Décomposition de la variance dans le modèle de classification de trajectoires de biomarqueurs

Amna Abichou Klich

N°d'ordre NNT : 2019LYSE1199

THESE de DOCTORAT DE L'UNIVERSITE DE LYON
opérée au sein de
l'Université Claude Bernard Lyon 1

Ecole Doctorale E2M2
(Evolution Ecosystème Microbiologie Modélisation)

Spécialité de doctorat : **Biostatistique**

Soutenue publiquement le 17/10/2019, par :
(Amna ABICHOU KLICH)

Décomposition de la variance dans le modèle de classification de trajectoires de biomarqueurs

---

Devant le jury composé de :

Benichou, Jacques, Professeur/Université de Rouen	Rapporteur
Hardouin, Jean-Benoît, Maître de conférence/Université de Nantes	Rapporteur
Dufour, Anne-Béatrice, Maître de conférence/Université Lyon 1	Examinatrice
Chalvet-Monfray, Karine, Professeur/Ecole vétérinaire de Lyon	Examinatrice
Ecochard, René, Professeur/Université Lyon 1	Directeur de thèse
Subtil, Fabien, Maître de conférence/Université Lyon 1	Co-directeur

Résumé

L'analyse de mesures longitudinales –appelées trajectoires– est de plus en plus fréquente en recherche médicale. L'un des intérêts de cette analyse est d'identifier des groupes d'individus ayant des trajectoires similaires. La classification obtenue peut être utilisée pour mieux comprendre l'hétérogénéité des évolutions entre individus. La classification peut être déterminée à partir d'un modèle pour lequel les trajectoires des individus correspondent à la trajectoire du groupe auquel ils sont affectés. L'objectif de la thèse est de développer une extension de ce modèle de classification standard permettant une meilleure prise en compte de la variabilité au sein des groupes, (i) variabilité des valeurs du marqueur (variance résiduelle) et (ii) variabilité des profils d'évolution (variance inter-individuelle).

Deux modèles de classification sont développés : 1) un premier modèle qui prend en compte une variance résiduelle au sein de chaque groupe variable d'un groupe à l'autre, et 2) un deuxième modèle qui prend en compte une variabilité des trajectoires au sein des groupes au lieu de prédire la même trajectoire pour tous les individus d'un même groupe, variabilité qui peut être identique ou variable d'un groupe à l'autre. L'intérêt de ces deux modèles a été montré par des travaux de simulations et par des applications cliniques. Globalement, lorsque le nombre de mesures et de trajectoires est suffisant, ces modèles donnent de meilleures classifications que celles du modèle de classification standard. Par ailleurs, en dehors de plans expérimentaux très contrôlés, les deux sources de variabilité sont inhérentes à la recherche en santé. Ces modèles sont donc très pertinents d'un point de vue clinique.

Mots clés: classification, trajectoires, variance inter-individuelle, variance résiduelle, algorithme CEM, données longitudinales

Hospices Civils de Lyon - Service de Biostatistique-Bioinformatique
Laboratoire de Biométrie et Biologie Evolutive - Équipe Biostatistiques-Santé
162 Avenue Lacassagne
69424 LYON Cedex 03

Variance decomposition in classification models
for biomarker trajectories

Abstract

The analysis of longitudinal measures –called trajectories– is more and more frequent in clinical research. One of the interests of this analysis is to identify groups of individuals with similar trajectories. The obtained classification is used to understand and explore the heterogeneity of trajectories among subjects. The classification can be performed by a model that predicts the same trajectory for all the subjects that are classified in the same group. The objective of this thesis is to develop an extension to the standard classification model that gives greater consideration to the variability within groups, (i) the variability of marker values (residual variance), and (ii) the variability of the individual trajectories inside a group (between-individual variance).

Two classification models were developed: 1) a first model that allows unequal residual variance across groups, and 2) a second model that takes into account a between-individual variance within each group instead of predicting the same trajectory for all subjects in the same group, a variance that can be equal or unequal across groups.

The interest of these two models has been studied by simulations and through clinical applications. Overall, when the number of trajectories and measurements per trajectory is sufficient, these models give better classification compared to the standard classification model. Moreover, except for highly controlled experimental designs, the two sources of variability are inherent to research in health. Therefore, these models are very relevant from a clinical point of view.

Keywords: classification, trajectories, between-individual variance, residual variance, CEM algorithm, longitudinal measures

Remerciements

Mes remerciements vont tout d'abord à Fabien Subtil et René Ecochard, qui m'ont encadré tout au long de cette thèse. Merci de m'avoir donné l'opportunité de travailler sur ce sujet de thèse passionnant et de m'avoir faire confiance durant ces quatre années.

Mes remerciements vont aussi aux rapporteurs de cette thèse, Messieurs Jacques Benichou et Jean-Benoît Hardouin, ainsi qu'à Madame Anne-Béatrice Dufour et Madame Karine Chalvet-Monfray pour avoir acceptés de faire partie du jury. Je remercie également les membres de comité de pilotage : M. Thomas Pommier, M. Vivian Viallon, M. Julien Jacques, et M. Christophe Genolini pour leurs remarques très constructives lors des deux réunions, en fin de première et seconde année.

Un immense merci à tous mes collègues du service de biostatistique des Hospices Civils de Lyon pour leurs conseils et leur amitié, et plus particulièrement à ceux situés à proximité de mon bureau.

Enfin, merci infiniment à mon mari, mes parents, mes deux sœurs et mon frère, pour leur soutien constant. Je remercie également mes amis pour les bons moments passés ensemble.

A mes enfants Zohra et Idriss

Table des matières

Tables des figures

Liste des tableaux

1 Introduction

La thématique de ce travail concerne l'analyse des biomarqueurs mesurés de manière répétée au cours de temps, pour le diagnostic de pathologies ou le pronostic mais également pour évaluer l'efficacité des traitements. Les méthodes statistiques usuelles dédiées à ce type d'analyse font l'hypothèse que la population d'étude est homogène, c'est-à-dire que, à caractéristiques connues comparables, l'évolution peut être décrite par un unique profil moyen avec des variations individuelles centrées autour de ce profil moyen. En pratique, il n'est pas rare qu'une hétérogénéité soit suspectée, et malgré la prise en compte des caractéristiques connues, plusieurs profils moyens peuvent être identifiés. Ainsi, cette hétérogénéité n'est pas expliquée par les variables explicatives connues : ces profils correspondent à des groupes latents ou classes latentes d'individus.

Les modèles à classes latentes, qui étendent la théorie des modèles de régression classique à l'étude des populations hétérogènes, permettent d'identifier des groupes latents en regroupant les profils qui se ressemblent en un petit nombre de groupes, caractérisés chacun par un profil moyen. Les modèles à classes latentes les plus courants ne sont pas purement dédiés à la classification, mais à la modélisation des profils d'évolution en présence d'hétérogénéité. En effet, ils se basent sur une approche de mélange : le profil d'un individu est modélisé par un mélange des profils moyens selon des proportions variant d'un individu à l'autre. A posteriori, les individus peuvent être classés en se basant sur ces proportions. D'autres modèles à classes latentes, appelés modèles de classification, sont dédiés principalement à la classification : le profil d'évolution d'un individu correspond au profil moyen du groupe auquel il est affecté, les profils moyens et la classification étant déterminés simultanément.

L'objectif de la thèse est de développer un modèle de classification de profils d'évolution permettant une meilleure prise en compte de la variabilité au sein des groupes, (i) variabilité des valeurs du marqueur (variabilité intra-individuelle) et (ii) variabilité des profils d'évolution (variabilité inter-individuelle).

La première partie de cette thèse décrit le contexte et les problématiques à la fois cliniques et méthodologiques de ce travail. Un état des lieux est fait concernant les différentes approches dédiées à la modélisation de données longitudinales, en particulier les modèles à

classes latentes, ainsi que les algorithmes d'estimations des paramètres proposés dans le cadre de ces modèles.

La deuxième partie présente un nouveau modèle de classification de profils d'évolution qui prend en compte une variabilité du marqueur biologique (variabilité intra-individuelle) variable d'un groupe à l'autre. Dans cette partie, des travaux de simulation sont réalisés pour évaluer l'impact de cette prise en compte sur la classification. Ce modèle est ensuite appliqué aux données d'un essai clinique qui compare les profils de biomarqueur de la fonction rénale selon la dose d'immunosuppresseur (cyclosporine) administré après une transplantation cardiaque.

La troisième partie présente un deuxième modèle de classification qui prend en compte une variabilité des profils d'évolution au sein des groupes au lieu de prédire la même évolution pour tous les individus d'un même groupe, variabilité qui peut être identique ou variable d'un groupe à l'autre. Dans cette partie, des travaux de simulation sont également réalisés pour évaluer l'impact de cette prise en compte sur la classification. Ce modèle est ensuite appliqué aux données d'une étude de cohorte pour caractériser les profils de biomarqueur d'hCG après un premier curetage pour une môle hydatiforme.

Bien que ce document corresponde à une thèse d'articles, des compléments sont apportés à ceux-ci, afin d'insister sur des détails particuliers, ou de donner un éclairage complémentaire.

2 Partie I : Contexte et problématique

2.1 Contexte et problématique en santé

2.1.1 Les trajectoires d'un biomarqueur

Selon le National Institute of Health (USA), la définition générale d'un biomarqueur qui est actuellement répandue dans la communauté scientifique, est « une caractéristique biologique mesurée de façon objective et évaluée comme un indicateur soit de processus biologiques normaux ou pathologiques, soit de réponses pharmacologiques résultant d'une intervention thérapeutique»[1]. Les biomarqueurs les plus accessibles et les plus connus sont sans doute les paramètres physiologiques, biochimiques ou moléculaires qui peuvent être détectés dans un tissu ou un fluide biologique (ex : sang, urine...).

Pour certaines pathologies, une seule mesure du biomarqueur permet d'effectuer le diagnostic. Par exemple, la présence d'un produit synthétisé spécifiquement en réponse à la présence d'un virus est révélatrice de la maladie. En effet, ce produit ne peut pas être détecté chez une personne que si le virus est présent, quelle que soit la quantité de produit mesurée.

Dans d'autres cas, notamment pour le diagnostic précoce ou le dépistage, le biomarqueur est mesuré de manière répétée au cours du temps et c'est l'évolution de ces valeurs qui peut être le reflet du développement de la maladie. Pour Zolg et Langen (2004)[2], un biomarqueur est une molécule qui indique une altération de l'état physiologique d'un individu en relation avec son état de santé ou de maladie. D'après cette définition, la valeur d'un biomarqueur n'est pas fixe, mais elle change au cours du temps, il s'agit d'un biomarqueur longitudinal, et la suite de ses valeurs successives est communément appelée trajectoire individuelle, ou plus simplement trajectoire.

Étudier les trajectoires de biomarqueurs ne permet pas seulement d'établir le diagnostic ou le pronostic de pathologies, mais également d'évaluer l'efficacité des traitements. Dans le cas du cancer de la prostate, l'augmentation du taux sérique de l'antigène spécifique de la prostate (PSA) est un indicateur de la progression de la maladie après un premier traitement (prostatectomie radicale ou radiothérapie)[3–5]. Le taux de PSA sérique doit

généralement retomber à des taux très faibles. Une augmentation du taux de PSA chez ces patients indique une rechute de la maladie ; la trajectoire du taux de PSA a donc une utilité clinique importante.

Etudier les trajectoires de biomarqueurs peut aussi permettre de créer une typologie des individus, en identifiant des groupes de sujets aux trajectoires similaires. Les groupes obtenus peuvent dans certains cas être reliés au pronostic. Par exemple, chez des patients traités par antirétroviraux (ARV), trois profils ont été établis : (i) observance moyenne en début et fin de suivi, avec une période de mauvaise observance entre deux et trois mois, (ii) bonne observance qui diminue au cours du temps et (iii) bonne observance tout au long du suivi[6]. Le groupe des patients avec une « bonne observance » était associé à un risque de décès plus faible que le groupe de patients avec une « mauvaise observance ».

Dans le cadre de cette thèse, deux exemples d'études sur des trajectoires des biomarqueurs ont été analysés, avec des objectifs différents.

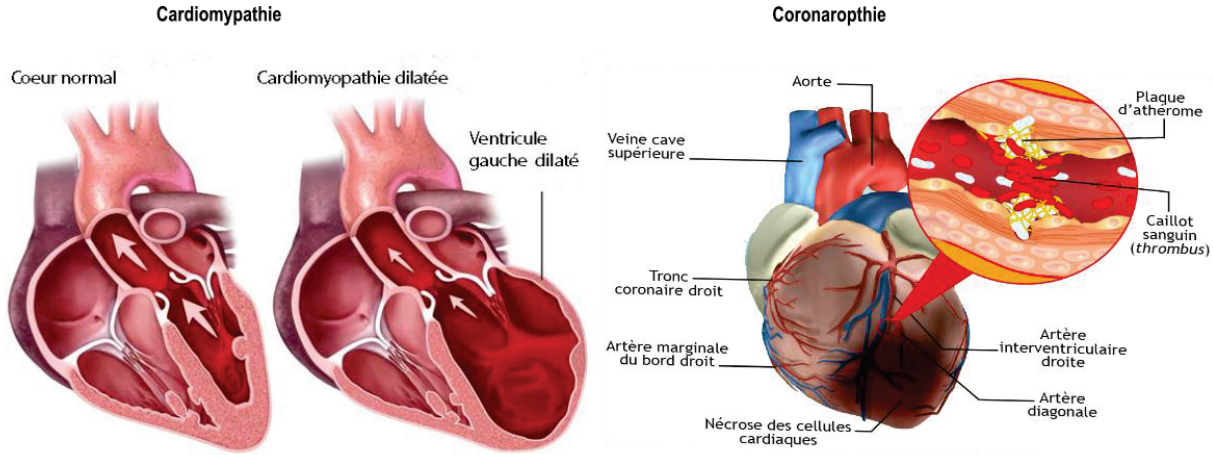
2.1.2 Les trajectoires de la créatinine et la greffe cardiaque

Transplantation cardiaque

La greffe cardiaque est une intervention chirurgicale qui consiste à remplacer le cœur gravement malade ou très endommagé d'une personne par un cœur en bonne santé provenant d'un donneur humain. On procède à la greffe s'il est impossible de traiter une insuffisance cardiaque congestive ou des lésions cardiaques au moyen de médicaments ou d'une autre intervention chirurgicale. La greffe est réservée aux personnes qui présentent un risque élevé de décès au cours de l'année ou des deux années à venir en raison de leur maladie du cœur.

Généralement, la greffe de cœur est envisagée dans deux types de cas (Figure 1). En premier lieu, en cas de lésions cardiaques irréversibles associées à une coronaropathie ayant entraîné des crises cardiaques graves (infarctus du myocarde). En second lieu, en cas de maladie du muscle cardiaque, ou cardiomyopathie, laquelle empêche le cœur de se contracter normalement à cause de lésions aux cellules musculaires. Plus rarement, d'autres formes de maladie du cœur exigent une greffe du cœur, comme des cardiopathies congénitales qui se caractérisent par des problèmes structuraux présents à la naissance.

Figure 1: Insuffisance cardiaque



Les immunosuppresseurs et la néphrotoxicité

L'introduction de la cyclosporine (inhibiteurs de la calcineurine) dans les protocoles immunosuppresseurs a révolutionné le pronostic des greffes cardiaques en réduisant significativement l'incidence des rejets aigus. Cependant son emploi chronique est associé à des effets secondaires, principalement la néphrotoxicité. Plusieurs études rétrospectives[7,8] situent le risque d'insuffisance rénale terminale après transplantation cardiaque entre 5 et 10 % sous un traitement standard de cyclosporine. Cette atteinte rénale apparaît précocement après l'introduction de la cyclosporine et la détérioration initiale de la fonction rénale serait associée à la survenue d'une insuffisance rénale terminale.

La recherche d'associations alternatives d'immunosuppresseurs, visant à réduire l'altération rénale sans altérer la fonction et la survie du greffon, ont été menées en transplantation rénale, hépatique et cardiaque[9–12]. Le schéma thérapeutique le plus fréquent consiste à l'introduction du mycophenolate mofetil (Cellcept®), associé à une diminution progressive parallèle de la cyclosporine. Les résultats de ces évaluations convergent vers un ralentissement du déclin de la fonction rénale, une amélioration des profils métaboliques, et l'absence de majoration significative du risque de rejet.

Les équipes de transplantation cardiaque ont toutes intégré la prescription de MMF dans leur stratégie immunosuppressive, mais ont conservé une posologie classique de cyclosporine. Une nouvelle approche de l'immunosuppression après transplantation cardiaque est basée sur une réduction immédiate et a priori de la cyclosporine sans attendre le moyen terme ou la survenue d'une insuffisance rénale. Cette stratégie de réduction de la dose de

cyclosporine a déjà été formellement évaluée en transplantation rénale, par comparaison à la stratégie de dose classique. La supériorité de cette stratégie a été évaluée dans le cadre de l'étude « Lowcyclo» dont nous utiliserons les données dans le cadre de ce travail.

Essai clinique « Lowcyclo »

L'étude Low-cyclo[13] est un essai multicentrique, prospectif et randomisé évaluant, au cours de la première année après greffe cardiaque, l'impact sur la fonction rénale de deux doses de cyclosporine : faible dose versus dose standard (toutes deux associées au mycophenolate et à des corticoïdes) chez des patients transplantés cardiaques de novo. Cette étude a visé à tester le bénéfice pour le rein, d'utiliser une dose plus faible de cyclosporine, jugée suffisante pour le cœur, mais susceptible d'être moins nocive pour le rein. Le dosage de cyclosporine à utiliser dans le bras « faible dose » a été ainsi défini: réduction d'un tiers du dosage résiduel de cyclosporine par rapport aux dosages standards.

Pour l'évaluation de la fonction rénale après la greffe cardiaque, la créatinine plasmatique (biomarqueur de la fonction rénale) a été mesurée tous les 15 jours durant les 3 premiers mois puis tous les mois jusqu'à 6 mois et tous les 3 mois jusqu'à 12 mois. La première mesure est faite à l'inclusion. 95 patients ont été inclus et randomisés. Un groupe de 47 patients a reçu la posologie standard de cyclosporine et un groupe de 48 patients a reçu la posologie réduite.

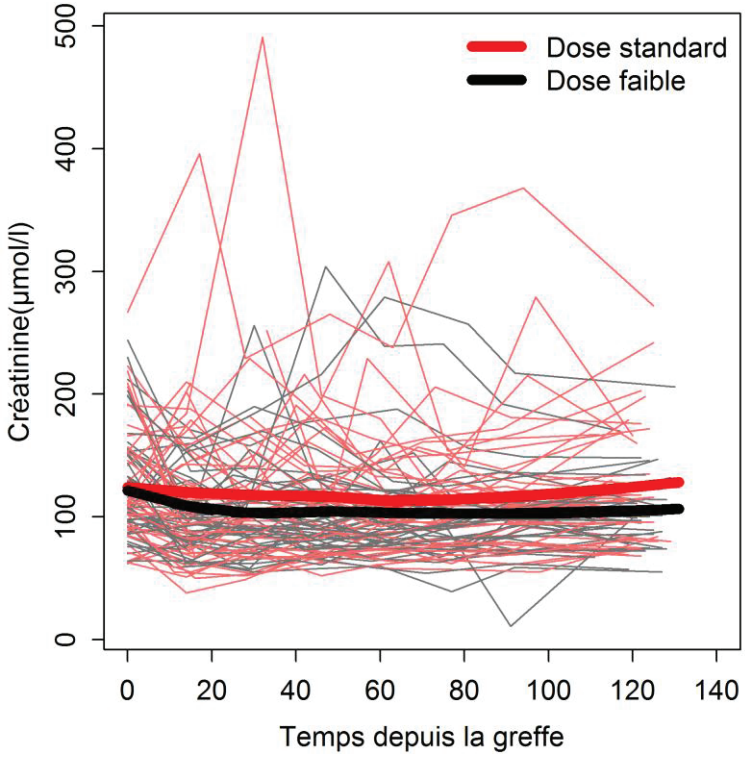
Les caractéristiques des patients de l'étude sont décrites dans le tableau 1.

Tableau 1 : Description des patients à l'inclusion dans les deux groupes de traitement

Caractéristiques	Dose réduite (n=47)	Dose standard (n=48)
Age moyen (sd), année	49.3 (10.2)	48.2 (11.5)
Sexe		
Hommes n (%)	36 (76.6%)	29 (60.4%)
Femmes n (%)	11 (23.4%)	19 (39.6%)
Poids moyen (sd), kg	72 (14)	69 (16)
Créatinine moyenne (sd), µmol/L	124.7 (44.2)	126.5 (47.8)
Type d'étiologie		
Cardiomyopathie idiopathique, n (%)	18 (38.3%)	18 (37.5%)
Coronaropathie, n (%)	16 (34%)	18 (37.5%)
Cardiomyopathie congénitale, n (%)	2 (4.3%)	3 (6.3%)
Pathologie valvulaire, n (%)	1 (2.1%)	3 (6.3%)
Autre, n (%)	10 (21.3%)	6 (12.5%)

Les trajectoires individuelles de la créatinine en fonction de temps après la greffe, ainsi que les trajectoires moyennes selon le bras de traitement (lissage loess) sont représentées sur la Figure 2. Le bénéfice de la réduction de la dose de cyclosporine sur l'évolution de la fonction rénale semble très modeste. Boissonnat et al[13] ont montré que ce bénéfice n'est pas statistiquement significatif, en se basant sur un modèle mixte classique.

Figure 2 : Trajectoires individuelles de créatinine par groupe de dose de cyclosporine.



Indépendamment de la dose, on constate une très grande variabilité d'évolution de créatinine entre les patients. Certaines trajectoires ont un niveau initial normal (100 µmol/L) puis restent stables, d'autres trajectoires ont un niveau initial plus élevé et diminuent par la suite, enfin quelques trajectoires augmentent dès la transplantation. Cette grande variabilité pourrait masquer l'effet du traitement, d'où l'intérêt de méthodes de modélisation de données longitudinales très hétérogènes, supposant l'existence de plusieurs profils moyens d'évolution.

2.1.3 Les trajectoires d'hCG et la môle hydatiforme

Maladies gestationnelles trophoblastiques

Les maladies gestationnelles trophoblastiques constituent un ensemble hétérogène de pathologies rares et mal connues de la grossesse. Parmi ces pathologies du placenta survenant après une fécondation normale ou anormale et caractérisées par un marqueur tumoral, l'hCG (human Chorionic Gonadotrophin), plusieurs groupes distincts ont été décrits sur la base de l'analyse histologique, cytogénétique et clinique : la môle hydatiforme complète ou partielle, la môle invasive, le choriocarcinome, et la tumeur de site d'implantation. Les môles

hydatiformes sont les formes bénignes de ces pathologies. Elles se présentent sous la forme d'une prolifération anormale du trophoblaste en début de grossesse, la môle pouvant être complète ou partielle selon que l'embryon est absent ou présent.

Le diagnostic de môle hydatiforme est en général suspecté sur une échographie en début de grossesse qui objective des anomalies assez caractéristiques du placenta. Le curetage permet l'analyse histologique du placenta et le diagnostic de môle, et plus précisément du type de môle complète ou partielle.

Environ 10 % à 20 % des môles complètes et 0,5 % des môles partielles évolueraient vers une maladie gestationnelle trophoblastique persistante se comportant comme une tumeur maligne[14]. Tous les cas de môles doivent donc avoir une surveillance systématique et régulière basée sur un test hautement spécifique : les hCG. L'évolution anormale des hCG doit être dépistée au plus vite car le retard au diagnostic est un élément déterminant du pronostic[15]. La concentration en hCG diminue après curetage jusqu'à son élimination totale. Cependant, chez certaines femmes, le tissu môlaire résiduel peut à nouveau proliférer et évoluer vers une forme maligne, ce qui se traduit par une remontée des taux sanguins d'hCG.

Le Centre de Référence des Maladies trophoblastiques, créé en 1999, a pour objectif d'optimiser la prise en charge des patientes atteintes de môle hydatiforme ou de tumeur trophoblastique. Il s'est doté d'un registre renseignant les cas de môles hydatiformes en France, avec les données relatives à ces patientes au diagnostic et lors de leur suivi. Ce registre a permis la réalisation de plusieurs études, dont l'une sur la caractérisation des évolutions d'hCG après un premier curetage.

Etude de cohorte du Centre de Référence des Tumeurs Trophoblastiques

L'étude prospective a été menée chez des patientes enregistrées du 1er janvier 2010 au 31 décembre 2012 au Centre de Référence des Tumeurs Trophoblastiques, et qui ont eu un curetage pour une môle hydatiforme. Après le curetage, les patientes étaient suivies avec des mesures hebdomadaires d'hCG jusqu'à des taux indétectables, puis toutes les 2 ou 3 semaines pour les môles partielles, ou tous les mois pendant 6 mois pour les môles complètes.

Sur 1629 patientes de la cohorte n'ont été retenues que les 1440 patientes, ayant au moins deux mesures, en restreignant le suivi aux dosages d'hCG effectués avant le 2nd

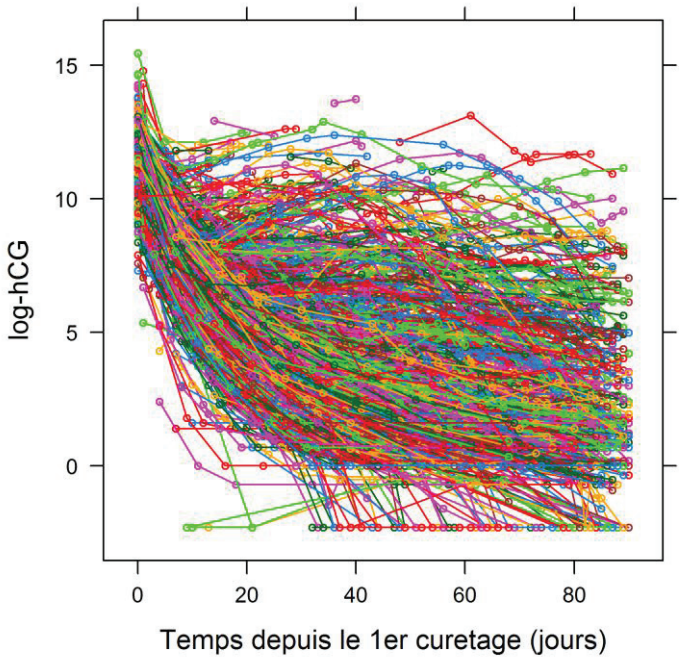
curetage pour les patientes ayant eu plus d'un curetage. Les caractéristiques des patientes de l'étude sont décrites dans le Tableau 2.

Tableau 2 : Description des caractéristiques des patientes à l'inclusion

Caractéristiques	Total (n=1440)
Age moyen (sd), année	31 (7.95)
Nombre d'enfants, médiane (Q1-Q3)	1 (0-2)
Nombre des grossesses avant la môle, médiane (Q1-Q3)	1 (0-2)
hCG médiane (Q1-Q3), mUI/mL	4195.50 (347.75 -59551.50)

Les trajectoires individuelles de logarithme d'hCG (Figure 3) montrent une très grande variabilité d'évolution après le curetage.

Figure 3 : Trajectoires individuelles de log-hCG après un curetage de môle hydatiforme



L'objectif de l'étude est d'identifier des groupes d'évolution d'hCG ayant un sens biologique, afin de mieux caractériser la diversité des profils après un premier curetage.

2.2 Contexte et problématique méthodologiques

Cette partie présente un état des connaissances des modèles statistiques développées dans la modélisation des trajectoires et la prise en compte des trajectoires hétérogènes. Le modèle à effets mixtes qui est le plus utilisé pour analyser les données longitudinales est tout d'abord présenté.

2.2.1 Modèle linéaire à effets mixtes

Le modèle linéaire à effets mixtes[16] est une extension de modèle de régression linéaire standard permettant, dans le cadre de l'analyse de trajectoires, d'appréhender et d'évaluer la diversité des trajectoires au cours du temps, tout en tenant compte de la corrélation entre les mesures d'un même individu.

