, C_G with $n_g = |C_g|$, let $d_{i,i'}$ be the Euclidean distance between curves i and i' , $i, i' = 1, \dots, N$. Let $D_g = \sum_{i,i' \in C_g} d_{ii'}$ be the sum of the pairwise distances for all curves in cluster C_g . Let μ^g and μ^* be, respectively, the mean

curve associated with cluster C_g , and the overall mean curve associated with the ensemble of N curves. The inter-gap and intra-gap measures are given by the following expressions :

$$\text{InterGap} = \sum_{g=1}^G \pi_g d_{\mu^g, \mu^*} \quad \text{and} \quad \text{IntraGap} = \sum_{g=1}^G \frac{1}{2n_g} D_g. \quad (3.3.1)$$

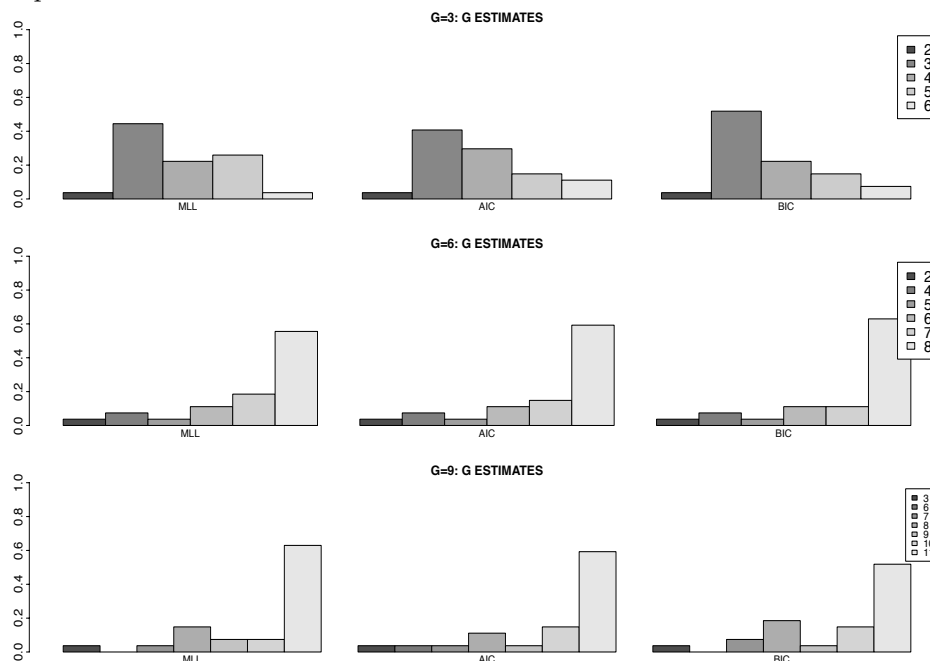
We measured the gap between the true clusters using the inter-gap measure, and the the gap within each true cluster with the intra-gap measure. Another ANOVA, this time with these two gap measures in the model, shows that after adjusting for the gaps, the sample size N is not significant (p -value = 0.20). However, G remains very significant (p -value < 0.005). We conclude that the ARI scores depend only on the number of clusters G , and the complexity of the data (given by the gap measures). Figure 3.4 shows how often the

FIGURE 3.3. One-dimensional model. Boxplots of the ARI scores for the models selected by MLL (top), AIC (middle) and BIC (bottom). The light grey boxes correspond to $G = 3$, whereas the darker grey ones correspond to $G = 9$. The middle grey boxes correspond to $G = 6$. N stands for sample size. The boxplots are organized first by N and then by G .



model selection criteria chose the right number of clusters. Again, all three criteria seem to have performed similarly : there is a clear tendency to overestimate the number of clusters.

FIGURE 3.4. One-dimensional model. Proportion of times each criteria chose a particular number of clusters.



3.3.2. Simulation study for the two-dimensional model

Another simulation study is carried out on the two-dimensional model. The generation of the simulated data is analogous to that described in Section 3.3.1. The only difference is in the generation of the parameters θ_μ , Θ , \mathbf{D} and Γ_μ for which the dimensions change. Since the results from the one-dimensional model indicate that the dimension parameters k and q as well as the sample size N are not significantly affecting the ARI scores, we fixed the values of the dimension parameters q and k ($q = 10; k = 2$), and generated five data sets of size $N = 100$ with varying number of clusters $G = 2, 4, 6, 8$ and $G = 10$, respectively. We repeat this procedure ten times.

Figure 3.6 shows a simulated dataset with $N = 100$ curves generated from the two-dimensional model with $q = 10$, $k = 2$ and $G = 6$. As in the one-dimensional model simulation, the assessment of the clustering performance is based on the ARI score between the known true partition and the estimated partition from the three model selection criteria (MLL, AIC and BIC). For example, for the dataset depicted in Figure 3.6, the three criteria perfectly selected the correct model with $q = 10$, $k = 2$, $G = 6$ yielding a very high ARI of 0.931. The box plots in Figure 3.5 show the simulation results. The ARI scores are

relatively high especially for smaller number of clusters. As in the one-dimensional case, the ARI scores also decrease with the number of clusters.

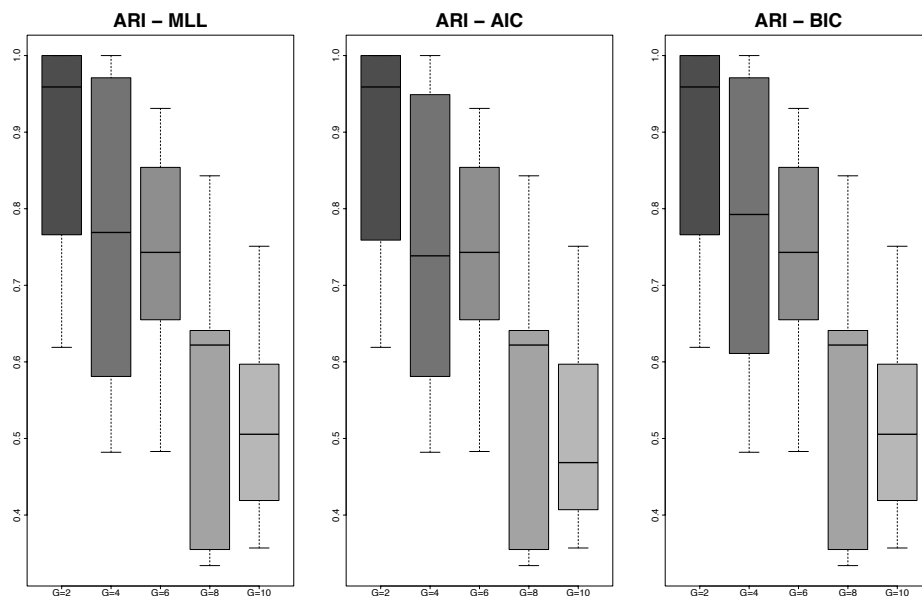


FIGURE 3.5. Two-dimensional model. Box plots of the ARI scores for the models selected by MLL (left), AIC (middle) and BIC (right).

3.3.3. Comparison study with real datasets

In this section we test our functional model-based clustering model on real datasets which have already been analyzed by several authors such as James and Sugar [23], McNicholas and Murphy [30] and Jacques and Preda [20, 21]. These authors have applied their own models for clustering longitudinal data. Here we report our results as well as theirs. We also present an application to the well-known Yeast Cell Cycle data Cho et al. [7] and compare our results to those of other researchers that have also used these data for comparison purposes.

3.3.3.1. *The Rats data*

The Rats dataset has been studied in a setting of longitudinal model-based clustering by McNicholas and Murphy [30]. The data were published in Crowder and Hand [10]. They consist of the body weights of rats on one of three different dietary supplement treatments. There were eight rats on Diet 1, four on Diet 2 and four on Diet 3. Weights were recorded first after a settling-in period, and then weekly for a period of nine weeks. An extra measurement was taken at 44 days to help calculate the effect of another treatment that occurred during

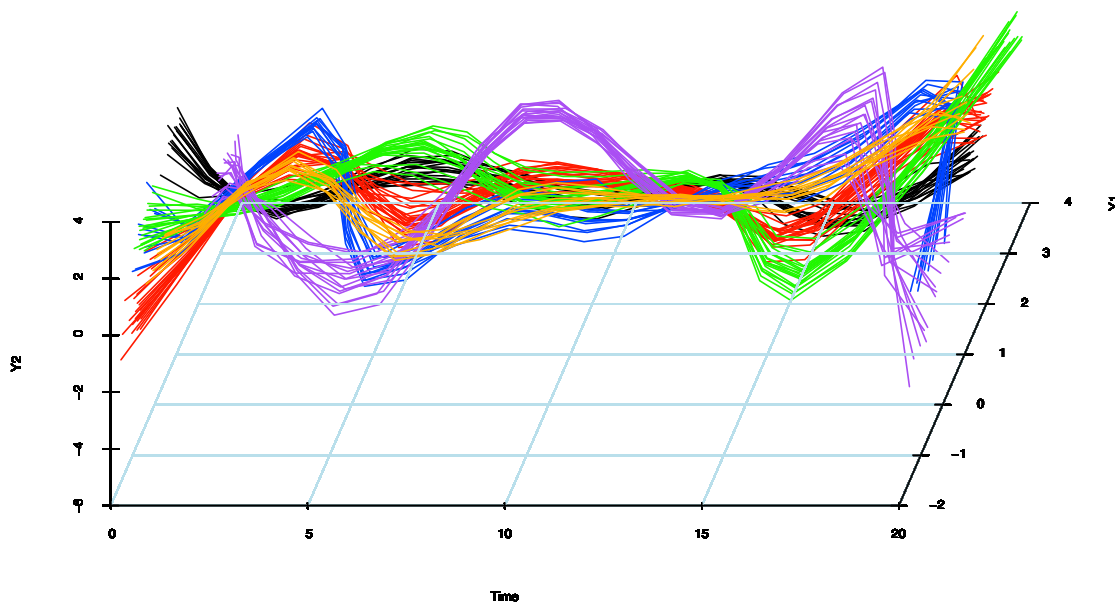


FIGURE 3.6. Two-dimensional model. Example of a dataset with six clusters

the sixth week. A total of eleven measurements were recorded for each rat. Following our proposed model selection methodology, we found that the best model consists of four groups of sizes eight, three, one and four. Our model yielded an ARI of 0.94. For the same dataset, McNicholas and Murphy [30] reported an ARI of 0.88. Figure 3.7 shows the original and predicted curves together with the original and estimated clusters. The estimated one-rat cluster represents an « estimated » outlier rat with a much larger weight than the other rats in the original cluster.

3.3.3.2. Growth data

The Growth data, available in the *fda* package of the software R, comes from the Berkeley growth study (Tuddenham and Snyder [38]). In this dataset, the heights of 54 girls and 39 boys were measured at 31 stages, from 1 to 18 years. The goal is to cluster the growth curves and to determine whether the resulting clusters reflect gender differences. Our selection procedure chose a model with two clusters. The dimensions (q, k) chosen by the different criteria were different. MLL and BIC chose a model with dimensions $(q = 18, k = 2)$, whereas AIC chose a model with $(q = 19, k = 2)$. Overall, these models were very similar. That is, the value of the dimension q did not make a big difference. Therefore, we adopted the smaller model which yielded an ARI of 0.513 and a correct classification rate of 86.02%. Figure 3.8 shows the original and predicted curves together with the original and estimated

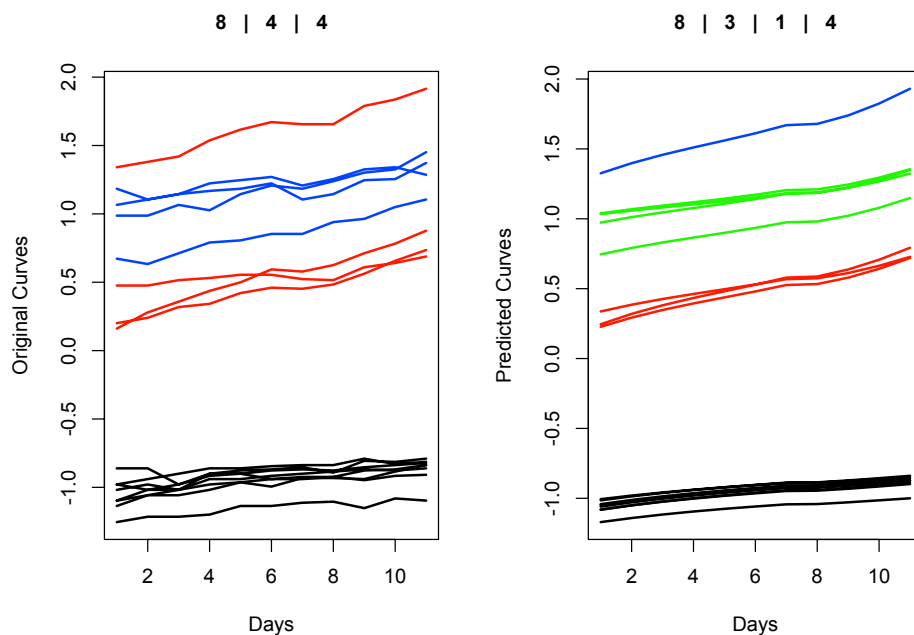


FIGURE 3.7. Original (left) and Predicted (right) curves for the Rats dataset

clusters. For the original partition, the black curves represent the 39 boys and the red curves represent the 54 girls. The clustering algorithm yields two clusters of sizes 50 and 43. With regard to the partition matrix (see Table 3.1) we can say that the estimated cluster 1 of size 50 (predicted curves in black) consists mostly of boys and the estimated cluster 2 of size 43 (predicted curves in red) consists mostly of girls. A partition matrix obtained from two partitions \mathcal{C}_1 and \mathcal{C}_2 is a matrix PM where an element PM_{ij} represents the number of individuals in cluster i of \mathcal{C}_1 that also fall in cluster j of \mathcal{C}_2 . For the same dataset, Jacques

TABLE 3.1. Partition matrix for Growth data

True clusters	Estimated clusters	
	1	2
Boys	38	1
Girls	12	42

and Preda [20, 21] using their *Funclust* procedure report a correct classification rate of 69.89%. They also evaluated the *fclust* procedure developed by James and Sugar [23]. This also yielded a a correct classification rate of 69.89%.

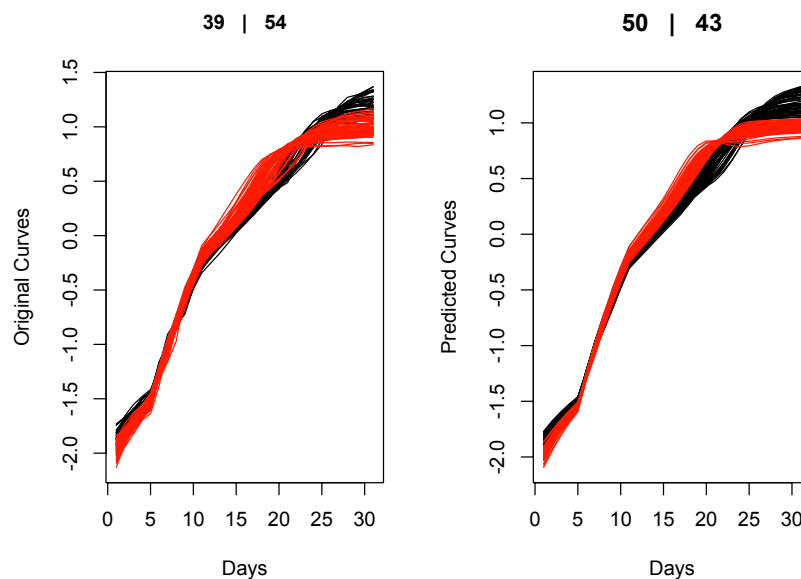


FIGURE 3.8. Original (left) and Predicted (left) curves for the Growth dataset

3.3.3.3. ECG data

The electrocardiogram (ECG) database studied in Olszewski [32] contains measurements of cardiac electrical activity as recorded from electrodes at various locations on the body. The ECG database was taken from <http://www.cs.ucr.edu/eamonn/timeseriesdata/>, which is the website of the UCR Time Series Classification and Clustering. It contains 200 data sets sampled at 96 time instants where 133 were identified as normal and 67 were identified as abnormal. Jacques and Preda [20, 21] reported a correct classification rate of 84% with the *Funclust* procedure and a correct classification rate of 74.5% for the *fclust* procedure. Our analysis of the ECG dataset found three groups which yielded an ARI of 0.84 and a correct classification rate of 78%.

3.3.3.4. The yeast cycle data

Longitudinal data are in general sparse (unbalanced) and the functional model-based clustering proposed herein is suitable for the analysis of all types of longitudinal data. The model is particularly well suited for the analysis of time-course data sets such as gene expression data. We present the results of the application of our model to the yeast cell cycle dataset. This records the fluctuations of the expression levels of about 6000 genes over two cell cycles comprising 17 time points. We consider the 5-phase subset of the data in Cho et al. [7]. It

consists of 386 genes which have been assigned to one of the five phases of the cells cycle. The five phases are estimated by experts. The clustering results should reveal five groups of genes associated with the five phases. However, to our knowledge, there is no clustering procedure that can automatically reproduce these phases adequately. Our model selection results for each of the three criteria are shown in Figure 3.9.

Each bar on the Figure 3.9 indicates the value on the y-axis of the criterion (AIC, BIC or MLL from top to bottom) for the triplet (q, k, G) where the couple (q, k) is shown on the x-axis and the number of clusters G is displayed by the color of the bar. The objective of the figure is to reveal, for a given criterion, the best model (or the best triplet (q, k, G)) through the highest bar (the one with the highest y-axis value).

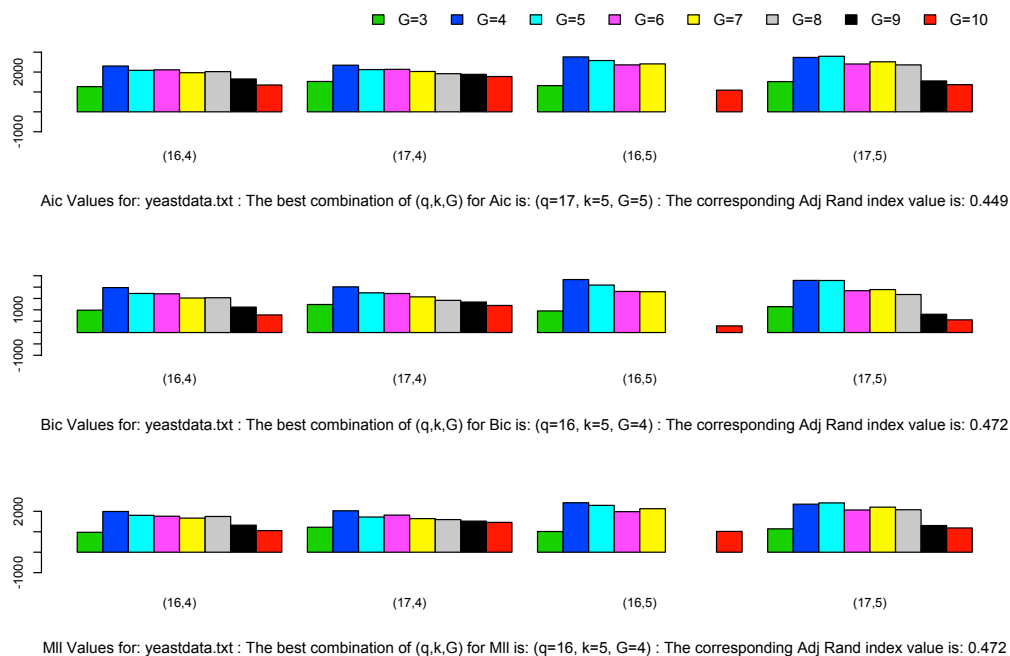


FIGURE 3.9. Model selection results for yeast cell cycle data

The AIC criterion indicates a five-cluster partition with an associated ARI 0.45, whereas both criteria MLL and BIC suggest a four-cluster partition with an associated ARI of slightly over 0.47. These results are highly comparable to those obtained by other studies on the same data set. Indeed, The Potts model clustering of Murua et al. [31] yielded nine clusters with an ARI of 0.45. Yeung et al. [42] analyzed the same subset of these data using model-based clustering based on Gaussian mixtures developed in Banfield and Raftery [2]. They reported four clusters with an ARI of about 0.43. Figure 3.10 shows the four mean

curves associated with the four-cluster solution yielded by the MLL and BIC criteria. The figure displays the observed and the estimated mean curves. For comparison purposes, we show in Figure 3.11 the five mean curves associated with the five original clusters proposed by Cho et al. [7]. The overall mean of the curves is displayed in the left panel of Figure 3.12.

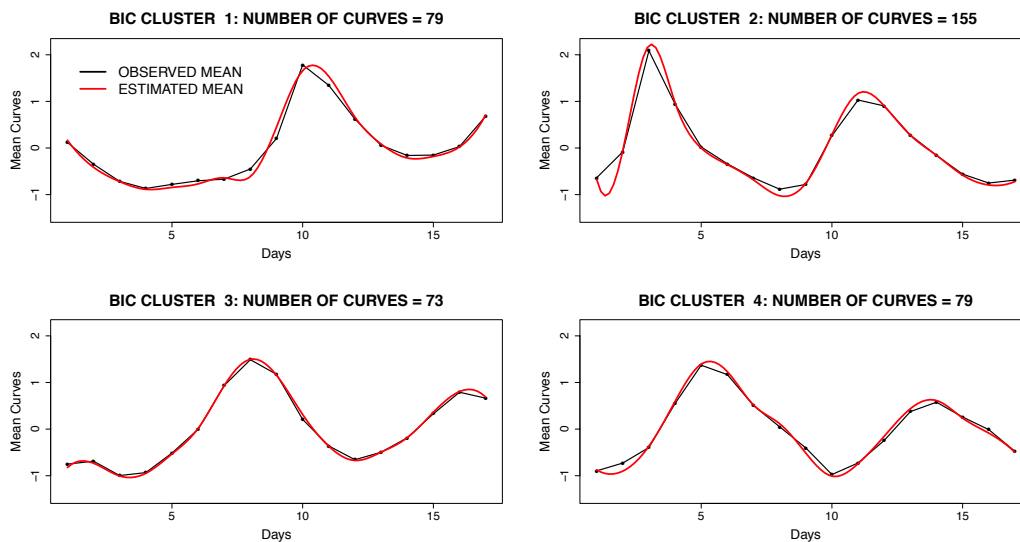


FIGURE 3.10. Yeast cycle data. Observed and model-estimated mean curves for the four clusters yielded by the MLL and BIC criteria

Observe that the mean curves associated with the four estimated clusters are very similar to the first four clusters found by Cho et al. [7]. The fifth cluster of Cho et al. [7] lies between clusters two and four of the estimated clusters. The covariance structure of the four estimated clusters is displayed in Figure 3.1. Observe that there is high correlation between time points close to valleys and peaks in the mean curves. One can also observe a slightly negative correlation for points further away from the valleys and peaks. This pattern is also observed in the overall covariance structure displayed in Figure 3.2 with respect to the overall mean curve of the data (see left panel of Figure 3.12). The estimated degrees of freedom were $\nu_0 = 3$. The right panel of Figure 3.12 displays the distribution of the ν_i variables associated with each observation. Recall that the error terms ϵ_i were modeled as a convolution of Normal and Inverse- χ^2 distributions, so that small values of ν_i give evidence of non-normally distributed errors.

