UNIVERSITÉ MONTPELLIER II
SCIENCES ET TECHNIQUES DU LANGUEDOC

THÈSE

pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ MONTPELLIER II

Formation Doctorale : Biostatistique
École Doctorale : Information, Structures, Systèmes

présentée par

Carine VÉRA

Modèles linéaires mixtes multiphasiques pour l'analyse de données longitudinales - Application à la croissance des plantes

Soutenue publiquement le 15 novembre 2004

devant le jury composé de :

M. Gilles DUCHARME	Président
Mme Hélène JACQMIN-GADDA	Rapporteur
M. Hervé MONOD	Rapporteur
M. François HOULLIER	Examinateur
M. Yann GUÉDON	Co-Directeur de thèse
M. Christian LAVERGNE	Directeur de thèse

Table des matières

Introduction générale

Des mesures effectuées sur des individus de manière répétée au cours du temps sont appelées données longitudinales au sens des méthodes statistiques. Les données longitudinales peuvent être collectées au cours d'un suivi des individus dans le temps, ou de manière rétrospective en extrayant des mesures sur chaque individu à partir d'enregistrements passés. La grosse majorité d'exemples de ce type de données est issue des domaines de la biologie et de la santé (essais cliniques), mais il en existe dans d'autres domaines, comme la sociologie (études de panel). Ces données requièrent des méthodes statistiques spécifiques car l'ensemble des observations relatives à un individu est généralement corrélé, et cette corrélation doit être prise en compte pour faire correctement de l'inférence statistique. Un ouvrage complet est dédié à l'analyse de données longitudinales (Diggle *et al.*, 2002). Le modèle linéaire mixte est le modèle de base pour l'analyse de ce type de données (Diggle *et al.*, 2002 ; Verbeke et Molenberghs, 2000). L'introduction d'effets aléatoires permet de séparer les différentes sources de variation : celle due à la sélection aléatoire d'un échantillon d'individus pour effectuer l'étude et celle due aux erreurs engendrées par le processus de mesure.

L'objectif de cette thèse est de développer des méthodes d'analyse de données longitudinales qui présentent un certain nombre de caractéristiques particulières :
• les données sont structurées en phases successives,
• les données sont influencées par des covariables pouvant varier dans le temps,
• les données présentent une hétérogénéité inter-individuelle.
Cette problématique statistique est issue d'une problématique biologique qui va être détaillée, mais elle aurait très bien pu être suscitée par des données provenant d'autres domaines, comme des données de suivi de patients dans le domaine biomédical.

En foresterie, pour la gestion des plantations, il est intéressant de connaitre les intéractions entre les arbres et l'environnement. Autrement dit, il est important de comprendre comment pousse un peuplement d'arbres en fonction des conditions environnementales. Les questions qui viennent immédiatement en tête sont entre autres : quel est l'effet des conditions climatiques ? Quelle est son importance selon l'âge des individus ? Y a t-il une hétérogénéité de la croissance plus ou moins forte entre les différents individus ? Si oui, durant quelle période de croissance ? Les données botaniques étudiées pour répondre à cette problématique sont des données relatives à la croissance en longueur d'arbres forestiers. Ces données longitudinales ont la particularité d'être récoltées rétrospectivement, la lecture de la croissance s'effectuant une fois les arbres abattus. D'un point de vue biologique,

l'objectif est donc d'analyser la croissance en longueur d'arbres forestiers en fonction de facteurs climatiques, et ceci pour différents âges des arbres.

L'idée la plus immédiate est de modéliser les données avec la famille des modèles linéaires mixtes sur toute la période de croissance. Après comparaison de divers modèles sur la base de critères de sélection tels que les critères AIC et BIC (Burnham et Anderson, 2002), il s'avère que le modèle linéaire mixte qui réalise le meilleur compromis entre ajustement aux données et parcimonie possède une structure complexe, avec de nombreux paramètres : tendance polynômiale d'ordre élevé, effet aléatoire sophistiqué avec un intercept et une pente aléatoires, matrice de variance-covariance complexe avec une structure de corrélation autorégressive d'ordre 1,...Il s'avère que la structure du modèle sélectionné est beaucoup trop complexe pour être interprétée d'un point de vue biologique, de manière claire.

Toutefois, les botanistes sont en mesure de suggérer des hypothèses pour la modélisation des données, hypothèses également mises en évidence par l'analyse exploratoire des données botaniques et des données climatiques. La croissance d'un arbre est constituée d'une succession de phases liées à sa morphogenèse[1] : "effet de base" où la croissance augmente rapidement, stabilisation quand elle atteint sa phase adulte, puis "dérive" caractérisée par une diminution progressive de la croissance. De plus, la croissance est sensible à des facteurs climatiques et principalement aux stress hydriques. Cependant cette sensibilité est plus ou moins importante selon l'âge de l'arbre. De même des facteurs non observés, modélisés par des effets aléatoires, doivent être pris en compte, comme les attaques de parasites dont peuvent être victimes de jeunes arbres. Cela peut entraîner au sein d'un même peuplement où tous les individus sont soumis aux mêmes conditions environnementales, une hétérogénéité de la croissance entre ces individus. À partir de ces considérations biologiques, nous souhaitons donc modéliser des données longitudinales qui sont structurées en phases successives, qui sont soumises à l'influence de covariables climatiques et qui présentent une hétérogénéité inter-individuelle. De plus, l'influence des covariables climatiques et l'hétérogénéité inter-individuelle varient selon la phase de croissance.

Certaines méthodes existantes pour l'analyse multiphasique ne sont pas appropriées dans notre cas. Par exemple en zootechnie, pour l'analyse multiphasique de l'efficacité de la reproduction chez des vaches laitières (Grossman *et al.*, 1995) ou pour l'étude de courbes de croissance multiphasique chez le vison (Sorensen *et al.*, 2003), des fonctions logistiques multiphasiques sont utilisées. Une fonction logistique est définie sur chacune des phases et les paramètres sont estimés avec un algorithme itératif, basé sur une adaptation de l'algorithme des moindres carrés dans un cadre non-linéaire. Cette méthode de modélisation d'un phénomène multiphasique n'est pas satisfaisante dans notre cas car la modélisation n'est axée que sur les phases et aucune hétérogénéité entre les individus n'est modélisée.

Par conséquent, nous proposons une nouvelle famille de modèles statistiques : les **modèles linéaires mixtes multiphasiques**. Le modèle linéaire mixte multiphasique est un modèle de type Markov caché qui combine :

[1] La morphogenèse correspond au développement des formes et des structures d'un organisme.

• une chaîne de Markov pour modéliser la succession de phases,
• des modèles linéaires mixtes associés aux états de la chaîne de Markov sous-jacente.
Pour chacune des phases, la tendance et les covariables sont modélisées par des effets fixes, et un effet aléatoire modélise l'hétérogénéité entre les individus.
Le modèle linéaire mixte multiphasique que l'on pourrait également traduire par "Markov switching linear mixed model" dans le contexte Markov caché est une nouvelle combinaison markovienne de modèles. Il existe, en effet, dans la littérature de nombreux modèles de type Markov caché qui combinent de manière markovienne des modèles, comme les "Markov switching autoregressive model" pour lesquels les observations dans chacun des états sont modélisées par un processus autorégressif (Ephraim et Merhav, 2002). De même, Churchill (1989) a introduit des modèles de type Markov caché basés sur une combinaison markovienne de chaînes de Markov d'ordre m. Ces modèles supposent une dépendance markovienne d'ordre m entre les observations conditionnellement aux états cachés, et ils ont été utilisés pour la détection de régions homogènes dans les séquences d'ADN (Muri, 1997). Martin (2002) a également proposé des modèles de mélange de modèles linéaires mixtes (cas particulier où la chaîne de Markov est d'ordre 0) pour l'analyse de données de puces d'ADN.

Organisation du document :

Le chapitre 1 est dédié à la présentation et à l'analyse exploratoire des jeux de données botaniques étudiés et des données climatiques correspondant à la période de croissance des données botaniques. L'analyse exploratoire des données botaniques met en évidence les hypothèses décrites précédemment pour la modélisation statistique proposée. Prendre en compte des covariables climatiques dans la modélisation soulève un problème de changement d'échelle entre le pas de temps annuel des données botaniques constituant la variable réponse du modèle et le pas de temps journalier des covariables climatiques. Pour résoudre ce problème, nous proposons un modèle bioclimatologique relativement rudimentaire, résultant de considérations à la fois biologiques et climatologiques afin d'obtenir des covariables climatiques ayant le même pas de temps annuel que les données botaniques.

Dans le chapitre 2, sont présentés le modèle linéaire mixte et le modèle de chaîne de Markov cachée, qui composent le modèle linéaire mixte multiphasique. Les méthodes d'estimation des paramètres de ces deux modèles sont exposées, en vue de leur utilisation pour l'estimation des paramètres du modèle linéaire mixte multiphasique au chapitre 3. Les différentes méthodes d'estimation sont comparées et leurs propriétés sont soulignées.

Le chapitre 3 est consacré au modèle linéaire mixte multiphasique et aux méthodes d'estimation proposées pour l'estimation de ses paramètres. Deux familles de modèles linéaires mixtes multiphasiques sont présentées. Elles diffèrent par la modélisation de l'effet aléatoire : la séquence observée peut être modélisée avec un unique effet aléatoire ou avec un effet aléatoire différent sur chacun des états.
L'estimation des paramètres de ces modèles n'est pas aisée. En effet, le modèle linéaire mixte et la chaîne de Markov étant deux modèles à structure cachée, la combinaison des deux structures cachées (les effets aléatoires et les états) ne permet pas l'écriture

analytique de l'étape E de l'algorithme EM. Nous envisageons également un algorithme de type EM qui prend en compte les effets aléatoires. Cet algortihme itératif est composé de trois étapes : restauration, maximisation et prédiction. L'étape de restauration est probabiliste et s'implémente par un algorithme "avant-arrière" sachant les effets aléatoires. L'étape de maximisation s'écrit sans difficulté. Cependant il est nécessaire de prédire à chaque itération une nouvelle valeur des effets aléatoires. Cette étape de prédiction pose problème car l'étape de restauration probabiliste détermine l'ensemble des séquences d'états possibles.

Pour contourner ces difficultés, nous proposons comme alternative à l'algorithme EM un algorithme itératif en trois étapes : restauration, maximisation et prédiction. L'étape de restauration peut être déterministe ou effectuée par simulation. L'algorithme basé sur une restauration déterministe est inspiré de l'algorithme de Baum-Viterbi (Jelinek, 1976), particulièrement adapté pour l'estimation des modèles ayant une structure "gauche-droite", utilisés pour les applications à la croissance des plantes. L'algorithme basé sur une restauration par simulation est inspiré de l'algorithme SEM (McLachlan et Krishnan, 1997). Compte tenu des relations d'indépendance conditionnelles spécifiques au modèle linéaire mixte multiphasique, l'étape de restauration déterministe avec l'algorithme de Viterbi nécessite l'hypothèse supplémentaire d'intégrer les effets aléatoires au calcul de la séquence d'états optimale. L'étape de restauration par simulation utilise un algorithme "avant-arrière" sachant les effets aléatoires. De même, l'étape de maximisation nécessite cette même hypothèse pour le calcul de la probabilité jointe de la séquence observée et de la séquence d'états restaurée. Par conséquent une étape de prédiction est nécessaire pour calculer les valeurs prédites des effets aléatoires, valeurs utilisées pour les étapes de restauration et de maximisation.

Le chapitre 4 est consacré à l'application de la modélisation et des méthodes d'estimation proposées pour l'analyse de la croissance en longueur d'arbres forestiers. Deux jeux de données botaniques présentés au chapitre 1 sont modélisés avec le modèle linéaire mixte multiphasique ayant un effet aléatoire différent sur chaque état. Les paramètres des modèles sont estimés par l'algorithme de type Baum-Viterbi. Pour chaque jeu de données, le nombre d'états (c'est-à-dire de phases de croissance) est sélectionné sur la base de critères de sélection de type AIC et BIC. Il est intéressant de déterminer, de manière globale, lequel des trois éléments intervenant dans la modélisation – le nombre de phases, les covariables climatiques et l'effet aléatoire individuel – est prépondérant. Une fois le nombre d'états sélectionné, pour chaque phase de croissance, on compare l'influence des différentes sources de variation : covariables climatiques et effet aléatoire. Un travail de validation est effectué en vérifiant que la segmentation met bien en évidence des phases séparées par des ruptures nettes.

Chapitre 1

Présentation et analyse exploratoire des données botaniques et climatiques

1.1 Introduction

Ce premier chapitre a pour but à la fois de présenter les jeux de données étudiés et de mettre en évidence des hypothèses pour la modélisation statistique proposée. Nous avons travaillé avec deux types de données : des données botaniques et des données climatiques correspondant à la période de croissance des données botaniques. La variable d'intérêt sera représentée par les données botaniques qui constitueront la variable réponse du modèle statistique développé. Les données climatiques permettront de construire des covariables a priori supposées pertinentes pour la modélisation statistique.

Dans un premier temps, nous décrirons les jeux de données botaniques et les jeux de données climatiques après avoir introduit quelques définitions de botanique. Après la présentation des jeux de données, une analyse exploratoire est nécessaire afin de mettre en évidence certaines propriétés et suggérer des hypothèses pour la modélisation statistique. Une analyse exploratoire des données botaniques sera réalisée en précisant les méthodes employées pour l'exploration de ce type de données spécifiques. Enfin, nous présenterons l'analyse exploratoire des données climatiques et nous proposerons un modèle bioclimatologique pour obtenir des covariables climatiques ayant le même pas de temps annuel que les données botaniques.

1.2 Présentation des jeux de données botaniques et climatiques

1.2.1 Quelques définitions de botanique

La plante est une structure organisée dans l'espace et le temps. La structure globale d'une plante peut être décomposée en un certain nombre d'entités botaniques, qui correspondent à différents niveaux d'organisation emboîtés les uns dans les autres. Le développement de la plante se construit par répétition de ces entités botaniques élémentaires au cours de trois processus fondamentaux : la croissance, la ramification et la réitération[1].

L'objectif biologique de cette étude étant d'analyser la croissance d'arbres forestiers, en fonction de facteurs environnementaux, et pour différents stades ontogéniques[2] des arbres, nous précisons, à partir de la revue critique de Caraglio et Barthélémy (1997), les modalités d'expression du processus de croissance des plantes :

On distingue deux types de croissance : la croissance primaire et la croissance secondaire. Pour toutes les plantes, la croissance primaire correspond à la croissance en longueur de la tige. Uniquement pour certaines plantes (les dicotylédons[3]), la croissance secondaire, aussi appelée croissance cambiale, correspond à la croissance en épaisseur de la tige et résulte du fonctionnement du cambium[4] mettant en place des structures ou formations anatomiques secondaires. Même si dans ce travail nous n'étudions que la croissance primaire, nous reparlerons dans les perspectives de la croissance secondaire.

La croissance primaire ou croissance en longueur d'une tige est le résultat de deux mécanismes : l'*organogenèse* et l'*allongement*. L'organogenèse se déroule à l'apex ou extrémité de la tige, au niveau du *méristème terminal*[5]. Celui-ci est constitué de cellules à intense activité mitotique, et initie de nouvelles portions de tige à sa base, et des ébauches foliaires sur ses flancs. L'allongement d'une tige est la manifestation directement observable de la croissance primaire. Il est essentiellement le résultat d'un allongement cellulaire qui prend naissance un peu en arrière du dôme apical.

Si aucune phase de repos prolongée n'est observée au cours de l'allongement de la tige, alors la croissance est qualifiée de *continue* (Figure 1.1). C'est le cas du Palmier à huile (*Elaeis guineensis* Jacq., Arecaceae) ou du Cocotier (*Cocos nucifera* L., Arecaceae). Au

[1] Duplication d'une partie ou de la totalité de l'architecture (Oldeman, 1974).

[2] L'ontogenèse est le déroulement de l'histoire d'un individu, à partir du germe jusqu'à l'aboutissement de son cycle vital.

[3] Les dicotylédons sont des végétaux à fleurs à feuilles larges, caractérisés par la présence de deux cotylédons (petites feuilles primaires de l'embryon) opposés.

[4] L'épaississement du tronc est causé par deux assises génératrices périphériques. La première externe se situe dans l'écorce et s'appelle l'assise génératrice subéro -phellodermique. Elle assure le renouvellement de l'écorce qui se desquame en vieillissant (cela est très visible sur le platane par exemple). La seconde est l'assise génératrice ligneuse, c'est le cambium. Elle assure la croissance en épaisseur du tronc en fabriquant le bois.

[5] Les méristèmes, localisés aux extremités des axes (tiges et racines) se définissent comme un ensemble de cellules embryonnaires à forte activité mitotique qui génèrent de nouveaux tissus (épiderme, tissus conducteurs...) et organes (Nougarède, 2001 ; Heuret, 2002).

contraire, si une tige se met en place par une alternance de périodes d'allongement et de périodes de repos, la croissance est qualifiée de *rythmique* (Figure 1.1). La portion de tige mise en place au cours d'une période d'allongement ininterrompue est appelée *unité de croissance*, dénotée UC. La portion de tige mise en place au cours d'une saison de végétation est appelée *pousse annuelle*, dénotée PA. Si la PA est allongée en une seule vague de croissance, elle est alors constituée d'une seule unité de croissance et elle est qualifiée de *monocyclique*. Si elle est allongée en plusieurs vagues de croissance successives, elle est alors constituée de plusieurs unités de croissance et elle est dite *polycyclique* (Figure 1.2). Une PA, composée de deux, trois et parfois même quatre UC, sera respectivement qualifiée de bicyclique, tricyclique et tétracyclique.

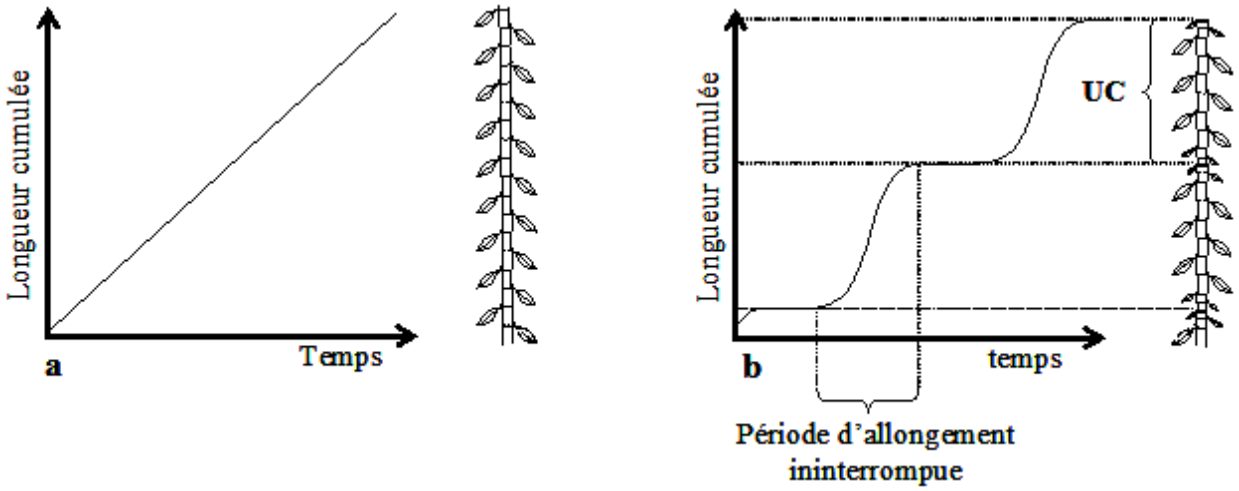

FIG. 1-1 – Croissance continue (a) ou rythmique (b) de l'axe feuillé (Heuret, 2002).

Remarquons que tous les arbres forestiers étudiés ont une croissance rythmique. Cette rythmicité temporelle se traduit par une rythmicité structurelle. Des marqueurs morphologiques, qui traduisent le fonctionnement passé des méristèmes, permettent le plus souvent de repérer a posteriori les arrêts de croissance. Ces marqueurs sont les cicatrices laissées par les cataphylles, petites écailles protégeant les bourgeons. Ils peuvent aussi être caractérisés par la diminution de la longueur des entre-noeuds[6]. Selon les espèces, la reconnaissance de ces marqueurs est possible sur un nombre d'années variable, ce qui permet de reconstituer a posteriori la croissance de l'arbre sur des périodes plus ou moins longues.

[6]Un axe est constitué d'une succession d'entre-noeuds séparés par des noeuds. Les noeuds sont les lieux d'insertion des feuilles (Heuret, 2002).

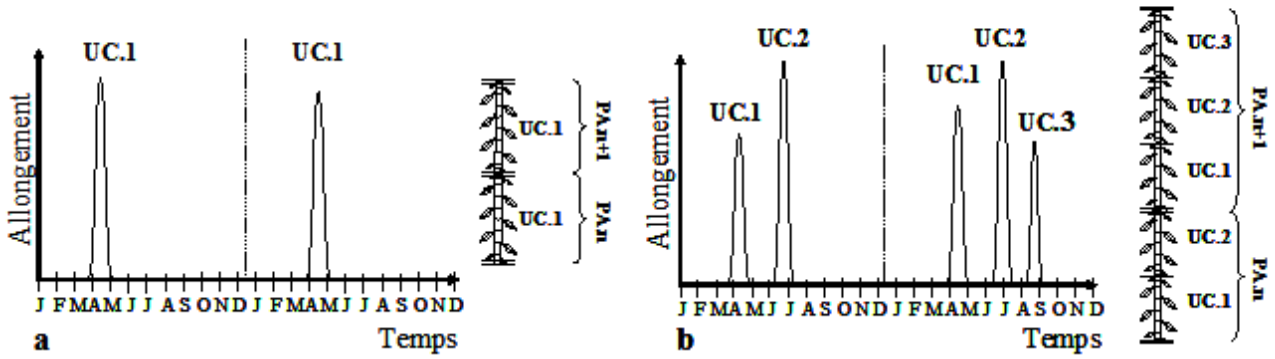

FIG. 1-2 – Croissance annuelle monocyclique (a) ou polycyclique (b) de l'axe feuillé (Heuret, 2002).

1.2.2 Les jeux de données relatifs au chêne sessile (*Quercus petraea* (Matt.) Liebl., Fagaceae)

Deux jeux de données sur le chêne sessile ont fait l'objet d'une étude botanique précise (Heuret *et al.*, 2000). L'objectif de cette étude était d'établir un modèle descriptif des différentes composantes de la croissance en hauteur du chêne sessile.

Tous les arbres observés proviennent de la forêt privée de Louppy-le-château, dans le département de la Meuse. Les deux peuplements, âgés respectivement de 15 et de 29 ans, sont issus de régénération naturelle et sont situés non loin l'un de l'autre. Pour les deux peuplements, les arbres étudiés ont subi les mêmes interventions sylvicoles[7] pour des âges identiques.

Les arbres du premier peuplement sont issus de la glandée de 1983. Ils ont été abattus en janvier 1998 et les observations portent sur un échantillon de 46 individus. Les arbres du second peuplement sont issus de la glandée de 1969. Ils ont été abattus en mars 1998 et les observations portent sur un échantillon de 19 individus. Comme beaucoup de ligneux des régions tempérées, le chêne a une croissance rythmique et son système aérien est constitué d'une succession d'UC. Le chêne sessile est polycyclique et peut former de une à quatre UC au cours d'une même saison végétative. Entre chaque phase d'allongement, il y a une période de repos durant laquelle le méristème, qui va générer l'UC suivante est protégé par les cataphylles. Ainsi, sur des portions d'axes relativement jeunes, les

[7]La sylviculture peut être définie comme l'ensemble des interventions humaines qui orientent la croissance d'un peuplement et des individus qui le composent, vers des objectifs qui peuvent être : la production de matière ligneuse (bois d'oeuvre, bois d'industrie...), la protection du sol contre l'érosion, la conservation de la biodiversité...(Meredieu, 1998).

cicatrices laissées par les cataphylles permettent de repérer les arrêts de croissance et donc de délimiter les UC (Figure 1.3). Sur des portions de tige plus âgées, les cicatrices laissées par les différents organes foliaires s'estompent et il est parfois difficile de localiser tous les arrêts de croissance.

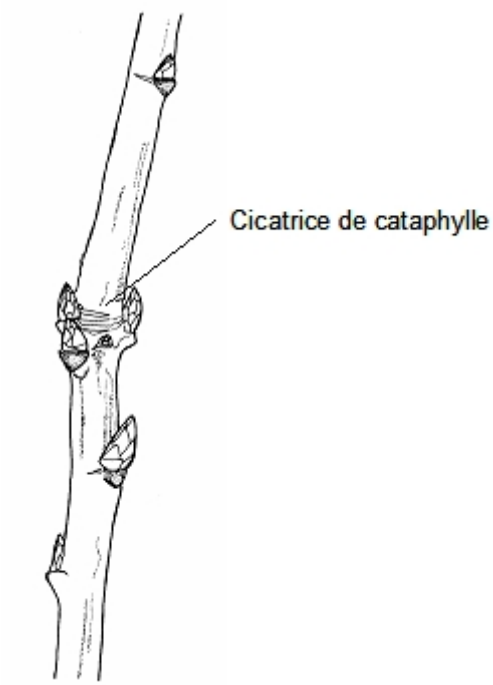

FIG. 1-3 – Arrêt de croissance inter-annuel délimité par les cicatrices de cataphylles chez le chêne sessile.

Le protocole de mesure est le suivant. Sur l'arbre abattu, les UC de l'axe principal sont délimitées par l'observation des marqueurs morphologiques. La prise de mesures s'effectue de la base vers le sommet de l'arbre, et une coupe transversale est effectuée à la base de chacune des UC repérées, afin de déterminer son âge par lecture de son nombre de cernes[8]. Le tronçon de bois séparant deux coupes, et comportant logiquement un arrêt de croissance, est ensuite fendu longitudinalement en passant par la moëlle. Cette opération permet de vérifier la présence d'un arrêt de croissance. Si sur un tronçon donné, un arrêt supplémentaire n'ayant pas été identifié à partir des marqueurs morphologiques externes est repéré par l'analyse de la moëlle, la longueur des UC est alors mesurée entre deux arrêts successifs visualisés sur la moëlle, et les données sont corrigées. Le caractère monocyclique ou polycyclique d'une PA est déduit de son nombre d'UC constitutives. Les longueurs des PA sont obtenues en faisant la somme des longueurs de toutes leurs UC.

[8]Un cerne (ou anneau de croissance) est un cercle concentrique formé chaque année par le cambium, visible sur la section radiale d'un tronc grâce à la différence d'aspect et de coloration entre le bois final (ou d'été) et le bois initial (ou de printemps).

Notons que contrairement à ce qui s'observe pour d'autres espèces tempérées, pour le chêne sessile les marqueurs morphologiques utilisables dans une analyse rétrospective de la croissance s'estompent rapidement du fait de la croissance en épaisseur. Par conséquent, pour les arbres âgés de 15 ans, les données de longueurs de PA relatives aux premières années de croissance de certains individus manquent. Pour les arbres de 29 ans, compte tenu de la difficulté de repérage des UC à la base des arbres, la longueur des UC n'a été mesurée qu'à partir de 1.5 m du sol. C'est pourquoi nous n'avons au mieux les premières longueurs de PA qu'à partir de 1974.

La figure 1.4 (resp. figure 1.5) représente l'évolution de la longueur des PA au cours du temps pour les individus âgés de 15 ans (resp. 29 ans).

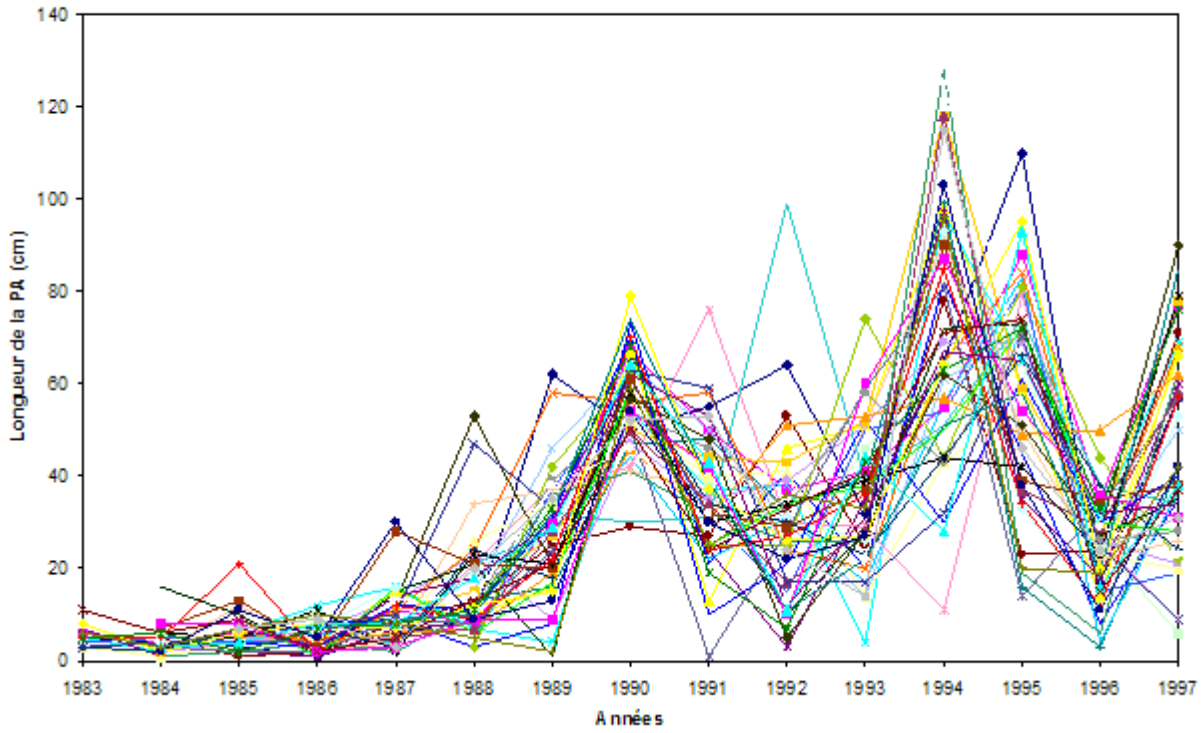

FIG. 1-4 – Longueur des PA en fonction des années chez les chênes sessiles âgés de 15 ans.

1.2.3 Les jeux de données relatifs au pin laricio (*Pinus nigra* Arnold ssp. *laricio* Poiret Maire var. *corsicana*)

Quatre jeux de données sur le pin laricio on été étudiés dans le cadre d'une thèse en biologie forestière (Meredieu, 1998) portant sur la croissance et la ramification du

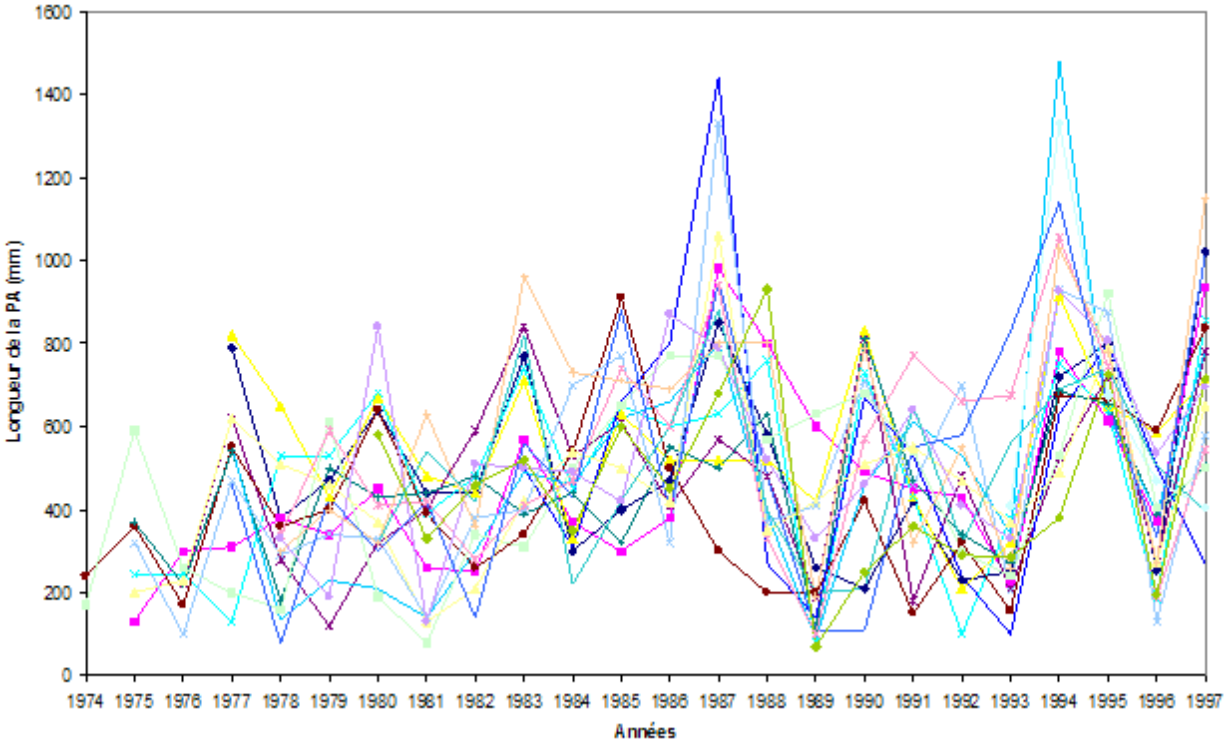

FIG. 1-5 – Longueur des PA en fonction des années chez les chênes sessiles âgés de 29 ans.

pin laricio. L'objectif était d'élaborer pour la sous-espère pin laricio un ensemble de modèles permettant de prévoir l'évolution des caractéristiques d'un peuplement donné et des arbres qui le composent. L'étude de ces données a permis en particulier d'approfondir la connaissance du développement architectural aérien du pin laricio.

Ces jeux de données correspondent à quatre peuplements de pin laricio de Corse, sélectionnés dans la forêt domaniale d'Orléans, et très semblables du point de vue de la sylviculture. Avant d'être plantés en forêt, les arbres ont été élevés en pépinière jusqu'à l'âge de 2 et 3 ans. Lorsqu'ils ont été abattus, ils étaient âgés respectivement de 6 ans (dont 2 ans en pépinière), 12 ans (dont 3 ans en pépinière), 18 ans (dont 3 ans en pépinière) et 23 ans (dont 3 ans en pépinière). Les hauteurs successives des UC pour chaque arrêt de croissance visible sur le terrain ont été mesurées pour chacun des individus. Ainsi, le premier jeu de données est composé d'un échantillon de 31 individus âgés de 6 ans dont on a pu relire la croissance sur toutes les années. Le second jeu de données comporte 29 individus âgés de 12 ans dont on a pu relire la croissance sur au mieux 11 ans. Les troisième et quatrième jeux de données contiennent respectivement 30 arbres âgés de 18 ans et 13 arbres âgés de 23 ans. Leur croissance a été relue au mieux respectivement sur 17 et 21 ans.

Précisons enfin que le pin laricio est monocyclique, sa pousse annuelle étant allongée

en une seule vague de croissance sur les mois d'avril et mai.

Les figures 1.6, 1.7, 1.8 et 1.9 représentent l'évolution de la longueur de la PA en fonction du temps chez les pins laricios âgés respectivement de 6, 12, 18 et 23 ans.

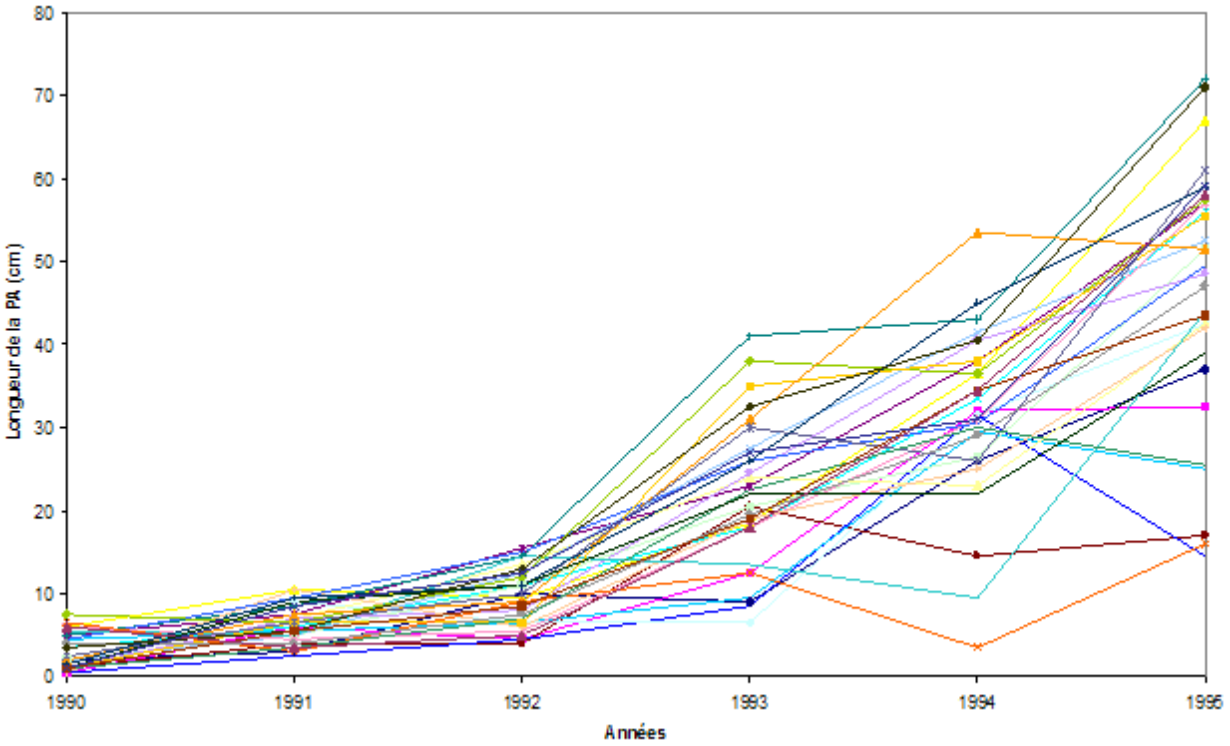

FIG. 1-6 – Longueur des PA en fonction des années chez les pins laricios âgés de 6 ans.

1.2.4 Mise en forme des jeux de données

À partir des données brutes observées sur les arbres forestiers, des séries contenant des informations sur la croissance des troncs des arbres ont été extraites avec le logiciel AMAPmod. Ce logiciel a été créé dans le cadre du projet AMAPmod au sein de l'UMR AMAP (botAnique et bioinforMatique de l'Architecture des Plantes) du CIRAD (Godin *et al.*, 1997, 1999). Ce projet vise à développer un ensemble de méthodes et d'outils logiciels destinés à l'analyse de la structure et de la croissance des plantes.

Remarquons également que le nombre de branches par pousse est disponible pour les pins laricios et que le nombre de cycles de croissance l'est pour les chênes sessiles. Pour chaque individu, une série multivariée contenant les informations relatives au tronc de l'arbre observé peut donc être construite. Nous verrons dans les perspectives comment utiliser toute l'information contenue dans ces séries multivariées.

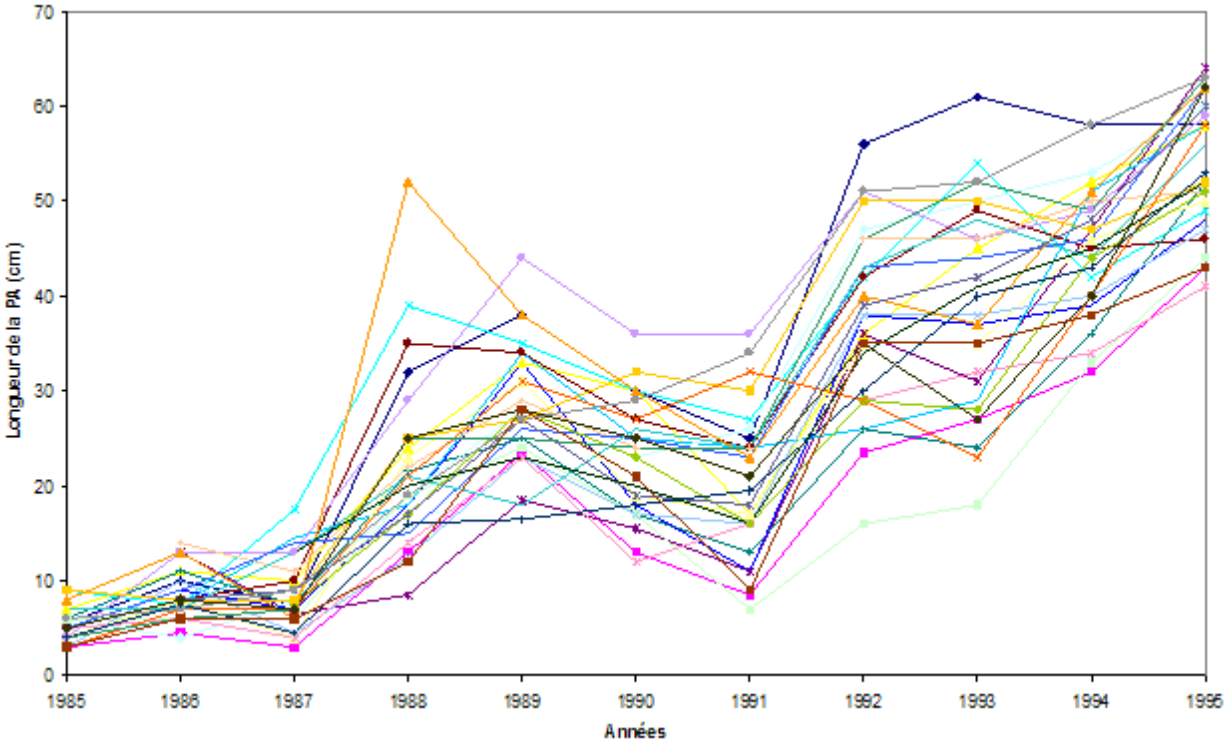

FIG. 1-7 – Longueur des PA en fonction des années chez les pins laricios âgés de 12 ans.