Modèle

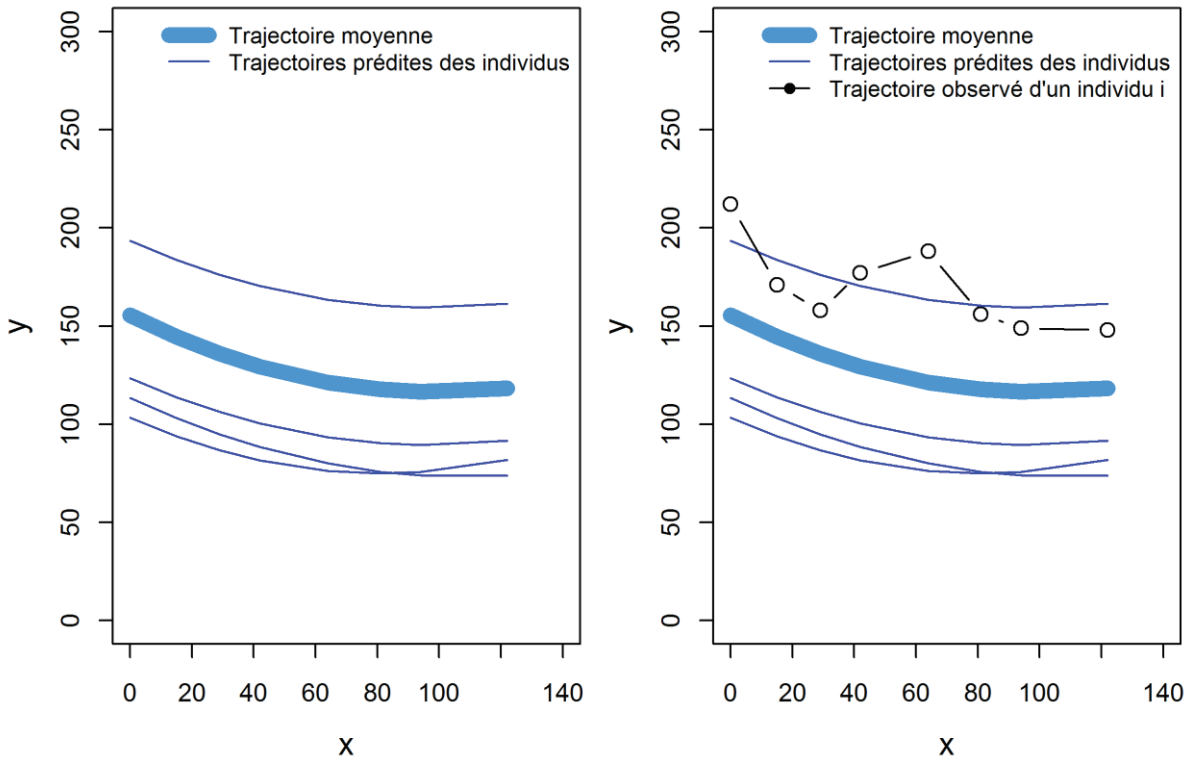
Considérons le vecteur $\mathbf{y}_i = (y_{i1},...,y_{iT_i})$ un vecteur de réponse de T_i mesures successives pour un individu i, avec $i = 1,..,N$, y_{it} correspond à la $t^{\text{ème}}$ mesure de l'individu i. Le modèle linéaire à effets mixtes s'écrit de la façon suivante :

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i \qquad (1.1)$$

$\mathbf{X}_i$ est la matrice de taille $T_i \times p$ correspondant aux p covariables pour lesquels un effet fixe $\boldsymbol{\beta}$ est considéré (e.g. le temps, des caractéristiques des individus et d'éventuelles interactions entre le temps et ces caractéristiques). $\mathbf{Z}_i$ est la matrice de taille $T_i \times q$ associée aux q paramètres à effets aléatoires $\mathbf{b}_i$ pour l'individu i, et $\boldsymbol{\varepsilon}_i$ correspond au vecteur des résidus (écart entre l'observé et le prédit). Les effets aléatoires $\mathbf{b}_i$ sont supposés distribués suivant une loi normale multivariée de moyenne 0 et de matrice variance-covariance **D.** Les résidus sont supposés indépendants et suivant identiquement une loi normale multivariée de moyenne **0** et de matrice de variance-covariance résiduelle $\sigma^2\mathbf{I}_{T_i}$, c'est-à-dire avec une variance résiduelle identique quelle que soit la mesure. Des extensions de ce modèle existent pour le cas où les résidus n'ont pas une variance constante, ou lorsque ceux-ci ne sont pas indépendants ; auquel cas une matrice de variance-covariance résiduelle générale $\boldsymbol{\Sigma}$ est définie.

Dans ce modèle, la variance totale $\left(Tr\left(\mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^{'}+\sigma^2\mathbf{I}_{T_i}\right)\right)$ se décompose en deux composantes : (i) la variance inter-individuelle $\left(Tr\left(\mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^{'}\right)\right)$ qui correspond à la variabilité entre les trajectoires prédites des individus, caractérisée par **D**, et (ii) la variance intra-individuelle ou résiduelle qui correspond aux écarts entre les trajectoires prédites par individu et leurs trajectoires observées, caractérisée par $\sigma^2\mathbf{I}_{T_i}$. Ces deux composantes de la variance sont illustrées dans la Figure 4.

Figure 4 : Variance inter-individuelle et variance intra-individuelle



Le modèle linéaire mixte suppose donc que le vecteur des réponses pour un individu *i* est une variable gaussienne multivariée $\mathbf{y}_i \sim N(\mathbf{X}_i\boldsymbol{\beta}\ ,\mathbf{V}_i=\mathbf{Z}_i\mathbf{D}\ \mathbf{Z}_i^t+\sigma^2\mathbf{I}_{n_i})$ avec une matrice de variance-covariance dont la structure est déterminée par les effets aléatoires et la structure des résidus.

Estimation des paramètres

Les paramètres du modèle linéaire mixte sont estimés soit par maximisation de la vraisemblance, soit par maximisation de la vraisemblance restreinte. La log-vraisemblance du modèle s'écrit :

$$L(\boldsymbol{\beta},\sigma,\mathbf{D}) = -\frac{1}{2}\sum_{i=1}^{N}\left\{T_i \log(2\pi) + \log\left|\mathbf{V}_i\right| - \left(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}\right)^t \mathbf{V}_i^{-1}\left(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}\right)\right\} \qquad (1.2)$$

Etant donné la complexité de la maximisation de cette vraisemblance selon $(\boldsymbol{\beta},\sigma,\mathbf{D})$, une estimation par étapes successives est proposée: les paramètres $\boldsymbol{\beta}$ sont estimés conditionnellement à la matrice $\mathbf{V}_i$ par :

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{N}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\right)^{-1}\left(\sum_{i=1}^{N}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\right) \qquad (1.3)$$

Puis, $(\sigma,\mathbf{D})$ sont estimés en remplaçant $\boldsymbol{\beta}$ dans (1.2) par (1.3) et en maximisant la fonction ainsi obtenue – dite vraisemblance profilée[17] – selon $(\sigma,\mathbf{D})$ par un algorithme de Newton-Raphson ou Quasi-Newton. Ces étapes sont répétées jusqu'à convergence. Dans le cadre des modèles à effets mixtes, la méthode du maximum de vraisemblance conduit avec cet algorithme à des estimations biaisées de $\mathbf{V}_i$; la maximisation de la vraisemblance restreinte permet dans ce cas de corriger ce biais[18,19].

Plusieurs procédures d'estimation sont disponibles dans un grand nombre de logiciels de statistiques. Parmi les plus connues, signalons la fonction lme du package nlme sous R, et la procédure MIXMED sous SAS.

Limites du modèle mixte

Le modèle mixte fait l'hypothèse que la population est homogène, c'est-à-dire qu'elle peut être décrite par un unique profil moyen d'évolution – fonction des caractéristiques connues des individus $\mathbf{X}_i$ – avec des écarts individuels gaussiens centrés autour de ce profil moyen. En santé, une hétérogénéité supplémentaire dans la population peut être suspectée, sans que les déterminants de cette hétérogénéité soient forcément connus, et ainsi plusieurs profils d'évolution latents peuvent être identifiés. Les trajectoires sont donc issues de

plusieurs sous-groupes, sans pour autant que ces groupes soient connus. D'où l'intérêt d'un modèle permettant de tenir compte de cette variabilité latente dans les données longitudinales.

2.2.2 Modélisation des données hétérogènes

La modélisation des données hétérogènes, supposant l'existence de plusieurs profils moyens, a fait l'objet de beaucoup de travaux ces dernières décennies. Le principe est de supposer que l'ensemble d'individus proviennent de K groupes latents. Dans ce cadre, deux grands types d'approches ont été proposées : le modèle de mélange et le modèle de classification, qui sont des extensions des modèles de régression standards. Dans cette section, ces deux modèles sont tout à d'abord présentés, ainsi que leurs algorithmes d'estimations des paramètres. Comme le nombre des groupes est considéré comme un paramètre qu'il faut traiter à part, la détermination de ce nombre est ensuite présentée.

2.2.2.1 Les modèles de mélange

Les modèles de mélange sont usuellement employés pour tenir compte d'une variabilité latente dans des données. Plusieurs auteurs[20,21] ont utilisé ces modèles de mélange dans le cadre de données longitudinales.

Modèle de mélange standard

Le modèle de mélange standard considère que l'ensemble des trajectoires peut être modélisée à partir d'un petit nombre de trajectoires, dites trajectoires typiques ou latentes. Chaque trajectoire individuelle est un mélange de ces trajectoires typiques avec des poids variables d'un sujet à l'autre. Les trajectoires typiques sont définies par une fonction paramétrique du temps (polynômes, polynômes quadratiques, polynômes fractionnaires, splines).

Le modèle de mélange décompose la distribution f des données $\mathbf{y}_i$ pour la trajectoire i en K distributions correspondant aux K groupes :

$$f(\mathbf{y}_i, \boldsymbol{\pi}, \boldsymbol{\beta}, \sigma) = \sum_{k=1}^{K} \pi_k f\left(\mathbf{y}_i, \boldsymbol{\beta}_k, \sigma^2\right) \tag{1.4}$$

$f\left(\mathbf{y}_i, \boldsymbol{\beta}_k, \sigma^2\right)$ correspond à la vraisemblance de la trajectoire *i* lorsque celle-ci est considérée relativement à l'appartenance au groupe *k*, $\boldsymbol{\beta}_k$ correspondant aux paramètres de la $k^{\text{ième}}$ trajectoire typique. π_k est une probabilité a priori mesurant l'influence de la trajectoire typique *k* sur l'ensemble des trajectoires observées ($\boldsymbol{\pi} = \left(\pi_1, \ldots, \pi_K\right)$, $\sum_{k=1}^{K} \pi_k = 1$).

La vraisemblance $f\left(\mathbf{y}_i, \boldsymbol{\beta}_k, \sigma^2\right)$ est calculée en considérant que les mesures de chaque trajectoire sont indépendantes conditionnellement à l'appartenance au groupe[20]. Par conséquent, $f\left(\mathbf{y}_i, \boldsymbol{\beta}_k, \sigma^2\right) = \prod_{t=1}^{T_i} f\left(y_{it}, \boldsymbol{\beta}_k, \sigma^2\right)$

Plusieurs méthodes ont été proposées pour estimer les paramètres du modèle de mélange. On trouve la méthode des moments, la méthode de maximum de vraisemblance et des approches bayésiennes. La méthode la plus utilisée aujourd'hui est la méthode de maximum de vraisemblance à l'aide de l'algorithme EM[22].

Algorithme EM

La vraisemblance d'un modèle de mélange est de la forme suivante :

$$L(\boldsymbol{\pi}, \boldsymbol{\beta}, \sigma^2) = \prod_{i=1}^{N} \sum_{k=1}^{K} \pi_k f(\mathbf{y}_i, \boldsymbol{\beta}_k, \sigma^2) \tag{1.5}$$

L'algorithme EM repose sur la maximisation de l'espérance de la vraisemblance complétée L_C, c'est-à-dire la vraisemblance des données en ajoutant dans celle-ci la composante $z_i = (z_{i1}, ..., z_{iK})$ avec $z_{ik} = 1$ si la trajectoire *i* appartient au groupe *k* et 0 sinon. La vraisemblance complétée s'écrit alors comme un produit des vraisemblances dans les différents groupes :

$$L_C(\mathbf{P},\boldsymbol{\pi},\boldsymbol{\beta},\sigma^2)=\prod_{i=1}^{N}\prod_{k=1}^{K}\left(\pi_k f(\mathbf{y}_i,\boldsymbol{\beta}_k,\sigma^2)\right)^{z_{ik}} \quad (1.6)$$

où **P** est la partition de $(\mathbf{y}_1,...,\mathbf{y}_N)$ définie par z_{ik}. En notant $post.prob_{ik}$ la probabilité a posteriori définie par $E(z_{ik}\left|\mathbf{y}_i,\boldsymbol{\beta},\sigma^2\right.)=P(z_{ik}=1\left|\mathbf{y}_i,\boldsymbol{\beta},\sigma^2\right.)$, les étapes de l'algorithme EM sont les suivantes :

- ✓ Initialisation *:* choix arbitraire d'une solution $\left(\boldsymbol{\pi}^{(0)},\boldsymbol{\beta}^{(0)},\sigma^{2(0)}\right)$
- ✓ Répétition jusqu'à la convergence des 2 étapes suivantes :
  - o étape E (estimation) : calcul des probabilités a posteriori des trajectoires dans les différents groupes conditionnellement aux valeurs de paramètres estimées à l'itération *m* :

$$post.prob_{ik}^{(m)}=\frac{\pi_k^{(m)} f(\mathbf{y}_i,\boldsymbol{\beta}_k^{(m)},\sigma^{2\,(m)})}{\sum_{k=1}^{K}\pi_k^{(m)} f(\mathbf{y}_i,\boldsymbol{\beta}_k^{(m)},\sigma^{2\,(m)})} \quad (1.7)$$

  - o étape M (maximisation) : maximisation de l'espérance du logarithme de la vraisemblance complétée selon $\left(\boldsymbol{\pi},\boldsymbol{\beta},\sigma^2\right)$; ceci équivaut à remplacer dans la vraisemblance complétée les valeurs de z_{ik} par les valeurs de $post.prob_{ik}^{(m)}$:

$$LL(\boldsymbol{\pi}^{(m+1)},\boldsymbol{\beta}_k^{(m+1)},\sigma^{2\,(m+1)})=\sum_{i=1}^{N}\sum_{k=1}^{K} post.prob_{ik}^{(m)}\ log\left(\pi_k^{(m+1)} f(\mathbf{y}_i,\boldsymbol{\beta}_k^{(m+1)},\sigma^{2\,(m+1)})\right) \quad (1.8)$$

  - o pour les paramètres π_k, on obtient l'estimation suivante : $\pi_k^{(m+1)}=\frac{1}{N}\sum_{i=1}^{N} post.prob_{ik}^{(m)}$.

Dans ce modèle, la trajectoire prédite pour un individu *i* est la somme des *K* trajectoires typiques pondérée par les probabilités a posteriori (la part de contribution de chaque trajectoire typique à la trajectoire de l'individu) :

$$\hat{\mathbf{y}}_i=\sum_{k=1}^{K} post.prob_{ik}\times\hat{\mathbf{y}}_k \quad (1.9)$$

Les modèles de mélange standards supposent que les mesures réalisées sur les mêmes sujets à des temps différents sont indépendantes conditionnellement au groupe, alors que la

répétition des mesures d'une variable sur un même sujet induit une corrélation intra-sujet. Une extension de ces modèles a été proposée, qui autorise une variabilité inter-individuelle au sein de chaque groupe grâce à des effets aléatoires ; ces effets aléatoires permettent également de prendre en compte la corrélation entre les mesures d'un même individu. Ceci constitue un modèle de mélange à effets mixtes

Le modèle de mélange à effets mixtes

En reprenant les notations précédemment introduites, le modèle de mélange à effets mixtes[23–26] pour la trajectoire $\mathbf{y}_i$ s'écrit :

$$\mathbf{y}_i = \sum_{k=1}^{K} \pi_k \left(\mathbf{X}_i \boldsymbol{\beta}_k \right) + \sum_{k=1}^{K} \pi_k \left(\mathbf{Z}_i \mathbf{b}_{ik} \right) + \boldsymbol{\varepsilon}_i \tag{1.10}$$

Les effets aléatoires $\mathbf{b}_{ik}$ sont supposés distribués selon une loi normale multivariée de moyenne $\mathbf{0}$ et de matrice variance-covariance $\mathbf{D}_k$ **.** Les résidus sont supposés suivre une loi normale multivariée de moyenne $\mathbf{0}$ et de matrice de variance-covariance résiduelle $\sigma^2 \mathbf{I}_{T_i}$. Pour des raisons d'identifiabilité, la matrice de variance-covariance des effets aléatoires $\mathbf{D}_k$ est souvent supposée identique d'un groupe à l'autre.

La méthode du maximum de vraisemblance est souvent utilisée pour estimer les paramètres du modèle de mélange à effets mixtes. Conditionnellement au groupe *k*, la vraisemblance marginale de $\mathbf{y}_i$ par rapport aux effets aléatoires est donnée par $N(\mathbf{X}_i \boldsymbol{\beta}_k, \mathbf{V}_{ik} = \mathbf{Z}_i \mathbf{D}_k \mathbf{Z}_i^t + \sigma^2 \mathbf{I}_{T_i})$. La vraisemblance globale pour la trajectoire *i* est la somme pondérée des vraisemblances conditionnellement à chacun des groupes. Celle-ci est maximisée par un algorithme itératif de type EM ou Newton-Raphson. Des approches bayésiennes sont elles aussi possibles[27,28].

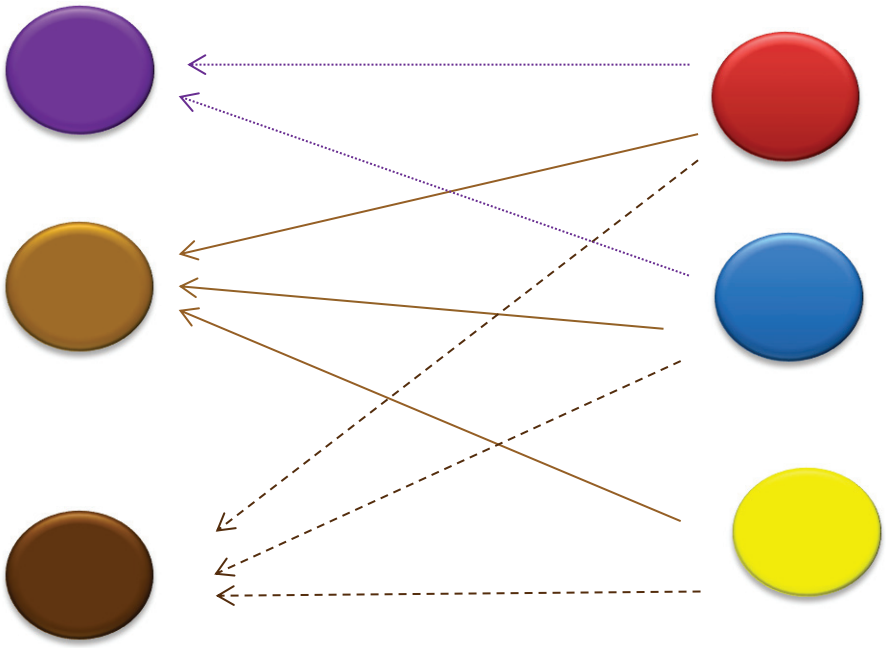
Limites des modèles de mélange

La particularité des modèles de mélange (standard ou à effets mixtes) est qu'ils n'attribuent pas un groupe à chaque trajectoire individuelle, mais utilisent un mélange pondéré de trajectoires latentes communes pour décrire ces trajectoires. Ils ne classent pas les

individus pour déterminer les trajectoires typiques : la partition des individus dans les groupes ne fait pas partie des paramètres du modèle. Chaque individu participe donc, plus ou moins, à l'estimation des paramètres de toutes les trajectoires typiques d'après l'équation (1.8), et chaque trajectoire typique contribue à la trajectoire prédite de chaque individu avec des poids différents selon l'équation (1.9). La signification de ces trajectoires typiques est difficile à établir puisqu'elles ne représentent pas des individus particuliers. Ainsi, les trajectoires typiques ne sont que des construits utilisés pour mieux modéliser les données en tenant compte d'une variabilité inter-individuelle latente.

Les modèles de mélange peuvent être illustrés par l'exemple des couleurs (Figure 5). Les couleurs secondaires ou tertiaires ne peuvent pas être classées dans une des trois couleurs primaires (rouge, bleu, jaune), car elles sont un mélange de ces 3 couleurs, avec des proportions différentes. Par exemple, le couleur « marron clair » est un mélange de 20 % de bleu, 20 % de rouge, et 60 % de jaune ; ainsi, le marron clair n'appartient à aucune des couleurs primaires, même s'il se rapproche plus du jaune que les deux autres couleurs. Le mélange ne sert donc pas à classer mais à reconstituer une diversité.

Figure 5 : Illustration du modèle de mélange



Une fois les trajectoires typiques estimées, il reste néanmoins possible de classer les individus en les affectant au groupe pour lequel la probabilité a posteriori est maximale (MAP, maximum a posteriori).

Une approche alternative aux modèles de mélange est l'utilisation de modèles de classification.

2.2.2.2 Le modèle de classification standard

Contrairement aux modèles de mélange, les modèles de classification consistent à rechercher une partition de l'échantillon de telle sorte que chaque groupe k soit assimilable à un sous-échantillon issu de la loi $f(\mathbf{y}_i, \boldsymbol{\beta}_k, \sigma^2)$. Les groupes ont une signification réelle en termes d'individus car à chaque groupe est associé un ensemble d'individus.

Celeux et Govaert[29] ont défini la notion de vraisemblance classifiante pour des données non longitudinales, par exemple pour classer des individus selon les mesures de différents marqueurs. La vraisemblance classifiante correspond à la vraisemblance complétée (1.6), dont l'espérance est utilisée dans les algorithmes EM pour l'estimation des paramètres d'un modèle de mélange.

Les paramètres sont à la fois les paramètres des trajectoires typiques et la partition des individus dans les groupes. Leur estimation est effectuée en maximisant la vraisemblance classifiante par l'algorithme de « classification » EM, noté CEM[30].

Algorithme CEM

L'algorithme CEM correspond à une version classifiante de l'algorithme EM. Une étape de classification est ajoutée pour affecter les individus à un groupe par la règle du MAP. Les étapes de l'algorithme CEM sont les suivantes :

- ✓ Initialisation *:* choix arbitraire d'une solution $\left(\boldsymbol{\pi}^{(0)}, \boldsymbol{\beta}^{(0)}, \sigma^{2(0)}\right)$
- ✓ Répétition jusqu'à la convergence des 2 étapes suivantes :

- étape E (estimation) : calcul des probabilités a posteriori de la trajectoire *i* dans les différents groupes conditionnellement aux valeurs de paramètres estimées à l'itération *m* :

$$post.prob_{ik}^{(m)} = \frac{\pi_k^{(m)} f(\mathbf{y}_i, \boldsymbol{\beta}_k^{(m)}, \sigma^{2\,(m)})}{\sum_{k=1}^{K} \pi_k^{(m)} f(\mathbf{y}_i, \boldsymbol{\beta}_k^{(m)}, \sigma^{2\,(m)})} \tag{1.11}$$

- étape C : affecter la trajectoire *i* au groupe avec la valeur de $post.prob_{ik}$ la plus élevée (MAP) ; ceci défini les valeurs de $z_{ik}^{(m)}$ ainsi que la partition $\mathbf{P}^{(m)}$ des trajectoires.
- étape M (maximisation) : maximisation du logarithme de la vraisemblance classifiante selon $\left(\boldsymbol{\pi}, \boldsymbol{\beta}, \sigma^2\right)$ à partir de la partition $\mathbf{P}^{(m)}$:

$$LL\left(\boldsymbol{\pi}^{(m+1)}, \boldsymbol{\beta}_k^{(m+1)}, \sigma^{2\,(m+1)}\right) = \sum_{i=1}^{N}\sum_{k=1}^{K} z_{ik}^{(m)} \log\left(\pi_k^{(m+1)} f\left(\mathbf{y}_i, \boldsymbol{\beta}_k^{(m+1)}, \sigma^{2\,(m+1)}\right)\right) \tag{1.12}$$

pour les paramètres π_k, on obtient l'estimation suivante : $\pi_k^{(m+1)} = \frac{1}{N}\sum_{i=1}^{N} z_{ik}^{(m)}$

Dans cette vraisemblance (1.12), on constate que les paramètres d'un groupe donné ne sont estimés qu'à partir des trajectoires qui lui sont affectées, contrairement aux modèles de mélanges. Ainsi, les groupes obtenus ne sont plus des construits, mais des groupes d'individus auxquels il est possible de donner une signification réelle.

Dans ce modèle, la trajectoire prédite d'un individu correspond à la trajectoire typique à laquelle il est affecté, c'est-à-dire $\hat{\mathbf{y}}_i = \hat{\mathbf{y}}_k$.

Pour réduire le risque de converger vers un optimum local lors de la maximisation de la vraisemblance, l'algorithme CEM est répété un grand nombre de fois à partir de valeurs initiales différentes. La solution retenue est celle qui conduit à la vraisemblance classifiante la plus élevée, supposée être le maximum global. Une autre solution consiste à utiliser des versions stochastiques de l'algorithme CEM, notées SEM et CAEM[30,31].

Algorithmes stochastiques de CEM

L'algorithme SEM, « stochastique EM », propose d'intercaler une étape stochastique de classification entre les étapes E et M. Après le calcul des probabilités d'appartenance aux groupes $post.prob_{ik}^{(m)}$, les valeurs de $z_{ik}^{(m)}$ tirées selon une loi multinomiale $\mathcal{M}\left(post.prob_{i1}^{(m)},\ldots,post.prob_{iK}^{(m)}\right)$. Ces tirages aléatoires limitent le risque de converger vers un maximum local instable de la vraisemblance. En revanche, contrairement à l'algorithme CEM, l'algorithme SEM peut ne pas converger au sens strict vers une partition stable. En pratique, Celeux et Diebolt (1985)[32] proposent d'utiliser l'algorithme SEM sur un certain nombre d'itérations, puis d'utiliser l'algorithme CEM pour obtenir une partition et une estimation des paramètres.

L'algorithme CAEM, « classification annealing EM », est basé sur l'algorithme SEM, mais il permet d'obtenir directement les estimations des paramètres du modèle, sans recours à l'algorithme CEM par la suite. Lors de l'étape E, la probabilité a posteriori est calculée de la manière suivante :

$$post.prob_{ik} = \frac{\left(\pi_k^{(m)} f(\mathbf{y}_i, \boldsymbol{\beta}_k^{(m)}, \sigma^{2\,(m)})\right)^{1/\tau_{(m)}}}{\sum_{k=1}^{K}\left(\pi_k^{(m)} f(\mathbf{y}_i, \boldsymbol{\beta}_k^{(m)}, \sigma^{2\,(m)})\right)^{1/\tau_{(m)}}} \qquad (1.13)$$

$\tau_{(m)}, m = 1,\ldots,M$ est une suite de valeurs décroissantes. Si $\tau_{(m)} = 1$, cet algorithme correspond à l'algorithme SEM ; si $\tau_{(m)}$ tend vers 0, il correspond à l'algorithme CEM. La valeur initiale recommandée de τ est $\tau_{(0)} = 1$ et la suite de valeurs est définie par $\tau_{(m+1)} = a\tau_{(m)}$, avec $0.9 \le a \le 1$. Ceci revient, dans un même algorithme, à utiliser un algorithme SEM au début, et finir par un algorithme CEM.

Celeux et Govaert[30] ont comparé ces deux algorithmes stochastiques sur des données simulées non longitudinales à partir d'un modèle de mélange. Ils recommandent d'utiliser l'algorithme CAEM plutôt que l'algorithme SEM pour les données à effectif faible, et l'algorithme SEM plutôt que l'algorithme CAEM pour les données à effectif élevé.

k-means, un cas particulier du modèle de classification

L'algorithme de type k-means[33] vise à déterminer la partition des individus et les trajectoires typiques de sorte à minimiser une fonction de perte :

$$W = \sum_{i=1}^{N} \sum_{k=1}^{K} z_{ik} d\left(\mathbf{y}_i, \mathbf{y}_k\right) \tag{1.14}$$

où d est une métrique de distance (par exemple la distance euclidienne). Cette approche correspond à un cas particulier simple du modèle de classification. En effet, maximiser la vraisemblance classifiante L_C revient à minimiser W pour des données gaussiennes, avec une matrice de variance-covariance résiduelle commune entre les groupes, et avec des probabilités a priori égales.

L'algorithme du k-means est une méthode de classification très géométrique. Elle nécessite que les individus aient des mesures effectuées aux même temps. De plus, il ne permet pas de tenir compte de covariables pouvant influencer les trajectoires individuelles. Par exemple, l'analyse des trajectoires d'un biomarqueur mesurant la réponse à un traitement peut faire apparaître des groupes de réponse : bons répondeurs et mauvais répondeurs. Mais une variable extérieure, telle que l'âge, peut entraîner une baisse du niveau du biomarqueur chez les personnes âgées quel que soit le groupe. Ne pas tenir compte de cette variable dans la classification peut entraîner une mauvaise prise en compte de la variabilité inter-individuelle, et faire apparaître des groupes qui ne sont pas forcément reliés au phénomène d'intérêt (ici la réponse au traitement).

2.2.2.3 Sélection de nombre des groupes

Dans la plupart des cas, le nombre de groupes est inconnu, alors que celui-ci est un paramètre d'entrée des modèles à classe latente. Bien souvent, les résultats obtenus avec des modèles à différents nombres de groupes sont comparés. Le choix final du nombre de groupes est complexe, et repose sur un faisceau d'arguments. Tout d'abord, l'adéquation du modèle aux données, pour laquelle plusieurs critères ont été proposés. Dans le contexte du modèle de mélange, le critère d'information bayésien (BIC) est le plus fréquemment utilisé. Sa formule est donnée par : $BIC = -2L(\pi, \theta_k) + p\log(N)$, où p est le nombre des paramètres estimés. Ce critère est préféré au critère AIC (Akaike Information Criterion), car il pénalise plus les modèles complexes. Dans le contexte d'un modèle de classification, l'ICL (Integrated

Complete Likelihood) et son approximation l'ICL-BIC, qui sont des critères basés sur la vraisemblance classifiante pénalisée, sont les principaux critères utilisés[18,19].

Le choix du nombre de groupes ne repose pas que sur des critères d'adéquation du modèle aux données. On préfère en général un nombre de groupes pour lequel les probabilités a posteriori d'appartenir au groupe dans lesquels les trajectoires ont effectivement été classées sont élevées ; ainsi, la trajectoire typique d'un groupe est définie par des trajectoires individuelles qui appartiennent « véritablement » au groupe. De plus, on évite les classifications avec des groupes vides. Enfin, le choix du modèle doit également avoir une justification clinique et les groupes obtenus un sens biologique.

3 Partie II : Prise en compte d'une variance résiduelle variable d'un groupe à l'autre

L'objectif de cette partie est d'implémenter un modèle de classification qui prend en compte une variance résiduelle au sein de chaque groupe variable d'un groupe à l'autre, et évaluer l'impact de cette prise en compte sur la classification à l'aide de données simulées. Ce modèle est ensuite appliqué aux données de l'essai clinique « Lowcyclo ».