3.4. APPLICATION TO THE PRRS VIREMIA DATASET

This section shows the analysis of the results of the application of the two-dimensional (2D) model to the porcine reproductive and respiratory syndrome virus (PRRSV) described in

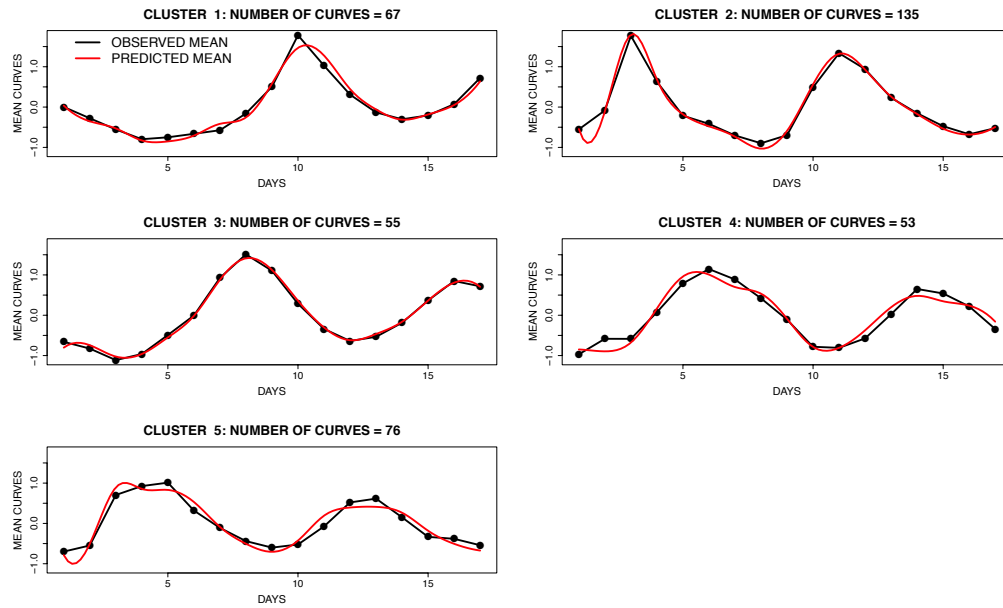


FIGURE 3.11. Yeast cycle data. Observed and model-estimated mean curves for the five clusters found by Cho et al. [7]

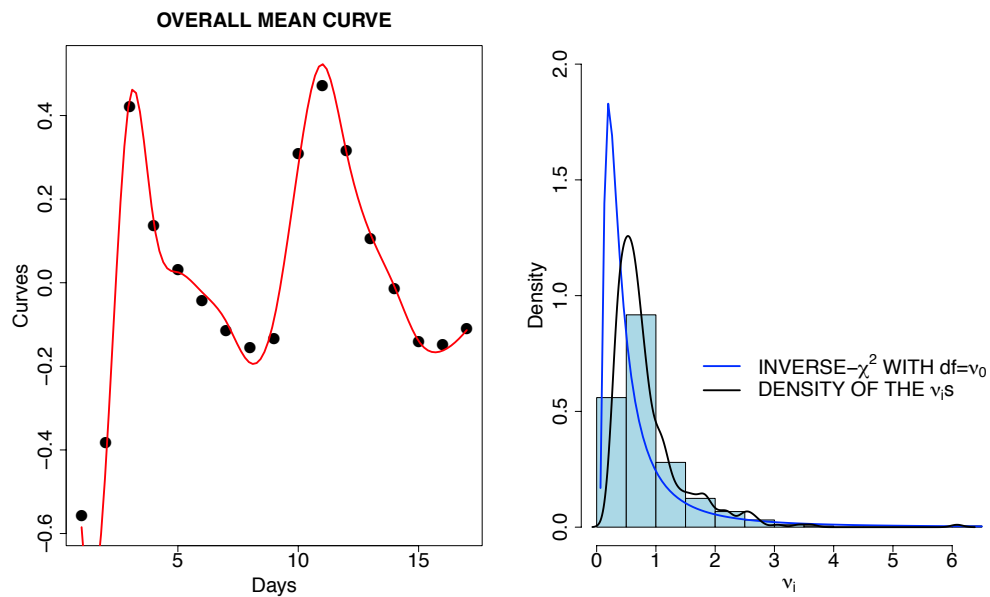


FIGURE 3.12. Yeast data. Overall mean curves (left) and distribution of the ν_i associated with the error term distribution of the data.

the introduction. Recall that the dataset consists of longitudinal data from pigs experimentally infected with the virus PRRSV. Each pig (1331 total, coming from high health farms that were free of PRRSV, *Mycoplasma hypopneumoniae* and swine influenza virus) has two responses measured, not necessarily, in fact often not, at the same time points : *Virus load* and *Weight*. We worked with the logarithmic transform of *Virus load*, which we will be denoted by $lVirus$. The records of each individual have been standardized by overall mean and overall standard deviation. A somehow strange feature in the dataset is the presence of 175 pigs for which the last measure of *Weight* has been recorded at day 42 instead of day 40 as it is the case for the other 1156 pigs. A consequence of this feature is the presence of a sharp negative slope in the mean curves from day 40 to day 42 which can be seen in Figure (3.13). The left panel of Figure (3.13) shows the mean curve of all 1331 pigs with the emphasis on the slope. The middle panel shows the mean curve for the 175 pigs with the last measurement at day 42, and the right panel displays the mean curve of the 1156 pigs whose last measurement was at day 40.

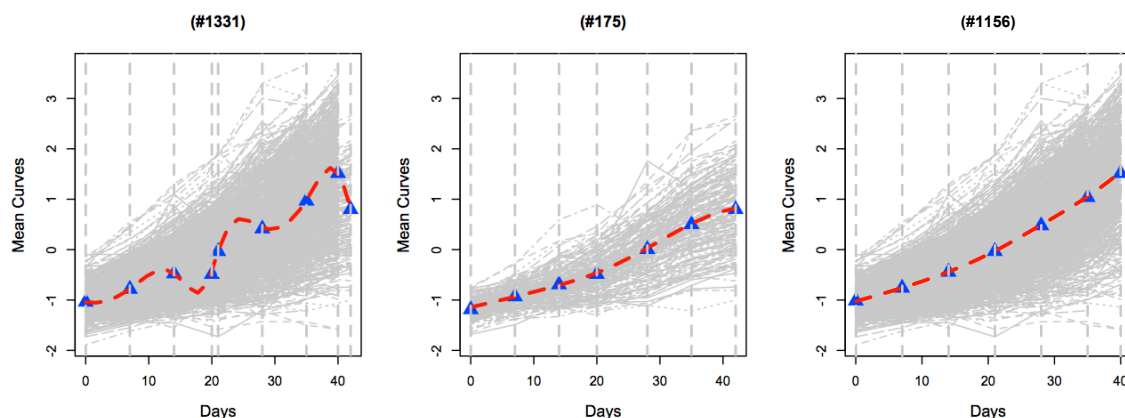


FIGURE 3.13. Illustration of the difference at days 40 and 42

In addition to the existing model and as specifically requested by the data structure, we develop a version of the two-dimensional model which explicitly incorporates some fixed effects. We will refer in this section to these versions of our model as mixed-effects models. For the dataset in study, the two fixed effects considered are the variables : Wur (for locus WUR10000125, a major genetic determinant of PRRSV disease studied in Boddicker et al. [3]) with 3 categories $\{0,1,2\}$ (associated to the genotypes 1, 2, 3, respectively), and $Experiment$ with 7 categories $\{1,2,3,4,5,6,7\}$. They can be introduced in the model separately or simultaneously. The main difference from the original 2D model lies in the treatment of the overall mean in the new model. Unlike the model presented in Equation (3.2.11), where the

vector $\boldsymbol{\theta}_\mu$ is constant, the mixed-effects model assumes a decomposition of that vector into vectors representing the fixed effects. Indeed, $\boldsymbol{\theta}_\mu$ is now an individual-indexed parameter which incorporates the the fixed effects. The following four models have been considered for analyzing the PRRSV data :

- (1) The model M_0 is the original 2D model including no fixed effects.
- (2) The model M_1 is the 2D model including the *Wur* fixed effect. The overall mean $\boldsymbol{\theta}_\mu$ in the M_0 model is now an individual-indexed parameter which incorporates the *Wur* effect :

$$\mathbb{Y}_i = \mathbb{B}_i \underbrace{(\boldsymbol{\theta}_\mu^{gen} + \mathbf{S}_{w_i})}_{\boldsymbol{\theta}_\mu^i} + \mathbb{B}_i \boldsymbol{\Theta} \boldsymbol{\mu}_{z_i} + \mathbb{B}_i \boldsymbol{\Theta} \boldsymbol{\gamma}_i^{z_i} + \boldsymbol{\Upsilon}_i,$$

where $w_i \in \{0, 1, 2\}$ are effects indicators. The category $w_i = 0$ is considered to be reference, meaning that the vector \mathbf{S}_0 is null. Thus, for individuals with $w_i = 0$, $\boldsymbol{\theta}_\mu^i = \boldsymbol{\theta}_\mu^{gen}$.

- (3) The model M_2 is the 2D model including the *Experiment* fixed effect wich takes values ranging from 1 to 8. The overall mean $\boldsymbol{\theta}_\mu$ in the M_0 model is now an individual-indexed parameter which incorporates the *Experiment* effect :

$$\mathbb{Y}_i = \mathbb{B}_i \underbrace{(\boldsymbol{\theta}_\mu^{gen} + \mathbf{T}_{e_i})}_{\boldsymbol{\theta}_\mu^i} + \mathbb{B}_i \boldsymbol{\Theta} \boldsymbol{\mu}_{z_i} + \mathbb{B}_i \boldsymbol{\Theta} \boldsymbol{\gamma}_i^{z_i} + \boldsymbol{\Upsilon}_i,$$

where $e_i \in \{0, 1, 2, 3, 4, 5, 6, 7\}$ are the effect indicators. The category $e_i = 0$ which corresponds to *Experiment* = 8 in the study, is considered to be reference, meaning that the vector \mathbf{T}_0 is null. Thus, for individuals with $e_i = 0$, $\boldsymbol{\theta}_\mu^i = \boldsymbol{\theta}_\mu^{gen}$.

- (4) The model M_{12} is the 2D model including both *Wur* and *Experiment* fixed effects. The model is expressed as

$$\mathbb{Y}_i = \mathbb{B}_i \underbrace{(\boldsymbol{\theta}_\mu^{gen} + \mathbf{S}_{w_i} + \mathbf{T}_{e_i})}_{\boldsymbol{\theta}_\mu^i} + \mathbb{B}_i \boldsymbol{\Theta} \boldsymbol{\mu}_{z_i} + \mathbb{B}_i \boldsymbol{\Theta} \boldsymbol{\gamma}_i^{z_i} + \boldsymbol{\Upsilon}_i,$$

where $w_i \in \{0, 1, 2\}$ and $e_i \in \{0, 1, 2, 3, 4, 5, 6, 7\}$ are the effects indicators. The categories $w_i = 0$ and $e_i = 0$ (corresponding to *Experiment* = 8) are considered to be references ($\mathbf{S}_0 = \mathbf{T}_0 = \mathbf{0}$). For individuals with $w_i = 0$ and $e_i = 0$, $\boldsymbol{\theta}_\mu^i = \boldsymbol{\theta}_\mu^{gen}$.

In the models M_1 , M_2 and M_{12} the additional parameters $\boldsymbol{\theta}_\mu$; $\mathbf{S}_1, \mathbf{S}_2$; $\mathbf{T}_1, \mathbf{T}_2, \dots, \mathbf{T}_7$ are to be estimated. There's no substantial change in the EM steps except in the equations involving $\boldsymbol{\theta}_\mu$. For illustration purposes, the parameter estimation calculations associated with the model M_{12} are shown in Appendix A.3. For each model considered, the first step is the identification of the optimal model through model selection based mainly on the criterion MLL (or AIC or BIC). Recall that a model is identified by the three parameters (q, k, G) ,

respectively the dimension of the B-spline basis, the dimension of the principal component function and the number of clusters. For a given criterion, optimal models are compared using either the adjusted Rand index (ARI) or the kappa coefficient (Cohen [8]). This latter index is also a measure of the agreement between partitions but requires that the partitions have equal number of classes (see Youness and Saporta [43], Reilly et al. [34] and Warrens [40] for studies related to the kappa coefficient and some comparison with the ARI in cluster analysis). As the partitions we compare are not necessarily equal in the number of classes, we use a transformation of the partition matrix (obtained from two partition tables) for the computation of the kappa statistic. The purpose of using this index is to find out if a smaller partition could be viewed as a partition formed by merging clusters from a larger partition, or inversely, if a larger partition could be viewed as a partition formed by splitting clusters in the smaller partition. Let \mathcal{C}_1 and \mathcal{C}_2 be two partitions (clustering) of the same dataset with r_1 and r_2 clusters, respectively. The associated $r_1 \times r_2$ partition matrix PM is a matrix with entries PM_{ij} = number of individuals in cluster i of \mathcal{C}_1 and cluster j of \mathcal{C}_2 . The transformation of PM makes it a square matrix in the following manner. Let us suppose, without loss of generality, that $r_1 < r_2$. If $r_1 = r_2$, then there is nothing to transform. For each column PM_j ($1 \leq j \leq r_2$), the row with the maximum number of individuals denoted j_{max} ($1 \leq j_{max} \leq r_1$) is identified. After a sweep of all columns, all those clusters with equal j_{max} are merged. This ensures that the transformed partition \mathcal{C}_2 contains at most r_1 clusters. Table 2 show an illustration of this procedure. From now on we will use the notation xG -partition to denote a partition with x clusters (recall that G stands for the number of clusters).

TABLE 3.2. Illustration of the partition matrices for the computation of the kappa coefficient

Real partition matrix						Transformed partition matrix			
3G-partition	5G-partition						3G-partition		
	1	2	3	4	5		1	2	3
1	142	25	105	0	0	1	142	130	0
2	0	155	228	0	272	2	0	655	0
3	0	9	0	181	214	3	0	223	181

For the interpretation of the kappa coefficient values, we use a commonly cited scale based on the table of interpretation given by Landis and Koch [26]. Next, we first explore the best models selected and study the effect, if any, of *Wur* and *Experiment* in the clustering results. Once the best models are selected, we study the effect of each fixed effect variable in the logarithm of the *Virus load* and *Weight*.

The M_0 model : The best three M_0 models based on the MLL criterion are, in order :

(8, 5, 3), (8, 5, 11) and (8, 4, 7). In addition, the kappa coefficient values between the 3G-partition and the other two (7G-partition and 11G-partition) are respectively 0.60 and 0.82, indicating substantial agreement or similarity between the small partition and the two larger ones. The 3G-partition is the optimal M_0 model according to the MLL (it is also the best choice for the other two criteria AIC and BIC). The related original and estimated mean curves are presented in Figure (3.14). The smoothed curves in the left panels (and in each

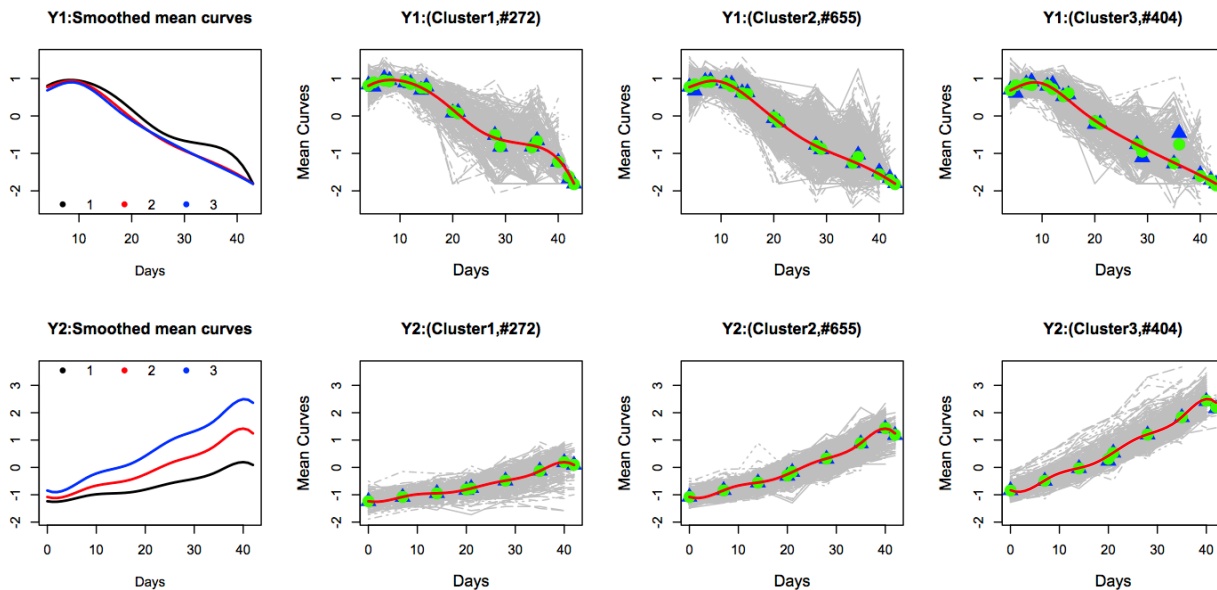


FIGURE 3.14. The M_0 3G-partition mean curves

cluster) are the estimated mean curves for each cluster defined as : $[C_g = B_i\theta_\mu + B_i\Theta\mu_g]$.

The M_1 model : The best three M_1 models based on the MLL criterion are, in order : (8, 5, 14), (8, 5, 3) and (8, 5, 12). The MLL criterion clearly chooses the 14G-partition as optimal. The mean curves presented in Figure (3.15), are very similar for all categories of Wur within each cluster. In these plots, the colors indicate the cluster and the line types indicate the effect. However, the kappa coefficient values between the 3G-partition and the two larger ones (14G-partition and 12G-partition) are respectively 0.8 and 0.75, indicating high similarity between the partitions. The 3G-partition is the optimal M_1 model according to the other criteria AIC and BIC. The ARI and kappa coefficient between the M_1 3G-partition and the M_0 3G-partition are respectively 0.98 and 0.99, suggesting that Wur has no effect on the clustering. This result seems to be confirmed in Figure (3.16), where the cluster mean curves are very similar for all categories of Wur . In these plots, the colors indicate the clusters and the line types indicate the effect.

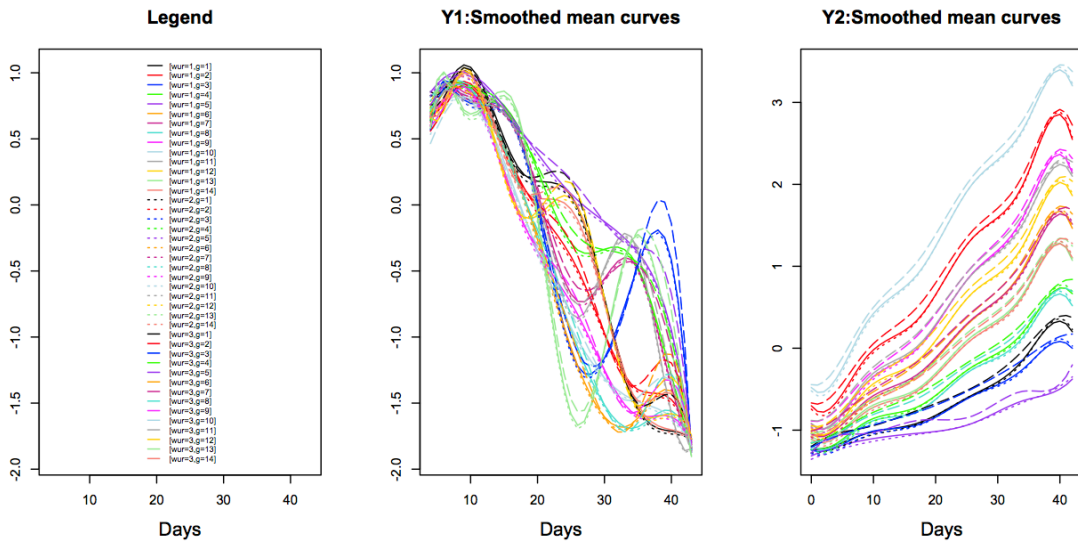


FIGURE 3.15. The M_1 14G-partition mean curves by Wur category

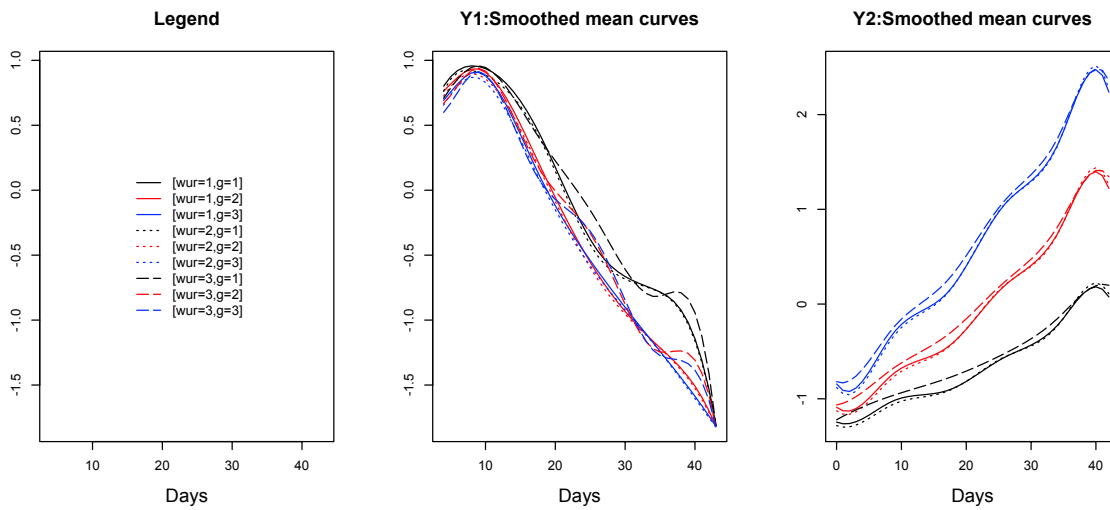


FIGURE 3.16. The M_1 3G-partition mean curves by Wur category

The M_2 model : The best three M_2 models based on the MLL criterion are, in order :

(8, 5, 8), (8, 5, 11) and (8, 5, 3). The MLL criterion clearly chooses the 8G-partition as optimal. The kappa coefficient values between the 3G-partition and the two larger ones (8G-partition and 11G-partition) are respectively 0.6 and 0.7, indicating substantial similarity between the partitions. The ARI between the M_2 8G-partition and the M_0 3G-partition is 0.14, suggesting that the *Experiment* has an effect on the clustering (the large number of clusters and categories makes it difficult to appropriately represent the mean curves). The 3G-partition is the optimal M_2 model for the other criteria AIC and BIC. The ARI between the M_2 3G-partition and the M_0 3G-partition is 0.32, again suggesting, as with the M_2 8G-partition, that the *Experiment* has an effect on the clustering results. This fact is confirmed in Figure (3.17), where the cluster g ($g = 1, 2, 3$) mean curves are clearly separated from one *Experiment* category to another, especially according to *Weight* (the \mathbf{Y}_2 variable in the plot). In these plots, the colors indicate the effect and the line types indicate the clusters.

The M_{12} model : The best three M_{12} models based on the MLL criterion are, in order : (8, 5, 3), (8, 5, 12) and (8, 4, 4). The kappa coefficient values between the 3G-partition and the other two (12G-partition and 4G-partition) are respectively 0.77 and 0.88, indicating substantial similarity. The 3G-partition is then considered the optimal M_{12} model (it is, in addition, the best choice for the other criteria AIC and BIC). The ARI between the M_{12} 3G-partition and the M_0 3G-partition is only 0.32, suggesting that both variables have an effect on the clustering results. However, the M_{12} 3G-partition is strongly similar to the M_2 3G-partition with an ARI of 0.81 and a kappa coefficient value of 0.89. This confirms that *Wur* has no effect on the clustering. The cluster mean curves related to the M_{12} 3G-partition are presented in Figures (3.18), (3.19) and (3.20). In these plots, the colors indicate the *Experiment* category and the line types indicate the *Wur* category. It can be seen from the plots that for each cluster, and for each *Experiment* category, all the curves with different line types are strongly similar. However, the curves are distinct according to the color, indicating the effect on clustering of the *Experiment* category.