À noter que toutes les séries sont décrites avec un pas de temps annuel, puisque la pousse de l'arbre est annuelle. Des difficultés s'ensuivent pour faire correspondre les jeux de données climatiques avec un pas de temps journalier.

1.2.5 Les jeux de données climatiques

En plus des jeux de données botaniques, nous disposons des jeux de données climatiques correspondant à la période de croissance des données botaniques. Toutes les données climatiques ont été fournies par Météo-France. Les données relatives au chêne sessile proviennent du poste de St Dizier pour les années allant de 1969 à 1997. Les données se rapportant au pin laricio sont issues du poste de Chambon-la-forêt qui est localisé en ambiance forestière, pour les années allant de 1976 à 1996. Ces deux jeux de données contiennent, à l'échelle journalière, les informations suivantes : la hauteur totale des précipitations cumulées sur la journée en millimètres, les températures minimale et maximale en degrés Celsius, ainsi que la durée d'ensoleillement en minutes uniquement pour le poste de St Dizier.

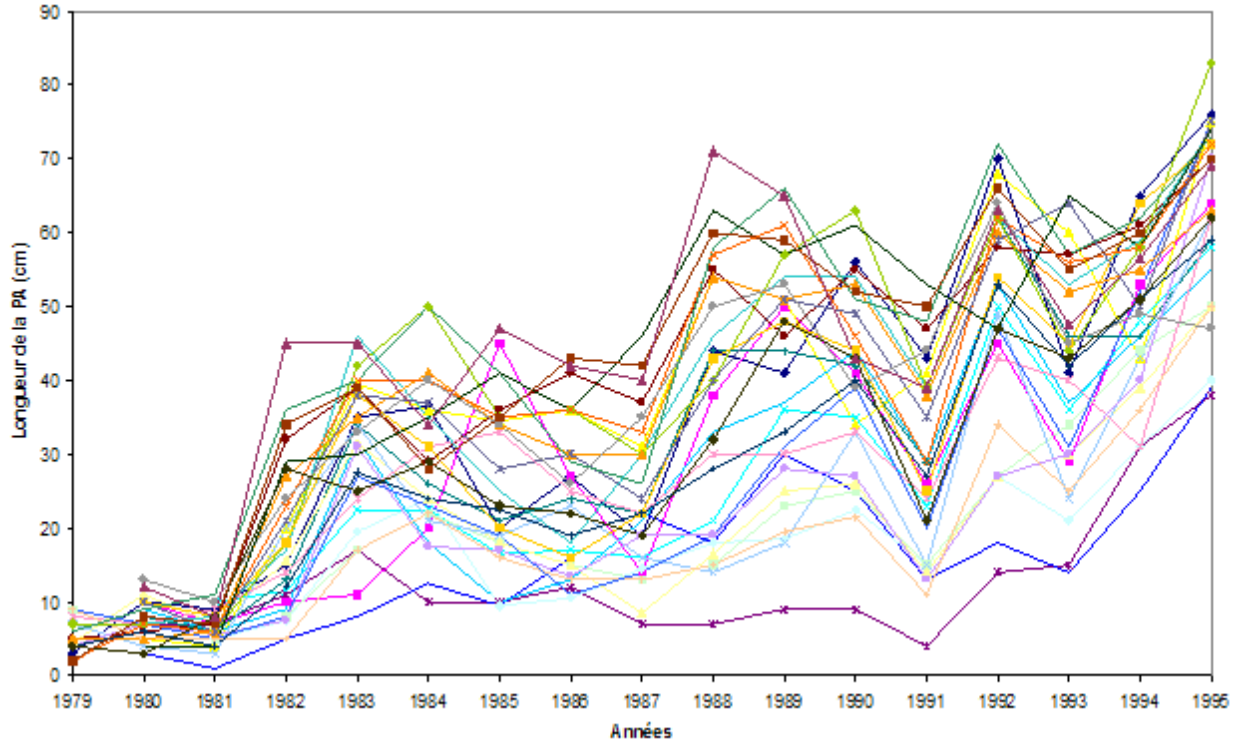

FIG. 1-8 – Longueur des PA en fonction des années chez les pins laricios âgés de 18 ans.

1.3 Analyse exploratoire des données

Avant tout travail de modélisation statistique des données, un travail d'analyse exploratoire est indispensable, d'une part pour résumer et structurer l'information contenue dans les données, d'autre part pour mettre en évidence certaines propriétés de l'échantillon et suggérer des hypothèses sur le modèle statistique.

1.3.1 Analyse exploratoire des données botaniques

Les données botaniques étudiées ont comme point commun de pouvoir être qualifiées de données longitudinales au sens des méthodes statistiques. Toutefois, contrairement au domaine biomédical où les données longitudinales sont récoltées au cours d'un suivi médical régulier, nos données de croissance de plantes sont collectées rétrospectivement, la croissance des arbres étant reconstituée à l'aide de mesures faites a posteriori, une fois les arbres abattus.

Dans l'analyse de données longitudinales, les exemples classiques de la littérature (Diggle *et al.*, 2002 ; Verbeke et Molenberghs, 2000) sont constitués d'un grand nombre de séries très courtes (beaucoup de sujets et seulement entre 5 et 10 observations par

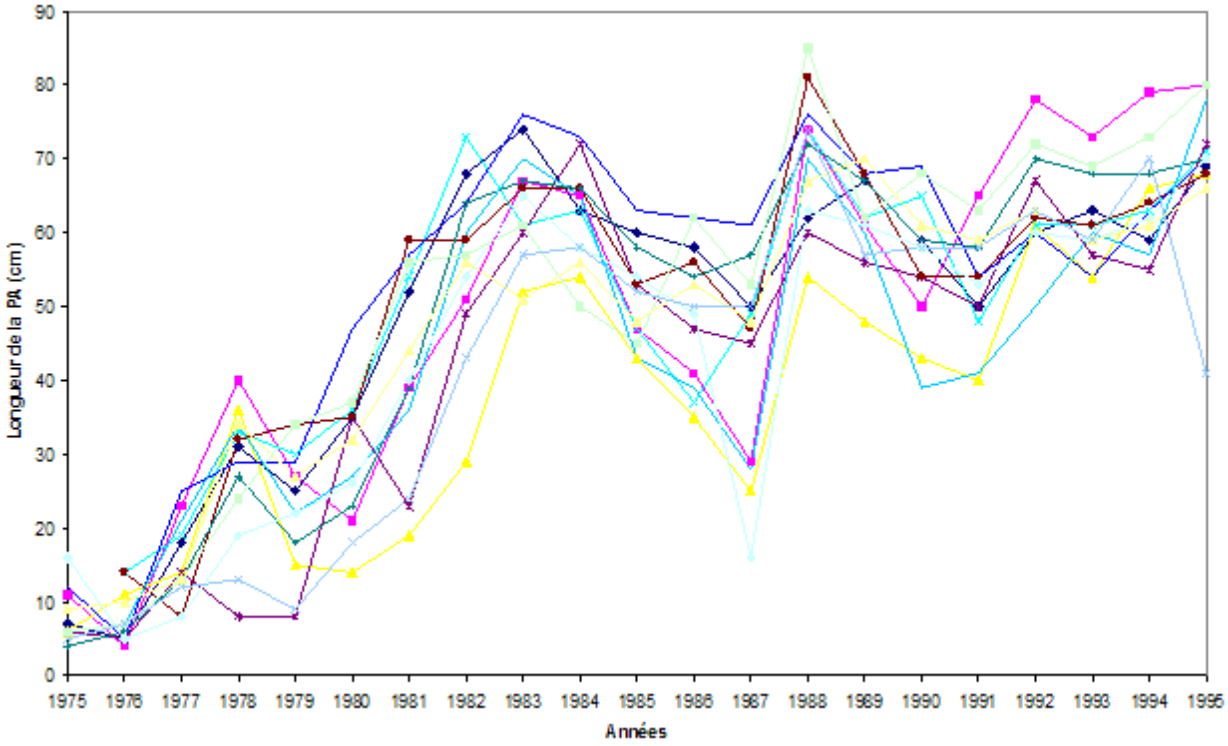

FIG. 1-9 – Longueur des PA en fonction des années chez les pins laricios âgés de 23 ans.

sujet). Ici nous disposons parfois d'un nombre faible de séries (du fait de contraintes expérimentales) qui sont en revanche observées sur une période plus importante (de 15 à 30 observations). Ainsi, à l'aide de techniques de filtrage linéaire utilisées notamment pour l'étude de séries chronologiques (Chatfield, 2003 ; Diggle, 1990), on souhaite savoir quelles informations peuvent ressortir d'une décomposition entre une tendance et des fluctuations.

Un filtre linéaire transforme une série chronologique $\{x_t\}$ en une autre série chronologique $\{y_t\}$ par l'opération linéaire :

$$y_t = \sum_{r=-q}^{+s} a_r x_{t+r},$$

avec $\{a_r\}$ un ensemble de poids.

On souhaite séparer les différentes sources de variation, à savoir ce qui varie lentement, et que l'on appelle *la tendance*, de ce qui varie rapidement, et que l'on nomme *les fluctuations locales* ou encore *les résidus*. La tendance est obtenue en appliquant aux données un filtre dit passe-bas (ne laissant passer que les basses fréquences). Il s'agit d'un filtre linéaire tel que $\sum_r a_r = 1$, symétrique ($s = q$ et $a_r = a_{-r}$) avec des poids décroissant

à partir de la valeur centrale. Les fluctuations locales sont obtenues en retranchant la tendance aux données, ce qui revient à appliquer aux données un filtre dit passe-haut (ne laissant passer que les hautes fréquences). Si z_t représente la valeur filtrée à l'instant t, alors :

$$z_t = x_t - \sum_{r=-q}^{+q} a_r x_{t+r}.$$

Dans ce chapitre, trois jeux de données sont étudiés : deux sur les chênes sessiles, et un autre sur les pins laricios âgés de 6 ans (Figures 1.4, 1.5 et 1.6). Pour ces trois exemples, nous avons choisi un filtre symétrique de demi-largeur r égale à 2. Les coefficients du filtre sont les probabilités d'une loi binomiale de paramètres 4 et 0.5 ; x_t est la longueur de la PA à l'instant t, et y_t est la valeur filtrée à l'instant t. Les effets de bords sont gérés en donnant aux valeurs manquantes la valeur de la plus proche extrémité (Brockwell et Davis, 2002).

Les figures 1.10, 1.11 et 1.12 (resp. 1.13, 1.14 et 1.15) représentent la tendance (resp. les fluctuations locales) extraite des données sur les chênes âgés de 15 ans, les chênes âgés de 29 ans, et les pins âgés de 6 ans.

La tendance reflète l'aspect endogène de la croissance des arbres. Pour les pins de 6 ans, la tendance traduit l'effet de base (Figure 1.12). Cela correspond à la phase d'établissement de l'arbre qui se caractérise par une forte augmentation de la longueur des PA au cours des premières années de croissance. L'effet de base est également présent au cours des premières années sur le graphe de la tendance des chênes de 15 ans (Figure 1.10). Pour les chênes âgés de 29 ans, l'effet de base n'est pas présent, car, contrairement au cas des chênes de 15 ans, les données pour les premières années de croissance manquent.

Les fluctuations locales reflètent l'influence des facteurs externes sur la croissance des arbres. Sur les deux graphes des fluctuations locales relatifs aux chênes (Figures 1.13 et 1.14), les courbes sont souvent synchrones entre individus, signifiant qu'une année est bonne ou mauvaise en terme de croissance, pour la majorité des arbres. Cela traduit un effet année très certainement dû au climat. En revanche, pour les pins de 6 ans, l'effet année n'est pas observable, le nombre de pics vers le haut étant à peu près égal au nombre de pics vers le bas pour une année donnée. Enfin, sur ces graphes des fluctuations locales, nous constatons que l'amplitude des fluctuations n'est pas constante ; elle paraît même proportionnelle à l'accroissement de la tendance chez les chênes de 15 ans. Cela traduit une variance non constante. Il faudra par conséquent tenir compte de cette caractéristique en introduisant de l'hétéroscédasticité dans la modélisation de la structure de variance-covariance.

À la suite de cette analyse exploratoire des données botaniques, nous pouvons émettre un certain nombre d'hypothèses pour la modélisation statistique :
La croissance d'un arbre semble se décomposer en phases successives bien distinctes. Pour les jeunes pins, seule une phase de "jeunesse" est observée alors qu'une phase de "jeunesse"

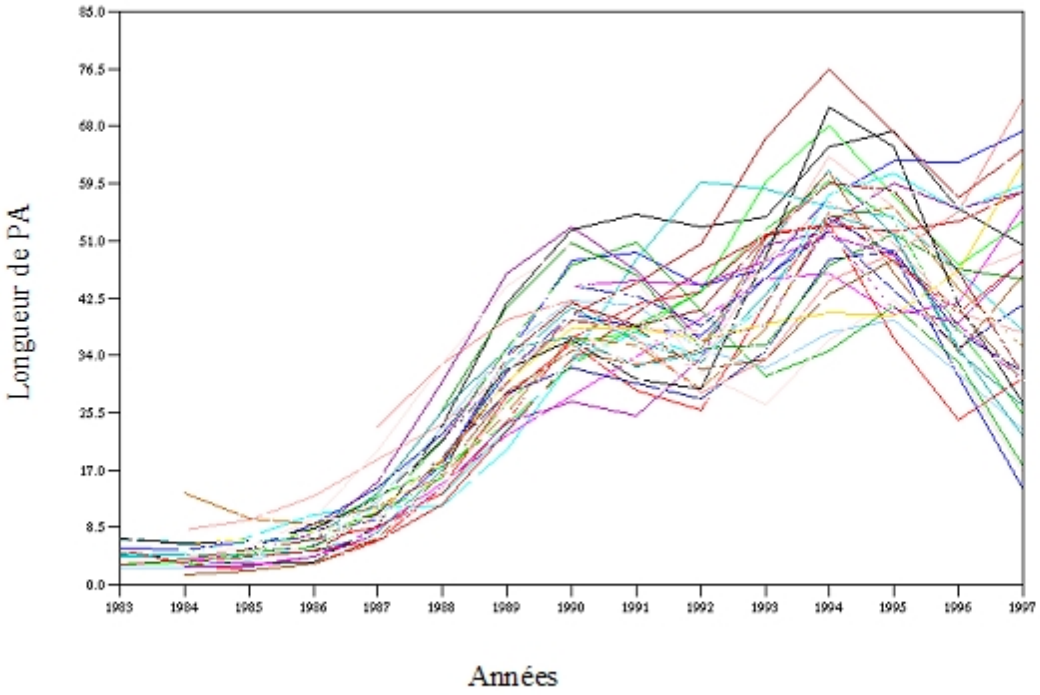

FIG. 1-10 – Tendance extraite des données des chênes sessiles âgés de 15 ans.

et une phase "adulte" sont observées pour les chênes sessiles. La phase de jeunesse est marquée par l'effet de base, et les fluctuations locales, non synchrones entre individus, ne permettent pas de mettre en évidence un effet année. A contrario, au cours de la phase "adulte", les fluctuations locales souvent synchrones entre individus traduisent un effet année très certainement lié au climat. Des covariables climatiques doivent donc être prises en compte dans la modélisation.

Enfin, des effets aléatoires doivent être introduits dans la modélisation. En effet, ils permettent de séparer les différentes sources de variation (celle due aux effets aléatoires et celle due au terme d'erreur) et, par la même, ils permettent de mieux structurer la matrice de variance-covariance des observations. Ils permettent également de prendre en compte l'hétérogénéité entre les individus en modélisant des facteurs non observés comme les maladies ou les attaques de parasites que peuvent connaître par exemple les jeunes arbres.

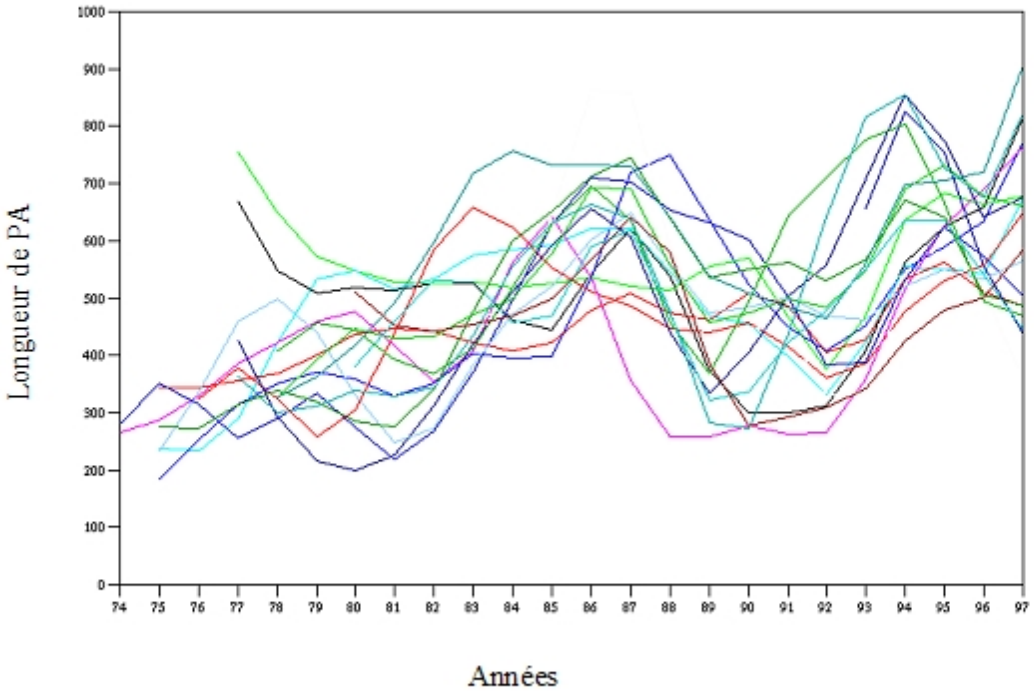

FIG. 1-11 – Tendance extraite des données des chênes sessiles âgés de 29 ans.

Nous souhaitons donc modéliser des données longitudinales qui ont trois caractéristiques principales :

· elles sont structurées en différentes phases successives,

· elles sont soumises à l'influence de covariables climatiques dans certaines phases,

· elles présentent une hétérogénéité inter-individuelle et une hétéroscédasticité dans la structure de variance-covariance.

Le modèle statistique doit donc modéliser la succession de phases de croissance et, sur chacune des phases, les observations doivent être modélisées avec un modèle linéaire mixte qui combine effets fixes et effets aléatoires. Sur chacune des phases, la tendance reflétant un niveau de croissance moyen et des covariables climatiques seront modélisées par des effets fixes, et un effet aléatoire "individu" modélisera l'hétérogénéité inter-individuelle.

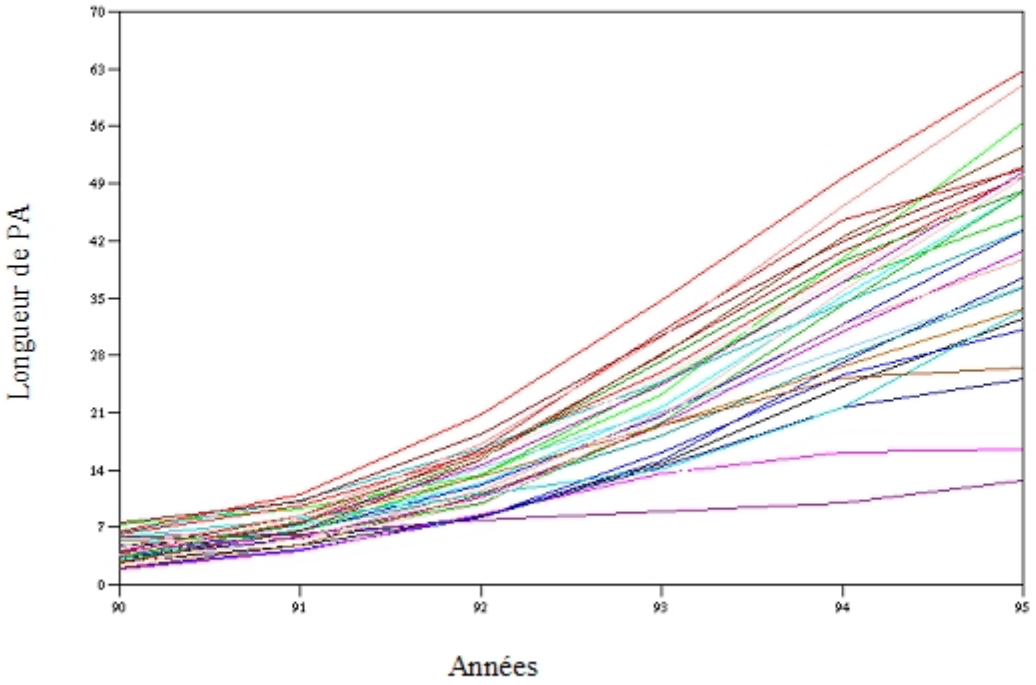

FIG. 1-12 – Tendance extraite des données des pins laricios âgés de 6 ans.

Toutefois, prendre en compte des covariables climatiques dans la modélisation soulève un problème de changement d'échelle entre le pas de temps annuel de la variable réponse du modèle et le pas de temps journalier qu'il est nécessaire d'adopter pour les covariables climatiques. Pourquoi un pas de temps journalier ? Nous allons à présent discuter de ce choix d'échelle ainsi que de la méthode proposée pour calculer des covariables climatiques au pas de temps annuel.

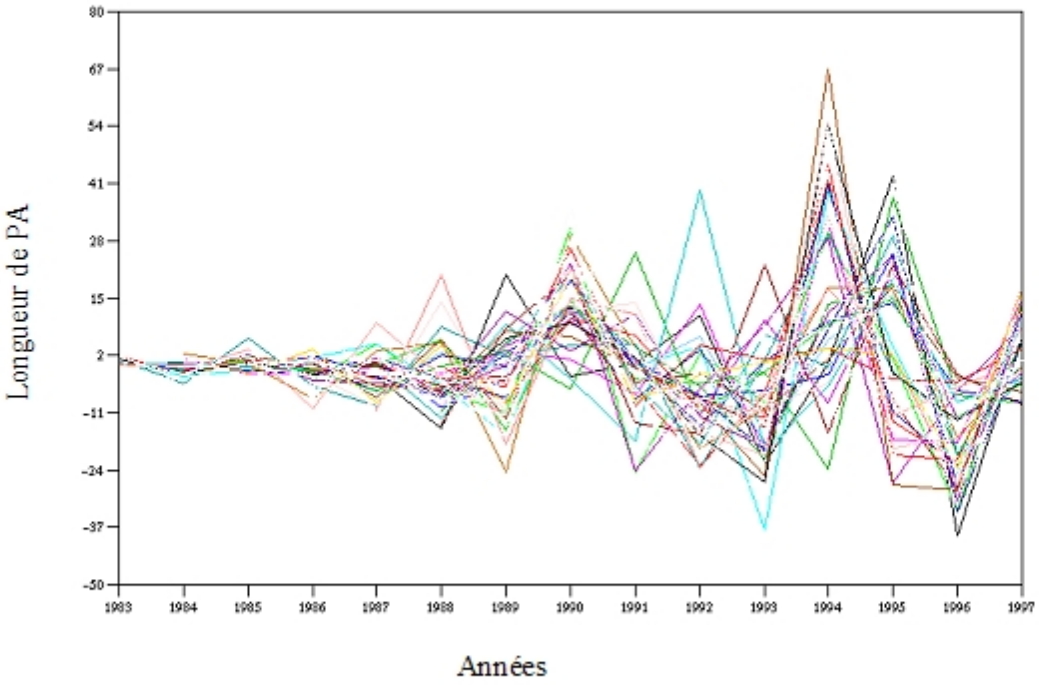

FIG. 1-13 – Fluctuations locales extraites des données des chênes sessiles âgés de 15 ans.

1.3.2 Analyse exploratoire des données climatiques

Le choix de l'échelle

Plusieurs méthodes pour modéliser la croissance secondaire (c'est-à-dire la croissance cambiale) en fonction du climat ont été proposées en dendrochronologie[9] (Monserud, 1986, Guiot, 1986). Ce sont des méthodes d'analyse pour séries chronologiques qui sont principalement basées sur des modèles ARMA. En 1989, Fritts et Swetnam ont proposé de modéliser l'influence du climat avec des régressions multiples et des analyses en composantes principales. Ils calculent des covariables climatiques ayant un pas de temps annuel à partir de données climatiques ayant un pas de temps mensuel.

[9] La dendrochronologie est l'étude de chronoséquences des cernes de croissance d'arbres.

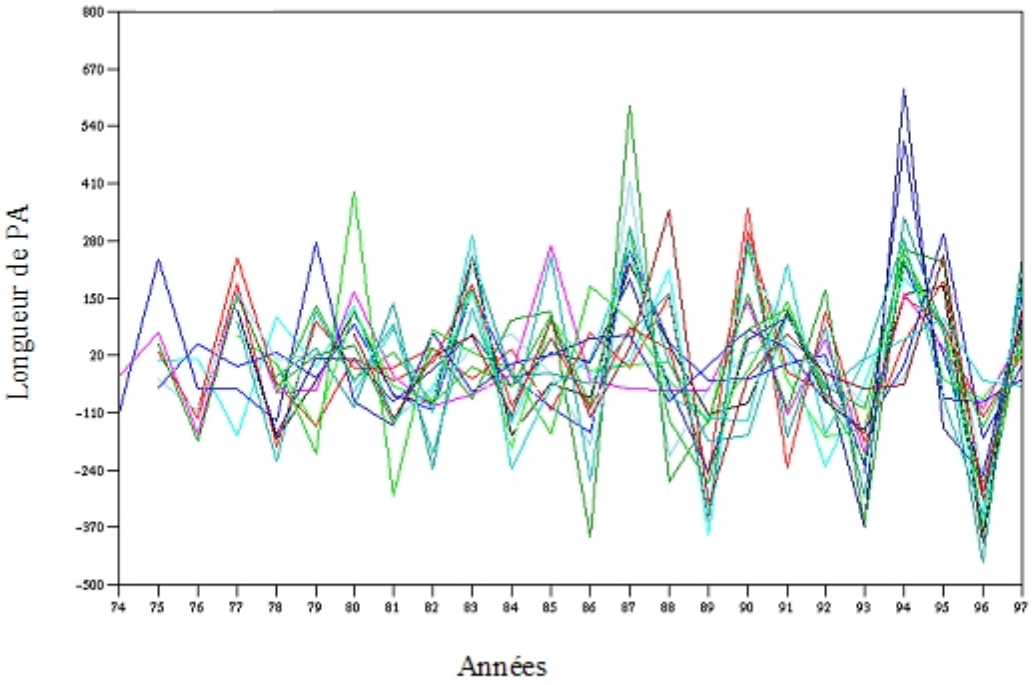

FIG. 1-14 – Fluctuations locales extraites des données des chênes sessiles âgés de 29 ans.

Cependant, les mois de l'année ne sont pas liés aux cycles biologiques de l'arbre ; de plus, une moyenne mensuelle lisse trop les données. Par exemple, une moyenne mensuelle des précipitations peut facilement cacher une importante sécheresse suivie de quelques jours de très importante pluviométrie, ou même dissimuler une importante sécheresse à cheval sur deux mois. De même, une moyenne mensuelle des températures minimales ne permet pas d'observer un gel, ou tout autre phénomène ponctuel. Certaines étapes du cycle biologique des arbres, comme le débourrement – moment d'épanouissement des bourgeons et période où la croissance de l'arbre est très sensible aux conditions climatiques – peuvent avoir lieu sur des périodes de deux ou trois semaines, à cheval sur deux mois. Pour toutes ces raisons, le mois n'est pas une unité de temps appropriée pour prendre en compte les facteurs climatiques pouvant influer sur la croissance des arbres. Pour des raisons biologiques et physiologiques, le jour reste la seule unité de temps pertinente pour

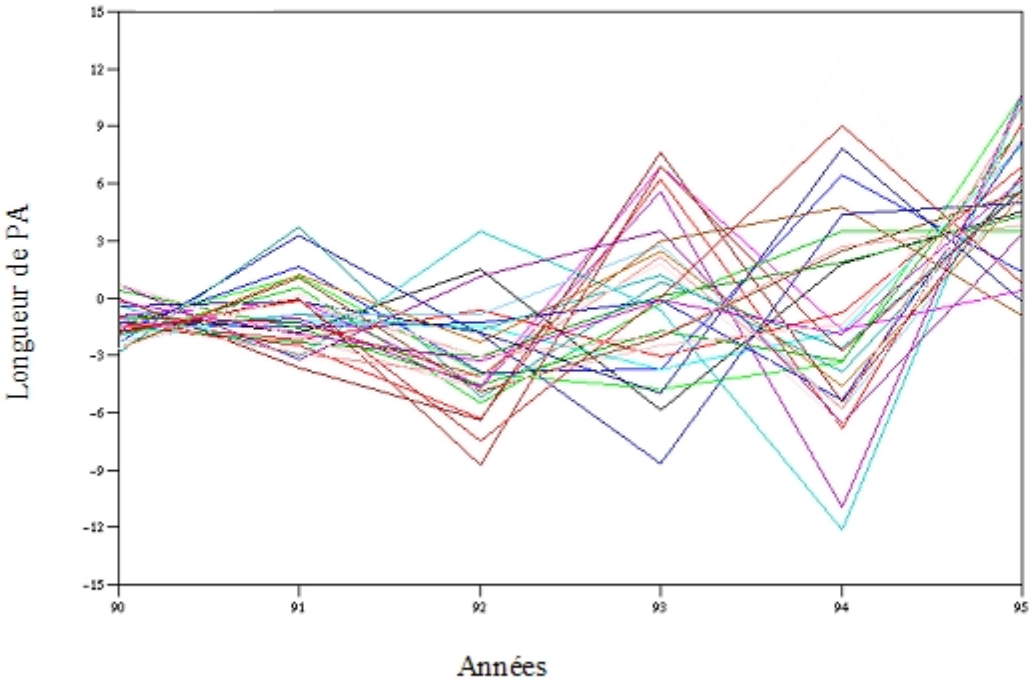

FIG. 1-15 – Fluctuations locales extraites des données des pins laricios âgés de 6 ans.

détecter des évènements climatiques ou tout évènement ponctuel. Nous proposons donc ce choix d'échelle pour les données climatiques.

Les covariables climatiques

L'eau est un facteur limitant pour la plante : si la plante manque d'eau, à partir d'un certain temps, sa croissance va plus ou moins en être affectée. D'après les botanistes, la pluviométrie est le principal facteur susceptible d'influencer de façon significative le développement des arbres forestiers tels que le chêne sessile ou le pin laricio. Toutefois, un stress hydrique sera d'autant plus préjudiciable pour la pousse de l'arbre qu'il se produit lorsque les températures maximales sont élevées. Par conséquent, nous avons choisi d'étudier le facteur pluviométrie couplé aux températures maximales. Mais il pourrait éga-

lement être intéressant d'étudier l'effet des températures minimales car un gel printanier peut également affecter le début de l'allongement de la PA.

Le facteur pluviométrie couplé aux températures maximales

Le chêne sessile et le pin laricio commencent leur premier cycle de croissance approximativement fin avril et ce dernier s'étale ensuite sur deux à trois semaines. Le chêne sessile, qui peut avoir jusqu'à quatre cycles annuels, séparés par des phases de repos plus ou moins longues, les répartira jusqu'à début septembre. Les matières carbonées issues de la photosynthèse serviront alors à alimenter les réserves de l'arbre en vue de la pousse printanière de l'année suivante. Le pin laricio est monocyclique : lorsqu'il aura développé les éléments qu'il avait préformés l'année précédente, il entamera une phase d'organogenèse durant laquelle il préformera les éléments de la pousse suivante en mettant notamment en place les ébauches foliaires qui deviendront des feuilles le printemps suivant. Le pin laricio est par conséquent beaucoup plus sensible aux conditions printanières que le chêne sessile.

Il faut distinguer deux périodes durant lesquelles un stress hydrique peut être préjudiciable à la croissance de l'arbre : les périodes correspondant à l'allongement et à l'organogenèse. Si un stress hydrique est détecté pendant la période d'allongement, alors il est susceptible d'influencer la PA de l'année en cours. En revanche, s'il est détecté au cours de la période d'organogenèse pendant laquelle l'arbre préforme ce qu'il allongera au moment du prochain débourrement, alors il est susceptible d'influencer la PA de l'année suivante.

La détermination précise de chacune des périodes n'est pas aisée et est matière à discussion. Selon l'avis des botanistes, nous choisirons pour les chênes sessiles les mois allant de mars à septembre comme première période critique pour un stress hydrique. La fin du dernier cycle d'allongement et l'organogenèse pour la PA de l'année suivante se chevauchant, nous choisirons la période allant de fin août à fin septembre pour la seconde période critique. Pour les pins laricios, les mois de mars, avril et mai constitueront la période critique pour un stress susceptible d'influencer la PA de l'année en cours. Les mois de juillet et août seront choisis comme période critique pour un stress se produisant au moment de l'organogenèse et susceptible d'influencer la longueur de la PA de l'année suivante.

Dans un premier temps, il nous paraît important de définir différents points de vue de la covariable climatique stress hydrique, quitte ensuite à en éliminer un certain nombre pour ne conserver que les plus pertinents. Les deux premières covariables auxquelles on pense naturellement sont les deux indicatrices qui signalent la présence ou l'absence d'un stress hydrique pour chacune des deux périodes considérées. Nous proposons également des covariables qui quantifient le stress hydrique sur chacune des périodes, ou encore qui décomptent le nombre de jours pendant lesquels il y a stress hydrique. Cependant, les données climatiques ayant un pas de temps journalier, la difficulté majeure est de calculer des covariables ayant un pas de temps annuel, et pouvant être mises en correspondance avec la variable réponse du modèle au pas de temps annuel.

Le modèle bioclimatologique proposé

Les techniques de statistique exploratoire usuellement employées lorsque le nombre de variables explicatives est très élevé relèvent des analyses multivariées telles que l'analyse en composantes principales (ACP) ou l'analyse en composantes multiples (ACM). Cependant, ces outils sont ici inapplicables car l'arbre a une importante inertie, et n'est donc sensible, hormis pour le gel, qu'à des cumuls climatiques sur un nombre variable de jours, et à une date variable. Appliquer une ACP à des données au jour le jour n'est absolument pas approprié. Nous proposons par conséquent un modèle bioclimatologique relativement rudimentaire, résultant de considérations à la fois biologiques et climatologiques. Le but ici n'est pas de construire un modèle bioclimatologique sophistiqué mais juste un modèle prenant en compte les principales contraintes biologiques de l'arbre et permettant de calculer des covariables climatiques au pas de temps annuel. La formalisation de ce modèle et le calcul des différents points de vue de la covariable stress hydrique sont présentés ci-dessous :

Présence/Absence

Dans un premier temps, nous détectons la présence ou l'absence d'un stress hydrique pour chacune des années de croissance de l'arbre, et ce pour les deux périodes qui nous intéressent. Pour cela, nous avons choisi de détecter tous les stress survenus au cours d'une année calendaire et de sélectionner ensuite ceux qui peuvent potentiellement influencer la longueur des PA en fonction de la date à laquelle ils sont détectés.

Pour écrire de façon mathématique ce que représente un stress hydrique pour l'arbre, nous devons tenir compte de l'inertie de l'arbre : un manque d'eau sur une courte période n'est pas nuisible à la croissance de l'arbre car il va puiser dans ses réserves ; en revanche au delà d'un certain seuil où la quantité d'eau est insuffisante, il est en situation de stress hydrique et cela peut devenir préjudiciable pour son développement. Un stress hydrique peut ainsi être traduit mathématiquement par une fonction affine par morceaux. Cette fonction est basée sur des cumuls de précipitations exprimés en millimètres, selon les jours.
• Le premier morceau est une fonction constante du temps sur une certaine durée, et traduit la période où l'arbre ne reçoit pas d'eau et durant laquelle il puise dans ses réserves.
• Le second morceau est une fonction affine croissante, de pente variable. Elle traduit la période où l'arbre reçoit à nouveau une certaine quantité d'eau qui s'avère insuffisante : il est alors en situation de stress hydrique. La période d'inertie et la pente de cette fonction seuil peuvent varier. En effet, si les températures maximales sont supérieures à une température seuil, alors la fonction seuil est contractée d'un facteur dépendant de la température seuil.

On note f, la fonction traduisant un stress hydrique et jouant le rôle de seuil pour la détection des stress ; g désigne la fonction du cumul des précipitations, et t est l'indice temporel désignant un jour.

La fonction f dépend de :
- t le temps,

- α le nombre de jours pendant lesquels le cumul des précipitations reste constant (absence de pluie),
- ρ la pente de la fonction affine croissante,
- τ le nombre de jours pour lesquels la fonction affine croissante est définie et pendant lesquels nous effectuons la détection de stress,
- T la température seuil au delà de laquelle les températures maximales sont susceptibles d'aggraver un stress hydrique,
- η la constante intervenant dans le facteur de contraction appliqué à f lorsque les températures maximales sont supérieures à T.

La fonction du cumul des précipitations g ne dépend que du temps. Pour simplifier, on se ramène à $g(1) = 0$.

Sur l'intervalle $[1, \tau + \alpha]$, la fonction f s'écrit sous la forme :

$$f(t; \alpha, \rho, \tau) = \begin{cases} 0 & t = 1, ..., \alpha, \\ \rho(t - \alpha) & t = \alpha + 1, ..., \alpha + \tau. \end{cases}$$

Une pente égale à 1 ou 2, signifiant que l'arbre doit recevoir au moins 1 ou 2 millimètres d'eau par jour pour ne pas être hydriquement stressé, paraît raisonnable.

En outre, pour détecter un stress hydrique pour une année donnée, nous traçons g la fonction du cumul des précipitations en fonction des jours et nous la "balayons" avec la fonction seuil f. Cela revient en quelque sorte à passer un filtre sur les données de cumuls de précipitation, en décalant au fur et à mesure d'une unité la fonction f.
Si nous notons f_k la fonction seuil correspondant à un décalage de k, alors pour tout k :

$$f_k(t; \alpha, \rho, \tau) = \begin{cases} g(k+1) & t = k+1, ..., k+\alpha, \\ g(k+1) + \rho(t - k - \alpha) & t = k+\alpha+1, ..., k+\alpha+\tau. \end{cases}$$

De plus, si les températures maximales sont supérieures à T sur $[k+1, k+\alpha+\tau]$, on note N le nombre de degrés moyen au dessus de T sur cette période, et f_k est alors contractée d'un facteur $\eta \times \lceil N \rceil$ où $\lceil N \rceil$ est la partie entière supérieure de N. La fonction seuil est alors $f_k\left(t; \frac{\alpha}{\eta\lceil N \rceil}, \rho\eta\lceil N \rceil, \frac{\tau}{\eta\lceil N \rceil}\right)$.

Il suffit ensuite de comparer pour chaque k la position relative des fonctions g et f afin de détecter s'il y a stress hydrique ou pas. S'il existe au moins un k pour lequel la fonction g se trouve en dessous de f sur la période de détection, alors il y a eu stress hydrique et la valeur 1 est attribuée à l'indicatrice.