3.1 Variabilité des trajectoires au sein de chaque groupe

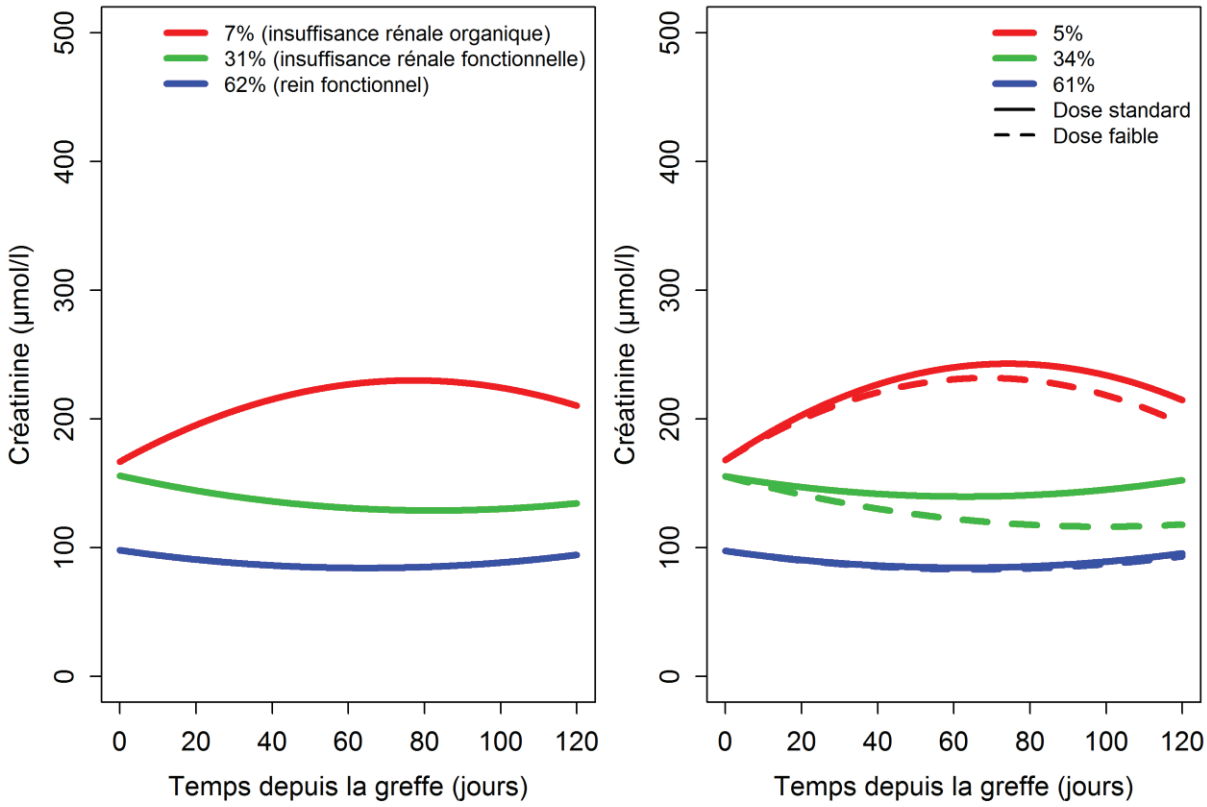
Sur les données de l'étude « Lowcyclo », l'analyse des trajectoires, dans une première analyse, a été effectuée par un modèle de classification de trajectoires standard en considérant trois groupes. Indépendamment de la dose de cyclosporine reçue, les groupes de trajectoires identifiés sont représentés sur la Figure 6A :

- ✓ un groupe de patients dont le niveau de créatinine est normal lors de la transplantation, et reste relativement stable au cours du temps ; ces patients sont considérés comme des patients ayant un rein fonctionnel lors de la transplantation,
- ✓ un groupe de patients dont le niveau de créatinine est élevé lors de la transplantation, mais diminue vers un niveau normal au cours du temps ; ces patients étaient en insuffisance rénale lors de la transplantation, avec une amélioration par la suite ; on peut parler d'insuffisance rénale fonctionnelle ;
- ✓ un groupe de patients dont le niveau de créatinine est élevé lors de la transplantation, et qui reste élevé tout au long du suivi ; ces patients sont considérés comme des patients en insuffisance rénale organique.

Ainsi la très grande variabilité des trajectoires entre les patients est en partie expliquée par l'existence naturelle de ces groupes de trajectoires. Une fois cette variabilité prise en compte – c'est-à-dire à profil typique comparable – le bénéfice de la réduction de dose de cyclosporine devient plus net (Figure 6B) :

- pour les patients ayant un niveau de créatinine initial élevé (insuffisance rénale organique ou fonctionnelle), on constate une diminution progressive de la créatinine dans le groupe de patients ayant reçu une dose faible de cyclosporine par rapport au groupe de patients ayant reçu une dose standard ; le rein de ces patients étant en mauvais état lors de la transplantation, il est supposé ne pas pouvoir supporter une dose forte de cyclosporine.
- chez les patients ayant eu un taux de créatinine normal, il n'y a pas de différence d'évolution de la créatinine au cours du temps entre les deux groupes de dose de cyclosporine : le rein de ces patients étant en bon état lors de la transplantation, il est supposé pouvoir supporter une dose forte de cyclosporine.

Figure 6 : Trajectoires typiques de créatinine globalement et par bras de traitement

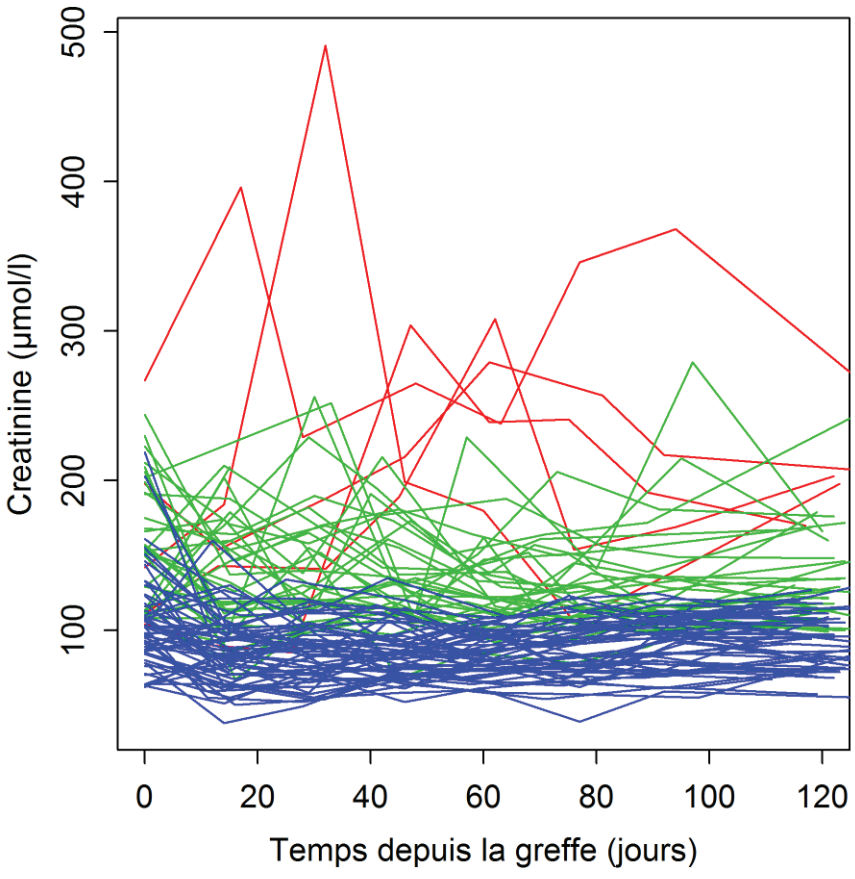


Ces résultats montrent l'intérêt des modèles de classification de trajectoires dans le cadre d'un essai clinique.

Ces analyses reposaient sur un modèle de classification de trajectoires standard, qui suppose que la variabilité des trajectoires au sein des groupes est la même d'un groupe à l'autre. Néanmoins, il semble que la variabilité des trajectoires au sein des groupes d'insuffisance rénale (fonctionnelle ou organique) soit plus importante que celle qui est observée dans le groupe des patients avec un rein fonctionnel (Figure 7). La variabilité des trajectoires au sein des groupes correspond à la combinaison de deux variances : (i) la variance inter-individuelle des trajectoires au sein des groupes et (ii) la variance intra-individuelle – ou variance résiduelle – des trajectoires au sein des groupes.

Dans cette étude, il a été supposé que la variabilité des trajectoires au sein des groupes est le reflet d'une variance résiduelle différente d'un groupe à l'autre. Le premier objectif de la thèse a été de développer et implémenter un modèle de classification de trajectoires prenant en compte une variance résiduelle variable d'un groupe à l'autre, et de l'appliquer aux trajectoires observées de créatinine de l'étude Lowcyclo.

Figure 7 : Classification des trajectoires de créatinine obtenue par un modèle de classification standard



3.2 Le modèle

En reprenant les notations précédemment introduites, si la trajectoire $\mathbf{y}_i$ de l'individu *i* appartient au groupe *k*, le modèle de classification avec une variance résiduelle variable d'un groupe à l'autre s'écrit :

$$\mathbf{y}_i = \mathbf{X}_{1i}\boldsymbol{\gamma} + \mathbf{X}_{2i}\boldsymbol{\beta}_k + \boldsymbol{\varepsilon}_i \tag{2.1}$$

$\mathbf{X}_{1i}$ est la matrice de taille $T_i \times p$ correspondant aux *p* covariables supposées avoir un effet identique d'un groupe à l'autre ($\boldsymbol{\gamma}$), $\mathbf{X}_{2i}$ est la matrice de taille $T_i \times q$ correspondant aux *q* covariables supposées avoir un effet différent d'un groupe à l'autre ($\boldsymbol{\beta}_k$). Les résidus $\boldsymbol{\varepsilon}_i$ sont supposés suivre une loi normale multivariée de moyenne **0** et de matrice de variance-covariance résiduelle $\sigma_k{}^2\mathbf{I}_{T_i}$ (pour un modèle de classification standard, la matrice de variance-covariance aurait été $\sigma^2\mathbf{I}_{T_i}$).

Les méthodes d'estimation des paramètres de ce modèle et algorithmes associés, ainsi que sa comparaison par rapport au modèle de classification standard sont décrites dans l'article publié dans la revue « Statistics in Medicine ».

3.3 Article publié dans la revue Statistics in Medicine

Received: 20 September 2017 | Accepted: 3 July 2018

DOI: 10.1002/sim.7921

RESEARCH ARTICLE

WILEY Statistics in Medicine

Unequal intra-group variance in trajectory classification

Amna Klich[1,2] | René Ecochard[1,2] | Fabien Subtil[1,2]

[1]Service de Biostatistique-Bioinformatique, Pôle Santé Publique, Hospices Civils de Lyon, Lyon, France

[2]Université de Lyon, Université Lyon 1, CNRS, Laboratoire de Biométrie et Biologie Evolutive UMR 5558, Villeurbanne, France

Correspondence
Amna Klich, Service de Biostatistique-Bioinformatique, Hospices Civils de Lyon 162, Avenue Lacassagne - F-69003 Lyon, France.
Email: amna.klich@chu-lyon.fr

Funding information
Agence Nationale de la Recherche, Grant/Award Number: 2012 BLAN SVSE 1

Classifying patients into groups according to longitudinal series of measurements (ie, trajectory classification) has become frequent in clinical research. Most classification models suppose an equal intra-group variance across groups. This assumption is sometimes inappropriate because measurements in diseased subjects are often more heterogeneous than in healthy ones. We developed a new classification model for trajectories that uses unequal intra-group variance across groups and evaluated its impact on classification using simulations and a clinical study. The classification and typical trajectories were estimated using the classification Expectation Maximization (EM) algorithm to maximize the classification likelihood, the log-likelihood being profiled during the Maximization (M) step of the algorithm. The simulations showed that assuming equal intra-group variance resulted in a high misclassification rate (up to 50%) when the real intra-group variances were different. This rate was greatly reduced by allowing intra-group variances to be different. Similar classification was obtained when the real intra-group variances were equal, except when the total sample size and the number of repeated measurements were small. In a randomized trial that compared the effect of low vs standard cyclosporine A dose on creatinine levels after cardiac transplantation, the classification model with unequal intra-group variance led to more meaningful groups than with equal intra-group variance and showed distinct benefits of low dose. In conclusion, we recommend the use of a classification model for trajectories that allows for unequal intra-group variance across groups except when the number of repeated measurements and total sample size are small.

KEYWORDS

classification, ECM algorithm, heterogeneity, intra-group variance, longitudinal measure, trajectories

1 | INTRODUCTION

Unsupervised classification aims to identify homogenous groups without any prior knowledge of these groups. In medicine, unsupervised classification involving a biomarker with repeated measurements (ie, biomarker trajectory) gathers biomarker trajectories into a small number of groups of similar trajectories. The obtained classification allows a better understanding of the heterogeneity of biomarker trajectory among patients.

The trajectory classification is often performed by a mixture model[1-5]; the predicted trajectory of each subject is a mixture of the different typical trajectories weighted by the posterior probabilities of belonging to them. The first objective of

this model is not to classify the trajectories, but to model appropriately each trajectory. Classification models[6-8] present alternatives to the mixture model: each trajectory belongs to one and only one group; thus, the predicted trajectory of a given subject is the typical trajectory of the group to which he/she belongs. These two models (mixture and classification) do not necessarily give the same results.[9,10] Classification models are models where the parameters of the typical trajectories and the classification are jointly estimated. The estimation process relies on the maximization of the classification likelihood. KmL[11] or k-means[12] for longitudinal data are particular classification models; they rely on the minimization of some within-group distance metric between trajectories. Celeux and Govaert[13] demonstrated that, in case of continuous data, that k-means is equivalent to maximizing the classification likelihood, when (i) a multivariate Gaussian distribution with an identity covariance matrix is used for the model and (ii) the proportion of trajectories, called prior probability in the following, is constrained to be equal for each group. However, the k-means algorithm requires an equal number of measurements for all trajectories and it does not allow the inclusion of covariate effects; the classification model does allow this and was therefore considered in the present study.

In the mixture model, most methods assume an equal intra-group variance across groups whereas this assumption is sometimes inappropriate. For example, in diseased subjects, the measurements are more heterogeneous than in healthy ones. Some software programs allow distinct intra-group variance.[14,15] Elsensohn et al[16] developed a graphical diagnostic tool based on drawing an envelope to check the assumption of an equal intra-group variance. However, little work has been done on the impact of considering unequal intra-group variance across groups for both mixture models and classification models. Moreover, in classification models, the current software programs often suppose an equal intra-group variance. Thus, methods for trajectory classification that allow unequal intra-group variance are needed.

This paper presents an implementation of a classification model for trajectories that allows unequal intra-group variance. One feature of this implementation is that it is possible to specify covariates with the same effect on all the groups and other covariates with distinct effects across groups. Simulations were performed to evaluate the impact on the classification of considering unequal intra-group variance across groups vs assuming an equal intra-group variance. The model was then applied to the data of a clinical trial, and the change of classification and typical trajectories according to the assumption is presented.

2 | THE CLASSIFICATION MODEL WITH UNEQUAL INTRA-GROUP VARIANCE

2.1 | The model

Let $\mathrm{y}_i = (y_{i1}, y_{i2}, \ldots, y_{iT_i})$ be the response vector of the T_i successive measurements in individual i, with $i = 1, \ldots, N$, y_{it} corresponding to the tth measurement of individual i. A linear model was used to model the typical trajectory of each group. If the ith individual trajectory belongs to the kth group, the model that assumes an equal intra-group variance across groups (denoted Mod_equal) may be written as

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta}_k + \boldsymbol{\varepsilon}_i, \tag{1}$$

where $\mathbf{X}_i$ is the $T_i \times p$ design matrix of p columns of variables, including a time function (such as a polynomial function of time or a natural cubic spline), characteristics of subjects, and potentially interactions between time and characteristics. $\boldsymbol{\beta}_k$ are the parameters of the model for the kth group. Errors $\boldsymbol{\varepsilon}_i$ are assumed to be normally distributed with mean zero and covariance matrix $\sigma^2\mathbf{I}_{T_i}$ common to all groups, where $\mathbf{I}$ denotes the identity matrix. σ^2 corresponds to the intra-group variance.

Let $z_{ik} = 1$ when the ith trajectory belongs to the kth group (= 0 otherwise) and let $\boldsymbol{\theta}_k$ be the vector $(\boldsymbol{\beta}_k, \sigma^2)$. The likelihood of an individual trajectory i that belongs to the kth group is $f(\mathbf{y}_i, \boldsymbol{\theta}_k)$; in usual classification models, it is calculated by assuming that the measurements of each individual are conditionally independent given the group membership.[8] Thus,

$$f\left(\mathbf{y}_i, \boldsymbol{\theta}_k\right) = \prod_{t=1}^{T_i} f\left(y_{it}, \boldsymbol{\theta}_k\right). \tag{2}$$

Over all individuals, the log classification likelihood is given by

$$L\left(P, \boldsymbol{\pi}, \boldsymbol{\theta}\right) = \sum_{k=1}^{K}\sum_{i=1}^{N} z_{ik} \log\left(\pi_k f\left(\mathbf{y}_i, \boldsymbol{\theta}_k\right)\right), \tag{3}$$

where P is the partition of $(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N)$ defined by z_{ik} and $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ is the vector of prior probabilities belonging to the groups ($0 \leq \pi_k \leq 1$ for all $k = 1, \ldots, K$ with $\sum_k \pi_k = 1$).

Most software programs that use a classification model for classification of trajectories employ the latter model; this supposes an equal intra-group variance across groups. In the present work, we propose a more general formulation in which the intra-group variance may be distinct across groups. In this model (called hereafter Mod_unequal), the intra-group variance of the kth group is denoted σ_k^2. Moreover, some covariates may have the same effect on all groups, and others may have different effects across groups. Hence, the $\mathbf{X}_i$ design matrix is split into two matrices: a matrix $\mathbf{X}_{1i}$ of covariates that have the same effect across groups (effect noted $\boldsymbol{\gamma}$) and a matrix $\mathbf{X}_{2i}$ of covariates that have distinct effects across groups (effects noted $\boldsymbol{\beta}_k$ for the kth group). If the ith individual trajectory belongs to the kth group, the model may then be written as

$$\mathbf{y}_i = \mathbf{X}_{1i}\boldsymbol{\gamma} + \mathbf{X}_{2i}\boldsymbol{\beta}_k + \boldsymbol{\varepsilon}_i. \tag{4}$$

In this model, the parameter vector $\boldsymbol{\theta}_k$ of Equation (3) becomes the vector $(\boldsymbol{\gamma}, (\boldsymbol{\beta}_k)_{k=1,\ldots,K}, (\sigma_k)_{k=1,\ldots,K})$. $\boldsymbol{\theta}_k$ is estimated by maximizing the log classification likelihood $L(P, \boldsymbol{\pi}, \boldsymbol{\theta})$.

2.2 | The CEM algorithm

The "classification" version of the Expectation Maximization (EM)[17] algorithm, called CEM,[13] was used to estimate $\boldsymbol{\pi}$, $\boldsymbol{\theta}$, and the partition P. It incorporates a classification step between the E-step and the M-step of the EM algorithm.

Starting from an initial partition of the trajectories, the mth iteration of the CEM algorithm is defined as follows.

The E- step computes the posterior probability of belonging to group k for each trajectory i

$$post.prob_{ik}^{(m)} = \frac{\pi_k^{(m)} f\left(\mathbf{y}_i, \boldsymbol{\theta}_k^{(m)}\right)}{\sum_{j=1}^{K} \pi_j^{(m)} f\left(\mathbf{y}_i, \boldsymbol{\theta}_j^{(m)}\right)} \propto \pi_k^{(m)} f\left(\mathbf{y}_i, \boldsymbol{\theta}_k^{(m)}\right). \tag{5}$$

In the case of Mod_unequal, the posterior probability is given by

$$post.prob_{ik}^{(m)} \propto \pi_k^{(m)} \prod_{t=1}^{T_i} \frac{1}{\sigma_k^{(m)}\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{y_{it} - \mathbf{X}_{1it}\boldsymbol{\gamma}_k^{(m)} + \mathbf{X}_{2it}\boldsymbol{\beta}_k^{(m)}}{\sigma_k^{(m)}}\right)^2\right). \tag{6}$$

The C- step assigns each trajectory i to the group that provides the maximum $post.prob_{ik}^{(m)}, 1 \leq k \leq K$. Let $P^{(m)}$ denote the resulting partition.

The M-step computes, for each $k = 1, \ldots, K$, the estimate $(\pi_k^{(m+1)}, \boldsymbol{\theta}_k^{(m+1)})$ that maximizes the log classification likelihood given $P_k^{(m)}$.

In the log classification likelihood, π_k is independent from the other parameters and the maximum-likelihood estimate is given explicitly by $\pi_k^{(m+1)} = \frac{\#P_k^{(m)}}{N}$. The estimates of parameters $\boldsymbol{\theta}_k$ were obtained by ordinary least squares in Mod_equal and by maximization of the profiled log-likelihood[18] in Mod_unequal. These estimates were computed using function *gls* in R package *nlme*. The R program is available in the Supporting Information.

The CEM algorithm was stopped when the difference in the log classification likelihood $L(P^{(m)}, \pi_k^{(m)}\theta_k^{(m)})$ after the C-step of two consecutive iterations was less than a given threshold value or when there was no change in the partition. The solution provided by the CEM algorithm depends on the initial partition.[13] Consequently, the CEM algorithm was repeated several times from different initial partitions and the classification that provided the highest value of log classification likelihood was finally retained.

For a model with unequal intra-group variance, the CEM algorithm leads to infinite log classification likelihood when the number of repeated measurements for a trajectory is lower than the number of parameters per group and there is only one trajectory in a group as the intra-group variance is equal to 0. In this case, the CEM algorithm was stopped and the solution obtained was discarded.

2.3 | The choice of the number of groups

The choice of the number of groups in unsupervised classification is difficult. This choice should be motivated by different factors. First is the adequacy of the model to the data, for which several criteria have been proposed. In the context of

a mixture model, the most frequently employed is the Bayesian information criterion (BIC), and in the context of a classification model, the classification likelihood criteria (CLC) and its BIC-type approximation called integrated classification likelihood-Bayesian information criterion (ICL-BIC) are the main criteria used[19,20]; the ICL-BIC was used in the present study. Second, the posterior probability of belonging to the group in which a trajectory was classified should be high for all trajectories; thus, the typical trajectory of a group should be defined by individual trajectories that "truly" belong to the group. Furthermore, one should avoid classifications with empty groups. Finally, the model choice should also have a clinical rationale and the obtained groups a meaningful clinical interpretation.

2.4 | The factors influencing classification

According to formula (6), the posterior probability that trajectory i belongs to group k depends on the prior probability of the kth group π_k, and the Gaussian density of trajectory y_i that has mean $\mathbf{X}_{1i}\hat{\boldsymbol{\gamma}} + \mathbf{X}_{2i}\hat{\boldsymbol{\beta}}_k$ and variance σ_k^2.

Let us consider the simple case of Gaussian trajectories with only one measurement per subject and two groups. The means of Groups 1 and 2 are 60 and 68, respectively, and their same standard deviation is 5. With equal prior probabilities across the two groups, a subject with measurement 63.5 is closer to the mean of Group 1 than to that of Group 2 (Euclidian distance); besides, he/she is classified in Group 1 because the density of this trajectory is higher for Group 1 than for Group 2 (Figure 1, panel A). However, when the standard deviations are distinct (say, 1 for Group 1 and 8 for Group 2), value 63.5 is classified in Group 2 (Figure 1, panel B) even if the Euclidean distance to Group 1 is smaller. This example illustrates the effect of intra-group variance on the classification. Consequently, the assumption of an equal intra-group variance between groups imposed by Mod_equal may bias the classification.

Let us consider now a case where the standard deviations relative to Groups 1 and 2 are 3 and 8, respectively. Panel C in Figure 1 represents the distribution densities of two groups that share equal prior probability and the distribution densities weighted by the unequal prior probability 30% and 70%. With equal prior probability, the individual with mean trajectory 63.5 is classified into Group 1, but with unequal prior probability, he/she is classified into Group 2 because the density of this trajectory weighted by the corresponding prior probability is higher in Group 2 than in Group 1. Hence, the partition in classification models is based not only on Euclidean distance but also on variance and prior probability.

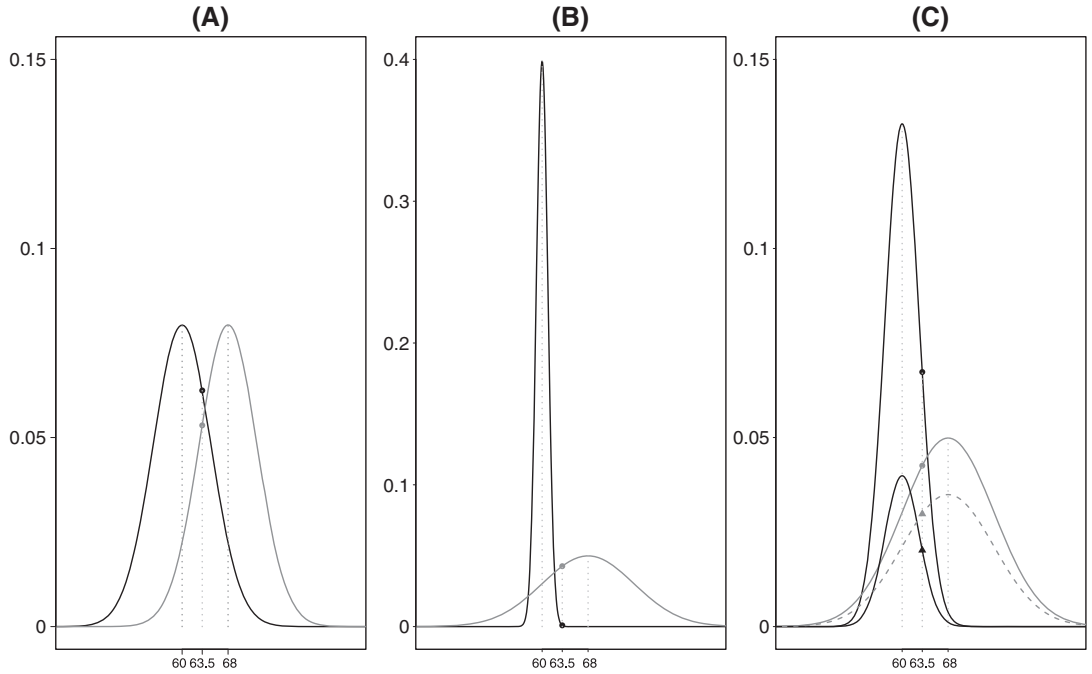


FIGURE 1 Distribution densities of Groups 1 and 2 in the illustration of the factors influencing the classification. For the three panels, the means of Groups 1 and 2 are 60 and 68, respectively. Panel A: two Gaussian densities with equal variance ($\sigma_1 = \sigma_2 = 5$) and equal prior probability ($\pi_1 = \pi_2 = 50\%$). Panel B: two Gaussian densities with unequal variance ($\sigma_1 = 1$, $\sigma_2 = 8$) and equal prior probability ($\pi_1 = \pi_2 = 50\%$). Panel C: two Gaussian densities with unequal variance ($\sigma_1 = 3$, $\sigma_2 = 8$,) and equal prior probability ($\pi_1 = \pi_2 = 50\%$) in solid lines, and two others with unequal variance ($\sigma_1 = 3$, $\sigma_2 = 8$), and unequal prior probability ($\pi_1 = 30$, $\pi_2 = 70\%$) in dashed lines

TABLE 1 Misclassification rates of Mod_unequal and Mod_equal for the reference simulation design and for the design with the proportion of the trajectories in each group given by P2

Proportions and Model	S1 (2, 2, 2)	S2 (2, 3, 4)	S3 (4, 3, 2)	S4 (4, 2, 3)	S5 (0.5, 3, 6)	S6 (0.6, 1, 1.33)
P1: (22%, 31%, 47%) - Mod_equal						
MCR (%)						
Overall	6.3	57.5	21.9	19.0	61.5	0.11
Group 1	15.5	8.2	71.7	70.4	6.8	0.00
Group 2	7.6	85.4	19.2	3.8	92.8	0.33
Group 3	0.8	61.6	0.5	5.1	66.1	0.07
P1: (22%, 31%, 47%) - Mod_unequal						
MCR (%)						
Overall	6.4	13.8	20.4	4.0	2.6	0.01
Group 1	16	14.6	49.8	7.3	0	0.09
Group 2	8	19.2	23.6	3.5	2.8	0.1
Group 3	0.8	9.6	4.2	2.5	3.6	0.007
P2: (33%, 33%, 33%) - Mod_equal						
MCR (%)						
Overall	7.5	51.5	33.9	28.4	54.4	0.13
Group 1	9.7	3.0	79.8	74.7	2.9	0.00
Group 2	11.5	89.5	21.0	4.1	96.3	0.37
Group 3	1.3	62.0	0.4	6.1	64.1	0.02
P2: (33%, 33%, 33%) - Mod_unequal						
MCR (%)						
Overall	7.6	12.0	28.4	4.4	2.1	0.01
Group 1	10.6	6.8	41.8	6.3	0	0.02
Group 2	10.8	16.9	37.4	3.5	1.9	0.02
Group 3	1.2	12.2	5.8	3.3	4.2	0.0009

MCR: misclassification rate; S1-S6: scenarios of intra-group variances (square root of intra-group variances given in brackets); P1 = (22%, 31%, 47%), P2 = (33%, 33%, 33%): proportion of the trajectories in each group; Mod_equal: model with equal intra-group variance between groups; Mod_unequal: model with unequal intra-group variance between groups.