All the optimal partitions from all the models (M_0 , M_1 , M_2 and M_{12}) have estimated $q = 8$ and $k = 5$ with various number of clusters. For the purpose of comparison, Figure (3.21) shows the variation of the MLL values according to the number of clusters. The best model chosen by MLL is the 3G-partition model M_{12} with $q = 8$ and $k = 5$. To investigate if *Wur* and/or *Experiment* are significant in the model, we use the approximation to the logarithm of the Bayes factors given by the difference between MLL associated to the corresponding models. Following the results above on clustering, we only compare the 3G-partitions associated with each model. We see that *Experiment* is significant when adjusted by the presence

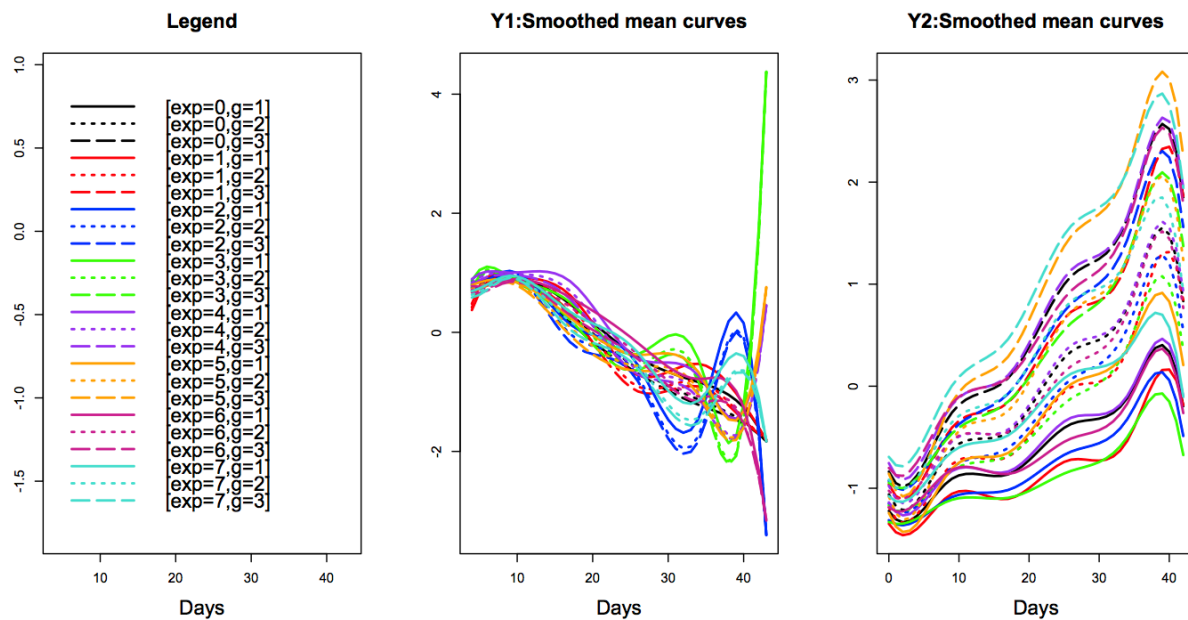


FIGURE 3.17. The M_2 3G-partition mean curves by *Experiment* category

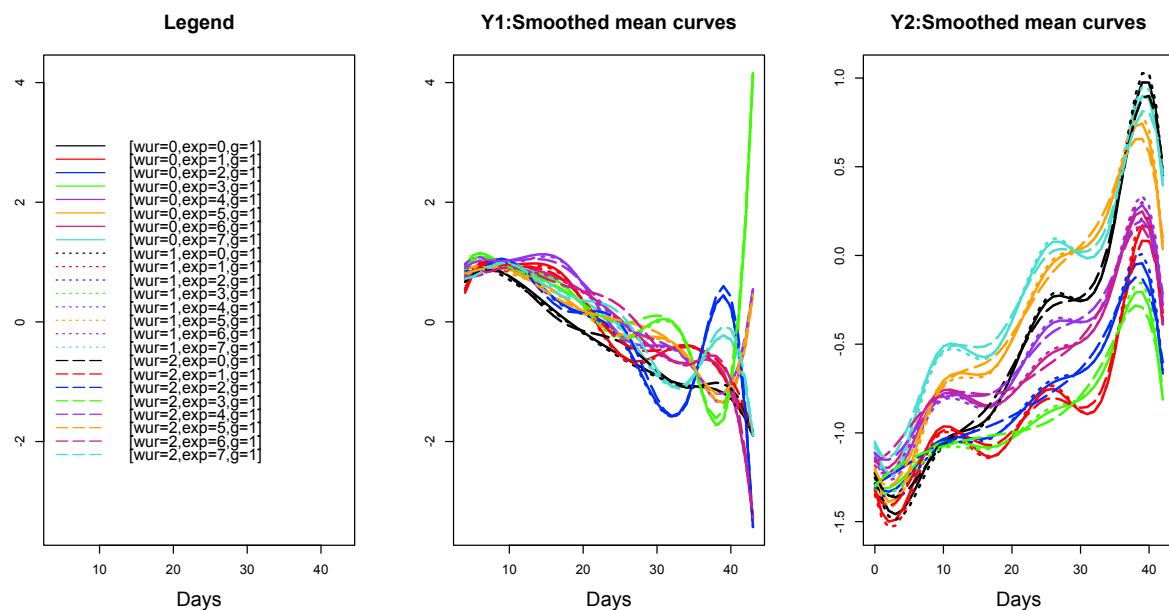


FIGURE 3.18. The M_{12} 3G-partition mean curves by *Wur* and *Experiment* for cluster 1

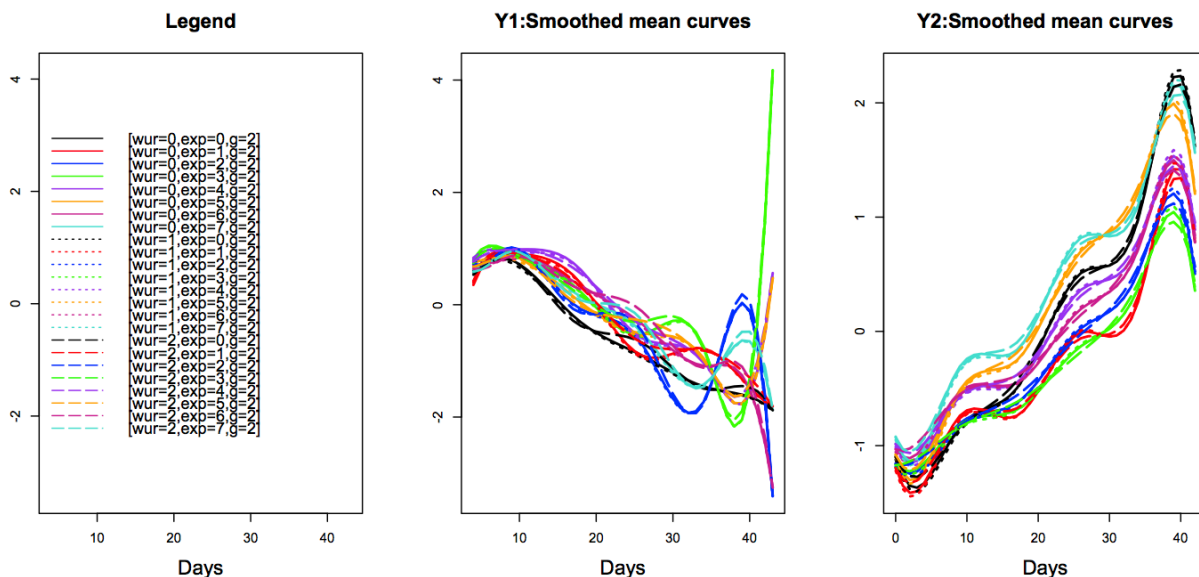


FIGURE 3.19. The M_{12} 3G-partition mean curves by Wur and $Experiment$ for cluster 2

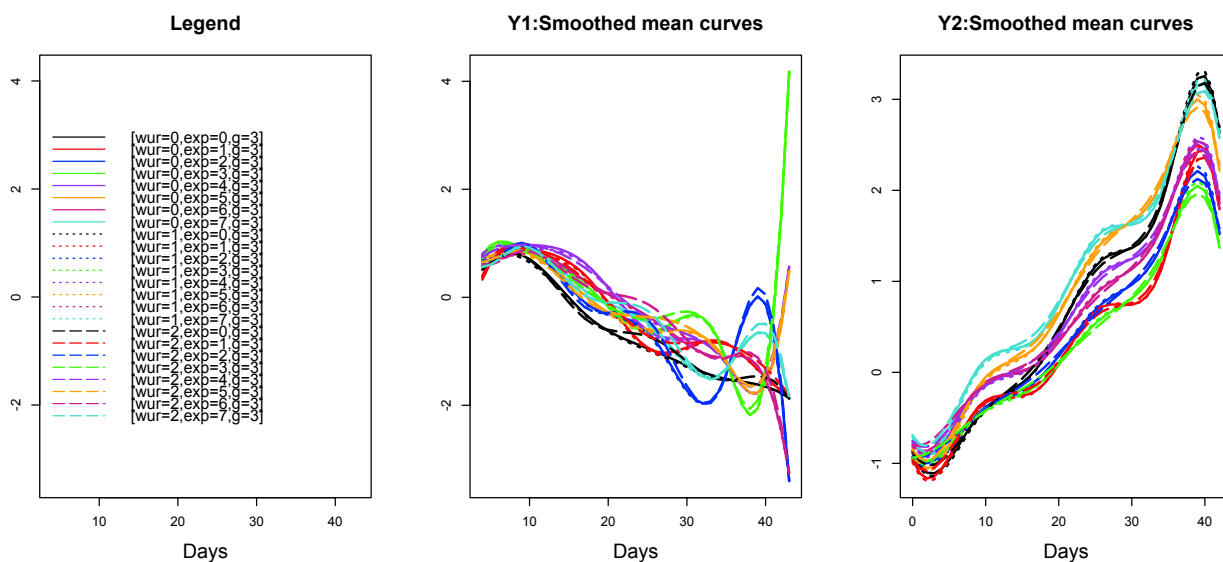


FIGURE 3.20. The M_{12} 3G-partition mean curves by Wur and $Experiment$ for cluster 3

of Wur with $MLL(M_{12}) - MLL(M_1) = 649.40$, and also that Wur is significant after adjusting for $Experiment$, with $MLL(M_{12}) - MLL(M_2) = 237.42$. Therefore, there is strong evidence that both variables have an effect on the response.

However, as noted above, *WUR* has no effect on the clustering, while *Experiment* does. So whatever is driving the clustering, this is somehow related and interacting with *Experiment* but not with *WUR*. Therefore, it seems that the effect of *WUR* is purely additive. Looking at the figures with the curves by cluster, it appears that the strongest effect of *Wur* is associated with pigs with genotype 3 (that is, $Wur = 2$).

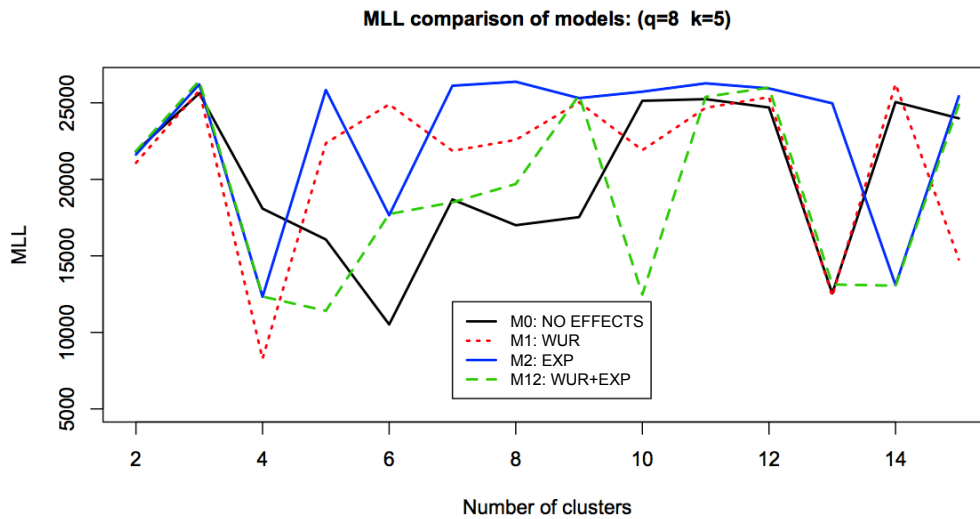


FIGURE 3.21. MLL Comparison for model selection

3.5. CONCLUSIONS AND DISCUSSION

The PHGC trials were undertaken to identify host genetic determinants of PRRSV infection outcome. There was hope that a few genetic loci would strongly control the host response, so that PRRSV resistance could be bred into the next generation of pigs. If a few alleles control PRRSV infection and these alleles segregate in the tested populations, then the pigs should separate into clusters with distinct disease trajectories. Of course, there could be other factors, both known and unknown, that drive clustering.

Our analysis of the eight trials finds at least three clusters, possibly attributable to, as yet, unknown factors. In addition, we find that experimental trial has a large effect on disease trajectories. Since experimental trials differed by time of year, breeding company supplier of pigs, breed of pig, and more, there are many possible and often confounded biological determinants of this experiment-to-experiment variation. Finally, in agreement with repeated evidence across these trials that locus WUR10000125 is a major genetic determinant of PRRSV disease (Boddicker et al. [3]; Boddicker et al. [4]; Boddicker et al. [5]), we confirm a major effect of this locus. The agreement on the WUR10000125 finding is not surprising since this locus was identified as highly significant from among 56118 SNPs after correction for multiple testing (Boddicker et al. [3]). Previous work has only considered a univariate response, either area under the virus load curve between days 0 and 21 (AUC21) or weight gain between days 0 and 42 (WG42). Our model shows a clear effect of WUR10000125 between days 7 and 21, which matches quite well with their chosen univariate measure. However, as shown in Figure 5 of Boddicker et al. [3], which analyzed only trials 1 through 3, the effect of WUR10000125 on weight gain trajectories was small compared to the differences in our cluster mean weight curves across eight trials.

Previous researchers have grouped the pigs from these trials by virus load trajectories. Trial 1 pigs were separated into four “extreme” groups of High or Low virus load crossed with High or Low weight gain by separating on the uncorrelated linear combinations (principal components) of the variables AUC21 and WG42 (Arceo et al. [1]). Trial 1-3 pigs were grouped visually as rebound pigs and non-rebound pigs (Boddicker et al. [3]). Islam et al. [19] visually identified three types of virus load curves from all eight trials and identified parametric curves appropriate for each type, which allowed statistically driven classification of pigs into two curve types, rebound and non-rebound. Our analysis supports at least three distinct groups of pigs, but “rebound pigs” appear in all three groups (Figure (3.16)), likely because weight trajectories seem to be more important in determining clusters. The advantage of our approach for identifying clusters is that it does not rely on arbitrary boundaries to define groups (as in Arceo et al. [1]) or preliminary visual detection of groups (as in

Boddicker et al. [3] and Islam et al. [19]).

There are a number of potential further directions for this research. In the short term, since virus load was moderately more associated with WUR than weight gain (Boddicker et al. [3]), it would also be of interest to test the virus load time series alone for effect of WUR. Previous trials found little evidence of a difference between WUR genotypes 1 and 2, so we might also reduce this effect to a binary trait and retest. In the long term, extending the model to handle additional random effects and, particularly, fixed effects from the many other loci measured in such a study would make this model available as a tool for Genome-wide association studies (GWAS).

The advantage of this model is that it allows simultaneous analysis of multivariate responses, which typically have been reduced to a single univariate measure before use in GWAS (see Visscher et al. [39]). There is always trouble choosing a univariate response or defending the one chosen against questions of hidden optimization (how many measures did the researcher test before settling on the one published in the study?). In addition, extension of the model to handle covariation between time series could reveal relevant biological associations.

In this work, we proposed a multi-dimensional functional model-based clustering procedure for the analysis of time-course and longitudinal data. We developed a criterion based on an approximation of the marginal log-likelihood for model selection. The MLL criterion proved to be very effective. The simulation study and numerical experiments on real datasets (from gene expression field, bioinformatics or biology) showed that the model performs very well and challenges other models committed to the same analysis. One of the important aspects of the model is its ability to incorporate mixed effects in the clustering analysis. The model has been implemented in JAVA code and it is available from the second author's web-site. For the comparison study, we were hindered by the non-availability of easy to implement codes from competing algorithms.

Bibliographie

- [1] Arceo, M. E., C. W. Ernst, J. K. Lunney, I. Choi, N. E. Raney, T. Huang, C. K. Tuggle, R. R. Rowland, and J. P. Steibel (2012). Characterizing differential individual response to porcine reproductive and respiratory syndrome virus infection through statistical and functional analysis of gene expression. *Front Genet* 3, 321.
- [2] Banfield, J. D. and A. E. Raftery (1993). Model-based gaussian and non-gaussian clustering. *Biometrics* 49, 803–821.
- [3] Boddicker, N., E. H. Waide, R. R. Rowland, J. K. Lunney, D. J. Garrick, J. M. Reecy, and J. C. M. Dekkers (2012). Evidence for a major qtl associated with host response to porcine reproductive and respiratory syndrome virus challenge. *J Anim Sci* 90, 1733–1746.
- [4] Boddicker, N. J., A. Bjorkquist, R. R. Rowland, J. K. Lunney, J. M. Reecy, and J. C. M. Dekkers (2014). Genome-wide association and genomic prediction for host response to porcine reproductive and respiratory syndrome virus infection. *Genet Sel Evol* 46, 18.
- [5] Boddicker, N. J., D. J. Garrick, R. R. Rowland, J. K. Lunney, J. M. Reecy, and J. C. M. Dekkers (2014). Validation and further characterization of a major quantitative trait locus associated with host response to experimental infection with porcine reproductive and respiratory syndrome virus. *Anim Genet.* 45, 48–58.
- [6] Brumback, B. and J. Rice (1998). Smoothing spline models for the analysis of nested and crossed samples of curves. *Journal of the American Statistical Association* 93, 961–976.
- [7] Cho, R. J., M. J. Campbell, E. A. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T. G. Wolfsberg, A. E. Gabrielian, D. Landsman, D. J. Lockhart, and R. W. Davis (1998). A genome-wide transcriptional analysis of the mitotic cell cycle. *Molecular Cell* 2, 65–73.
- [8] Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20, 37–46.
- [9] Collins, J. E., D. A. Benfield, W. T. Christianson, L. Harris, J. C. Hennings, D. P. Shaw, S. M. Goyal, S. McCullough, R. B. Morrison, H. S. Joo, D. Goreyca, and D. Chladek (1991). Isolation of swine infertility and respiratory syndrome virus (isolate atcc vr-2332) in north america and experimental reproduction of the disease in gnotobiotic pigs. *Journal of Veterinary Diagnostic Investigation* 4, 117–126.

- [10] Crowder, M. J. and D. J. Hand (1990). *Analysis of Repeated Measures*. London : Chapman and Hall.
- [11] Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the em algorithm (with discussion). *Journal of the Royal Statistical Society 39*, 1–38.
- [12] Diggle, P. J., P. J. Heagerty, K.-Y. Liang, and S. L. Zeger (2002). *Analysis of Longitudinal Data, 2nd Ed.* Oxford : Oxford University Press.
- [13] Fraley, C. and A. E. Raftery (1999). Mclust : Software for model-based cluster analysis. *Journal of Classification 16(2)*, 297–306.
- [14] Fraley, C. and A. E. Raftery (2003). Enhanced software for model-based clustering, density estimation, and discriminant analysis : Mclust. *Journal of Classification 20*, 263–286.
- [15] Fraley, C. and A. E. Raftery (2006). Mclust version 3 for r : Normal mixture modeling and model-based clustering. Technical Report 504. Department of Statistics, University of Washington.
- [16] Golub, G. H. and C. F. Van Loan (1996). *Parameter estimation*. Baltimore, MD : Johns Hopkins.
- [17] Holtkamp, D. J., J. B. Kliebenstein, J. J. Zimmerman, E. Neumann, H. Rotto, T. K. Yoder, C. Wang, P. Yeske, C. L. Mowrer, and C. Haley (2012). Economic impact of porcine reproductive and respiratory syndrome virus on u.s. pork producers. *Animal Industry Report AS 658*.
- [18] Hubert, L. and P. Arabie (1985). Comparing partitions. *Journal of Classification 2*, 193–218.
- [19] Islam, Z. U., S. C. Bishop, N. J. Savill, R. R. R. Rowland, J. K. Lunney, B. Tribble, and A. B. Doeschl-Wilson (2013). Quantitative analysis of porcine reproductive and respiratory syndrome (prrs) viremia profiles from experimental infection : a statistical modelling approach. *PLoS One 8(12) :e83567*.
- [20] Jacques, J. and C. Preda (2013a). Funclust : A curves clustering method using functional random variable density approximation. *Neurocomputing 112*, 164–171.
- [21] Jacques, J. and C. Preda (2013b). Model-based clustering for multivariate functional data. *Computational Statistics and Data Analysis*. In press.
- [22] James, G., T. G. Hastie, and C. A. Sugar (2001). Principal component models for sparse functional data. *Biometrika 87*, 587–602.
- [23] James, G. and C. A. Sugar (2003). Clustering for sparsely sampled functional data. *Journal of the American Statistical Association 98*, 397–408.
- [24] Kubista, M., J. M. Andrade, M. Bengtsson, A. Frootan, J. Jonák, K. Lind, R. Sindelka, R. Sjöback, B. Sjögreen, L. Strömbom, A. Ståhlberg, and N. Zoric (2006). The real-time

- polymerase chain reaction. *Mol Aspects Med.* 27, 95–125.
- [25] Laird, N. M. and J. H. Ware (1982). Random-effects models for longitudinal data. *Biometrics* 38, 963–974.
- [26] Landis, J. and G. Koch (1977). The measurement of observer agreement for categorical data. *Biometrics* 33, 159–174.
- [27] Lin, D. and Z. Ying (2001). Semiparametric and nonparametric regression analysis of longitudinal data. *Journal of the American Statistical Association* 96, 103–126.
- [28] Loula, T. (1991). Mystery pig-disease. *Agri-Practice* 12, 23–34.
- [29] Lunney, J. K., J. P. Steibel, J. M. Reecy, E. Fritz, M. F. Rothschild, M. Kerrigan, B. Tribble, and R. R. Rowland (2011). Probing genetic control of swine responses to prrsv infection : current progress of the prrs host genetics consortium. *BMC Proc.* 2011 Jun 3;5 Suppl 4 :S30.
- [30] McNicholas, P. D. and T. B. Murphy (2010). Model-based clustering of longitudinal data. *Canadian Journal of Statistics* 38, 153–168.
- [31] Murua, A., L. Stanberry, and W. Stuetzle (2008). On potts model clustering, kernel k-means and density estimation. *Journal of Computational and Graphical Statistics* 17, 629–658.
- [32] Olszewski, R. (2001). Generalized feature extraction for structural pattern recognition in time-series data. Ph.D. thesis, Carnegie Mellon University.
- [33] Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association* 66, 846–850.
- [34] Reilly, C., C. Wang, and M. Rutherford (2005). A rapid method for the comparison of cluster analyses. *Statistica Sinica* 15, 19–33.
- [35] Rowland, R., J. Lunney, and J. Dekkers (2012). Control of porcine reproductive and respiratory syndrome (prrs) through genetic improvements in disease resistance and tolerance. *Front. in Gen.* 3 :260.
- [36] Tian, K., X. Yu, T. Zhao, Y. Feng, Z. Cao, and C. e. a. Wang (2007). Emergence of fatal prrsv variants : unparalleled outbreaks of atypical prrs in china and molecular dissection of the unique hallmark. *PLoS ONE.* 2 :e526, 23–34.
- [37] Tibshirani, R., G. Walther, and T. Hastie (2001). Estimating the number of clusters in a data set via the gap statistics. *J. R. Statist. Soc. B* 63, 411–423.
- [38] Tuddenham, R. and M. Snyder (1954). Physical growth of california boys and girls from birth to eighteen years. *Universities of California Public Child Development* 1, 188–364.
- [39] Visscher, P. M., M. A. Brown, M. I. McCarthy, and J. Yang (2012). Five years of gwas discovery. *The American Journal of Human Genetics* 90(1), 7–24.
- [40] Warrens, M. J. (2008). On the equivalence of cohen’s kappa and the hubert-arabic adjusted rand index. *Journal of Classification* 25, 177–183.

- [41] Wensvoort, G., C. Terpstra, and J. M. A. e. a. Pol (1991). Mystery swine disease in the netherlands : The isolation of lelystad virus. *The Veterinary quarterly* 13, 121–130.
- [42] Yeung, K. Y., C. Fraley, A. Murua, A. E. Raftery, and W. L. Ruzzo (2001). Model-based clustering and data transformations for gene expression data. *Bioinformatics* 17, 977–987.
- [43] Youness, G. and G. Saporta (2004). Some measures of agreement between close partitions. *Student* 5, 1–12.
- [44] Zeger, S. L. and P. J. Diggle (1994). Semiparametric models for longitudinal data with application to cd4 cell numbers in hiv seroconverters. *Biometrics* 50, 689–699.