À titre d'illustration, la figure 1.16 (resp. 1.17) représente la détection d'un stress hydrique pour les mois de mars, avril et mai de l'année 1996 (resp. de l'année 1997) avec les données climatiques relatives aux chênes sessiles. Sur chacune des figures, la fonction g est tracée en bleu et la fonction f est tracée en rouge. Le seuil f correspond à un choix des paramètres, α égal à 21, ρ égal à 1, T égal à 28, et η égal à 1. Puisque, sur chacun des

graphes, la fonction g se trouve à un moment donné au dessous de f, cela signifie qu'après une période sans eau de 21 jours, l'arbre n'a pas reçu une quantité d'eau suffisante et connaît donc une situation de stress hydrique pour la période en question des années 1996 et 1997.

Remarquons que les stress détectés par le modèle sont validés a posteriori en vérifiant qu'ils sont bien cohérents avec les observations météorologiques.

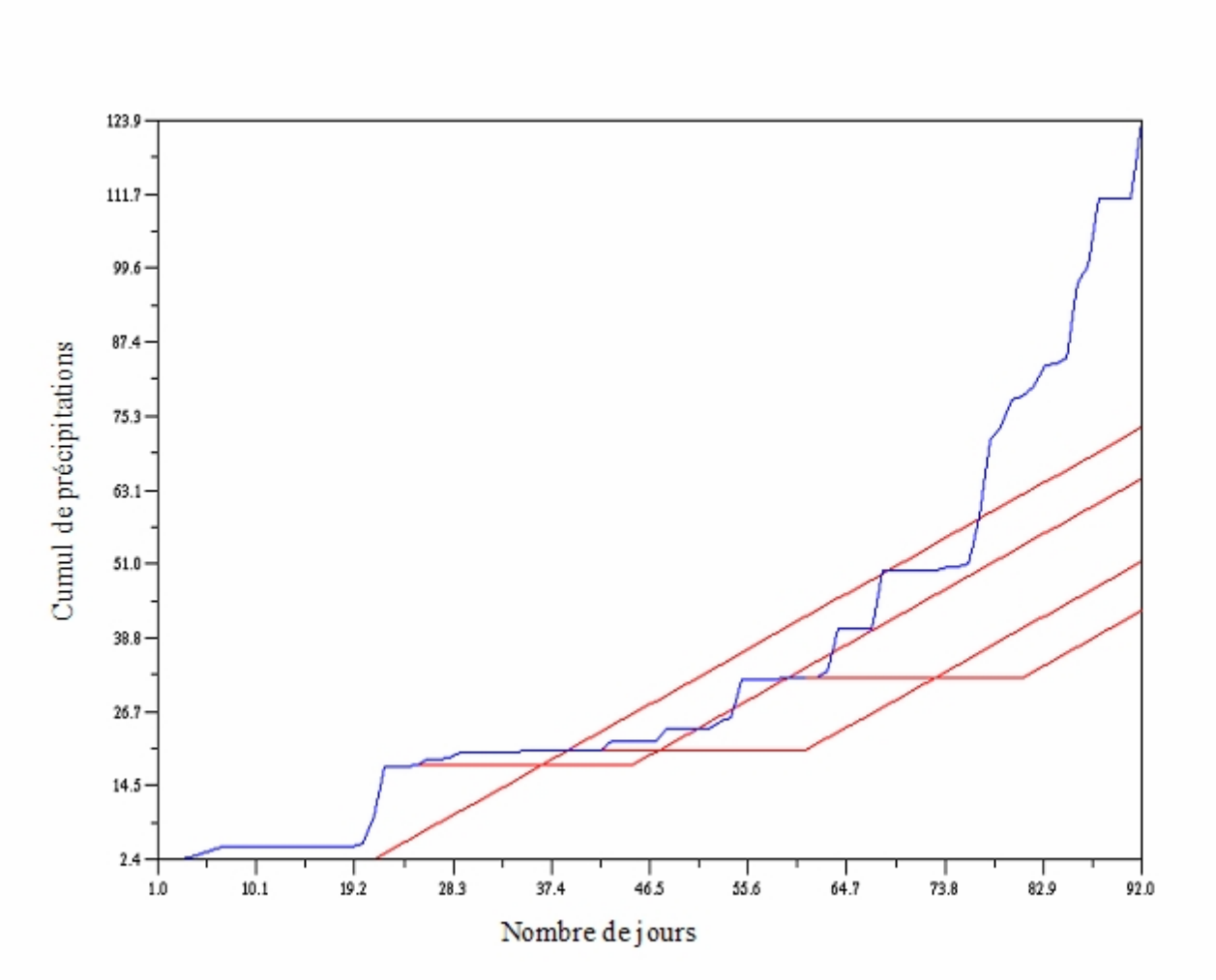

FIG. 1-16 – Détection de stress hydriques pour les mois de mars, avril et mai de l'année 1996 pour les chênes sessiles.

Quantification

Le modèle permet de construire une variable présence/absence d'un stress hydrique, mais il est aussi possible de construire d'autres covariables climatiques qui sont des points

de vue différents sur la même covariable stress hydrique. Par exemple, le stress peut être quantifié avec un pas de temps annuel. Pour cela, pour chacun des stress détectés pour les différents décalages k au cours d'une année, nous calculons un taux de déficit moyen d'eau par jour par rapport à une situation de non stress sur la période de détection. Autrement dit, pour un stress détecté correspondant à un décalage de k, nous calculons d'abord l'aire A_k comprise entre les fonctions g et f. Le calcul de A_k peut s'effectuer de deux manières différentes.

• La première méthode consiste à ne calculer que la partie de l'aire lorsque g est au dessous de f.
Pour tout k,

$$A_k = \sum_{t=k+\alpha+1}^{k+\alpha+\tau} \{f_k(t;\alpha,\rho) - g(t)\} I(f_k(t;\alpha,\rho) \geqslant g(t)),$$

où $I()$ désigne la fonction indicatrice.

• La seconde méthode consiste à tenir compte sur la plage de détection des passages de la fonction g au dessus de f avant de repasser au dessous. On calcule toute l'aire comprise entre f et g sur la plage de détection, en comptant positivement (resp. négativement) l'aire lorsque g est au dessous de f (resp. g au dessus de f), ce qui s'écrit :

$$A_k = \sum_{t=k+\alpha+1}^{k+\alpha+\tau} (f_k(t;\alpha,\rho) - g(t)).$$

Le taux de déficit moyen d'eau sur la période de détection est alors donné dans les deux cas par :

$$D_k = \frac{A_k}{\tau}.$$

Nombre de jours de stress

Le nombre de jours de stress correspondant à D_k est donné par :

$$N_k = \sum_{t=k+\alpha+1}^{k+\alpha+\tau} I(f_k(t;\alpha,\rho) \geqslant g(t))$$

Nous pouvons aussi quantifier avec un pas de temps annuel un stress hydrique ayant lieu au cours de la période d'allongement (resp. d'organogenèse), en prenant le maximum des taux de déficit moyens calculés pour les stress détectés durant la période où a lieu l'allongement (resp. l'organogenèse). Si D_{k_0} est le plus grand des D_k, le nombre de jours correspondant au stress quantifié est alors N_{k_0}.

Nous sommes donc en mesure de calculer des tableaux de covariables climatiques relatives au stress hydrique et ayant un pas de temps annuel, identique à celui de la variable réponse. Rappelons que l'analyse exploratoire des données botaniques a suggéré de modéliser une succession de phases, l'influence de covariables climatiques et une hétérogénéité entre les individus. D'une part la succession de phases peut être modélisée avec une chaîne de Markov, et d'autre part sur chacune des phases, les données peuvent être modélisées par un modèle linéaire mixte. La tendance reflétant le niveau moyen de croissance sur la phase et les covariables climatiques peuvent être modélisées par des effets fixes, et l'hétérogénéité inter-individuelle peut être modélisée par un effet aléatoire "individu". Nous proposons donc de combiner une chaîne de Markov et des modèles linéaires mixtes associés aux états de la chaîne de Markov sous-jacente de sorte que le modèle obtenu, de type Markov caché, satisfasse les trois hypothèses mises en évidence par l'analyse exploratoire. Ce nouveau modèle sera appelé modèle linéaire mixte multiphasique.

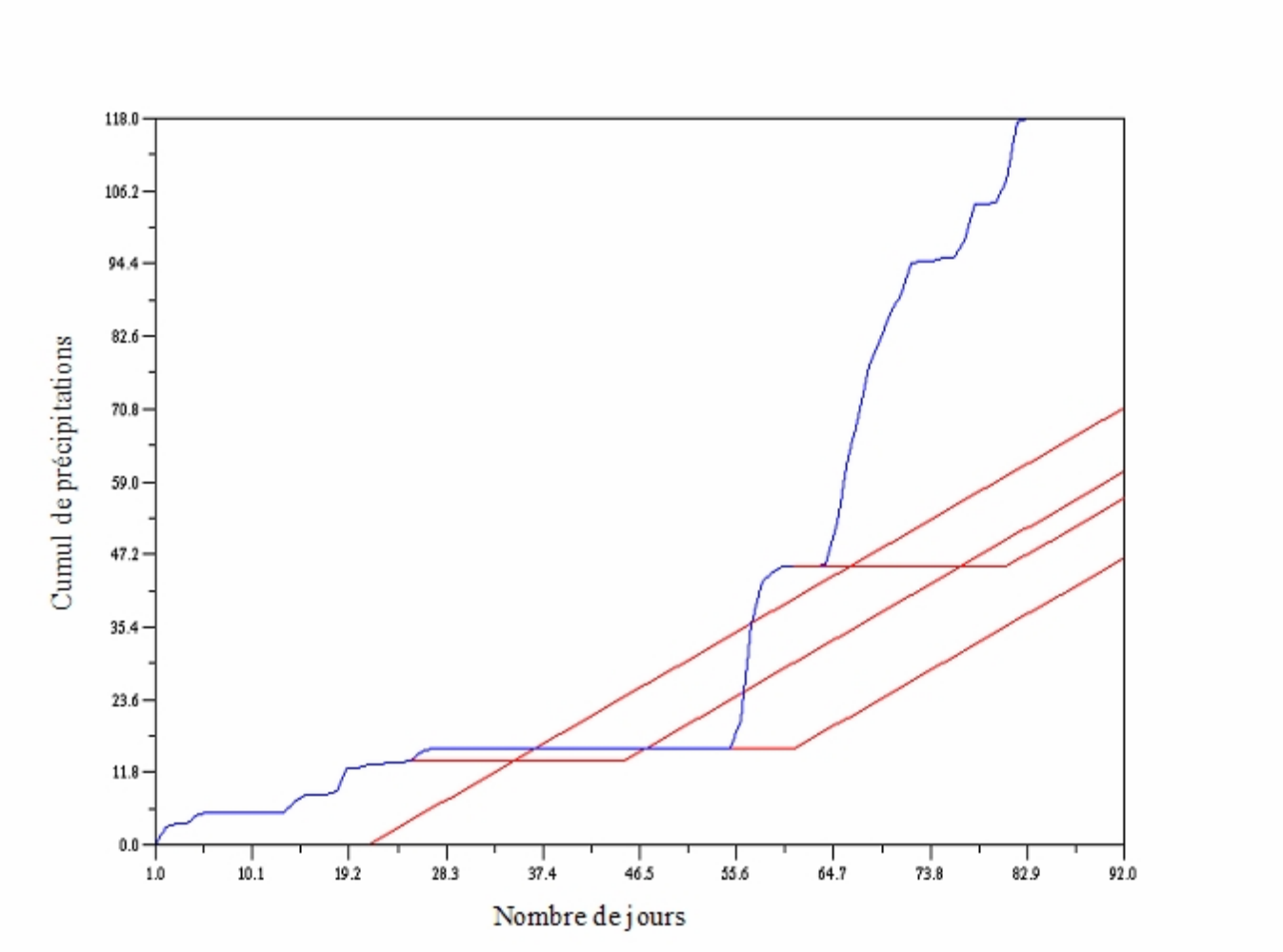

FIG. 1-17 – Détection de stress hydriques pour les mois de mars, avril et mai de l'année 1997 pour les chênes sessiles.

Chapitre 2

Modèles statistiques étudiés et méthodes d'estimation associées

2.1 Introduction

Le modèle linéaire mixte multiphasique résulte de la combinaison de deux types de modèles statistiques : **la chaîne de Markov cachée** et **le modèle linéaire mixte**. Avant de définir de manière précise le modèle linéaire mixte multiphasique et d'exposer les méthodes d'estimation proposées pour les paramètres de ce modèle au chapitre 3, nous présentons dans ce chapitre les deux modèles statistiques qui le composent et les méthodes d'estimation des paramètres de ces deux modèles. Une attention particulière sera portée à l'estimation des paramètres par l'algorithme EM. En effet, la chaîne de Markov cachée et le modèle linéaire mixte étant deux modèles à structure cachée, l'algorithme EM – bien que ne constituant pas la seule méthode d'estimation possible – est bien adapté pour l'estimation des paramètres de ces modèles. Il sera de plus appliqué au chapitre 3 pour l'estimation des paramètres du modèle linéaire mixte multiphasique.

2.2 Le modèle linéaire mixte

Le modèle linéaire mixte, noté L2M (pour Linear Mixed Model) est une extension du modèle linéaire classique où l'on autorise l'introduction d'effets aléatoires. L'histoire du modèle mixte remonte à Ronald Fisher, et plus précisément, à ses travaux sur l'analyse de variance et sur le coefficient de corrélation intra-classe. C'est le papier d'Eisenhart dans Biometrics en 1947 qui a établi une distinction claire entre effets fixes et effets aléatoires. La modélisation des effets pouvant intervenir dans l'explication du phénomène étudié s'est alors enrichie, la partie explicative du modèle s'est raffinée, combinant linéairement ces deux types d'effets. Le modèle linéaire mixte s'est développé par la suite autour de différentes influences parmi lesquelles la génétique animale (C. R. Henderson, D. Gianola, R. L. Quaas, R. Thompson). Les L2M sont en effet très largement utilisés pour des problèmes de génétique quantitative animale et des problèmes d'élevage, pour lesquels les effets aléatoires sont corrélés du fait que des individus ont des gènes en commun. Des

auteurs comme P. Diggle, K. Y. Liang, S. Zeger et N. Laird ont développé les L2M pour la biostatistique. D'autres auteurs ont développé des méthodes pour l'estimation des paramètres de ces modèles : H. O. Hartley et J. N. K. Rao (1967) ont proposé la méthode d'estimation par maximum de vraisemblance pour une large classe de modèles englobant les L2M ; Patterson et Thompson (1971) ont mis au point l'estimation par maximum de vraisemblance restreint (REML). Cette nouvelle méthode d'estimation pour les paramètres des L2M fait disparaître momentanément les effets fixes pour ne maximiser que la vraisemblance concernant les composantes de la variance et indépendante des effets fixes. Actuellement, le modèle linéaire mixte a des applications dans de nombreux domaines, notamment en biologie, en médecine et en agronomie. Il est de plus l'outil de base pour l'analyse de données corrélées telles que les données longitudinales (Diggle *et al.*, 2002 ; Verbeke et Molenberghs, 2000).

Nous définirons dans un premier temps la notion d'effet aléatoire. Suivra un exemple de modélisation mixte combinant des effets fixes et des effets aléatoires. Ensuite, deux approches pour introduire le modèle linéaire mixte seront présentées et illustrées par des applications à l'analyse de données longitudinales.

2.2.1 La modélisation avec effets aléatoires

Les modèles linéaires mixtes permettent d'étudier la variabilité que présentent des données, de façon beaucoup plus précise et plus élaborée que le simple modèle linéaire classique (Searle *et al.*, 1992). Au cours d'une expérience, diverses sources de variations peuvent influencer les valeurs de la variable réponse. Ces différentes sources de variation sont modélisées par des facteurs qui peuvent avoir deux natures : fixe ou aléatoire. On parle alors de facteur à effets fixes et de facteur à effets aléatoires :
• les facteurs à effets fixes ont en général un nombre fini de niveaux et les données se répartissent sur ces différents niveaux. On souhaite en retirer une information concernant l'effet de chaque niveau sur la variable d'intérêt.
• les facteurs à effets aléatoires ont un nombre potentiellement infini de niveaux et les données (en nombre fini) se répartissent sur un échantillon aléatoire de ces niveaux. La façon dont chacun des niveaux influe sur le résultat ne présente pas d'intérêt. En revanche, on souhaite connaître la part de variabilité induite par ces effets.

Prenons l'exemple de notre problématique botanique où l'on souhaite modéliser des longueurs de pousses annuelles en prenant en compte l'influence du climat et l'hétérogénéité inter-individuelle. La covariable climatique est représentée par le facteur stress hydrique. Il s'agit d'un facteur à effets fixes (même s'il n'a pas un nombre fini de niveaux) car les précipitations ne sont pas sélectionnées de manière aléatoire à partir d'une distribution des valeurs des précipitations. En revanche, les individus d'un jeu de données ont été sélectionnés de manière aléatoire. Il est donc justifié de tenir compte dans la modélisation de la variabilité induite par cette sélection aléatoire en introduisant des effets aléatoires individuels.

L'introduction d'effets aléatoires permet de séparer la variation totale en deux parties : la variation due aux effets aléatoires et la variation affectée aux erreurs. L'introduction de différentes composantes de la variance permet ainsi d'apporter plus de précisions sur la variation totale.

Un exemple de modélisation mixte

Pour illustrer la notion d'effet aléatoire, nous présentons un autre exemple de modélisation statistique mixte faisant intervenir à la fois des effets fixes et des effets aléatoires. Cet exemple est inspiré de l'ouvrage de Searle *et al.* (1992).

On considère un essai clinique ayant pour objectif de tester l'efficacité d'un placébo et de 3 médicaments prescrits pour réduire la pression artérielle. Un échantillon de 24 individus ayant une pression artérielle élevée est choisi pour cette étude. Les 24 individus sont répartis en 4 groupes de 6 personnes. On administre le placebo aux individus de l'un des groupes et chacun des 3 médicaments aux individus de chacun des 3 groupes restants. Pour chaque personne, on mesure pour 5 dates différentes la pression artérielle après la prise du médicament. On dispose donc d'un ensemble de données longitudinales, composé de 24 séries de longueur 5. Dans cet exemple, deux facteurs peuvent avoir un effet sur le résultat de l'expérience : le médicament administré et l'individu concerné. On souhaite mesurer l'efficacité des médicaments sur la diminution de la pression artérielle. Le médicament est un facteur à effets fixes à 4 niveaux. De plus, les 24 individus ont été choisis aléatoirement parmi l'ensemble des individus ayant une pression artérielle élevée. L'effet de chacun des individus sur le résultat a peu d'importance, et il importe de mesurer la variabilité des données induite par ces individus. Ceci représente une des composantes de la variation totale et le facteur "individu" est considéré comme un facteur à effets aléatoires.

Dans cet exemple simple, la nature de chacun des facteurs est évidente. Dans la pratique, les modélisations sont souvent bien plus compliquées et la nature des facteurs n'est pas toujours facile à identifier (Searle *et al.*, 1992).

Le modèle linéaire mixte peut être introduit de plusieurs façons ; deux approches sont présentées ici. Chacune de ces approches sera illustrée par un exemple relatif à l'analyse de données longitudinales. Cette présentation s'appuie largement sur les bases théoriques du modèle linéaire mixte décrites par J. L. Foulley (2002).

2.2.2 Définition d'un modèle linéaire mixte

Approche de Rao et Kleffe (1988)

Un modèle linéaire mixte est un modèle linéaire s'écrivant sous la forme :

$$Y = X\beta + \varepsilon, \tag{2.1}$$

où

· Y de dimension $N \times 1$ est le vecteur aléatoire à expliquer,
· β de dimension $p \times 1$ est le vecteur des paramètres à estimer,
· X de dimension $N \times p$ est la matrice d'incidence connue de β,
· $\varepsilon \sim N(0, \Gamma)$ de dimension $N \times 1$ est le vecteur aléatoire d'erreur.
La variable aléatoire ε se décompose en une combinaison linéaire de variables aléatoires structurales (i.e. liées à la structure du dispositif expérimental) non observables $\xi_k; k = 0, 1, \ldots, K$:

$$\varepsilon = \sum_{k=0}^{K} U_k \xi_k = U\xi,$$

où
· $U = (U_0, U_1, \ldots U_k, \ldots U_K)$ de dimension $N \times q$, est une concaténation de matrices d'incidence connues U_k de dimension $(N \times q_k)$, avec $\sum_{k=0}^{K} q_k = q$,
· $\xi = (\xi_0', \xi_1', \ldots \xi_k', \ldots, \xi_K')'$ de dimension $q \times 1$, est le vecteur correspondant des variables structurales $\xi_k = \{\xi_{kl}\}; l = 1, 2, \ldots, q_k$, tel que $\xi \sim N(0, \Sigma_\xi)$. Les ξ_k sont de dimension $q_k \times 1$.
On suppose que Σ_ξ est une fonction linéaire de paramètres $\theta_m; m = 1, \ldots, M$, s'écrivant sous la forme $\Sigma_\xi = \sum_{m=1}^{M} \theta_m F_m$, où les matrices F_m sont des matrices carrées d'ordre q. Dans le cas général, θ_m et F_m ne sont pas soumis à des contraintes particulières si ce n'est qu'ils doivent assurer la positivité de Σ_ξ. On obtient alors une structure linéaire pour la matrice de variance-covariance $\Gamma = U\Sigma_\xi U'$. Cette propriété est caractéristique de ce que l'on entend sous le vocable de "modèle linéaire mixte" qui est tel qu'à la fois, son espérance $\mu = X\beta$, et sa variance $\Gamma = \sum_{m=1}^{M} \Gamma_m \theta_m$, avec $\Gamma_m = UF_mU'$, sont des fonctions linéaires des paramètres.
Les variables aléatoires ξ_k apparaissent en fait comme un moyen de structurer la matrice de variance-covariance Γ de la variable réponse.

En pratique cette modélisation correspond par exemple à un modèle linéaire mixte à K facteurs aléatoires indépendants :

$$y = X\beta + \sum_{k=0}^{K} U_k \xi_k,$$

avec
· $\xi_0 = e$ (le terme d'erreur) ; $U_0 = I_N$ où I_N désigne la matrice identité de dimension N,
· $\xi_k \sim N(0, \sigma_k^2 I_{q_k})$ et $E(\xi_k, \xi_l') = 0, \forall k \neq l$ (autrement dit, les ξ_k sont indépendants entre eux). Ainsi $\Sigma_\xi = \sum_{k=0}^{K} \sigma_k^2 I_{q_k}$.
Par conséquent

$$E(y) = X\beta,$$

et

$$\Gamma = \sum_{k=0}^{K} \sigma_k^2 U_k U_k'$$

sont linéaires en les paramètres β et σ_k^2.

Illustration

Cette approche a été reprise par Diggle *et al.* (2002). Pour mieux modéliser les différentes sources de variation aléatoire présentes dans des données longitudinales, ils prennent en compte dans leur modèle trois sources de variation aléatoire :
• les effets aléatoires : quand les individus sont sélectionnés de manière aléatoire, cela peut entraîner une hétérogénéité entre les individus.
• l'autocorrélation : pour chaque individu, au moins une part de la mesure effectuée peut être interprétée comme une réponse à un processus stochastique variant dans le temps. Ce type de variation est vu comme une corrélation entre deux mesures faites sur le même individu, dépendant de l'intervalle de temps entre les deux mesures. Typiquement, la corrélation diminue quand l'intervalle de temps augmente.
• l'erreur de mesure : le processus de mesure peut induire lui même une variation dans les données.

Il existe plusieurs manières d'incorporer ces trois sources de variation dans différents modèles. Dans le cadre de la décomposition d'un profil individuel, Diggle *et al.* proposent la formulation suivante :

$$Y_i = X_i\beta + \varepsilon_i.$$

Les notations de ce modèle sont identiques à celles de (2.1), si ce n'est qu'elles sont relatives à l'individu i ; Y_i est alors de dimension $n_i \times 1$.
Le terme $\varepsilon_i \sim N(0, \Gamma_i)$ peut se décomposer sous la forme :

$$\varepsilon_i = U_i\xi_i + \varepsilon_{(1)i} + \varepsilon_{(2)i},$$

où
· $U_i\xi_i$ est la partie des effets aléatoires relatifs à l'individu i, $\xi_i \sim N(0, D)$,

· $\varepsilon_{(2)i}$ est une composante d'autocorrélation traduisant qu'une partie au moins du profil individuel observé est la réponse d'un processus stochastique temporel ; $\varepsilon_{(2)i} \sim N(0, \tau^2 H_i)$ où H_i est la matrice d'autocorrélation qui dépend de i uniquement à travers le nombre n_i d'observations et à travers les instants t_{ij} auxquels les mesures sont faites. On suppose de plus que le $(j,k)^{ème}$ élément h_{ijk} de H_i est modélisé sous la forme

$$h_{ijk} = g\left(|t_{ij} - t_{ik}|\right),$$

avec $g(.)$ une fonction décroissante du temps telle que $g(0) = 1$. Cela signifie que la corrélation entre $\varepsilon_{(2)ij}$ et $\varepsilon_{(2)ik}$ dépend uniquement de l'intervalle de temps entre les mesures y_{ij} et y_{ik}, et qu'elle décroît si la longueur de cet intervalle augmente. Habituellement,

les fonctions $g(.)$ sont les fonctions dites d'autocorrélation exponentielle et gaussienne définies respectivement par $g(u) = \exp(-\phi u)$ et $g(u) = \exp(-\phi u^2)$ où ϕ est connu et $\phi > 0$.

· $\varepsilon_{(1)i} \sim N(0, \sigma^2 I_{n_i})$ est l'erreur de mesure. Elle traduit la variation aléatoire supplémentaire liée au processus de mesure.
On suppose que ξ_i, $\varepsilon_{(1)i}$ et $\varepsilon_{(2)i}$ sont indépendants. Par conséquent $\Gamma_i = U_i D U_i' + V_i$ avec $V_i = \tau^2 H_i + \sigma^2 I_{n_i}$.

Selon les caractéristiques des données que l'on souhaite modéliser et selon l'importance que l'on souhaite leur donner, le terme ε_i peut se décomposer de différentes manières, ce qui donne lieu à différents modèles linéaires simples ou différents modèles linéaires mixtes.

. 3 modèles linéaires :

$$Y_i = X_i\beta + \varepsilon_{(1)i},$$

$$Y_i = X_i\beta + \varepsilon_{(2)i},$$

$$Y_i = X_i\beta + \varepsilon_{(1)i} + \varepsilon_{(2)i}.$$

. 3 modèles linéaires mixtes :

$$Y_i = X_i\beta + U_i\xi_i + \varepsilon_{(1)i},$$

$$Y_i = X_i\beta + U_i\xi_i + \varepsilon_{(2)i},$$

$$Y_i = X_i\beta + U_i\xi_i + \varepsilon_{(1)i} + \varepsilon_{(2)i}.$$

Les trois premiers modèles supposent l'absence d'effets aléatoires. Le second de ces modèles suppose une unique source de variation aléatoire traduite par une autocorrélation que l'on modélise par un processus purement aléatoire. A contrario, les trois derniers modèles prennent en compte des effets aléatoires. Le premier de ces trois modèles est le modèle linéaire mixte classique, au sens où il est le plus utilisé y compris pour des applications autres que l'analyse de donnés longitudinales. Le dernier des six modèles, proposé par Diggle *et al.* (2002), est spécifique à l'analyse de données longitudinales. Il a pour originalité de prendre en compte une structure d'autocorrélation, usuellement utilisée pour l'analyse de séries chronologiques. Toutefois, d'autres types de données qui ne sont pas corrélées dans le temps, comme des données de cluster, seraient modélisés en supprimant de ce modèle le processus temporel.

Dans leur ouvrage, les auteurs envisagent différents modèles en combinant les différentes sources de variation, mais ils n'estiment aucun modèle qui comprend à la fois une autocorrélation et des effets aléatoires qui ne sont pas de simples intercepts aléatoires.

Selon eux, dans les applications, l'effet de l'autocorrélation est dominé par la combinaison des effets aléatoires et de l'erreur de mesure. En revanche, Verbeke et Molenberghs (2000) discutent de structures de covariance appropriées lorsque le modèle possède des effets aléatoires autres que de simples intercepts.

Approche marginale de modèles hiérarchiques

Cette approche en deux étapes est due à Lindley et Smith (1972) et a fortement inspiré Verbeke et Molenberghs (2000) dans leur ouvrage dédié aux modèles linéaires mixtes pour les données longitudinales. Notre présentation s'appuie largement sur celle de Verbeke et Molenberghs.

De par leur caractère déséquilibré[1], beaucoup de données longitudinales ne peuvent pas être analysées avec des techniques de régression multivariée. Verbeke et Molenberghs proposent une alternative en deux étapes inspirée du fait que souvent des fonctions de régression linéaire ajustent bien les profils longitudinaux. Dans une première étape, pour chacun des individus, on résume le vecteur des données répétées par le vecteur des coefficients estimés par régression linéaire. Dans une seconde étape, les techniques de régression multivariée sont utilisées pour mettre en relation ces coefficients estimés avec des covariables connues (telles que le traitement...). Le modèle linéaire mixte pour données longitudinales résulte alors de la combinaison des deux modèles obtenus au cours des deux étapes.

Étape 1
On note $Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{in_i})'$ le vecteur aléatoire de dimension $n_i \times 1$, contenant les variables aléatoires relatives aux n_i mesures pour l'individu i. On suppose que Y_i vérifie le modèle de régression linéaire suivant :

$$Y_i = U_i\beta_i + \varepsilon_i, \tag{2.2}$$

où
· U_i de dimension $n_i \times q$ est une matrice de covariables connues, modélisant la manière dont la réponse évolue dans le temps pour l'individu i,
· β_i de dimension $q \times 1$ est le vecteur des coefficients inconnus de la régression linéaire,
· ε_i de dimension $n_i \times 1$ est le vecteur des termes d'erreur. En général, on suppose que les ε_i suivent une loi normale $N(0, \sigma^2 I_{n_i})$ et qu'ils sont indépendants.

Étape 2
On modélise la variabilité observée entre les individus avec un modèle de régression multivariée de la forme :

$$\beta_i = K_i\beta + \xi_i, \tag{2.3}$$

[1] au sens où l'on ne dispose pas du même nombre de mesures pour tous les individus, et/ou au sens où les mesures ne sont pas effectuées à des dates fixes.

où
· β_i est le vecteur des coefficients estimés de la régression linéaire spécifique à l'individu i. Notons qu'en pratique, c'est $\hat{\beta}_i$ qui est utilisé dans (2.3) ; il est le vecteur des coefficients estimés obtenus en ajustant le modèle (2.2) avec une régression linéaire.
· K_i de dimension $q \times p$ est une matrice de covariables connues,
· β de dimension $p \times 1$ est le vecteur des paramètres inconnus de la régression,
Les ξ_i de loi normale $N(0, D)$, avec D une matrice de variance-covariance de forme générale symétrique, sont de dimension $q \times 1$ et sont supposés indépendants.

En remplaçant (2.3) dans (2.2), on obtient

$$Y_i = U_i K_i \beta + U_i \xi_i + \varepsilon_i,$$

ou encore,

$$Y_i = X_i \beta + U_i \xi_i + \varepsilon_i, \tag{2.4}$$

avec $X_i = U_i K_i$, de dimension $n_i \times p$.

Le modèle (2.4) est appelé modèle linéaire mixte où β est le vecteur des effets fixes et ξ_i est le vecteur des effets aléatoires relatifs à l'individu i. Il s'agit également de la définition d'un modèle linéaire mixte donnée par Laird et Ware (1982) qui, dans un cadre général, supposent que $\varepsilon_i \sim N(0, \Sigma_i)$, où Σ_i est une matrice de variance-covariance qui ne dépend de i qu'à travers sa dimension n_i (i.e. l'ensemble des paramètres inconnus de Σ_i ne dépend pas de i). Ils supposent également que tous les ξ_i et tous les ε_i sont indépendants entre eux.

Illustration

Une illustration immédiate de cette approche réside dans la modélisation de profils longitudinaux par les modèles à coefficients aléatoires (Laird et Ware, 1982 ; Verbeke et Molenberghs, 2000). Ce modèle peut être présenté à partir des données de croissance faciale mesurées chez 11 filles et 16 garçons à 4 âges équidistants (8, 10, 12 et 14 ans) et présentées par Pothoff et Roy (1964). Ces données ont été analysées en détail par Verbeke et Molenbergs (1997) et par Foulley *et al.* (2000).

Pour analyser ces données, on ajuste une droite de régression pour chacun des individus. Le modèle de régression linéaire propre à chaque individu s'écrit :

$$y_{ijk} = A_{ik} + B_{ik} t_j + \varepsilon_{ijk},$$

où i désigne l'indice du sexe ($i = 1, 2$ selon que l'individu est de sexe féminin ou masculin), j est celui de la mesure ($j = 1, 2, 3, 4$), t_j est l'âge correspondant, et k désigne l'indice de l'individu intra sexe ($k = 1, \ldots, 11$ pour $i = 1$; $k = 1, \ldots 16$ pour $i = 2$).
Ainsi y_{ijk} est la $j^{ème}$ mesure faite sur le $k^{ème}$ individu de sexe i et A_{ik} et B_{ik} sont respectivement l'intercept et la pente propres à l'individu k de sexe i.

Si on considère à présent que les individus représentent un échantillon aléatoire des enfants de chaque sexe, les A_{ik} et B_{ik} sont des variables aléatoires qui suivent une loi normale :

$$\begin{pmatrix} A_{ik} \\ B_{ik} \end{pmatrix} \sim N\left(\begin{pmatrix} \alpha_i \\ \beta_i \end{pmatrix}, \begin{pmatrix} \sigma_a^2 & \sigma_{ab} \\ \sigma_{ab} & \sigma_b^2 \end{pmatrix}\right).$$

Cela revient à décomposer l'intercept et la pente en la somme de deux parties :

$$A_{ik} = \alpha_i + a_{ik},$$

et

$$B_{ik} = \beta_i + b_{ik},$$

c'est-à-dire en une espérance propre à chaque sexe (respectivement α_i, β_i) et en un écart individuel (respectivement a_{ik}, b_{ik}) considéré comme aléatoire par rapport à l'échantillonnage des individus, et propre à l'individu k de sexe i.
Le modèle se réécrit sous la forme :

$$y_{ijk} = (\alpha_i + \beta_i t_j) + (a_{ik} + b_{ik} t_j) + \varepsilon_{ijk},$$

où $(\alpha_i + \beta_i t_j)$ est la partie fixe et $(a_{ik} + b_{ik} t_j)$ est la partie aléatoire.

Avec cette formulation, un modèle linéaire mixte se définit alors comme un modèle linéaire dans lequel tout ou partie des paramètres associés à certaines unités expérimentales sont traités comme des variables aléatoires du fait de l'échantillonnage de ces unités dans une population plus large.

Ce type d'analyse en deux étapes présente toutefois deux inconvénients. Tout d'abord il y a une perte d'information en résumant le vecteur y_i des mesures pour l'individu i par le vecteur $\widehat{\beta}_i$. Ensuite, on introduit une variabilité aléatoire en remplaçant dans le modèle (2.3) le vecteur β_i par le vecteur des paramètres estimés $\widehat{\beta}_i$. De plus, la matrice de variance-covariance de $\widehat{\beta}_i$ dépend du nombre de mesures disponibles pour l'individu i ainsi que des dates auxquelles les mesures ont été effectuées, et ceci n'est pas pris en compte dans l'étape 2 de l'analyse.

Par la suite, nous retiendrons la première définition d'un L2M, dans laquelle la variation totale est décomposée en différentes sources de variation.

Définissons à présent le modèle de chaîne de Markov cachée qui donne sa structure au modèle linéaire mixte multiphasique.

2.3 Le modèle de chaîne de Markov cachée

Les modèles de chaînes de Markov cachées ont été introduits en 1966 par Baum et Petrie. Cette classe de modèles est construite à partir de la classe des chaînes de Markov, en faisant l'hypothèse qu'une séquence n'est pas directement générée par une chaîne de Markov mais indirectement par des lois de probabilité attachées aux états de la chaîne

de Markov. Ces modèles ont une utilisation dans de nombreux domaines. L'une des premières applications de ces modèles et l'une des plus importantes, à partir de 1970, est la reconnaissance automatique de la parole. Jelinek *et al.* (1975) ont utilisé ces modèles pour la modélisation acoustique de phonèmes. L'utilisation des chaînes de Markov cachées en reconnaissance de la parole est également très largement traitée dans l'article de Rabiner (1989). Les premières applications à l'analyse des séquences d'ADN ont été réalisées par Churchill (1989). Depuis quelques années, ces modèles sont très employés pour les problèmes d'alignements multiples de séquences d'ADN (Krogh *et al.*, 1994 ; Churchill et Lazavera, 1999 ; Durbin *et al.*, 1998). Les arbres de Markov cachés, définis à l'origine par Crouse *et al.* (1988), qui sont une extension simple des chaînes de Markov cachées, ont été utilisés en traitement du signal basé sur l'analyse en ondelettes.

Une chaîne de Markov cachée est un couple de processus stochastiques $\{S_t, Y_t; t = 1, 2, ...\}$:
• le processus sous-jacent $\{S_t; t = 1, 2, ...\}$, est appelé régime, ou processus d'état ou encore processus caché car il est non observable. Le processus caché est une chaîne de Markov homogène (autrement dit on suppose que $P(S_t = j|S_{t-1} = i)$ est indépendant du temps t), d'ordre 1, à valeurs dans l'espace d'états fini $S = \{1, ..., J\}$, de vecteur des probabilités initiales Π, et de matrice des probabilités de transition P :

$$\Pi = (\pi_j = P(S_1 = j); j = 1, \ldots, J),$$

avec $\sum_{j=1}^{J} \pi_j = 1$ et,

$$P = (p_{ij} = P(S_t = j|S_{t-1} = i); i, j = 1, ..., J),$$

avec $\sum_{j=1}^{J} p_{ij} = 1$.

• le second processus $\{Y_t; t = 1, 2, ...\}$, appelé processus d'observation ou d'émission, est une séquence de variables aléatoires conditionnellement indépendantes sachant les états cachés. Dans un cadre général, le processus $\{Y_t\}$ peut être continu ou discret, univarié ou multivarié. Ici, nous supposons qu'il est discret et univarié. Chaque observation Y_t est à valeurs dans l'espace des observations $\{1, ..., U\}$. Conditionnellement aux états s_t, les observations Y_t sont générées de manière indépendante selon la matrice des probabilités d'observation B :

$$B = (b_j(u) = P(Y_t = u|S_t = j); u = 1, ..., U; j = 1, ..., J),$$

avec $\sum_{u=1}^{U} b_j(u) = 1$.

Dans le contexte de notre problématique botanique, le processus caché cherche à modéliser la succession des phases de croissance, et le processus d'observation est la séquence botanique étudiée, à savoir la séquence des longueurs de pousses annuelles.

Notation :
Par la suite, $S_1^t = s_1^t$ (resp. $Y_1^t = y_1^t$) désignera la suite des réalisations $S_1 = s_1, ..., S_t = s_t$ (resp. $Y_1 = y_1, ..., Y_t = y_t$).

Le couple de processus $\{S_t, Y_t; t = 1, 2, ...\}$ est une chaîne de Markov cachée d'ordre 1 si la relation de dépendance suivante est vérifiée :

$$\begin{aligned} P\left(S_t = s_t, Y_t = y_t | S_1^{t-1} = s_1^{t-1}, Y_1^{t-1} = y_1^{t-1}\right) &= P\left(Y_t = y_t, S_t = s_t | S_{t-1} = s_{t-1}\right) \\ &= b_{s_t}\left(y_t\right) p_{s_{t-1}s_t}. \end{aligned}$$

À tout instant, la distribution de chacune des variables aléatoires Y_t ne dépend donc de la chaîne de Markov qu'à travers l'état s_t. Ceci peut se représenter sous forme d'un graphe des indépendances conditionnelles. Donnons au préalable la définition d'un graphe d'indépendance conditionnelle.

Définition d'un graphe d'indépendance conditionnelle

Le graphe d'indépendance conditionnelle (Lauritzen, 1996 ; Smyth *et al.*, 1997) est un modèle graphique qui représente les relations d'indépendance conditionnelle entre des variables aléatoires. Chaque sommet du graphe est une variable aléatoire. S'il existe un arc ayant pour origine Y et pour extrémité Z, alors de manière naïve, nous disons que Y "influence" Z.

Si A, B et C sont trois variables aléatoires sommets d'un graphe orienté et sans cycle, alors A est parent de B s'il existe un arc ayant pour origine A pointant sur B. Le sommet C est enfant de B s'il existe un arc ayant pour origine B et pointant sur C. Le sommet A est dit ancêtre de B s'il est, soit parent de B, soit l'ancêtre d'un parent de B (définition récursive). Le sommet C est dit descendant de B s'il est, soit enfant de B, soit enfant d'un descendant de B (définition récursive).
La notation

$$A \perp B | C$$

signifie que A est indépendant de B sachant C.

La structure d'une chaîne de Markov cachée d'ordre 1 est définie par le graphe d'indépendance conditionnelle de la Figure 2.1.