3 | SIMULATIONS

3.1 | Designs

Simulation studies were performed to evaluate the impact on the classification of unequal vs equal intra-group variance across groups according to the total sample size, the number of repeated measurements, and the real intra-group variances.

A reference simulation design was defined: 320 trajectories ($N = 320$), 11 repeated measurements ($T = 11$), three groups, and proportion of trajectories in each group is given by P1 = (22%, 31%, and 47%). Another design was considered with the proportion of trajectories given by P2 = (33%, 33%, and 33%). For these two designs, six different scenarios (S1 to S6) were used by varying the square roots of the intra-group variances (Table 1). The typical trajectories were defined by a second-order polynomial that differed from one group to another by the initial value and the slope (Figure 2). The typical trajectories were the same for the six scenarios.

In scenario S1, the square root of the intra-group variance was the same for all groups (2, 2, and 2). In scenarios S2 to S4, the square root of the intra-group variances were (2, 3, and 4), (4, 3, and 2), and (4, 2, and 3), respectively, to reflect increasing, decreasing, and random order of the intra-group variances. Scenarios S5 and S6 were similar to the scenario S2, but the differences between the intra-group variances were higher in S5 and smaller in S6 than in S2.

Other designs were considered by varying the total sample size ($N = 50$, $N = 100$, and $N = 200$), and the number of repeated measurements ($T = 5$) compared with the reference design. Scenarios S1, S2, and S5 were considered for these designs.

In addition, to evaluate the impact of allowing unequal intra-group variance on the number of groups chosen by the ICL-BIC, the two models were estimated for the reference design with different K values ($K = 2$, $K = 3$, and $K = 4$).

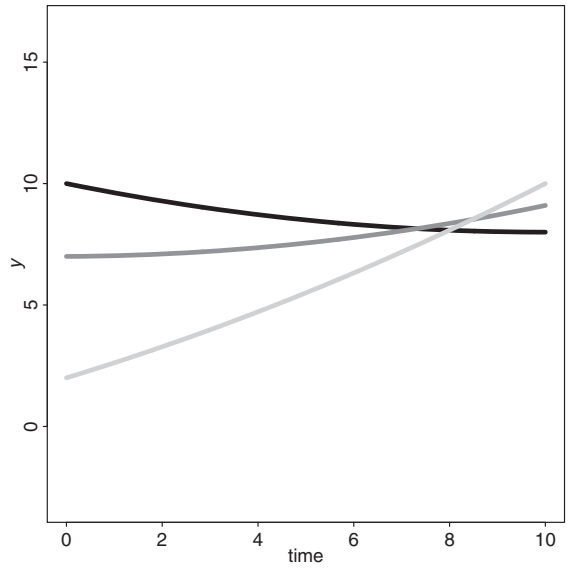


FIGURE 2 Typical trajectories of the three groups used for the simulation study. Black line: typical trajectory of Group 1; medium gray line: typical trajectory of Group 2; light gray line: Group 3

Scenarios S1, S2, and S5 were considered. The proportion of simulations for which two, three, or four groups were chosen was calculated.

For each scenario, 1000 datasets were simulated. Mod_equal and Mod_unequal were applied to each dataset. A second-order polynomial relationship was used to model the link between the outcome y and time. The CEM algorithm was stopped when the difference in the log classification likelihood after the C-step of two consecutive iterations was less than 0.001. It was repeated 100 times from different initial partitions. The model performance was assessed by the overall misclassification rate (MCR), the bias (difference between the estimated and the real typical trajectories), the bias in intra-group variance estimates, and the number of times the algorithm found an empty group.

3.2 | Results

For the reference design and for the design with the proportions of the trajectories given by P2, the MCR was smaller in Mod_unequal than in Mod_equal except for scenario S1 in which the real intra-group variances were equal and the MCR was similar (Table 1). With Mod_equal and an unequal real intra-group variance, the MCR was sometimes very high (around 50%).

In Mod_equal, the MCR was higher in scenarios S2 to S5 than in S1. For example, in S2, the square root of this variance was estimated to be 3.11, so the intra-group variances of Groups 1 and 2 were overestimated. Consequently, Group 1 attracted some trajectories of the other groups (especially from Group 2), and Group 2 attracted some trajectories of Group 3, and consequently, the MCRs of Groups 2 and 3 were high. Due to the appropriate intra-group variance estimates, Mod_unequal corrected this misclassification. The amplitude of the correction depended on (i) the distances between the typical trajectories; (ii) the values of the intra-group variances; and (iii) the differences between intra-group variances. The amplitude of the correction was higher in S2 than in S3 and S4. In S3, the two closest typical trajectories (Groups 1 and 2) had high square root intra-group variances (3 and 4) relative to the distances between typical trajectories. Hence, these two groups overlapped and neither of these models was able to classify correctly the trajectories. Conversely, with S2, the two closest typical trajectories (Groups 1 and 2) had lower square root intra-group variances (2 and 3) than Group 3; the two groups did not overlap too much and the Mod_unequal was able to identify the groups. The amplitude of the correction between the two models was as important as the groups were separated owing to the real intra-group variances. In scenario S6, the groups were well separated even when assuming wrongly equal intra-group variance; consequently, both models were able to classify correctly the trajectories.

The groups with high underestimation of the square root of the intra-group variances were the groups with high MCR in Mod_equal (Table S1). Moreover, the overall MCR of Mod_equal was higher with proportion vector P1 than with P2

TABLE 2 Misclassification rates of Mod_unequal and Mod_equal according to the total sample size and the number of repeated measurements

Design and Model	S1 (2, 2, 2)	S2 (2, 3, 4)	S5 (0.5, 3, 6)
Reference design			
$N = 320$, P1 = (22%, 31%, 47%), $T = 11$			
MCR (%) Mod_equal	6.3	57.5	61.5
MCR (%) Mod_unequal	6.4	13.8	2.6
Change of total sample size for $T = 11$			
$N = 200$			
MCR (%) Mod_equal	6.7	57.4	61.6
MCR (%) Mod_unequal	7.2	15.5	2.6
$N = 100$			
MCR (%) Mod_equal	8.0	55.3	61.3
MCR (%) Mod_unequal	9.4	19.5	2.7
$N = 50$			
MCR (%) Mod_equal	10.7	53.7	61.5
MCR (%) Mod_unequal	12.1	23.5	3.1
Change of total sample size for $T = 5$			
$N = 320$			
MCR (%) Mod_equal	7.9	57.2	61.3
MCR (%) Mod_unequal	10.1	40.8	13.9
$N = 100$			
MCR (%) Mod_equal	11.2	56.1	61.3
MCR (%) Mod_unequal	18.4	37.8	8.9
$N = 50$			
MCR (%) Mod_equal	13.0	54.7	61.1
MCR (%) Mod_unequal	23.1	37.9	12.9

MCR: misclassification rate; S1, S2, S5: scenarios of intra-group variances (square root of intra-group variances given in brackets); P1 = (22%, 31%, 47%): proportion of the trajectories in each group; *N*: total sample size; *T*: number of repeated measurements; Mod_equal: model with equal intra-group variance between groups; Mod_unequal: model with unequal intra-group variance between groups.

in scenarios S2 and S5, and smaller in scenarios S3 and S4. Indeed, in S2 and S5, the groups with the highest MCRs were those with the highest proportions; this resulted in a high overall MCR. The bias in the estimated typical trajectories was often high when the MCR was high and low when the MCR was low. Finally, the number of times the algorithm found an empty group was lower in Mod_unequal than in Mod_equal.

When the total sample size was smaller than the one of the reference design ($N = 200$, $N = 100$, and $N = 50$) but the number of repeated measurements still equals to 11, the MCRs of the Mod_unequal were much lower (at least 30% less) than those of the Mod_equal for S2 and S5, as found for the reference design, and a little higher for S1 (less than 2% higher) (Table 2). When the number of repeated measurements decreased to 5, the MCRs of the Mod_unequal increased compared with the reference design, but they were still lower than those of the Mod_equal for S2 and S5. For S1, the MCR for the Mod_unequal was higher than that of the Mod_equal in case of small total sample size. Indeed, the intra-group variances of the Mod_unequal were greatly underestimated in this case, leading to an increase in the MCR.

The proportion of simulations for which three groups (ie, the real number of groups) was chosen by the Mod_unequal was higher than 85% irrespective of the scenario (Table 3); for Mod_equal, a comparable proportion of simulations found three groups for S1, but a considerably lower proportion found three groups for S2 and S5. For the latter model, two groups were selected in the majority of simulations in S2, and this was the result of trajectories of predefined Groups 1 and 2 being regrouped. This was because their typical trajectories were close and the real intra-group variance of each group was quite high. When these two groups were merged, the estimated intra-group variance of the resultant group was close to that of Group 3, which is expected for a model that assumes equal intra-group variance.

TABLE 3 Proportion (%) of simulations for which two, three, or four groups were chosen by ICL-BIC

Model	*K*	S1 (2, 2, 2)	S2 (2, 3, 4)	S5 (0.5, 3, 6)
Mod_equal	2	0.1	91.6	17.1
	3*	99.8	7.8	47.4
	4	0.1	0.6	35.5
Mod_unequal	2	1.2	14.4	0
	3*	98.4	85.4	100
	4	0.4	0.2	0

S1, S2, and S5: scenarios of intra-group variances (square root of intra-group variances given in brackets); Mod_equal: model with equal intra-group variance between groups; Mod_unequal: model with unequal intra-group variance between groups; *K*: number of groups assumed by the model; ICL-BIC: integrated classification likelihood-Bayesian information criterion.

* the real number of groups.

4 | APPLICATION

4.1 | Context

In a prospective, multicenter, open-label, parallel-group controlled trial called "Low-cyclo",[21] 95 patients aged 18 to 65 years old undergoing de novo heart transplantation were centrally randomized to receive either a low dose or a standard dose of Cyclosporine A (CsA concentrations 130-200 μg/L, $n = 47$ vs 200-300 μg/L, $n = 48$) during the first three months post-transplant (along with mycophenolate mofetyl and corticosteroids). The assessment of renal function after transplantation called for eight measurements of serum creatinine: one every 50 days for the first three months, one per month until six months, and quarterly thereafter until end of follow-up. A mixed model did not find significant differences in creatinine over time between the two arms. The representation of the individual trajectories of creatinine over time had high heterogeneity (Figure 3). One hypothetical explanation was the presence of groups within the population, with an effect of low-dose CsA only in some of them; this motivated the use of a classification model to identify groups of creatinine trajectories and estimate the effect of low-dose CsA dose on each group.

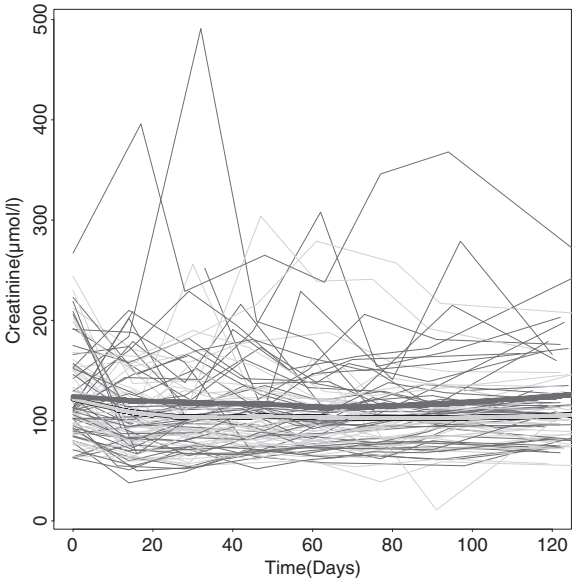


FIGURE 3 Typical and observed trajectories by treatment arm. Bold dark gray line: typical trajectories of patients with standard dose; bold light gray line: typical trajectories of patients with low dose; straight lines: observed creatinine trajectories

TABLE 4 Proportions (%) of trajectories and parameters estimates of Mod_unequal and Mod_equal in low-cyclo study

	Proportions (%)			Square Root of Intra-Group Variance			Mean Difference*		
Model	**Gr. 1**	**Gr. 2**	**Gr. 3**	**Gr. 1**	**Gr. 2**	**Gr. 3**	**Gr. 1**	**Gr. 2**	**Gr. 3**
Mod_equal	7.4	61	31.6	32	32	32	28.9	28.2	1.89
Mod_unequal	33.7	44.2	22.1	56.4	17.9	11.7	22.5	2.4	2

Mod_equal: model with equal intra-group variance between groups; Mod_unequal: model with unequal intra-group variance between groups.
* between low-dose and standard-dose typical trajectories (μg/L).

4.2 | Classification models

The 95 patients were classified using the Mod_equal and Mod_unequal. In each model, a second-order polynomial relationship was used to model the link between creatinine and time. Knowing that the effect of the low dose of CsA on creatinine is progressive, an interaction between treatment and time was added to the model in order to estimate the treatment effect. The whole model may be written as

$$\mathbf{Creat}_i = \beta_{0k} + \beta_{1k}\mathbf{Time}_i + \beta_{2k}\left(\mathbf{Time}_i^2\right) + \beta_{3k}\left(\mathbf{Time}_i \times \mathbf{dose\_group}_i\right) + \boldsymbol{\varepsilon}_i.$$

As in the simulations, the CEM algorithm was repeated 100 times from different initial partitions of the trajectories.

A classification with three groups was finally selected. The three groups retained by the two models may be clinically relevant. Group 1 corresponded to patients with low kidney function; these had high creatinine levels during follow-up and even an increase at the beginning of follow-up in Mod_equal. These patients were likely to have an organic renal insufficiency. Group 2 corresponded to patients with low kidney function just after heart transplantation and improvement after transplantation (return to normal levels around 100 μmol/L). These patients had probably a renal insufficiency related to the heart disease that required transplantation. Group 3 corresponded to patients with normal levels of creatinine at the time of transplantation, which remained stable after which, and therefore with a good kidney function.

The square root intra-group variances, the proportions of trajectories, and the mean difference between low dose and standard dose were different between Mod_equal and Mod_unequal (Table 4, panels A and B in Figure 4).

With Mod_equal, Group 1 included only seven patients. Despite the hypothesis of an equal intra-group variance, the observed individual trajectories of Groups 1 and 2 were more heterogeneous than those of Group 3 (panel C in Figure 4). Hence, Mod_equal may be not appropriate for this application.

The benefit of low-dose CsA was observed in Groups 1 and 2 with Mod_equal and only in Group 1 with Mod_unequal. Group 1 of Mod_unequal, which had the highest intra-group variance, attracted 23 patients classified in Group 2 by Mod_equal (all 16 patients with standard-dose CsA and seven patients with low-dose CsA, who had high creatinine values; Table 5), and Group 2 of Mod_unequal attracted 35 patients classified in Group 3 by Mod_equal (15 patients with standard-dose CsA and 20 patients with low-dose CsA, who had low creatinine values). In consequence, Group 2 of Mod_unequal included 83% of patients classified in Group 3 with Mod_equal for which no benefit of low-dose CsA was found. That is why we found no benefit of low-dose CsA in Group 2 with Mod_unequal.

5 | DISCUSSION

The present work presents a classification model for trajectories that allows unequal intra-group variance across groups. The simulation results showed that this model gives better classification results than the one with equal intra-group variance when the real intra-group variances are different across groups even when the total sample size is small.

When the real intra-group variances are equal and the total sample size and the number of repeated measurements are large, similar classifications are obtained with Mod_unequal and Mod_equal; however, when the real intra-group variances are equal and the total sample size and the number of repeated measurements are small, Mod_equal may provide classifications a little better than Mod_unequal. Thus, choosing the most appropriate model will depend on the study design (ie, the number of subjects included and the measurements made) but also on the biological mechanism analyzed. In observational studies, unequal intra-group variance is commonplace, and therefore, Mod_unequal should be recommended. However, in the case of an experimental design in which all conditions are controlled, it may be assumed that the real intra-group variances are equal, and thus, Mod_equal should be recommended.

The CEM algorithm was used to maximize the classification likelihood. Its solution is known to depend on the initial partition of the trajectories between groups. In the present work, the CEM algorithm was repeated several times starting

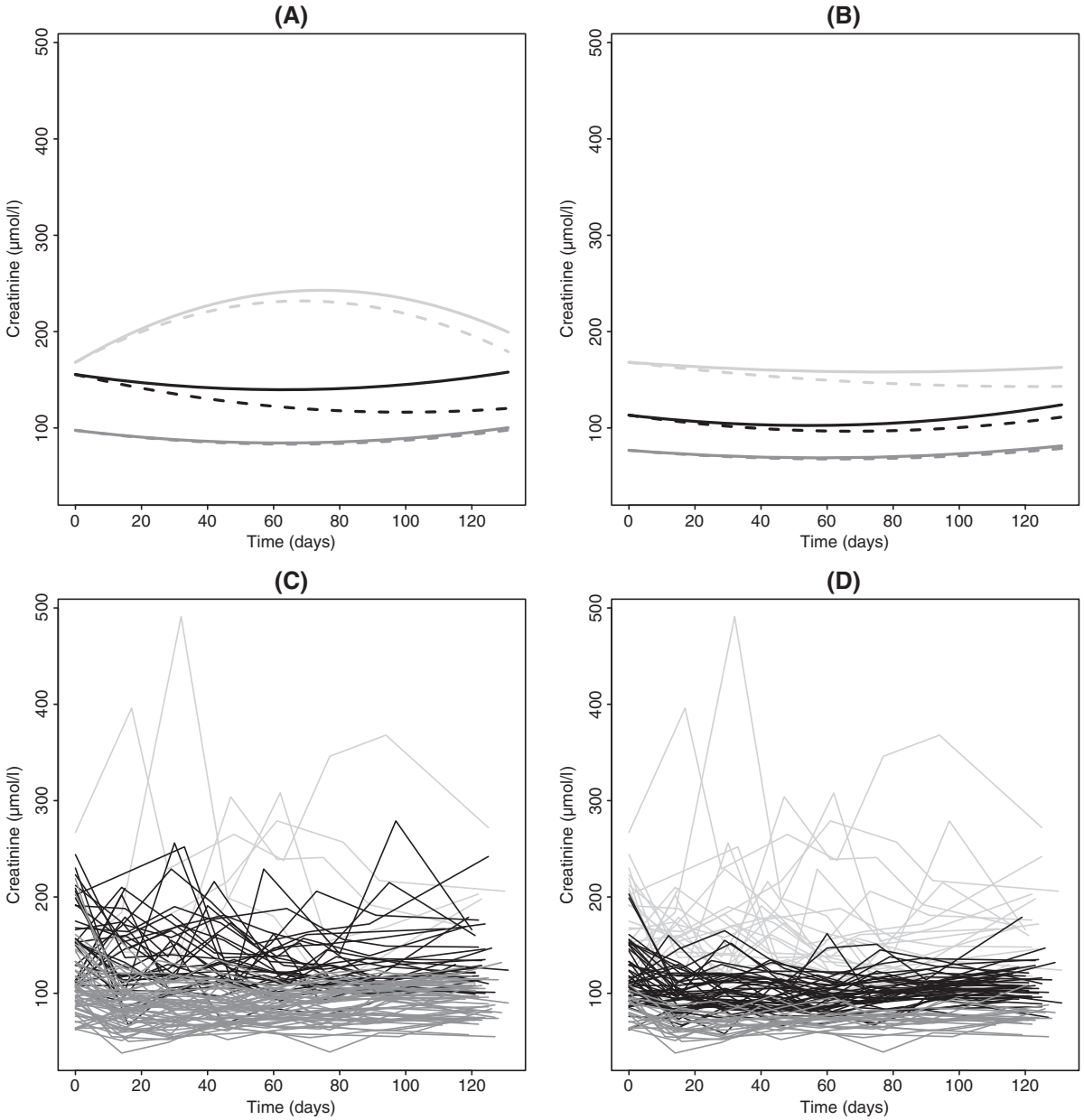


FIGURE 4 Classification of creatinine trajectories obtained with the classification model with equal intra-group variance across groups and with the classification model with unequal intra-group variance. Panels A and B: typical trajectories obtained with the two classification models (Mod_equal) and (Mod_unequal), respectively. Panels C and D: observed individual trajectories colored according to the group membership obtained with the two classification models (Mod_equal) and (Mod_unequal), respectively. Dark gray line: typical trajectories of Group 1; medium gray line: typical trajectories of Group 2; light gray line: typical trajectories of Group 3; solid lines: standard dose arm; dashed lines: low-dose arm

with different initial partitions and the clustering that provided the greatest value of the log classification likelihood was retained. This proved to be time consuming. Other algorithms that depend weakly on the initial partitions could have been used instead (such as the stochastic EM (SEM) algorithm[13]), but these algorithms need a large number of iterations to converge.

To perform a statistical test or to obtain confidence interval of parameters estimated at the last iteration of the CEM algorithm, one could calculate the standard errors from the inverse observed information matrix evaluated at the maximum-likelihood estimates[22] irrespective of the structure of variance. However, this approach does not take into account the uncertainty of the classification obtained during the C-step of the CEM algorithm. To address this issue, we suggest an approach to calculate these standard errors based on the method proposed by Rubin[23] and the SEM algorithm.

TABLE 5 Contingency tables of patient classification according to the two models for the two arms

Standard-Dose CsA					Low-Dose CsA				
	Mod_unequal					Mod_unequal			
	Gr. 1	Gr. 2	Gr. 3	Total		Gr. 1	Gr. 2	Gr. 3	Total
Mod_equal					Mod_equal				
Gr. 1	4	0	0	4	Gr. 1	3	0	0	3
Gr. 2	16	0	0	16	Gr. 2	7	7	0	14
Gr. 3	1	15	12	28	Gr. 3	1	20	9	30
Total	21	15	12	48	Total	11	27	9	47

Mod_equal: model with equal intra-group variance between groups; Mod_unequal: model with unequal intra-group variance between groups.

For this, a SEM algorithm is run on the last iteration of the CEM algorithm. The standard error of a parameter is then calculated by the square root of the sum (1) of the mean of the estimated variances of the parameter over all SEM iterations and (2) of the variance of the estimated parameters over all SEM iterations. Further work is needed to investigate and verify the validity of this method, independently of the structure of variance.

In the present work, we used the unrestricted version of the classification likelihood[9]; the "prior" probabilities are estimated using the study data, as with empirical Bayes methods.[24] This means that an individual trajectory is classified not only according to its own values but also according to the other individual trajectories through the prior probabilities. Classifying a trajectory using only the individual values can be made using the restricted version of the classification likelihood where the prior probabilities are constrained to be equal.[9] An interesting work would be to investigate the restricted classification version in the case of distinct intra-group variances. Indeed, as seen in Equation (6), the "prior" probabilities and the intra-group variances play similar roles in the classification, and therefore, fixing the prior probabilities may improve the identifiability of the classification model.

Mod_unequal and Mod_equal rely on the assumption of conditional independence between the measurements of the same trajectory given the group membership. The assumption that trajectory classification removes the serial dependence between same-subject measurements was already raised in the mixture model.[1] Within the context of the mixture model, Muthén and Shedden[25] have proposed to take into account this serial dependence by adding individual random effects within each group. The impact of random-effect addition on classification results has not yet been analyzed. This addition would allow the distinction of inter-individual from intra-individual variances within each group; herein, the two types of variances were pooled into the "intra-group variance".

In the Low-cyclo study, Group 1 obtained with Mod_equal included only seven extreme trajectories that were furthermore very heterogeneous. With Mod_unequal, this group kept 23 trajectories that were also heterogeneous. The three typical trajectories obtained with Mod_unequal differed only by the level of creatinine at transplantation. However, the individual trajectories were not only classified according to this level, they were also classified according to the heterogeneity of their values.

6 | CONCLUSION

On the basis of the aforementioned simulation study, we recommend using the classification model with unequal intra-group variance across groups, except when the total sample size and the number of repeated measurements are small. In the latter situation, the choice of model should depend on the study design and prior knowledge of the biological mechanism analyzed.

ACKNOWLEDGEMENTS

The authors would like to thank Pascale Boissonnat for providing Low-cyclo data. We also thank Dr Philip Robinson for helpful comments, suggestions, and revisions of the final manuscript.

FUNDING AND COMPETING INTERESTS

The authors declare to have no conflict of interest. This work was partially funded by the Agence Nationale de la Recherche (ANR Grant 2012 BLAN SVSE 1).

ORCID

Amna Klich http://orcid.org/0000-0002-3052-7335

REFERENCES

1. Nagin DS, Odgers CL. Group-based trajectory modeling in clinical research. *Annu Rev Clin Psychol*. 2010;6:109-138.
2. Pickles A, Croudace T. Latent mixture models for multivariate and longitudinal outcomes. *Stat Methods Med Res*. 2010;19:271-289.
3. Einbeck J, Darnell J, Hinde J. Nonparametric maximum likelihood estimation for random effect models. The npmlreg Package. 2014.
4. Formann AK. Mixture analysis of longitudinal binary data. *Stat Med*. 2006;25(9):1457-1469.
5. Legler JM, Davis WW, Potosky AL, Hoffman RM. Latent variable modelling of recovery trajectories: sexual function following radical prostatectomy. *Stat Med*. 2004;23(18) :2875-2893.
6. Symons MJ. Clustering criteria and multivariate normal mixtures. *Biometrics*. 1981;37(1):35-43.
7. Subtil F, Boussari B, Bastard M, et al. An alternative classification to mixture modeling for longitudinal counts or binary measures. *Stat Methods Med Res*. 2017;26:453-470.
8. James GM, Sugar CA. Clustering for sparsely sampled functional data. *J Am Stat Assoc*. 2003;98(462):397-408.
9. Celeux G, Govaert G. Comparison of the mixture and the classification maximum likelihood in cluster analysis. *J Stat Comput Simul*. 1993;47(3-4):127-146.
10. Govaert G, Nadif M. Comparison of the mixture and the classification maximum likelihood in cluster analysis with binary data. *Comput Stat Data Anal*. 1996;23(1):65-81.
11. Genolini C, Falissard B. KmL: k-means for longitudinal data. *Comput Stat*. 2010;25(2):317-328.
12. Hartigan JA, Wong MA. Algorithm AS 136: a k-means clustering algorithm. *J R Stat Soc Ser C Appl Stat*. 1979;28(1):100-108.
13. Celeux G, Govaert G. A classification EM algorithm for clustering and two stochastic versions. *Comput Stat Data Anal*. 1992;14:315-332.
14. Buyske S. R Package mmlcr: Mixed-Mode Latent Class Regression. 2003.
15. Gaffney S, Smyth P. Trajectory clustering with mixtures of regression models. In: Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 1999; San Diego, CA. https://doi.org/10.1145/312129.312198
16. Elsensohn M-H, Klich A, Ecochard R, et al. A graphical method to assess distribution assumption in group-based trajectory models. *Stat Methods Med Res*. 2016;25:968-982.
17. Dempster A, Laird N, Rubin D. Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc Ser B Methodol*. 1977;39(1):1-38.
18. Gałecki A, Burzykowski T. *Linear Mixed-Effects Models Using R: A Step-by-Step Approach*. New York, NY: Springer; 2013.
19. McLachlan G, Peel D. *Finite Mixture Models*. Hoboken, NJ: John Wiley & Sons; 2004.
20. Biernacki C, Celeux G, Govaert G. Assessing a Mixture Model for Clustering with the Integrated Classification Likelihood. INRIA; 1998.
21. Boissonnat P, Gaillard S, Mercier C, et al. Impact of the early reduction of cyclosporine on renal function in heart transplant patients: a French randomised controlled trial. *Trials*. 2012;13:231
22. Jones BL, Nagin DS. Advances in group-based trajectory modeling and an SAS procedure for estimating them. *Sociol Methods Res*. 2007;35:542-571.
23. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. Hoboken, NJ: John Wiley & Sons; 2004.
24. Carlin BP, Louis TA. *Bayes and Empirical Bayes Methods for Data Analysis*. 2nd ed. New York, NY: Taylor & Francis; 2010.
25. Muthén B, Shedden K. Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics*. 1999;55(2):463-469.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Klich A, Ecochard R, Subtil F. Unequal intra-group variance in trajectory classification. *Statistics in Medicine*. 2018;37:4155–4166. https://doi.org/10.1002/sim.7921

Supporting information

Supplementary Table S1 – mean bias of the estimated typical trajectories and number of times the algorithm found an empty group for Mod_unequal and Mod_equal, for the reference simulation design and for the design with the proportions of the trajectories given by P2.