Chapitre 4

FUNCTIONAL MODEL-BASED CLUSTERING WITH LASSO-TYPE PENALIZATION FOR LONGITUDINAL DATA

ABSTRACT

We propose a new time-course or longitudinal data analysis framework that aims at combining functional model-based clustering and the Lasso penalization to identify groups of individuals with similar patterns. An EM algorithm-based approach is used on a functional modeling where the individual curves are approximated into a space spanned by a finite basis of B-splines which dimension as well as the required number of functional principal components and the number of clusters are determined by penalizing a mixture of Student's t-distributions with unknown degrees of freedom. The first penalty term shrinks the coefficients of the cluster means on each functional principal component and helps to estimate the optimal number of principal components. The second penalty term shrinks the pairwise distances of cluster means and leads to an optimal estimation of the unknown number of clusters. The penalization or tuning parameters are set by Latin Hypercube Sampling (LHS) and their optimal values are found either by cross-validation or through the Bayesian Lasso functional clustering model developed herein. We use simulation study with few numerical examples and we apply the new methodology to a chronic obstructive pulmonary disease real dataset in a study of the microarray transcriptome of rats exposed to cigarette smoke.

Key words : Longitudinal data, model-based clustering, sparse longitudinal data, functional data analysis, mixture student, gene expression, Lasso penalization.

4.1. INTRODUCTION

Longitudinal data are usually analyzed using the very flexible class of linear mixed models (Laird and Ware [21]; Verbeke and Molenberghs [31]) which explicitly decompose the variation in the data into between and within-subject variability. However, the work of [34] has shown that another flexible methodology, Functional data analysis (FDA), can be a very useful complementary tool. Functional data analysis, which was primarily designed for the analysis of random trajectories and infinite-dimensional data, is rapidly evolving. Many interesting procedures incorporating this approach have recently emerged in statistics and bioinformatics in order to analyze time-course gene expression data. Clustering and classification techniques are two of the major applications of the functional approach with this type of genomic data (Ullah and Finch [30]). By definition, functional data clustering is used to search for natural groupings of data with similar characteristics. Recently, [19] reviewed the main literature on functional data clustering. They noted that most approaches fall within three broad categories : (a) the two-stage method that consists of applying dimension reduction techniques to the data before performing clustering; (b) the machine learning approach that uses nonparametric techniques on specifically defined distances or dissimilarities between curves; and (c) the model-based clustering approach which assumes a probabilistic mixture distribution on either the principal components (the FPCA *scores*) or the *expansion coefficients* associated with a functional data expansion into a finite dimensional basis of functions. Our present work falls into this latter category.

James and Sugar [20] seems to have been the first authors to introduce a functional model-based clustering method. They incorporated in the functional model a Gaussian-mixture distribution for the expansion coefficients associated with a finite spline basis. For rougher curves, Giacomini et al. [11] proposed a Gaussian-mixture model on a wavelet decomposition of the curves. Recently, Adjogou et al. [1] introduced a Bayesian model based on splines in which the clusters are modeled by a mixture of Student's t-distributions. An advantage of this latter model lies in the introduction of a principled way based on Bayes factors to choose the number of clusters, a problem that is the focus of the present paper. A different approach has been proposed by [26]. These authors assume that the curves arise from a mixture of regressions on a polynomial basis, with possible changes in regime at each instant of time.

In this work, we are particularly interested in shedding light into the mechanisms underlying critical exposure to tobacco smoke. Exposition to tobacco smoke at long-term chronic levels as well as at acute high levels represent known risks to human health. In order to understand the initial molecular events of chronic obstructive pulmonary disease (COPD) that leads to smoking related symptoms, [28] studied the microarray transcriptome of rats exposed to

cigarette smoke. During a period of 34 weeks, male Spague-Dawley rats were examined in a time-course study. This consisted of triplicate measures at 12 precise time points. The interval periods were chosen so that time-points 2 to 5 may be considered as early stage exposure with acute symptoms, while time-points 8 to 13 may be considered as prolonged exposure with chronic symptoms. The initial study analyzed the data with t-statistics associated with gene expression differences between exposed and control rats. These revealed a strong presence of upregulation of metabolic processes accompanied by stress response and genes involved in inflammation. During the later phase of smoke exposure the expression of genes related with immunity, and defense progressively increased.

In this paper, we extend the analysis of these data using the functional approach married with model based clustering. The idea is to find groups or classes of genes/proteins that are either upregulated or downregulated at the different stages of the exposure. The functional approach is needed since the data consist of time-course expressions; the model-based approach is used to cluster these time-courses. This is roughly done by reducing the dimension of the functional data to a set of latent variables which are in turn used for clustering. The discovering of the latent variables and the clustering are done simultaneously in the proposed model using a Lasso-type penalization approach embedded into a Bayesian framework.

Our method is also useful for the analysis and clustering of general complete or sparse time-course or longitudinal data. It is inspired by recent works in variable selection for clustering of high-dimensional data (see for example Bouveyron and Brunet [2] for a nice review). Nowadays, penalizing criteria for clustering are the preferred methods of variable selection for high-dimensional data. Since the pioneering work of Tibshirani [29], where the Lasso was introduced, several works on model based clustering have introduced L_1 or L_∞ penalty terms in the log-likelihood function (Pan and Shen [24], Wang and Zhou [32]). This is done to yield model sparsity in the form of variable selection (which may also be seen as a form of dimension reduction). Traditionally, the Lasso penalizes the absolute values (L_1 -norm) of coefficients that are key to the model. Our procedure uses a double Lasso-penalty in the clustering criterion in order to yield optimal choices for the reduced dimension of the data (similar to variable selection in the regression context) and the number of clusters. The strength of the regularization is then determined by two penalization parameters whose optimal values are unknown. Usually, these parameters are *tuning* parameters, that is, the model is estimated for some particular chosen values of these parameters. Their optimal values are usually determined by cross-validation techniques. In this work, we introduced a

Bayesian Lasso penalization model. Cross-validation is not necessarily needed. The regularization parameters are incorporated in the model through a Lasso-driven prior distribution.

The paper is organized as follows. Section 2 introduces the model-based clustering with Lasso penalty (model, parameter estimation, implementation). Section 3 discusses the model selection method. Section 4 describes a simulation study and comparison with existing methods. In Section 5 we apply our methodology to the analysis of the tobacco exposure data.

4.2. FUNCTIONAL MODEL-BASED CLUSTERING WITH LASSO PENALIZATION

We introduce in this section our penalized functional model-based clustering method and the inference procedure including the expectation-maximization (EM) algorithm (Dempster et al. [4]). The fundamentals of the model are similar to the one developed in Adjogou et al. [1]. It combines functional principal components analysis and model-based clustering in a mixed effects model setting and a Bayesian framework for the estimation of the parameters.

4.2.1. Fundamentals of the model

We recall below the main characteristics of the basis model. If $Y_i(\cdot)$ denotes the source function that originally generates the n_i observed measurements $\mathbf{Y}_i = (y_{i1}, \dots, y_{in_i})$ at time points $\mathbf{t}_i = (t_{i1}, \dots, t_{in_i})$ for the individual i in a longitudinal study, its evaluation at a specific time t is assumed to be decomposed in the form :

$$\begin{cases} Y_i(t) = \hat{\mu}(t) + \hat{\mathbf{f}}(t)^\top \boldsymbol{\alpha}_i + \epsilon_i(t) & (i = 1, \dots, N) \\ \boldsymbol{\alpha}_i = \boldsymbol{\mu}_{\mathbf{z}_i} + \boldsymbol{\gamma}_i^{\mathbf{z}_i} \end{cases} \quad (4.2.1)$$

where $\hat{\mu}(t)$ is an overall mean function; the functions $\hat{f}_j(t)$ are the k functional principal components (FPC) with $\hat{\mathbf{f}}(t)^\top = (\hat{f}_1(t), \dots, \hat{f}_k(t))$ and the $\epsilon_i(t)$ are error terms. The k -dimensional vectors $\boldsymbol{\alpha}_i = (\alpha_{i1}, \dots, \alpha_{ik})$ are the component scores representing the coefficients of $Y_i(t)$ on the FPCs. Furthermore, a mixed effects framework is imposed on the clustering model through the component scores. Indeed, the component scores of each individual i are expressed as the sum of the individual i 's cluster mean $\boldsymbol{\mu}_{\mathbf{z}_i}$ and his own effect $\boldsymbol{\gamma}_i^{\mathbf{z}_i}$ (or the deviation from its cluster effect). The variables \mathbf{z}_i are the cluster membership indicators. Thus, the combination of the two expressions of Equation (4.2.1) yields a 3-term decomposition for the curve $Y_i(t)$ in addition to the error term $\epsilon_i(t)$: the overall mean $[\hat{\mu}(t)]$, the cluster effect $[\hat{\mathbf{f}}(t)^\top \boldsymbol{\mu}_{\mathbf{z}_i}]$ and the individual-specific effect $[\hat{\mathbf{f}}(t)^\top \boldsymbol{\gamma}_i^{\mathbf{z}_i}]$. All those terms are rewritten in a matrix form using the specification of the model in a finite-dimensional

basis $\mathbf{b}(t)^\top = (b_1(t), \dots, b_q(t))$ of B-splines to obtain the following expression :

$$\mathbf{Y}_i = \mathbf{B}_i \boldsymbol{\theta}_\mu + \mathbf{B}_i \boldsymbol{\Theta} \boldsymbol{\mu}_{\mathbf{z}_i} + \mathbf{B}_i \boldsymbol{\Theta} \boldsymbol{\gamma}_i^{\mathbf{z}_i} + \boldsymbol{\epsilon}_i. \quad (4.2.2)$$

In Equation (4.2.2), the n_i -dimensional vector \mathbf{Y}_i contains the observed measurements at time points \mathbf{t}_i and $\mathbf{B}_i = [\mathbf{b}(t_{i1}), \dots, \mathbf{b}(t_{in_i})]^\top$ is the matrix of the spline basis evaluated at those time points. The q -dimensional vector $\boldsymbol{\theta}_\mu$ and the matrix $\boldsymbol{\Theta}$ represent, respectively, the coefficients in the basis of the overall mean function $\hat{\mu}(t) = \mathbf{b}(t)^\top \boldsymbol{\theta}_\mu$ and the principal components functions $\hat{\mathbf{f}}(t)^\top = \mathbf{b}(t)^\top \boldsymbol{\Theta}$. The measurement errors $\boldsymbol{\epsilon}_i$ are assumed to follow a multivariate Student's t distribution with unknown degrees of freedom ν_0 . The functional model-based clustering in Equation (4.2.2) is embedded into a Bayesian framework and the following assumptions are made to complete the setup.

$$\left\{ \begin{array}{l} \mathbf{z}_i \sim \text{Multinomial}(1; \pi_1, \dots, \pi_G) \quad \text{with} \quad (\pi_1, \dots, \pi_G) \sim \text{Dirichlet}(a_1, \dots, a_G) \\ \boldsymbol{\mu}_g \sim \mathcal{N}_k(\mathbf{0}, \boldsymbol{\Gamma}_\mu) \quad \text{and} \quad \boldsymbol{\Gamma}_\mu \sim \text{InvWishart}(m, (m-k-1)\mathbf{I}_k) \\ \boldsymbol{\gamma}_i^g \sim \mathcal{N}_k(\mathbf{0}, \boldsymbol{\Gamma}_g) \quad \text{and} \quad \boldsymbol{\Gamma}_g \sim \text{InvWishart}(m, (m-k-1)\mathbf{D}) \\ \mathbf{D} = \text{diag}(d_{11}, d_{22}, \dots, d_{kk}) \quad \text{with} \quad d_{jj} \sim \text{Inv}\chi^2(m) \quad \text{and i.i.d.} \quad (j = 1, \dots, k) \\ [\boldsymbol{\epsilon}_i | \nu_i \sim \mathcal{N}_{n_i}(\mathbf{0}, \sigma^2 \nu_i \mathbf{I}_{n_i}) \quad \text{and} \quad \nu_i \sim \text{Inv}\chi^2(\nu_0)] \Rightarrow [\boldsymbol{\epsilon}_i \sim t_{\nu_0}(\mathbf{0}, \sigma^2 \mathbf{I}_{n_i})] \\ \text{with} \quad \sigma^2 \sim \text{InvGamma}(\alpha_\sigma, \beta_\sigma). \end{array} \right. \quad (4.2.3)$$

In general, the clustering of the individuals based on this model relies essentially on a space of much reduced dimension than the original longitudinal trajectories through the decomposition on the functional principal components. In the perspective of the EM algorithm used to estimate the parameters, the log-likelihood of the model as stated in Equations (4.2.2) and (4.2.3) is obtained by considering the « complete data » (\mathbf{Y}, \mathbf{W}) where $\mathbf{Y} = \{\mathbf{Y}_1, \dots, \mathbf{Y}_N\}$ denotes the observed data composed of the N longitudinal trajectories ($i = 1, \dots, N$) and $\mathbf{W} = \{\mathbf{z}_1, \dots, \mathbf{z}_N, \boldsymbol{\gamma}_1^{\mathbf{z}_1}, \dots, \boldsymbol{\gamma}_N^{\mathbf{z}_N}\} = \{\bar{\mathbf{z}}, \vec{\boldsymbol{\gamma}}^{\mathbf{z}}\}$ denotes the missing data composed of the cluster indicators and the individual-specific effects. Let $\boldsymbol{\Pi}$ denote the set of the model parameters to be estimated and, let $\mathfrak{L}(\mathbf{Y}, \mathbf{W}; \boldsymbol{\Pi}) = \log[p(\mathbf{Y}, \mathbf{W}; \boldsymbol{\Pi})]$ denote the log-likelihood derived from the distributions involved in the model. Note that $\boldsymbol{\Pi} = \{\vec{\nu}, \vec{\boldsymbol{\mu}}, \vec{\boldsymbol{\Gamma}}, \boldsymbol{\Lambda}\}$ where

$$\left\{ \begin{array}{l} \vec{\nu} = \{\nu_1, \dots, \nu_N\}; \quad \vec{\boldsymbol{\mu}} = \{\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_G\}; \quad \vec{\boldsymbol{\Gamma}} = \{\boldsymbol{\Gamma}_1, \dots, \boldsymbol{\Gamma}_G\}; \\ \boldsymbol{\Lambda} = \{\boldsymbol{\theta}_\mu, \boldsymbol{\Theta}, \mathbf{D}, \boldsymbol{\Gamma}_\mu, \pi_1, \pi_2, \dots, \pi_G, \nu_0, \sigma^2\}. \end{array} \right. \quad (4.2.4)$$

The expression of $\mathfrak{L}(\mathbf{Y}, \mathbf{W}; \boldsymbol{\Pi})$ is presented in Appendix A.2 as well as the details leading to its computation.

4.2.2. The penalized log-likelihood

One of the main goals of the model is to adequately determine the two characteristic model parameters : the number of clusters G and the dimension k of the functional principal components $\hat{\mathbf{f}}(t)$. Unlike in Adjogou et al. [1], the dimension q of the B-splines basis is not

considered as a parameter to be estimated. The number of basis functions is either set to a specific value with respect to the measurement time points of all individuals, or indirectly defined by supplying the break points or knots. The motivation for this decision comes from one of the conclusions of the simulation study in Adjogou et al. [1] which indicates that the value of the parameter q has very little influence on the clustering results especially on the Adjusted Rand Index (ARI) scores.

In this new framework, we choose to estimate k and G by penalizing the log-likelihood function. The two penalizations are Lasso-type ones. The main objective is to obtain a sparse solution with many estimates of cluster means basis coefficients automatically shrunk, thus realizing dimension reduction and with many inter-cluster distances shrunk, thus merging homogeneous clusters. The proposed penalized log-likelihood function is defined as

$$\begin{aligned}\mathfrak{L}^{pen}(\mathbf{Y}, \mathbf{W}; \mathbf{\Pi}) &= \mathfrak{L}(\mathbf{Y}, \mathbf{W}; \mathbf{\Pi}) - P_\lambda(\mathbf{\Pi}) \\ &= \log[p(\mathbf{Y}, \mathbf{W}; \mathbf{\Pi})] - P_\lambda(\vec{\boldsymbol{\mu}})\end{aligned}\tag{4.2.5}$$

where $\vec{\boldsymbol{\mu}} = \{\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_G\}$ and $P_\lambda(\cdot)$ denotes a Lasso-type penalty function with tuning parameter λ . In our model, a Lasso-type penalty function is applied to the cluster means and another one is applied to the distances between those cluster means. The penalization $P_\lambda(\vec{\boldsymbol{\mu}})$ then takes the form :

$$\begin{aligned}P_\lambda(\vec{\boldsymbol{\mu}}) &= \sum_{j=1}^k P_{\lambda_1} \left(\sum_{g=1}^G |\mu_g^j| \right) + \sum_{g=1}^G \sum_{h=1}^g P_{\lambda_2} \left(D^{ist}(\boldsymbol{\mu}_g, \boldsymbol{\mu}_h) \right) \\ &= \lambda_1 \sum_{j=1}^k \left(\sum_{g=1}^G |\mu_g^j| \right) + \lambda_2 \sum_{g=1}^G \sum_{h=1}^g D^{ist}(\boldsymbol{\mu}_g, \boldsymbol{\mu}_h)\end{aligned}\tag{4.2.6}$$

where the function $D^{ist}(\cdot, \cdot)$ can optionally be the L_1 norm distance or the L_2 norm distance. The first term of the penalty function which is associated with the hyperparameter λ_1 is used to shrink towards zero, for a given j , the estimates $(\sum_{g=1}^G |\mu_g^j|)$ which are close to zero. As a consequence, the model will reduce the dimension k of the cluster space by eliminating those principal components j that are irrelevant for the model. The second term of the penalty function which is associated with the hyperparameter λ_2 is used to shrink towards zero the distances of very similar estimated cluster means. As a consequence, any two clusters with very similar cluster means will be forced to merge. This will reduce the initial assumed number of clusters G . Only clusters with very different means are expected to remain.

4.2.3. Choosing the penalty parameters by cross-validation

Most of the literature concerning Lasso-type penalization suggest using cross-validation in order to estimate and fix the value of the penalty parameters λ_1 and λ_2 . Note that this

procedure basically amounts to estimating the model for a given pair of optimal values of (λ_1, λ_2) , ignoring the fact that the data (and the model) have been previously used to choose the pair (λ_1, λ_2) . Another issue with cross-validation is its computational cost. This may be large for large data and complex models such as the one considered in this paper. Note that in order to find an optimal pair (λ_1, λ_2) , a grid of values in the two-dimensional space of penalty parameters must be chosen. Therefore, a third issue with this procedure relates to how to choose the grid. If a simple uniform grid is to be chosen, then most of the time, the size of the grid would be too large. For example, for a 20×20 grid, one already needs to fit 400 models times the number of the cross-validation folds; if one performs a 5-fold cross-validation, the number of times one would need to fit the model would be 2000.

In a K -fold cross-validation for each pair (λ_1, λ_2) , the dataset is randomly split into K mutually exclusive subsets of approximately equal size, called the folds. The model is estimated and a measure of the goodness of fit such as the log-likelihood is computed K times. Each time, $(K - 1)$ subsets are put together and used as training set to estimate the model. The other remaining subset is used as validation set to compute the log-likelihood. The cross-validation log-likelihood is the overall mean from the K folds.

We suggest to choose a grid by using techniques from design of experiments. One particular useful technique for computer experiments is the use of Latin hypercube sampling (LHS). These are designs that try to fill in the search space much more efficiently than a uniform grid. For example, a uniform grid of size 10×10 may be too coarse to really find the optimal pair. But a LHS array of size 100 would cover the space of the penalty parameters in an efficient way. The LHS technique has been applied to many different computer models since 1975 (Steck et al. [27], Iman et al. [16, 17], Iman and Conover [14, 15], Iman and Helton [18], Wyss and Jorgensen [33]).

We suggest to use a hierarchical search with small LHS arrays. This sampling approach is computationally cheap and ensures that each of the input variables has all portions of its range represented, which is a very interesting requirement in order to evaluate as many values of the tuning parameters as possible. First, in order to find the right order of magnitude of the penalty parameters, a search over a Latin array of size 50 is made. Once the optimal values in this initial search is found, a second search is directed in a reduced space of values (that hopefully covers the region where the true optimal parameters are found) given again by a small-size LHS array. This search strategy works well in our experiments. For the implementation of LHS in our procedure, we use the package *DiceDesign* (Dupuy et al. [5]) developed in the software R.

The cross-validation criterion is entirely based on the penalized log-likelihood presented in Equation (4.2.5). The most commonly used cross-validation practices are based on either a 3-fold or a 10-fold. Depending on the size of the dataset to be analyzed, one may prefer the 3-fold for some gain in the execution time.

4.2.4. The Bayesian Lasso functional clustering model

Even though we managed to reduce the number of model fits considerably by using LHS procedure, the cost of the search is sometimes still too high for large datasets. To alleviate this cost and to make sure we have obtained the optimal penalty parameters, we propose a model where the penalty function is part of the likelihood. This allows us to consider the pair (λ_1, λ_2) as model parameters, just as the rest of the parameters. Since the form of the penalty function is essential, we simply propose to normalize the penalized likelihood, that is, to make it a density. Therefore, this solution requires finding the normalizing constant of the penalized likelihood function.

Indeed, the penalized likelihood function derived from the penalized log-likelihood in Equation (4.2.5) can be expressed as $\left[\mathbf{L}^{pen}(\mathbf{Y}, \mathbf{W}; \mathbf{\Pi}) = p(\mathbf{Y}, \mathbf{W}; \mathbf{\Pi}) \cdot e^{-P_\lambda(\vec{\mu})} \right]$ and we search an appropriate transformation to be applied to that expression in order to make it a density function. The complete expression of the density function $p(\mathbf{Y}, \mathbf{W}; \mathbf{\Pi})$ and details on its computation are presented in Appendix A.2. We simply recall here that $p(\mathbf{Y}, \mathbf{W}; \mathbf{\Pi})$ satisfies :

$$\begin{aligned}
 p(\mathbf{Y}, \mathbf{W}; \mathbf{\Pi}) &= p(\vec{\mathbf{Y}}, \vec{\mathbf{z}}, \vec{\gamma}^{\mathbf{z}}; \vec{\nu}, \vec{\mu}, \vec{\Gamma}, \Lambda) & (4.2.7) \\
 &= \prod_{i=1}^N p(\mathbf{Y}_i, \mathbf{z}_i, \gamma_i^{\mathbf{z}_i} | \nu_i, \vec{\mu}, \vec{\Gamma}, \Lambda) \cdot p(\nu_i | \vec{\mu}, \vec{\Gamma}, \Lambda) \cdot \prod_{g=1}^G p(\boldsymbol{\mu}_g) \cdot p(\Gamma_g) \cdot p(\Lambda) \\
 &= \bar{p}(\mathbf{Y}, \mathbf{W}; \mathbf{\Pi}) \cdot \prod_{g=1}^G p(\boldsymbol{\mu}_g).
 \end{aligned}$$

where $\left[\bar{p}(\mathbf{Y}, \mathbf{W}; \mathbf{\Pi}) = \prod_{i=1}^N p(\mathbf{Y}_i, \mathbf{z}_i, \gamma_i^{\mathbf{z}_i} | \nu_i, \vec{\mu}, \vec{\Gamma}, \Lambda) \cdot p(\nu_i | \vec{\mu}, \vec{\Gamma}, \Lambda) \cdot \prod_{g=1}^G p(\Gamma_g) \cdot p(\Lambda) \right]$. The only component in Equation (4.2.7) that involves explicitly the cluster means $\boldsymbol{\mu}_g$ as in the penalty function $P_\lambda(\vec{\mu})$ is the term $p(\boldsymbol{\mu}_g)$ or more precisely the distributions of the $\boldsymbol{\mu}_g$ which are $\mathcal{N}_k(\mathbf{0}, \Gamma_\mu)$. The other components (gathered in $\bar{p}(\mathbf{Y}, \mathbf{W}; \mathbf{\Pi})$) are known density functions that do not depend on $\vec{\mu}$. In order to normalize $\mathbf{L}^{pen}(\mathbf{Y}, \mathbf{W}; \mathbf{\Pi})$, we only need to normalize the induced prior on $\vec{\mu}$, given by $\left[\prod_{g=1}^G p(\boldsymbol{\mu}_g) \cdot e^{-P_\lambda(\vec{\mu})} \right]$. That is, we need to

compute the integral

$$C(\lambda_1, \lambda_2, \mathbf{\Gamma}_\mu) = \left[\int \left(\prod_{g=1}^G \frac{e^{-\frac{1}{2}\boldsymbol{\mu}_g^T \mathbf{\Gamma}_\mu^{-1} \boldsymbol{\mu}_g}}{(2\pi)^{k/2} |\mathbf{\Gamma}_\mu|^{1/2}} \right) e^{-P_\lambda(\bar{\boldsymbol{\mu}})} d(\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_G) \right]. \quad (4.2.8)$$

The normalized penalized log-likelihood is given by

$$\mathcal{L}_c^{pen}(\mathbf{Y}, \mathbf{W}; \mathbf{\Pi}) = \log[p(\mathbf{Y}, \mathbf{W}; \mathbf{\Pi})] - P_\lambda(\bar{\boldsymbol{\mu}}) - \log C(\lambda_1, \lambda_2, \mathbf{\Gamma}_\mu). \quad (4.2.9)$$

We refer to the model based on this normalized penalized log-likelihood as a *Bayesian Lasso functional clustering model* or *Bayesian Lasso FCM* for short. We use Monte Carlo numerical integration to estimate the integral in (4.2.8). We sample values $\{\bar{\boldsymbol{\mu}}_m\}_{m=1}^M$ according to a kG -Multivariate Normal distribution with mean zero and block-diagonal variance-covariance matrix with G blocks equal to $\mathbf{\Gamma}_\mu$. The sampling is done with respect to the prior distribution of $\bar{\boldsymbol{\mu}}$, which is exactly the term in parentheses in the integral above. Consequently, our estimator $\tilde{C}(\lambda_1, \lambda_2, \mathbf{\Gamma}_\mu)$ is given by

$$\tilde{C}(\lambda_1, \lambda_2, \mathbf{\Gamma}_\mu) = \left[\frac{1}{M} \sum_{m=1}^M e^{-\lambda_1 \sum_{j=1}^k (\sum_{g=1}^G |\mu_{g(m)}^j|) - \lambda_2 \sum_{g=1}^G \sum_{h=1}^g Dist(\mu_{g(m)}, \mu_{h(m)})} \right]. \quad (4.2.10)$$

4.2.5. EM algorithm : Expectation and Maximization steps

As mentioned in Section 4.2.1, the iterative EM algorithm is used to estimate the parameters of the model. The function $S(\mathbf{\Pi}|\mathbf{\Pi}^{(t)})$ to be maximized by the EM algorithm for maximum likelihood methods is defined as :

$$S(\mathbf{\Pi}|\mathbf{\Pi}^{(t)}) = \begin{cases} Q(\mathbf{\Pi}|\mathbf{\Pi}^{(t)}) - P_\lambda(\bar{\boldsymbol{\mu}}) - \log C(\lambda_1, \lambda_2, \mathbf{\Gamma}_\mu) & \text{for the Bayesian Lasso FCM,} \\ Q(\mathbf{\Pi}|\mathbf{\Pi}^{(t)}) - P_\lambda(\bar{\boldsymbol{\mu}}) & \text{for cross-validation of the penalized clustering.} \end{cases} \quad (4.2.11)$$

where $[Q(\mathbf{\Pi}|\mathbf{\Pi}^{(t)}) = E_{\mathbf{W}|\mathbf{Y}; \mathbf{\Pi}^{(t)}}[\log p(\mathbf{Y}, \mathbf{W}; \mathbf{\Pi})]]$. The observed data \mathbf{Y} and the missing data \mathbf{W} are as previously defined in Section 4.2.1. The computation of $Q(\mathbf{\Pi}|\mathbf{\Pi}^{(t)})$ at the expectation step is identical to that of Adjogou et al. [1]. All the analytical developments leading to its calculus are the same as presented in Appendix A.2.