Le caractère directionnel du paramètre d'index (le temps t) fait que le graphe est naturellement un graphe orienté. Les sommets de ce graphe sont les variables aléatoires S_t ou Y_t. On paramètre les arcs verticaux par les probabilités d'observation et les arcs horizontaux par les probabilités de transition. D'après l'orientation des arcs, la variable aléatoire S_t influence directement Y_t et la variable aléatoire S_{t-1} influence directement S_t. Les réciproques ne sont pas vraies. Dans ce type de graphe, l'absence d'arc entre deux sommets signifie que les deux variables aléatoires concernées sont indépendantes conditionnellement aux autres variables. Ainsi on a les deux relations :

$$\begin{aligned} S_t \perp \{S_1, Y_1, ..., S_{t-2}, Y_{t-2}, Y_{t-1}\} | S_{t-1}, \qquad t > 0, \\ Y_t \perp \{S_1, Y_1, ..., S_{t-1}, Y_{t-1}\} | S_t, \qquad t \geqslant 0. \end{aligned} \tag{2.5}$$

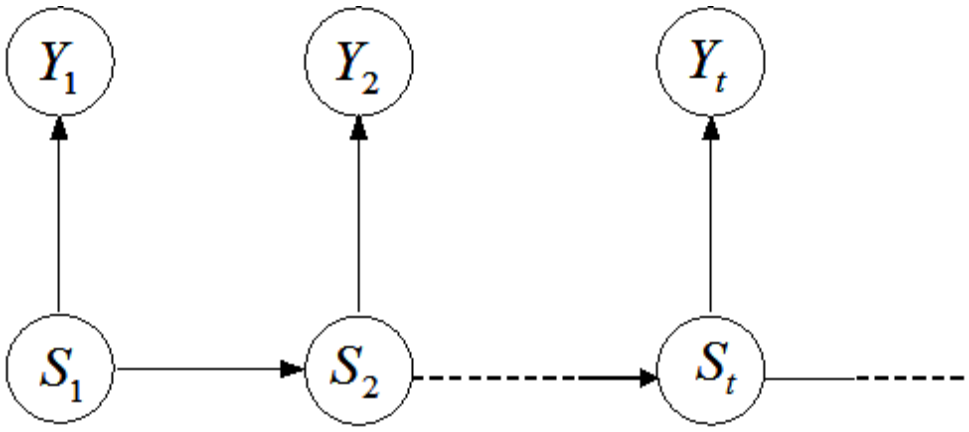

FIG. 2-1 – Graphe d'indépendance conditionnelle d'une chaîne de Markov cachée d'ordre 1.

Une variable aléatoire est donc indépendante de toutes les autres variables aléatoires à l'exception de ses descendants connaissant ses parents. La relation d'indépendance conditionnelle (2.5) signifie que si l'on supprime le sommet S_t, les variables aléatoires $S_{t-\nu}$ ($\nu > 0$) sont séparées des variables aléatoires $S_{t+\nu}$ ($\nu > 0$) et la variable aléatoire Y_t est isolée. Cela peut s'exprimer en disant que l'état à l'instant t étant connu, la connaissance des valeurs prises par les variables aléatoires $S_{t-\nu}$ n'influe pas sur les valeurs prises par les variables aléatoires Y_t et $S_{t+\nu}$. Cette propriété peut se résumer par la factorisation suivante, écrite dans le cas particulier d'une séquence de longueur T :

$$\begin{aligned} P\left(S_1^T = s_1^T, Y_1^T = y_1^T\right) &= P\left(S_1 = s_1\right) P\left(Y_1 = y_1 | S_1 = s_1\right) \\ &\quad \times \prod_{t=2}^{T} P\left(S_t = s_t | S_{t-1} = s_{t-1}\right) P\left(Y_t = y_t | S_t = s_t\right) \\ &= \pi_{s_1} b_{s_1}\left(y_1\right) \prod_{t=2}^{T} p_{s_{t-1} s_t} b_{s_t}\left(y_t\right). \end{aligned}$$

Il s'agit de la factorisation de la loi jointe d'une séquence observée de longueur T.

Remarquons enfin qu'un modèle de mélange où il n'y a pas de dépendance entre les variables aléatoires non observables successives est une chaîne de Markov cachée d'ordre 0.

2.4 Méthodes d'estimation des paramètres des modèles statistiques étudiés

Une fois les deux modèles statistiques qui composent le modèle linéaire mixte multiphasique introduits, intéressons nous à présent à l'estimation des paramètres inconnus de ces modèles, en vue de la présentation des méthodes d'estimation pour le modèle linéaire mixte multiphasique au chapitre suivant. Comme le modèle linéaire mixte multiphasique combine les paramètres de la chaîne de Markov sous-jacente et ceux des modèles linéaires mixtes associés aux états de la chaîne de Markov, il suppose deux types de variables non-observables : les états de la chaîne de Markov et les effets aléatoires. Il donne par conséquent un cadre propice à l'utilisation de l'algorithme EM (Expectation-Maximization) pour l'estimation des paramètres par maximum de vraisemblance. Après une présentation de l'algorithme EM dans un cadre général de données incomplètes, nous rappellerons d'abord les méthodes d'estimation des paramètres d'un modèle linéaire mixte et ensuite les méthodes d'estimation des paramètres d'une chaîne de Markov cachée. Alors que l'algorithme EM constitue la méthode d'estimation classique pour une chaîne de Markov cachée, il est une des méthodes d'estimation possibles pour le modèle linéaire mixte.

2.4.1 Présentation de l'algorithme EM

L'algorithme EM (McLachlan et Krishnan, 1997) est une méthodologie statistique qui s'applique principalement à l'estimation par maximum de vraisemblance pour des problèmes aux données incomplètes, le terme de "données incomplètes" englobant de nombreuses situations : données manquantes, données censurées, variables latentes, paramètres aléatoires. Cet algorithme a été introduit par Baum *et al.* (1970) dans le contexte des chaînes de Markov cachées, puis il a été étendu à des modèles plus généraux par Dempster, Laird, et Rubin (1977). Cet algorithme itératif déterministe permet d'approcher l'estimateur du maximum de vraisemblance lorsque la maximisation directe de la log-vraisemblance du modèle est difficile et se trouve simplifiée par l'augmentation des données. Il est ainsi appelé car chaque itération est composée de deux étapes : une étape E (pour Expectation) et une étape M (pour Maximization).

Nous noterons Y la variable aléatoire ayant pour réalisation y le vecteur des données observées. On suppose que la loi de Y dépend d'un paramètre θ appartenant à l'espace des paramètres Θ. On souhaite estimer θ par maximum de vraisemblance. Le vecteur des données observées y est interprété comme vecteur de données incomplètes. Il peut être complété par le vecteur z des données manquantes (non-observées) pour former le vecteur x des données complètes au sens où une solution simple existe pour l'estimation de θ par maximum de vraisemblance.

Si $L(\theta)$ désigne la vraisemblance du paramètre θ associée au vecteur y et si $f(y, z; \theta)$ désigne la vraisemblance du paramètre associée aux données complètes, alors dans un cadre discret, les deux vraisemblances sont liées par la relation

$$L(\theta) = \sum_z f(y, z; \theta).$$

Remarque : Par la suite, nous parlerons de (log-) vraisemblance pour les données complètes, en omettant de préciser qu'il s'agit de la (log-) vraisemblance du paramètre associée aux données complètes.
Nous noterons $\theta^{(k)}$ la valeur du paramètre θ à l'itération k.

L'itération k de l'algorithme EM se définit par les deux étapes suivantes :

Étape E (Expectation)
L'étape E consiste à calculer $Q\left(\theta|\theta^{(k)}\right)$, l'espérance conditionnelle de la log-vraisemblance du paramètre θ pour les données complètes à l'itération k, connaissant le vecteur des observations (les données incomplètes) et la valeur des paramètres à l'étape courante $\theta^{(k)}$:

$$Q\left(\theta|\theta^{(k)}\right) = E\left\{\log f\left(y, z; \theta\right)|y; \theta^{(k)}\right\}.$$

Étape M (Maximization)
On choisit $\theta^{(k+1)}$ de telle sorte que

$$Q\left(\theta^{(k+1)}|\theta^{(k)}\right) \geq Q\left(\theta|\theta^{(k)}\right), \text{ pour tout } \theta \in \Theta.$$

Cela revient à actualiser la valeur courante du paramètre en maximisant par rapport à θ la fonction obtenue à l'étape E :

$$\theta^{(k+1)} = \arg\max_{\theta\in\Theta} Q\left(\theta|\theta^{(k)}\right).$$

La propriété d'accroissement monotone de la fonction de vraisemblance au cours des itérations est caractéristique de l'algorithme EM et a été montrée par Dempster *et al.* (1977). Pour toute suite $\left(\theta^{(k)}\right)_{k\geq 0}$ générée par l'algorithme EM,

$$L\left(\theta^{(k+1)}\right) \geq L\left(\theta^{(k)}\right), \ \forall k,$$

ce qui signifie qu'à chaque itération k, la nouvelle estimation du paramètre $\theta^{(k+1)}$ améliore l'estimation précédente $\theta^{(k)}$ du point de vue de la vraisemblance.

En général, la vraisemblance des données observées possède plusieurs points stationnaires qui peuvent être des maxima locaux, ou bien un maximum global, ou encore des points selle. La convergence vers l'un d'entre eux dépend de la valeur initiale du paramètre $\theta^{(0)}$. Si la fonction de vraisemblance $L(\theta)$ est unimodale, alors toute suite d'itérations EM converge vers l'unique maximum global, et ce quelle que soit la valeur initiale $\theta^{(0)}$.

Remarquons également que la convergence de la séquence de vraisemblance $\left\{L\left(\theta^{(k)}\right)\right\}$ vers un point stationnaire L^* n'implique pas automatiquement la convergence de la séquence $\left\{\theta^{(k+1)}\right\}$ vers un point stationnaire θ^*.

Enfin, la vitesse de convergence de l'algorithme EM dépend principalement de la proportion d'information manquante sur θ, étant donné que l'on observe uniquement y au lieu d'observer conjointement y et z. Comme cette proportion peut varier pour les différentes composantes de θ, cela explique une vitesse de convergence variable pour ces composantes. Si la log-vraisemblance des données observées se calcule aisément, elle constitue le meilleur moyen pour évaluer la convergence de l'algorithme.

Diverses variantes de l'algorithme EM ont été développées pour contourner certaines difficultés rencontrées dans la mise en oeuvre des étapes E ou M (McLachlan et Krishnan, 1997 ; Foulley, 2002). L'algorithme GEM (Generalized EM) (McLachlan et Krishnan, 1997) et l'algorithme OSL (One Step Late) (Green, 1990 ; McLachlan et Krishnan, 1997) conviennent entre autres lorsque l'étape M de maximisation pose problème. En effet, lorsqu'il est impossible de maximiser globalement la fonction $Q\left(\theta|\theta^{(k)}\right)$ par rapport à θ, autrement dit lorsque l'étape M n'admet pas de solution explicite, une version généralisée de l'algorithme EM, appelé GEM, choisit $\theta^{(k+1)}$ de sorte que $Q\left(\theta^{(k+1)}|\theta^{(k)}\right) \geq Q\left(\theta|\theta^{(k)}\right), \forall k$. La valeur actualisée du paramètre $\theta^{(k+1)}$ ne maximise pas la fonction $Q\left(\theta|\theta^{(k)}\right)$, elle l'augmente simplement. Lorsqu'à l'étape M, $\theta^{(k+1)}$ doit s'obtenir par exemple en résolvant un système d'équations complexe en le paramètre θ, alors l'algorithme OSL propose d'évaluer le système d'équations en $\theta^{(k)}$, valeur du paramètre à l'étape courante.

Lorsque l'étape E ne s'écrit pas de manière analytique et qu'une intégration numérique est difficile à mettre en oeuvre, on a recours à des algorithmes itératifs de simulation tel que l'algorithme SEM (Stochastic EM) (McLachlan et Krishnan, 1997 ; Foulley, 2002). Cet algorithme a été introduit par Celeux et Diebolt (1985) en vue de l'estimation par maximum de vraisemblance des paramètres d'une loi de mélange. Le principe réside dans la maximisation de la log-vraisemblance des données complètes à partir, non pas de son expression analytique, mais d'une évaluation numérique de celle-ci via le calcul de $\log f\left(y, z^{(k)}; \theta\right)$ où $z^{(k)}$ est un échantillon simulé de données manquantes, tiré dans leur distribution conditionnelle. Outre la simplicité du procédé, celui-ci offre l'avantage d'éviter le blocage de l'algorithme en des points stationnaires stables mais indésirables (Celeux *et al.*, 1996). Cette idée a été reprise par Wei et Tanner (1990). Ces auteurs ont proposé l'algorithme EM de Monte Carlo, dénoté MCEM pour calculer la fonction $Q\left(\theta|\theta^{(k)}\right)$ de l'étape E – qui ne s'écrit pas de manière analytique – par le biais d'une approximation de Monte-Carlo. Notons que l'algorithme SEM, qui simule une seule séquence d'états par individu, est un cas particulier de l'algorithme MCEM qui en simule plusieurs par individu. La propriété de croissance monotone de la fonction de vraisemblance n'est plus valable pour ce type d'algorithmes par simulation.

2.4.2 Estimation des paramètres d'un modèle linéaire mixte

Rappelons brièvement les hypothèses d'un modèle linéaire mixte dans un cadre général :

$$Y = X\beta + U\xi + \varepsilon, \tag{2.6}$$

où
· $\varepsilon \sim N(0, \sigma_0^2 V_0)$,
· $\forall j \in \{1, ..., K\}, \xi_j \sim N(0, \sigma_j^2 A_j)$. Les ξ_j sont indépendants entre eux et indépendants de ε. Ainsi $\xi \sim N(0, D)$ avec D diagonale par blocs, $D = \left\{\sigma_j^2 A_j\right\}_{j=1,...,K}$.

Par conséquent,

$$E(Y) = X\beta,$$

et

$$\text{var}(Y) = \sigma_0^2 V_0 + UDU' = \sum\nolimits_{j=0}^{K} \sigma_j^2 V_j = \Gamma,$$

avec $V_j = U_j A_j U_j'$, $\forall j \in \{1, ..., K\}$.
Nous noterons $\sigma^2 = (\sigma_0^2, \sigma_1^2, ..., \sigma_K^2)$ le vecteur des composantes de la variance et $|\Gamma|$ le déterminant de la matrice Γ.

La fonction de vraisemblance pour le vecteur des observations y (de taille N) s'écrit :

$$f\left(y; \beta, \sigma^2\right) = \frac{1}{(2\pi)^{N/2} |\Gamma|^{1/2}} \exp\left\{-\frac{1}{2}(y - X\beta)' \Gamma^{-1} (y - X\beta)\right\},$$

et la log-vraisemblance :

$$\log f\left(y; \beta, \sigma^2\right) = -\frac{N}{2}\log(2\pi) - \frac{1}{2}\log(|\Gamma|) - \frac{1}{2}(y - X\beta)' \Gamma^{-1} (y - X\beta). \tag{2.7}$$

Nous souhaitons estimer à la fois le vecteur des effets fixes β, et les $K + 1$ composantes de la variance, $\sigma_j^2, j = 0, ..., K$.

Les méthodes d'estimation des paramètres d'un modèle linéaire mixte sont exposées dans l'ouvrage de Searle *et al.* (1992) consacré aux modèles linéaires mixtes. Nous présentons ici deux algorithmes d'estimation : l'algorithme EM, et la méthode dite de Henderson. Ils permettent de déterminer les estimations des paramètres d'un L2M par maximum de vraisemblance (ML – Maximum Likelihood) et par maximum de vraisemblance restreint (REML – Restricted Maximum Likelihood). Après une brève présentation des méthodes ML et REML, nous présentons également brièvement la méthode de Henderson et nous

décrivons de manière précise l'algorithme EM dans le cas particulier de l'estimation ML des paramètres d'un L2M dans un contexte de données longitudinales.

Estimation par maximum de vraisemblance :

On dérive la log-vraisemblance (2.7) par rapport à β et par rapport aux $K+1$ paramètres σ_j^2 ; $j = 0, ...K$. Les équations obtenues n'étant pas linéaires en les paramètres σ_j^2, un algorithme itératif est mis en oeuvre. À partir de valeurs initiales de σ_j^2, on itère la résolution des $K+1$ équations relatives aux composantes de la variance jusqu'à convergence. Puis, à l'aide des valeurs des composantes de la variance obtenues, on résout l'équation relative aux effets fixes pour avoir l'estimation de β.

Estimation par maximum de vraisemblance restreint :

Cette méthode d'estimation spécifique aux L2M a été développée par Patterson et Thompson (1971). On supprime provisoirement les effets fixes pour ne maximiser que la partie de la vraisemblance concernant les composantes de la variance. Pour éliminer la partie des effets fixes, on projette le modèle sur l'orthogonal du sous-espace vectoriel engendré par les colonnes de X. L'estimation par maximum de vraisemblance restreint n'est autre que l'estimation par maximum de vraisemblance dans le modèle projeté. Après une estimation itérative des composantes de la variance, on estime le vecteur des effets fixes β.
L'estimation REML a également une interprétation bayésienne (Harville, 1974). Elle repose sur le concept de vraisemblance marginale. Après élimination de β par intégration de la fonction de vraisemblance, on obtient la vraisemblance marginale des σ_j^2. Le système des $K+1$ équations obtenues par dérivation de cette log-vraisemblance marginale est identique au système d'équations obtenues par projection du modèle sur l'orthogonal de X.
Cette méthode d'estimation a l'avantage sur la méthode ML de tenir compte de la perte de degrés de liberté occasionnée par l'estimation des effets fixes. Toutefois, dans notre étude, nous privilégierons l'estimation ML car nous sommes amenés à comparer des modèles qui diffèrent par la structure de leurs effets fixes et la comparaison de la log-vraisemblance de tels modèles n'est pas fondée pour une estimation REML.

Remarquons que les méthodes d'estimation par ML et REML ne se situent pas sur un même plan que la méthode dite de Henderson et l'algorithme EM. En fait, l'algorithme EM et la méthode de Henderson (basée sur la résolution des équations de Henderson) sont des alternatives à la résolution itérative des systèmes d'équation obtenus dans le cadre de l'estimation ML et REML.

La méthode de Henderson

Henderson *et al.* (1959) ont proposé des équations permettant d'obtenir simultanément la prédiction BLUP (Best Linear Unbiased Predictor) de ξ, et l'estimation BLUE

(Best Linear Unbiased Estimator) de β (estimation équivalente au maximum de vraisemblance sous les hypothèses de normalité adéquates). Ces équations sont également appelées équations du modèle mixte (MME – Mixed Model Equation). Toutefois, la résolution des équations de Henderson nécessite de connaître les valeurs des composantes de la variance σ_j^2. Pour calculer les estimations ML et REML, Harville (1977) a proposé un schéma itératif alternant, pour des valeurs de σ_j^2, la résolution des équations de Henderson et, pour des valeurs de β et ξ, la résolution des équations relatives aux composantes de la variance. Laird (1981) a montré que si $V_0 = I_N$ et si la matrice D est diagonale par blocs, alors l'algorithme EM est équivalent à l'algorithme de Henderson décrit par Harville (1977).

L'algorithme EM

Le L2M étant un modèle à structure cachée (le vecteur des effets aléatoires ξ est non-observé), l'algorithme EM, présenté à la section 2.4.1 peut également s'appliquer à l'estimation des paramètres de ce modèle. Il permet de déterminer les estimations ML et REML (Laird et Ware, 1982). Le vecteur $x = (y', \xi')'$ est alors le vecteur des données complètes. Notons qu'il existe deux schémas itératifs possibles pour l'estimation ML avec l'algorithme EM. Le premier donne à chaque itération des nouvelles valeurs à chacun des paramètres, alors que le deuxième itère uniquement pour des estimations successives des σ_j^2, et estime β à la convergence. Pour l'estimation REML où on ne se préoccupe du β qu'après l'estimation des composantes de la variance, on utilise le second schéma.

Comme l'algorithme EM sera présenté au chapitre 3 pour l'estimation d'un modèle linéaire mixte multiphasique, nous choisissons d'exposer en détail l'algorithme EM pour l'estimation d'un L2M. De plus, comme nous étudions des données longitudinales, nous allons écrire l'algorithme EM dans le cas particulier de l'estimation ML des paramètres d'un L2M modélisant la séquence d'observations y_a relative à l'individu a ($a = 1, ...N$ où N est le nombre d'individus) de longueur T_a.

On suppose que l'observation $y_{a,t}$ relative à l'individu a à l'instant t est modélisée par le L2M suivant :

$$y_{a,t} = \beta + \xi_a + \varepsilon_{a,t} \qquad t = 1, ..., T_a, \tag{2.8}$$

où
· β est le terme des effets fixes,
· $\xi_a \sim N(0, \tau^2)$ est l'effet aléatoire relatif à l'individu a,
· $\varepsilon_{a,t} \sim N(0, \sigma^2)$ est le terme d'erreur. Les $\varepsilon_{a,t}$ sont supposés indépendants entre eux et indépendants de ξ_a.
Par conséquent, $y_{a,t} \sim N(\beta, \tau^2 + \sigma^2)$.
Remarquons que τ^2 (resp. σ^2) est ici l'équivalent du σ_1^2 (resp. σ_0^2) décrit pour le modèle (2.6). On est de plus dans le cas particulier où il n'y a qu'un seul effet aléatoire ($K = 1$).

Si $Y_a = (Y_{a,1}, ..., Y_{a,T_a})'$ est le vecteur aléatoire correspondant à une séquence de longueur T_a et ayant pour réalisation le vecteur $y_a = (y_{a,1}, ..., y_{a,T_a})'$, alors (2.8) s'écrit sous la forme matricielle suivante :

$$Y_a = X_a\beta + U_a\xi_a + \varepsilon_a,$$

où
· X_a et U_a de dimension $T_a \times 1$ sont respectivement les matrices d'incidence de β et ξ_a. X_a et U_a sont identiques et sont composés uniquement de 1.
· $\varepsilon_a = (\varepsilon_{a,1}, \varepsilon_{a,2}, ..., \varepsilon_{a,T_a})'$ de dimension $T_a \times 1$ est le vecteur aléatoire d'erreur, $\varepsilon_a \sim N(0, V_a)$ où $V_a = \sigma^2 I_{T_a}$.

Avec les notations définies ci-dessus, la matrice de variance-covariance de Y_a s'écrit sous la forme :

$$\text{var}\,(Y_a) = \tau^2 U_a U_a' + V_a.$$

Nous noterons $\Gamma_a = \tau^2 U_a U_a' + V_a$.

Écriture de la densité des données complètes
Par la suite $\theta = (\beta, \tau^2, \sigma^2)$ désigne le vecteur des paramètres à estimer.

Pour la spécification du problème aux données complètes, nous supposons que la séquence y_a est observée ainsi que l'effet aléatoire ξ_a. La densité des données complètes pour une séquence de longueur T_a s'écrit :

$$f\,(y_a, \xi_a; \theta) = f\,(y_a|\xi_a; \theta)\, f\,(\xi_a; \theta)$$

Explicitons le terme de la densité de la séquence des observations conditionnellement à l'effet aléatoire : $f\,(y_a|\xi_a; \theta)$.
Nous savons que

$$Y_a|\xi_a \sim N(X_a\beta + U_a\xi_a, \sigma^2 I_{T_a}). \tag{2.9}$$

La covariance entre deux observations sachant l'effet aléatoire est donc donnée par :

$$cov(y_{a,t}, y_{a,t'}|\xi_a) = \begin{cases} \sigma^2 & \text{si } t = t', \\ 0 & \text{sinon.} \end{cases}$$

Par conséquent, les observations de la séquence sont conditionnellement indépendantes sachant l'effet aléatoire et,

$$f\,(y_a|\xi_a; \theta) = \prod_{t=1}^{T_a} f\,(y_{a,t}|\xi_a; \theta)\,.$$

De plus, d'après (2.9), $y_{a,t}|\xi_a \sim N(\beta + \xi_a, \sigma^2)$, alors,

$$f\,(y_{a,t}|\xi_a; \theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_{a,t} - \beta - \xi_a)^2}{2\sigma^2}\right).$$

La densité des données complètes pour une séquence de longueur T_a s'écrit donc :

$$f\left(y_a, \xi_a; \theta\right) = \left\{ \prod_{t=1}^{T_a} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(y_{a,t} - \beta - \xi_a\right)^2}{2\sigma^2} \right) \right\} \frac{1}{\sqrt{2\pi}\tau} \exp\left(-\frac{\xi_a^2}{2\tau^2} \right).$$

Il s'ensuit que la log-vraisemblance associée aux données complètes s'écrit :

$$\log f\left(y_a, \xi_a; \theta\right) = -\frac{\left(T_a + 1\right)}{2} \log 2\pi - \log \tau - \frac{\xi_a^2}{2\tau^2} + \sum_{t=1}^{T_a} \left(-\log \sigma - \frac{\left(y_{a,t} - \beta - \xi_a\right)^2}{2\sigma^2} \right).$$

Étape E

Cette étape consiste à calculer l'espérance conditionnelle de la log-vraisemblance des données complètes connaissant la séquence des observations et la valeur des paramètres à l'itération k.

$$\begin{aligned} E\left(\log f\left(y_a, \xi_a; \theta\right) | Y_a = y_a; \theta^{(k)} \right) = & -\frac{\left(T_a + 1\right)}{2} \log 2\pi - \log \tau - \frac{1}{2\tau^2} E\left(\xi_a^2 | Y_a = y_a; \theta^{(k)} \right) \\ & + \sum_{t=1}^{T_a} \left(-\log \sigma - \frac{1}{2\sigma^2} E\left(\left(y_{a,t} - \beta - \xi_a\right)^2 | Y_a = y_a; \theta^{(k)} \right) \right), \end{aligned} \tag{2.10}$$

avec

$$\begin{aligned} E\left(\left(y_{a,t} - \beta - \xi_a\right)^2 | Y_a = y_a; \theta^{(q)} \right) = & \, y_{a,t}^2 - 2y_{a,t}\beta - 2y_{a,t} E(\xi_a | Y_a = y_a; \theta^{(k)}) + \beta^2 \\ & + E(\xi_a^2 | Y_a = y_a; \theta^{(k)}) + 2\beta E(\xi_a | Y_a = y_a; \theta^{(k)}). \end{aligned}$$

Étape M

Cette étape consiste à maximiser l'espérance conditionnelle de la log-vraisemblance des données complètes en chacun des paramètres à estimer.

Estimation du paramètre β

L'annulation de la dérivée de (2.10) par rapport à β conduit à :

$$\beta^{(k+1)} = \frac{\sum_{t=1}^{T_a} \left(y_{a,t} - E(\xi_a | Y_a = y_a; \theta^{(k)}) \right)}{T_a}.$$

De plus, la loi du vecteur $\begin{pmatrix} \xi_a \\ Y_a \end{pmatrix}$ est donnée par :

$$\begin{pmatrix} \xi_a \\ Y_a \end{pmatrix} \sim N_{T_a+1}\left(\begin{pmatrix} 0 \\ X_a\beta \end{pmatrix}, \begin{pmatrix} \tau^2 & \tau^2 U_a' \\ \tau^2 U_a & \Gamma_a \end{pmatrix}\right).$$

Ainsi

$$\xi_a | Y_a = y_a \sim N\left(\tau^2 U_a' \Gamma_a^{-1}(Y_a - X_a\beta), \tau^2 - \tau^4 U_a' \Gamma_a^{-1} U_a\right). \tag{2.11}$$

Il s'ensuit que :

$$E\left(\xi_a | Y_a = y_a\right) = \tau^2 U_a' \Gamma_a^{-1}\left(Y_a - X_a\beta\right), \tag{2.12}$$

et

$$\beta^{(k+1)} = \frac{\sum_{t=1}^{T_a}\left(y_{a,t} - \tau^{2(q)} U_a' \Gamma_a^{-1(k)}\left(Y_a - X_a\beta^{(k)}\right)\right)}{T_a}. \tag{2.13}$$

Estimation du paramètre τ^2
L'annulation de la dérivée de (2.10) par rapport à τ^2 conduit à :

$$\tau^{2(k+1)} = E\left(\xi_a^2 | Y_a = y_a; \theta^{(k)}\right).$$

De (2.11) et (2.12), nous déduisons :

$$E\left(\xi_a^2 | Y_a = y_a\right) = \tau^2 - \tau^4 U_a' \Gamma_a^{-1} U_a + \tau^4 \left(Y_a - X_a\beta\right)' \Gamma_a^{-1} U_a U_a' \Gamma_a^{-1}\left(Y_a - X_a\beta\right). \tag{2.14}$$

D'où

$$\begin{aligned} \tau^{2(k+1)} &= \tau^{2(k)} - \tau^{4(k)} U_a' \Gamma_a^{-1(k)} U_a \\ &\quad + \tau^{4(k)}\left(Y_a - X_a\beta^{(k)}\right)' \Gamma_a^{-1(k)} U_a U_a' \Gamma_a^{-1(k)}\left(Y_a - X_a\beta^{(k)}\right). \end{aligned} \tag{2.15}$$

Estimation du paramètre σ^2
L'annulation de la dérivée de (2.10) par rapport à σ^2 conduit à :

$$\sigma^{2(k+1)} = \frac{\sum_{t=1}^{T_a} E\left(\left(y_{a,t} - \beta - \xi_a\right)^2 | Y_a = y_a; \theta^{(k)}\right)}{T_a}.$$

Remarquons que $y_{a,t} - \beta - \xi_a = \varepsilon_{a,t}$. Nous souhaitons donc déterminer $E\left(\varepsilon_{a,t}^2 | Y_a = y_a; \theta^{(k)}\right)$.
La loi du couple $\begin{pmatrix} \varepsilon_{a,t} \\ Y_a \end{pmatrix}$ est donnée par :

$$\begin{pmatrix} \varepsilon_{a,t} \\ Y_a \end{pmatrix} \sim N_{T_a+1}\left(\begin{pmatrix} 0 \\ X_a\beta \end{pmatrix}, \begin{pmatrix} \sigma^2 & C' \\ C & \Gamma_a \end{pmatrix}\right)$$

où

$C = \begin{pmatrix} 0 \\ \vdots \\ \sigma^2 \\ 0 \\ \vdots \end{pmatrix}$ est le vecteur de dimension $T_a \times 1$, composé de $T_a - 1$ zéros et de la valeur σ^2 sur sa $t^{ème}$ ligne.
Ainsi

$$E(\varepsilon_{a,t}|Y_a = y_a) = C'\Gamma_a^{-1}\left(Y_a - X_a\beta\right).$$

Il s'ensuit que :

$$E(\varepsilon_{a,t}^2|Y_a = y_a) = \sigma^2 - C'\Gamma_a^{-1}C + \left(Y_a - X_a\beta\right)'\Gamma_a^{-1}CC'\Gamma_a^{-1}\left(Y_a - X_a\beta\right), \tag{2.16}$$

et

$$\sigma^{2(k+1)} = \frac{\sum_{t=1}^{T_a}\sigma^{2(k)} - C'^{(k)}\Gamma_a^{-1(k)}C^{(k)} + \left(Y_a - X_a\beta^{(k)}\right)'\Gamma_a^{-1(k)}C^{(k)}C'^{(k)}\Gamma_a^{-1(k)}\left(Y_a - X_a\beta^{(k)}\right)}{T_a} \tag{2.17}$$

Remarquons qu'en supposant que les N individus sont indépendants, la généralisation de l'algorithme EM pour l'ensemble des N séquences est immédiate. Il suffit de rajouter au numérateur et au dénominateur de (2.13), (2.15) et (2.17) une sommation sur l'indice a variant de 1 à N.

L'algorithme EM est un algorithme performant car il se généralise aisément à des contextes difficiles. Toutefois il peut s'avérer très lent et même si on est sûr de faire croître la valeur de la fonction de vraisemblance et de rester dans l'espace des paramètres, il peut tout à fait rester coincé en un maximum local de la fonction. La méthode de Henderson reste sans doute la méthode la plus performante pour l'estimation des paramètres d'un L2M car elle ne nécessite pas d'inversion lourde de matrices, et en particulier pour la matrice de variance-covariance Γ (la résolution des équations de Henderson nécessite l'inversion de matrices souvent diagonales).

2.4.3 Estimation des paramètres d'une chaîne de Markov cachée

Les états d'une chaîne de Markov cachée n'étant pas observables, le problème de l'estimation des paramètres peut être considéré comme un problème aux données incomplètes. La principale méthode d'estimation des paramètres par maximum de vraisemblance est l'algorithme EM qui, pour les chaînes de Markov cachées, est également connu sous le nom d'algorithme de Baum-Welch (Baum *et al.*, 1970). Toutefois, d'autres algorithmes peuvent

fournir des estimations par maximum de vraisemblance. Deux versions stochastiques de l'algorithme EM, l'algorithme SEM (Stochastic EM) (Celeux et Dielbolt, 1985 ; Foulley, 2002) et l'algorithme EM à la Gibbs – introduit par Robert *et al.* (1993) dans le cadre des chaînes de Markov cachées – consistent à simuler les données non observées (c'est-à-dire les états cachés), sachant la valeur courante du paramètre et la séquence observée, puis à maximiser la log-vraisemblance des données complètes. Ils diffèrent par la manière de simuler les états cachés : l'algorithme SEM simule les états selon leur loi jointe alors que l'algorithme EM à la Gibbs simule les états cachés état par état selon la loi conditionnelle d'un état sachant les autres.

L'algorithme SEM consiste à :

· simuler la séquence des états cachés $s_1^{T(k+1)}$ selon la loi jointe conditionnelle

$$P\left(S_1^T = s_1^T | Y_1^T = y_1^T; \theta^{(k)}\right),$$

· choisir $\theta^{(k+1)} = \arg\max_{\theta\in\Theta} \log f\left(s_1^{T(k+1)}, y_1^T; \theta\right)$.

L'algorithme EM à la Gibbs consiste à :

· simuler chaque état caché $s_t^{(k+1)}$ selon la loi

$$P\left(S_t = s_t | S_1^{t-1} = s_1^{t-1(k+1)}, S_{t+1}^T = s_{t+1}^{T(k)}, Y_1^T = y_1^T; \theta^{(k)}\right),$$

· choisir $\theta^{(k+1)} = \arg\max_{\theta\in\Theta} \log f\left(s_1^{T(k+1)}, y_1^T; \theta\right)$.

Ces deux algorithmes peuvent être considérés comme des versions "maximum de vraisemblance" de l'échantillonnage de Gibbs (Geman et Geman, 1984) bien que, dans ce cadre, il n'y ait pas de loi invariante connue ou postulée à l'avance.

Un autre algorithme itératif permet également d'estimer les paramètres d'une chaîne de Markov cachée. Il s'agit de l'algorithme de Baum-Viterbi qui restaure la séquence d'états globalement optimale et estime les paramètres par une procédure de maximisation. L'équivalence asymptotique des estimateurs obtenus avec l'algorithme EM et l'algorithme de Baum-Viterbi a été montrée par Merhav et Ephraim (1991). De plus, Juang et Rabiner (1990) ont prouvé la propriété de croissance monotone de la log-vraisemblance des séquences d'états optimales associées aux séquences observées.

Nous débutons cette section en présentant l'algorithme EM pour l'estimation ML des paramètres d'une chaîne de Markov cachée d'ordre 1 (Baum *et al.*, 1970 ; Ephraim et Merhav, 2002). Par la suite, nous détaillons l'algorithme SEM et l'algorithme de Baum-Viterbi car l'algorithme itératif proposé au chapitre suivant pour l'estimation des paramètres d'un modèle linéaire mixte multiphasique est une alternative à l'algorithme EM dont l'étape de restauration est soit déterministe (l'algorithme est alors de type Baum-Viterbi), soit effectuée par simulation (l'algorithme est alors de type SEM).

Estimation avec l'algorithme EM

Rappelons les hypothèses d'une chaîne de Markov cachée d'ordre 1 $\{S_t, Y_t; t = 1, 2, ...\}$:
· $\{S_t; t = 1, 2, ...\}$ est le processus d'état à valeurs dans l'espace d'états fini $\{1, ..., J\}$,
· $\{Y_t; t = 1, 2, ...\}$ est le processus d'observation (ou d'émission) à valeurs dans l'espace des observations fini $\{1, ..., U\}$.
· θ l'ensemble des paramètres du modèle comprend :
- les probabilités initiales $(\pi_j; j = 1, ..., J)$,
- les probabilités de transition $(p_{ij}; i, j = 1, ..., J)$,
- les probabilités d'observation $(b_j(u); j = 1, ..., J; u = 1, ..., U)$.

Nous restreignons la présentation au cas d'une seule séquence d'observations y_1^T de longueur T.

Écriture de la densité des données complètes
Pour la spécification du problème aux données complètes, nous supposons qu'à la fois la séquence y_1^T et la séquence des états s_1^T sont obervées. La vraisemblance des données complètes s'écrit donc :

$$\begin{aligned} f(s_1^T, y_1^T; \theta) &= P\left(S_1^T = s_1^T, Y_1^T = y_1^T; \theta\right) \\ &= P\left(S_1 = s_1; \theta\right) P\left(Y_1 = y_1 | S_1 = s_1; \theta\right) \\ &\quad \times \prod_{t=2}^{T} P\left(S_t = s_t | S_{t-1} = s_{t-1}; \theta\right) P\left(Y_t = y_t | S_t = s_t; \theta\right) \\ &= \pi_{s_1} b_{s_1}\left(y_1\right) \prod_{t=2}^{T} p_{s_{t-1} s_t} b_{s_t}\left(y_t\right). \end{aligned}$$

La log-vraisemblance des données complètes s'écrit :

$$\begin{aligned} \log f(s_1^T, y_1^T; \theta) &= \sum_j I\left(s_1 = j\right) \log \pi_j + \sum_{i,j} \left\{ \sum_{t=2}^{T} I\left(s_{t-1} = i, s_t = j\right) \right\} \log p_{ij} \\ &\quad + \sum_j \sum_u \left\{ \sum_{t=1}^{T} I\left(s_t = j, y_t = u\right) \right\} \log b_j(u), \end{aligned} \tag{2.18}$$

où $I\left(\right)$ désigne la fonction indicatrice.

Étape E

On calcule l'espérance conditionnelle de la log-vraisemblance des données complètes connaissant la séquence des observations et la valeur des paramètres à l'itération k.

$$
\begin{aligned}
Q(\theta|\theta^{(k)}) &= E\left\{\log f(s_1^T, y_1^T; \theta)|Y_1^T = y_1^T; \theta^{(k)}\right\} \\
&= \sum_j P(S_1 = j|Y_1^T = y_1^T; \theta^{(k)}) \log \pi_j \\
&\quad + \sum_{i,j} \left\{\sum_{t=2}^{T} P\left(S_{t-1} = i, S_t = j|Y_1^T = y_1^T; \theta^{(k)}\right)\right\} \log p_{ij} \\
&\quad + \sum_j \sum_u \left\{\sum_{t=1}^{T} P\left(Y_t = u, S_t = j|Y_1^T = y_1^T; \theta^{(k)}\right)\right\} \log b_j(u). \quad (2.19)
\end{aligned}
$$

L'espérance conditionnelle $Q(\theta|\theta^{(k)})$ s'écrit comme la log-vraisemblance des données complètes (2.18) pour laquelle les variables indicatrices sont remplacées par les probabilités conditionnelles sachant les données observées et la valeur courante des paramètres.

Les quantités $L_j^{(k)}(t) = P\left(S_t = j|Y_1^T = y_1^T; \theta^{(k)}\right)$, $P\left(S_t = j, S_{t-1} = i|Y_1^T = y_1^T; \theta^{(k)}\right)$ et $P\left(Y_t = u, S_t = j|Y_1^T = y_1^T; \theta^{(k)}\right)$ sont calculées de manière récursive par un algorithme "avant-arrière" (Devijver, 1985 ; Ephraim et Merhav, 2002), appelé algorithme "forward-backward", et décrit ci-dessous.

Algorithme "avant-arrière"

Cet algorithme a initialement été décrit par Baum *et al.* (1970). Les récurrences "avant" et "arrière" n'étant pas numériquement stables, Levinson *et al.* (1983) ont proposé d'utiliser des facteurs d'échelle pour renormaliser les quantités "avant", de sorte que leur somme vaille un, puis d'utiliser les mêmes facteurs d'échelle dans la phase "arrière". Ensuite Devijver (1985) a montré que cet algorithme pouvait être justifié par la méthodologie des modèles à espace d'états.