Proportions and model	S1 (2,2,2)	S2 (2,3,4)	S3 (4,3,2)	S4 (4,2,3)	S5 (0.5,3,6)	S6 (0.6,1,1.33)
P1: (22%,31%,47%) - Mod_equal						
Bias						
Overall	0.0245	-1.0305	0.3952	0.3131	-1.3545	-0.0008
Group 1	0.0738	-0.7751	0.6430	0.7739	-1.0400	-0.0016
Group 2	0.0025	-2.1223	0.4421	0.2165	-2.8128	-0.0028
Group 3	-0.0026	-0.1940	0.1009	-0.0510	-0.2100	-0.0012
Bias in sigma						
Group 1	-0.0156	1.3312	-1.1688	-1.0331	3.8363	0.4191
Group 2	-0. 0156	0.3312	-0.1688	0.9668	1.3364	0.0101
Group 3	-0. 0156	-0.6668	0.8311	-0.0331	-1.6636	-0.3108
Nf	32	30	32	41	16	50
P1: (22%,31%,47%) - Mod_unequal						
Bias						
Overall	0.0260	-0.0154	0.1358	0.0153	-0.0127	0.0004
Group 1	0.0728	0.0185	0.3251	0.0580	0.0007	0.0001
Group 2	0.0071	-0.0176	0.0574	0.0018	0.0015	-0.0001
Group 3	0.0013	-0.0473	0.0248	-0.0144	-0.0037	0.0007
Bias in sigma						
Group 1	-0.0417	-0.0618	-0.0543	0.0196	0.0001	-0.0018
Group 2	-0.0215	-0.0722	-0.1076	-0.0044	-0.0107	-0.0001
Group 3	-0.0054	-0.0639	-0.0195	-0.0048	0.0228	-0.0007
Nf	16	16	14	6	14	36
P2: (33%,33%,33%) - Mod_equal						
Bias						
Overall	-0.0032	-0.9511	0.4824	0.3578	-1.2586	-0.0017
Group 1	0.0278	-0.5002	0.7287	0.8297	-0.6915	-0.0021
Group 2	-0.0345	-2.1301	0.5744	0.3139	-2.8168	-0.0031
Group 3	-0.0031	-0.2232	0.1439	-0.0707	-0.2672	$-2.2\ 10^{-20}$
Bias in sigma						
Group 1	-0.0191	1.1043	-0.9077	-0.9141	3.2906	0.4179
Group 2	-0.0191	0.1043	0.0922	1.0858	0.7906	0.0179
Group 3	-0.0191	-0.8956	1.0922	0.0858	-2.2093	-0.3120
Nf	24	31	30	33	17	31
P2: (33%,33%,33%) - Mod_unequal						
Bias						
Overall	0.0008	-0.0722	0.0900	0.0079	-0.0165	-0.0005
Group 1	0.0352	0.0053	0.2193	0.0415	0.0010	-0.0059
Group 2	-0.0296	-0.0842	0.0253	-0.0014	-0.0040	-0.0006
Group 3	-0.0033	-0.1382	0.0262	-0.0164	-0.0449	-0.0005
Bias in sigma						
Group 1	-0.0243	-0.0134	-0.2002	0.0171	-0.0008	-0.0009

	Group 2	-0.0315	0.0017	-0.2483	-0.0030	-0.0002	-0.0013
	Group 3	-0.0053	-0.0072	-0.0144	-0.0074	0.0278	-0.0033
Nf		13	12	13	5	9	34

Bias: difference between the estimated and the real typical trajectory- Bias in sigma: bias in the square root of intra-group variance - Nf: number of times over 1000 simulations that the algorithm found an empty groups - S1-S6: scenarios of intra-group variances (square root of intra-group variances given in brackets) - P1=(22%, 31%, 47%), P2=(33%,33%,33%): proportion of the trajectories in each group - Mod_equal: model with equal intra-group variance between groups - Mod_unequal: model with unequal intra-group variance between groups.

3.4 Principaux résultats de l'article

Les résultats des simulations ont montré que le modèle de classification avec une variance résiduelle variable d'un groupe à l'autre donne de meilleurs résultats de classification que celui avec une variance résiduelle supposée identique entre les groupes lorsque les variances résiduelles réelles sont différentes d'un groupe à l'autre, même lorsque l'effectif total de l'échantillon est faible. L'écart entre les deux modèles exprimé en pourcentage d'individus mal classés peut atteindre 40 % à 50 %, ce qui entraîne un biais dans l'estimation des trajectoires typiques.

Lorsque la variance résiduelle réelle est identique d'un groupe à l'autre, des classifications similaires sont obtenues avec les deux modèles dans le cas où nombre de trajectoires et le nombre de mesures répétées sont élevés. Dans le cas contraire (dans les simulations, moins de 5 mesures par trajectoire et moins de 100 trajectoires), le modèle de classification standard conduit à un pourcentage de trajectoires mal classées plus faible.

Ainsi, le choix du modèle le plus approprié dépendra du plan d'étude (nombre de sujets inclus et de mesures effectuées), mais aussi du type d'étude. Dans les études observationnelles dans le domaine de la santé, il est probable que la variance résiduelle soit différente d'un groupe à l'autre, et le modèle avec variance résiduelle différente d'un groupe à l'autre devrait donc être recommandé. Cependant, dans le cas d'un plan expérimental dans lequel toutes les conditions sont contrôlées, les variances résiduelles réelles sont supposées identiques ; le modèle standard pourra donc être recommandé. Notons cependant que, dans un grand nombre de cas, un groupe à pathologie sévère présente une variance plus grande qu'un groupe à pathologie bénigne ou qu'un groupe de sujets normaux.

Sur les données de l'étude de « Lowcyclo », la classification des trajectoires de la créatinine obtenue par le modèle de classification avec une variance résiduelle variable d'un groupe à l'autre est différente de celle obtenue par le modèle de classification standard.

3.5 Estimation des paramètres

Dans le modèle de classification de trajectoires avec variance résiduelle variable d'un groupe à l'autre (noté Mod_unequal dans l'article), les paramètres $\left(\boldsymbol{\gamma},(\boldsymbol{\beta}_k)_{k=1,\ldots,K},(\sigma_k)_{k=1,\ldots,K}\right)$ sont estimés à l'étape M de l'algorithme CEM par la méthode de maximisation de vraisemblance profilée : le principe est de maximiser la vraisemblance par étapes successives. Par soucis de simplicité, cette méthode est détaillée dans la suite en considérant un modèle de classification de trajectoires avec uniquement des covariables étant supposées avoir un effet différent d'un groupe à l'autre ($\boldsymbol{\beta}_k$).

Avant de détailler les étapes de la méthode, il faut préciser la paramétrisation du modèle. La variance résiduelle est décomposée en un terme commun à tous les groupes (σ^2) et des termes spécifiques à chaque groupe latent ($\delta_k, k=1,\ldots,K$) :

$$\text{si } i \in k,\ \ \text{var}(\boldsymbol{\varepsilon}_i) = \sigma_k^2 = \sigma^2 \delta_k \tag{2.2}$$

La modélisation de variance (2.2) utilise $K + 1$ paramètres pour estimer K variances, et n'est donc pas identifiable. Des contraintes sur les paramètres $\boldsymbol{\delta}$ doivent être imposées pour que le modèle soit identifiable. La plus classique consiste à fixer à 1 le paramètre spécifique au premier groupe ($\delta_1 = 1$). Dans ce cas, les paramètres suivants ($\delta_k, k=2,\ldots,K$) représentent les ratios de variance résiduelle par rapport au premier groupe, c'est-à-dire :

$$\text{si } i \in k,\ \ \text{var}(\boldsymbol{\varepsilon}_i) = \sigma_1^2 \delta_k,\ \text{avec } \delta_k = \frac{\sigma_k^2}{\sigma_1^2} \tag{2.3}$$

Dans ce cas, la log-vraisemblance s'écrit:

$$LL(\boldsymbol{\beta}, \sigma_1^2, \boldsymbol{\delta}) \propto -\frac{n}{2}\log(\sigma_1^2) - \frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\log(\delta_k^2) - \frac{1}{2\sigma_1^2}\sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\delta_k^{-2}\left(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}_k\right)^2 \tag{2.4}$$

Si $\boldsymbol{\delta}$ est fixé, il existe une solution explicite à la valeur de $\boldsymbol{\beta}$ maximisant la log-vraisemblance (2.4) :

$$\boldsymbol{\beta}(\boldsymbol{\delta}) = \left(\sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \delta_k^{-2} \mathbf{X}_i \mathbf{X}_i^t \right)^{-1} \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \delta_k^{-2} \mathbf{X}_i \mathbf{y}_i \tag{2.5}$$

En remplaçant $\boldsymbol{\beta}$ par l'estimation $\boldsymbol{\beta}(\boldsymbol{\delta})$ dans l'équation (2.4), la log-vraisemblance, appelée log-vraisemblance profilée selon $\boldsymbol{\beta}$, s'écrit :

$$\begin{aligned} LL(\sigma_1^2, \boldsymbol{\delta}) &= LL(\boldsymbol{\beta}(\boldsymbol{\delta}), \sigma_1^2, \boldsymbol{\delta}) \\ &= -\frac{n}{2} \log(\sigma_1^2) - \frac{1}{2} \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log(\delta_k^2) - \frac{1}{2\sigma_1^2} \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \delta_k^{-2} \mathbf{r}_i^2 \end{aligned} \tag{2.6}$$

avec

$$\begin{aligned} \mathbf{r}_i \equiv \mathbf{r}_i(\boldsymbol{\delta}) &= \mathbf{y}_i - \mathbf{X}_i \hat{\boldsymbol{\beta}}(\boldsymbol{\delta}) \\ &= \mathbf{y}_i - \mathbf{X}_i \left(\sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \delta_k^{-2} \mathbf{X}_i \mathbf{X}_i^t \right)^{-1} \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \delta_k^{-2} \mathbf{X}_i \mathbf{y}_i \end{aligned}$$

Maximiser la vraisemblance profilée (2.6) selon le paramètre σ_1^2 pour une valeur de $\boldsymbol{\delta}$ fixée conduit à un optimum σ_1^2 fonction de $\boldsymbol{\delta}$:

$$\hat{\sigma}_1^2(\boldsymbol{\delta}) = \frac{1}{n} \sum_{i=1}^{n} z_{i1} \mathbf{r}_i^2 \tag{2.7}$$

En remplaçant σ_1^2 dans la vraisemblance profilée (2.6) par son estimation fonction de $\boldsymbol{\delta}$ (2.7), la nouvelle vraisemblance profilée qui ne dépend plus que de $\boldsymbol{\delta}$ est de la forme suivante :

$$LL(\boldsymbol{\beta}(\boldsymbol{\delta}), \hat{\sigma}_1^2(\boldsymbol{\delta}), \boldsymbol{\delta}) = -\frac{n}{2} \log(\hat{\sigma}_1^2(\boldsymbol{\delta})) - \frac{1}{2} \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log(\delta_k^2) - \frac{1}{2\hat{\sigma}_1^2(\boldsymbol{\delta})} \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \delta_k^{-2} \mathbf{r}_i^2 \tag{2.8}$$

L'estimateur $\hat{\boldsymbol{\delta}}$ de $\boldsymbol{\delta}$ est obtenu en maximisant la vraisemblance profilée donnée par (2.8) selon $\boldsymbol{\delta}$, à l'aide d'un algorithme de maximisation numérique. Ensuite, les estimateurs $\boldsymbol{\beta}$ et $\hat{\sigma}_1^2$ sont obtenus en remplacent le paramètre $\boldsymbol{\delta}$ par son estimateur $\hat{\boldsymbol{\delta}}$ dans les formules (2.5) et (2.7). La méthode du maximum de vraisemblance conduit avec cet algorithme à des estimations biaisées de variance résiduelle σ_k^2 ; la maximisation de la vraisemblance restreinte permet dans ce cas de corriger ce biais[18,19].

3.6 Modèles de classification pour des données binomiales

Dans la partie 3.3, l'extension du modèle de classification standard est appliquée à des données continues mesurées de façon répétée au cours de temps, en prenant en compte une variance résiduelle variable d'un groupe à l'autre.

Dans le cas des données binaires ou binomiales, cas dans lequel la variance théorique est déterminée par l'espérance, la régression logistique est le modèle le plus souvent employé. Si la variance estimée à partir des écarts entre les valeurs prédites et les valeurs observées est supérieure à la variance théorique, il y a sur-dispersion. Une sur-dispersion peut être mise en évidence en estimant le paramètre de dispersion. Ce paramètre est donné par le rapport entre la somme des carrées des résidus de Pearson et le nombre de degrés de liberté.

Dans le cadre du modèle de classification, la prise en compte d'une variabilité résiduelle différente d'un groupe à l'autre revient à prendre en compte une sur-dispersion variable d'un groupe à l'autre dans le cas des données binomiales. La prise en compte de cette variabilité est nécessaire, car une sur-dispersion mal prise en compte peut fausser la classification et les estimations des paramètres de trajectoires typiques.

Dans cette partie, le modèle de classification des données longitudinales binomiales en prenant en compte une sur-dispersion variable d'un groupe à l'autre est présenté, et l'impact de cette prise en compte sur la classification est évalué à l'aide de données simulées et sur un exemple de l'observance aux traitements chez des patients séropositifs au VIH et traités par antirétroviraux.

3.6.1 Modèle

Pour prendre en compte la sur-dispersion dans chaque groupe pour les données binomiales, une solution consiste à supposer que la variable $\mathbf{y}_i$, conditionnement à l'appartenance au groupe, suit une loi beta binomiale. Cette loi est une composition de la loi binomiale et la loi de beta, c'est-à-dire que $\left[y_{it} | k \right]$ suit une loi de binomiale de paramètre (n_{it}, p_{itk}), et p_{itk} est aléatoire et suit une loi beta de paramètre a_{itk} et b_{itk}. Le modèle peut être re-paramétré de telle sorte que les paramètres a_{itk} et b_{itk} s'expriment en un paramètre d'espérance μ et un paramètre de dispersion σ de la manière suivante :

$$\begin{cases} a_{itk} = \dfrac{\mu_{itk}}{\sigma_{itk}} \\ b_{itk} = \dfrac{1-\mu_{itk}}{\sigma_{itk}} \end{cases} \tag{2.9}$$

Dans ce cas, l'espérance est donnée par $\frac{a_{itk}}{a_{itk}+b_{itk}}$ et la variance par $\mu_{itk}(1-\mu_{itk})\left(1+(n_{it}-1)\frac{\sigma_{itk}}{\sigma_{itk}+1}\right)$. Le paramètre de dispersion est donné par : $\sigma_{itk} = \frac{1}{a_{itk}+b_{itk}}$; s'il est nul, il n'y a pas de sur-dispersion, et la variance correspond à la variance classique de la loi binomiale.

En reprenant les notations précédemment introduites dans la partie 3.2, l'espérance peut être modélisée par des variables explicatives $\mathbf{X}$ incluant le temps (e.g. des caractéristiques des individus et d'éventuelles interactions entre le temps et ces caractéristiques). Si la trajectoire *i* appartient au groupe *k,* le modèle s'écrit de la façon suivante :

$$\text{logit}(\boldsymbol{\mu}_{ik}) = \mathbf{X}_i\boldsymbol{\beta}_k \tag{2.10}$$

Le modèle est contraint de telle sorte que la somme des paramètres a_{itk} et b_{itk} soit fixe, mais puisse être variable d'un groupe à l'autre ($\sigma_{itk} = \sigma_k$), seul le paramètre d'espérance varie en fonction du temps et de groupe. Le modèle obtenu est appelé Mod_Unequal. Lorsque le paramètre de dispersion est contraint à être égal entre les groupes (c'est-à-dire, $\sigma_1 = \sigma_2 = ... = \sigma_K$), le modèle est appelé Mod_Equal.

En supposant que les mesures de chaque individu sont indépendantes conditionnellement à l'appartenance au groupe, la vraisemblance de $\left[\mathbf{y}_i \middle| k\right]$ s'écrit :

$$f\left(\mathbf{y}_i, \boldsymbol{\beta}_k, \sigma_k\right) = \prod_{t=1}^{T_i} \frac{B\left(y_{it} + \frac{\mu_{itk}\left(\boldsymbol{\beta}_k\right)}{\sigma_k}, n_{it} - y_{it} + \frac{1-\mu_{itk}\left(\boldsymbol{\beta}_k\right)}{\sigma_k}\right)}{B\left(\frac{\mu_{itk}\left(\boldsymbol{\beta}_k\right)}{\sigma_k}, \frac{1-\mu_{itk}\left(\boldsymbol{\beta}_k\right)}{\sigma_k}\right)} \tag{2.11}$$

Sur l'ensemble des individus, la vraisemblance classifiante s'écrit :

$$L_C(\mathbf{P},\boldsymbol{\pi},\boldsymbol{\beta},\boldsymbol{\sigma}) = \prod_{i=1}^{N}\prod_{k=1}^{K}\left(\pi_k f(\mathbf{y}_i,\boldsymbol{\mu}_{ik}(\boldsymbol{\beta}_k),\sigma_k)\right)^{z_{ik}} \quad (2.12)$$

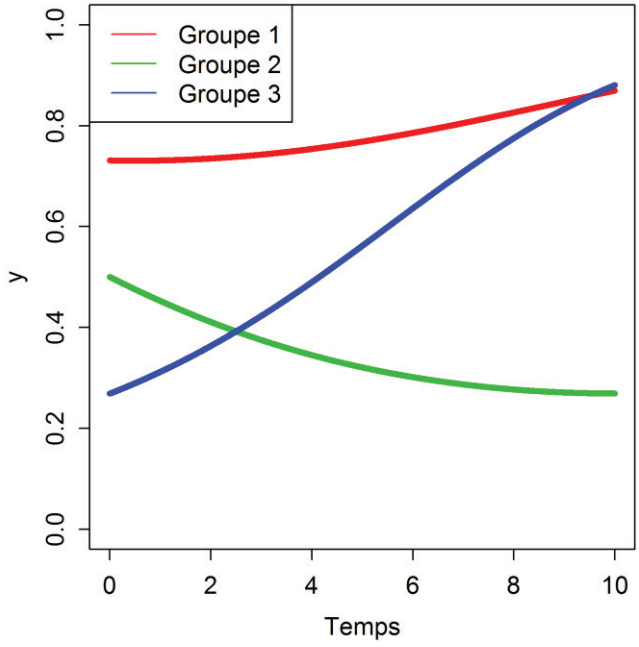
Les paramètres des trajectoires typiques $(\boldsymbol{\beta},\boldsymbol{\sigma})$ et la partition des individus dans les groupes sont estimés en maximisant la vraisemblance classifiante par l'algorithme CEM.

3.6.2 Simulations

Design

Des études de simulation ont été réalisées pour évaluer l'impact de la prise en compte de la sur-dispersion variable d'un groupe à l'autre sur la classification dans le modèle de classification de données binomiales. Le design de simulation correspond au design utilisé dans le cas des données gaussiennes (partie 3.3) : 320 trajectoires, 11 mesures répétées ($T_i = 11$), trois groupes avec les proportions de trajectoires dans chaque groupe suivantes : 22 %, 31 % et 47 %. Les données ont été simulées à partir d'une loi binomiale de paramètre $n_{it} = 5$ et d'espérance dépendant des trajectoires typiques. Les trajectoires typiques étaient définies par un polynôme du second ordre qui différait d'un groupe à l'autre par la valeur initiale et la pente (Figure 8).

Figure 8 : Trajectoires typiques des trois groupes pour l'étude de simulation



Quatre scénarios ont été utilisés en faisant varier l'écart entre les paramètres de dispersion des groupes.

100 jeux de données ont été simulés pour chaque scénario. Mod_Equal et Mod_Unequal ont été appliqués à chaque jeu de données. Une relation polynomiale de second ordre a été utilisée pour modéliser le lien entre le résultat **y** et le temps. L'algorithme CEM a été arrêté lorsque la différence du logarithme de vraisemblance classifiante après l'étape C entre deux itérations consécutives était inférieure à 0.001. L'algorithme a été répété 100 fois à partir de différentes partitions initiales, la solution retenue était celle qui conduisait à la vraisemblance classifiante la plus élevée. La performance du modèle a été évaluée par le pourcentage de trajectoires mal classées (MCR) moyen sur l'ensemble des jeux de données.

Résultats

Le modèle de classification avec une sur-dispersion variable d'un groupe à l'autre (Mod_Unequal) donne de meilleurs résultats de classification que celui avec une sur-dispersion supposée identique entre les groupes (Mod_Equal) lorsque la sur-dispersion est réellement différente d'un groupe à l'autre. L'écart entre les deux modèles exprimé en pourcentage d'individus mal classés peut atteindre 20 % (Tableau 3).

Tableau 3 : Pourcentages d'individus mal classés pour les deux modèles de classification

Scénario	S1	S2	S3	S4
	(0.2, 0.2, 0.2)*	(1.0, 1.4, 2.3)*	(1.4, 2.0, 3.3)*	(2.0, 3.0, 4.0)*
MCR (%) Mod_Equal	4.8	35.5	49.9	43.6
	(10, 2 , 4) †	(59, 25, 33) †	(64, 39, 51)	(64, 31, 42) †
MCR (%) Mod_Unequal	4.7	14.8	16.9	28.7
	(9, 2, 4) †	(31, 10, 11) †	(37, 11, 11) †	(85, 22, 19) †

† pourcentages d'individus mal classés dans chacun des groupes ; * paramètres de dispersion σ pour chacun des groupes

Pour S1, scénario pour lequel les paramètres de dispersion sont égaux dans les trois groupes, le MCR est similaire entre Mod_Equal et Mod_Unequal, indiquant qu'il n'y a pas de perte à utiliser Mod_Unequal même lorsque le paramètre de dispersion est égal pour tous les groupes. Pour S2 et S3 pour lesquels les paramètres de dispersion sont différents entre les groupes, le MCR de Mod_Equal est plus élevé que celui pour S1. Pour S2, le paramètre de

dispersion est estimé à 1.6 pour tous les groupes, et donc le paramètre de dispersion des groupes 1 et 2 est surestimé. En conséquence, le groupe 1 attire quelques trajectoires des autres groupes, surtout des trajectoires du groupe 3. Mod_Unequal corrige cette erreur de classification, en estimant correctement le paramètre de dispersion de chaque groupe. Pour S3, l'écart de MCR entre Mod_Unequal et Mod_Equal est légèrement plus élevé que pour S2, ce qui s'explique par le fait que l'écart entre les paramètres de dispersion des trois groupes était plus élevé. Pour S4, les paramètres de dispersion sont élevés par rapport aux distances entre trajectoires typiques ce qui créé beaucoup de chevauchement entre les groupes ; ainsi le bénéfice de Mod_Unequal par rapport à Mod_Equal est plus modeste que celui pour S2 et S3.

Ces résultats de simulation montrent l'intérêt du modèle de classification beta-binomial avec un paramètre de dispersion différent d'un groupe à l'autre. Ce modèle est d'autant plus pertinent que la sur-dispersion est très fréquente dans les études cliniques. Il sera néanmoins nécessaire de vérifier son intérêt en fonction du nombre d'individus et du nombre de mesures, comme cela a été fait dans le cas des données gaussiens.

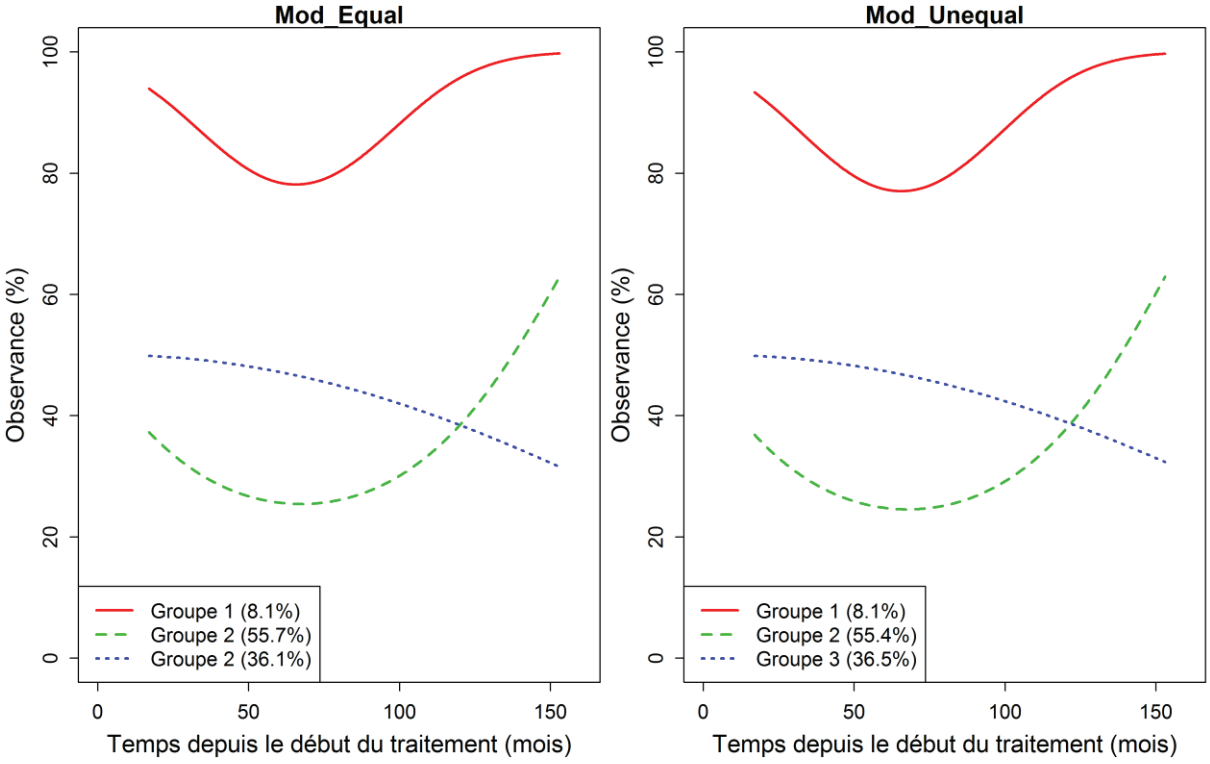
3.6.3 Application

L'étude « ISAARV » a été lancée au Sénégal pour fournir un traitement antirétroviral extrêmement actif aux patients séropositifs pour le VIH[6]. Dans le cadre de ce projet, 330 patients infectés par le VIH et traités par des antirétroviraux ont été suivis au moins tous les deux mois pour évaluer leur observance au traitement et déterminer les causes de non-observance. L'estimation de l'observance faite par le pharmacien est basée sur un décompte des comprimés retournés en présence du patient et complété par un questionnaire. Pour chaque médicament antiviral, l'observance a été calculée comme le rapport entre le nombre de comprimés pris et le nombre de comprimés prescrits.

Dans ce travail, seulement 296 patients sur les 330 patients de l'étude sont retenus pour l'analyse. En effet, 34 patients avaient à toutes les visites des données manquantes sur la date de visite ou sur le nombre de comprimés pris ou prescrit. La durée de suivi médian est de 75 mois (IQR 56-95). L'observance globale par mois est définie comme le rapport entre la somme du nombre de comprimés pris et la somme du nombre de comprimés prescrits sur l'ensemble des traitements antirétroviraux.

Les trajectoires de l'observance sont classées en trois groupes en utilisant les deux modèles Mod_Equal et Mod_Unequal ; un polynôme du second ordre sur le temps a été retenu pour modéliser les trajectoires typiques. Les trajectoires typiques obtenues avec les deux modèles, ainsi que les pourcentages de patients dans chaque groupe, sont présentés dans la Figure 9.

Figure 9 : Trajectoires typiques d'observance obtenues avec les deux modèles



Les effectifs et les paramètres de dispersion pour les trois groupes sont donnés dans le Tableau 4 :

Tableau 4 : Estimations des effectifs et des paramètres de dispersion des deux modèles

	Effectifs			Paramètres de dispersion		
Model	Gr. 1	Gr. 2	Gr. 3	Gr. 1	Gr. 2	Gr. 3
Mod_Equal	24	165	107	0.44	0.44	0.44
Mod_Unequal	24	164	108	0.51	0.34	0.48

Les trajectoires typiques obtenues par les deux modèles sont similaires, conduisant à (i) un groupe de patients avec une bonne observance au début du traitement (observance proche de 100 %), mais qui diminue légèrement avant de ré-augmenter à la fin, (ii) un groupe de patients avec une mauvaise observance au début du traitement (autour de 50 %), et qui continue à diminuer au cours du suivi, et (iii) un groupe de patients avec une mauvaise observance au début du traitement (autour de 40 %), qui diminue légèrement en début de suivi avant de ré-augmenter progressivement.

Dans le Tableau 4, les écarts entre les estimations des paramètres de dispersion obtenus par Mod_Unequal ne sont pas très élevés, ce qui explique que les classifications des patients des deux modèles sont proches (Tableau 5).

Tableau 5 : Croisement des classifications obtenues par les deux modèles

	Mod_unequal			
	Gr. 1	Gr. 2	Gr. 3	Total
Mod_equal				
Gr. 1	24	0	0	24
Gr. 2	0	160	5	165
Gr. 3	0	4	103	107
Total	24	164	108	

Les changements de groupes entre les deux modèles n'interviennent que pour les groupes 2 et 3, groupes dont les trajectoires typiques sont proches.

3.7 Perspectives

3.7.1 Intervalle de confiance des paramètres

Les intervalles de confiance des estimations de paramètres définissant les trajectoires typiques peuvent être obtenus à partir de la distribution des estimateurs du maximum de vraisemblance restreinte. Ces estimateurs sont supposés distribués asymptotiquement suivant une loi normale multivariée centrée sur les vraies valeurs des paramètres et de matrice de variance-covariance égale à l'inverse de la matrice d'information de Fisher (l'espérance négative de la matrice Hessienne de la log-vraisemblance restreinte) évaluée aux valeurs estimées des paramètres.

Cependant, cette stratégie ne tient pas compte de l'incertitude sur la classification obtenue lors de l'étape C de l'algorithme CEM. L'algorithme SEM peut être utilisé afin de décrire l'incertitude sur ces paramètres. Pour cela, à partir de la dernière itération de l'algorithme CEM, l'algorithme SEM est répété sur M itérations :

$$\begin{cases} \text{itération } 1: & \hat{\boldsymbol{\beta}}^{(1)}, \text{var}\left(\hat{\boldsymbol{\beta}}^{(1)}\right) \\ & \vdots \\ \text{itération } 2: & \hat{\boldsymbol{\beta}}^{(m)}, \text{var}\left(\hat{\boldsymbol{\beta}}^{(m)}\right) \\ & \vdots \\ \text{itération } M: & \hat{\boldsymbol{\beta}}^{(M)}, \text{var}\left(\hat{\boldsymbol{\beta}}^{(M)}\right) \end{cases} \tag{2.13}$$

A chaque itération m, les paramètres et leurs variances sont estimés (les variances sont estimées à partir de l'inverse de matrice d'information de Fisher).