The maximization step is similar to the one in Adjogou et al. [1] apart from a few differences. Indeed, the maximization of $S(\mathbf{\Pi}|\mathbf{\Pi}^{(t)})$ is identical to the one of $Q(\mathbf{\Pi}|\mathbf{\Pi}^{(t)})$ for all the parameters except for the cluster means $\boldsymbol{\mu}_g$ because of their additional appearance in the penalty function. The results of the M-step for the other parameters do not depend on the model chosen, since the normalizing constant of the Bayesian Lasso FCM depends only on the current value of the parameters. The M-step equations are presented below. Note

that Appendix A.2 includes details on the expressions involved in the following equations such as P_{ig} , $\hat{\gamma}_{ig}$ and \hat{V}_i^g .

$$\left\{ \begin{array}{l}
\mathbf{\Gamma}_g^{(t+1)} = \left[\left(\sum_{i=1}^N P_{ig} \right) + (m+k+1) \right]^{-1} \left[\left(\sum_{i=1}^N P_{ig} (\hat{\gamma}_{ig} \hat{\gamma}_{ig}^\top + \hat{V}_i^g) \right) + (m-k-1) \mathbf{D}^{(t)} \right] \\
\pi_g^{(t+1)} = \frac{\left(\sum_{i=1}^N P_{ig} \right) + (a_g - 1)}{N + \left(\sum_{g=1}^G a_g \right) - G} \quad (0 \leq \pi_g^{(t+1)} \leq 1) \\
\boldsymbol{\theta}_\mu^{(t+1)} = \left[\sum_{i=1}^N \left(\mathbf{B}_i^\top \mathbf{B}_i \right) \nu_i^{-1(t)} \right]^{-1} \left[\sum_{i=1}^N \sum_{g=1}^G \frac{P_{ig}}{\nu_i^{(t)}} \mathbf{B}_i^\top \left(\mathbf{Y}_i - \mathbf{B}_i \boldsymbol{\theta}_{(t)} \boldsymbol{\mu}_g^{(t)} - \mathbf{B}_i \boldsymbol{\theta}_{(t)} \hat{\gamma}_i^g \right) \right] \\
\boldsymbol{\Theta}_j^{(t+1)} = \left\{ \sum_{i=1}^N \sum_{g=1}^G \frac{P_{ig}}{\nu_i^{(t)}} \left[(\boldsymbol{\alpha}_{ig})_j^2 + (\hat{V}_i^g)_{jj} \right] \left(\mathbf{B}_i^\top \mathbf{B}_i \right) \right\}^{-1} \{ \Omega_1 - \Omega_2 \} \quad ; \quad \text{for } j = 1, \dots, k. \\
\Omega_1 = \sum_{i=1}^N \sum_{g=1}^G \frac{P_{ig}}{\nu_i^{(t)}} \left[(\boldsymbol{\alpha}_{ig})_j \mathbf{B}_i^\top \left(\mathbf{Y}_i - \mathbf{B}_i \boldsymbol{\theta}_{(t)} \right) \right]; \quad (\boldsymbol{\alpha}_{ig})_j = (\boldsymbol{\mu}_g + \hat{\gamma}_i^g)_j \\
\Omega_2 = \sum_{i=1}^N \sum_{g=1}^G \frac{P_{ig}}{\nu_i^{(t)}} \left[\sum_{h \neq j}^k \left((\boldsymbol{\alpha}_{ig})_j (\boldsymbol{\alpha}_{ig})_h + (\hat{V}_i^g)_{hj} \right) \left(\mathbf{B}_i^\top \mathbf{B}_i \right) \boldsymbol{\Theta}_h \right] \\
\sigma_{(t+1)}^2 = \frac{\frac{1}{2} \left\{ \sum_{i=1}^N \sum_{g=1}^G \frac{P_{ig}}{\nu_i^{(t)}} \left[\left\| \mathbf{Y}_i - \mathbf{B}_i \boldsymbol{\theta}_{(t)} \boldsymbol{\mu}_g^{(t)} - \mathbf{B}_i \boldsymbol{\theta}_{(t)} \hat{\gamma}_i^g \right\|^2 + \text{trace} \left(\mathbf{B}_i \boldsymbol{\theta}_{(t)} \hat{V}_{ig} \boldsymbol{\theta}_{(t)}^\top \mathbf{B}_i^\top \right) \right] \right\} + \beta_\sigma}{\frac{1}{2} \left[\sum_{i=1}^N n_i \right] + (\alpha_\sigma + 1)} \\
\nu_0^{(t+1)} = \frac{b_{\nu_0} + 1 + \sqrt{2b_{\nu_0} + 1}}{b_{\nu_0}} \quad \text{with } b_{\nu_0} = \exp \left(\frac{1}{N} \sum_{i=1}^N (\log \nu_i^{(t)} + 1/\nu_i^{(t)}) - 1 \right) \\
\nu_i^{(t+1)} = \frac{\left\{ \nu_0^{(t)} + \sum_{g=1}^G \frac{P_{ig}}{\sigma_{(t)}^2} \left[\left\| \mathbf{Y}_i - \mathbf{B}_i \boldsymbol{\theta}_{(t)} \boldsymbol{\mu}_g^{(t)} - \mathbf{B}_i \boldsymbol{\theta}_{(t)} \hat{\gamma}_i^g \right\|^2 + \text{trace} \left(\mathbf{B}_i \boldsymbol{\theta}_{(t)} \hat{V}_{ig} \boldsymbol{\theta}_{(t)}^\top \mathbf{B}_i^\top \right) \right] \right\}}{n_i + 2 + \nu_0^{(t)}} \\
d_{jj}^{(t+1)} = \frac{1}{2} \left[\frac{(mG - m - 2) + \sqrt{(mG - m - 2)^2 + 4m \times (m - k - 1) \times \sum_{g=1}^G \left\{ \mathbf{\Gamma}_g^{-1(t)} \right\}_{jj}}}{(m - k - 1) \times \sum_{g=1}^G \left\{ \mathbf{\Gamma}_g^{-1(t)} \right\}_{jj}} \right] \quad (j = 1, \dots, k) \\
\left\{ \mathbf{\Gamma}_\mu^{(t+1)} \right\}_{jj} = \frac{1}{m + k + 1 + G} \left[\left\{ \sum_{g=1}^G \boldsymbol{\mu}_{g(t)} \boldsymbol{\mu}_{g(t)}^\top \right\}_{jj} + (m - k - 1) \right].
\end{array} \right.$$

The results of the M-step for the cluster means $\boldsymbol{\mu}_g$ are presented below for each of the two options of the $D^{ist}(\cdot, \cdot)$ function in the second penalty term from Equation (4.2.6).

$$\left\{ \begin{array}{l} A_1 = \left[\sum_{i=1}^N \frac{P_{ig}}{\nu_i^{(t)} \sigma_{(t)}^2} \left((\mathbf{Y}_i - \mathbf{B}_i \boldsymbol{\theta}_\mu^{(t)} - \mathbf{B}_i \boldsymbol{\Theta}_{(t)} \hat{\gamma}_i^g)^\top \mathbf{B}_i \boldsymbol{\Theta}_{(t)} \right)_j \right] \\ \quad + \left[\sum_{i=1}^N \frac{-P_{ig}}{2\nu_i^{(t)} \sigma_{(t)}^2} \sum_{r \neq j}^k \mu_g^{r(t)} \left(\boldsymbol{\Theta}_{(t)}^\top \mathbf{B}_i^\top \mathbf{B}_i \boldsymbol{\Theta}_{(t)} \right)_{jr} \right] \\ B_1 = \left[\sum_{i=1}^N \left\{ \frac{P_{ig}}{\nu_i^{(t)} \sigma_{(t)}^2} \left(\boldsymbol{\Theta}_{(t)}^\top \mathbf{B}_i^\top \mathbf{B}_i \boldsymbol{\Theta}_{(t)} \right)_{jj} \right\} + \left\{ \boldsymbol{\Gamma}_\mu^{-1(t)} \right\}_{jj} \right] \\ \mu_g^{j(t+1)} = \left[\frac{A_1 - \lambda_1 \left\{ \text{sign}(\mu_g^{j(t)}) \right\} + \left\{ \lambda_2 \sum_{h \neq g}^G \frac{\mu_h^{j(t)}}{D^{ist}(\mu_g, \mu_h)} \right\}}{B_1 + \left\{ \lambda_2 \sum_{h \neq g}^G \frac{1}{D^{ist}(\mu_g, \mu_h)} \right\}} \right] \quad \text{for the } L_2 \text{ norm distance} \\ \mu_g^{j(t+1)} = \left[\frac{A_1 - \lambda_1 \left\{ \text{sign}(\mu_g^{j(t)}) \right\} - \lambda_2 \left\{ \sum_{h=1}^G \text{sign}(\mu_g^{j(t)} - \mu_h^{j(t)}) \right\}}{B_1} \right] \quad \text{for the } L_1 \text{ norm distance.} \end{array} \right.$$

Note that $\text{sign}(x)$ equals -1 if $x < 0$, equals 1 if $x > 0$ and equals 0 if $x = 0$. It is worth noting that all the constraints imposed on the parameters in Adjogou et al. [1] (such as \mathbf{D} being diagonal) still prevail here in the estimation of the parameters.

4.3. MODEL SELECTION

In this section, we describe the steps toward the selection of the optimal model. While Adjogou et al. [1] performs model selection by computing criteria such as AIC, BIC or MLL (Marginal log-likelihood) developed therein, we consider in this new framework two Lasso-type penalizations and select the optimal model through cross-validation or preferably, the Bayesian Lasso FCM which is equivalent to a 1-fold cross-validation. For either a K -fold cross-validation criterion (where, usually, $K = 3$ or 10) or the Bayesian Lasso FCM, the first step of model selection consists of identifying the optimal values of the two tuning parameters $(\lambda_1^{opt}, \lambda_2^{opt})$ from a grid of proposed values defined through Latin hypercube sampling (see Section 4.2.3). The second step consists of identifying the optimal number of functional principal components k^{opt} and the optimal number of clusters G^{opt} . The following are the steps for model selection :

- (1) **Postulate values for k and G** : Let k^{pos} and G^{pos} be the postulated values of the number of functional components and the number of clusters, respectively. As mentioned in Section 4.2.2, the B-splines basis dimension q is not estimated but rather set manually. However, the value of q is not chosen at random. Some rules govern its

choice. For a B-splines basis of order m_B (with polynomials of degree $d_B = m_B - 1$), the number of basis functions can be expressed as $q = m_B + i_B$ where i_B is the number of interior knots. And q must satisfy $q \geq m_B$. For example, q must be at least 4 in the case of cubic splines. Furthermore, the value of q must be large enough to ensure a significant number of interior knots that will be equally spaced within the range of the measurement time points $[\min t_{ij}, \max t_{ij}]$ ($1 \leq i \leq N; 1 \leq j \leq n_i$) in order to span the individual curves. Also, as the columns of the (q, k) -dimensional matrix Θ are orthonormal due to the orthogonality constraint $\Theta^\top \Theta = I_k$, the value of q must also satisfy $q \geq k^{pos}$. The values (q, k^{pos}, G^{pos}) are used to initialize each single run of the EM-algorithm.

- (2) **Perform the EM algorithm with (λ_1, λ_2)** : For each couple (λ_1, λ_2) and for each postulated (initial) values of the parameters (k^{pos}, G^{pos}) , either the Bayesian Lasso FCM is fitted using the whole dataset, or a K -fold cross-validation is performed on the training/validation sets as described in Section 4.2.3.

The parameters Θ , θ_μ and σ^2 are initialized by assuming a model with a single cluster. The cluster parameters such as μ_g , Γ_g and π_g as well as the cluster membership indicators \mathbf{z}_i may be initialized by applying any clustering procedure to the scores α_i yielded by the single-cluster model. In our experiments, we used the Gaussian model-based clustering procedure implemented in the *mclust package* (Fraley and Raftery [6, 7, 8]).

- (3) **Find $(\lambda_1^{opt}, \lambda_2^{opt})$** : For the Bayesian Lasso FCM, step 2 ends up with the computed value of the penalized log-likelihood, the estimated number of valid clusters (that is, the number of non-empty clusters), denoted G^{valid} , and the estimated parameters. Recall that step 2 is performed for each couple (λ_1, λ_2) in the LHS grid. The optimal values $(\lambda_1^{opt}, \lambda_2^{opt})$ are defined as the couple (λ_1, λ_2) that maximizes the Bayesian Lasso FCM log-likelihood. The next step in this case is to determine k^{opt} and G^{opt} based on the estimators yielded by the corresponding FCM with $(\lambda_1^{opt}, \lambda_2^{opt})$.

In the case of a K -fold cross-validation, the penalized log-likelihood value associated with each (λ_1, λ_2) is computed as the mean of the K penalized likelihood values from the validation sets of the data folds (as opposed to the training sets of the data folds used for estimation). The optimal values of the tuning parameters are defined as the couple (λ_1, λ_2) that maximizes the cross-validated penalized log-likelihood. Then, a final run of the algorithm with the chosen $(\lambda_1^{opt}, \lambda_2^{opt})$ is launched using the whole

dataset. The estimators from this run are used to compute k^{opt} and G^{opt} .

- (4) **Find** (k^{opt}, G^{opt}) : The optimal values of G and k are identified by examining, respectively, the matrix of the between distances of the cluster means, \mathbf{D}_M , and the vector of elements $v_j = \left[\sum_{g=1}^G |\mu_g^j| \right]$ with $j = 1, \dots, k^{pos}$, which will be denoted by $\mathbf{V}_M = (v_j)_{j=1}^{k^{pos}}$.

• **Determining** k^{opt} : The optimal number of functional principal components is obtained by reducing k^{pos} . The elements associated with very small values of v_j are dropped from the model. For that purpose, two criteria are proposed and we consider as k^{opt} the minimum of the values provided by these two criteria.

- (a) The first criterion is inspired by the notion of inertia in classical principal components analysis. Recall that the inertia of a factor corresponds to the information it carried. In our setup, the inertia of a the j^{th} component is associated with its v_j value. Similarly to the criterion of cumulative proportion of total inertia, we search among the top ranked v_j values for the minimum number of principal components contributing to at least 80% of the cumulative sum ; that is, we look for the smallest $k^* \leq k^{pos}$ such that $\sum_{j=1}^{k^*} v_{(j)} \geq 0.80 \sum_{j=1}^{k^{pos}} v_j$, where $v_{(1)} \geq v_{(2)} \geq \dots \geq v_{(k^{pos})}$ are the ranked statistics associated with the components of the vector \mathbf{V}_M .
- (b) The other criterion is based on an approximate multiple testing procedure. Consider $\{\mu_1^j, \dots, \mu_G^j\}$, as a sample, and $v_j = \left[\sum_{g=1}^G |\mu_g^j| \right]$ as an associated statistic. We would like to test the null hypothesis of zero posterior expectation $E(\mu_g^j) = 0$, for all $g = 1, \dots, G$. As a heuristic, we suppose that under the null hypothesis the posterior of each μ_g^j follows a mean-zero Normal distribution with a common variance σ_k^2 . Under the null hypothesis, v_j is distributed as a sum of G independent half-Normal($0, \sigma_k^2$). The half-normal distribution is a fold at the mean of an ordinary normal distribution with mean zero. It is essentially the distribution of the absolute value of a normal distribution with mean zero. Although the distribution of our test statistic is known, its density is not known in closed form. However, we could easily estimate percentiles from this distribution by Monte Carlo simulation if σ_k^2 were known. Note that under the null hypothesis, $E(|\mu_g^j|) = \sigma_k \sqrt{2/\pi}$. Therefore, up to a constant, the mean of the components of the vector \mathbf{V}_M is an estimate of σ_k . In practice, we expect only a few components to be negligible ; so the estimate of σ_k may be taken from only the smallest v_j s, or even just the minimum of the v_j s. Since k^{pos} simultaneous tests need to be performed (one for each v_j), we apply a Bonferroni correction and work with a threshold $R_{\alpha/k}$ so that we do not reject the null hypothesis if the observed value of v_j is smaller or equal to this

threshold. Note that the threshold is given by the equation $P(T \leq R_{\alpha/k}) = \alpha/k$, where T is distributed as σ_k times the sum of G independent half-Normal(0, 1). That is,

$$R_{\alpha/k} = \sigma_k q_{\bar{T}, \alpha/k}$$

where $q_{\bar{T}, \alpha}$ is the 100 α^{th} percentile of the sampled distribution of the sum of G independent half-Normal(0, 1).

• **Determining G^{opt}** : We recall that the fundamental idea in the identification of the optimal number of clusters is the merging of clusters with identical characteristics, that is, with very small between distances. We suppose that the distances between vectors given by the matrix $\mathbf{D}_M = (D_{gh})_{1 \leq g, h \leq G}$ are Euclidean distances. Our heuristic assumes that under the null hypothesis of null distance (that is $\boldsymbol{\mu}_g = \boldsymbol{\mu}_h$), the posterior distribution of each pair of means is given by $(\boldsymbol{\mu}_g - \boldsymbol{\mu}_h) \sim \mathcal{N}_k(0, \sigma_G^2 I_k)$ with a common scale parameter σ_G^2 . In this case, we have :

$$D_{gh}^2 = \left[\sum_{l=1}^k (\mu_g^l - \mu_h^l)^2 \right] \sim \sigma_G^2 \chi_k^2$$

We estimate the scale parameter σ_G^2 with the mean of the squared distances as $\hat{\sigma}_G^2 = \left[\frac{2}{G(G-1)} \sum_{g < h} D_{gh}^2 \right]$, and plug this estimator in the above equation. As in the case of k^{opt} , in practice, we only use the smallest elements D_{gh} in the estimation of $\hat{\sigma}_G^2$. Using the Bonferroni correction for multiple testing, the null hypothesis is not rejected if the observed value D_{gh}^2 is not larger than $\hat{\sigma}_G^2$ times the lower $100(\alpha/G)^{th}$ percentile of a χ_k^2 distribution. Note that for simplicity, we have assumed mutual independence between the distances.

4.4. SIMULATION STUDY

We conduct a simulation study to examine the performance of the proposed methodology. We investigate specifically the ability of the method to reproduce and cluster original curves by correctly estimating the key parameters from postulated values. Following the described model selection procedure, the simulation study also concentrates on identifying the most relevant threshold to consider for the determination of G^{opt} .

Simulations setup

The process used to generate the simulated curves is the same as described in Adjogou et al. [1]. Various curves are generated based on different values of the sample size $N \in \{100, 500, 900, 4000\}$, the spline basis dimension $q \in \{10, 12, 14, 15, 20\}$, the number of functional principal components $k \in \{2, 4, 5, 6, 8\}$ and the number of clusters $G \in \{3, 6, 9, 15, 20, 40\}$. Overall, for each combination of (N, q, k, G) , the parameters of the model $\{\boldsymbol{\theta}_\mu, \boldsymbol{\Theta}, \mathbf{D}, \boldsymbol{\Gamma}_\mu, \pi_1, \pi_2, \dots, \pi_G, \nu_0, \sigma^2, \boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_G, \boldsymbol{\Gamma}_1, \dots, \boldsymbol{\Gamma}_G\}$ are generated at random according to the distributions assumed in the proposed model. We consider individual curves measured at 12 period times. This choice reflects the size of data from the microarray transcriptome of rats exposed to cigarette smoke (see Section 4.5).

For the simulations, $n^{LHS} = 50$ couples (λ_1, λ_2) are randomly obtained from LHS with bounds $(a^{LHS} = 0.1$ and $b^{LHS} = 50)$. As the B-splines basis dimension q is not estimated, its value is set to the one used to generate the curves. We postulate $k^{pos} = 8$ for the number of functional principal components. For the sake of efficiency in the simulations, the postulated values G^{pos} for the number of clusters are data dependent, and depend on the true number of clusters G . The corresponding set of values are given in Table 4.1.

Three different thresholds were proposed to estimate the optimal number of clusters. As mentioned in the previous section, we consider only the smallest distances for the estimation of $\hat{\sigma}_G^2$. The first criterion, **1low**, uses the smallest distance in \mathbf{D}_M ; the second criterion, **25low**, uses the mean of the distances falling below the first quartile of the distances in \mathbf{D}_M ; and the third criterion, **50low**, uses the mean of the distances falling below the median of the distances in \mathbf{D}_M .

TABLE 4.1. Values of postulated number of clusters according to G

G	G^{pos} values				
3	12	10	8	6	4
6	16	14	12	10	8
9	19	17	15	13	11
15	26	24	22	20	18
20	30	28	26	24	22
40	50	48	46	44	42

Simulation analysis tools

The simulation is based on 12 datasets generated as described above, each one corresponding to a specific (and different) combination of (N, q, k, G) . For each dataset, the couples (λ_1, λ_2) were generated and the procedure was launched for each value of G^{pos} , with $k^{pos} = 8$. The quality of the results is assessed by comparing the partitions (clustering) created by the model and the original (true) cluster memberships. The comparison was made through the Adjusted Rand Index (ARI) (Rand [25], Hubert and Arabie [13]). A perfect agreement between the two partitions yields an ARI score of 1. The closest the score is to 1, the more similar the partitions are. The ARI has become the standard measure of comparison in the statistical literature on clustering. Therefore, in addition to the model parameters estimates, the relevant quantities yielded by each simulation run are $\{k^{opt}, G_{1low}^{opt}, ARI_{1low}, G_{25low}^{opt}, ARI_{25low}, G_{50low}^{opt}, ARI_{50low}\}$, where the subindex represents the criterion used to choose G^{opt} .