L'algorithme "avant-arrière" se décompose en deux calculs récursifs, le premier de 1 à T (passe "avant") et le second de T à 1 (passe "arrière"). Il repose sur la décomposition suivante :

$$
\begin{aligned}
L_j^{(k)}(t) &= P\left(S_t = j|Y_1^T = y_1^T; \theta^{(k)}\right) \\
&= \frac{P\left(Y_{t+1}^T = y_{t+1}^T|S_t = j; \theta^{(k)}\right)}{P\left(Y_{t+1}^T = y_{t+1}^T|Y_1^t = y_1^t; \theta^{(k)}\right)} P\left(S_t = j|Y_1^t = y_1^t; \theta^{(k)}\right) \\
&= B_j^{(k)}(t) F_j^{(k)}(t), \quad (2.20)
\end{aligned}
$$

avec

$$B_j^{(k)}(t) = \frac{P\left(Y_{t+1}^T = y_{t+1}^T | S_t = j; \theta^{(k)}\right)}{P\left(Y_{t+1}^T = y_{t+1}^T | Y_1^t = y_1^t; \theta^{(k)}\right)},$$

et

$$F_j^{(k)}(t) = P\left(S_t = j | Y_1^t = y_1^t; \theta^{(k)}\right).$$

Les quantités $F_j^{(k)}(t)$ (appelées probabilités filtrées) sont calculées dans la passe "avant" alors que les quantités $B_j^{(k)}(t)$ ou les quantités $L_j^{(k)}(t)$ (appelées probabilités lissées) sont calculées dans la passe "arrière".

Remarque : Pour alléger l'écriture des diverses expressions, dans la présentation de l'algorithme, nous ne préciserons pas que toutes les probabilités sont calculées sachant $\theta^{(k)}$, la valeur des paramètres à l'étape k. Nous omettrons également l'exposant (k) dans les écritures de $L_j(t)$, $F_j(t)$ et $B_j(t)$.

Récurrence "avant"

Elle est initialisée pour $t = 1$ et $j = 1, ..., J$ par

$$\begin{aligned} F_j(1) &= P(S_1 = j | Y_1 = y_1) \\ &= \frac{P(Y_1 = y_1 | S_1 = j)}{P(Y_1 = y_1)} P(S_1 = j) \\ &= \frac{b_j(y_1)}{N_1} \pi_j, \end{aligned} \tag{2.21}$$

où $N_1 = P(Y_1 = y_1)$, facteur de normalisation, est égal à :

$$\begin{aligned} N_1 &= \sum_{j=1}^{J} P(Y_1 = y_1 | S_1 = j) P(S_1 = j) \\ &= \sum_{j=1}^{J} b_j(y_1) \pi_j. \end{aligned}$$

Pour $t = 2, ..., T$ et $j = 1, ..., J$, la récurrence "avant" s'écrit :

$$\begin{aligned} F_j(t) &= P\left(S_t = j | Y_1^t = y_1^t\right) \\ &= \frac{\sum_{i=1}^{J} P\left(S_t = j, S_{t-1} = i, Y_t = y_t | Y_1^{t-1} = y_1^{t-1}\right)}{P\left(Y_t = y_t | Y_1^{t-1} = y_1^{t-1}\right)} \\ &= \frac{P\left(Y_t = y_t | S_t = j\right)}{P\left(Y_t = y_t | Y_1^{t-1} = y_1^{t-1}\right)} \sum_{i=1}^{J} P\left(S_t = j | S_{t-1} = i\right) P\left(S_{t-1} = i | Y_1^{t-1} = y_1^{t-1}\right) \\ &= \frac{b_j\left(y_t\right)}{N_t} \sum_{i=1}^{J} p_{ij} F_i(t-1), \end{aligned} \tag{2.22}$$

où le facteur de normalisation $N_t = P\left(Y_t = y_t | Y_1^{t-1} = y_1^{t-1}\right)$, est égal à :

$$\begin{aligned} N_t &= \sum_{j=1}^{J} P\left(S_t = j, Y_t = y_t | Y_1^{t-1} = y_1^{t-1}\right) \\ &= \sum_{j=1}^{J} b_j\left(y_t\right) \sum_{i=1}^{J} p_{ij} F_i(t-1). \end{aligned}$$

Ainsi pour implémenter cette récurrence "avant", il faut d'abord calculer les quantités $P\left(S_t = j, Y_t = y_t | Y_1^{t-1} = y_1^{t-1}\right) = b_j\left(y_t\right) \sum_{i=1}^{J} p_{ij} F_i(t-1)$; ensuite nous déduisons par sommation le facteur de normalisation N_t, et enfin nous obtenons les quantités "avant" en effectuant la normalisation $F_j(t) = P\left(S_t = j, Y_t = y_t | Y_1^{t-1} = y_1^{t-1}\right) / N_t$.

Récurrence "arrière"

Elle consiste à calculer soit $B_j(t) = P\left(Y_{t+1}^T = y_{t+1}^T | S_t = j\right) / P\left(Y_{t+1}^T = y_{t+1}^T | Y_1^t = y_1^t\right)$ soit $L_j(t) = P\left(S_t = j | Y_1^T = y_1^T\right)$ pour chaque état j en reculant de T à 1.

La récurrence "arrière" est initialisée pour $t = T$ et $j = 1, ..., J$ par

$$L_j(T) = P\left(S_T = j | Y_1^T = y_1^T\right) = F_j(T)$$

Par conséquent, $B_j(T) = 1$.

Pour $t = T-1, ..., 1$ et $j = 1, ..., J$, nous avons la récurrence suivante :

$$
\begin{aligned}
B_j(t) &= \frac{P\left(Y_{t+1}^T = y_{t+1}^T | S_t = j\right)}{P\left(Y_{t+1}^T = y_{t+1}^T | Y_1^t = y_1^t\right)} \\
&= \frac{\sum_{k=1}^J P\left(Y_{t+2}^T = y_{t+2}^T, Y_{t+1} = y_{t+1}, S_{t+1} = k | S_t = j\right)}{P\left(Y_{t+2}^T = y_{t+2}^T, Y_{t+1} = y_{t+1} | Y_1^t = y_1^t\right)} \\
&= \frac{\sum_{k=1}^J P\left(Y_{t+2}^T = y_{t+2}^T | S_{t+1} = k\right) P\left(Y_{t+1} = y_{t+1} | S_{t+1} = k\right) P(S_{t+1} = k | S_t = j)}{P\left(Y_{t+2}^T = y_{t+2}^T | Y_1^{t+1} = y_1^{t+1}\right) P\left(Y_{t+1} = y_{t+1} | Y_1^t = y_1^t\right)} \\
&= \frac{1}{N_{t+1}} \sum_{k=1}^J B_k(t+1) b_k\left(y_{t+1}\right) p_{jk}
\end{aligned}
$$

De (2.20), nous déduisons :

$$
\begin{aligned}
L_j(t) &= \frac{1}{N_{t+1}} \left\{ \sum_{k=1}^J \frac{L_k(t+1)}{F_k(t+1)} b_k\left(y_{t+1}\right) p_{jk} \right\} F_j(t) \\
&= \left\{ \sum_{k=1}^J \frac{L_k(t+1)}{G_k(t+1)} p_{jk} \right\} F_j(t), \qquad (2.23)
\end{aligned}
$$

avec

$$
\begin{aligned}
G_k(t+1) &= \frac{F_k(t+1) N_{t+1}}{b_k\left(y_{t+1}\right)} \\
&= P\left(S_{t+1} = k | Y_1^t = y_1^t\right).
\end{aligned}
$$

De plus, d'après (2.22) :

$$
G_k(t+1) = \sum_{j=1}^J p_{jk} F_j(t).
$$

Cette quantité est appelée probabilité prédite et peut être extraite et stockée en mémoire lors de la récurrence "avant".

Étape M

Cette étape consiste à maximiser l'espérance conditionnelle de la log-vraisemblance des données complètes qui vient d'être calculée à l'étape E. Nous cherchons la valeur des paramètres qui maximise cette quantité.

Estimation du paramètre π_j

Nous cherchons à maximiser $\sum_{j=1}^J L_j^{(k)}(1) \log \pi_j$ sous la contrainte $\sum_{j=1}^J \pi_j = 1$. Nous notons $Q_\pi = \sum_{j=1}^J L_j^{(k)}(1) \log \pi_j$.

La condition

$$\left(\frac{\partial Q_\pi}{\partial \pi_1}, \frac{\partial Q_\pi}{\partial \pi_2}, \ldots, \frac{\partial Q_\pi}{\partial \pi_J}\right) = (0, 0, \ldots, 0)$$

s'écrit

$$\frac{L_1^{(k)}(1)}{\pi_1} = \frac{L_2^{(k)}(1)}{\pi_2} = \ldots = \frac{L_J^{(k)}(1)}{\pi_J}.$$

Nous en déduisons :

$$\frac{L_j^{(k)}(1)}{\pi_j} = \frac{\sum_{i=1}^J L_i^{(k)}(1)}{\sum_{i=1}^J \pi_i} = 1 \qquad j = 1, ..., J,$$

car

$$\sum_{i=1}^J L_i^{(k)}(1) = \sum_{i=1}^J P\left(S_1 = i | Y_1^T = y_1^T; \theta^{(k)}\right) = 1.$$

De plus

$$\frac{\partial^2 Q_\pi}{\partial \pi_j^2} = -\frac{L_j^{(k)}(1)}{\pi_j^2} < 0$$

Donc

$$\pi_j^{(k+1)} = L_j^{(k)}(1) \qquad j = 1, ..., J.$$

Estimation du paramètre p_{ij}

Nous cherchons à maximiser $\sum_{j=1}^J \sum_{t=2}^T P\left(S_t = j, S_{t-1} = i | Y_1^T = y_1^T; \theta^{(k)}\right) \log p_{ij}$ sous la contrainte $\sum_{j=1}^J p_{ij} = 1$. Selon le même principe de maximisation que pour le paramètre π_j, nous obtenons :

$$p_{ij}^{(k+1)} = \frac{\sum_{t=1}^{T-1} P\left(S_{t+1} = j, S_t = i | Y_1^T = y_1^T; \theta^{(k)}\right)}{\sum_{t=1}^{T-1} P\left(S_t = i | Y_1^T = y_1^T; \theta^{(k)}\right)} \qquad i, j = 1, ..., J.$$

De plus $L_i^{(k)}(t) = P\left(S_t = i | Y_1^T = y_1^T; \theta^{(k)}\right) = \sum_{j=1}^J P\left(S_{t+1} = j, S_t = i | Y_1^T = y_1^T; \theta^{(k)}\right)$, et par identification avec (2.23), nous déduisons :

$$P\left(S_{t+1} = j, S_t = i | Y_1^T = y_1^T; \theta^{(k)}\right) = \frac{L_j^{(k)}(t+1)}{G_j^{(k)}(t+1)} p_{ij}^{(k)} F_i^{(k)}(t).$$

Par conséquent

$$p_{ij}^{(k+1)} = \frac{\sum_{t=1}^{T-1} L_j^{(k)}(t+1) p_{ij}^{(k)} F_i^{(k)}(t) / G_j^{(k)}(t+1)}{\sum_{t=1}^{T-1} L_i^{(k)}(t)} \qquad i, j = 1, ..., J.$$

Estimation du paramètre $b_j(u)$

Nous cherchons à maximiser $\sum_{u=1}^{U}\sum_{t=1}^{T} P\left(Y_t = u, S_t = j | Y_1^T = y_1^T; \theta^{(k)}\right)$ sous la contrainte $\sum_{u=1}^{U} b_j(u) = 1$. Selon le même principe de maximisation que pour les paramètres π_j et p_{ij}, nous obtenons :

$$\begin{aligned} b_j^{(k+1)}(u) &= \frac{\sum_{t=1}^{T} P\left(Y_t = u, S_t = j | Y_1^T = y_1^T; \theta^{(k)}\right)}{\sum_{v=1}^{U}\sum_{t=1}^{T} P\left(Y_t = v, S_t = j | Y_1^T = y_1^T; \theta^{(k)}\right)} \\ &= \frac{\sum_{t=1}^{T} P\left(Y_t = u, S_t = j | Y_1^T = y_1^T; \theta^{(k)}\right)}{\sum_{t=1}^{T} P\left(S_t = j | Y_1^T = y_1^T; \theta^{(k)}\right)} \\ &= \frac{\sum_{t=1}^{T} L_j^{(k)}(t)\, I\left(y_t = u\right)}{\sum_{t=1}^{T} L_j^{(k)}(t)} \qquad j = 1, ..., J; u = 1, ..., U \end{aligned}$$

Estimation avec l'algorithme SEM

L'algorithme SEM est une version stochastique de l'algorithme EM introduite par Celeux et Diebolt (1985) pour des modèles de mélange de lois exponentielles, visant à pallier les inconvénients de EM, comme la dépendance du point de départ et la convergence vers des points stationnaires qui ne sont pas des maxima locaux de la log-vraisemblance. L'algorithme SEM consiste à introduire une perturbation aléatoire à chaque itération de l'algorithme EM pour tenter d'éviter une mauvaise convergence (points selles, maxima locaux "peu intéressants",...). En revanche, la propriété d'accroissement monotone de la fonction de vraisemblance de l'EM n'est plus vraie pour SEM.

Cet algorithme est une procédure itérative alternant deux étapes. Si $\theta^{(k)}$ est la valeur courante du paramètre, l'itération k de l'algorithme SEM est décrite par les deux étapes suivantes :

Étape S : on simule la séquence des états cachés $s_1^{T(k)} = \left(s_1^{(k)}, ..., s_T^{(k)}\right)$ selon la loi jointe conditionnelle $P\left(S_1^T = s_1^T | Y_1^T = y_1^T; \theta^{(k)}\right)$.

Étape M : on maximise par rapport à θ la log-vraisemblance des données complètes pour réactualiser la valeur du paramètre. Autrement dit, on choisit $\theta^{(k+1)}$ de sorte que

$$\theta^{(k+1)} = \arg\max_{\theta\in\Theta} \log f\left(s_1^{T(k)}, y_1^T; \theta\right).$$

Explicitons de manière plus précise ces deux étapes :

Étape S

L'étape S de l'algorithme SEM consiste en premier lieu à calculer la loi jointe conditionnelle des états cachés sachant la séquence observée y_1^T et la valeur courante du paramètre

$\theta^{(k)}$. Cette loi se calcule, dans notre cas, à nouveau par une récurrence "avant-arrière" analogue à celle décrite par Baum *et al.* (1970) pour l'algorithme EM.

La loi jointe des états cachés $P\left(S_1^T = s_1^T | Y_1^T = y_1^T\right)$ se déduit de la décomposition des lois conditionnelles :

$$\begin{aligned} P\left(S_1^T = s_1^T | Y_1^T = y_1^T\right) &= P\left(S_1 = s_1 | Y_1^T = y_1^T\right) P\left(S_2 = s_2 | S_1 = s_1, Y_1^T = y_1^T\right) \\ &\quad \times \ldots P\left(S_t = s_t | S_1^{t-1} = s_1^{t-1}, Y_1^T = y_1^T\right) \\ &\quad \times \ldots P\left(S_T = s_T | S_1^{T-1} = s_1^{T-1}, Y_1^T = y_1^T\right), \end{aligned} \tag{2.24}$$

ou de manière analogue

$$\begin{aligned} P\left(S_1^T = s_1^T | Y_1^T = y_1^T\right) &= P\left(S_1 = s_1 | S_2^T = s_2^T, Y_1^T = y_1^T\right) \\ &\quad \times P\left(S_2 = s_2 | S_3^T = s_3^T, Y_1^T = y_1^T\right) \\ &\quad \times \ldots P\left(S_t = s_t | S_{t+1}^T = s_{t+1}^T, Y_1^T = y_1^T\right) \\ &\quad \times \ldots P\left(S_T = s_T | Y_1^T = y_1^T\right). \end{aligned} \tag{2.25}$$

La loi jointe sera ainsi complètement déterminée si l'on connaît, pour chaque instant t, les probabilités conditionnelles $P\left(S_t = s_t | S_1^{t-1} = s_1^{t-1}, Y_1^T = y_1^T\right)$ ou les probabilités conditionnelles $P\left(S_t = s_t | S_{t+1}^T = s_{t+1}^T, Y_1^T = y_1^T\right)$.

Qian et Titterington (1990) utilisent la décomposition (2.24) et fournissent une formule de récurrence "avant-arrière" permettant de calculer les probabilités conditionnelles de chaque état s_t sachant les états précédents s_1^{t-1}, la séquence y_1^T et la valeur du paramètre θ.

Chib (1996) présente un algorithme "avant-arrière" pour calculer les probabilités $P\left(S_t = s_t | S_{t+1}^T = s_{t+1}^T, Y_1^T = y_1^T\right)$ à l'aide des probabilités prédites $P\left(S_t = j | Y_1^{t-1} = y_1^{t-1}\right)$, et des probabilités filtrées $P\left(S_t = j | Y_1^t = y_1^t\right)$ calculées au cours de la récurrence "avant" décrite à l'étape E de l'algorithme EM. Nous utiliserons donc la décomposition (2.25) de la loi jointe.

Les probabilités conditionnelles de chaque état s_t sachant les états suivants s_{t+1}^T, la séquence y_1^T et la valeur du paramètre θ, ne dépendent que de l'état suivant s_{t+1}, des t premières observations y_1^t, et du paramètre θ. En effet,

$$
\begin{aligned}
& P\left(S_t = j | S_{t+1}^T = s_{t+1}^T, Y_1^T = y_1^T; \theta\right) \\
= & \frac{P\left(S_t = j, S_{t+1}^T = s_{t+1}^T, Y_1^T = y_1^T; \theta\right)}{P\left(S_{t+1}^T = s_{t+1}^T, Y_1^T = y_1^T, \theta\right)} \\
= & \frac{P\left(S_{t+2}^T = s_{t+2}^T, Y_{t+1}^T = y_{t+1}^T | S_{t+1} = s_{t+1}; \theta\right) P\left(S_{t+1} = s_{t+1} | S_t = j; \theta\right) P\left(S_t = j, Y_i^t = y_1^t; \theta\right)}{P\left(S_{t+2}^T = s_{t+2}^T, Y_{t+1}^T = y_{t+1}^T | S_{t+1} = s_{t+1}; \theta\right) P\left(S_{t+1} = s_{t+1}, Y_1^t = y_1^t; \theta\right)} \\
= & \frac{P\left(S_{t+1} = s_{t+1} | S_t = j; \theta\right) P\left(S_t = j | Y_1^t = y_1^t; \theta\right)}{P\left(S_{t+1} = s_{t+1} | Y_1^t = y_1^t; \theta\right)} \\
= & \frac{p_{j s_{t+1}} F_j(t)}{\sum_i p_{i s_{t+1}} F_i(t)},
\end{aligned}
$$

où $F_j(t)$ est la probabilité filtrée et $P\left(S_{t+1} = s_{t+1} | Y_1^t = y_1^t; \theta\right) = \sum_i p_{i s_{t+1}} F_i(t)$ est la probabilité prédite, toutes deux issues de la récurrence "avant" de l'algorithme "avant-arrière" présenté à l'étape E de l'EM.

L'itération k de l'étape S de l'algorithme SEM est décrite par l'algorithme "avant-arrière" suivant (il s'agit d'une récurrence "arrière" qui se calcule à l'aide de la récurrence "avant" présentée dans le cadre EM) :

Pour $t = T$ et $j = 1, ..., J$, on calcule grâce à l'algorithme "avant" décrit dans le cadre EM

$$P\left(S_T = j | Y_1^T = y_1^T; \theta^{(k)}\right) = L_j^{(k)}(T) = F_j^{(k)}(T).$$

On génère ensuite un nombre aléatoire entre 0 et 1 et on regarde à quel état il correspond sur la fonction de répartition de la loi de probabilité $\left\{P\left(S_T = j | Y_1^T = y_1^T; \theta^{(k)}\right); j = 1, ...J\right\}$.

Pour $t = T - 1, ..., 1$ et $j = 1, ..., J$, on calcule

$$P\left(S_t = j | S_{t+1}^T = s_{t+1}^{T^{(k)}}, Y_1^T = y_1^T; \theta^{(k)}\right) = \frac{P\left(S_{t+1} = s_{t+1}^{(k)} | S_t = j\right) F_j^{(k)}(t)}{\sum_i P\left(S_{t+1} = s_{t+1}^{(k)} | S_t = i\right) F_i^{(k)}(t)}.$$

On génère ensuite un nombre aléatoire entre 0 et 1 et on regarde à quel état il correspond sur la fonction de répartition de la loi de probabilité

$$\left\{P\left(S_t = j | S_{t+1}^T = s_{t+1}^{T^{(k)}}, Y_1^T = y_1^T; \theta^{(k)}\right); j = 1, ...J\right\}.$$

Étape M

Sachant la suite des états cachés $s_1^{T(k)}$, l'étape M consiste à maximiser en θ la log-vraisemblance des données complètes donnée par

$$
\begin{aligned}
& \log f(s_1^{T(k)}, y_1^T; \theta) \\
= & \sum_j I\left(s_1^{(k)} = j\right) \log \pi_j + \sum_{i,j} \left\{ \sum_{t=2}^{T} I\left(s_{t-1}^{(k)} = i, s_t^{(k)} = j\right) \right\} \log p_{ij} \\
& + \sum_j \sum_u \left\{ \sum_{t=1}^{T} I\left(s_t^{(k)} = j, y_t = u\right) \right\} \log b_j(u).
\end{aligned}
$$

Selon les mêmes principes de maximisation que pour l'étape M de l'algorithme EM, nous en déduisons les formules de réestimation suivantes :

$$\pi_j^{(k+1)} = I\left(s_1^{(k)} = j\right) \qquad j = 1, ..., J,$$

$$p_{ij}^{(k+1)} = \frac{\sum_{t=1}^{T-1} I\left(s_t^{(k)} = i, s_{t+1}^{(k)} = j\right)}{\sum_{t=1}^{T-1} I\left(s_t^{(k)} = i\right)} \qquad i, j = 1, ..., J,$$

$$b_j^{(k+1)}(u) = \frac{\sum_{t=1}^{T} I\left(s_t^{(k)} = j, y_t = u\right)}{\sum_{t=1}^{T} I\left(s_t^{(k)} = j\right)} \qquad j = 1, ..., J; u = 1, ..., U.$$

Les formules de réestimation, obtenues à l'étape M de l'algorithme EM, sont similaires à celles obtenues à l'étape M de l'algorithme SEM, à la différence près que les probabilités conditionnelles des variables aléatoires S_t et Y_t sachant y_1^T et $\theta^{(k)}$ pour l'algorithme EM sont remplacées par les indicatrices des valeurs des états simulés à l'itération k pour l'algorithme SEM. L'étape M de SEM consiste donc à faire de simples comptages : nombre d'états simulés prenant une valeur particulière, nombre de transitions entre états simulés et nombre de fois où une observation est émise depuis un état simulé prenant une valeur particulière.

Estimation avec l'algorithme de Baum-Viterbi

Souvent utilisé pour la reconnaissance de parole par chaînes de Markov cachées, l'algorithme de Baum-Viterbi a été introduit pour la première fois par Jelinek (1976). Il est, comme les algorithmes EM et SEM, un algorithme itératif de type restauration-maximisation. Il restaure la séquence d'états globalement optimale et estime les paramètres par une procédure de maximisation. Il porte ce nom car chaque itération fait intervenir l'algorithme de Viterbi (Forney, 1973) ainsi qu'une étape de réestimation de type Baum (Baum *et al.*, 1970). Dans une première étape, la séquence d'états globalement optimale est restaurée par l'algorithme de Viterbi, et dans une seconde étape, les paramètres sont estimés par maximisation de la probabilité jointe de la séquence observée et de la séquence d'états globalement optimale qui a été restaurée.

Le principe de l'algorithme de Baum-Viterbi pour l'estimation du paramètre θ repose sur une double maximisation

$$\max_{\theta} \max_{s_1^T} P\left(S_1^T = s_1^T, Y_1^T = y_1^T; \theta\right),$$

alternant une maximisation par rapport à la séquence d'états s_1^T, et une maximisation par rapport au paramètre θ.

Si $\theta^{(k)}$ est la valeur courante du paramètre, l'itération k de l'algorithme de Baum-Viterbi est composée d'une étape de restauration déterministe et d'une étape de maximisation de type Baum.

Étape 1 : restauration déterministe

La séquence d'états globalement optimale, notée $s_1^{T(k)}$ est estimée en maximisant $P\left(S_1^T = s_1^T, Y_1^T = y_1^T; \theta^{(k)}\right)$ par rapport à s_1^T avec l'algorithme de Viterbi détaillé ci-dessous.

L'algorithme de Viterbi

Le premier algorithme de restauration globale dans le cas de processus à structure cachée est dû à Viterbi (1967). La justification de cet algorithme a ensuite été donnée par Forney (1973) et s'appuie sur la théorie des graphes. L'algorithme de Viterbi permet de déterminer la séquence d'états globalement optimale, c'est-à-dire celle qui explique au mieux la séquence observée pour un modèle donné. Cet algorithme de programmation dynamique repose sur une propriété de décomposabilité de la fonction à optimiser.
Comme le processus d'état $\{S_t\}$ est une chaîne de Markov, nous avons la décomposition suivante :

$$\begin{aligned}\max_{s_1,\ldots,s_T} P\left(S_1^T = s_1^T, Y_1^T = y_1^T\right) &= \max_{s_t}\left\{\max_{s_{t+1},\ldots,s_T} P\left(Y_{t+1}^T = y_{t+1}^T, S_{t+1}^T = s_{t+1}^T | S_t = s_t\right)\right. \\ &\left.\times \max_{s_1,\ldots,s_{t-1}} P\left(S_1^t = s_1^t, Y_1^t = y_1^t\right)\right\}. \qquad (2.26)\end{aligned}$$

Si nous notons

$$\alpha_j(t) = \max_{s_1,\ldots,s_{t-1}} P\left(S_t = j, S_1^{t-1} = s_1^{t-1}, Y_1^t = y_1^t\right),$$

alors, (2.26) se réécrit :

$$\max_{s_1,\ldots,s_T} P\left(S_1^T = s_1^T, Y_1^T = y_1^T\right) = \max_{j}\left\{\max_{s_{t+1},\ldots,s_T} P\left(Y_{t+1}^T = y_{t+1}^T, S_{t+1}^T = s_{t+1}^T | S_t = j\right)\alpha_j(t)\right\}.$$

L'algorithme est initialisé pour $t = 1$ et $j = 1, ..., J$ par

$$\begin{aligned}\alpha_j(1) &= P\left(Y_1 = y_1 | S_1 = j\right) P\left(S_1 = j\right) \\ &= b_j\left(y_1\right) \pi_j. \end{aligned} \tag{2.27}$$

L'équation de programmation dynamique s'écrit pour $t = 2, ..., T$ et $j = 1, ..., J$:

$$\begin{aligned}\alpha_j(t) &= \max_{s_1,\ldots,s_{t-1}} P\left(S_t = j, S_1^{t-1} = s_1^{t-1}, Y_1^t = y_1^t\right) \\ &= P\left(Y_t = y_t | S_t = j\right) \max_i \left\{P\left(S_t = j | S_{t-1} = i\right)\right. \\ &\quad \left. \times \max_{s_1,\ldots,s_{t-2}} P\left(S_{t-1} = i, S_1^{t-2} = s_1^{t-2}, Y_1^{t-1} = y_1^{t-1}\right)\right\} \\ &= b_j\left(y_t\right) \max_i \left\{p_{ij} \alpha_i(t-1)\right\}. \end{aligned} \tag{2.28}$$

Les sous-séquences d'états optimales s_1^t se terminant dans un état donné se déduisent ainsi des J sous-séquences d'états optimales s_1^{t-1} se terminant dans les différents états calculés à l'étape précédente. L'état précédent optimal est donné par

$$\psi_j(t) = \arg\max_i \left\{p_{ij} \alpha_i(t-1)\right\}.$$

La vraisemblance de la séquence d'états optimale associée à la séquence obervée y_1^T est égale à

$$\max_j \left\{\alpha_j(T)\right\},$$

et l'état final optimal est donné par

$$\widetilde{s}_T = \arg\max_j \left\{\alpha_j(T)\right\}.$$

La séquence d'états optimale est alors extraite par une procédure de chaînage arrière ("backtracking"). Pour $t = T - 1, ..., 1$

$$\widetilde{s}_t = \psi_{\widetilde{s}_{t+1}}\left(t+1\right).$$

Remarquons que l'algorithme de Viterbi est l'équivalent de l'algorithme "avant" (décrit dans le cadre EM) en termes de programmation dynamique (en oubliant l'étape de normalisation dans l'algorithme "avant"). En effet, les équations (2.27) et (2.28) sont similaires aux équations (2.21) et (2.22), si ce n'est que les sommations ont été remplacées par des maximisations.

Étape 2 : maximisation de type Baum
La nouvelle valeur du paramètre $\theta^{(k+1)}$ est ensuite obtenue en maximisant $P\left(S_1^T = s_1^{T(k)}, Y_1^T = y_1^T; \theta\right)$ par rapport à θ :

$$\theta^{(k+1)} = \arg\max_{\theta\in\Theta} P\left(S_1^T = s_1^{T(k)}, Y_1^T = y_1^T; \theta\right).$$

L'étape de maximisation de l'algorithme de Baum-Viterbi est similaire à l'étape de maximisation de l'algorithme SEM, à la différence près que la séquence $s_1^{T(k)}$ est une séquence d'états simulée pour l'algorithme SEM alors que $s_1^{T(k)}$ est la séquence d'états globalement optimale pour l'algorithme de Baum-Viterbi.

Principe et propriété

Le principe de l'algorithme de Baum-Viterbi consiste à itérer les étapes 1 et 2 jusqu'à ce qu'un point fixe soit atteint ou qu'un critère d'arrêt soit satisfait. La propriété caractéristique de cet algorithme est la propriété de croissance monotone de la log-vraisemblance des séquences d'états optimales associées aux séquences observées. Cette propriété a été démontrée par Juang et Rabiner (1990). La preuve est principalement basée sur le théorème de convergence globale de Zangwill (1969). Ce théorème de convergence globale est un résultat général applicable au cas où l'espace d'états caché est un espace de variables indépendantes, ce qui est notre cas. Avant d'énoncer ce théorème et de démontrer la propriété, nous donnons quelques définitions et notations.

Soit Θ un sous espace ouvert de l'espace euclidien $\mathbb{R}^p$. Un modèle de Markov caché θ est un point de Θ et pour tout $\theta \in \Theta$, nous avons la relation $\theta \longrightarrow (\pi(\theta), P(\theta), B(\theta))$ où π, P et B désignent respectivement le vecteur des probabilités initiales, la matrice des probabilités de transition et la matrice des probabilités d'observation. De plus, on suppose que Θ est un compact et que $f\left(s_1^T, y_1^T; \theta\right)$ est continue sur Θ et différentiable sur son intérieur de sorte que $f\left(s_1^T, y_1^T; \theta\right)$ est bornée supérieurement.

Un algorithme T sur Θ est un tracé de points de Θ dans des sous ensembles de Θ. Quand le tracé est point par point, T est simplement une transformation. On dit que l'algorithme T sur Θ est fermé si $\theta \in \Theta$, $\rho \in \Theta$, $\theta_n \longrightarrow \theta$, $\rho_n \longrightarrow \rho$ et $\rho_n \in T(\theta_n)$ impliquent que $\rho \in T(\theta)$. La fermeture est une généralisation de la continuité. Pour un tracé point par point, la continuité implique la fermeture.

Soit Ω l'ensemble des points fixes de T. Une fonction g sur Θ est appelée fonction ascendante pour l'algorithme T si les trois conditions suivantes sont vérifiées :

1) $g : \Theta \longrightarrow \mathbb{R}' \subset \mathbb{R}$ est continue,

2) $g(\rho) > g(\theta)$ pour $\rho \in T(\theta)$ et $\theta \notin \Omega$,

3) $g(\rho) \geqslant g(\theta)$ pour $\rho \in T(\theta)$ et $\theta \in \Omega$.

Théorème de convergence globale (Zangwill, 1969) :

Soit $\{\theta_i\}_{i=0}^{\infty}$ la séquence générée par un algorithme T telle que $\theta_{i+1} \in T(\theta_i)$, avec $\theta_0 \in \Theta$. On suppose que T est fermé et que $\Omega \subset \Theta$ est l'ensemble des points fixes de T. Alors :

i) Ω est fermé,

ii) tous les points d'accumulation de $\{\theta_i\}$ sont dans Ω et $g(\theta_i)$ converge de manière monotone vers $g(\theta^*)$ avec $\theta^* \in \Omega$, si g est une fonction ascendante.

Démonstration de la propriété de convergence monotone de l'algorithme de Baum-Viterbi

Nous notons

$$\max_{s_1^T} f\left(s_1^T, y_1^T; \theta\right) = \max_{s_1^T} \left\{ \pi_{s_1} \prod_{t=2}^{T} p_{s_{t-1}s_t} b_{s_t}\left(y_t\right) \right\},$$

et

$$\overline{\theta} = \arg\max_{\theta} \left\{ \max_{s_1^T} f\left(s_1^T, y_1^T; \theta\right) \right\}.$$

Pour montrer que l'algorithme de Baum-Viterbi converge de manière monotone au sens de la (log-) vraisemblance de la séquence d'états optimale associée à la séquence observée, nous appliquons le théorème de convergence globale de Zangwill. Il suffit de prouver que l'algorithme $T : \theta \longrightarrow \overline{\theta}$ est fermé et que la fonction $g\left(\theta\right) = \max_{s_1^T} f\left(s_1^T, y_1^T; \theta\right)$ est une fonction ascendante pour l'algorithme.

L'algorithme T est fermé car nous supposons que

$$f\left(y_1^T; \theta\right) = \sum_{s_1^T} f\left(s_1^T, y_1^T; \theta\right),$$

et par conséquent $\max_{s_1^T} f\left(s_1^T, y_1^T; \theta\right)$ est continûment différentiable en θ pour presque tous les y_1^T dans un espace mesurable totalement fini.

Reste à prouver que la fonction g est une fonction ascendante pour l'algorithme T. Comme $\overline{\theta} \in T\left(\theta\right)$, il suffit de montrer que $g\left(\overline{\theta}\right) \geqslant g\left(\theta\right)$.
Soient $\tilde{s}_1^T$ et $\overline{s}_1^T$ les deux séquences d'états optimales telles que

$$\tilde{s}_1^T = \arg\max_{s_1^T} f\left(s_1^T, y_1^T; \theta\right),$$

$$\overline{s}_1^T = \arg\max_{s_1^T} f\left(s_1^T, y_1^T; \overline{\theta}\right).$$

On a

$$g\left(\overline{\theta}\right) = \max_{s_1^T} f\left(s_1^T, y_1^T; \overline{\theta}\right) \geqslant f\left(\tilde{s}_1^T, y_1^T; \overline{\theta}\right). \tag{2.29}$$

De plus, nous pouvons écrire

$$\begin{aligned} f\left(\tilde{s}_1^T, y_1^T; \overline{\theta}\right) &= \max_{\theta'} f\left(\tilde{s}_1^T, y_1^T; \theta'\right) \\ &= \max_{\theta'} \left\{ \max_{s_1^T} f\left(s_1^T, y_1^T; \theta'\right) \right\} \qquad (2.30) \\ &\geqslant \max_{s_1^T} f\left(s_1^T, y_1^T; \theta\right) = g\left(\theta\right). \qquad (2.31) \end{aligned}$$

L'inégalité (2.29) est stricte sauf si $\overline{s}_1^T = \tilde{s}_1^T$, lorsque $\overline{\theta} \in T\left(\overline{\theta}\right)$. L'inégalité(2.31) est stricte à moins que θ soit le maximum de (2.30) ou que $\theta \in T\left(\theta\right)$.

Par conséquent le théorème de convergence globale de Zangwill est vérifié : la fonction $g(\theta) = \max_{s_1^T} f\left(s_1^T, y_1^T; \theta\right)$ converge de manière monotone vers $g(\theta^*)$ où θ^* est un point fixe de T.

Pour conclure ce paragraphe sur l'algorithme de Baum-Viterbi, notons que cet algorithme est également utilisé pour l'identification de mélanges indépendants ; il est alors connu sous le nom d'algorithme CEM (Classification EM) (Celeux et Govaert, 1992). L'algorithme CEM converge en un nombre fini d'itérations et toujours rapidement. Toutefois, il a le désavantage de fournir des estimations biaisées, même avec des échantillons de grande taille, en particulier si les composantes du mélange sont peu séparées ou si les probabilités marginales des états cachés sont assez différentes, dans le cas de modèles stationnaires (Celeux et Clairambault, 1991).

Commentaires

Les trois algorithmes présentés pour l'estimation par maximum de vraisemblance des paramètres d'une chaîne de Markov cachée sont des algorithmes itératifs de type restauration-maximisation. Ils diffèrent principalement par la nature de l'étape de restauration : restauration probabiliste, restauration par simulation et restauration déterministe. En effet, au cours de l'étape E de l'algorithme EM, un algorithme "avant-arrière" permet une restauration probabiliste de toutes les séquences d'états possibles. A contrario, l'étape S de l'algorithme SEM simule une séquence d'états, et la première étape de l'algorithme de Baum-Viterbi restaure de manière déterministe la séquence d'états globalement optimale. L'étape M de l'algorithme EM maximise, en chacun des paramètres à estimer, l'espérance conditionnelle de la log-vraisemblance des données complètes connaissant le vecteur des observations et la valeur des paramètres à l'étape courante. L'étape M des algorithmes SEM et Baum-Viterbi consiste à maximiser, en chacun des paramètres à estimer, la log-vraisemblance des données complètes, les données manquantes étant remplacées par la séquence d'états simulée pour l'algorithme SEM et par la séquence d'états globalement optimale pour l'algorithme de Baum-Viterbi.

Contrairement à l'algorithme SEM, l'algorithme EM et l'algorithme de Baum-Viterbi convergent de façon monotone : la fonction de log-vraisemblance des données observées croît de manière monotone pour l'algorithme EM, et la fonction de log-vraisemblance des séquences d'états optimales associées aux séquences observées croît également de manière monotone pour l'algorithme de Baum-Viterbi.

À noter que l'algorithme "avant-arrière" décrit dans le cadre EM peut être utilisé pour une autre fonction que sa fonction première décrite précédemment. En effet, comme alternative à l'algorithme de Viterbi qui calcule la séquence d'états $\widetilde{s}_1^T$ maximisant globalement $P\left(S_1^T = s_1^T, Y_1^T = y_1^T\right)$, il est possible de calculer une séquence d'états optimale sur la base d'un critère local, c'est-à-dire de déterminer pour chaque instant t l'état le plus vraisemblable

$$\widetilde{s}_t = \arg\max_j P\left(S_t = j | Y_1^T = y_1^T\right) = \arg\max_j L_j\left(t\right).$$

L'algorithme "avant-arrière" peut directement être appliqué pour calculer la séquence d'états optimale sur la base d'un tel critère local. Le calcul de la séquence d'états par l'algorithme de Viterbi a comme propriété d'optimalité de maximiser la probabilité de la séquence d'états entière, alors que le calcul de la séquence d'états par l'algorithme "avant-arrière" a comme propriété d'optimalité de maximiser le nombre moyen d'états "corrects".

Chapitre 3

Le modèle linéaire mixte multiphasique : présentation et méthodes d'estimation proposées

3.1 Introduction

D'un point de vue statistique, notre but est d'analyser des données longitudinales présentant les caractéristiques particulières suivantes :

- les données sont a priori structurées en phases successives,
- les données sont influencées par des covariables pouvant varier dans le temps,
- les données présentent une hétérogénéité inter-individuelle.

Pour modéliser ce type particulier de données, nous proposons une nouvelle famille de modèles appelés **modèles linéaires mixtes multiphasiques**. Ce modèle a la particularité d'avoir une double structure cachée car il résulte du "mariage" des deux modèles à structure cachée présentés dans le chapitre 2 : la chaîne de Markov cachée et le modèle linéaire mixte. Deux familles de modèles linéaires mixtes multiphasiques qui diffèrent par le choix de la modélisation de l'effet aléatoire seront présentées. Par la suite, nous nous intéresserons au problème de l'estimation des paramètres de ce modèle. Après avoir mis en évidence les difficultés liées à la double structure cachée du modèle pour l'estimation des paramètres par l'algorithme EM, nous envisagerons un autre algorithme de type EM prenant en compte les effets aléatoires. Cet algorithme itératif, sachant les effets aléatoires se compose de trois étapes : restauration probabiliste, maximisation et prédiction. Même si l'étape de restauration probabiliste et l'étape de maximisation sachant les effets aléatoires s'écrivent sans difficulté majeure, se pose le problème de la prédiction des effets aléatoires. Nous proposerons comme alternative à l'algorithme EM un algorithme itératif en trois étapes : restauration, maximisation et prédiction. L'étape de restauration pourra être déterministe et alors l'algorithme sera de type Baum-Viterbi ou bien elle pourra être effectuée par simulation et l'algorithme sera alors de type SEM.