En se basant sur la méthode de Rubin[34], la variance globale d'un paramètre sera calculée par la somme 1) de la moyenne des variances estimées des paramètres sur toutes les itérations SEM, et 2) de la variance des estimations des paramètres sur les itérations :

$$\text{var}(\hat{\boldsymbol{\beta}}) = \underbrace{\frac{1}{M}\sum_{m=1}^{M}\text{var}(\hat{\boldsymbol{\beta}}^{(m)})}_{\text{variance Intra}} + \underbrace{\frac{M+1}{M}\frac{\sum_{m=1}^{M}\left(\hat{\boldsymbol{\beta}}^{(m)} - moy(\hat{\boldsymbol{\beta}}^{(m)})\right)^2}{M-1}}_{\text{variance Inter}} \tag{2.14}$$

Des travaux supplémentaires sont nécessaires pour étudier et vérifier la validité de cette méthode, indépendamment de la structure de la variance résiduelle.

3.7.2 Modélisation conjointe

Il est très fréquent d'employer plusieurs biomarqueurs pour évaluer la condition de santé ou de maladie d'une personne. En effet, analyser conjointement plusieurs biomarqueurs permet d'une part de mieux comprendre le lien qui existe entre eux et d'autre part d'exploiter leur structure de dépendance. Roy et Lin[35] ont proposé un modèle à variable latente, appelé aussi modèle conjoint, permettant de modéliser conjointement plusieurs marqueurs dans le cas où les marqueurs sont des mesures de la même quantité non-observée. Par exemple, dans l'étude Lowcyclo, la fonction rénale peut être évaluée non seulement par la créatinine plasmatique, mais par d'autres biomarqueurs tels que la cystatine C et l'urée.

Dans le cadre des trajectoires de biomarqueurs hétérogènes, un modèle de mélange conjoint[36,37] a été développé pour analyser le lien entre les trajectoires des différent marqueurs. Comme déjà mentionné dans la partie 2.2.2.1, la particularité des modèles de mélange de façon globale est qu'ils ne classent pas les individus pour déterminer les profils typiques. Il sera donc intéressant de développer dans un premier temps un modèle de classification conjoint (qui prend en compte plusieurs biomarqueurs), permettant d'identifier K groupes des sujets ayant des trajectoires similaires sur chacun de marqueur, et ensuite d'étendre ce modèle en prenant en compte une variance résiduelle différente d'un groupe à l'autre.

Dans un modèle de classification conjoint, le principe sera la même que le modèle de mélange conjoint : les distributions de *J* biomarqueurs sont supposées être indépendantes conditionnement à l'appartenance au groupe, c'est-à-dire

$$f(\mathbf{y}_i^1, \mathbf{y}_i^2, .., \mathbf{y}_i^J) = f(\mathbf{y}_i^1) f(\mathbf{y}_i^2) ... f(\mathbf{y}_i^J) \tag{2.15}$$

avec $\mathbf{y}_i^j$ le vecteur de réponse des T_i mesures successives pour un individu *i* du $j^{\text{ème}}$ biomarqueur, et $f(\mathbf{y}_i^j)$ la distribution de ce vecteur conditionnellement à l'appartenance au groupe.

La distribution $f\left(\mathbf{y}_i^j, \boldsymbol{\beta}_k^j, \sigma_k^j\right)$ est calculée en considérant que les mesures de chaque trajectoire de chaque biomarqueur *j* sont indépendantes conditionnellement à l'appartenance au groupe. Par conséquence, $f\left(\mathbf{y}_i^j, \boldsymbol{\beta}_k^j, \sigma_k^j\right) = \prod_{t=1}^{T_i} f\left(y_{it}^j, \boldsymbol{\beta}_k^j, \sigma_k^j\right)$.

Ainsi, la vraisemblance des trajectoires de biomarqueurs pour un individu *i* s'écrit de la façon suivante :

$$f(\mathbf{y}_i^1, \mathbf{y}_i^2, .., \mathbf{y}_i^J) = \prod_{k=1}^{K} \pi_k \prod_{j=1}^{J} \left(f(\mathbf{y}_i^j, \boldsymbol{\beta}_k^j, \sigma_k^j) \right)^{z_{ik}} \tag{2.16}$$

En se basant sur cette vraisemblance (2.16), un modèle de classification conjoint peut être implémenté avec une variance résiduelle σ_k^2 soit variable d'un groupe à l'autre, soit identique d'un groupe à l'autre.

Il serait intéressant d'implémenter ce modèle, et d'en évaluer l'intérêt dans le cadre d'applications concrètes.

4 Partie III : Prise en compte d'une variance inter-individuelle

4.1 Variabilité inter-individuelle des trajectoires

Dans le modèle de classification décrit dans la deuxième partie, il a été supposé qu'il n'y a pas de variabilité inter-individuelle au sein de chaque groupe. La trajectoire prédite pour tous les individus d'un même groupe est la même. Cela signifie que les différences entre les trajectoires individuelles observées au sein d'un groupe sont considérées comme des fluctuations résiduelles. L'hypothèse de trajectoire unique au sein d'un groupe est réductrice, car il est peu probable que la variabilité inter-individuelle soit nulle. Le modèle de classification standard est donc inapproprié sur le plan conceptuel, ce qui peut fausser la classification, mais il l'est également en termes d'inférence. Comme pour le modèle de mélange décrit dans la partie 2.2.2.1, ce modèle suppose que, conditionnellement à l'appartenance au groupe, les mesures d'un même individu sont indépendantes les unes des autres, ce qui n'est pas le cas ; ceci peut fausser les erreurs standards des paramètres estimés (et pour certains modèles, les estimations ponctuelles en elles-mêmes).

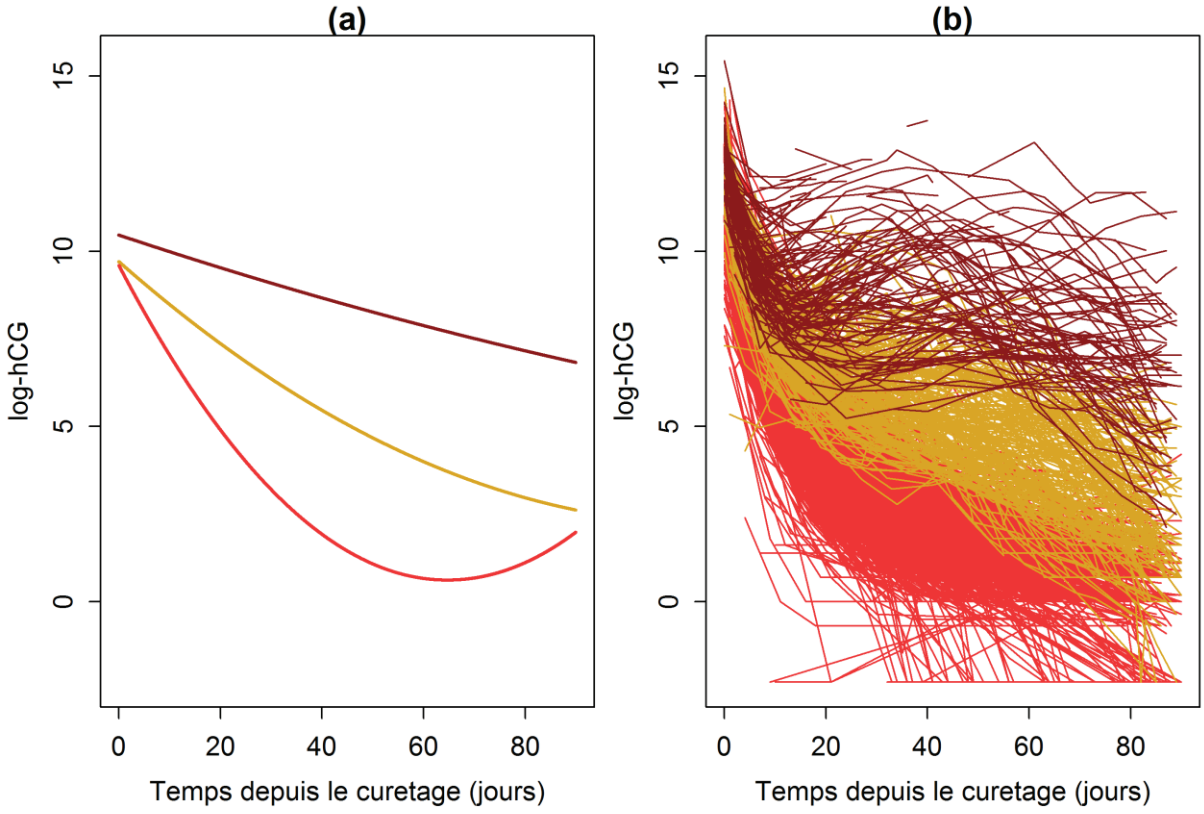
En présence de variabilité inter-individuelle au sein d'un groupe, la variance résiduelle du modèle de classification standard intègre la vraie variance résiduelle et la variance inter-individuelle ; elle est donc surestimée. Au-delà des estimations des paramètres, ceci peut également fausser la classification obtenue.

Sur les données de l'étude « Môle hydatiforme », l'analyse des trajectoires d'hCG a été effectuée par un modèle de classification de trajectoires standard en considérant trois groupes. Les groupes de trajectoires identifiés sont représentés sur la Figure 10 (a) :

- un groupe de patients dont le niveau d'hCG est élevé lors de curetage, puis diminue légèrement au cours du temps jusqu'à stabilisation, conduisant à des valeurs élevées d'hCG ;

- un groupe de patients dont le niveau d'hCG est élevé lors du curetage, bien que plus faible que celui du groupe 1, puis diminue au cours du temps, conduisant à des valeurs intermédiaires d'hCG.
- un groupe de patients dont le niveau d'hCG est élevé lors de curetage, similaire au groupe 2, puis diminue rapidement au cours du temps, conduisant à des valeurs très faibles à la fin du suivi.

Figure 10 : Trajectoires typiques d'hCG et classification des trajectoires observées obtenue par un modèle de classification standard



Ces analyses reposaient sur un modèle de classification de trajectoires standard, qui suppose qu'il n'y a pas de variance inter-individuelle au sein de chaque groupe. Néanmoins, il semble qu'il y ait une très grande variabilité des trajectoires individuelles au sein de chaque groupe (Figure 10b). Par exemple, dans le groupe 1, certaines trajectoires d'hCG diminuent après le curetage puis restent stable, alors que d'autres diminuent après le curetage, mais cette diminution est suivie de deux périodes : une période avec une augmentation des valeurs d'hCG, et une deuxième avec une diminution pour atteindre des valeurs d'hCG intermédiaires. Ainsi, la variabilité entre patientes au sein des groupes est manifeste. Ces

différences ont une signification clinique, une augmentation des valeurs d'hCG étant évocatrice d'une récidive avec possible cancérisation de la môle.

L'objectif de ce deuxième chapitre est d'implémenter un modèle de classification prenant en compte la variabilité inter-individuelle au sein de chaque groupe, et d'évaluer l'impact de cette prise en compte sur la classification à l'aide de données simulées et sur l'exemple de la môle hydatiforme.

4.2 Le modèle

En reprenant quelques notations précédemment introduites, si la trajectoire $\mathbf{y}_i$ de l'individu *i* appartient au groupe *k*, le modèle de classification qui prend en compte la variance inter-individuelle au sein de chaque groupe, appelé modèle de classification à effets mixtes, s'écrit :

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta}_k + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i \tag{2.17}$$

$\mathbf{Z}_i$ est la matrice de taille $T_i \times q$ correspondant aux q covariables supposées avoir un effet aléatoire ($\mathbf{b}_i$). Ces effets aléatoires sont supposés distribués suivant une loi normale multivariée de moyenne 0 et de matrice variance-covariance $\mathbf{D}_k$**.** Cette matrice $\mathbf{D}_k$ peut être variable d'un groupe à l'autre, ou identique (c'est-à-dire, $\mathbf{D}_1 = \mathbf{D}_2 = ... = \mathbf{D}_K$). Ce modèle suppose donc que le vecteur des réponses pour un individu *i* est une variable gaussienne multivariée dont la matrice de variance-covariance est $\mathbf{V}_i$:

$$\mathbf{y}_i \sim N(\mathbf{X}_i\boldsymbol{\beta}_k, \mathbf{V}_i = \mathbf{Z}_i\mathbf{D}_k\mathbf{Z}_i^t + \sigma^2\mathbf{I}_{T_i})$$

Les méthodes d'estimation des paramètres de ce modèle et algorithmes associés, ainsi que sa comparaison par rapport au modèle de classification standard sont décrites dans l'article rédigé pour la revue « Statistics in Medicine ».

4.3 Article rédigé pour la revue Statistics in Medicine

Trajectory clustering using mixed classification models

Short title: Trajectory clustering with mixed classification model

Amna Klich[1,2], René Ecochard[1,2], Fabien Subtil[1,2]

[1] Service de Biostatistique-Bioinformatique, Pôle Santé Publique, Hospices Civils de Lyon, Lyon, France;

[2] Université de Lyon, Université Lyon 1, CNRS, Laboratoire de Biométrie et Biologie Évolutive UMR 5558, Villeurbanne, France;

Author to whom correspondence should be addressed:

Amna Klich

Service de Biostatistique-Bioinformatique

Hospices Civils de Lyon

162, avenue Lacassagne - F-69003, Lyon, France.

E-mail: amna.klich@chu-lyon.fr

Abstract

Trajectory classification has become frequent in clinical research to understand and explore the heterogeneity of trajectory among subjects. The standard classification model assumes no between-individual variance within groups. However, this assumption is often not appropriate, and in this case the error variance of the model may be overestimated, leading to a biased classification. Hence, two extensions of the standard classification model for trajectories were developed through a mixed model. A first one considers an equal between-individual variance for all groups, and a second one considers unequal between-individual variance between groups. Simulations were performed to evaluate the impact of these considerations on the obtained classification.

The simulation results showed that the first extended model gives a lower misclassification percentage (with differences from 15% to 50%) than the standard classification one in case of a true variance between individuals inside groups. The second extended model decreases the misclassification percentage compared to the first one (up to 11%) when the between-individual variance is unequal between groups, but this second model, to be adjusted, requires a higher number of trajectories than the first one.

Using hCG trajectories after a curettage for hydatidiform mole from a clinical cohort, the standard classification model mainly classified trajectories according to their level whereas the two extended models classified them according to their pattern, leading to more clinically relevant groups.

In conclusion, for studies with a non-negligible number of individual trajectories, the use, in first instance, of a classification model that considers equal between-individual variance for all groups rather than a standard classification model, appears more appropriate. A model that considers unequal between-individual variance may find its place thereafter.

Keywords: classification; longitudinal data; between-individual variance; mixed model; ECM algorithm; trajectories

1 Introduction

Unsupervised classification involving a biomarker with repeated measurements allows gathering of biomarker trajectories into a small number of groups of similar trajectories, without any prior knowledge of these groups. It is used to understand and explore the heterogeneity of biomarker trajectory among patients.

Different methods have been proposed for trajectory classification. For finite mixture models,[1–5] each individual trajectory is modelled by a mixture of the different typical trajectories weighted by the posterior probabilities of belonging to them. The emphasis is on the estimation of the parameters of the model rather than on the classification: the main objective is not to classify trajectories, but to make valid inference in a context of heterogeneity. The trajectory predicted by the model, for a given individual, corresponds to a mixture of the typical trajectory of the different groups. A classification of the trajectories may still be obtained a posteriori, by allocating each individual trajectory to the group with the highest posterior probability. The classification model[6–9] is an alternative to the mixture model. Its main objective is to classify trajectories into groups, the partition of the trajectories into the groups being a parameter of the model. Each individual trajectory therefore belongs to one and only one group, and the trajectory predicted by the model for a given individual is the typical trajectory of the group to which he/she belongs. These two models (mixture and classification) do not necessarily give the same results.[10,11]

The standard classification model – hereafter referred to as the "fixed effects classification model" – assumes that all individual trajectories are homogenous within a group, i.e. there is no between-individual variability within groups. This means that the differences between individual trajectories within a group are considered as error variance, i.e. the random variability which is unexplained by the model. However, in life sciences, heterogeneity is much more common than homogeneity. Without accounting for the

variability within a group, the error variance of the model which combines the true error variance and the between-individual variance may thus be overestimated, leading to a biased classification. A possible solution is to take into account, in the standard classification model, the between-individual variance within each group to allow an individual trajectory to deviate from the typical trajectory of its group. Concerning mixture models, many authors[12–15] have proposed to extend the standard mixture model by supplementing it with a mixed model. [16] In this extended model, because trajectories are not classified in groups during the estimation of the parameters, the variance-covariance matrix associated with the random effects does not reflect the true variability of individuals in a group. This is contrary to what occurs when using classification models.

The integration of mixed models into classification models is not well developed. Moreover, whatever the context (mixture or classification model), the interest of adding random-effects has not yet been analysed.

This paper presents two extensions of the standard classification model called herein "mixed effects classification models" which take into account the between-individual variance within each group using a mixed model. In these extended models, the first one assumes an equal between-individual variance for all groups, and the second one considers unequal between-individual variance between groups. Simulations were then performed to evaluate the impact of assuming no between-individual variance vs. considering between-individual variance within groups (with equal or unequal between-individual variance between groups).

The models were then applied to a clinical cohort of women followed for changes in hCG levels after a curettage for hydatidiform mole. The classifications and typical trajectories obtained using the different models were compared.

2 The mixed effects classification model

2.1 The model

Three classification models are defined herein: the standard one (the fixed effects classification model), and its two extensions (the mixed effects classification models). Let $\mathbf{y}_i = (y_{i1}, y_{i2}, ..., y_{iT_i})$ be the response vector of the T_i successive measurements in individual *i,* with $i = 1, .., N$, y_{it} corresponding to the t^{th} measurement of individual *i*. If the i^{th} individual trajectory belongs to the k^{th} group, the fixed effects classification model (called hereafter Mod_Fix), which does not allow for between-individual variability within groups, may be written as:

$$\begin{aligned} &\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta}_k + \boldsymbol{\varepsilon}_i, \qquad i = 1, ..., N, \\ &\boldsymbol{\varepsilon}_i \sim N(0, \sigma^2\mathbf{I}_{T_i}) \end{aligned} \tag{1}$$

which gives, given the group membership, the following conditional density

$$f(\mathbf{y}_i \,|\, \mathbf{X}_i, \boldsymbol{\beta}_k, \sigma^2) \sim N(\mathbf{X}_i\boldsymbol{\beta}_k, \sigma^2\mathbf{I}_{T_i}) \tag{2}$$

where $\mathbf{X}_i$ is the $T_i \times p$ design matrix with p columns of variables, including: a time function (such as a polynomial function of time or a natural cubic spline), characteristics of subjects, and potential interactions between time and characteristics. $\boldsymbol{\beta}_k$ represent the parameters of the model for the k^{th} group. The errors $\boldsymbol{\varepsilon}_i$ are assumed to be normally distributed with mean zero and variance σ^2 equal for all groups.

Let $z_{ik} = 1$ when the i^{th} individual trajectory belongs to the k^{th} group (0 otherwise). $\boldsymbol{\theta}_k$ is the whole parameter vector of the model for the k^{th} group: $(\boldsymbol{\beta}_k, \sigma^2)$. The classification

likelihood[10] is calculated by assuming that the measurements of each individual are independent given the group membership.[17] Thus, $f(\mathbf{y}_i|\boldsymbol{\theta}_k)=\prod_{t=1}^{T_i} f(y_{it}|\boldsymbol{\theta}_k)$

Over all individuals, the classification likelihood is given by:

$$L(\mathbf{P},\boldsymbol{\pi},\boldsymbol{\theta})=\prod_{i=1}^{N}\prod_{k=1}^{K}\left(\pi_k f(\mathbf{y}_i|\boldsymbol{\theta}_k)\right)^{z_{ik}} \tag{3}$$

where $\mathbf{P}$ is the partition of $(\mathbf{y}_1,\mathbf{y}_2,..,\mathbf{y}_N)$ defined by z_{ik} and $\boldsymbol{\pi}=(\pi_1,..,\pi_K)$ is the vector of the proportion of individual trajectories, called "prior probabilities" ($0\le\pi_k\le 1$ for $k = 1,..,K$ with $\sum_{k=1}^{K}\pi_k=1$).

As stated previously, in the model defined by (1), the errors for the same individual, given group membership, are assumed independent. But, this independence hypothesis is unlikely for trajectories. One way to deal with the non-independence of individual errors is to alter the model by introducing random effects. Moreover, the use of random effects allows an individual trajectory to deviate the typical trajectory of its group, thus accounting for the between-individual variance within groups. This model is the mixed effects classification model. If the i^{th} individual trajectory belongs to the k^{th} group, the model may be written:

$$\begin{aligned}&\mathbf{y}_i=\mathbf{X}_i\boldsymbol{\beta}_k+\mathbf{Z}_i\mathbf{b}_{ik}+\boldsymbol{\varepsilon}_i,\quad i=1,...,N,\\&\boldsymbol{\varepsilon}_i\sim N(0,\sigma^2\mathbf{I}_{T_i}),\\&\mathbf{b}_{ik}\sim N(0,\mathbf{D}_k)\end{aligned} \tag{4}$$

where $\mathbf{Z}_i$ is the design matrix with q columns of covariates which effects are assumed to be random across individuals. The random effects $\mathbf{b}_{ik}$ are assumed to be normally distributed with mean zero and variance-covariance matrix $\mathbf{D}_k$. Hereafter, the model obtained will be

called Mod_Mix_unequal. When the variance-covariance matrix $\mathbf{D}_k$ is constrained to be equal across groups (i.e. $\mathbf{D}_1 = \mathbf{D}_2 = ... = \mathbf{D}_K$), the model is called Mod_Mix_equal.

Given the group membership, the marginal mean and variance-covariance matrix of $\mathbf{y}_i$ are given as follows:

$$\begin{aligned} E(\mathbf{y}_i \mid i \in k) &= \mathbf{X}_i \boldsymbol{\beta}_k \\ \text{var}(\mathbf{y}_i \mid i \in k) &= \mathbf{Z}_i \mathbf{D}_k \mathbf{Z}_i^{'} + \sigma^2 \mathbf{I}_{T_i} \end{aligned} \tag{5}$$

which implies the following marginal normal distribution, conditional on group membership:

$$f(\mathbf{y}_i \mid \mathbf{X}_i, \mathbf{Z}_i, \boldsymbol{\beta}_k, \mathbf{D}_k, \sigma^2) \sim N(\mathbf{X}_i \boldsymbol{\beta}_k, \mathbf{Z}_i \mathbf{D}_k \mathbf{Z}_i^{'} + \sigma^2 \mathbf{I}_{T_i}) \tag{6}$$

The classification likelihood of this mixed effects classification model is close to the one defined for the fixed effects classification model described in equation (3), with a change in the distribution conditional to the group membership, and a change in the $\boldsymbol{\theta}_k$ parameters which are now $((\boldsymbol{\beta}_k)_{k=1,..,K}, (\mathbf{D}_k)_{k=1,..,K}, \sigma^2)$.

The parameter vectors for the models $(\mathbf{P}, \boldsymbol{\pi}, \boldsymbol{\theta})$ are estimated by maximising the log classification likelihood using the CEM algorithm.

The choice of the number of groups is based on two factors; (i) a statistical criterion, the integrated classification likelihood-Bayesian information criterion (ICL-BIC),[18,19] which evaluates the adequacy of the models with the data, and (ii) a clinical expertise, so that the groups obtained have a meaningful clinical interpretation.

2.2 The CEM algorithm

The CEM algorithm[20] was used to estimate $\boldsymbol{\pi}$, $\boldsymbol{\theta}$, and the partition $\mathbf{P}$. It incorporates a classification step between the E-step and the M-step of the classical EM algorithm.

Starting from an initial partition of the trajectories, the m^{th} iteration of the CEM algorithm is defined as follows:

The E-step computes the posterior probability of belonging to group k for each individual trajectory i given the parameter values estimated at the previous iteration:

$$\begin{aligned} & post.prob_{ik}^{(m)} = P(i \in k \mid \mathbf{y}_i, \boldsymbol{\pi}^{(m)}, \boldsymbol{\theta}^{(m)}) \\ & \propto \pi_k^{(m)} \int_{\mathbf{b}_{ik}} f(\mathbf{y}_i \mid \boldsymbol{\beta}_k^{(m)}, \mathbf{b}_{ik}, \sigma^{2\,(m)}) f(\mathbf{b}_{ik} \mid \mathbf{D}_k^{(m)}) d\mathbf{b}_{ik} \\ & \propto \pi_k^{(m)} f(\mathbf{y}_i \mid \boldsymbol{\beta}_k^{(m)}, \mathbf{D}_k^{(m)}, \sigma^{2\,(m)}) \end{aligned} \quad (7)$$

For the mixed effects classification model, the posterior probability is given by:

$$post.prob_{ik}^{(m)} \propto \pi_k^{(m)} \frac{1}{(2\pi)^{\frac{T_i}{2}} \sqrt{\det(\mathbf{V}_{ik}^{(m)})}} \exp\left(-\frac{1}{2} \left(y_{it} - \mathbf{X}_{it} \boldsymbol{\beta}_k^{(m)} \right)^t \left(\mathbf{V}_{ik}^{(m)} \right)^{-1} \left(y_{it} - \mathbf{X}_{it} \boldsymbol{\beta}_k^{(m)} \right) \right) \quad (8)$$

with $\mathbf{V}_{ik} = \mathbf{Z}_i \mathbf{D}_k \mathbf{Z}_i^{'} + \sigma^2 \mathbf{I}_{T_i}$

The C-step assigns each trajectory i to the group that provides the maximum value of $post.prob_{ik}^{(m)}$. Let $\mathbf{P}^{(m)}$ denote the resulting partition.

The M-step computes, for each $k = 1,..,K$, the estimate $(\pi_k^{(m+1)}, \boldsymbol{\theta}_k^{(m+1)})$ that maximises the log classification likelihood given $\mathbf{P}_k^{(m)}$.

In the log classification likelihood, π_k is independent from the other parameters and the maximum-likelihood estimate is given explicitly by: $\pi_k^{(m+1)} = \frac{\#\mathbf{P}_k^{(m)}}{N}$. The estimates of $\boldsymbol{\theta}_k$ were obtained by ordinary least squares in Mod_Fix and by maximisation of the marginal restricted profiled log-likelihood[21] in Mod_Mix_equal.

The CEM algorithm was stopped when the difference in the log classification likelihood $L(\mathbf{P}^{(m)}, \pi_k^{(m)}, \boldsymbol{\theta}_k^{(m)})$, after the C-step of two consecutive iterations, was less than a given threshold value or when there was no change in the partition. The solution provided by the CEM algorithm depends on the initial partition.[20] Consequently, the CEM algorithm was

repeated several times from different initial partitions and the classification that provided the highest value of log classification likelihood was kept.

To estimate the parameters of the typical trajectories and the partition of the trajectory classification model using the CEM algorithm, three generic functions were created in R. A main function performs the three steps of the CEM algorithm, given an initial partition. A second function is used inside the main function which permits to estimate the parameters in the M-step, using the *lme* function of the *nlme* package. Another function repeats the CEM algorithm (main function) several times from different initial partitions and provides the best solution.

3 Simulations

3.1 Designs

The general design of the simulations was defined as follows: 320 trajectories (N=320), 11 repeated measurements (T=11), 3 groups with proportion of trajectories in each group given by 22%, 31%, and 47%. The typical trajectories of the 3 groups were defined by second-order polynomials shown in Figure 1. In each group, the individual trajectories were simulated by taking into account a random intercept and a random slope centred on the coefficients of the group $\boldsymbol{\beta}_k$ with variance-covariance matrix $\mathbf{D}_k$, and by considering a residual variance of σ^2.

A total of 7 simulation scenarios were defined by changing $\mathbf{D}_k$. The correlation between the intercept and slope was fixed to 0.5 whatever the scenario. At first, 4 scenarios (S1 to S4) were defined considering an equal variance-covariance matrix $\mathbf{D}$ for all groups. This matrix was chosen according to the ratio of the between-individual variance ($Tr(\mathbf{Z}_i\mathbf{D}_k\mathbf{Z}_i^{'})$) over the total variance within groups ($Tr(\mathbf{Z}_i\mathbf{D}_k\mathbf{Z}_i^{'}+\sigma^2\mathbf{I}_{T_i})$). S1 is a scenario

with a very low ratio, and S2 to S4 are scenarios with increasing ratios of between-individual variance over total variance (Table 1). Then, another 3 scenarios (S5 to S7) were established to consider an unequal variance-covariance matrix $\mathbf{D}_k$ between groups. S5 is a scenario with a very low ratio of between-individual variance over total variance for all groups, resulting in well-separated groups. Scenario S6 considers a very high ratio of between-individual variance over total variance for all groups, resulting in 3 overlapped groups. S7 is a scenario with a high ratio of between-individual variance over total variance for all groups, but the variance-covariance matrices were defined such as the overlap between groups 2 and 3 was smaller than the overlap for scenario S6 (Table 2).

For the two mixed classification models (Mod_Mix_equal and Mod_Mix_unequal), the intercept and slope were considered as random effects, and the second-order polynomial as fixed effects.

The CEM algorithm was stopped when the absolute difference in the log classification likelihood after the C-step of two consecutive iterations was less than 0.001. It was repeated 100 times from different initial partitions.

For each scenario, 1000 datasets were simulated. The Mod_Fix and Mod_Mix_equal were compared for scenarios S1 to S4. The Mod_Mix_equal and Mod_Mix_unequal were compared for scenarios S3 and S5 to S7.