In order to calibrate the ARI index with the difficulty of the problem, we also report a measure of data complexity as presented by Chen et al. [3]. Let N be the total number of curves in the data, n_g be the number of curves in cluster g , and MC_g be the mean curve of cluster g . Consider the following quantities of *Homogeneity* and *Separation* given respectively by

$$H.ave = \left[\frac{1}{N} \sum_{i=1}^N D^{ist}(Y_i, MC_{z_i}) \right], \quad S.ave = \left[\frac{1}{\sum_{g \neq h}^G n_g n_h} \sum_{g \neq h}^G n_g n_h D^{ist}(MC_g, MC_h) \right].$$

The Homogeneity is calculated as the average distance between each curve and the mean curve of the cluster it belongs to. It reflects the compactness of the clusters. The Separation is calculated as the weighted average distance between the cluster mean curves. It reflects the overall distance between clusters. As the indices $H.ave$ and $S.ave$ are closely related to respectively within-cluster and between-cluster variances, the similarity ratio

$$Ratio = \left[1 - \left(\frac{N}{N-1} \right) \left(\frac{H.ave}{H.ave + S.ave} \right) \right]$$

serves as a measure of homogeneity. Therefore, datasets with large Ratios are easier to cluster than those with small Ratios.

Simulation results

The first element of the simulation study is the comparison of the three criteria proposed to estimate the optimal number of clusters. For each dataset, we computed, for each criterion,

the average of the ARIs from the five different postulated G^{pos} . The results of an analysis of variance indicate that the criteria are significantly different. As shown in Figure 4.1, the criterion **low** appears to perform best. In Figure 4.2, we compare the similarity ratios and

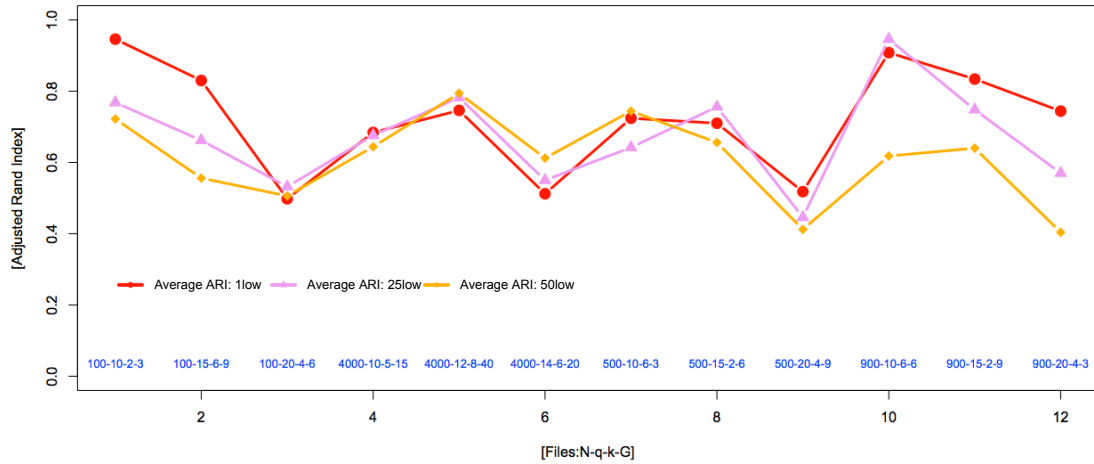


FIGURE 4.1. Comparison of the three criteria for model selection.

the criterion **low** average ARI_{1low} for every dataset. In addition, in this figure we also compare the performance of the current model with the one presented in Adjogou et al. [1]. This latter model uses an approximate Bayes-factor type criterion to choose the model parameters k , q and G . Because this model evaluates all possible models in a grid of values of (k, q, G) , its performance might be better than the one of our model. However, for the same reason, its computational cost is much larger than ours. The boxplots in Figure 4.2 are associated with the ARI values obtained from the five different postulated G^{pos} . The figure shows that the results from the two different functional model-based clustering models are comparable. However, our model has found the clusterings with much less computational cost. Also, note that the trend in the similarity ratio is also depicted in the ARI averages. This observation reflects that the clustering strength of the models are highly related to the degree of complexity of the data.

Another element analyzed in the simulation is the impact of the proposed number of cluster G^{pos} . The question addressed here is whether there is a significant difference in the clustering results if G^{pos} is far or close to the real number of clusters. The answer would give an indication on how to select G^{pos} in practice. For that purpose, we draw in Figure 4.3 a scatterplot of the values $(\sqrt{G^{pos}} - \sqrt{G})$ (representing the gap between proposed and real

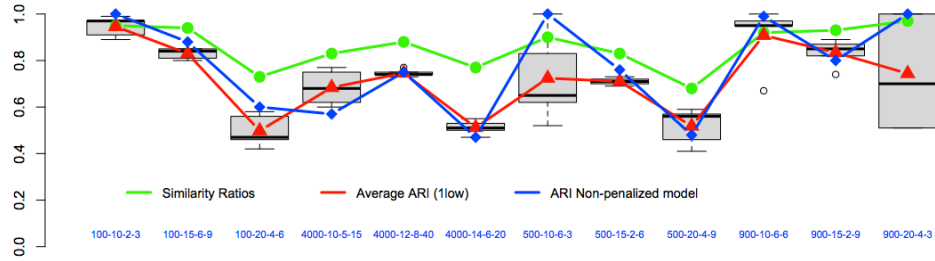


FIGURE 4.2. Similarity ratio and model performance.

G) against the corresponding ARI_{low} criterion). There is no structure nor trend observable from this figure. The conclusion is that there appears to be no relationship between the clustering performance and the proposed G in the algorithm. No matter how close or far G^{pos} is to the real number of clusters, the model performs similarly.

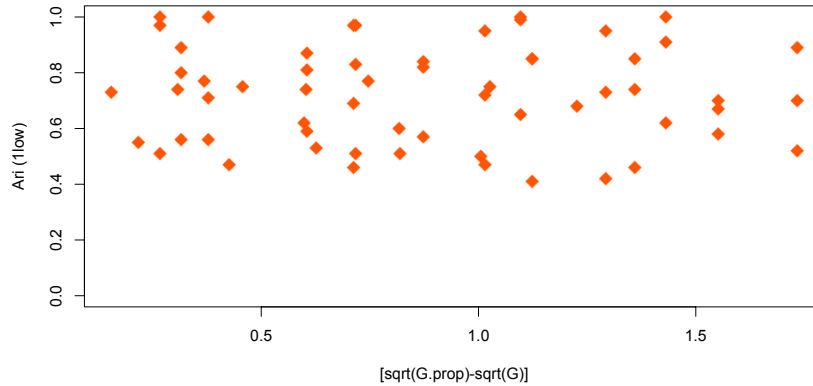


FIGURE 4.3. Influence of postulated G

Finally, we evaluate the capacity of the model to replicate the real number of clusters in the data. Consider the gap variable between estimated and real G given by $(\sqrt{G^{est}} - \sqrt{G})$. For the non-penalized functional model of [1], the gap is a single value for each dataset. For the penalized functional model presented here, the gap variable is represented by the set of G^{est} obtained through the postulated values for G . This is reflected by the boxplots in Figure 4.4. Note that the average gap values for the current model are very small and comparable to the ones from the non-penalized functional model.

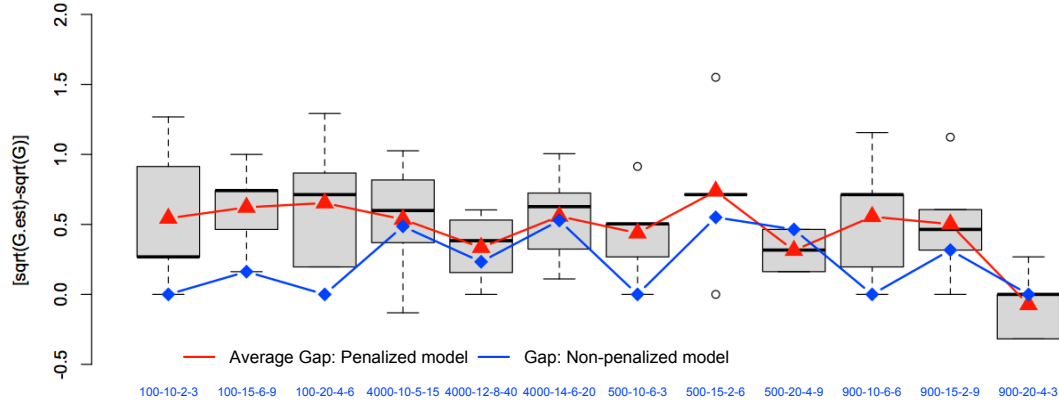


FIGURE 4.4. Comparison of True and estimated number of clusters

4.5. CHRONIC OBSTRUCTIVE PULMONARY DISEASE

We applied the Bayesian Lasso functional clustering model to shed light into the initial molecular events linked to chronic obstructive pulmonary disease (COPD). The dataset, described previously in the introduction section, relates to time-course genetic expression difference between tobacco-smoke exposed rats (the treatment group) and a control group of non-exposed rats (Stevenson et al. [28]). The dataset comes from the project GEO GSE7079 (Gene Expression Omnibus [9]) and is related to a study of molecular changes due to the exposure of rats to tobacco smoke. It is a time-course data with 12 day time-points : 1, 3, 5, 14, 21, 28, 42, 56, 84, 112, 182 and 238. Probesets (genes) without any GO annotation were discarded (Gene Ontology Consortium [10]). The 3464 probesets considered in our study correspond to 39.4% of the original 8799 probesets in the dataset.

Analysis

We set the initial proposed number of clusters G^{pos} equal to 50, 20, and 10. These choices led respectively to models with 22, 11 and 7 clusters. Despite the difference in the number of clusters, all three partitions are very similar in the sense that those partitions with smaller number of clusters are basically formed by merging of clusters in the larger partitions. Figure 4.5 displays the cluster means from all three partitions. The bottom row shows

the estimated cluster-specific mean curves, while the plots on the first row show the two-dimensional graphical representation of the functional principal components scores. Recall that the cluster mean curves are given by linear combinations of the functional principal components. For this particular data, these two-dimensional representations are exact because the estimated dimension k of the curves is exactly 2.

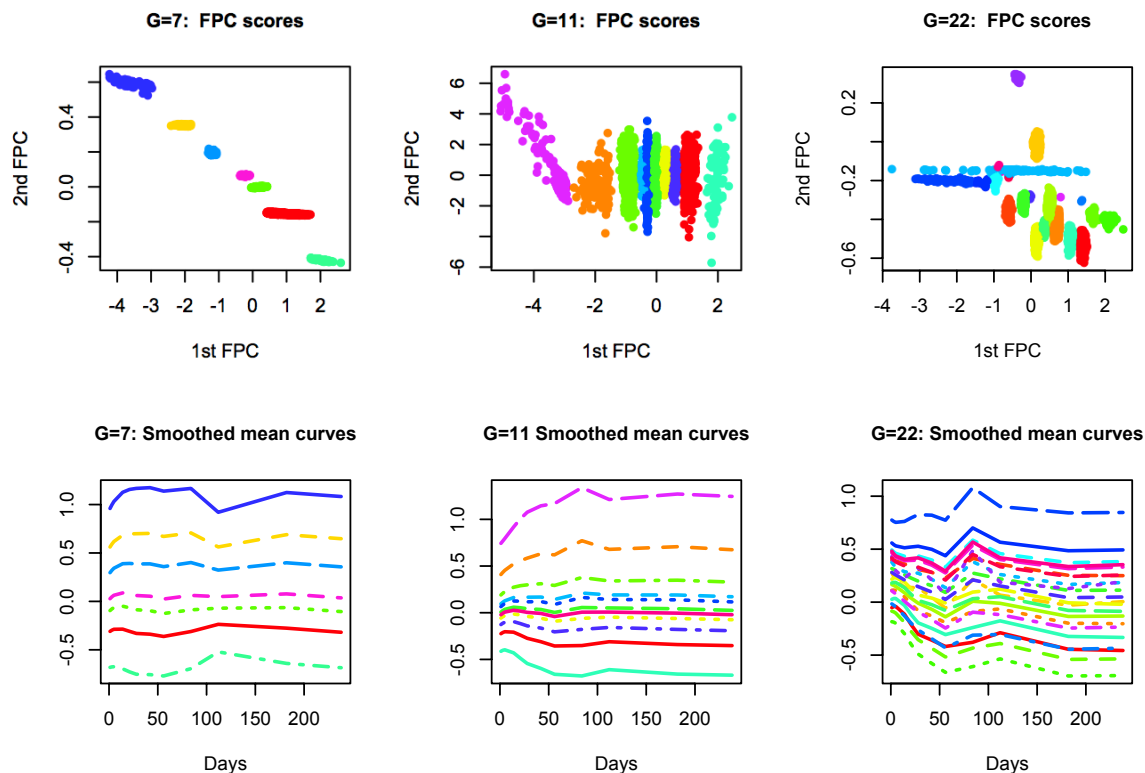


FIGURE 4.5. Cluster mean curves for the three partitions found by the Bayesian Lasso clustering model. The first row displays the clusters in the two-dimensional space of functional principal components (FPC) scores. The bottom row shows the cluster mean-curves.

The three partitions of genes given by the Bayesian Lasso functional clustering model were inspected for enrichment of functionalities with the DAVID platform (Huang et al. [12], Maere et al. [22], National Institute of Allergy and Infectious Diseases, NIH [23]). The partition of 22 clusters revealed several clusters highly enriched in functions previously attributed to acute and chronic exposure to cigarette smoke : immune response/immune system (clusters 1, 8, and partially 12), inflammation (clusters 15, 16), and apoptosis (clusters 3, 5, 6, 11). Surprisingly, these latter clusters, which are associated to late/prolonged exposure

with cigarette smoke, do not share general expression patterns such as global up or down-regulation profiles. However, clusters 15 and 16, associated with early exposure, do share a common upregulated expression pattern. In the 7-cluster partition, four clusters may be characterized by enrichment of gene functions directly related to early and late phases of tobacco smoke exposure. Among them, there is a cluster which despite its large size still has an interesting expression profile and excellent GO ontology enrichment scores : it represents genes that are gradually and increasingly repressed during the entire process of exposure to smoke. Curiously, all clusters enriched in gene functions associated with early phase of smoke exposure are also enriched in functions associated to long/chronic exposure. Among these, genes in cluster 2 do not show major expression changes during the late phase. Probably, these genes are not genes triggering chronic symptoms, but are genes that when activated « set the stage », that is, they may be associated with acute sensitivity for developing symptoms at a prolonged smoke exposure. In contrast, the cluster of genes specifically upregulated in the late phase has a more fuzzy profile, that is, there is no simple or clear tendency in the gene expressions. The 11-cluster partition presents three clusters identified as predominantly characterizing gene-functions associated with exposure to smoke. In all three clusters, the expression profile are in perfect accord to whether genes functions are associated with early or late phases of smoking.

4.6. CONCLUSION

In this paper, we introduced a model-based Bayesian Lasso functional clustering method for the analysis of longitudinal data. The model combines dimension reduction and clustering through functional principal component analysis, and model-based clustering. Model selection is done through a Lasso driven prior for the cluster means. Latin Hypercube Sampling was used to efficiently explore the space of penalty parameters.

The analysis of gene expression from smoke exposure showed that many deregulation events are associated with relevant gene-functions. This suggests that gene-repression may be a very common effect associated with biological effects of smoke exposure. We note that gene-repression is typically more difficult to find by classical data analysis approaches, and in consequence, it is frequently less regarded. The case of upregulated genes may thus be more punctual for specific aspects. In summary, one may conclude that the clustering approach allowed for identification of large groups of gradually deregulated genes that otherwise might be difficult to capture using traditional statistical approaches such as multiple testing of two groups (e.g., smoke-exposed versus control groups).

Bibliographie

- [1] Adjogou, F., K. Dorman, and A. Murua (2017). Functional model-based clustering for longitudinal data. Article to be submitted.
- [2] Bouveyron, C. and C. Brunet (2014). Model-based clustering of high-dimensional data : A review. *Computational Statistics and Data Analysis* 71, 52–78.
- [3] Chen, G., S. A. Jaradat, N. Banerjee, T. S. Tanaka, M. S. H. Ko, and M. Q. Zhang (2002). Evaluation and comparison of clustering algorithms in analyzing es cell gene expression data. *Statistica Sinica*, 241–262.
- [4] Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the em algorithm (with discussion). *Journal of the Royal Statistical Society* 39, 1–38.
- [5] Dupuy, D., C. Helbert, and J. Franco (2015). DiceDesign and DiceEval : Two R packages for design and analysis of computer experiments. *Journal of Statistical Software* 65(11), 1–38.
- [6] Fraley, C. and A. E. Raftery (1999). Mclust : Software for model-based cluster analysis. *Journal of Classification* 16(2), 297–306.
- [7] Fraley, C. and A. E. Raftery (2003). Enhanced software for model-based clustering, density estimation, and discriminant analysis : Mclust. *Journal of Classification* 20, 263–286.
- [8] Fraley, C. and A. E. Raftery (2006). Mclust version 3 for r : Normal mixture modeling and model-based clustering. Technical Report 504. Department of Statistics, University of Washington.
- [9] Gene Expression Omnibus (2007). Chronic rat exposure to cigarette smoke. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE7079>.
- [10] Gene Ontology Consortium (1999-2015). GO. <http://www.geneontology.org/>.
- [11] Giacomini, M., S. Lambert-Lacroix, G. Marot, and F. Picard (2013). Wavelet-based clustering for mixed-effects functional models in high dimension. *Biometrics* 69(1), 31–40.
- [12] Huang, D. W., B. T. Sherman, and R. A. Lempicki (2009). Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nature Protoc.* 4(1), 44–57.

- [13] Hubert, L. and P. Arabie (1985). Comparing partitions. *Journal of Classification* 2, 193–218.
- [14] Iman, R. and W. Conover (1982a). A distribution-free approach to inducing rank correlation among input variables. *Communications in Statistics B11*(3), 311–334.
- [15] Iman, R. and W. Conover (1982b). Sensitivity analysis techniques : Self-teaching curriculum. Nuclear Regulatory Commission Report, NUREG/CR-2350, Technical Report SAND81-1978, Sandia National Laboratories, Albuquerque, NM.
- [16] Iman, R., J. Helton, and J. Campbell (1981a). An approach to sensitivity analysis of computer models, part 1. introduction, input variable selection and preliminary variable assessment. *Journal of Quality Technology* 13(3), 174–183.
- [17] Iman, R., J. Helton, and J. Campbell (1981b). An approach to sensitivity analysis of computer models, part 2. ranking of input variables, response surface validation, distribution effect and technique synopsis. *Journal of Quality Technology* 13(4), 232–240.
- [18] Iman, R. L. and J. C. Helton (1985). A comparison of uncertainty and sensitivity analysis techniques for computer models. NUREGKR-3904, SAND84-1461. Albuquerque, NM : Sandia National Laboratories.
- [19] Jacques, J. and C. Preda (2014). Functional data clustering : a survey. *Advances in Data Analysis and Classification, Springer Verlag* 8(3).
- [20] James, G. and C. A. Sugar (2003). Clustering for sparsely sampled functional data. *Journal of the American Statistical Association* 98, 397–408.
- [21] Laird, N. M. and J. H. Ware (1982). Random-effects models for longitudinal data. *Biometrics* 38, 963–974.
- [22] Maere, S., K. Heymans, and M. Kuiper (2005). Bingo : a cytoscape plugin to assess over-representation of gene ontology categories in biological networks. *Bioinformatics* 21(16), 3448–3449.
- [23] National Institute of Allergy and Infectious Diseases, NIH (2017). DAVID bioinformatics resources. <https://david.ncifcrf.gov/>.
- [24] Pan, W. and X. Shen (2007). Penalized model-based clustering with application to variable selection. *Journal of Machine Learning Research* 8, 1145–1164.
- [25] Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association* 66, 846–850.
- [26] Same, A., F. Chamroukhi, G. Govaert, and P. Aknin (2011). Model-based clustering and segmentation of times series with changes in regime. *Advances in Data Analysis and Classification* 5(4), 301–322.
- [27] Steck, G. P., R. L. Iman, and D. A. Dahlgren (1976). Probabilistic analysis of loca , annual report for 1976. SAND76-0535, Sandia National Laboratories, Albuquerque, NM.

- [28] Stevenson, C., D. C., R. Webster, C. Battram, D. Hynx, J. Giddings, P. Cooper, P. Chakravarty, I. Rahman, J. Marwick, P. Kirkham, C. Charman, D. Richardson, N. Nirmala, P. Whittaker, and K. Butler (2007). Comprehensive gene expression profiling of rat lung reveals distinct acute and chronic responses to cigarette smoke inhalation. *Am. J. Physiol Lung Cell Mol. Physiol.* 293(5), L1183–93.
- [29] Tibshirani, R. (2011). Regression shrinkage and selection via the lasso : a retrospective. *Journal of the Royal Statistical Society Series B* 73, 273–282.
- [30] Ullah, S. and C. F. Finch (2013). Applications of functional data analysis : A systematic review. *BMC Medical Research Methodology* 1471-2288, 13–43.
- [31] Verbeke, G. and G. Molenberghs (2000). *Linear Mixed Models for Longitudinal Data*. Springer series in statistics. New York.
- [32] Wang, S. and J. Zhou (2008). Variable selection for model-based high dimensional clustering and its application to microarray data. *Biometrics* 64, 440–448.
- [33] Wyss, G. D. and K. H. Jorgensen (1998). A user’s guide to lhs : Sandia’s latin hypercube sampling software. AND98-0210 Distribution Unlimited Release Category UC-505. Risk Assessment and Systems Modeling Department Sandia National Laboratories.
- [34] Zhao, X., J. Marron, and M. Wells (2004). The functional data analysis view of longitudinal data. *Statistica Sinica* 14, 789–808.

Chapitre 5

CONCLUSION

Devenues un outil majeur dans l'étude de l'évolution temporelle d'un phénomène, les données longitudinales sont essentiellement caractérisées par la dépendance entre les mesures répétées prises sur un même individu et requièrent une analyse statistique spécifique et appropriée. Le développement de telles méthodologies statistiques constitue un domaine d'étude en pleine expansion dans la majorité des disciplines scientifiques notamment en sciences sociales et médicales où les données longitudinales sont de plus en plus utilisées.

Cette thèse regroupe nos contributions dans le domaine des méthodologies d'analyse et de partitionnement des données longitudinales, qu'elles soient équilibrées ou non équilibrées. Nous y présentons deux méthodologies nouvelles, performantes et compétitives de classification non supervisée basée sur l'approche de l'analyse des données fonctionnelles (ADF) ainsi qu'une revue de la littérature sur l'ensemble des méthodes de regroupement en classes de données longitudinales. Selon l'approche de l'ADF qui est de plus en plus exploitée en analyse de données longitudinales ou de données de grande dimension en général, les mesures répétées recueillies auprès des individus sont considérées comme des fonctions ou courbes partiellement observées sur un axe temporel. Ces méthodes de regroupement en classes basées sur des modèles que nous proposons dans le cadre de la présente thèse utilisent conjointement l'analyse fonctionnelle en composantes principales et la modélisation à effets mixtes mais se distinguent par leur flexibilité, la pertinence des hypothèses sous-jacentes ou encore leur construction dans un cadre bayésien.

Le premier chapitre de cette thèse a été consacré à l'introduction. Dans le deuxième chapitre, nous avons présenté un aperçu général des différentes méthodes d'analyse des données longitudinales avec une emphase particulière sur les méthodes de regroupement en classes selon les approches les plus communément utilisées. La grande majorité des procédures existantes dans ce domaine proviennent de l'extension de méthodes disponibles pour l'analyse

de données indépendantes.

Dans le deuxième chapitre de cette thèse, nous avons présenté le modèle flexible développé pour l'analyse et le partitionnement de tout type de données longitudinales, que les mesures soient uniformément ou non uniformément obtenues sur l'axe temporel pour les différents individus. Le modèle combine l'analyse fonctionnelle en composantes principales et le regroupement en classes qui repose sur l'espace des coefficients dans la base des splines et un modèle de mélange de distributions de Student de degrés de liberté inconnus. Un nouveau critère de sélection de modèle basé sur une approximation de la log-vraisemblance marginale (MLL) a été développé. Les études de simulations réalisées et les applications sur des données réelles d'expression génétique, ainsi que les comparaisons effectuées avec d'autres procédures existantes dans ce domaine en plein essor confirment la pertinence et l'efficacité du nouveau modèle proposé. Une extension de ce modèle au cas multidimensionnel a été proposée et également évaluée.