3.2 Une nouvelle famille de modèles : le modèle linéaire mixte multiphasique

Nous définissons un modèle linéaire mixte multiphasique comme un modèle de type Markov caché qui combine :
• une chaîne de Markov pour modéliser la succession de phases,
• des modèles linéaires mixtes associés aux états de la chaîne de Markov sous-jacente. Pour chacune des phases, la tendance et les covariables sont modélisées par des effets fixes, et un effet aléatoire modélise l'hétérogénéité entre les individus.

Si nous appelons modèle linéaire multiphasique, un modèle de chaîne de Markov cachée dans le cas particulier où les observations sont modélisées avec des modèles linéaires, alors un modèle linéaire mixte multiphasique est un modèle linéaire multiphasique pour lequel on "rajoute" des effets aléatoires aux modèles linéaires. Une autre façon de définir ce modèle consiste à dire qu'il résulte de la combinaison markovienne de modèles linéaires mixtes. Il pourrait également s'appeler "Markov switching linear mixed model" dans le contexte Markov caché où il existe de nombreux modèles de type Markov caché combinant une chaîne de Markov et divers modèles (Ephraim et Merhav, 2002). Un modèle linéaire mixte multiphasique hérite donc de la structure cachée de chacun des deux modèles qui le composent : les effets aléatoires et les états cachés.

La figure 3.1 résume les liens qui existent (en terme d'effets aléatoires et de combinaison markovienne) entre quatre modèles : le modèle linéaire (LM), le modèle linéaire mixte (L2M), le modèle linéaire multiphasique (LM multiphasique) et le modèle linéaire mixte multiphasique (L2M multiphasique).

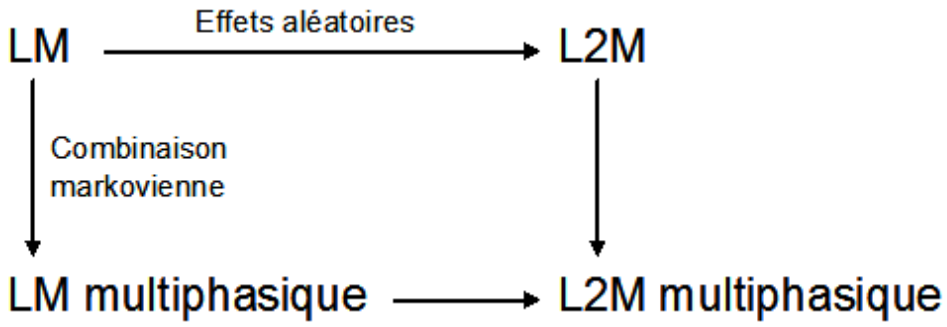

FIG. 3-1 – Liens en termes d'effets aléatoires et de combinaison markovienne entre les modèles LM, L2M, LM multiphasique et L2M multiphasique.

- les flèches horizontales représentent l'introduction d'effets aléatoires,
- les flèches verticales représentent l'introduction d'une combinaison markovienne (c'est-à-dire les modèles sont combinés de manière markovienne).

Nous présentons deux familles de modèles linéaires mixtes multiphasiques qui diffèrent par le choix de la modélisation de l'effet aléatoire : la séquence observée peut être modélisée avec un unique effet aléatoire ou avec un effet aléatoire différent sur chacun des états.

Notations :
Par la suite :
· a désigne l'individu et N est le nombre d'individus,
· T_a est la longueur de la séquence observée relative à l'individu a,
· $y_{a,t}$ est l'observation relative à l'individu a à l'instant t,
· $\{S_{a,t}; a = 1, \ldots, N; t = 1, \ldots, T_a\}$ est une chaîne de Markov d'ordre 1, à J états, indexée par le temps t.

3.2.1 Un seul effet aléatoire pour toute la séquence d'observations

L'observation $y_{a,t}$, relative à l'individu a se trouvant dans l'état $s_{a,t}$ à l'instant t, est modélisée par le modèle linéaire mixte suivant :

$$y_{a,t}|_{S_{a,t}=s_{a,t}} = \beta_{s_{a,t}} + \tau_{s_{a,t}}\xi_a + \varepsilon_{a,t}, \qquad t = 1, ..., T_a, \tag{3.1}$$

où
· $\beta_{s_{a,t}}$ est le terme des effets fixes pour l'état $s_{a,t}$,
· $\xi_a \sim N(0,1)$ est l'effet aléatoire relatif à l'individu a pour toute la séquence,
· $\tau_{s_{a,t}}$ est le coefficient multiplicateur de l'effet aléatoire ξ_a, relatif à l'état $s_{a,t}$. Le coefficient $\tau_{s_{a,t}}$ est strictement positif et $\tau_{s_{a,t}}\xi_a \sim N(0, \tau^2_{s_{a,t}})$.
· $\varepsilon_{a,t}|S_{a,t} = s_{a,t} \sim N(0, \sigma^2_{s_{a,t}})$ est le terme d'erreur relatif à l'individu a se trouvant dans l'état $s_{a,t}$ à l'instant t. Les $\varepsilon_{a,t}$ sont supposés indépendants entre eux et indépendants de ξ_a.
Par conséquent $y_{a,t}|S_{a,t} = s_{a,t} \sim N(\beta_{s_{a,t}}, \tau^2_{s_{a,t}} + \sigma^2_{s_{a,t}})$.

La séquence observée est modélisée avec un unique effet aléatoire, la variance induite par cet effet aléatoire pouvant différer pour chacune des phases. Ce choix de modélisation donne plus d'importance à l'individu qu'à l'état.

La figure 3.2 représente le graphe d'indépendance conditionnelle d'un modèle linéaire mixte multiphasique avec un unique effet aléatoire pour la séquence relative à l'individu a. Remarquons que d'après les notations définies à la section 2.3, la relation d'indépendance conditionnelle entre les observations, l'effet aléatoire et les états, s'écrit sous la forme :

$$Y_{a,t} \perp \left\{S_{a,1}^{t-1}, Y_{a,1}^{t-1}, S_{a,t+1}^{T_a}, Y_{a,t+1}^{T_a}\right\} | S_{a,t}, \xi_a.$$

Si on supprime les sommets $S_{a,t}$ et ξ_a du graphe d'indépendance conditionnelle, alors la variable aléatoire $Y_{a,t}$ est isolée de toutes les autres variables aléatoires. Autrement dit, sachant $S_{a,t}$ et ξ_a, la variable aléatoire $Y_{a,t}$ est indépendante de toutes les autres variables aléatoires.

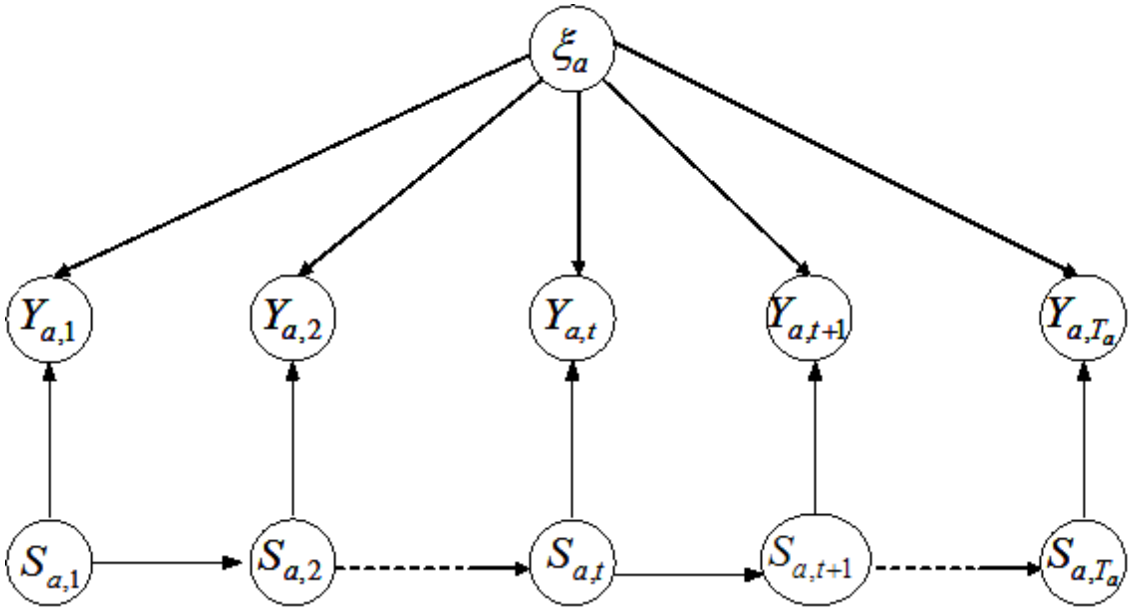

FIG. 3-2 – Graphe d'indépendance conditionnelle d'un modèle linéaire mixte multiphasique avec un unique effet aléatoire pour la séquence relative à l'individu a, et de longueur T_a.

Du point de vue de l'application botanique, cette modélisation correspond par exemple à un arbre qui, d'une façon générale, a une bonne croissance par rapport à la population sur toute sa période de croissance, mais modulée sur ses différentes phases de croissance.

3.2.2 Un effet aléatoire différent pour chaque état

Une seconde famille de modèles linéaires mixtes multiphasiques peut aussi être proposée. Dans le modèle (3.1), la séquence observée est modélisée avec un unique effet aléatoire. Il est possible de modéliser chaque phase avec un effet aléatoire différent. L'équation (3.1) devient alors :

$$y_{a,t}|_{S_{a,t}=s_{a,t}} = \beta_{s_{a,t}} + \xi_{a,s_{a,t}} + \varepsilon_{a,t}, \qquad t = 1, ..., T_a, \tag{3.2}$$

où les notations sont identiques à (3.1) si ce n'est que $\xi_{a,s_{a,t}}|S_{a,t} = s_{a,t} \sim N(0, \tau^2_{s_{a,t}})$ est l'effet aléatoire relatif à l'individu a se trouvant dans l'état $s_{a,t}$ à l'instant t.
On suppose que $\xi_{a,s_{a,t}}$ et $\xi_{a,s_{a,t'}}$ sont indépendants dès lors que $s_{a,t} \neq s_{a,t'}$. Autrement dit, on suppose que les $\xi_{a,j}, j = 1, ...J$ sont mutuellement indépendants.

Comme pour le modèle avec un unique effet aléatoire, on a

$$y_{a,t}|S_{a,t} = s_{a,t} \sim N(\beta_{s_{a,t}}, \tau^2_{s_{a,t}} + \sigma^2_{s_{a,t}}).$$

La loi d'une observation sachant l'état est donc identique pour les deux modèles. En revanche, la loi d'une observation sachant l'état et l'effet aléatoire est différente pour chacun des deux modèles :
- pour le modèle avec un unique effet aléatoire, noté par la suite **modèle 1** :

$$y_{a,t}|S_{a,t} = s_{a,t}, \xi_a \sim N(\beta_{s_{a,t}} + \tau^2_{s_{a,t}}\xi_a, \sigma^2_{s_{a,t}}),$$

- pour le modèle avec autant d'effets aléatoires que d'états, noté par la suite **modèle 2** :

$$y_{a,t}|S_{a,t} = s_{a,t}, \xi_{a,s_{a,t}} \sim N(\beta_{s_{a,t}} + \xi_{a,s_{a,t}}, \sigma^2_{s_{a,t}}).$$

On peut remarquer que :

$$f\left(\xi_{a,s_{a,t}}\right) = \sum_j f\left(\xi_{a,s_{a,t}}|S_{a,t} = j\right) P\left(S_{a,t} = j\right),$$

ce qui signifie que l'effet aléatoire $\xi_{a,s_{a,t}}$ suit un mélange de lois gaussiennes. Si $S_{a,t} = j$, alors $\xi_{a,j} \sim N(0, \tau^2_j), j = 1, ...J$.

Contrairement au cas du modèle 1, il n'est pas possible d'établir un graphe d'indépendance conditionnelle pour le modèle 2. En effet, pour une phase j donnée, l'effet aléatoire $\xi_{a,j}$ est identique sur toute la phase et lie toutes les observations de la phase entre elles. Par conséquent, les observations sont liées par phases. Mais les instants délimitant les J phases n'étant pas connus, il est impossible de représenter sur un graphe les relations d'indépendance conditionnelle entre les différentes variables aléatoires du modèle 2.

La figure 3.3 représente la structure d'un modèle linéaire mixte multiphasique à trois états pour la séquence relative à l'individu a, de longueur T_a, et modélisée par un effet aléatoire différent sur chaque état. Nous avons choisi de montrer un modèle "gauche-droite", c'est-à-dire constitué d'une succession d'états transitoires et d'un état final absorbant. Cette structure caractérisée par une matrice des probabilités de transition triangulaire supérieure est particulièrement adaptée pour les applications à la croissance des plantes.

Les $(\pi_j = P\left(S_1 = j\right); j = 1, 2, 3)$ sont les probabilités initiales associées aux trois états et les $(p_{ij} = P\left(S_{t+1} = j|S_t = i\right); i = 1, 2, 3, j = i, \ldots, 3)$ sont les probabilités de transition entre les états.

Ce choix de modélisation, avec un effet aléatoire différent sur chaque état, met plus l'accent sur l'état que sur l'individu et correspond, par exemple en botanique, à un arbre qui peut bien pousser par rapport à la population sur certaines périodes de croissance, et qui ensuite peut moins bien pousser ou très mal pousser sur d'autres périodes.

3.3 Méthodes d'estimation

Étudions à présent le problème de l'estimation des paramètres d'un modèle linéaire mixte multiphasique. Ces paramètres sont de deux types. Il y a, d'une part, les paramètres

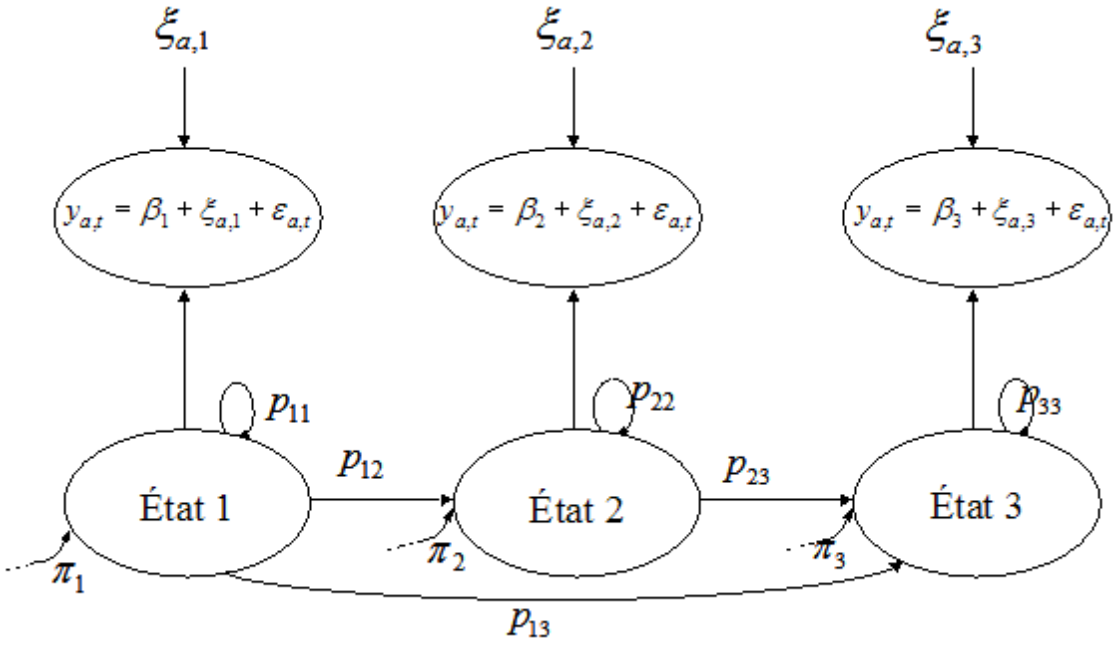

FIG. 3-3 – Structure "gauche-droite" d'un modèle linéaire mixte multiphasique à trois états pour la séquence relative à l'individu a, de longueur T_a, et modélisée avec un effet aléatoire différent sur chaque état.

liés à la chaîne de Markov sous-jacente et, d'autre part, les paramètres associés aux J modèles linéaires mixtes.

Les paramètres markoviens comprennent :

· les probabilités initiales associées aux J états : $(\pi_j = P(S_1 = j); j = 1, \ldots, J)$,

· les probabilités de transition entre les états : $(p_{ij} = P(S_{t+1} = j | S_t = i); i, j = 1, \ldots, J)$.

Les paramètres relatifs aux modèles linéaires mixtes sont constitués des J ensembles $\left\{\beta_j, \tau_j^2, \sigma_j^2; j = 1, \ldots, J\right\}$.

Le modèle linéaire mixte multiphasique suppose ainsi deux types de variables non-observables : les états de la chaîne de Markov et les effets aléatoires des modèles linéaires mixtes. L'algorithme EM est donc a priori approprié pour l'estimation des paramètres par maximum de vraisemblance. Au chapitre 2, est présenté l'algorithme EM pour l'estimation des paramètres de chacun des deux modèles qui composent le modèle linéaire mixte multiphasique et qui ne comportent chacun qu'un type de variable non-observable. Avant de mettre en évidence les difficultés qui apparaissent à l'étape E de l'EM pour chacun des deux modèles linéaires mixtes multiphasiques, et qui sont liées à la double structure cachée des modèles, présentons dans un premier temps l'algorithme EM dans le cadre particulier d'un modèle linéaire multiphasique.

3.3.1 Estimation des paramètres d'un modèle linéaire multiphasique avec l'algorithme EM

Cas d'une seule séquence d'observations

Rappelons que le modèle linéaire multiphasique est une chaîne de Markov cachée pour laquelle les observations sont modélisées avec des modèles linéaires. L'algorithme EM a été présenté au chapitre 2 (section 2.4.3), dans un cadre non-paramétrique pour l'estimation des paramètres d'une chaîne de Markov cachée. Appliquons les résultats obtenus au cas particulier où l'observation $y_{a,t}$ se trouvant dans l'état $s_{a,t}$ à l'instant t est modélisée par le modèle linéaire suivant :

$$y_{a,t}|_{S_{a,t}=s_{a,t}} = \beta_{s_{a,t}} + \varepsilon_{a,t}, \qquad t = 1, ..., T_a, \tag{3.3}$$

où

· $\beta_{s_{a,t}}$ est le terme des effets fixes pour l'état $s_{a,t}$ dans lequel se trouve l'individu a à l'instant t,

· $\varepsilon_{a,t}$ est le terme d'erreur relatif à l'individu a se trouvant dans l'état $s_{a,t}$ à l'instant t et $\varepsilon_{a,t}|S_{a,t} = s_{a,t} \sim N(0, \sigma^2_{s_{a,t}})$. Les $\varepsilon_{a,t}$ sont supposés indépendants entre eux.

Par conséquent $y_{a,t}|S_{a,t} = s_{a,t} \sim N(\beta_{s_{a,t}}, \sigma^2_{s_{a,t}})$.

Notations :

Par la suite :

· $y_{a,1}^{T_a}$ désigne la séquence des observations $y_{a,1}, ..., y_{a,T_a}$ relative à l'individu a aux dates successives $t = 1, ..., T_a$,

· $s_{a,1}^{T_a}$ désigne la séquence des états cachés relative à l'individu a aux dates successives $t = 1, ..., T_a$,

· $\theta = (\pi_j, p_{ij}, \beta_j, \sigma_j^2)$, $i, j = 1, ..., J$ est le vecteur des paramètres à estimer,

· $\gamma = (\beta_j, \sigma_j^2)$, $j = 1, ..., J$ est le vecteur des paramètres à estimer relatifs aux J modèles linéaires,

· $\lambda = (\pi_j, p_{ij})$, $i, j = 1, ..., J$ est le vecteur des paramètres markoviens à estimer.

L'espérance conditionnelle de la log-vraisemblance des données complètes connaissant la séquence des observations et la valeur des paramètres à l'itération k, dans le cadre d'une chaîne de Markov cachée, s'écrit dans notre cas particulier où la loi d'observation est la loi normale $N\left(\beta_j, \sigma_j^2\right)$, sous la forme :

$$
\begin{aligned}
& E\left(\log f(y_{a,1}^{T_a}, s_{a,1}^{T_a}; \theta) | Y_{a,1}^{T_a} = y_{a,1}^{T_a}; \theta^{(k)}\right) \\
= & \sum_{j=1}^{J} P(S_{a,1} = j | Y_{a,1}^{T_a} = y_{a,1}^{T_a}; \theta^{(k)}) \log \pi_j \\
& + \sum_{i,j=1}^{J} \left\{ \sum_{t=2}^{T_a} P\left(S_{a,t-1} = i, S_{a,t} = j | Y_{a,1}^{T_a} = y_{a,1}^{T_a}; \theta^{(k)}\right) \right\} \log p_{ij} \\
-\frac{T_a}{2} \log 2\pi + \sum_{j=1}^{J} \sum_{t=1}^{T_a} P(S_{a,t} & = j | Y_{a,1}^{T_a} = y_{a,1}^{T_a}; \theta^{(k)}) \left(-\log \sigma_j - \frac{1}{2\sigma_j^2} \left(y_{a,t} - \beta_j\right)^2 \right), \qquad (3.4)
\end{aligned}
$$

car

$$
E\left(\left(y_{a,t} - \beta_j\right)^2 | Y_{a,1}^{T_a} = y_{a,1}^{T_a}, S_{a,t} = j; \theta^{(k)} \right) = \left(y_{a,t} - \beta_j\right)^2 .
$$

Les quantités $L_{a,j}^{(k)}(t) = P(S_{a,t} = j | Y_{a,1}^{T_a} = y_{a,1}^{T_a}; \theta^{(k)})$ et $P\left(S_{a,t} = j, S_{a,t-1} = i | Y_{a,1}^{T_a} = y_{a,1}^{T_a}; \theta^{(k)}\right)$ sont calculées de manière récursive par l'algorithme "avant-arrière" décrit à la section 2.4.3 (chapitre 2).

Estimation des paramètres markoviens π_j et p_{ij}

L'expression des estimateurs des paramètres π_j et p_{ij} est identique à celle donnée à la section 2.4.3. Voici pour rappel les expressions des estimateurs à l'itération k :

$$
\pi_j^{(k+1)} = L_{a,j}^{(k)}(1) \qquad j = 1, ..., J,
$$

$$
p_{ij}^{(k+1)} = \frac{\sum_{t=1}^{T_a - 1} L_{a,j}^{(k)}(t+1) p_{ij}^{(k)} F_{a,i}^{(k)}(t) / G_{a,j}^{(k)}(t+1)}{\sum_{t=1}^{T_a - 1} L_{a,i}^{(k)}(t)} \qquad i, j = 1, ..., J,
$$

où
$F_{a,i}^{(k)}(t) = P\left(S_{a,t} = i | Y_{a,1}^{t} = y_{a,1}^{t}; \theta^{(k)}\right)$ est la probabilité filtrée,
et $G_{a,j}^{(k)}(t+1) = P\left(S_{a,t+1} = k | Y_{a,1}^{t} = y_{a,1}^{t}; \theta^{(k)}\right)$ est la probabilité prédite.

Estimation des paramètres liés aux modèles linéaires β_j et σ_j^2

L'annulation successive de la dérivée de (3.4) par rapport à β_j et à σ_j^2 conduit aux formules de réestimation suivantes :

$$
\beta_j^{(k+1)} = \frac{\sum_{t=1}^{T_a} L_{a,j}^{(k)}(t)\, y_{a,t}}{\sum_{t=1}^{T_a} L_{a,j}^{(k)}(t)} \qquad j = 1, ..., J,
$$

$$\sigma_j^{2(k+1)} = \frac{\sum_{t=1}^{T_a} L_{a,j}^{(k)}(t) \left(y_{a,t} - \beta_j^{(k+1)}\right)^2}{\sum_{t=1}^{T_a} L_{a,j}^{(k)}(t)} \qquad j = 1, ..., J.$$

Généralisation au cas de N séquences

Supposons que les N individus sont indépendants entre eux. La généralisation de l'estimation des paramètres du modèle linéaire multiphasique pour l'ensemble des N séquences observées est immédiate et conduit aux formules de réestimation suivantes :

Estimation des paramètres markoviens π_j et p_{ij}

$$\pi_j^{(k+1)} = \frac{\sum_{a=1}^{N} L_{a,j}^{(k)}(1)}{N} \qquad j = 1, ..., J.$$

$$p_{ij}^{(q+1)} = \frac{\sum_{a=1}^{N} \sum_{t=1}^{T_a-1} L_{a,j}^{(k)}(t+1) p_{ij}^{(k)} F_{a,i}^{(k)}(t) / G_{a,j}^{(k)}(t+1)}{\sum_{a=1}^{N} \sum_{t=1}^{T_a-1} L_{a,i}^{(k)}(t)} \qquad i, j = 1, ..., J.$$

Estimation des paramètres liés aux modèles linéaires β_j et σ_j^2

$$\beta_j^{(k+1)} = \frac{\sum_{a=1}^{N} \sum_{t=1}^{T_a} L_{a,j}^{(k)}(t)\, y_{a,t}}{\sum_{a=1}^{N} \sum_{t=1}^{T_a} L_{a,j}^{(k)}(t)} \qquad j = 1, ..., J,$$

$$\sigma_j^{2(k+1)} = \frac{\sum_{a=1}^{N} \sum_{t=1}^{T_a} L_{a,j}^{(k)}(t) \left(y_{a,t} - \beta_j^{(k+1)}\right)^2}{\sum_{a=1}^{N} \sum_{t=1}^{T_a} L_{a,j}^{(k)}(t)} \qquad j = 1, ..., J.$$

L'estimation des paramètres du modèle linéaire multiphasique s'écrit sans difficulté avec l'algorithme EM. À présent, transformons le modèle linéaire multiphasique en modèle linéaire mixte multiphasique en lui ajoutant une seconde structure cachée, à savoir un ou plusieurs effets aléatoires non-observé(s), et intéressons nous à l'estimation des paramètres avec l'algorithme EM.

3.3.2 Estimation des paramètres d'un modèle linéaire mixte multiphasique avec l'algorithme EM

Nous présentons, dans ce paragraphe, les difficultés rencontrées pour l'estimation des paramètres des modèles 1 et 2, avec l'algorithme EM. Nous nous restreignons au cas d'une seule séquence d'observations.

Un seul effet aléatoire pour toute la séquence d'observations

Rappelons que l'observation $y_{a,t}$, relative à l'individu a se trouvant dans l'état $s_{a,t}$ à l'instant t, est modélisée par le modèle linéaire mixte suivant :

$$y_{a,t}|_{S_{a,t}=s_{a,t}} = \beta_{s_{a,t}} + \tau_{s_{a,t}}\xi_a + \varepsilon_{a,t}, \qquad t = 1, ..., T_a.$$

Remarque :
Jusqu'à présent, nous avons noté $y_{a,1}^{T_a}$ la séquence des observations $y_{a,1}, ..., y_{a,T_a}$ relative à l'individu a aux dates successives $t = 1, ..., T_a$. Nous réserverons cette notation pour l'écriture de la (log-) vraisemblance, ou pour des écritures qui nécessitent une décomposition de la séquence en plusieurs morceaux, comme pour la présentation de l'algorithme "avant-arrière". En revanche, pour désigner le vecteur des observations dans une écriture matricielle, nous choisirons pour alléger l'écriture la notation y_a.

Si $Y_a = (Y_{a,1}, ..., Y_{a,T_a})'$ est le vecteur aléatoire correspondant à la séquence relative à l'individu a de longueur T_a, et ayant pour réalisation le vecteur $y_a = (y_{a,1}, ..., y_{a,T_a})'$, alors (3.1) s'écrit sous forme matricielle de la manière suivante :

$$Y_a|_{S_{a,1}^{T_a}=s_{a,1}^{T_a}} = X_a\beta + U_a\tau\xi_a + \varepsilon_a, \tag{3.5}$$

où
· $\beta = (\beta_1, \beta_2, ..., \beta_J)'$, de dimension $J \times 1$ est le vecteur des paramètres inconnus relatifs aux effets fixes dans les J différents états,
· X_a de dimension $T_a \times J$ est la matrice d'incidence de β. Chaque colonne de X_a est composée de valeurs nulles et de $N_{a,j}, j = 1, ..., J$ valeurs égales à 1, où $N_{a,j}$ est le nombre d'observations de la séquence relative à l'individu a se trouvant dans l'état j, et vérifie $\sum_{j=1}^{J} N_{a,j} = T_a$,
· ξ_a de dimension 1×1 est l'effet aléatoire unique pour toute la séquence,
· U_a de dimension $T_a \times J$ est la matrice d'incidence de ξ_a. U_a est identique à X_a,
· $\tau = (\tau_1, \tau_2, ..., \tau_J)'$ de dimension $J \times 1$ est le vecteur des coefficients multiplicateurs de l'effet aléatoire, autrement dit, c'est le vecteur des écarts-type engendrés par l'effet aléatoire sur les J différents états,
· $\varepsilon_a = (\varepsilon_{a,1}, \varepsilon_{a,2}, ..., \varepsilon_{a,T_a})'$ de dimension $T_a \times 1$ est le vecteur aléatoire d'erreurs. Comme tous les termes d'erreurs sont supposés indépendants entre eux, $\varepsilon_a|S_{a,1}^{T_a} = s_{a,1}^{T_a} \sim N(0, V_a)$ où V_a est une matrice diagonale par blocs, de dimension $T_a \times T_a$, composée de J blocs eux mêmes diagonaux : $V_{a,j} = \sigma_j^2 I_{Na,j}$.

Avec les notations définies ci-dessus, la matrice de variance-covariance de Y_a sachant la séquence des états s'écrit sous la forme :

$$\text{var}\left(Y_a|S_{a,1}^{T_a} = s_{a,1}^{T_a}\right) = U_a\tau\tau'U_a' + V_a.$$

La covariance entre deux observations sachant la séquence d'états s'écrit :

$$\text{cov}\left(y_{a,t}, y_{a,t'}|S_{a,1}^{T_a} = s_{a,1}^{T_a}\right) = \begin{cases} \tau^2_{s_{a,t}} + \sigma^2_{s_{a,t}} & \text{si } t = t', \\ \tau^2_{s_{a,t}} & \text{si } s_{a,t} = s_{a,t'} \text{ et } t \neq t', \\ \tau_{s_{a,t}}\tau_{s_{a,t'}} & \text{sinon.} \end{cases}$$

Écriture de la densité des données complètes
Nous noterons :
· $\theta = (\pi_j, p_{ij}, \beta_j, \tau_j, \sigma_j^2)$, $i, j = 1, ..., J$, le vecteur des paramètres à estimer,
· $\gamma = (\beta_j, \tau_j, \sigma_j^2)$, $j = 1, ..., J$, le vecteur des paramètres à estimer relatifs aux J modèles linéaires mixtes,
· $\lambda = (\pi_j, p_{ij})$, $i, j = 1, ..., J$, le vecteur des paramètres markoviens à estimer.

Pour la spécification du problème aux données complètes, supposons que les séquences $y_{a,1}^{T_a}$ et $s_{a,1}^{T_a}$ sont observées ainsi que l'effet aléatoire ξ_a. Comme ξ_a est l'unique effet aléatoire, il est indépendant de la séquence des états $s_{a,1}^{T_a}$. Ceci se traduit sur le graphe d'indépendance conditionnelle (figure 3.2) par le fait qu'aucune flèche ne lie directement la variable aléatoire ξ_a à l'une des variables aléatoires $S_{a,1}, ..., S_{a,T_a}$. La densité des données complètes s'écrit donc :

$$\begin{aligned} f\left(y_{a,1}^{T_a}, s_{a,1}^{T_a}, \xi_a; \theta\right) &= f\left(y_{a,1}^{T_a}|\xi_a, s_{a,1}^{T_a}; \gamma\right) f\left(\xi_a, s_{a,1}^{T_a}; \gamma, \lambda\right) \\ &= f\left(y_{a,1}^{T_a}|\xi_a, s_{a,1}^{T_a}; \gamma\right) f\left(\xi_a; \gamma\right) f\left(s_{a,1}^{T_a}; \lambda\right) \end{aligned} \tag{3.6}$$

Détaillons l'écriture des termes de (3.6).
Comme $\{S_{a,t}\}$ est une chaîne de Markov d'ordre 1, la loi de probabilité jointe de la séquence des états est donnée par :

$$f\left(s_{a,1}^{T_a}; \lambda\right) = P\left(S_{a,1}^{T_a} = s_{a,1}^{T_a}; \lambda\right) = \pi_{s_{a,1}} \prod_{t=2}^{T_a} p_{s_{a,t-1}s_{a,t}}.$$

Il reste à expliciter le terme de la densité de la séquence des observations conditionnellement à la séquence des états et à l'effet aléatoire : $f\left(y_{a,1}^{T_a}|s_{a,1}^{T_a}, \xi_a; \gamma\right)$.
Nous savons que $Y_a|S_{a,1}^{T_a} = s_{a,1}^{T_a} \sim N(X_a\beta, U_a\tau\tau'U_a' + V_a)$. Par conséquent,

$$Y_a|S_{a,1}^{T_a} = s_{a,1}^{T_a}, \xi_a \sim N(X_a\beta + U_a\tau\xi_a, V_a). \tag{3.7}$$

La covariance entre deux observations sachant la séquence des états et l'effet aléatoire est donnée par :

$$\text{cov}(y_{a,t}, y_{a,t'}|S_{a,1}^{T_a} = s_{a,1}^{T_a}, \xi_a) = \begin{cases} \sigma^2_{s_{a,t}} & \text{si } t = t', \\ 0 & \text{sinon.} \end{cases}$$

Par conséquent, les observations sont conditionnellement indépendantes sachant la séquence des états et l'effet aléatoire et,

$$f\left(y_{a,1}^{T_a}|s_{a,1}^{T_a},\xi_a;\gamma\right)=\prod_{t=1}^{T_a}f\left(y_{a,t}|s_{a,t},\xi_a;\gamma\right).$$

D'après (3.7), $y_{a,t}|S_{a,t}=s_{a,t},\xi_a\sim N\left(\beta_{s_{a,t}}+\tau_{s_{a,t}}\xi_a,\sigma^2_{s_{a,t}}\right)$, alors :

$$f\left(y_{a,t}|s_{a,t},\xi_a;\gamma\right)=\frac{1}{\sqrt{2\pi}\sigma_{s_{a,t}}}\exp\left(-\frac{\left(y_{a,t}-\beta_{s_{a,t}}-\tau_{s_{a,t}}\xi_a\right)^2}{2\sigma^2_{s_{a,t}}}\right).$$

La densité des données complètes s'écrit donc :

$$\begin{aligned}f(y_{a,1}^{T_a},s_{a,1}^{T_a},\xi_a;\theta) &= \left\{\prod_{t=1}^{T_a}\frac{1}{\sqrt{2\pi}\sigma_{s_{a,t}}}\exp\left(-\frac{\left(y_{a,t}-\beta_{s_{a,t}}-\tau_{s_{a,t}}\xi_a\right)^2}{2\sigma^2_{s_{a,t}}}\right)\right\}\\ &\quad\times\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{\xi_a^2}{2}\right)\pi_{s_{a,1}}\prod_{t=2}^{T_a}p_{s_{a,t-1}s_{a,t}}.\end{aligned}$$

Il s'ensuit que la log-vraisemblance des données complètes s'écrit :

$$\begin{aligned}\log f(y_{a,1}^{T_a},s_{a,1}^{T_a},\xi_a;\theta) &= \log\pi_{s_{a,1}}+\sum_{t=2}^{T_a}\log p_{s_{a,t-1}s_{a,t}}-\frac{(T_a+1)}{2}\log 2\pi-\frac{\xi_a^2}{2}\\ &\quad+\sum_{t=1}^{T_a}\left(-\log\sigma_{s_{a,t}}-\frac{\left(y_{a,t}-\beta_{s_{a,t}}-\tau_{s_{a,t}}\xi_a\right)^2}{2\sigma^2_{s_{a,t}}}\right),\end{aligned}$$

ce qui peut se réécrire :

$$\begin{aligned}\log f(y_{a,1}^{T_a},s_{a,1}^{T_a},\xi_a;\theta) &= \sum_{j=1}^{J}I(s_{a,1}=j)\log\pi_j+\sum_{i,j=1}^{J}\sum_{t=2}^{T_a}I\left(s_{a,t-1}=i,s_{a,t}=j\right)\log p_{ij}\\ &\quad-\frac{(T_a+1)}{2}\log 2\pi-\frac{\xi_a^2}{2}\\ &\quad+\sum_{j=1}^{J}\sum_{t=1}^{T_a}I\left(s_{a,t}=j\right)\left(-\log\sigma_j-\frac{\left(y_{a,t}-\beta_j-\tau_j\xi_a\right)^2}{2\sigma_j^2}\right).\end{aligned}\tag{3.8}$$

Étape E de l'algorithme EM

Cette étape consiste à calculer l'espérance conditionnelle de la log-vraisemblance des données complètes connaissant la séquence des observations et la valeur des paramètres à l'itération k.

$$
\begin{aligned}
&E\left(\log f(y_{a,1}^{T_a}, s_{a,1}^{T_a}, \xi_a; \theta)|Y_{a,1}^{T_a} = y_{a,1}^{T_a}; \theta^{(k)}\right) \\
&= \sum_{j=1}^{J} L_{a,j}^{(k)}(1) \log \pi_j + \sum_{i,j=1}^{J} \sum_{t=2}^{T_a} P\left(S_{a,t} = j, S_{a,t-1} = i|Y_{a,1}^{T_a} = y_{a,1}^{T_a}; \theta^{(k)}\right) \log p_{ij} \\
&- \frac{(T_a+1)}{2} \log 2\pi - \frac{1}{2} E\left(\xi_a^2|Y_{a,1}^{T_a} = y_{a,1}^{T_a}; \theta^{(k)}\right) \\
&+ E\left\{\sum_{j=1}^{J} \sum_{t=1}^{T_a} I\left(s_{a,t} = j\right) \left(-\log \sigma_j - \frac{\left(y_{a,t} - \beta_j - \tau_j \xi_a\right)^2}{2\sigma_j^2}\right) |Y_{a,1}^{T_a} = y_{a,1}^{T_a}; \theta^{(k)}\right\},
\end{aligned}
$$

avec

$$
L_{a,j}^{(k)}(1) = P\left(S_{a,1} = j|Y_{a,1}^{T_a} = y_{a,1}^{T_a}; \theta^{(k)}\right).
$$

De plus,

$$
\begin{aligned}
& E\left\{\sum_{j=1}^{J} \sum_{t=1}^{T_a} I\left(s_{a,t} = j\right) \left(-\log \sigma_j - \frac{\left(y_{a,t} - \beta_j - \tau_j \xi_a\right)^2}{2\sigma_j^2}\right) |Y_{a,1}^{T_a} = y_{a,1}^{T_a}; \theta^{(k)}\right\} \\
= & \sum_{j=1}^{J} \sum_{t=1}^{T_a} L_{a,j}^{(k)}(t)\, E\left(-\log \sigma_j - \frac{\left(y_{a,t} - \beta_j - \tau_j \xi_a\right)^2}{2\sigma_j^2} |Y_{a,1}^{T_a} = y_{a,1}^{T_a}, S_{a,t} = j; \theta^{(k)}\right) \\
= & \sum_{j=1}^{J} \sum_{t=1}^{T_a} L_{a,j}^{(k)}(t) \left(-\log \sigma_j - \frac{1}{2\sigma_j^2} E\left(\left(y_{a,t} - \beta_j - \tau_j \xi_a\right)^2 |Y_{a,1}^{T_a} = y_{a,1}^{T_a}, S_{a,t} = j; \theta^{(k)}\right)\right).
\end{aligned}
$$

Par conséquent

$$
\begin{aligned}
&E\left(\log f(y_{a,1}^{T_a}, s_{a,1}^{T_a}, \xi_a; \theta)|Y_{a,1}^{T_a} = y_{a,1}^{T_a}; \theta^{(k)}\right) \\
&= \sum_{j=1}^{J} L_{a,j}^{(k)}(1) \log \pi_j + \sum_{i,j=1}^{J} \sum_{t=2}^{T_a} P\left(S_{a,t} = j, S_{a,t-1} = i|Y_{a,1}^{T_a} = y_{a,1}^{T_a}; \theta^{(k)}\right) \log p_{ij} \\
&- \frac{(T_a+1)}{2} \log 2\pi - \frac{1}{2} E\left(\xi_a^2|Y_{a,1}^{T_a} = y_{a,1}^{T_a}; \theta^{(k)}\right) \\
&+ \sum_{j=1}^{J} \sum_{t=1}^{T_a} L_{a,j}^{(k)}(t) \left(-\log \sigma_j - \frac{1}{2\sigma_j^2} E\left(\left(y_{a,t} - \beta_j - \tau_j \xi_a\right)^2 |Y_{a,1}^{T_a} = y_{a,1}^{T_a}, S_{a,t} = j; \theta^{(k)}\right)\right),
\end{aligned}
\tag{3.9}
$$

avec

$$
\begin{aligned}
&E\left(\left(y_{a,t}-\beta_j-\tau_j\xi_a\right)^2|Y_{a,1}^{T_a}=y_{a,1}^{T_a},S_{a,t}=j;\theta^{(k)}\right)\\
&=y_{a,t}^2-2y_{a,t}\beta_j-2y_{a,t}\tau_jE\left(\xi_a|Y_{a,1}^{T_a}=y_{a,1}^{T_a},S_{a,t}=j;\theta^{(k)}\right)+\beta_j^2\\
&+\tau_j^2E\left(\xi_a^2|Y_{a,1}^{T_a}=y_{a,1}^{T_a},S_{a,t}=j;\theta^{(k)}\right)\\
&+2\beta_j\tau_jE\left(\xi_a|Y_{a,1}^{T_a}=y_{a,1}^{T_a},S_{a,t}=j;\theta^{(k)}\right).
\end{aligned}
$$

À ce stade, apparaissent trois difficultés que nous allons expliciter les unes après les autres :
(1) le calcul de $L_{a,j}^{(k)}(t)$,
(2) le calcul de $E\left(\xi_a^2|Y_{a,1}^{T_a}=y_{a,1}^{T_a};\theta^{(k)}\right)$,
(3) le calcul de $E\left(\xi_a|Y_{a,1}^{T_a}=y_{a,1}^{T_a},S_{a,t}=j;\theta^{(k)}\right)$ et de $E\left(\xi_a^2|Y_{a,1}^{T_a}=y_{a,1}^{T_a},S_{a,t}=j;\theta^{(k)}\right)$.