The performance of the models was assessed by the overall misclassification rate (MCR) and the bias (mean difference between the estimated and the true typical trajectories).

3.2 Results

The MCR of Mod_Mix_equal and Mod_Fix increased with the ratio of between-individual variance over total variance (Table 1). However, the MCR was smaller with Mod_Mix_equal than with Mod_Fix for scenarios S2-S4. The difference in MCR between the

two models increased as the ratio of between-individual variance over total variance increased, sometimes reaching about 50% difference between the two models. For example, in scenarios S3 and S4, the trajectories of the 3 groups overlapped very much due to the high between-individual variance over total variance ratio, mainly due to the slope variance. Mod_Fix was unable to differentiate the groups, leading to a very high MCR (around 60%), but Mod_Mix_equal corrected this misclassification. In scenario S1, in which the groups were well-separated due to the very low ratio of the between-individual variance over total variance, both models were able to classify correctly the trajectories. Hence, there is no loss in using Mod_Mix _homo even when there is no between-individual variance.

For scenarios with an unequal variance-covariance matrix between groups, the MCR was similar between Mod_Mix_equal and Mod_Mix_unequal for scenarios S3 and S5 (Table 2). For scenario S3, the variance-covariance matrix was equal for the 3 groups, demonstrating that there is no loss in using Mod_Mix_unequal even with an equal variance-covariance matrix. In scenario S5, the groups were well-separated even when assuming wrongly an equal variance-covariance matrix for the three groups due to the very low ratio of the between-individual variance over total variance used for each group. Consequently, both models were able to classify correctly the trajectories.

In S6 and S7, the MCR for Mod_Mix_equal were higher (29.9 and 22.9% respectively) compared to Mod_Mix_unequal (26.33 and 11.93, respectively). The Mod_Mix_unequal decreases this misclassification as it estimates appropriately the variance-covariance matrix. In S6, groups 2 and 3 had high between-individual variances relative to the distances between typical trajectories; the groups overlapped, and neither model was able to correctly classify the trajectories. Conversely, in S7, there was less overlap between groups 2 and 3, and Mod_Mix_unequal was able to identify correctly the 3rd group, leading to a lower

MCR than in S6. The bias in the estimated typical trajectories followed that of MCR, i.e. bias was high when MCR was high and conversely (Supplementary Table S1 and table S2).

4 Application

4.1 Context

A prospective study involved women registered to the French Trophoblastic Disease Reference Centre (TDRC, Lyon, France), from the 1st of January 2010 to the 31st of December 2012, who underwent curettage for hydatidiform mole. [22] After curettage, the women were followed with weekly measurements of total hCG until undetectable levels, then every 2-3 weeks for partial moles, or every month, during 6 months, for complete moles.

The objective was to identify groups of hCG trajectories that make sense from a biological point of view, and to present the changes in classification and typical trajectories according to the assumptions made concerning the between-individual variance within groups. This analysis was restricted to the 1440 women who had at least two hCG measurements, and only took into account hCG measurements performed before a potential second curettage, in the case of patients with more than one curettage. The median follow-up was 6 months, with a mean of 8 measurements per women. Individual trajectories of the logarithm-hCG after mole curettage are presented in Figure 2. Trajectories are characterized by an initial quick decline after curettage, followed by a stabilisation phase, sometimes showing an increase after.

4.2 Classification models

The 1440 patients were classified using Mod_Fix, Mod_Mix_equal and Mod_Mix_unequal. A second-order polynomial evolution was considered for the log-hCG measurements. For mixed models, only the intercept and slope were considered as random for

identifiability purposes. If the i^{th} patient trajectory belongs to the k^{th} group, Mod_Fix may be written:

$$\mathbf{log-hCG}_i = \beta_{0k} + \beta_{1k}\mathbf{Time}_i + \beta_{2k}(\mathbf{Time}_i^2) + \boldsymbol{\varepsilon}_i$$

The two models Mod_Mix_equal and Mod_Mix_unequal may be written:

$$\mathbf{log\text{-}hCG}_i = \beta_{0k} + \beta_{1k}\mathbf{Time}_i + \beta_{2k}(\mathbf{Time}_i^2) + b_{0ik} + b_{1ik}\mathbf{Time}_i + \boldsymbol{\varepsilon}_i$$

where $(\mathbf{b_{0k}}, \mathbf{b_{1k}}) \sim N\left(\begin{bmatrix}0\\0\end{bmatrix}, \mathbf{D}_k = \begin{bmatrix}\sigma_{0k}^2 & \sigma_{01k}\\ \sigma_{01k} & \sigma_{1k}^2\end{bmatrix}\right)$

Similarly to simulations, the CEM algorithm was repeated 100 times from different initial partitions of the trajectories. Classifications with 2, 3, or 4 groups were considered.

4.3 Results

Table 3 gives the ICL-BIC values according to different numbers of groups and models. From a statistical point of view, 4 groups seem to give the best classification in Mod_Fix and Mod_Mix_equal, and 3 groups in Mod_Mix_unequal. However, groups 3 and 4 had similar typical trajectories according to Mod_Fix and Mod_Mix_equal. Hence, a 3-group classification was kept whatever the model.

The classification and typical trajectories obtained using the 3 models as well as the percentages of patients in each group are shown in Figure 3 and Figure 4. The classification and the typical trajectories were slightly different between the 3 models especially for group 1. Group 1 had different trajectories according to Mod_Fix (2 periods) and Mod_Mix_equal (3 periods). In both models group 1 was first characterised by a short decrease in hCG. However, the second period according to Mod_Fix was relatively stable while Mod_Mix_equal characterised 2 following periods: a first one with stable or increasing hCG

levels, and a second one with decreasing hCG levels. Group 2 was characterised by a rather lower initial hCG value compared to group 1 in Mod_Fix, followed by a slow decrease leading to intermediate hCG values. In Mod_Mix_equal, group 2 trajectories show 2 periods: a first quick decrease followed either by stabilisation or slow decrease, leading at the end of the follow-up to either high or intermediate values of hCG. Group 3 for Mod_Fix corresponds to trajectories with a very quick decrease, leading to very low hCG values at the end of the follow-up. For Mod_Mix_equal, this group corresponds to trajectories with 3 periods: a short period with very quick decrease in hCG, followed by a period with intermediate decrease, and a last period with slow decrease or stabilisation.

43 individual trajectories switched from group 2 with Mod_Fix to group 1 with Mod_Mix_equal (Table 4, Figure 4): they had a low initial hCG value, explaining their classification into group 2 by Mod_Fix. However, the random intercept in Mod_Mix_equal concealed this gap at the beginning, and the trajectories were classified in group 1 to reflect the decrease following the period of increase or stabilization.

65 individual trajectories switched from group 1 with Mod_Fix to group 2 with Mod_Mix_equal (Figure 4): they had a low slope after the first quick decline compared to the slope of trajectories of group 2 with Mod_Fix, but the random slope allows to classify them in a group with only one period of evolution after the first quick decline.

62 individual trajectories switched from group 2 with Mod_Fix to group 3 with Mod_Mix_equal (Figure 4): they had a rather high initial hCG value and ended with intermediate hCG values. Overall, they correspond to trajectories with intermediate hCG values, and were classified in group 2 with Mod_Fix. Due to the random intercept and slope, Mod_Mix_equal allowed trajectories with intermediate hCG values in group 3 and identified within these 62 trajectories, a pattern with 3 periods of decrease.

51 individual trajectories switched from group 3 with Mod_Fix to group 2 with Mod_Mix_equal (Figure 4): they correspond to trajectories with low initial hCG values and very low values at the end of the follow-up, and are consequently classified in group 3 with Mod_Fix. Mod_Mix_equal identifies trajectories with only 2 periods of decrease, and classified them in group 2; the low initial hCG level is compensated by the random intercept.

Globally, Mod_Fix mainly classifies trajectories according to their level, whereas Mod_Mix_equal classifies them according to their pattern (more specifically, the number of periods of decrease).

The typical trajectories obtained with Mod_Mix_unequal were similar to the ones obtained with Mod_Mix_equal. There was a switch of trajectories from group 2 with Mod_Mix_equal to group 1 with Mod_Mix_unequal, due to the higher variance of the random intercept estimated by Mod_Mix_unequal that helped to attract some trajectories from group 2. There was also a switch of trajectories from group 3 with Mod_Mix_equal to group 2 with Mod_Mix_unequal, explained by the higher variance of the random slope estimated by Mod_Mix_unequal for group 2 that helped again to attract some trajectories from group 3.

5 Discussion

The present work presents two extensions of the standard classification model for trajectories that take into account the between-individual variance within each group through the use of a mixed model. The simulation results showed that the first extension – assuming an equal between-individual variance between groups, Mod_Mix_equal – gives a lower misclassification percentage (with differences from 15% to 50%) than the standard classification one in case of a non-negligible ratio of the between-individual variance over the total variance within groups (between-individual and residual variance). Furthermore, there is

no loss in using this model when there is no between-individual variance, even if this is an unlikely case. Therefore, the use of Mod_Mix_equal should be recommended over Mod_Fix. The second extension, Mod_Mix_unequal, that allows a unequal between-individual variance between groups, decreases the misclassification percentage compared to Mod_Mix_equal when the variance-covariance matrix was unequal between groups (up to 11% in the simulations). The magnitude of the correction increases when (i) the difference in between-individual variances between groups increases relative to the overall between-individual variance, (ii) the overall between-individual variance increases relative to the residual variance and (iii) the typical trajectories are well separated. Moreover, there is no loss in using Mod_Mix_unequal when the between-individual variance is equal for all groups.

The contribution of Mod_Mix_unequal over Mod_Mix_equal is modest in the simulation study. Moreover, the simulations were performed in scenarios with a high number of trajectories and measurements per trajectory, a necessary condition for the convergence of this model. Hence, there is no obvious recommendation for Mod_Mix_unequal. For Mod_Mix_equal, the benefit over the standard classification model is higher. Moreover, in a clinical study, there is always between-individual variability within a group. The implementation of Mod_Mix_equal in a statistical software is easy since it allows for mixed models; the formulae for the posterior probabilities is given in the present paper. Mod_Mix_equal should generally be recommended for trajectory classification. The simulation study did not explore the impact of the number of trajectories on the respective performance of Mod_Mix_equal and standard classification models. It is expected that the convergence for Mod_Mix_equal requires a higher number of trajectories than for Mox_Fix. More simulations should thus be performed to investigate this point.

The solution given by the CEM algorithm used to estimate the parameters is known to depend on the initial partition of the trajectories between groups. In the present work, the

CEM algorithm was repeated several times, starting with different initial partitions, and the solution that provided the highest log classification likelihood was kept. This proved to be time consuming. An alternative would be to give, as initial partition, the solution obtained by the kmeans[23] algorithm applied to the individual trajectory coefficients, estimated separately for each trajectory using the least square algorithm. The distance used in the kmeans algorithm should be the Euclidean distance relative to the covariance matrix within groups, as suggested by James and Sugar. [12] Further work would be needed to compare this alternative approach to the one used in the present work.

Mod_Mix_unequal requires a large number of trajectories to estimate all its parameters (K variance-covariance matrices needed compared to only one for Mod_Mix_equal). To reduce the number of parameters, Proust and Jacqmin-Gadda [12] proposed, in the context of the mixture model, a variance-covariance matrix $\mathbf{D}$ proportional across groups. Another approach was proposed by many authors[9,24,25] in the case of individual feature classification: the variance-covariance matrix is re-parameterised in terms of its eigenvalue decomposition as $\mathbf{D}_k = \lambda_k \mathbf{\Omega}_k \mathbf{A}_k \mathbf{\Omega}_k'$, where the parameter λ_k determines the volume of the k^{th} group, $\mathbf{\Omega}_k$ its orientation, and $\mathbf{A}_k$ its shape. One can force some of these quantities (volume, orientation, or shape) to be equal across groups.

The present study proposed the implementation of the mixed effect classification models in the case of Gaussian longitudinal data. It would be interesting to extend this model for other types of longitudinal data (binary, count…), and to evaluate the impact of taking into account the between-individual variance. In these cases, the posterior probability cannot be calculated analytically due to the integral over the random effects. Techniques, such as the Laplace approximation, could be used to approximate this integral.

In the simulation study, it would be interesting to evaluate the impact of allowing between-individual variance on the number of groups chosen by criteria such as ICL-BIC. In the case of a non-negligible between-individual variance, we believe that, compared to Mod_Mix_equal, classifying trajectories using the standard classification model would require increasing the number of groups, which may lose their clinical interpretation. Further work would also be needed to investigate the ability of ICL-BIC to choose the appropriate number of groups in case of between-individual variance.

Regarding the application of these methods to hCG data, the models taking into account between-individual variances led to a classification based mainly on the pattern of the trajectories rather than the levels, which is clinically more relevant.

Based on the simulation study presented herein, using the mixed effects classification model that considers equal between-individual variance for all groups appears more clinically relevant than the standard classification model for studies with a non-negligible number of trajectories.

Acknowledgments

The authors would like to thank François Golfier for providing “hydatidiform mole” data. They also thank Mrs Véréna Landel for helpful comments, suggestions, and revisions of the final manuscript.

Funding and competing interests

The authors declare to have no conflict of interest. This work was partially funded by the *Agence Nationale pour la Recherche* (ANR Grant 2012 BLAN SVSE 1)

References

1. Nagin, D. S. & Odgers, C. L. Group-based trajectory modeling in clinical research. *Annu. Rev. Clin. Psychol.* **6**, 109–138 (2010).
2. Pickles, A. & Croudace, T. Latent mixture models for multivariate and longitudinal outcomes. *Stat. Methods Med. Res.* **19**, 271–289 (2010).
3. Einbeck, J. & Hinde, R. D. and J. *npmlreg: Nonparametric maximum likelihood estimation for random effect models*. (2014).
4. Formann, A. K. Mixture analysis of longitudinal binary data. *Stat. Med.* **25**, 1457–1469 (2006).
5. Legler, J. M., Davis, W. W., Potosky, A. L. & Hoffman, R. M. Latent variable modelling of recovery trajectories: sexual function following radical prostatectomy. *Stat. Med.* **23**, 2875–2893 (2004).
6. Klich, A., Ecochard, R. & Subtil, F. Unequal intra-group variance in trajectory classification. *Stat. Med.* **37**, (2018).
7. Symons, M. J. Clustering Criteria and Multivariate Normal Mixtures. *Biometrics* **37**, 35–43 (1981).
8. Subtil, F. *et al.* An alternative classification to mixture modeling for longitudinal counts or binary measures. *Stat. Methods Med. Res.* **26**, 453–470 (2017).
9. James, G. M. & Sugar, C. A. Clustering for Sparsely Sampled Functional Data. *J. Am. Stat. Assoc.* **98**, 397–408 (2003).
10. Celeux, G. & Govaert, G. Comparison of the mixture and the classification maximum likelihood in cluster analysis. *J. Stat. Comput. Simul.* **47**, 127–146 (1993).
11. Govaert, G. & Nadif, M. Comparison of the mixture and the classification maximum likelihood in cluster analysis with binary data. *Comput. Stat. Data Anal.* **23**, 65–81 (1996).

12. Proust, C. & Jacqmin-Gadda, H. Estimation of linear mixed models with a mixture of distribution for the random effects. *Comput. Methods Programs Biomed.* **78**, 165–173 (2005).
13. Gaffney, S. & Smyth, P. Trajectory Clustering with Mixtures of Regression Models. in *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 63–72 (ACM, 1999). doi:10.1145/312129.312198
14. Muthén, B. & Shedden, K. Finite Mixture Modeling with Mixture Outcomes Using the EM Algorithm. *Biometrics* **55**, 463–469 (1999).
15. DeSarbo, W. S. & Cron, W. L. A maximum likelihood methodology for clusterwise linear regression. *J. Classif.* **5**, 249–282 (1988).
16. Laird, N. M. & Ware, J. H. Random-effects models for longitudinal data. *Biometrics* **38**, 963–974 (1982).
17. Muthén, B. & Shedden, K. Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics* **55**, 463–469 (1999).
18. McLachlan, G. & Peel, D. *Finite Mixture Models*. (John Wiley & Sons, 2004).
19. Biernacki, C., Celeux, G. & Govaert, G. *Assessing a Mixture Model for Clustering with the Integrated Classification Likelihood*. (INRIA, 1998).
20. Celeux, G. & Govaert, G. A classification EM algorithm for clustering and two stochastic versions. *Comput. Stat. Data Anal.* **14**, 315–332 (1992).
21. Gałecki, A. & Burzykowski, T. *Linear Mixed-Effects Models Using R - A Step-by-Step Approach*. (2013).
22. Schmitt, C. *et al.* Risk of gestational trophoblastic neoplasia after hCG normalisation according to hydatidiform mole type. *Gynecol. Oncol.* **130**, 86–89 (2013).
23. Hartigan, J.A & Wong, M.A. Algorithm AS 136: A k-Means Clustering Algorithm. *J. R. Stat. Soc. Ser. C Appl. Stat.* 100–108 (1979).

24. Banfield, J. D. & Raftery, A. E. Model-Based Gaussian and Non-Gaussian Clustering. *Biometrics* **49**, 803 (1993).

25. Celeux, G. & Govaert, G. Gaussian parsimonious clustering models. *Pattern Recognit.* **28**, 781–793 (1995).

Table 1 – Misclassification rates of Mod_Fix and Mod_Mix_equal for scenarios S1 to S4.

Scenario	Random effects covariance matrix	Between/ Total ratio (%)	MCR (%)	
			Mod_Fix	Mod_Mix_equal
S1	$\begin{pmatrix} 0.01 & 0.001 \\ 0.001 & 0.001 \end{pmatrix}$	1.3	0.21 (0.4,0.3,0.05)†	0.20 (0.4,0.3,0.05)†
S2	$\begin{pmatrix} 0.4 & 0.1 \\ 0.1 & 0.1 \end{pmatrix}$	55	21.04 (24,22,19)†	5.24 (3,8,4)†
S3	$\begin{pmatrix} 0.8 & 0.25 \\ 0.25 & 0.3 \end{pmatrix}$	77	62.43 (76,55,60)†	8.05 (6,12,5)†
S4	$\begin{pmatrix} 2 & 0.6 \\ 0.6 & 0.7 \end{pmatrix}$	89	64.18 (75,54,65)†	20.19 (20,29,14)†

MCR: misclassification rate - S1, S2, S3, S4: scenarios for the random effects variance-covariance matrix - Mod_Fix: fixed effects classification model - Mod_Mix_equal: mixed effects classification model which assumes an equal between-individual variance for all groups. Between/Total ratio: ratio of the between-individual variance over the total variance - †: the MCR in each of the three groups.

Table 2 – Misclassification rates of the three models for scenarios S3 and S5 to S7.

Scenario	Random effects covariance matrix	Between/Total ratio (%)	MCR (%)	
			Mod_Mix_equal	Mod_Mix_unequal
S3	$\begin{pmatrix} 0.8 & 0.25 \\ 0.25 & 0.3 \end{pmatrix}$	77 (77,77,77)*	8.05 (6,12,5)†	8.03 (6,13,5)†
S5	$\begin{pmatrix} 0.008 & 0.002 & & & \\ 0.002 & 0.003 & & & \\ & & 0.02 & 0.006 & \\ & & 0.006 & 0.007 & \\ & & & & 0.0001 & 0.00001 \\ & & & & 0.00001 & 0.00001 \end{pmatrix}$	3 (3,7,0.01)*	0.48 (0.5,1.1,0.02)†	0.43 (0.7,0.8,0.05)†
S6	$\begin{pmatrix} 3 & 0.61 & & & \\ 0.61 & 0.5 & & & \\ & & 0.8 & 0.25 & \\ & & 0.25 & 0.3 & \\ & & & & 4 & 0.63 \\ & & & & 0.63 & 0.4 \end{pmatrix}$	82 (86,77,85)*	29.90 (27,26,34)†	26.33 (23,24,29)†
S7	$\begin{pmatrix} 3 & 0.61 & & & \\ 0.61 & 0.5 & & & \\ & & 0.8 & 0.25 & \\ & & 0.25 & 0.3 & \\ & & & & 0.7 & 0.13 \\ & & & & 0.13 & 0.1 \end{pmatrix}$	76 (86,77,58)*	22.90 (31,46,3)†	11.93 (16,21,4)†

MCR: misclassification rate – S3, S5- S7: scenarios with an unequal variance-covariance matrix of the random effects between groups- Mod_Mix_equal: mixed effects classification model which assumes an equal between-individual variance for all groups - Mod_Mix_unequal: mixed effects classification model which considers unequal between-individual variance between groups. Between/Total ratio: the ratio of the between-individual variance over the total variance -* : the ratio of the between-individual variance over the total variance for each group - †: the MCR in each of the three groups

Table 3: Selection of the number of groups for the 3 models

Model	K	ICL-BIC
Mod_Fix	2	43075.95
	3	41048.26
	4	39456.56*
Mod_Mix_equal	2	35555.93
	3	35056.70
	4	34836.55*
Mod_Mix_unequal	2	35523.05
	3	35270.12*
	4	35705.88

Mod_Fix: fixed effects classification model - Mod_Mix_equal: mixed effects classification model which assumes an equal between-individual variance for all groups- Mod_Mix_unequal: mixed effects classification model which considers unequal between-individual variance between groups. K: the number of groups- *: the the highest ICL-BIC

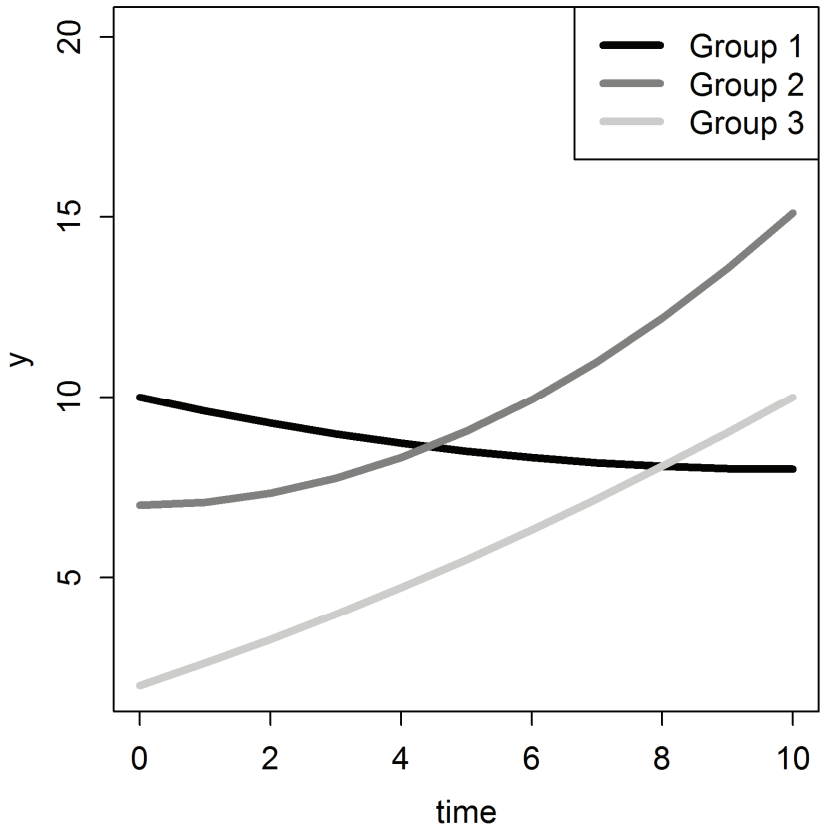
Table 4 – Contingency tables of patient classification according to the different models

Mod_Fix vs Mod_Mix_equal				Mod_Mix_unequal vs Mod_Mix_equal				
	Mod_Fix				Mod_Mix_unequal			
	Gr. 1	Gr. 2	Gr. 3		Gr. 1	Gr. 2	Gr. 3	Total
Mod_Mix_equal				Mod_Mix_equal				
Gr. 1	63	43	1	Gr. 1	107	0	0	107
Gr. 2	65	157	51	Gr. 2	35	238	0	273
Gr. 3	9	62	989	Gr. 3	4	50	1006	1060
Total	137	262	1041		146	288	1006	1440

Mod_Fix: fixed effects classification model - Mod_Mix_equal: mixed effects classification model which assumes an equal between-individual variance for all groups - Mod_Mix_unequal: mixed effects classification model which considers unequal between-individual variance between groups.

Figure legends

Figure 1 - Typical trajectories of the 3 groups used for the simulation study



Footnote to Figure 1: Black line: typical trajectory of group 1, medium grey line: typical trajectory of group 2, light grey line: typical trajectory of group 3.

Figure 2: Trajectories of logarithm-hCG values after hydatidiform mole curettage

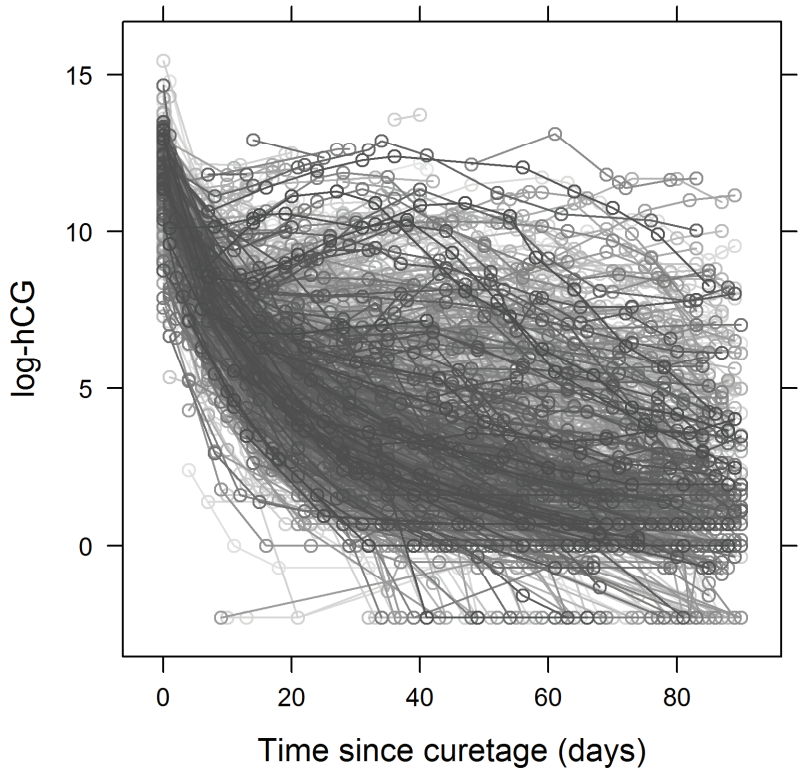
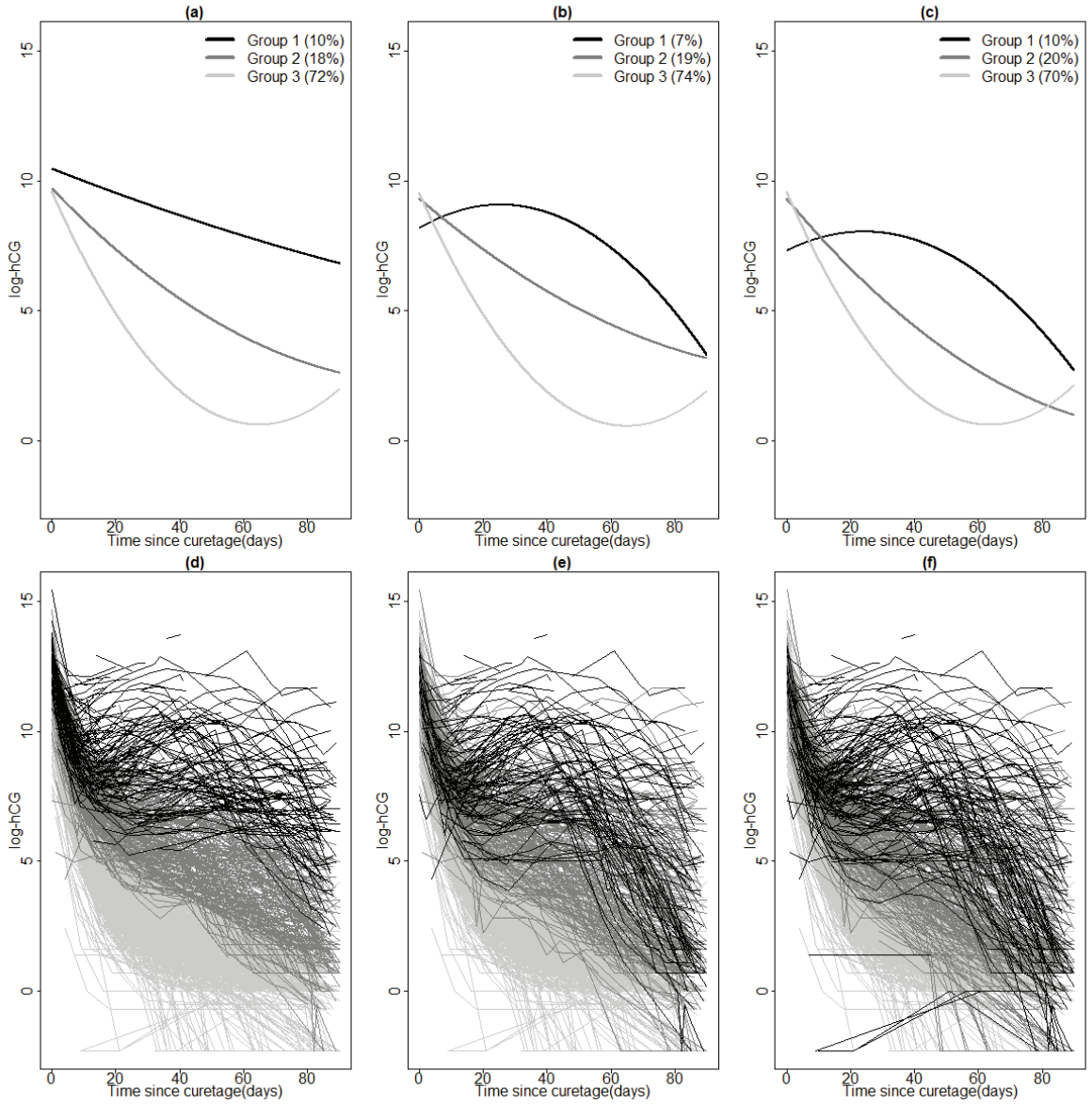
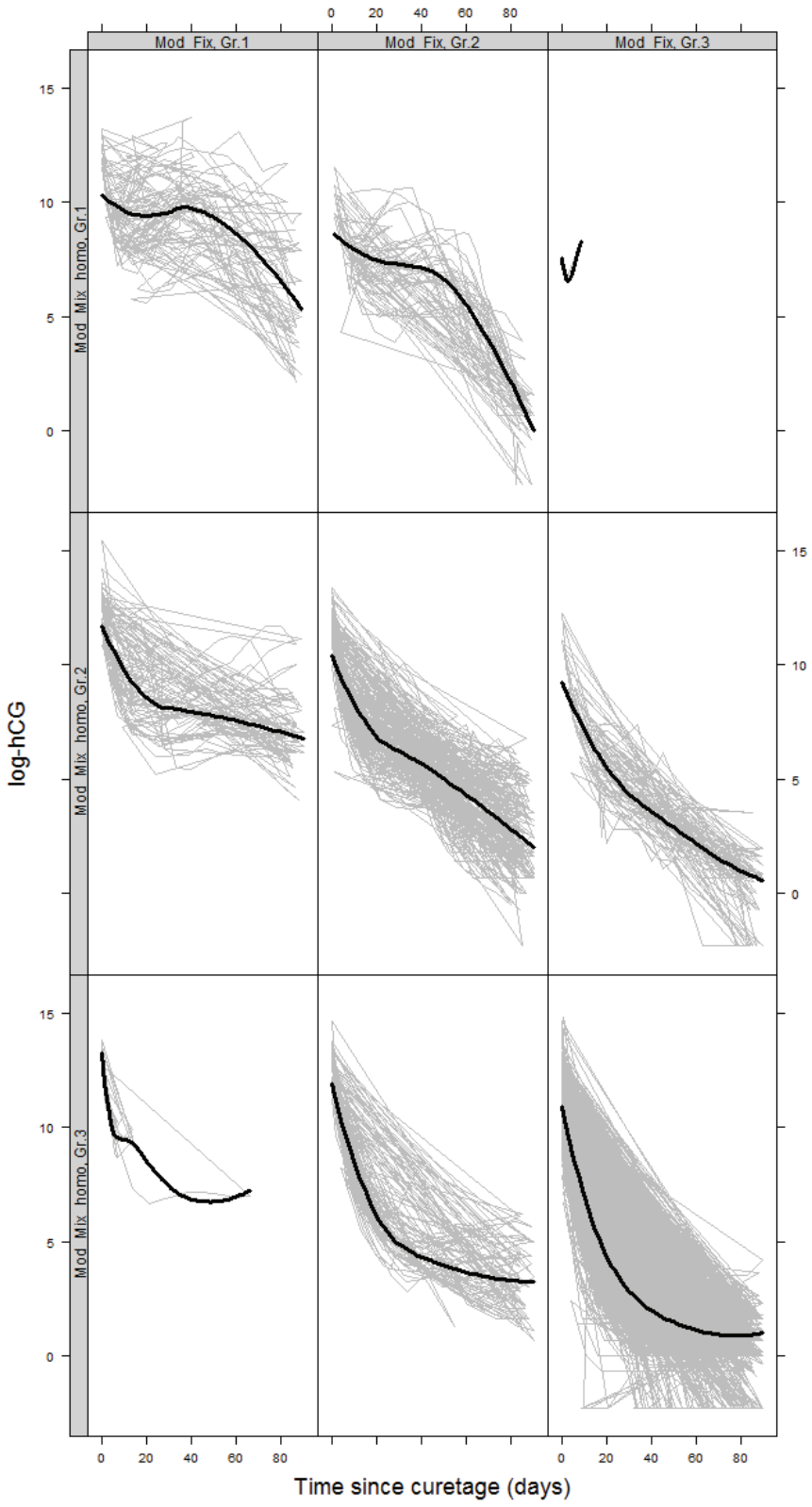