Nous avons présenté dans le troisième chapitre de cette thèse, une autre nouvelle procédure de partitionnement pour l'analyse des données longitudinales, qui utilise l'approche fonctionnelle et une pénalisation double du type Lasso dans la fonction de log-vraisemblance pour simultanément déterminer la dimension appropriée de la base finie de fonctions (réduction de la dimension) et le nombre approprié de groupes homogènes (partitionnement). La performance et l'utilité de la procédure ont été démontrées par la simulation et l'application sur des données réelles. Il faut rappeler qu'un des aspects novateurs dans cette méthodologie est lié à l'utilisation du « Latin Hypercube Sampling (LHS) » pour le choix de la grille de valeurs des deux paramètres d'ajustement du modèle.

Une caractéristique importante et commune aux deux méthodologies de regroupement proposées est le développement d'outils spécifiques et efficaces pour la sélection de modèles. D'une part, un nouveau critère basé sur une approximation de log-vraisemblance marginale (MLL) a été proposé et se compare efficacement à d'autres critères usuels similaires tels que AIC et BIC. D'autre part, outre l'option de la validation croisée, l'approche directe d'optimisation utilise une transformation particulière appliquée à la log-vraisemblance pénalisée et basée sur une méthode d'intégration numérique de Monte Carlo dans le but de réaliser la sélection de modèle optimal.

Une quantité de travail non négligeable associée à la réalisation de cette thèse réside dans le développement des programmes automatisés ou codes pour l'exécution des méthodologies proposées. Ces codes écrits en *R* et *Java* feront très prochainement l'objet d'une mise à jour

pour la création d'une application logicielle ou progiciel spécialisé dans le partitionnement de données longitudinales, qu'elles soient constituées d'une seule ou de plusieurs variables réponses, qu'elles soient balancées ou non balancées, et avec prise en compte d'un ou de plusieurs effets fixes.

Quelques aspects des éléments novateurs proposés dans cette thèse peuvent faire valablement l'objet d'amélioration ou d'extension. Par exemple, étant donné que l'analyse des données fonctionnelles (ADF) s'intéresse aux données à dimension infinie telles que les courbes ou les images, les modèles proposés peuvent être étendus pour traiter également le regroupement en classes de données longitudinales lorsque les observations ne sont plus des courbes, mais des images. En effet, dans le domaine de la médecine, de la neuroscience et de la psychologie, les données d'imagerie par résonance magnétique fonctionnelle (IRMf) sont de plus en plus analysées par l'approche de l'analyse des données fonctionnelles (Lindquist [1]). Ces images mesurent les réponses hémodynamiques du cerveau et indiquent l'évolution de l'activité neuronale avec une grande résolution spatiale. Chaque image étant constituée de *voxels* qui partitionnent uniformément le cerveau, il est possible de suivre la variation de l'amplitude de l'activité neuronale à chaque voxel à travers le temps. De plus, cette expérience peut être répétée plusieurs fois sur le même sujet, aussi bien que sur d'autres sujets. Il serait donc intéressant d'étudier les modes de variation ou les caractéristiques communes de l'activité cérébrale de plusieurs individus à partir des données de IMRf. Notons que Tian [2] présente quelques techniques statistiques d'analyse de données fonctionnelles comme la réduction de dimension dans les études d'imagerie cérébrale.

Bibliographie

- [1] Lindquist, M. (2008). The statistical analysis of fmri data. *Statistical Science* 23(4), 439–464.
- [2] Tian, T. S. (2010). Functional data analysis in brain imaging studies. *Frontiers in Quantitative Psychology and Measurement* 1(35), 1–11.

Annexe A

SOME ANALYTICAL DETAILS ON PARTITIONING AND EM STEPS

A.1. PARTITIONING OF AN INCOMPLETE MULTIVARIATE GAUSSIAN DATA

If a p -dimensional random variable \mathbf{Y} is partitioned as

$$\begin{bmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{bmatrix} \sim MVN \left(\begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Gamma}_{11} & \boldsymbol{\Gamma}_{12} \\ \boldsymbol{\Gamma}_{21} & \boldsymbol{\Gamma}_{22} \end{bmatrix} \right) \quad (\text{A.1.1})$$

where \mathbf{Y}_1 is an r -dimensional vector and \mathbf{Y}_2 is an $(p - r)$ -dimensional vector, it can be shown that the conditional distribution of \mathbf{Y}_1 , given that $\mathbf{Y}_2 = \mathbf{y}_2$, is multivariate Gaussian with mean $[\boldsymbol{\mu}_1 + \boldsymbol{\Gamma}_{12}\boldsymbol{\Gamma}_{22}^{-1}(\mathbf{y}_2 - \boldsymbol{\mu}_2)]$ and covariance $[\boldsymbol{\Gamma}_{11} - \boldsymbol{\Gamma}_{12}\boldsymbol{\Gamma}_{22}^{-1}\boldsymbol{\Gamma}_{21}]$. See Shaikh et al. [1].

A.2. ANALYTICAL DEVELOPMENTS FOR EM EXPECTATION STEP :

The log-likelihood $\log[p(\mathbf{Y}, \mathbf{W}; \boldsymbol{\Pi})]$ is

$$\begin{aligned} & \sum_{i=1}^N \left\{ \begin{aligned} & -\frac{n_i}{2} \log(\nu_i \sigma^2) - \frac{1}{2\nu_i \sigma^2} \left\| \mathbf{Y}_i - (\mathbf{B}_i \boldsymbol{\theta}_\mu + \mathbf{B}_i \boldsymbol{\Theta} \boldsymbol{\mu}_{\mathbf{z}_i} + \mathbf{B}_i \boldsymbol{\Theta} \boldsymbol{\gamma}_i^{\mathbf{z}_i}) \right\|^2 \\ & -\frac{1}{2} \log(|\boldsymbol{\Gamma}_{\mathbf{z}_i}|) - \frac{1}{2} \boldsymbol{\gamma}_{i, \mathbf{z}_i}^T \boldsymbol{\Gamma}_{\mathbf{z}_i}^{-1} \boldsymbol{\gamma}_i^{\mathbf{z}_i} + \sum_{g=1}^G Z_{ig} \log(\pi_g) \\ & + \frac{\nu_o}{2} \log\left(\frac{\nu_o}{2}\right) - \log[\Gamma\left(\frac{\nu_o}{2}\right)] - \left(1 + \frac{\nu_o}{2}\right) \log(\nu_i) - \frac{\nu_o}{2\nu_i} \end{aligned} \right\} \\ & + \sum_{g=1}^G \left\{ \begin{aligned} & -\frac{1}{2} \log(|\boldsymbol{\Gamma}_\mu|) - \frac{1}{2} \boldsymbol{\mu}_g^T \boldsymbol{\Gamma}_\mu^{-1} \boldsymbol{\mu}_g + \frac{m}{2} \log(|(m - k - 1)\mathbf{D}|) \\ & - \frac{(m+k+1)}{2} \log(|\boldsymbol{\Gamma}_g|) - \frac{(m-k-1)}{2} \text{trace}[\mathbf{D}\boldsymbol{\Gamma}_g^{-1}] \end{aligned} \right\} \\ & + \left\{ \frac{km}{2} \log(m - k - 1) - \frac{(m+k+1)}{2} \log(|\boldsymbol{\Gamma}_\mu|) - \frac{(m-k-1)}{2} \text{trace}[\boldsymbol{\Gamma}_\mu^{-1}] \right\} \\ & + \sum_{j=1}^k \left\{ \begin{aligned} & + \frac{m}{2} \log\left(\frac{m}{2}\right) - \log[\Gamma\left(\frac{m}{2}\right)] - \left(1 + \frac{m}{2}\right) \log(d_{jj}) - \frac{m}{2d_{jj}} \end{aligned} \right\} \\ & + \left\{ \alpha_\sigma \log(\beta_\sigma) - \log[\Gamma(\alpha_\sigma)] - (\alpha_\sigma + 1) \log(\sigma^2) - \frac{\beta_\sigma}{\sigma^2} \right\} \\ & + \left\{ -\log[B(a_1, \dots, a_G)] + \sum_{g=1}^G (a_g - 1) \log(\pi_g) \right\} \\ & + \mathcal{C} \end{aligned} \quad (\text{A.2.1})$$

where \mathcal{C} is the normalizing constant, and $B(a_1, \dots, a_G) = B(\mathbf{a})$ is the multivariate Beta function which can be expressed in terms of the Gamma function $\Gamma(\cdot)$ as $B(\mathbf{a}) = \frac{[\prod_{g=1}^G \Gamma(a_g)]}{\Gamma(\sum_{g=1}^G a_g)}$.

We can rewrite the last expression of $\log[p(\mathbf{Y}, \mathbf{W}; \mathbf{\Pi})]$ as $\log[p(\mathbf{Y}, \mathbf{W}; \mathbf{\Pi})] = \mathcal{L} + \mathcal{H}$ where \mathcal{L} groups the terms depending on the individuals $i = 1, \dots, N$:

$$\begin{aligned} \mathcal{L} &= \sum_{i=1}^N l_i(\vec{\mu}, \vec{\Gamma}, \Lambda) = \sum_{i=1}^N -\frac{n_i}{2} \log(\nu_i \sigma^2) - \frac{1}{2\nu_i \sigma^2} \left\| \mathbf{Y}_i - (\mathbf{B}_i \boldsymbol{\theta}_\mu + \mathbf{B}_i \boldsymbol{\Theta} \boldsymbol{\mu}_{\mathbf{z}_i} + \mathbf{B}_i \boldsymbol{\Theta} \boldsymbol{\gamma}_i^{\mathbf{z}_i}) \right\|^2 \\ &\quad - \frac{1}{2} \log(|\Gamma_{\mathbf{z}_i}|) - \frac{1}{2} \boldsymbol{\gamma}_{i, \mathbf{z}_i}^\top \boldsymbol{\Gamma}_{\mathbf{z}_i}^{-1} \boldsymbol{\gamma}_i^{\mathbf{z}_i} + \sum_{g=1}^G z_{ig} \log(\pi_g) \\ &\quad + \frac{\nu_o}{2} \log\left(\frac{\nu_o}{2}\right) - \log\left[\Gamma\left(\frac{\nu_o}{2}\right)\right] - \left(1 + \frac{\nu_o}{2}\right) \log(\nu_i) - \frac{\nu_o}{2\nu_i} \end{aligned}$$

and \mathcal{H} groups the remainder terms. The function Q to be maximized is

$$\begin{aligned} Q(\mathbf{\Pi} | \mathbf{\Pi}^{(t)}) &= E_{\mathbf{W} | \mathbf{Y}; \mathbf{\Pi}^{(t)}} [\log p(\mathbf{Y}, \mathbf{W}; \mathbf{\Pi})] = E_{\mathbf{z}, \boldsymbol{\gamma}^{\mathbf{z}} | \mathbf{Y}, \vec{\nu}^{(t)}, \vec{\mu}^{(t)}, \vec{\Gamma}^{(t)}, \Lambda^{(t)}} [\mathcal{L} + \mathcal{H}] \\ &= \sum_{i=1}^N E_{\mathbf{z}_i, \boldsymbol{\gamma}_i^{\mathbf{z}_i} | \mathbf{Y}_i, \nu_i^{(t)}, \vec{\mu}^{(t)}, \vec{\Gamma}^{(t)}, \Lambda^{(t)}} [l_i(\vec{\mu}, \vec{\Gamma}, \Lambda)] + \mathcal{H} \\ &= \sum_{i=1}^N E_{\mathbf{z}_i | \mathbf{Y}_i, \nu_i^{(t)}, \vec{\mu}^{(t)}, \vec{\Gamma}^{(t)}, \Lambda^{(t)}} \left\{ E_{\boldsymbol{\gamma}_i^{\mathbf{z}_i} | \mathbf{z}_i, \mathbf{Y}_i, \nu_i^{(t)}, \vec{\mu}^{(t)}, \vec{\Gamma}^{(t)}, \Lambda^{(t)}} [l_i(\vec{\mu}, \vec{\Gamma}, \Lambda)] \right\} + \mathcal{H}. \quad (\text{A.2.2}) \end{aligned}$$

Let $m_i(\mathbf{z}_i, \vec{\mu}, \vec{\Gamma}, \Lambda) = E_{\boldsymbol{\gamma}_i^{\mathbf{z}_i} | \mathbf{z}_i, \mathbf{Y}_i, \nu_i^{(t)}, \vec{\mu}^{(t)}, \vec{\Gamma}^{(t)}, \Lambda^{(t)}} [l_i(\vec{\mu}, \vec{\Gamma}, \Lambda)]$. In order to compute $m_i(\mathbf{z}_i, \vec{\mu}, \vec{\Gamma}, \Lambda)$, we need to find the conditional distribution of $\boldsymbol{\gamma}_i^{\mathbf{z}_i} | \mathbf{Y}_i, \mathbf{z}_i, \nu_i^{(t)}, \vec{\mu}^{(t)}, \vec{\Gamma}^{(t)}, \Lambda^{(t)}$. Using Bayes rule, we have

$$p(\boldsymbol{\gamma}_i^{\mathbf{z}_i} | \mathbf{Y}_i, \mathbf{z}_i, \nu_i^{(t)}, \vec{\mu}^{(t)}, \vec{\Gamma}^{(t)}, \Lambda^{(t)}) = \frac{p(\mathbf{Y}_i | \boldsymbol{\gamma}_i^{\mathbf{z}_i}, \mathbf{z}_i, \nu_i^{(t)}, \vec{\mu}^{(t)}, \vec{\Gamma}^{(t)}, \Lambda^{(t)}) p(\boldsymbol{\gamma}_i^{\mathbf{z}_i} | \mathbf{z}_i, \nu_i^{(t)}, \vec{\mu}^{(t)}, \vec{\Gamma}^{(t)}, \Lambda^{(t)})}{p(\mathbf{Y}_i | \mathbf{z}_i, \nu_i^{(t)}, \vec{\mu}^{(t)}, \vec{\Gamma}^{(t)}, \Lambda^{(t)})}$$

Note that all distributions involved in this expression are Gaussian. Let $\mathcal{N}_r(\mu, \Sigma)$ denote an r -variate Gaussian distribution with mean μ and covariance Σ . We have

$$\begin{aligned} \mathbf{Y}_i | \boldsymbol{\gamma}_i^{\mathbf{z}_i}, \mathbf{z}_i, \nu_i^{(t)}, \vec{\mu}^{(t)}, \vec{\Gamma}^{(t)}, \Lambda^{(t)} &\sim \mathcal{N}_{n_i}([\mathbf{B}_i \boldsymbol{\theta}_\mu^{(t)} + \mathbf{B}_i \boldsymbol{\Theta}^{(t)} \boldsymbol{\mu}_{\mathbf{z}_i}^{(t)} + \mathbf{B}_i \boldsymbol{\Theta}^{(t)} \boldsymbol{\gamma}_i^{\mathbf{z}_i}], [\sigma_{(t)}^2 \nu_i^{(t)} I_{n_i}]) \\ \boldsymbol{\gamma}_i^{\mathbf{z}_i} | \mathbf{z}_i, \nu_i^{(t)}, \vec{\mu}^{(t)}, \vec{\Gamma}^{(t)}, \Lambda^{(t)} &\sim \mathcal{N}_k(0, \boldsymbol{\Gamma}_{\mathbf{z}_i}^{(t)}) \\ \mathbf{Y}_i | \mathbf{z}_i, \nu_i^{(t)}, \vec{\mu}^{(t)}, \vec{\Gamma}^{(t)}, \Lambda^{(t)} &\sim \mathcal{N}_{n_i}([\mathbf{B}_i \boldsymbol{\theta}_\mu^{(t)} + \mathbf{B}_i \boldsymbol{\Theta}^{(t)} \boldsymbol{\mu}_{\mathbf{z}_i}^{(t)}], [\sigma_{(t)}^2 \nu_i^{(t)} I_{n_i} + \mathbf{B}_i \boldsymbol{\Theta}^{(t)} \boldsymbol{\Gamma}_{\mathbf{z}_i}^{(t)} \boldsymbol{\Theta}^{(t)\top} \mathbf{B}_i^\top]) \end{aligned}$$

These simplifications lead to a multivariate Gaussian distribution

$$\boldsymbol{\gamma}_i^{\mathbf{z}_i} | \mathbf{Y}_i, \mathbf{z}_i, \nu_i^{(t)}, \vec{\mu}^{(t)}, \vec{\Gamma}^{(t)}, \Lambda^{(t)} \sim \mathcal{N}_k(\hat{\boldsymbol{\gamma}}_i^{\mathbf{z}_i}, \hat{\mathbf{V}}_i^{\mathbf{z}_i}), \quad (\text{A.2.3})$$

with

$$\hat{\boldsymbol{\gamma}}_i^{\mathbf{z}_i} = \left\{ \nu_i^{(t)} \sigma_{(t)}^2 \boldsymbol{\Gamma}_{\mathbf{z}_i}^{-1(t)} + \boldsymbol{\Theta}_{(t)}^\top \mathbf{B}_i^\top \mathbf{B}_i \boldsymbol{\Theta}_{(t)} \right\}^{-1} \boldsymbol{\Theta}_{(t)}^\top \mathbf{B}_i^\top \left\{ \mathbf{Y}_i - \mathbf{B}_i \boldsymbol{\theta}_\mu^{(t)} - \mathbf{B}_i \boldsymbol{\Theta}_{(t)} \boldsymbol{\mu}_{\mathbf{z}_i}^{(t)} \right\}$$

$$\hat{V}_i^{\mathbf{z}_i} = \left\{ \mathbf{\Gamma}_{\mathbf{z}_i}^{-1(t)} + \frac{\mathbf{\Theta}_{(t)}^\top \mathbf{B}_i^\top \mathbf{B}_i \mathbf{\Theta}_{(t)}}{\nu_i^i \sigma_{(t)}^2} \right\}^{-1}$$

In the expression of $l_i(\vec{\mu}, \vec{\Gamma}, \mathbf{\Lambda})$, the random variable $\gamma_i^{\mathbf{z}_i}$ only occurs in the terms $\left\| \mathbf{Y}_i - (\mathbf{B}_i \boldsymbol{\theta}_\mu + \mathbf{B}_i \mathbf{\Theta}_{\mu_{\mathbf{z}_i}} + \mathbf{B}_i \mathbf{\Theta}_{\gamma_i^{\mathbf{z}_i}}) \right\|^2$ and $\gamma_{i,\mathbf{z}_i}^\top \mathbf{\Gamma}_{\mathbf{z}_i}^{-1} \gamma_{i,\mathbf{z}_i}$. The other terms are left unchanged by the expectation. Consider the following identity that applies to any random vector U of dimension n .

$$E(U^\top U) = \text{trace}(E[U^\top U]) = E(\text{trace}[U^\top U]) = E(\text{trace}[UU^\top]) \quad (\text{A.2.4})$$

$$= \text{trace}(E[UU^\top]) = \hat{U}^\top \hat{U} + \text{trace}(\hat{V}_U), \quad (\text{A.2.5})$$

where $\hat{U} = E(U)$, and $\hat{V}_U = \text{Var}(U)$. Using this, we get

$$\begin{aligned} E_{\gamma_i^{\mathbf{z}_i} | \mathbf{Y}_i, \mathbf{z}_i, \nu_i^{(t)}, \vec{\mu}^{(t)}, \vec{\Gamma}^{(t)}, \mathbf{\Lambda}^{(t)}} \left\{ \left\| \mathbf{Y}_i - (\mathbf{B}_i \boldsymbol{\theta}_\mu + \mathbf{B}_i \mathbf{\Theta}_{\mu_{\mathbf{z}_i}} + \mathbf{B}_i \mathbf{\Theta}_{\gamma_i^{\mathbf{z}_i}}) \right\|^2 \right\} \\ = \left\| \mathbf{Y}_i - \mathbf{B}_i \boldsymbol{\theta}_\mu^{(t)} - \mathbf{B}_i \mathbf{\Theta}_{(t)} \boldsymbol{\mu}_{\mathbf{z}_i}^{(t)} - \mathbf{B}_i \mathbf{\Theta}_{(t)} \hat{\gamma}_i^{\mathbf{z}_i} \right\|^2 + \text{trace} \left[\mathbf{B}_i \mathbf{\Theta}_{(t)} \hat{V}_i^{\mathbf{z}_i} \mathbf{\Theta}_{(t)}^\top \mathbf{B}_i^\top \right], \end{aligned}$$

and

$$E_{\gamma_i^{\mathbf{z}_i} | \mathbf{Y}_i, \mathbf{z}_i, \nu_i^{(t)}, \vec{\mu}^{(t)}, \vec{\Gamma}^{(t)}, \mathbf{\Lambda}^{(t)}} \left\{ \gamma_{i,\mathbf{z}_i}^\top \mathbf{\Gamma}_{\mathbf{z}_i}^{-1} \gamma_{i,\mathbf{z}_i} \right\} = \hat{\gamma}_{i,\mathbf{z}_i}^\top \mathbf{\Gamma}_{\mathbf{z}_i}^{-1} \hat{\gamma}_{i,\mathbf{z}_i} + \text{trace} \left[\mathbf{\Gamma}_{\mathbf{z}_i}^{-1} \hat{V}_i^{\mathbf{z}_i} \right],$$

which leads to the computation of $m_i(\mathbf{z}_i, \vec{\mu}, \vec{\Gamma}, \mathbf{\Lambda})$.