(1) Calcul de $L_{a,j}^{(k)}(t)$

Le calcul de $L_{a,j}^{(k)}(t)$ avec l'algorithme "avant-arrière", décrit au chapitre 2 dans le cadre de l'estimation d'une chaîne de Markov cachée, n'est plus valable ici. En effet, pour une chaîne de Markov cachée, les observations sont conditionnellement indépendantes sachant les états alors que pour un modèle linéaire mixte multiphasique, où il y a en plus la présence de l'effet aléatoire, les observations sont conditionnellement indépendantes sachant les états et l'effet aléatoire.

Le calcul ne peut donc pas s'effectuer tel quel, car l'effet aléatoire lie toutes les observations entre elles, et il est impossible de découper la séquence en deux sous-séquences : l'une pour les instants allant de 1 à t et l'autre pour les instants allant de $t+1$ à T_a (cf figure 3.2).

(2) Calcul de $E\left(\xi_a^2|Y_{a,1}^{T_a}=y_{a,1}^{T_a};\theta^{(k)}\right)$

Comme

$$
E\left(\xi_a^2|Y_{a,1}^{T_a}=y_{a,1}^{T_a};\theta^{(k)}\right)=E\left(E\left(\xi_a^2|Y_{a,1}^{T_a}=y_{a,1}^{T_a},S_{a,1}^{T_a}=s_{a,1}^{T_a};\theta^{(k)}\right)|Y_{a,1}^{T_a}=y_{a,1}^{T_a};\theta^{(k)}\right),
$$

nous souhaitons tout d'abord déterminer $E\left(\xi_a^2|Y_{a,1}^{T_a}=y_{a,1}^{T_a},S_{a,1}^{T_a}=s_{a,1}^{T_a};\theta^{(k)}\right)$.

La loi du vecteur $\begin{pmatrix}\xi_a\\ Y_a|S_{a,1}^{T_a}=s_{a,1}^{T_a}\end{pmatrix}$ est donnée par :

$$
\begin{pmatrix}\xi_a\\ Y_a|S_{a,1}^{T_a}=s_{a,1}^{T_a}\end{pmatrix}\sim N_{T_a+1}\left(\begin{pmatrix}0\\ X_a\beta\end{pmatrix},\begin{pmatrix}1 & \tau'U_a'\\ \tau U_a & \Gamma_a\end{pmatrix}\right),
$$

avec $\Gamma_a = U_a \tau \tau' {U_a}' + V_a$.

Par conséquent,

$$\xi_a | Y_{a,1}^{T_a} = y_{a,1}^{T_a}, S_{a,1}^{T_a} = s_{a,1}^{T_a} \sim N(\tau' U_a' \Gamma_a^{-1} (Y_a - X_a \beta), 1 - \tau' U_a' \Gamma_a^{-1} U_a \tau), \tag{3.10}$$

et il s'ensuit que :

$$E\left(\xi_a^2 | Y_{a,1}^{T_a} = y_{a,1}^{T_a}, S_{a,1}^{T_a} = s_{a,1}^{T_a}\right) = 1 - \tau' U_a' \Gamma_a^{-1} U_a \tau + (Y_a - X_a\beta)' \Gamma_a^{-1} U_a \tau \tau' U_a' \Gamma_a^{-1} (Y_a - X_a \beta).$$

Nous noterons $Q_a = 1 - \tau' U_a' \Gamma_a^{-1} U_a \tau + (Y_a - X_a\beta)' \Gamma_a^{-1} U_a \tau \tau' U_a' \Gamma_a^{-1} (Y_a - X_a \beta)$.

Ainsi,

$$\begin{aligned} E\left(\xi_a^2 | Y_{a,1}^{T_a} = y_{a,1}^{T_a}; \theta^{(k)}\right) &= E\left(Q_a | Y_{a,1}^{T_a} = y_{a,1}^{T_a}; \theta^{(k)}\right) \\ &= \sum_{s_{a,1}^{T_a}} h\left(s_{a,1}^{T_a}\right) P\left(S_{a,1}^{T_a} = s_{a,1}^{T_a} | Y_{a,1}^{T_a} = y_{a,1}^{T_a}; \theta^{(k)}\right), \end{aligned}$$

où
· $\sum_{s_{a,1}^{T_a}}$ désigne la somme sur toutes les séquences d'états possibles (il y en a J^{T_a} si toutes les probabilités de transition sont strictement positives),
· $h\left(s_{a,1}^{T_a}\right)$ est la valeur de Q_a pour la séquence d'états $s_{a,1}^{T_a}$.

La probabilité conditionnelle $P\left(S_{a,1}^{T_a} = s_{a,1}^{T_a} | Y_{a,1}^{T_a} = y_{a,1}^{T_a}; \theta^{(k)}\right)$ se décompose sous la forme suivante :

$$P\left(S_{a,1}^{T_a} = s_{a,1}^{T_a} | Y_{a,1}^{T_a} = y_{a,1}^{T_a}\right) = \frac{P\left(Y_{a,1}^{T_a} = y_{a,1}^{T_a} | S_{a,1}^{T_a} = s_{a,1}^{T_a}\right) P\left(S_{a,1}^{T_a} = s_{a,1}^{T_a}\right)}{P\left(Y_{a,1}^{T_a} = y_{a,1}^{T_a}\right)}$$

Dans le cadre d'observations conditionnellement indépendantes sachant les états, la probabilité $P\left(Y_{a,1}^{T_a} = y_{a,1}^{T_a}\right)$ se calcule comme le produit des facteurs de normalisation obtenus au cours de l'algorithme "avant-arrière" décrit au chapitre 2. Ici, les observations étant conditionnellement indépendantes sachant les états et l'effet aléatoire, cette probabilité ne peut pas se calculer telle quelle.

Par conséquent, le calcul de $E\left(\xi_a^2 | Y_{a,1}^{T_a} = y_{a,1}^{T_a}; \theta^{(k)}\right)$ n'est pas possible.

(3) **Calcul de** $E\left(\xi_a | Y_{a,1}^{T_a} = y_{a,1}^{T_a}, S_{a,t} = j; \theta^{(k)}\right)$ et de $E\left(\xi_a^2 | Y_{a,1}^{T_a} = y_{a,1}^{T_a}, S_{a,t} = j; \theta^{(k)}\right)$

La difficulté vient du fait que l'on conditionne non pas par rapport à toute la séquence d'états, mais seulement par rapport à l'état à l'instant t. Connaissant la loi de l'effet

aléatoire ξ_a et la loi conditionnelle $Y_{a,1}^{T_a}|S_{a,1}^{T_a} = s_{a,1}^{T_a}$, il est possible de déterminer la loi conditionnelle de $\xi_a|Y_{a,1}^{T_a} = y_{a,1}^{T_a}, S_{a,1}^{T_a} = s_{a,1}^{T_a}$. Cependant, les lois conditionnelles de $\xi_a|Y_{a,1}^{T_a} = y_{a,1}^{T_a}, S_{a,1}^{T_a} = s_{a,1}^{T_a}$ et de $\xi_a|Y_{a,1}^{T_a} = y_{a,1}^{T_a}, S_{a,t} = j$ sont différentes car l'effet aléatoire est commun à toutes les observations et donc à tous les états. Par conséquent, l'effet aléatoire et les états ne sont pas conditionnellement indépendants sachant la séquence d'observations.

Nous pouvons décomposer $P\left(\xi_a|Y_{a,1}^{T_a} = y_{a,1}^{T_a}, S_{a,t} = j\right)$ sous la forme suivante :

$$P\left(\xi_a|Y_{a,1}^{T_a} = y_{a,1}^{T_a}, S_{a,t} = j\right) = \frac{P\left(S_{a,t} = j|Y_{a,1}^{T_a} = y_{a,1}^{T_a}, \xi_a\right) P\left(Y_{a,1}^{T_a} = y_{a,1}^{T_a}|\xi_a\right) f(\xi_a)}{P\left(S_{a,t} = j|Y_{a,1}^{T_a} = y_{a,1}^{T_a}\right) P\left(Y_{a,1}^{T_a} = y_{a,1}^{T_a}\right)}$$

Les quantités $L_{a,j}(t) = P\left(S_{a,t} = j|Y_{a,1}^{T_a} = y_{a,1}^{T_a}\right)$ et $P\left(Y_{a,1}^{T_a} = y_{a,1}^{T_a}\right)$ ne pouvant pas être calculées, le calcul des espérances conditionnelles $E\left(\xi_a|Y_{a,1}^{T_a} = y_{a,1}^{T_a}, S_{a,t} = j; \theta^{(k)}\right)$ et $E\left(\xi_a^2|Y_{a,1}^{T_a} = y_{a,1}^{T_a}, S_{a,t} = j; \theta^{(k)}\right)$ ne peut pas s'effectuer.

Après avoir explicité les trois difficultés rencontrées pour l'écriture de l'étape E de l'algorithme EM, nous pouvons conclure que cette étape ne s'écrit pas de manière analytique pour le modèle 1. Par conséquent, nous écartons l'algorithme EM comme méthode d'estimation pour les paramètres du modèle 1. Nous verrons à la section 3.3.4 la méthode d'estimation proposée comme alternative.

Un effet aléatoire différent pour chaque état

Rappelons que l'observation $y_{a,t}$, relative à l'individu a se trouvant dans l'état $s_{a,t}$ à l'instant t, est modélisée par le modèle linéaire mixte suivant :

$$y_{a,t}|_{S_{a,t}=s_{a,t}} = \beta_{s_{a,t}} + \xi_{a,s_{a,t}} + \varepsilon_{a,t}, \qquad t = 1, ..., T_a,$$

qui peut s'écrire sous forme matricielle de la manière suivante :

$$Y_a|_{S_{a,1}^{T_a}=s_{a,1}^{T_a}} = X_a\beta + U_a\xi_{a,1}^J + \varepsilon_a,$$

où les notations sont identiques à celles de (3.5) si ce n'est que $\xi_{a,1}^J = \left(\xi_{a,1}, \xi_{a,2}, ..., \xi_{a,J}\right)'$, de dimension $J \times 1$ est le vecteur des J effets aléatoires relatifs aux J états pour l'individu a. Pour $j = 1, ..., J$, $\xi_{a,j} \sim N(0, \tau_j^2)$.
Comme les $\xi_{a,j}$ sont supposés indépendants entre eux, alors $\xi_{a,1}^J|S_{a,1}^{T_a} = s_{a,1}^{T_a} \sim N_J(0, D)$ où D est une matrice diagonale de dimension $J \times J$ ayant pour termes diagonaux les τ_j^2 pour $j = 1, ..., J$.

Avec les notations définies ci-dessus, la matrice de variance-covariance de Y_a sachant la séquence des états s'écrit sous la forme :

$$\text{var}\left(Y_a|S_{a,1}^{T_a} = s_{a,1}^{T_a}\right) = U_aDU_a' + V_a.$$

La covariance entre deux observations sachant les états s'écrit :

$$\text{cov}\left(y_{a,t}, y_{a,t'}|S_{a,1}^{T_a} = s_{a,1}^{T_a}\right) = \begin{cases} \tau^2_{s_{a,t}} + \sigma^2_{s_{a,t}} & \text{si } t = t', \\ \tau^2_{s_{a,t}} & \text{si } s_{a,t} = s_{a,t'} \text{ et } t \neq t', \\ 0 & \text{sinon.} \end{cases}$$

Écriture de la densité des données complètes

Comme pour le modèle 1, nous noterons :

· $\theta = (\pi_j, p_{ij}, \beta_j, \tau_j^2, \sigma_j^2)$, $i, j = 1, ..., J$ le vecteur des paramètres à estimer,

· $\gamma = (\beta_j, \tau_j^2, \sigma_j^2)$, $j = 1, ..., J$ le vecteur des paramètres à estimer relatifs aux J modèles linéaires mixtes,

· $\lambda = (\pi_j, p_{ij})$, $i, j = 1, ..., J$ le vecteur des paramètres markoviens à estimer.

Pour la spécification du problème aux données complètes, supposons que les séquences $y_{a,1}^{T_a}$ et $s_{a,1}^{T_a}$ soient observées ainsi que le vecteur des effets aléatoires $\xi_{a,1}^J$. La densité des données complètes s'écrit :

$$\begin{aligned} f\left(y_{a,1}^{T_a}, s_{a,1}^{T_a}, \xi_{a,1}^J; \theta\right) &= f\left(y_{a,1}^{T_a}|\xi_{a,1}^J, s_{a,1}^{T_a}; \gamma\right) f\left(\xi_{a,1}^J, s_{a,1}^{T_a}; \gamma, \lambda\right) \\ &= f\left(y_{a,1}^{T_a}|\xi_{a,1}^J, s_{a,1}^{T_a}; \gamma\right) f\left(\xi_{a,1}^J|s_{a,1}^{T_a}; \gamma\right) f\left(s_{a,1}^{T_a}; \lambda\right). \end{aligned} \tag{3.11}$$

Détaillons l'écriture de chacun des termes de (3.11).

Comme pour le modèle 1, la loi de probabilité jointe de la séquence des états est donnée par :

$$f\left(s_{a,1}^{T_a}; \lambda\right) = \pi_{s_{a,1}} \prod_{t=2}^{T_a} p_{s_{a,t-1}s_{a,t}}.$$

Par hypothèse,

$$\text{cov}\left(\xi_{a,s_{a,t}}, \xi_{a,s_{a,t'}}|S_{a,1}^{T_a} = s_{a,1}^{T_a}\right) = \begin{cases} \tau^2_{s_{a,t}} & \text{si } s_{a,t} = s_{a,t'}, \\ 0 & \text{sinon.} \end{cases}$$

Autrement dit, les $\xi_{a,s_{a,t}}$ sont conditionnellement indépendants sachant la séquence des états : $\xi_{a,s_{a,t}}$ et $\xi_{a,s_{a,t'}}$ sont indépendants dès lors que $s_{a,t} \neq s_{a,t'}$. Les $\xi_{a,j}, j = 1, ..., J$ sont mutuellement indépendants et $\xi_{a,j} \sim N\left(0, \tau_j^2\right)$.

Par conséquent, la densité du vecteur des effets aléatoires conditionnellement à la séquence des états s'écrit simplement comme le produit des densités des J effets aléatoires $\xi_{a,j}$:

$$f\left(\xi_{a,1}^{J}|s_{a,1}^{T_a};\gamma\right) = f\left(\xi_{a,1}^{J}\right) = \prod_{j=1}^{J} f\left(\xi_{a,j}\right)$$
$$= \prod_{j=1}^{J} \frac{1}{\sqrt{2\pi}\tau_j} \exp\left(-\frac{\xi_{a,j}^2}{2\tau_j^2}\right).$$

Il reste à expliciter le terme de la densité de la séquence des observations conditionnellement au vecteur des effets aléatoires et à la séquence des états : $f\left(y_{a,1}^{T_a}|\xi_{a,1}^{J}, s_{a,1}^{T_a};\gamma\right)$. Nous savons que $Y_a|S_{a,1}^{T_a} = s_{a,1}^{T_a} \sim N\left(X_a\beta, U_aDU_a' + V_a\right)$. Par conséquent,

$$Y_a|\xi_{a,1}^{J}, S_{a,1}^{T_a} = s_{a,1}^{T_a} \sim N\left(X_a\beta + U_a\xi_{a,1}^{J}, V_a\right). \tag{3.12}$$

La covariance entre deux observations, sachant le vecteur des effets aléatoires et la séquence des états, est donnée par :

$$\text{cov}\left(y_{a,t}, y_{a,t'}|\xi_{a,1}^{J}, S_{a,1}^{T_a} = s_{a,1}^{T_a}\right) = \begin{cases} \sigma_{s_{a,t}}^2 & \text{si } t = t', \\ 0 & \text{sinon.} \end{cases}$$

Par conséquent, les observations sont conditionnellement indépendantes sachant le vecteur des effets aléatoires et la séquence des états, et

$$f\left(y_{a,1}^{T_a}|\xi_{a,1}^{J}, s_{a,1}^{T_a};\gamma\right) = \prod_{t=1}^{T_a} f\left(y_{a,t}|\xi_{a,s_{a,t}}, s_{a,t};\gamma\right).$$

De plus, d'après (3.12), $y_{a,t}|\xi_{a,s_{a,t}}, s_{a,t} \sim N\left(\beta_{s_{a,t}} + \xi_{a,s_{a,t}}, \sigma_{s_{a,t}}^2\right)$, alors :

$$f\left(y_{a,t}|\xi_{a,s_{a,t}}, s_{a,t};\gamma\right) = \frac{1}{\sqrt{2\pi}\sigma_{s_{a,t}}} \exp\left(-\frac{\left(y_{a,t} - \beta_{s_{a,t}} - \xi_{a,s_{a,t}}\right)^2}{2\sigma_{s_{a,t}}^2}\right).$$

La densité des données complètes s'écrit donc :

$$f\left(y_{a,1}^{T_a}, s_{a,1}^{T_a}, \xi_{a,1}^{J};\theta\right) = \left\{\prod_{t=1}^{T_a} \frac{1}{\sqrt{2\pi}\sigma_{s_{a,t}}} \exp\left(-\frac{(y_{a,t} - \beta_{s_{a,t}} - \xi_{s_{a,t}})^2}{2\sigma_{s_{a,t}}^2}\right)\right\}$$
$$\times \left\{\prod_{j=1}^{J} \frac{1}{\sqrt{2\pi}\tau_j} \exp\left(-\frac{\xi_{a,j}^2}{2\tau_j^2}\right)\right\} \pi_{s_{a,1}} \prod_{t=2}^{T_a} p_{s_{a,t-1}s_{a,t}}.$$

Il s'ensuit que la log-vraisemblance des données complètes s'écrit :

$$\begin{aligned}\log f\left(y_{a,1}^{T_a}, s_{a,1}^{T_a}, \xi_{a,1}^{J}; \theta\right) = \log \pi_{s_{a,1}} + \sum_{t=2}^{T_a} \log p_{s_{a,t-1}s_{a,t}} - \frac{(T_a + J)}{2} \log 2\pi \\ + \sum_{t=1}^{T_a} \left(- \log \sigma_{s_{a,t}} - \frac{\left(y_{a,t} - \beta_{s_{a,t}} - \xi_{s_{a,t}}\right)^2}{2\sigma_{s_{a,t}}^2} \right) \\ + \sum_{j=1}^{J} \left(- \log \tau_j - \frac{\xi_{a,j}^2}{2\tau_j^2} \right),\end{aligned}$$

ce qui peut se réécrire :

$$\begin{aligned}\log f\left(y_{a,1}^{T_a}, s_{a,1}^{T_a}, \xi_{a,1}^{J}; \theta\right) = \sum_{j=1}^{J} I\left(s_{a,1} = j\right) \log \pi_j + \sum_{i,j=1}^{J} \sum_{t=2}^{T_a} I\left(s_{a,t-1} = i, s_{a,t} = j\right) \log p_{ij} \\ - \frac{(T_a + J)}{2} \log 2\pi \\ + \sum_{j=1}^{J} \sum_{t=1}^{T_a} I\left(s_{a,t} = j\right) \left(- \log \sigma_j - \frac{\left(y_{a,t} - \beta_j - \xi_{a,j}\right)^2}{2\sigma_j^2} \right) \\ + \sum_{j=1}^{J} \left(- \log \tau_j - \frac{\xi_{a,j}^2}{2\tau_j^2} \right). \qquad (3.13)\end{aligned}$$

Étape E de l'algorithme EM

L'espérance conditionnelle de la log-vraisemblance des données complètes, connaissant la séquence d'observations et la valeur des paramètres à l'itération k, s'écrit de la manière suivante :

$$
\begin{aligned}
&E\left(\log f\left(y_{a,1}^{T_a}, s_{a,1}^{T_a}, \xi_{a,1}^{J}; \theta\right) | Y_{a,1}^{T_a} = y_{a,1}^{T_a}; \theta^{(k)}\right) \\
&= \sum_{j=1}^{J} L_{a,j}^{(k)}(1) \log \pi_j + \sum_{i,j=1}^{J} \sum_{t=2}^{T_a} P\left(S_{a,t} = j, S_{a,t-1} = i | Y_{a,1}^{T_a} = y_{a,1}^{T_a}; \theta^{(k)}\right) \log p_{ij} \\
&- \frac{(T_a + J)}{2} \log 2\pi \\
&+ \sum_{j=1}^{J} \sum_{t=1}^{T_a} L_{a,j}^{(k)}(t) \left(-\log \sigma_j - \frac{1}{2\sigma_j^2} E\left(\left(y_{a,t} - \beta_j - \xi_{a,j}\right)^2 | Y_{a,1}^{T_a} = y_{a,1}^{T_a}, S_{a,t} = j; \theta^{(k)}\right)\right) \\
&+ \sum_{j=1}^{J} \left(-\log \tau_j - \frac{1}{2\tau_j^2} E\left(\xi_{a,j}^2 | Y_{a,1}^{T_a} = y_{a,1}^{T_a}; \theta^{(k)}\right)\right),
\end{aligned}
$$

avec

$$
\begin{aligned}
&E\left(\left(y_{a,t} - \beta_j - \xi_{a,j}\right)^2 | Y_{a,1}^{T_a} = y_{a,1}^{T_a}, S_{a,t} = j; \theta^{(k)}\right) \\
&= y_{a,t}^2 - 2y_{a,t}\beta_j - 2y_{a,t} E\left(\xi_{a,j} | Y_{a,1}^{T_a} = y_{a,1}^{T_a}, S_{a,t} = j; \theta^{(k)}\right) + \beta_j^2 \\
&+ E(\xi_{a,j}^2 | Y_{a,1}^{T_a} = y_{a,1}^{T_a}, S_{a,t} = j; \theta^{(k)}) + 2\beta_j E\left(\xi_{a,j} | Y_{a,1}^{T_a} = y_{a,1}^{T_a}, S_{a,t} = j; \theta^{(k)}\right).
\end{aligned}
$$

Comme pour le modèle linéaire mixte multiphasique avec un seul effet aléatoire, trois difficultés apparaissent :
(1) le calcul de $L_{a,j}^{(k)}(t)$,
(2) le calcul des $E\left(\xi_{a,j}^2 | Y_{a,1}^{T_a} = y_{a,1}^{T_a}; \theta^{(k)}\right), j = 1..., J$,
(3) le calcul des $E\left(\xi_{a,j} | Y_{a,1}^{T_a} = y_{a,1}^{T_a}, S_{a,t} = j; \theta^{(k)}\right)$ et des $E\left(\xi_{a,j}^2 | Y_{a,1}^{T_a} = y_{a,1}^{T_a}, S_{a,t} = j; \theta^{(k)}\right), j = 1, ..., J$.

(1) **Calcul de** $L_{a,j}^{(k)}(t)$

Tout comme pour le modèle 1, le calcul de $L_{a,j}^{(k)}(t)$ ne peut plus s'effectuer avec l'algorithme "avant-arrière" décrit au chapitre 2 car, avec la présence des effets aléatoires, les observations sont conditionnellement indépendantes sachant le vecteur des effets aléatoires et la séquence des états. Le calcul n'est donc pas possible.

(2) **Calcul de** $E\left(\xi_{a,j}^2 | Y_{a,1}^{T_a} = y_{a,1}^{T_a}; \theta^{(k)}\right), j = 1..., J$

Nous souhaitons déterminer dans un premier temps $E\left(\xi_{a,j}^2 | Y_{a,1}^{T_a} = y_{a,1}^{T_a}, S_{a,1}^{T_a} = s_{a,1}^{T_a}; \theta^{(k)}\right)$.

La loi du vecteur $\begin{pmatrix} \xi_{a,j} \\ Y_a | S_{a,1}^{T_a} = s_{a,1}^{T_a} \end{pmatrix}$ est donnée par :

$$\begin{pmatrix} \xi_{a,j} \\ Y_a | S_{a,1}^{T_a} = s_{a,1}^{T_a} \end{pmatrix} \sim N_{T_a+1} \left(\begin{pmatrix} 0 \\ X_a\beta \end{pmatrix}, \begin{pmatrix} \tau_j^2 & C_j' U_a' \\ U_a C_j & \Gamma_a \end{pmatrix} \right),$$

où

· $\Gamma_a = U_a D U_a' + V_a$,

· $C_j = \begin{pmatrix} 0 \\ \vdots \\ \tau_j^2 \\ 0 \\ \vdots \end{pmatrix}$ est le vecteur de dimension $J \times 1$, composé de $J-1$ zéros et ayant la valeur τ_j^2 sur sa $j^{ème}$ ligne.

Par conséquent,

$$\xi_{a,j} | Y_{a,1}^{T_a} = y_{a,1}^{T_a}, S_{a,1}^{T_a} = s_{a,1}^{T_a} \sim N(C_j' U_a' \Gamma_a^{-1} (Y_a - X_a\beta), \tau_j^2 - C_j' U_a' \Gamma_a^{-1} U_a C_j), \qquad (3.14)$$

et il s'ensuit que :

$$E\left(\xi_{a,j}^2 | Y_{a,1}^{T_a} = y_{a,1}^{T_a}, S_{a,1}^{T_a} = s_{a,1}^{T_a}\right) = \tau_j^2 - C_j' U_a' \Gamma_a^{-1} U_a C_j + (Y_a - X_a\beta)' \Gamma_a^{-1} U_a C_j C_j' U_a' \Gamma_a^{-1} (Y_a - X_a\beta).$$

Nous noterons $R_a = \tau_j^2 - C_j' U_a' \Gamma_a^{-1} U_a C_j + (Y_a - X_a\beta)' \Gamma_a^{-1} U_a C_j C_j' U_a' \Gamma_a^{-1} (Y_a - X_a\beta)$.

Ainsi,

$$\begin{aligned} E\left(\xi_{a,j}^2 | Y_{a,1}^{T_a} = y_{a,1}^{T_a}; \theta^{(k)}\right) &= E\left(R_a | Y_{a,1}^{T_a} = y_{a,1}^{T_a}; \theta^{(k)}\right) \\ &= \sum_{s_{a,1}^{T_a}} g\left(s_{a,1}^{T_a}\right) P\left(S_{a,1}^{T_a} = s_{a,1}^{T_a} | Y_{a,1}^{T_a} = y_{a,1}^{T_a}; \theta^{(k)}\right) \end{aligned}$$

où

· $\sum_{s_{a,1}^{T_a}}$ désigne la somme sur toutes les séquences d'états possibles (il y en a J^{T_a} si toutes les probabilités de transition sont strictement positives),

· $g\left(s_{a,1}^{T_a}\right)$ est la valeur de R_a pour la séquence d'états $s_{a,1}^{T_a}$.

Tout comme pour le modèle 1, la probabilité conditionnelle $P\left(S_{a,1}^{T_a} = s_{a,1}^{T_a} | Y_{a,1}^{T_a} = y_{a,1}^{T_a}; \theta^{(k)}\right)$ ne se calcule pas. Par conséquent, le calcul des $E\left(\xi_{a,j}^2 | Y_{a,1}^{T_a} = y_{a,1}^{T_a}; \theta^{(k)}\right), j = 1, ..., J$ n'est pas possible.

(3) **Calcul des** $E\left(\xi_{a,j}|Y_{a,1}^{T_a}=y_{a,1}^{T_a},S_{a,t}=j;\theta^{(k)}\right)$ et des $E\left(\xi_{a,j}^2|Y_{a,1}^{T_a}=y_{a,1}^{T_a},S_{a,t}=j;\theta^{(k)}\right),j=1...,J$

Comme pour le modèle 1, la difficulté vient du fait que l'on conditionne seulement par rapport à l'état à l'instant t, et non pas par rapport à toute la séquence d'états. On peut décomposer $P\left(\xi_{a,j}|Y_{a,1}^{T_a}=y_{a,1}^{T_a},S_{a,t}=j\right)$ sous la forme suivante :

$$P\left(\xi_{a,j}|Y_{a,1}^{T_a}=y_{a,1}^{T_a},S_{a,t}=j\right)=\frac{P\left(S_{a,t}=j|Y_{a,1}^{T_a}=y_{a,1}^{T_a},\xi_{a,j}\right)P\left(Y_{a,1}^{T_a}=y_{a,1}^{T_a}|\xi_{a,j}\right)f(\xi_{a,j})}{P\left(S_{a,t}=j|Y_{a,1}^{T_a}=y_{a,1}^{T_a}\right)P\left(Y_{a,1}^{T_a}=y_{a,1}^{T_a}\right)}$$

Comme dans le cadre du modèle 1, les quantités $L_{a,j}(t)=P\left(S_{a,t}=j|Y_{a,1}^{T_a}=y_{a,1}^{T_a}\right)$ et $P\left(Y_{a,1}^{T_a}=y_{a,1}^{T_a}\right)$ ne pouvant pas être calculées, le calcul des espérances conditionnelles $E\left(\xi_{a,j}|Y_{a,1}^{T_a}=y_{a,1}^{T_a},S_{a,t}=j;\theta^{(k)}\right)$ et $E\left(\xi_{a,j}^2|Y_{a,1}^{T_a}=y_{a,1}^{T_a},S_{a,t}=j;\theta^{(k)}\right)$ n'est pas envisageable.

Ainsi l'étape E de l'algorithme EM ne s'écrit pas, non plus, de manière analytique pour le modèle 2. Par conséquent, nous écartons l'algorithme EM comme méthode d'estimation des paramètres des modèles 1 et 2.

Toutefois, les quantités $L_{a,j}(t)=P\left(S_{a,t}=j|Y_{a,1}^{T_a}=y_{a,1}^{T_a}\right)$ et $P\left(Y_{a,1}^{T_a}=y_{a,1}^{T_a}\right)$, qui posent problème pour l'écriture de l'étape E de l'algorithme EM, peuvent se calculer si on rajoute la présence de l'effet aléatoire dans le conditionnement. Par conséquent, nous avons envisagé d'écrire un autre algorithme de type EM, sachant l'effet aléatoire pour le modèle 1 et sachant le vecteur des effets aléatoires pour le modèle 2. Cet algorithme itératif est composé de trois étapes : restauration probabiliste, maximisation et prédiction.

3.3.3 Estimation par un algorithme itératif avec restauration probabiliste et sachant les effets aléatoires

Toutes les notations sont identiques à celles de la section 3.3.2, et nous restreignons la présentation au cas d'une seule séquence d'observations.
Dans cette section nous écrivons, pour les deux modèles, les étapes de restauration et de maximisation sachant les effets aléatoires, et nous présentons la difficulté liée à la prédiction des effets aléatoires qui nous amènera à proposer à la section suivante d'autres algorithmes itératifs, de type Baum-Viterbi et de type SEM.

Un seul effet aléatoire pour toute la séquence d'observations

Si $\theta^{(k)}$ est la valeur courante du paramètre et $\tilde{\xi}_a^{(k)}$ la valeur prédite courante de l'effet aléatoire, l'itération k de l'algorithme est décrite par les trois étapes suivantes :

Étape de restauration : on restaure, de manière probabiliste avec un algorithme "avant-arrière" sachant l'effet aléatoire, l'ensemble des séquences d'états possibles. On calcule

l'espérance conditionnelle de la log-vraisemblance des données complètes (3.8) sachant la séquence des observations, la valeur des paramètres à l'itération k, et la valeur prédite de l'effet aléatoire à l'itération k :

$$E\left(\log f(y_{a,1}^{T_a}, s_{a,1}^{T_a}, \xi_a; \theta) | Y_{a,1}^{T_a} = y_{a,1}^{T_a}, \widetilde{\xi}_a^{(k)}; \theta^{(k)}\right).$$

Étape de maximisation : on maximise par rapport à θ l'espérance conditionnelle calculée à l'étape de restauration pour réactualiser la valeur du paramètre. Autrement dit, on choisit $\theta^{(k+1)}$ de sorte que

$$\theta^{(k+1)} = \arg\max_{\theta\in\Theta} E\left(\log f(y_{a,1}^{T_a}, s_{a,1}^{T_a}, \xi_a; \theta) | Y_{a,1}^{T_a} = y_{a,1}^{T_a}, \widetilde{\xi}_a^{(k)}; \theta^{(k)}\right).$$

Étape de prédiction : on calcule la valeur prédite $\widetilde{\xi}_a^{(k+1)}$ à partir de $\theta^{(k+1)}$, la valeur du paramètre calculée à l'étape de maximisation :

$$\widetilde{\xi}_a^{(k+1)} = E\left(\xi_a | Y_{a,1}^{T_a} = y_{a,1}^{T_a}; \theta^{(k+1)}\right).$$

Nous allons à présent décrire en détail chacune de ces trois étapes.

Première étape : restauration probabiliste

D'après l'expression de la log-vraisemblance des données complètes (3.8), nous écrivons :

$$\begin{aligned}
&E\left(\log f(y_{a,1}^{T_a}, s_{a,1}^{T_a}, \xi_a; \theta) | Y_{a,1}^{T_a} = y_{a,1}^{T_a}, \widetilde{\xi}_a^{(k)}; \theta^{(k)}\right) \\
&= \sum_{j=1}^{J} L_{a,j}^{\prime(k)}(1) \log \pi_j + \sum_{i,j=1}^{J} \sum_{t=2}^{T_a} P\left(S_{a,t} = j, S_{a,t-1} = i | Y_{a,1}^{T_a} = y_{a,1}^{T_a}, \widetilde{\xi}_a^{(k)}; \theta^{(k)}\right) \log p_{ij} \\
&- \frac{(T_a+1)}{2} \log 2\pi - \frac{1}{2}\left(\widetilde{\xi}_a^{(k)}\right)^2 \\
&+ E\left\{\sum_{j=1}^{J} \sum_{t=1}^{T_a} I\left(s_{a,t} = j\right) \left(-\log \sigma_j - \frac{\left(y_{a,t} - \beta_j - \tau_j \xi_a\right)^2}{2\sigma_j^2}\right) | Y_{a,1}^{T_a} = y_{a,1}^{T_a}, \widetilde{\xi}_a^{(k)}; \theta^{(k)}\right\},
\end{aligned}$$

avec

$$L_{a,j}^{\prime(k)}(1) = P\left(S_{a,1} = j | Y_{a,1}^{T_a} = y_{a,1}^{T_a}, \widetilde{\xi}_a^{(k)}; \theta^{(k)}\right).$$

De plus,

$$\begin{aligned}
& E\left\{\sum_{j=1}^{J}\sum_{t=1}^{T_a} I\left(s_{a,t}=j\right)\left(-\log\sigma_j-\frac{\left(y_{a,t}-\beta_j-\tau_j\xi_a\right)^2}{2\sigma_j^2}\right)|Y_{a,1}^{T_a}=y_{a,1}^{T_a},\widetilde{\xi}_a^{(k)};\theta^{(k)}\right\}\\
= & \sum_{j=1}^{J}\sum_{t=1}^{T_a} L_{a,j}^{\prime(k)}(t)\,E\left(-\log\sigma_j-\frac{\left(y_{a,t}-\beta_j-\tau_j\xi_a\right)^2}{2\sigma_j^2}|Y_{a,1}^{T_a}=y_{a,1}^{T_a},S_{a,t}=j,\widetilde{\xi}_a^{(k)};\theta^{(k)}\right)\\
= & \sum_{j=1}^{J}\sum_{t=1}^{T_a} L_{a,j}^{\prime(k)}(t)\left(-\log\sigma_j-\frac{1}{2\sigma_j^2}\left(y_{a,t}-\beta_j-\tau_j\widetilde{\xi}_a^{(k)}\right)^2\right).
\end{aligned}$$

Par conséquent

$$\begin{aligned}
& E\left(\log f(y_{a,1}^{T_a},s_{a,1}^{T_a},\xi_a;\theta)|Y_{a,1}^{T_a}=y_{a,1}^{T_a},\widetilde{\xi}_a^{(k)};\theta^{(k)}\right)\\
& =\sum_{j=1}^{J} L_{a,j}^{\prime(k)}(1)\log\pi_j+\sum_{i,j=1}^{J}\sum_{t=2}^{T_a} P\left(S_{a,t}=j,S_{a,t-1}=i|Y_{a,1}^{T_a}=y_{a,1}^{T_a},\widetilde{\xi}_a^{(k)};\theta^{(k)}\right)\log p_{ij}\\
& -\frac{(T_a+1)}{2}\log 2\pi-\frac{1}{2}\left(\widetilde{\xi}_a^{(k)}\right)^2\\
& +\sum_{j=1}^{J}\sum_{t=1}^{T_a} L_{a,j}^{\prime(k)}(t)\left(-\log\sigma_j-\frac{1}{2\sigma_j^2}\left(y_{a,t}-\beta_j-\tau_j\widetilde{\xi}_a^{(k)}\right)^2\right). \qquad (3.15)
\end{aligned}$$

Le calcul de $L'_{a,j}(t)=P\left(S_{a,t}=j|Y_{a,1}^{T_a}=y_{a,1}^{T_a},\tilde{\xi}_a\right)$ se calcule avec un algorithme "avant-arrière" similaire à celui décrit au chapitre 2.

Il est possible de décomposer $L'_{a,j}(t)$ de la manière suivante :

$$\begin{aligned}
L'_{a,j}(t) &= P\left(S_{a,t}=j|Y_{a,1}^{T_a}=y_{a,1}^{T_a},\tilde{\xi}_a\right)\\
&= \frac{P\left(Y_{a,t+1}^{T_a}=y_{a,t+1}^{T_a}|S_{a,t}=j,\tilde{\xi}_a\right)}{P\left(Y_{a,t+1}^{T_a}=y_{a,t+1}^{T_a}|Y_{a,1}^{t}=y_{a,1}^{t},\tilde{\xi}_a\right)}P\left(S_{a,t}=j|Y_{a,1}^{t}=y_{a,1}^{t},\tilde{\xi}_a\right)\\
&= B'_{a,j}(t)F'_{a,j}(t), \qquad (3.16)
\end{aligned}$$

avec

$$B'_{a,j}(t)=\frac{P\left(Y_{a,t+1}^{T_a}=y_{a,t+1}^{T_a}|S_{a,t}=j,\tilde{\xi}_a\right)}{P\left(Y_{a,t+1}^{T_a}=y_{a,t+1}^{T_a}|Y_{a,1}^{t}=y_{a,1}^{t},\tilde{\xi}_a\right)},$$

et

$$F'_{a,j}(t) = P\left(S_{a,t} = j | Y_{a,1}^t = y_{a,1}^t, \tilde{\xi}_a\right).$$

Les quantités $F'_{a,j}(t)$ sont calculées dans la passe "avant", alors que les quantités $B'_{a,j}(t)$ ou les quantités $L'_{a,j}(t)$ sont calculées dans la passe "arrière".