Figure 3: Classification of logarithm-hCG trajectories obtained with the fixed effects classification model and with the mixed effects classification models



Footnote to Figure 3:
Panels (a), (b) and (c): typical trajectories obtained with the 3 classification models; Mod_Fix, Mod_Mix_equal and Mod_Mix_unequal, respectively.
Panels (e), (f) and (g): observed individual trajectories coloured according to the group membership obtained with the 3 classification models; Mod_Fix, Mod_Mix_equal and Mod_Mix_unequal, respectively.

Figure 4: Observed logarithm-hCG trajectories that change group between the fixed effects classification model and the mixed effects classification model



Grey line: observed individual trajectories; black line: smooth trajectory

Supporting information

Supplementary Table S1 – mean bias of the estimated typical trajectories for Mod_Fix and Mod_Mix_equal

Scenario	Random effects covariance matrix	Between/ Total ratio (%)	Bias (%)	
			Mod_Fix	Mod_Mix_equal
S1	$\begin{pmatrix} 0.01 & 0.001 \\ 0.001 & 0.001 \end{pmatrix}$	1.3	0.0014	0.0005
S2	$\begin{pmatrix} 0.4 & 0.1 \\ 0.1 & 0.1 \end{pmatrix}$	55	-0.2830	0.0012
S3	$\begin{pmatrix} 0.8 & 0.25 \\ 0.25 & 0.3 \end{pmatrix}$	77	-0.3678	-0.0166
S4	$\begin{pmatrix} 2 & 0.6 \\ 0.6 & 0.7 \end{pmatrix}$	89	-0.4252	-0.0046

S1, S2, S3, S4: scenarios for the random effects variance-covariance matrix - Mod_Fix: fixed effects classification model - Mod_Mix_equal: mixed effects classification model which assumes an equal between-individual variance for all groups. Between/Total ratio: ratio of the between-individual variance over the total variance

Supplementary Table S2 – mean bias of the estimated typical trajectories for Mod_Mix_equal and Mod_Mix_unequal

Scenario	Random effects covariance matrix	Between/Total ratio (%)	Bias (%)	
			Mod_Mix_equal	Mod_Mix_unequal
S3	$\begin{pmatrix} 0.8 & 0.25 \\ 0.25 & 0.3 \end{pmatrix}$	77	-0.0166	0.0370
S5	$\begin{pmatrix} 0.008 & 0.002 & & & & \\ 0.002 & 0.003 & & & & \\ & & 0.02 & 0.006 & & \\ & & 0.006 & 0.007 & & \\ & & & & 0.0001 & 0.00001 \\ & & & & 0.00001 & 0.00001 \end{pmatrix}$	3	0.0570	0.0011
S6	$\begin{pmatrix} 3 & 0.61 & & & & \\ 0.61 & 0.5 & & & & \\ & & 0.8 & 0.25 & & \\ & & 0.25 & 0.3 & & \\ & & & & 4 & 0.63 \\ & & & & 0.63 & 0.4 \end{pmatrix}$	82	-0.7136	-0.4763
S7	$\begin{pmatrix} 3 & 0.61 & & & & \\ 0.61 & 0.5 & & & & \\ & & 0.8 & 0.25 & & \\ & & 0.25 & 0.3 & & \\ & & & & 0.7 & 0.13 \\ & & & & 0.13 & 0.1 \end{pmatrix}$	76	0.0169	0.0977

S3, S5- S7: scenarios with an unequal variance-covariance matrix of the random effects between groups- Mod_Mix_equal: mixed effects classification model which assumes an equal between-individual variance for all groups - Mod_Mix_unequal: mixed effects classification model which considers unequal between-individual variance between groups. Between/Total ratio: the ratio of the between-individual variance over the total variance

4.4 Principaux résultats de l'article

Les résultats des simulations ont montré que le modèle de classification à effet mixtes – en supposant une variance inter-individuelle égale entre les groupes – donne de meilleurs résultats de classification que le modèle standard. Lorsque le ratio entre la variance inter-individuelle et la variance résiduelle est au moins de un, l'écart entre les deux modèles exprimé en pourcentage d'individus mal classés peut atteindre 15 à 50 %. Lorsque la variance inter-individuelle est plus faible que la variance résiduelle, le modèle de classification à effets mixtes conduit un pourcentage de trajectoires mal classées proche de celui de modèle de classification standard : l'utilisation du modèle à effets mixtes n'entraîne donc aucune perte. L'usage du modèle de classification à effets mixtes devrait donc être recommandé, même si des travaux complémentaires seraient nécessaires pour comparer les deux modèles en fonction du nombre d'individus ou du nombre de mesures.

Lorsque la vraie matrice de covariance de variance des effets aléatoires est différente entre les groupes – i.e., variabilité inter-individuelle n'est pas la même d'un groupe à l'autre – le modèle de classification à effets mixtes supposant une variance inter-individuelle différente entre les groupes conduit à un pourcentage d'individus mal classés plus faible que celui du modèle de classification supposant une variance inter-individuelle égale (écart de l'ordre de 11 %). Par ailleurs, il n'y a pas de perte à utiliser ce modèle lorsque la variance inter-individuelle est égale pour tous les groupes ; à nouveau, cette conclusion serait à affiner avec de nouvelles simulations faisant varier le nombre d'individus ou le nombre de mesures.

4.5 Perspectives

4.5.1 Initialisation de l'algorithme CEM

Les estimations des paramètres du modèle (2.17) fournies par l'algorithme CEM dépendent de la partition initiale. L'algorithme est répété un grand nombre de fois à partir de partitions initiales aléatoires différentes. La solution retenue est celle qui conduit à la vraisemblance classifiante la plus élevée, tel que mentionné dans l'article ci-dessus.

Cependant, cette stratégie d'initialisation est très coûteuse en temps de calcul. Une alternative d'initialisation serait de donner comme partition initiale la solution obtenue par l'algorithme du kmeans modifié. Le principe revient à appliquer l'algorithme du kmeans non

pas sur les trajectoires, mais sur les coefficients des trajectoires individuelles, estimés séparément pour chaque trajectoire en utilisant l'algorithme des moindres carrés. La distance utilisée dans l'algorithme kmeans devrait être la distance euclidienne :

- (i) par rapport à la matrice de covariance empirique de la différence entre les coefficients estimés par la méthode des moindres carrés pour chaque trajectoire et les coefficients de la trajectoire typique associée à chaque trajectoire : $\mathrm{cov}\left(\hat{\boldsymbol{\beta}}_i - \hat{\boldsymbol{\beta}}_k\right)$, lorsque la matrice de variance covariance des effets aléatoires est supposée identique d'un groupe à l'autre ;
- (ii) par rapport à la matrice de covariance empirique des coefficients estimée au sein de groupes lorsque la matrice de variance covariance des effets aléatoires est supposée différente d'un groupe à l'autre.

La matrice de variance covariance est ré-estimée à chaque itération de l'algorithme. La solution obtenue par cet algorithme du kmeans modifié est une bonne approximation de la solution finale de l'algorithme CEM du modèle de classification à effets mixtes dans le cas (i), et une approximation plus forte pour le cas (ii).

En effet, à l'étape C de l'algorithme CEM décrit dans l'article, la trajectoire *i* est classée dans le groupe avec la probabilité a posteriori la plus élevée sachant les valeurs estimées des paramètres à l'étape M. Le calcul de cette probabilité a posteriori pour le groupe *k* peut être décomposé de la façon suivante :

$$
\begin{aligned}
& post.prob_{ik} = P(i \in k \mid \mathbf{y}_i, \boldsymbol{\pi}, \boldsymbol{\theta}) \\
& \propto \pi_k f(\mathbf{y}_i \mid \boldsymbol{\beta}_k, \mathbf{D}_k, \sigma^2) \\
& \propto \log(\pi_k) - \frac{1}{2}\log|\mathbf{V}_{ik}| - \frac{1}{2}\left(\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}_k\right)^t \mathbf{V}_{ik}^{-1}\left(\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}_k\right) \\
& \propto \log(\pi_k) - \frac{1}{2}\log|\mathbf{V}_{ik}| - \frac{1}{2}\left(\mathbf{y}_i - \hat{\mathbf{y}}_i\right)^t \mathbf{V}_{ik}^{-1}\left(\mathbf{y}_i - \hat{\mathbf{y}}_i\right) - \frac{1}{2}\underbrace{\left(\mathbf{y}_i - \hat{\mathbf{y}}_i\right)^t \mathbf{V}_{ik}^{-1}\left(\hat{\mathbf{y}}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}_k\right)}_{\varepsilon_i^t \mathbf{V}_{ik}^{-1}\left(\mathbf{X}_i\hat{b}_i\right)} \\
& \quad - \frac{1}{2}\underbrace{\left(\hat{\mathbf{y}}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}_k\right)^t \mathbf{V}_{ik}^{-1}\left(\mathbf{y}_i - \hat{\mathbf{y}}_i\right)}_{\left(\mathbf{X}_i\hat{b}_i\right)^t \mathbf{V}_{ik}^{-1}\varepsilon_i} - \frac{1}{2}\left(\hat{\mathbf{y}}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}_k\right)^t \mathbf{V}_{ik}^{-1}\left(\hat{\mathbf{y}}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}_k\right), \text{ avec } \hat{\mathbf{y}}_i = \mathbf{X}_i\hat{\boldsymbol{\beta}}_k + \mathbf{X}_i\hat{\mathrm{b}}_{ik} \\
& \propto \quad \log(\pi_k) - \frac{1}{2}\log|\mathbf{V}_{ik}| - \frac{1}{2}\left(\mathbf{y}_i - \hat{\mathbf{y}}_i\right)^t \mathbf{V}_{ik}^{-1}\left(\mathbf{y}_i - \hat{\mathbf{y}}_i\right) - \frac{1}{2}\left(\hat{\mathbf{y}}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}_k\right)^t \mathbf{V}_{ik}^{-1}\left(\hat{\mathbf{y}}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}_k\right) \\
& \propto \quad \log(\pi_k) - \frac{1}{2}\log|\mathbf{V}_{ik}| - \frac{1}{2}\left(\mathbf{y}_i - \hat{\mathbf{y}}_i\right)^t \mathbf{V}_{ik}^{-1}\left(\mathbf{y}_i - \hat{\mathbf{y}}_i\right) - \frac{1}{2}\left(\hat{\boldsymbol{\beta}}_i - \hat{\boldsymbol{\beta}}_k\right)^t \mathbf{X}_i^t\mathbf{V}_{ik}^{-1}\mathbf{X}_i\left(\hat{\boldsymbol{\beta}}_i - \hat{\boldsymbol{\beta}}_k\right) \\
& \propto \quad \log(\pi_k) - \frac{1}{2}\log|\mathbf{V}_{ik}| - \frac{1}{2}\left(\mathbf{y}_i - \hat{\mathbf{y}}_i\right)^t \mathbf{V}_{ik}^{-1}\left(\mathbf{y}_i - \hat{\mathbf{y}}_i\right) - \frac{1}{2}\left(\hat{\boldsymbol{\beta}}_i - \hat{\boldsymbol{\beta}}_k\right)^t \mathrm{cov}(\hat{\boldsymbol{\beta}}_k)^{-1}\left(\hat{\boldsymbol{\beta}}_i - \hat{\boldsymbol{\beta}}_k\right)
\end{aligned}
\tag{2.18}
$$

avec $\mathrm{cov}(\hat{\boldsymbol{\beta}}_k)^{-1} = \mathbf{X}_i\left(\mathbf{X}_i\mathbf{D}_k\mathbf{X}_i^t + \sigma^2\mathbf{I}_{T_i}\right)^{-1}\mathbf{X}_i^t$. Pour cette décomposition, on considère que la matrice de design des effets aléatoires est égale à celle des effets fixes ($\mathbf{Z}_i = \mathbf{X}_i$).

Dans le cas où la matrice de variance covariance des effets aléatoires est supposée constante d'un groupe à l'autre (i), cette probabilité a posteriori s'écrit de la façon suivante :

$$
\log(\pi_k) - \frac{1}{2}\log|\mathbf{V}_i| - \frac{1}{2}\left(\mathbf{y}_i - \hat{\mathbf{y}}_i\right)^t \mathbf{V}_i^{-1}\left(\mathbf{y}_i - \hat{\mathbf{y}}_i\right) - \frac{1}{2}\left(\hat{\boldsymbol{\beta}}_i - \hat{\boldsymbol{\beta}}_k\right)^t \mathrm{cov}(\hat{\boldsymbol{\beta}})^{-1}\left(\hat{\boldsymbol{\beta}}_i - \hat{\boldsymbol{\beta}}_k\right) \tag{2.19}
$$

Ainsi, seul le premier et le quatrième terme de la probabilité interviennent dans la détermination de la partition des trajectoires dans les groupes. Dans le cas (i), la matrice $\mathrm{cov}(\hat{\boldsymbol{\beta}})$ est approximée par la matrice de variance covariance $\mathrm{cov}\left(\hat{\boldsymbol{\beta}}_i - \hat{\boldsymbol{\beta}}_k\right)$, ce qui revient à négliger le terme σ^2. Par conséquent, la solution obtenue par ce k-means utilisant une distance euclidienne pondérée doit être une bonne approximation des estimations des paramètres du modèle de classification, et donc une bonne initialisation.

Dans le cas où la matrice de variance covariance des effets aléatoires est supposée constante d'un groupe à l'autre (ii), les termes deux et trois de l'équation (2.19) sont négligés en plus du terme σ^2 dans l'utilisation du k-means modifié. L'approximation obtenue est vraisemblablement moins bonne.

Des travaux supplémentaires sont nécessaires afin d'étudier les performances de l'algorithme CEM en utilisant cette méthode d'initialisation.

4.5.2 Réduction de nombre des paramètres

Lorsque le nombre de trajectoires et le nombre de mesures est faible, le modèle (2.17) avec une matrice de variance-covariance variable d'un groupe à l'autre peut être non identifiable ou difficilement identifiable. En effet, le nombre des paramètres du modèle est élevé par rapport au nombre d'observations (produit du nombre de trajectoires et du nombre de mesures). Par conséquence, il devient nécessaire de diminuer le nombre des paramètres afin d'obtenir des modèles identifiables. Pour cela, il est possible de re-paramétrer les matrices variance-covariance[38,39] des effets aléatoires du modèle (2.17) en s'appuyant sur leur décomposition en valeurs propres.

$$\mathbf{D}_k = \mathbf{\Omega}_k \mathbf{\Lambda}_k \mathbf{\Omega}_k'$$

où $\mathbf{\Omega}_k$ est la matrice orthogonale de vecteurs propres, et $\mathbf{\Lambda}_k$ la matrice diagonale composée des valeurs propres, qui peut être décomposée en un nombre réel λ_k et une matrice $\mathbf{A}_k$ tels que $\mathbf{\Lambda}_k = \lambda_k \mathbf{A}_k$ avec $|\mathbf{A}_k| = 1$. Chaque matrice de variance-covariance des effets aléatoires est donc finalement décomposée sous la forme :

$$\mathbf{D}_k = \lambda_k \mathbf{\Omega}_k \mathbf{A}_k \mathbf{\Omega}_k'$$

où λ_k caractérise le volume du groupe, $\mathbf{A}_k$ caractérise la forme du groupe, et $\mathbf{\Omega}_k$ caractérise l'orientation du groupe.

Celeux et Govaert (1995)[39] ont proposé de nombreux modèle de mélange, allant du plus simple (volumes, orientations et formes considérés égaux) au modèle le plus compliqué (volumes, orientations et formes différents pour chaque groupe), en décrivant les algorithmes de maximum de vraisemblance associés. Ce travail a été réalisé sur des données non-longitudinales.

Il sera intéressant d'utiliser cette re-paramétrisation dans le cadre du modèle de classification de trajectoires à effets mixtes avec une matrice de variance-covariance variable

d’un groupe à l’autre lorsque le nombre d’observations n’est pas assez important pour estimer tous les paramètres.

5 Conclusion générale

Dans ce travail de thèse, deux modèles de classification de trajectoires sont proposés pour créer une typologie des individus, en identifiant des groupes d'individus aux trajectoires de biomarqueur similaires. Par rapport au modèle de classification de trajectoires standard, chaque modèle prend en compte une source de variabilité supplémentaire. Le premier modèle de classification prend en compte une variance résiduelle au sein de chaque groupe variable d'un groupe à l'autre, tout en supposant que la variance inter-individuelle est nulle, alors que le deuxième modèle prend en compte une variance inter-individuelle au sein de chaque groupe, qui peut être identique ou variable d'un groupe à l'autre. Malgré les limites abordées dans ce travail, l'intérêt de ces deux modèles a été montré par des travaux de simulations et par des applications cliniques. Globalement, lorsque le nombre de mesures ou de trajectoires est suffisant, ces modèles sont toujours avantageux en termes de classification par rapport au modèle de classification standard. Par ailleurs, en dehors de plans expérimentaux très contrôlés, les deux sources de variabilité mentionnées ci-dessus sont inhérentes à la recherche en santé. Ces modèles sont donc très pertinents d'un point de vue clinique.

Des travaux sont encore nécessaires pour vérifier en particulier l'intérêt du deuxième modèle en fonction du nombre des mesures ou du nombre de trajectoires lorsqu'une variance inter-individuelle différente d'un groupe à l'autre est considérée. De plus, il serait intéressant d'étendre ces modèles à des biomarqueurs de nature différente tels que ceux reflétant un processus de comptage. L'inférence associée à ces modèles est encore à développer.

Au final, selon les hypothèses effectuées dans le modèle de classification, les trajectoires typiques obtenues peuvent être différentes. Néanmoins, dans une étude clinique, la vraie classification n'est pas connue, et sauf dans de rares cas (étude diagnostique), il n'existe pas forcément de vraie classification : la classification obtenue par les modèles est un construit qui aide à l'interprétation des données. Ainsi, le choix final de la classification repose à la fois sur la validité des hypothèses des modèles, mais également sur la pertinence clinique de la classification, et sur l'aide à l'interprétation qu'elle apporte.

6 Production scientifique

Publications

Accepté : **Klich A**, Subtil F, Ecochard E. *Unequal intra-group variance in trajectory classification. Statistics in Medicine,* 4155-4166, 2018

Rédigé : **Klich A,** Subtil F, Ecochard E**.** *Trajectory clustering with mixed classification model.*

Communication orale

Klich A, Subtil F, Ecochard R. *Prise en compte de l'hétérogénéité interindividuelle dans la classification de trajectoires*. EPICLIN 2018, 30 mai-1 juin, Nice, France.

Communication affichée

Klich A, Subtil F, Ecochard R. *Trajectory clustering with mixed effects classification model*. ISCB 2017, 9-13 juillet, Vigo, Espagne.

7 Références

1. Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework. *Clinical Pharmacology & Therapeutics* **69**, 89–95 (2001).
2. Zolg, J. W. & Langen, H. How industry is approaching the search for new diagnostic markers and biomarkers. *Mol. Cell Proteomics* **3**, 345–354 (2004).
3. Taylor, J. M. G., Yu, M. & Sandler, H. M. Individualized predictions of disease progression following radiation therapy for prostate cancer. *J. Clin. Oncol.* **23**, 816–825 (2005).
4. Bellera, C. A., Hanley, J. A., Joseph, L. & Albertsen, P. C. Hierarchical changepoint models for biochemical markers illustrated by tracking postradiotherapy prostate-specific antigen series in men with prostate cancer. *Ann Epidemiol* **18**, 270–282 (2008).
5. You, B. *et al.* Advantages of prostate-specific antigen (PSA) clearance model over simple PSA half-life computation to describe PSA decrease after prostate adenomectomy. *Clin. Biochem.* **41**, 785–795 (2008).
6. Bastard, M. *et al.* Revisiting long-term adherence to highly active antiretroviral therapy in Senegal using latent class analysis. *J. Acquir. Immune Defic. Syndr.* **57**, 55–61 (2011).
7. Goldstein, D. J. *et al.* Cyclosporine-associated end-stage nephropathy after cardiac transplantation: incidence and progression. *Transplantation* **63**, 664–668 (1997).
8. Pattison, J. M. *et al.* The incidence of renal failure in one hundred consecutive heart-lung transplant recipients. *Am. J. Kidney Dis.* **26**, 643–648 (1995).
9. Aleksic, I. *et al.* Improvement of impaired renal function in heart transplant recipients treated with mycophenolate mofetil and low-dose cyclosporine. *Transplantation* **69**, 1586–1590 (2000).

10. Barkmann, A. *et al.* Improvement of acute and chronic renal dysfunction in liver transplant patients after substitution of calcineurin inhibitors by mycophenolate mofetil. *Transplantation* **69**, 1886–1890 (2000).
11. Schrama, Y. C. *et al.* Conversion to mycophenolate mofetil in conjunction with stepwise withdrawal of cyclosporine in stable renal transplant recipients. *Transplantation* **69**, 376–383 (2000).
12. Weir, M. R. *et al.* Long-term impact of discontinued or reduced calcineurin inhibitor in patients with chronic allograft nephropathy. *Kidney Int.* **59**, 1567–1573 (2001).
13. Boissonnat, P. *et al.* Impact of the early reduction of cyclosporine on renal function in heart transplant patients: a French randomised controlled trial. *Trials* **13**, 231 (2012).
14. Schmitt, C. *et al.* Risk of gestational trophoblastic neoplasia after hCG normalisation according to hydatidiform mole type. *Gynecol. Oncol.* **130**, 86–89 (2013).
15. Chechia, A. *et al.* [Molar pregnancy. Retrospective study of 60 cases in Tunisia]. *Tunis Med* **79**, 441–446 (2001).
16. Laird, N. M. & Ware, J. H. Random-effects models for longitudinal data. *Biometrics* **38**, 963–974 (1982).
17. Cox, D. R. & Reid, N. Parameter Orthogonality and Approximate Conditional Inference. *Journal of the Royal Statistical Society: Series B (Methodological)* **49**, 1–18 (1987).
18. Patterson, H. D. & Thompson, R. Recovery of inter-block information when block sizes are unequal. *Biometrika* **58**, 545–554 (1971).
19. Harville, D. A. Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems. *Journal of the American Statistical Association* **72**, 320–338 (1977).
20. Nagin, D. S. & Odgers, C. L. Group-based trajectory modeling in clinical research. *Annu Rev Clin Psychol* **6**, 109–138 (2010).

21. Pickles, A. & Croudace, T. Latent mixture models for multivariate and longitudinal outcomes. *Stat Methods Med Res* **19**, 271–289 (2010).
22. Dempster, A., Laird, N. & Rubin, D. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* **39**, 1–38 (1977).
23. Muthén, B. & Shedden, K. Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics* **55**, 463–469 (1999).
24. James, G. M. & Sugar, C. A. Clustering for Sparsely Sampled Functional Data. *Journal of the American Statistical Association* **98**, 397–408 (2003).
25. Proust, C. & Jacqmin-Gadda, H. Estimation of linear mixed models with a mixture of distribution for the random effects. *Comput Methods Programs Biomed* **78**, 165–173 (2005).
26. Verbeke, G. & Lesaffre, E. A Linear Mixed-Effects Model With Heterogeneity in the Random-Effects Population. *Journal of the American Statistical Association* **91**, 217–221 (1996).
27. Gaffney, S. J. Curve clustering with random effects regression mixtures. in *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics* (2003).
28. ELLIOTT, M. R., GALLO, J. J., TEN HAVE, T. R., BOGNER, H. R. & KATZ, I. R. Using a Bayesian latent growth curve model to identify trajectories of positive affect and negative events following myocardial infarction. *Biostatistics* **6**, 119–143 (2005).
29. Celeux, G. & Govaert, G. Comparison of the mixture and the classification maximum likelihood in cluster analysis. *Journal of Statistical Computation and Simulation* **47**, 127–146 (1993).
30. Celeux, G. & Govaert, G. *A classification EM algorithm for clustering and two stochastic versions*. (INRIA, 1991).

31. Celeux, G. & Diebolt, J. *Une version de type recuit simule de l'algorithme EM*. (1989).
32. Celeux, G. & Diebolt, J. The SEM Algorithm : a Probabilistic Teacher Algorithm Derived from the EM Algorithm for the Mixture Problem. *Computational Statistics Quarterly* **2**, 73–82 (1985).
33. MacQueen, J. Some methods for classification and analysis of multivariate observations. in (The Regents of the University of California, 1967).
34. Rubin, D. B. *Multiple Imputation for Nonresponse in Surveys*. (John Wiley & Sons, 2004).
35. Roy, J. & Lin, X. Latent variable models for longitudinal data with multiple continuous outcomes. *Biometrics* **56**, 1047–1054 (2000).
36. Nagin, D. S., Jones, B. L., Passos, V. L. & Tremblay, R. E. Group-based multi-trajectory modeling. *Stat Methods Med Res* **27**, 2015–2023 (2018).
37. Nagin, D. S. & Tremblay, R. E. Analyzing developmental trajectories of distinct but related behaviors: a group-based method. *Psychol Methods* **6**, 18–34 (2001).
38. Banfield, J. D. & Raftery, A. E. Model-Based Gaussian and Non-Gaussian Clustering. *Biometrics* **49**, 803–821 (1993).
39. Celeux, G. & Govaert, G. Gaussian parsimonious clustering models. *Pattern Recognition* **28**, 781–793 (1995).