Next, we compute $E_{\mathbf{z}_i | \mathbf{Y}_i, \nu_i^{(t)}, \vec{\mu}^{(t)}, \vec{\Gamma}^{(t)}, \mathbf{\Lambda}^{(t)}} \left\{ m_i(\mathbf{z}_i, \vec{\mu}, \vec{\Gamma}, \mathbf{\Lambda}) \right\}$. This requires finding the distribution of the discrete random variable $\left\{ \mathbf{z}_i | \mathbf{Y}_i, \nu_i^{(t)}, \vec{\mu}^{(t)}, \vec{\Gamma}^{(t)}, \mathbf{\Lambda}^{(t)} \right\}$. Let \mathbf{e}_g be the G -dimensional vector whose components are all zero, except for the g th component which is set to 1. We have,

$$p(\mathbf{z}_i = \mathbf{e}_g | \mathbf{Y}_i, \nu_i^{(t)}, \vec{\mu}^{(t)}, \vec{\Gamma}^{(t)}, \mathbf{\Lambda}^{(t)}) \quad (\text{A.2.6})$$

$$\begin{aligned} &= \frac{p(\mathbf{Y}_i | \mathbf{z}_i = \mathbf{e}_g, \nu_i^{(t)}, \vec{\mu}^{(t)}, \vec{\Gamma}^{(t)}, \mathbf{\Lambda}^{(t)}) p(\mathbf{z}_i = \mathbf{e}_g | \nu_i^{(t)}, \vec{\mu}^{(t)}, \vec{\Gamma}^{(t)}, \mathbf{\Lambda}^{(t)})}{p(\mathbf{Y}_i | \nu_i^{(t)}, \vec{\mu}^{(t)}, \vec{\Gamma}^{(t)}, \mathbf{\Lambda}^{(t)})} \\ &= \frac{\pi_g^{(t)} p(\mathbf{Y}_i | \mathbf{z}_i = \mathbf{e}_g, \nu_i^{(t)}, \vec{\mu}^{(t)}, \vec{\Gamma}^{(t)}, \mathbf{\Lambda}^{(t)})}{\sum_{h=1}^G \pi_h^{(t)} p(\mathbf{Y}_i | \mathbf{z}_i = \mathbf{e}_h, \nu_i^{(t)}, \vec{\mu}^{(t)}, \vec{\Gamma}^{(t)}, \mathbf{\Lambda}^{(t)})}. \end{aligned} \quad (\text{A.2.7})$$

The computation of the expression in (A.2.6) requires knowledge of the distribution of the random variable $\mathbf{Y}_i | \mathbf{z}_i, \nu_i^{(t)}, \vec{\mu}^{(t)}, \vec{\Gamma}^{(t)}, \mathbf{\Lambda}^{(t)}$, which is a n_i -variate Gaussian random variable with

$$\mathbf{E}_{i,\mathbf{z}_i} = \mathbf{B}_i \boldsymbol{\theta}_\mu^{(t)} + \mathbf{B}_i \mathbf{\Theta}_{(t)} \boldsymbol{\mu}_{\mathbf{z}_i}^{(t)} \quad (\text{A.2.8})$$

$$\boldsymbol{\Sigma}_{i,\mathbf{z}_i} = \nu_i^i \sigma_{(t)}^2 \left[I_{n_i} - \mathbf{B}_i \mathbf{\Theta}_{(t)} \left(\nu_i^i \sigma_{(t)}^2 \mathbf{\Gamma}_{\mathbf{z}_i}^{-1(t)} + \mathbf{\Theta}_{(t)}^\top \mathbf{B}_i^\top \mathbf{B}_i \mathbf{\Theta}_{(t)} \right)^{-1} \mathbf{\Theta}_{(t)}^\top \mathbf{B}_i^\top \right] \quad (\text{A.2.9})$$

This distribution has been obtained by integrating out the individual random effects

$$p(\mathbf{Y}_i | \mathbf{z}_i, \nu_i, \vec{\boldsymbol{\mu}}, \vec{\boldsymbol{\Gamma}}, \boldsymbol{\Lambda}) = \int_{\gamma_i} p(\mathbf{Y}_i, \gamma_i | \mathbf{z}_i, \nu_i, \vec{\boldsymbol{\mu}}, \vec{\boldsymbol{\Gamma}}, \boldsymbol{\Lambda}) d\gamma_i,$$

where

$$\begin{aligned} p(\mathbf{Y}_i, \gamma_i | \mathbf{z}_i, \nu_i, \vec{\boldsymbol{\mu}}, \vec{\boldsymbol{\Gamma}}, \boldsymbol{\Lambda}) &= p(\mathbf{Y}_i | \gamma_i, \mathbf{z}_i, \nu_i, \vec{\boldsymbol{\mu}}, \vec{\boldsymbol{\Gamma}}, \boldsymbol{\Lambda}) p(\gamma_i | \mathbf{z}_i, \nu_i, \vec{\boldsymbol{\mu}}, \vec{\boldsymbol{\Gamma}}, \boldsymbol{\Lambda}) \\ &= \mathcal{N}_{n_i}(\mathbf{B}_i \boldsymbol{\theta}_\mu + \mathbf{B}_i \boldsymbol{\Theta} \boldsymbol{\mu}_{\mathbf{z}_i} + \mathbf{B}_i \boldsymbol{\Theta} \gamma_i^{\mathbf{z}_i}; \sigma^2 \nu_i I_{n_i}) \times \mathcal{N}_k(\mathbf{0}; \boldsymbol{\Gamma}_{\mathbf{z}_i}). \end{aligned}$$

Let $P_{ig}^{(t)}$ be the probability that the individual i belongs to group g . We have from Equation (A.2.6),

$$\begin{aligned} P_{ig}^{(t)} &= \pi_g^{(t)} F_{ig}^{(t)} / \sum_{h=1}^G (\pi_h^{(t)} F_{ih}^{(t)}) \quad \text{with} \\ F_{ig}^{(t)} &= \exp \left\{ -\frac{1}{2} (\mathbf{Y}_i - \mathbf{E}_{ig}^{(t)})^\top \boldsymbol{\Sigma}_{ig}^{-1(t)} (\mathbf{Y}_i - \mathbf{E}_{ig}^{(t)}) \right\} / \left\{ (2\pi)^{n_i/2} |\boldsymbol{\Sigma}_{ig}^{(t)}|^{1/2} \right\} \quad g = 1, \dots, G. \end{aligned}$$

Therefore

$$\begin{aligned} &E_{\mathbf{z}_i | \mathbf{Y}_i, \nu_i^{(t)}, \vec{\boldsymbol{\mu}}^{(t)}, \vec{\boldsymbol{\Gamma}}^{(t)}, \boldsymbol{\Lambda}^{(t)}} \left\{ m_i(\mathbf{z}_i, \vec{\boldsymbol{\mu}}, \vec{\boldsymbol{\Gamma}}, \boldsymbol{\Lambda}) \right\} \\ &= \sum_{g=1}^G p(\mathbf{z}_i = \mathbf{e}_g | \mathbf{Y}_i, \nu_i^{(t)}, \vec{\boldsymbol{\mu}}^{(t)}, \vec{\boldsymbol{\Gamma}}^{(t)}, \boldsymbol{\Lambda}^{(t)}) \times m_i(\mathbf{z}_i = \mathbf{e}_g, \vec{\boldsymbol{\mu}}^{(t)}, \vec{\boldsymbol{\Gamma}}^{(t)}, \boldsymbol{\Lambda}^{(t)}) \\ &= \sum_{g=1}^G P_{ig}^{(t)} \times m_i(\mathbf{z}_i = \mathbf{e}_g, \vec{\boldsymbol{\mu}}^{(t)}, \vec{\boldsymbol{\Gamma}}^{(t)}, \boldsymbol{\Lambda}^{(t)}). \end{aligned}$$

Thus, from Equation (A.2.2), we have :

$$\begin{aligned} Q(\boldsymbol{\Pi} | \boldsymbol{\Pi}^{(t)}) &= \sum_{i=1}^N E_{\mathbf{z}_i | \mathbf{Y}_i, \nu_i^{(t)}, \vec{\boldsymbol{\mu}}^{(t)}, \vec{\boldsymbol{\Gamma}}^{(t)}, \boldsymbol{\Lambda}^{(t)}} \left\{ m_i(\mathbf{z}_i, \vec{\boldsymbol{\mu}}, \vec{\boldsymbol{\Gamma}}, \boldsymbol{\Lambda}) \right\} + \mathcal{H} \\ &= \sum_{i=1}^N \sum_{g=1}^G P_{ig}^{(t)} \times m_i(\mathbf{z}_i = \mathbf{e}_g, \vec{\boldsymbol{\mu}}^{(t)}, \vec{\boldsymbol{\Gamma}}^{(t)}, \boldsymbol{\Lambda}^{(t)}) + \mathcal{H}. \end{aligned}$$

Finally, after the expectation step, the expression of the function $Q(\boldsymbol{\Pi} | \boldsymbol{\Pi}^{(t)})$ where all the parameters are at their t^{th} updated value is given by :

$$\begin{aligned} Q(\boldsymbol{\Pi} | \boldsymbol{\Pi}^{(t)}) &= \sum_{i=1}^N \sum_{g=1}^G P_{ig} \times \left\{ \begin{aligned} &-\frac{n_i}{2} \log(\nu_i \sigma^2) - \frac{1}{2} \log(|\boldsymbol{\Gamma}_g|) + \log(\pi_g) \\ &-\frac{1}{2\nu_i \sigma^2} \left\{ \left\| \mathbf{Y}_i - \mathbf{B}_i \boldsymbol{\theta}_\mu - \mathbf{B}_i \boldsymbol{\Theta} \boldsymbol{\mu}_g - \mathbf{B}_i \boldsymbol{\Theta} \hat{\boldsymbol{\gamma}}_i^g \right\|^2 \right\} \\ &+\frac{1}{2\nu_i \sigma^2} \left\{ \text{trace} \left[\mathbf{B}_i \boldsymbol{\Theta} \hat{\mathbf{V}}_i^g \boldsymbol{\Theta}^\top \mathbf{B}_i^\top \right] \right\} \\ &-\frac{1}{2} \left\{ \hat{\boldsymbol{\gamma}}_{ig}^\top \boldsymbol{\Gamma}_g^{-1} \hat{\boldsymbol{\gamma}}_{ig} + \text{trace} \left[\boldsymbol{\Gamma}_g^{-1} \hat{\mathbf{V}}_i^g \right] \right\} \\ &+\frac{\nu_o}{2} \log\left(\frac{\nu_o}{2}\right) - \log[\Gamma\left(\frac{\nu_o}{2}\right)] - \left(1 + \frac{\nu_o}{2}\right) \log(\nu_i) - \frac{\nu_o}{2\nu_i} \end{aligned} \right\} \\ &+ \sum_{g=1}^G \left\{ \begin{aligned} &-\frac{1}{2} \log(|\boldsymbol{\Gamma}_\mu|) - \frac{1}{2} \boldsymbol{\mu}_g^\top \boldsymbol{\Gamma}_\mu^{-1} \boldsymbol{\mu}_g + \frac{m}{2} \log(|(m-k-1)\mathbf{D}|) \\ &-\frac{(m+k+1)}{2} \log(|\boldsymbol{\Gamma}_g|) - \frac{(m-k-1)}{2} \text{trace}[\mathbf{D}\boldsymbol{\Gamma}_g^{-1}] \end{aligned} \right\} \end{aligned}$$

$$\begin{aligned}
& + \frac{km}{2} \log(m-k-1) - \frac{(m+k+1)}{2} \log(|\mathbf{\Gamma}_\mu|) - \frac{(m-k-1)}{2} \text{trace}[\mathbf{\Gamma}_\mu^{-1}] \\
& + \sum_{j=1}^k \left\{ + \frac{m}{2} \log\left(\frac{m}{2}\right) - \log[\Gamma\left(\frac{m}{2}\right)] - \left(1 + \frac{m}{2}\right) \log(d_{jj}) - \frac{m}{2d_{jj}} \right\} \\
& + \alpha_\sigma \log(\beta_\sigma) - \log[\Gamma(\alpha_\sigma)] - (\alpha_\sigma + 1) \log(\sigma^2) - \frac{\beta_\sigma}{\sigma^2} \\
& + -\log[B(a_1, \dots, a_G)] + \sum_{g=1}^G (a_g - 1) \log(\pi_g) + \mathcal{C} \tag{A.2.10}
\end{aligned}$$

A.3. THE UPDATING EM EQUATIONS FOR THE MIXED-EFFECTS MODEL FOR PRRSV

In the models M_1 , M_2 and M_{12} the additional parameters $\boldsymbol{\theta}_\mu$; \mathbf{S}_1 , \mathbf{S}_2 ; $\mathbf{T}_1, \mathbf{T}_2, \dots, \mathbf{T}_7$ are to be estimated. There's no substantial change in the EM steps except in the equations involving $\boldsymbol{\theta}_\mu$. The parameter estimation calculations associated with the model M_{12} lead to the following equations for the two effects model M_{12} . In these equations, $\mathbf{S}_l = \mathbf{S}_{w_i|w_i=l}$ ($l = 1, 2$) and $\mathbf{T}_h = \mathbf{T}_{e_i|e_i=h}$ ($h = 1, 2, \dots, 7$). Let N_l^* be the subset of individuals with $\{w_i = l\}$ and N_h^* be the subset of individuals with $\{e_i = h\}$, and define

$$\begin{cases} \mathcal{A}_1 = \left[\sum_{i=1}^N (\mathbf{B}_i^\top \mathbf{B}_i) \nu_i^{-1(t)} \right]^{-1} \\ \mathcal{A}_2 = \left[\sum_{i=1}^N \sum_{g=1}^G \frac{P_{ig}}{\nu_i^t} \mathbf{B}_i^\top \left(\mathbf{Y}_i - \mathbf{B}_i \mathbf{S}_{w_i}^{(t)} - \mathbf{B}_i \mathbf{T}_{e_i}^{(t)} - \mathbf{B}_i \boldsymbol{\Theta}_{(t)} (\boldsymbol{\mu}_g^{(t)} + \hat{\gamma}_i^g) \right) \right], \end{cases}$$

$$\begin{cases} \mathcal{A}_3 = \left[\sum_{i \in N_l^*} (\mathbf{B}_i^\top \mathbf{B}_i) \nu_i^{-1(t)} \right]^{-1} \\ \mathcal{A}_4 = \left[\sum_{i \in N_l^*} \sum_{g=1}^G \frac{P_{ig}}{\nu_i^t} \mathbf{B}_i^\top \left(\mathbf{Y}_i - \mathbf{B}_i \boldsymbol{\theta}_\mu^{gen(t)} - \mathbf{B}_i \mathbf{T}_{e_i}^{(t)} - \mathbf{B}_i \boldsymbol{\Theta}_{(t)} (\boldsymbol{\mu}_g^{(t)} + \hat{\gamma}_i^g) \right) \right], \end{cases}$$

$$\begin{cases} \mathcal{A}_5 = \left[\sum_{i \in N_h^*} (\mathbf{B}_i^\top \mathbf{B}_i) \nu_i^{-1(t)} \right]^{-1} \\ \mathcal{A}_6 = \left[\sum_{i \in N_h^*} \sum_{g=1}^G \frac{P_{ig}}{\nu_i^t} \mathbf{B}_i^\top \left(\mathbf{Y}_i - \mathbf{B}_i \boldsymbol{\theta}_\mu^{gen(t)} - \mathbf{B}_i \mathbf{S}_{w_i}^{(t)} - \mathbf{B}_i \boldsymbol{\Theta}_{(t)} (\boldsymbol{\mu}_g^{(t)} + \hat{\gamma}_i^g) \right) \right]. \end{cases}$$

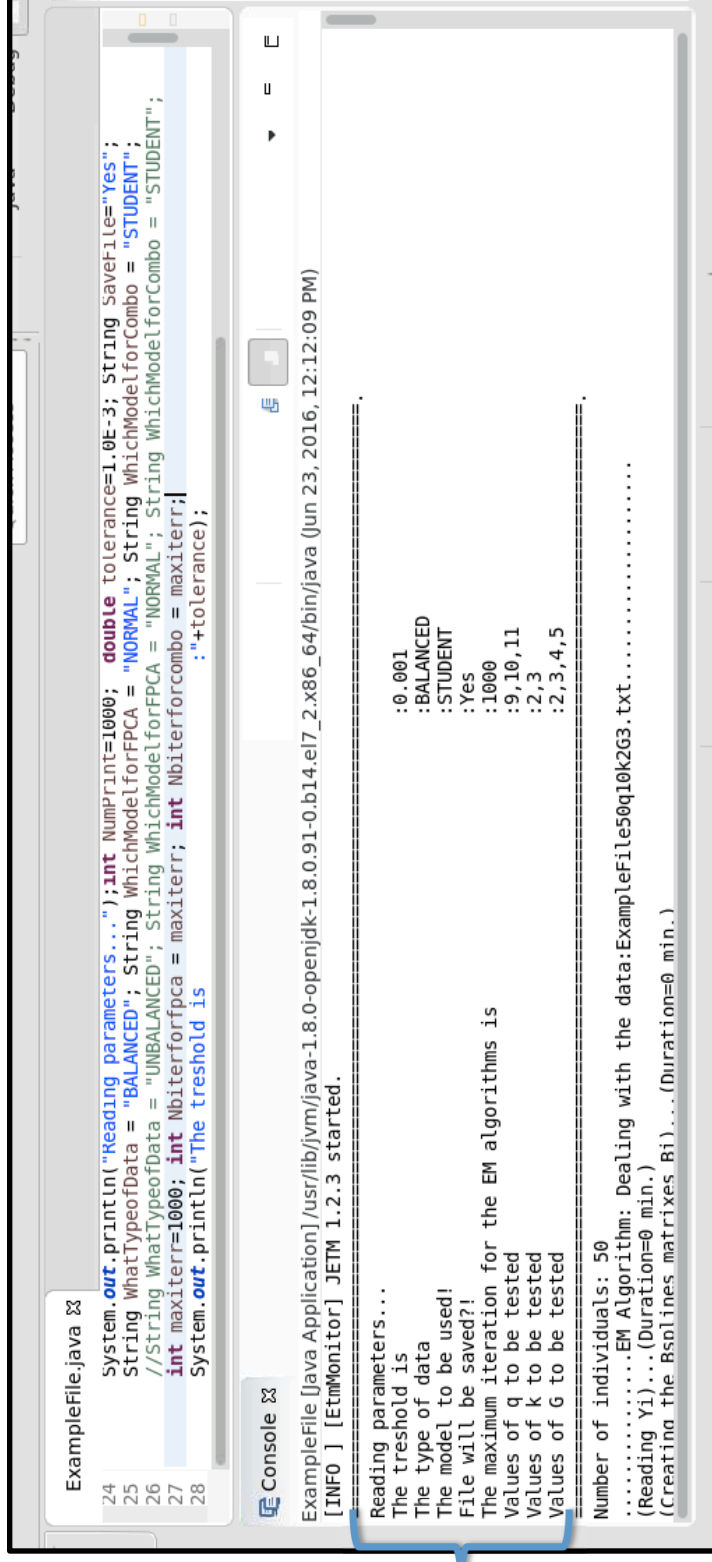
Then

$$\boldsymbol{\theta}_\mu^{gen(t)} = \mathcal{A}_1 \times \mathcal{A}_2, \quad \mathbf{S}_l^{(t)} = \mathcal{A}_3 \times \mathcal{A}_4, \quad \mathbf{T}_h^{(t)} = \mathcal{A}_5 \times \mathcal{A}_6.$$

Annexe B

AN ILLUSTRATION OF THE CODE FOR THE FUNCTIONAL MODEL-BASED CLUSTERING ANALYSIS

2. It is important to know whether the longitudinal data are balanced or unbalanced. One has the choice between Normal errors and Student errors. One also has the choice to save or not the results files by choosing SaveFile=Yes or No.



```
ExampleFile.java 24 System.out.println("Reading parameters...");int NumPriors=1000; double tolerance=1.0E-3; String SaveFile="Yes";
25 String WhatTypeofData = "BALANCED"; String WhichModelForFPCA = "NORMAL"; String WhichModelForCombo = "STUDENT";
26 //String WhatTypeofData = "UNBALANCED"; String WhichModelForFPCA = "NORMAL"; String WhichModelForCombo = "STUDENT";
27 int maxiterr=1000; int Nbiterforfpc = maxiterr; int Nbiterforcombo = maxiterr;
28 System.out.println("The threshold is
: "+tolerance);

Console
ExampleFile [Java Application] /usr/lib/jvm/java-1.8.0-openjdk-1.8.0.91-0.b14.el7_2.x86_64/bin/java (Jun 23, 2016, 12:12:09 PM)
[INFO ] [EtmMonitor] JETM 1.2.3 started.

Reading parameters...
The threshold is
:0.001
The type of data
:BALANCED
The model to be used!
:STUDENT
File will be saved?!
:Yes
The maximum iteration for the EM algorithms is
:1000
Values of q to be tested
:9,10,11
Values of k to be tested
:2,3
Values of G to be tested
:2,3,4,5

Number of individuals: 50
.....EM Algorithm: Dealing with the data:ExampleFile50q10k263.txt.....
(Reading Y1)...(Duration=0 min.)
(Creating the Bsplines matrixes B1)...(Duration=0 min.)
```

3. These characteristics are presented at the beginning of the execution, to recall the parameters defined.


```
ExampleFile.java
1 package A HOW TO RUN EXAMPLE;

Console
ExampleFile [Java Application] /usr/lib/jvm/java-1.8.0-openjdk-1.8.0.91-0.b14.el7_2.x86_64/bin/java (Jun 25, 2016, 12:10:00 AM)
(Values: q=9, Value of k=2, Value of G=5)(Checking values: q= 9, k= 2, G= 5)
The initial ari (just after Fpca) is : 0.803
-----
(Executing Combo Model)...
(At Iteration 0)
Initial value of LEMIN: 5.0
Highest Parameters-RV:6.44783
Likelihood-RV:0.33356
The Likelihood is.. 1501.82936
-----
Successful Execution of Combo: Total number of iterations: 7
At Iteration 7: Highest Parameters-RV:0.53161
At Iteration 7: Likelihood-RV:4.8E-4
The Likelihood is:1664.57489
For (N,q,k,G)=(50,9,2,5): Execution time for Combo 0 min. or 0 sec.
-----
1 2 3 4 5
```

4

4. This is an illustration of the model execution for $q=9$, $k=2$ and $G=5$. The initial ARI (Adjusted Rand Index) is the one calculated just after the model initialization with the *mclust* based on the functional principal scores after FPCA (Functional Principal Components Analysis). Further, the functional model-based analysis is executed.

```

<terminated> ExampleFile [Java Application] /usr/lib/jvm/java-1.8.0-openjdk-1.8.0.91-0.b14.e17_2.x86_64/bin/java (Jun 23, 2016, 12:
.....
No empty group
Writing files...
*****
For (q,k,G)=(11,3,5) we have (Aic,Bic,MLL,Ari,LogL)=(1645.132,1539.015,1346.527,0.722,1756.132 )
*****
=====
Table of AIC values:
=====
0.000      0.000      2.000      3.000      4.000      5.000
9.000      2.000      999999.999  1884.627  1829.500  1829.732
9.000      3.000      999999.999  1872.978  1625.748  1655.932
10.000     2.000      999999.999  2337.301  1964.484  2204.404
10.000     3.000      999999.999  2344.729  999999.999  1884.946
11.000     2.000      999999.999  2218.834  1904.205  2155.903
11.000     3.000      999999.999  2216.740  2017.130  1645.132

Table of BIC values:
=====
0.000      0.000      2.000      3.000      4.000      5.000
9.000      2.000      999999.999  1835.871  1774.052  1767.591
9.000      3.000      999999.999  1799.365  1539.707  1557.463
10.000     2.000      999999.999  2285.676  1906.167  2139.395
10.000     3.000      999999.999  2267.292  999999.999  1782.653
11.000     2.000      999999.999  2164.342  1843.021  2088.026

```

5

5. At the end of the execution for all the values in *Qvect*, *Kvect* and *Gvect* the criteria values are computed for the AIC, BIC, MLL, ARI, the Log-likelihood values and the number of iteration until convergence.

```

Console >> ExampleFile [Java Application] /usr/lib/jvm/java-1.8.0-openjdk-1.8.0.91-0.b14.e17_2.x86_64/bin/java (Jun 23, 2016, 14:
<terminated>
Table of MLL values:
0.000 0.000 0.000 2.000 3.000 4.000 5.000
9.000 2.000 1661.298 999999.999 1571.803 1594.544
9.000 3.000 1632.614 999999.999 1402.801 1435.295
10.000 2.000 2046.973 999999.999 1709.579 1916.724
10.000 3.000 2026.230 999999.999 999999.999 1602.090
11.000 2.000 1934.614 999999.999 1651.219 1866.297
11.000 3.000 1902.932 999999.999 1733.076 1346.527

Table of Adj Rand Index values:
0.000 0.000 0.000 2.000 3.000 4.000 5.000
9.000 2.000 1.000 999999.999 0.803 0.738
9.000 3.000 1.000 999999.999 0.773 0.587
10.000 2.000 1.000 999999.999 0.779 0.722
10.000 3.000 1.000 999999.999 999999.999 0.587
11.000 2.000 1.000 999999.999 0.779 0.738
11.000 3.000 1.000 999999.999 0.773 0.722

Table of Valid G:
0.000 0.000 0.000 2.000 3.000 4.000 5.000
9.000 2.000 3.000 1.000 4.000 4.000 5.000

```

```

Console [Java Application] /usr/lib/jvm/java-1.8.0-openjdk-1.8.0.91-0.b14.el7_2.x86_64/bin/java (Jun 23, 2016)
<terminated> ExampleFile [Java Application] /usr/lib/jvm/java-1.8.0-openjdk-1.8.0.91-0.b14.el7_2.x86_64/bin/java (Jun 23, 2016)
Table of TheLog values:
0.000      0.000      2.000      3.000      4.000      5.000
9.000      2.000      999999.999  1935.627  1887.500  1894.732
9.000      3.000      999999.999  1949.978  1715.748  1758.932
10.000     2.000      999999.999  2391.301  2025.484  2272.404
10.000     3.000      999999.999  2425.729  999999.999  1991.946
11.000     2.000      999999.999  2275.834  1968.205  2226.903
11.000     3.000      999999.999  2301.740  2115.130  1756.132

Table of iterations number:
0.000      0.000      2.000      3.000      4.000      5.000
9.000      2.000      0.000      9.000      1.000      7.000
9.000      3.000      0.000      10.000     627.000   1000.000
10.000     2.000      0.000      8.000      1000.000  8.000
10.000     3.000      0.000      10.000     0.000    1000.000
11.000     2.000      0.000      8.000      1000.000  11.000
11.000     3.000      0.000      10.000     12.000   1000.000

The whole execution time was 10 min.
-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
Measurement Point | # | Average | Min | Max | Total
-----|-----|-----|-----|-----|-----|

```

5

Bibliographie

- [1] Shaikh, M., P. D. McNicholas, and A. F. Desmond (2010). A pseudo-em algorithm for clustering incomplete longitudinal data. *The International Journal of Biostatistics, Issue 1, Article 8. 6.*