Récurrence "avant"

Elle est initialisée pour $t = 1$ et $j = 1, ..., J$ par

$$\begin{aligned} F'_{a,j}(1) &= P\left(S_{a,1} = j | Y_{a,1} = y_{a,1}, \tilde{\xi}_a\right) \\ &= \frac{P\left(Y_{a,1} = y_{a,1} | S_{a,1} = j, \tilde{\xi}_a\right) f\left(\tilde{\xi}_a\right) P\left(S_{a,1} = j\right)}{P\left(Y_{a,1} = y_{a,1}, \tilde{\xi}_a\right)} \\ &= \frac{f\left(y_{a,1} | S_{a,1} = j, \tilde{\xi}_a\right) f\left(\tilde{\xi}_a\right) \pi_j}{N'_{a,1}}, \end{aligned}$$

où

· $f(y_{a,1}|S_{a,1} = j, \tilde{\xi}_a)$, densité de l'observation $y_{a,1}|S_{a,1} = j, \tilde{\xi}_a$ est la densité de la loi normale $N(\beta_j + \tau_j \tilde{\xi}_a, \sigma_j^2)$,

· $N'_{a,1}$, le facteur de normalisation est égal à :

$$\begin{aligned} N'_{a,1} &= P(Y_{a,1} = y_{a,1}, \tilde{\xi}_a) \\ &= \sum_{j=1}^{J} f\left(y_{a,1} | S_{a,1} = j, \tilde{\xi}_a\right) f\left(\tilde{\xi}_a\right) \pi_j. \end{aligned}$$

Pour $t = 2, ..., T_a$ et $j = 1, ..., J$, la récurrence "avant" s'écrit :

$$
\begin{aligned}
F'_{a,j}(t) &= P\left(S_{a,t}=j|Y_{a,1}^{t}=y_{a,1}^{t},\tilde{\xi}_a\right)\\
&= \frac{\sum_{i=1}^{J} P\left(S_{a,t}=j,S_{a,t-1}=i,Y_{a,t}=y_{a,t}|Y_{a,1}^{t-1}=y_{a,1}^{t-1},\tilde{\xi}_a\right)}{P\left(Y_{a,t}=y_{a,t}|Y_{a,1}^{t-1}=y_{a,1}^{t-1},\tilde{\xi}_a\right)}\\
&= \frac{1}{P\left(Y_{a,t}=y_{a,t}|Y_{a,1}^{t-1}=y_{a,1}^{t-1},\tilde{\xi}_a\right)}\\
&\times \sum_{i=1}^{J} P\left(Y_{a,t}=y_{a,t}|S_{a,t}=j,\tilde{\xi}_a\right) P\left(S_{a,t}=j|S_{a,t-1}=i\right) P\left(S_{a,t-1}=i|Y_{a,1}^{t-1}=y_{a,1}^{t-1},\tilde{\xi}_a\right)\\
&= \frac{f(y_{a,t}|S_{a,t}=j,\tilde{\xi}_a)}{N'_{a,t}} \sum_{i=1}^{J} p_{ij} F'_{a,i}(t-1), \qquad (3.17)
\end{aligned}
$$

où $N'_{a,t}=P(Y_{a,t}=y_{a,t}|Y_{a,1}^{t-1}=y_{a,1}^{t-1},\tilde{\xi}_a)$, le facteur de normalisation est égal à :

$$
\begin{aligned}
N'_{a,t} &= \sum_{j=1}^{J} P\left(S_{a,t}=j,Y_{a,t}=y_{a,t}|Y_{a,1}^{t-1}=y_{a,1}^{t-1},\tilde{\xi}_a\right)\\
&= \sum_{j=1}^{J} f\left(y_{a,t}|S_{a,t}=j,\tilde{\xi}_a\right) \sum_{i=1}^{J} p_{ij} F'_{a,i}(t-1),
\end{aligned}
$$

avec $f\left(y_{a,t}|S_{a,t}=j,\tilde{\xi}_a\right)$, densité de la loi normale $N\left(\beta_j+\tau_j\tilde{\xi}_a,\sigma_j^2\right)$.

Remarquons que cette récurrence "avant" est similaire à la récurrence "avant" décrite au chapitre 2 (section 2.4.3), si ce n'est qu'il y a en plus la présence de l'effet aléatoire dans le conditionnement des probabilités d'observation et que l'on injecte, pour l'initialisation $f\left(\tilde{\xi}_a\right)$, la densité de l'effet aléatoire prédit. En effet, $f\left(\tilde{\xi}_a\right)$ est présent dans l'écriture de $F'_{a,j}(1)$ et de fait dans l'écriture de $N'_{a,1}$. Sans le conditionnement par l'effet aléatoire (récurrence "avant" classique), $f\left(y_{a,t}|S_{a,t}=j\right)$, densité de $y_{a,t}|S_{a,t}=j$ est la densité de la loi normale $N\left(\beta_j,\tau_j^2+\sigma_j^2\right)$, alors qu'en conditionnant par l'effet aléatoire, $f\left(y_{a,t}|S_{a,t}=j,\tilde{\xi}_a\right)$, densité de $y_{a,t}|S_{a,t}=j,\tilde{\xi}_a$ est la densité de la loi normale $N\left(\beta_j+\tau_j\tilde{\xi}_a,\sigma_j^2\right)$.

Récurrence "arrière"

Elle consiste à calculer soit

$$
B'_{a,j}(t)=P\left(Y_{a,t+1}^{T_a}=y_{a,t+1}^{T_a}|S_{a,t}=j,\tilde{\xi}_a\right)/P\left(Y_{a,t+1}^{T_a}=y_{a,t+1}^{T_a}|Y_{a,1}^{t}=y_{a,1}^{t},\tilde{\xi}_a\right),
$$

soit $L'_{a,j}(t) = P\left(S_{a,t} = j | Y_{a,1}^{T_a} = y_{a,1}^{T_a}, \tilde{\xi}_a\right)$ pour chaque état j, en reculant de T_a à 1.

La récurrence "arrière" est initialisée pour $t = T_a$ et $j = 1, ..., J$ par

$$L'_{a,j}(T_a) = P\left(S_{a,T_a} = j | Y_{a,1}^{T_a} = y_{a,1}^{T_a}, \tilde{\xi}_a\right) = F'_{a,j}(T_a).$$

Par conséquent, $B'_{a,j}(T_a) = 1$.

Pour $t = T_a - 1, ..., 1$ et $j = 1, ..., J$, nous avons la récurrence suivante :

$$\begin{aligned}
&B'_{a,j}(t) \\
&= \frac{P\left(Y_{a,t+1}^{T_a} = y_{a,t+1}^{T_a} | S_{a,t} = j, \tilde{\xi}_a\right)}{P\left(Y_{a,t+1}^{T_a} = y_{a,t+1}^{T_a} | Y_{a,1}^{t} = y_{a,1}^{t}, \tilde{\xi}_a\right)} \\
&= \frac{\sum_{k=1}^{J} P\left(Y_{a,t+2}^{T_a} = y_{a,t+2}^{T_a}, Y_{a,t+1} = y_{a,t+1}, S_{a,t+1} = k | S_{a,t} = j, \tilde{\xi}_a\right)}{P\left(Y_{a,t+2}^{T_a} = y_{a,t+2}^{T_a}, Y_{a,t+1} = y_{a,t+1} | Y_{a,1}^{t} = y_{a,1}^{t}, \tilde{\xi}_a\right)} \\
&= \frac{\sum_{k=1}^{J} P\left(Y_{a,t+2}^{T_a} = y_{a,t+2}^{T_a} | S_{a,t+1} = k, \tilde{\xi}_a\right) P\left(Y_{a,t+1} = y_{a,t+1} | S_{a,t+1} = k, \tilde{\xi}_a\right) P\left(S_{a,t+1} = k | S_{a,t} = j\right)}{P\left(Y_{a,t+2}^{T_a} = y_{a,t+2}^{T_a} | Y_{a,1}^{t+1} = y_{a,1}^{t+1}, \tilde{\xi}_a\right) P\left(Y_{a,t+1} = y_{a,t+1} | Y_{a,1}^{t} = y_{a,1}^{t}, \tilde{\xi}_a\right)} \\
&= \frac{1}{N'_{a,t+1}} \sum_{k=1}^{J} B'_{a,k}(t+1) f\left(y_{a,t+1} | S_{a,t+1} = k, \tilde{\xi}_a\right) p_{jk}.
\end{aligned}$$

La récurrence "arrière" est similaire à la récurrence "arrière" décrite au chapitre 2 si ce n'est qu'il y a en plus la présence de l'effet aléatoire dans le conditionnement des probabilités d'observation.

Sur le même principe que pour la récurrence "arrière" décrite au chapitre 2, nous déduisons de (3.16) que :

$$\begin{aligned}
L'_{a,j}(t) &= \frac{1}{N'_{a,t+1}} \left\{ \sum_{k=1}^{J} \frac{L'_{a,k}(t+1)}{F'_{a,k}(t+1)} f\left(y_{a,t+1} | S_{a,t+1} = k, \tilde{\xi}_a\right) p_{jk} \right\} F'_{a,j}(t) \\
&= \left\{ \sum_{k=1}^{J} \frac{L'_{a,k}(t+1)}{G'_{a,k}(t+1)} p_{jk} \right\} F'_{a,j}(t),
\end{aligned}$$

avec

$$\begin{aligned} G'_{a,k}(t+1) &= \frac{F'_{a,k}(t+1)N'_{a,t+1}}{f\left(y_{a,t+1}|S_{a,t+1}=k,\tilde{\xi}_a\right)} \\ &= P\left(S_{a,t+1}=k|Y^t_{a,1}=y^t_{a,1},\tilde{\xi}_a\right) \end{aligned}$$

Et, d'après (3.17) :

$$G'_{a,k}(t+1) = \sum_{j=1}^{J} p_{jk}F'_{a,j}(t),$$

cette quantité (la probabilité prédite) pouvant être extraite et stockée en mémoire lors de la récurrence "avant".

Seconde étape : maximisation

Cette étape consiste à maximiser l'espérance conditionnelle de la log-vraisemblance des données complètes qui vient d'être calculée à l'étape de restauration. Nous cherchons la valeur des paramètres qui maximise cette quantité.

Estimation des paramètres markoviens π_j et p_{ij}

L'étape de maximisation pour les paramètres relatifs à la chaîne de Markov est similaire à celle de l'étape M de l'algorithme EM, à la différence près que les probabilités filtrées, les probabilités lissées et les probabilités prédites sont en plus conditionnées par la valeur prédite de l'effet aléatoire $\tilde{\xi}_a$. Les formules de réestimation à l'itération k sont les suivantes :

$$\pi_j^{(k+1)} = L'^{(k)}_{a,j}(1) \qquad j=1,...,J,$$

$$p_{ij}^{(k+1)} = \frac{\sum_{t=1}^{T_a-1} L'^{(k)}_{a,j}(t+1)p_{ij}^{(k)}F'^{(k)}_{a,i}(t)/G'^{(k)}_{a,j}(t+1)}{\sum_{t=1}^{T_a-1} L'^{(k)}_{a,i}(t)} \qquad i,j=1,...,J,$$

où
$F'^{(k)}_{a,i}(t) = P\left(S_{a,t}=i|Y^t_{a,1}=y^t_{a,1},\tilde{\xi}_a;\theta^{(k)}\right)$ est la probabilité filtrée,
et $G'^{(k)}_{a,j}(t+1) = P\left(S_{a,t+1}=k|Y^t_{a,1}=y^t_{a,1},\tilde{\xi}_a;\theta^{(k)}\right)$ est la probabilité prédite.

Estimation des paramètres liés aux modèles linéaires mixtes β_j, τ_j et σ_j^2

L'annulation successive de la dérivée de (3.15) par rapport à β_j, τ_j et σ_j^2 conduit aux formules de réestimation suivantes :

$$\beta_j^{(k+1)} = \frac{\sum_{t=1}^{T_a} L'^{(k)}_{a,j}(t)\left(y_{a,t}-\tau_j^{(k)}\widetilde{\xi}_a^{(k)}\right)}{\sum_{t=1}^{T_a} L'^{(k)}_{a,j}(t)} \qquad j=1,...,J,$$

$$\tau_j^{(k+1)} = \frac{\sum_{t=1}^{T_a} L'^{(k)}_{a,j}(t) \left(y_{a,t} - \beta_j^{(k)}\right) \widetilde{\xi}_a^{(k)}}{\sum_{t=1}^{T_a} L'^{(k)}_{a,j}(t) \left(\widetilde{\xi}_a^{(k)}\right)^2} \qquad j = 1, ..., J,$$

$$\sigma_j^{2(k+1)} = \frac{\sum_{t=1}^{T_a} L'^{(k)}_{a,j}(t) \left(y_{a,t} - \beta_j^{(k)} - \tau_j^{(k)} \widetilde{\xi}_a^{(k)}\right)^2}{\sum_{t=1}^{T_a} L'^{(k)}_{a,j}(t)} \qquad j = 1, ..., J.$$

Troisième étape : prédiction

À l'itération k, une fois les étapes de restauration et de maximisation calculées, se pose la question de la réactualisation de la valeur prédite de l'effet aléatoire $\widetilde{\xi}_a$ servant à l'itération suivante. Il est donc nécessaire d'effectuer une étape de prédiction pour calculer la nouvelle valeur $\widetilde{\xi}_a^{(k+1)}$.

Nous avons

$$\begin{aligned} \widetilde{\xi}_a^{(k+1)} &= E\left(\xi_a | Y_{a,1}^{T_a} = y_{a,1}^{T_a}; \theta^{(k+1)}\right) \\ &= E\left(E\left(\xi_a | Y_{a,1}^{T_a} = y_{a,1}^{T_a}, S_{a,1}^{T_a} = s_{a,1}^{T_a}\right) | Y_{a,1}^{T_a} = y_{a,1}^{T_a}; \theta^{(k+1)}\right) \end{aligned}$$

D'après (3.10) :

$$\begin{aligned} \widetilde{\xi}_a^{(k+1)} &= E\left(\tau' U_a' \Gamma_a^{-1} \left(Y_a - X_a \beta\right) | Y_{a,1}^{T_a} = y_{a,1}^{T_a}; \theta^{(k+1)}\right) \\ &= \sum_{s_{a,1}^{T_a}} h\left(s_{a,1}^{T_a}\right) P\left(S_{a,1}^{T_a} = s_{a,1}^{T_a} | Y_{a,1}^{T_a} = y_{a,1}^{T_a}; \theta^{(k)}\right) \end{aligned}$$

où
· $\Gamma_a = U_a \tau \tau' U_a{}' + V_a$,
· $h\left(s_{a,1}^{T_a}\right)$ est la valeur de $\tau' U_a' \Gamma_a^{-1} \left(Y_a - X_a \beta\right)$ pour la séquence d'états $s_{a,1}^{T_a}$.

Mais comme nous l'avons vu à la section précédente, la probabilité conditionnelle $P\left(S_{a,1}^{T_a} = s_{a,1}^{T_a} | Y_{a,1}^{T_a} = y_{a,1}^{T_a}; \theta^{(k)}\right)$ ne se calcule pas. Par conséquent, le calcul de $\widetilde{\xi}_a^{(k+1)}$ pose problème.

Toutefois, nous pourrions envisager de simplifier le calcul de $\widetilde{\xi}_a^{(k+1)}$ en calculant simplement

$$\begin{aligned} \widetilde{\xi}_a^{(k+1)} &= E\left(\xi_a | Y_{a,1}^{T_a} = y_{a,1}^{T_a}, S_{a,1}^{T_a} = s_{a,1}^{T_a}; \theta^{(k+1)}\right) \\ &= \tau'^{(k+1)} U_a' \Gamma_a^{-1(k+1)} \left(Y_a - X_a \beta^{(k+1)}\right). \end{aligned}$$

Or, l'étape de restauration probabiliste détermine l'ensemble des séquences d'états possibles. Par conséquent, on ne connaît pas la séquence d'états $s_{a,1}^{T_a}$ et on ne peut donc pas déterminer les matrices U_a' et $\Gamma_a^{-1(k+1)}$.

Ainsi, envisager un algorithme de type EM sachant la valeur prédite de l'effet aléatoire permet l'écriture des étapes de restauration et de maximisation à la condition de savoir prédire à chaque itération une nouvelle valeur de l'effet aléatoire. Or cette étape pose problème du fait de la nature probabiliste de la restauration. Cet algorithme n'est donc pas possible pour le modèle 1.

Un effet aléatoire différent pour chaque état

Si $\theta^{(k)}$ est la valeur courante du paramètre et $\widetilde{\xi}_{a,1}^{J(k)}$ la valeur prédite courante du vecteur des effets aléatoires, l'itération k de l'algorithme est décrite par les trois étapes suivantes :

Étape de restauration : on restaure, de manière probabiliste avec un algorithme "avant-arrière" sachant le vecteur des effets aléatoires, l'ensemble des séquences d'états possibles. On calcule l'espérance conditionnelle de la log-vraisemblance des données complètes (3.13) sachant la séquence des observations, la valeur des paramètres à l'itération k, et la valeur prédite du vecteur des effets aléatoires à l'itération k :

$$E\left(\log f(y_{a,1}^{T_a}, s_{a,1}^{T_a}, \xi_{a,1}^{J}; \theta) | Y_{a,1}^{T_a} = y_{a,1}^{T_a}, \widetilde{\xi}_{a,1}^{J(k)}; \theta^{(k)}\right).$$

Étape de maximisation : on maximise par rapport à θ l'espérance conditionnelle calculée à l'étape de restauration pour réactualiser la valeur du paramètre. Autrement dit, on choisit $\theta^{(k+1)}$ de sorte que

$$\theta^{(k+1)} = \arg\max_{\theta\in\Theta} E\left(\log f(y_{a,1}^{T_a}, s_{a,1}^{T_a}, \xi_{a,1}^{J}; \theta) | Y_{a,1}^{T_a} = y_{a,1}^{T_a}, \widetilde{\xi}_{a,1}^{J(k)}; \theta^{(k)}\right).$$

Étape de prédiction : on calcule les valeurs prédites $\widetilde{\xi}_{a,j}^{(k+1)}, j = 1, ...J$ à partir de $\theta^{(k+1)}$, la valeur du paramètre calculée à l'étape de maximisation :

$$\widetilde{\xi}_{a,j}^{(k+1)} = E\left(\xi_{a,j} | Y_{a,1}^{T_a} = y_{a,1}^{T_a}; \theta^{(k+1)}\right).$$

Décrivons en détail chacune des trois étapes.

Première étape : restauration probabiliste

D'après l'expression de la log-vraisemblance des données complètes (3.13), nous écrivons :

$$
\begin{aligned}
&E\left(\log f\left(y_{a,1}^{T_a}, s_{a,1}^{T_a}, \xi_{a,1}^{J}; \theta\right) | Y_{a,1}^{T_a} = y_{a,1}^{T_a}, \widetilde{\xi}_{a,1}^{J(k)}; \theta^{(k)}\right) \\
&= \sum_{j=1}^{J} L_{a,j}^{\prime\prime(k)}(1) \log \pi_j + \sum_{i,j=1}^{J} \sum_{t=2}^{T_a} P\left(S_{a,t} = j, S_{a,t-1} = i | Y_{a,1}^{T_a} = y_{a,1}^{T_a}, \widetilde{\xi}_{a,1}^{J(k)}; \theta^{(k)}\right) \log p_{ij} \\
&- \frac{(T_a + J)}{2} \log 2\pi \\
&+ \sum_{j=1}^{J} \sum_{t=1}^{T_a} L_{a,j}^{\prime\prime(k)}(t) \left(-\log \sigma_j - \frac{1}{2\sigma_j^2}\left(y_{a,t} - \beta_j - \tilde{\xi}_{a,j}^{(k)}\right)^2\right) \\
&+ \sum_{j=1}^{J} \left(-\log \tau_j - \frac{1}{2\tau_j^2}\left(\tilde{\xi}_{a,j}^{(k)}\right)^2\right), \qquad (3.18)
\end{aligned}
$$

avec

$$
L_{a,j}^{\prime\prime(k)}(t) = P\left(S_{a,t} = j | Y_{a,1}^{T_a} = y_{a,1}^{T_a}, \widetilde{\xi}_{a,1}^{J(k)}; \theta^{(k)}\right).
$$

La probabilité $L_{a,j}''(t) = P\left(S_{a,t} = j | Y_{a,1}^{T_a} = y_{a,1}^{T_a}, \tilde{\xi}_{a,1}^{J}\right)$ se calcule avec un algorithme "avant-arrière" similaire à l'algorithme "avant-arrière" du chapitre 2 et à l'algorithme "avant-arrière" décrit dans le cadre du modèle 1.

Il est possible de décomposer $L_{a,j}''(t)$ de la manière suivante :

$$
\begin{aligned}
L_{a,j}''(t) &= P\left(S_{a,t} = j | Y_{a,1}^{T_a} = y_{a,1}^{T_a}, \tilde{\xi}_{a,1}^{J}\right) \\
&= \frac{P\left(Y_{a,t+1}^{T_a} = y_{a,t+1}^{T_a} | S_{a,t} = j, \tilde{\xi}_{a,1}^{J}\right)}{P\left(Y_{a,t+1}^{T_a} = y_{a,t+1}^{T_a} | Y_{a,1}^{t} = y_{a,1}^{t}, \tilde{\xi}_{a,1}^{J}\right)} P\left(S_{a,t} = j | Y_{a,1}^{t} = y_{a,1}^{t}, \tilde{\xi}_{a,1}^{J}\right) \\
&= B_{a,j}''(t) F_{a,j}''(t), \qquad (3.19)
\end{aligned}
$$

avec

$$
B_{a,j}''(t) = \frac{P\left(Y_{a,t+1}^{T_a} = y_{a,t+1}^{T_a} | S_{a,t} = j, \tilde{\xi}_{a,1}^{J}\right)}{P\left(Y_{a,t+1}^{T_a} = y_{a,t+1}^{T_a} | Y_{a,1}^{t} = y_{a,1}^{t}, \tilde{\xi}_{a,1}^{J}\right)},
$$

et

$$
F_{a,j}''(t) = P\left(S_{a,t} = j | Y_{a,1}^{t} = y_{a,1}^{t}, \tilde{\xi}_{a,1}^{J}\right).
$$

Les quantités $F_{a,j}''(t)$ sont calculées dans la passe "avant", alors que les quantités $B_{a,j}''(t)$ ou les quantités $L_{a,j}''(t)$ sont calculées dans la passe "arrière".

Récurrence "avant"
Elle est initialisée pour $t = 1$ et $j = 1, ..., J$ par

$$\begin{aligned} F''_{a,j}(1) &= P\left(S_{a,1} = j | Y_{a,1} = y_{a,1}, \tilde{\xi}^J_{a,1}\right) \\ &= \frac{P\left(Y_{a,1} = y_{a,1} | S_{a,1} = j, \tilde{\xi}^J_{a,1}\right) P\left(\tilde{\xi}^J_{a,1}\right) P\left(S_{a,1} = j\right)}{P\left(Y_{a,1} = y_{a,1}, \tilde{\xi}^J_{a,1}\right)} \\ &= \frac{P\left(Y_{a,1} = y_{a,1} | S_{a,1} = j, \tilde{\xi}_{a,j}\right) P\left(S_{a,1} = j\right) P\left(\tilde{\xi}^J_{a,1}\right)}{P\left(Y_{a,1} = y_{a,1}, \tilde{\xi}^J_{a,1}\right)} \\ &= \frac{f\left(y_{a,1} | S_{a,1} = j, \tilde{\xi}_{a,j}\right) \pi_j \prod_{j=1}^J f\left(\tilde{\xi}_{a,j}\right)}{N''_{a,1}} \end{aligned}$$

où
· $f(y_{a,1}|S_{a,1} = j, \tilde{\xi}_{a,j})$, densité de l'observation $y_{a,1}|S_{a,1} = j, \tilde{\xi}_{a,j}$ est la densité de la loi normale $N(\beta_j + \tilde{\xi}_{a,j}, \sigma^2_j)$,
· $f\left(\tilde{\xi}_{a,j}\right)$, densité de l'effet aléatoire relatif à l'individu a se trouvant dans l'état j est la densité de la loi normale $N\left(0, \tau^2_j\right)$,
· $N''_{a,1} = P\left(Y_{a,1} = y_{a,1}, \tilde{\xi}^J_{a,1}\right)$, le facteur de normalisation est égal à :

$$\begin{aligned} N''_{a,1} &= \sum_{j=1}^J P\left(Y_{a,1} = y_{a,1} | S_{a,1} = j, \tilde{\xi}^J_{a,1}\right) P\left(\tilde{\xi}^J_{a,1}\right) P\left(S_{a,1} = j\right) \\ &= \left\{\sum_{j=1}^J f\left(y_{a,1} | S_{a,1} = j, \tilde{\xi}_{a,j}\right) \pi_j\right\} \prod_{j=1}^J f\left(\tilde{\xi}_{a,j}\right). \end{aligned}$$

Sur le même principe que pour le modèle 1, pour $t = 2, ..., T_a$ et $j = 1, ..., J$, la récurrence "avant" s'écrit :

$$\begin{aligned} F''_{a,j}(t) &= P\left(S_{a,t} = j | Y^t_{a,1} = y^t_{a,1}, \tilde{\xi}^J_{a,1}\right) \\ &= \frac{1}{N''_{a,t}} f\left(y_{a,t} | S_{a,t} = j, \tilde{\xi}_{a,j}\right) \sum_{i=1}^J p_{ij} F''_{a,i}(t-1), \end{aligned} \tag{3.20}$$

où $N''_{a,t} = P\left(Y_{a,t} = y_{a,t} | Y^{t-1}_{a,1} = y^{t-1}_{a,1}, \tilde{\xi}^J_{a,1}\right)$, le facteur de normalisation est égal à :

$$N''_{a,t} = \sum_{j=1}^{J} P\left(S_{a,t} = j, Y_{a,t} = y_{a,t} | Y_{a,1}^{t-1} = y_{a,1}^{t-1}, \tilde{\xi}_{a,1}^{J}\right)$$
$$= \sum_{j=1}^{J} f\left(y_{a,t} | S_{a,t} = j, \tilde{\xi}_{a,j}\right) \sum_{i=1}^{J} p_{ij} F''_{a,i}(t-1),$$

avec $f\left(y_{a,t} | S_{a,t} = j, \tilde{\xi}_{a,j}\right)$ densité de la loi normale $N\left(\beta_j + \tilde{\xi}_{a,j}, \sigma_j^2\right)$.

Cette récurrence "avant" s'implémente de la même manière que dans le cadre du modèle 1. Elle est similaire à celle du modèle 1, si ce n'est qu'à l'initialisation la densité de l'effet aléatoire ξ_a est remplacée par le produit des J densités des effets aléatoires $\xi_{a,j}$, et que, dans le conditionnemment des probabilités d'observation, la valeur prédite de l'unique effet aléatoire $\tilde{\xi}_a$ est remplacée par $\tilde{\xi}_{a,j}$, la valeur prédite de l'effet aléatoire relatif à l'état j.

Récurrence "arrière"
Elle s'effectue sur le même principe que pour le modèle 1.
La récurrence "arrière" est initialisée pour $t = T_a$ et $j = 1, ..., J$ par

$$L''_{a,j}(T_a) = P\left(S_{a,T_a} = j | Y_{a,1}^{T_a} = y_{a,1}^{T_a}, \tilde{\xi}_{a,1}^{J}\right) = F''_{a,j}(T_a).$$

Par conséquent, $B''_{a,j}(T_a) = 1$.

Pour $t = T_a - 1, ..., 1$ et $j = 1, ..., J$, la récurrence "arrière" s'écrit :

$$B''_{a,j}(t) = \frac{P\left(Y_{a,t+1}^{T_a} = y_{a,t+1}^{T_a} | S_{a,t} = j, \tilde{\xi}_{a,1}^{J}\right)}{P\left(Y_{a,t+1}^{T_a} = y_{a,t+1}^{T_a} | Y_{a,1}^{t} = y_{a,1}^{t}, \tilde{\xi}_{a,1}^{J}\right)}$$
$$= \frac{1}{N''_{a,t+1}} \sum_{k=1}^{J} B''_{a,k}(t+1) f\left(y_{a,t+1} | S_{a,t+1} = k, \tilde{\xi}_{a,k}\right) p_{jk}.$$

Cette récurrence "arrière" est semblable à celle du modèle 1, à la différence près que dans le conditionnement des probabilités d'observation, la valeur prédite de l'unique effet aléatoire $\tilde{\xi}_a$ est remplacée par $\tilde{\xi}_{a,j}$, la valeur prédite de l'effet aléatoire relatif à l'état j.

Seconde étape : maximisation

On cherche la valeur des paramètres qui maximise l'espérance conditionnelle de la log-vraisemblance des données complètes qui vient d'être calculée à l'étape de restauration.

Estimation des paramètres markoviens π_j et p_{ij}

L'estimation des paramètres relatifs à la chaîne de Markov est similaire à celle décrite pour le modèle 1, à la différence près que les probabilités filtrées, les probabilités lissées et les probabilités prédites sont conditionnées par la valeur prédite du vecteur des effets aléatoires $\tilde{\xi}_{a,1}^{J}$. Les formules de réestimation à l'itération k sont les suivantes :

$$\pi_j^{(k+1)} = L_{a,j}^{\prime\prime(k)}(1) \qquad j = 1, ..., J,$$

$$p_{ij}^{(k+1)} = \frac{\sum_{t=1}^{T_a-1} L_{a,j}^{\prime\prime(k)}(t+1) p_{ij}^{(k)} F_{a,i}^{\prime\prime(k)}(t) / G_{a,j}^{\prime\prime(k)}(t+1)}{\sum_{t=1}^{T_a-1} L_{a,i}^{\prime\prime(k)}(t)} \qquad i, j = 1, ..., J,$$

où
$F_{a,i}^{\prime\prime(k)}(t) = P\left(S_{a,t} = i | Y_{a,1}^{t} = y_{a,1}^{t}, \tilde{\xi}_{a,1}^{J}; \theta^{(k)}\right)$ est la probabilité filtrée,
et $G_{a,j}^{\prime\prime(k)}(t+1) = P\left(S_{a,t+1} = k | Y_{a,1}^{t} = y_{a,1}^{t}, \tilde{\xi}_{a,1}^{J}; \theta^{(k)}\right)$ est la probabilité prédite.

Estimation des paramètres liés aux modèles linéaires mixtes β_j, τ_j^2 et σ_j^2
L'annulation successive de la dérivée de (3.18) par rapport à β_j, τ_j^2 et σ_j^2 conduit aux formules de réestimation suivantes :

$$\beta_j^{(k+1)} = \frac{\sum_{t=1}^{T_a} L_{a,j}^{\prime\prime(k)}(t) \left(y_{a,t} - \tilde{\xi}_{a,j}^{(k)}\right)}{\sum_{t=1}^{T_a} L_{a,j}^{\prime\prime(k)}(t)} \qquad j = 1, ..., J,$$

$$\tau_j^{2(k+1)} = \left(\tilde{\xi}_{a,j}^{(k)}\right)^2 \qquad j = 1, ..., J,$$

$$\sigma_j^{2(k+1)} = \frac{\sum_{t=1}^{T_a} L_{a,j}^{\prime\prime(k)}(t) \left(y_{a,t} - \beta_j^{(k)} - \tilde{\xi}_{a,j}^{(k)}\right)^2}{\sum_{t=1}^{T_a} L_{a,j}^{\prime\prime(k)}(t)} \qquad j = 1, ..., J.$$

Troisième étape : prédiction

Tout comme pour le modèle 1, envisager un algorithme de type EM sachant la valeur prédite du vecteur des effets aléatoires permet l'écriture des étapes de restauration et de maximisation à condition de savoir prédire à chaque itération une nouvelle valeur du vecteur des effets aléatoires. Cette étape pose problème car l'étape de restauration est probabiliste.

En effet, si nous voulons prédire les nouvelles valeurs $\tilde{\xi}_{a,j}^{(k+1)}, j = 1, ...J$, nous écrivons

$$\begin{aligned}
\tilde{\xi}_{a,j}^{(k+1)} &= E\left(\xi_{a,j} | Y_{a,1}^{T_a} = y_{a,1}^{T_a}; \theta^{(k+1)}\right) \\
&= E\left(E\left(\xi_{a,j} | Y_{a,1}^{T_a} = y_{a,1}^{T_a}, S_{a,1}^{T_a} = s_{a,1}^{T_a}\right) | Y_{a,1}^{T_a} = y_{a,1}^{T_a}; \theta^{(k+1)}\right)
\end{aligned}$$

D'après (3.14) :

$$\begin{aligned}\overset{\sim}{\xi}_{a,j}^{(k+1)} &= E\left(C_j' U_a' \Gamma_a^{-1}\left(Y_a - X_a\beta\right)|Y_{a,1}^{T_a} = y_{a,1}^{T_a}; \theta^{(k+1)}\right) \\ &= \sum_{s_{a,1}^{T_a}} g\left(s_{a,1}^{T_a}\right) P\left(S_{a,1}^{T_a} = s_{a,1}^{T_a}|Y_{a,1}^{T_a} = y_{a,1}^{T_a}; \theta^{(k)}\right)\end{aligned}$$

où
· $\Gamma_a = U_a D U_a' + V_a$,
· $C_j = \begin{pmatrix} 0 \\ \vdots \\ \tau_j^2 \\ 0 \\ \vdots \end{pmatrix}$ est le vecteur de dimension $J \times 1$, composé de $J-1$ zéros et ayant la valeur τ_j^2 sur sa $j^{ème}$ ligne,
· $g\left(s_{a,1}^{T_a}\right)$ est la valeur de $C_j' U_a' \Gamma_a^{-1}\left(Y_a - X_a\beta\right)$ pour la séquence d'états $s_{a,1}^{T_a}$.

La probabilité conditionnelle$P\left(S_{a,1}^{T_a} = s_{a,1}^{T_a}|Y_{a,1}^{T_a} = y_{a,1}^{T_a}; \theta^{(k)}\right)$ ne se calculant pas, $\overset{\sim}{\xi}_{a,j}^{(k+1)}$ ne se calcule pas non plus.
Toutefois, comme pour le modèle 1, nous pourrions envisager de simplifier le calcul de $\overset{\sim}{\xi}_{a,j}^{(k+1)}$ en calculant simplement

$$\begin{aligned}\overset{\sim}{\xi}_{a,j}^{(k+1)} &= E\left(\xi_{a,j}|Y_{a,1}^{T_a} = y_{a,1}^{T_a}, S_{a,1}^{T_a} = s_{a,1}^{T_a}; \theta^{(k+1)}\right) \\ &= C_j'^{(k+1)} U_a' \Gamma_a^{-1(k+1)}\left(Y_a - X_a\beta^{(k+1)}\right).\end{aligned}$$

L'étape de restauration étant probabiliste, on détermine l'ensemble des séquences d'états possibles, mais on ne connaît pas la séquence d'états $s_{a,1}^{T_a}$ et on ne peut donc pas déterminer les matrices U_a' et $\Gamma_a^{-1(k+1)}$.

Par conséquent, cet algorithme de type EM sachant le vecteur des effets aléatoires n'est pas, non plus, possible pour le modèle 2. Notons toutefois que certains calculs, comme ceux des algorithmes "avant-arrière" seront utilisés par la suite, dans les méthodes d'estimation proposées.

3.3.4 Méthodes d'estimation proposées

L'algorithme EM s'écrit pour les estimations d'un modèle linéaire mixte (section 2.4.2) et d'un modèle linéaire multiphasique (section 3.3.1), mais comme nous venons de le voir dans la section précédente, il ne se transpose pas pour l'estimation d'un modèle linéaire mixte multiphasique du fait de la combinaison des deux structures cachées (les effets aléatoires et les états). Pour contourner les difficultés liées à l'étape E de l'algorithme

EM et à la prédiction des effets aléatoires pour un algorithme de type EM sachant les effets aléatoires, nous proposons un algorithme itératif en trois étapes : restauration, maximisation et prédiction. Deux types de restauration sont envisagés : une restauration déterministe et une restauration par simulation.

Algorithme itératif avec restauration déterministe

L'algorithme itératif avec restauration déterministe proposé est inspiré de l'algorithme de Baum-Viterbi (Jelinek, 1976 ; Ephraim et Merhav, 2002), décrit à la section 2.4.3. En effet, l'algorithme de Baum-Viterbi est particulièrement approprié, lorsque le modèle possède une structure "gauche-droite" et que la log-vraisemblance de la séquence d'états optimale associée à la séquence observée représente une part importante de la log-vraisemblance de l'ensemble des séquences d'états possibles associées à la séquence observée. Cette propriété est souvent vérifiée pour les modèles "gauche-droite" utilisés pour les applications à la croissance des plantes. De plus, comme nous l'avons montré au chapitre 2, l'algorithme de Baum-Viterbi possède la "bonne" propriété de croissance monotone de la log-vraisemblance des séquences d'états optimales associées aux séquences observées (Juang et Rabiner, 1990).

Pour rappel, l'algorithme de Baum-Viterbi est une procédure d'estimation pour les paramètres d'un modèle de Markov caché. Dans une première étape, la séquence d'états globalement optimale est restaurée avec l'algorithme de Viterbi (Forney, 1973), et dans une seconde étape, les paramètres sont estimés par maximisation de la probabilité jointe de la séquence observée et de la séquence d'états globalement optimale qui a été restaurée. Comme nous allons l'expliquer, il est nécessaire de prendre en compte les effets aléatoires dans le calcul de la séquence d'états restaurée et dans la maximisation de la loi jointe de la séquence observée et de la séquence d'états restaurée.

Si $\theta^{(k)}$ est la valeur courante du paramètre et $\tilde{\xi}_a^{(k)}$ la valeur prédite courante de l'effet aléatoire, l'itération k de l'algorithme proposé est décrite par les trois étapes suivantes :

Modèle 1
Étape de restauration : on calcule avec l'algorithme de Viterbi $\tilde{s}_{a,1}^{T_a(k)} = \left(\tilde{s}_{a,1}^{(k)}, ..., \tilde{s}_{a,T_a}^{(k)}\right)$, la séquence d'états optimale associée à la séquence observée $y_{a,1}^{T_a}$ et à l'effet aléatoire prédit $\tilde{\xi}_a^{(k)}$. La restauration est basée sur la décomposition de

$$\max_{s_{a,1},...,s_{a,T_a}} P\left(S_{a,1}^{T_a} = s_{a,1}^{T_a}, Y_{a,1}^{T_a} = y_{a,1}^{T_a}, \tilde{\xi}_a^{(k)}; \theta^{(k)}\right).$$

Étape de maximisation : on maximise par rapport à θ la log-vraisemblance des données complètes (la séquence observée, complétée par la séquence d'états restaurée et l'effet aléatoire prédit) pour réactualiser la valeur du paramètre. Autrement dit, on choisit $\theta^{(k+1)}$ de sorte que

$$\theta^{(k+1)} = \arg\max_{\theta\in\Theta} \log f\left(\tilde{s}_{a,1}^{T_a(k)}, y_{a,1}^{T_a}, \tilde{\xi}_a^{(k)}; \theta\right).$$

Étape de prédiction : on calcule la valeur prédite $\widetilde{\xi}_a^{(k+1)}$ à partir de $\tilde{s}_{a,1}^{T_a(k)}$, la séquence d'états restaurée et à partir de $\theta^{(k+1)}$, la valeur du paramètre calculée à l'étape de maximisation :

$$\widetilde{\xi}_a^{(k+1)} = E\left(\xi_a | Y_{a,1}^{T_a} = y_{a,1}^{T_a}, S_{a,1}^{T_a} = \tilde{s}_{a,1}^{T_a(k)}; \theta^{(k+1)}\right).$$

Modèle 2

La description de l'algorithme est similaire à celle du modèle 1 si ce n'est que la valeur $\tilde{\xi}_a$ est remplacée par les valeurs $\tilde{\xi}_{a,j}, j = 1, ..., J$.

Nous allons à présent décrire en détail chacune des étapes de l'algorithme proposé pour les modèles 1 et 2.

Première étape : restauration déterministe

L'algorithme de Viterbi a été présenté à la section 2.4.3 dans le cadre d'une chaîne de Markov cachée où les observations sont conditionnellement indépendantes sachant les états. Toutefois cette présentation n'est plus valable telle quelle dans le cadre des modèles 1 et 2. En effet, pour le modèle 1, les observations sont conditionnellement indépendantes sachant l'effet aléatoire et la séquence d'états, et, pour le modèle 2, les observations sont conditionnellement indépendantes sachant le vecteur des effets aléatoires et la séquence d'états. Par conséquent, selon le même principe que pour le calcul des probabilités lissées $L'_{a,j}(t)$ ou $L''_{a,j}(t)$ où les effets aléatoires ont été intégrés à un algorithme "avant-arrière", nous intégrons ici la présence des effets aléatoires à l'algorithme de Viterbi. Nous déterminons la séquence d'états globalement optimale associée à la séquence observée et à l'effet aléatoire prédit $\tilde{\xi}_a$ pour le modèle 1, et nous déterminons la séquence d'états globalement optimale associée à la séquence observée et au vecteur des effets aléatoires prédits $\tilde{\xi}_{a,1}^{J}$ pour le modèle 2.

Modèle 1

On a la décomposition suivante :

$$\begin{aligned} & \max_{s_{a,1},...,s_{a,Ta}} P\left(S_{a,1}^{T_a} = s_{a,1}^{T_a}, Y_{a,1}^{T_a} = y_{a,1}^{T_a}, \tilde{\xi}_a\right) \\ = & \max_j \left\{ \max_{s_{a,t+1},...,s_{a,Ta}} P\left(Y_{a,t+1}^{T_a} = y_{a,t+1}^{T_a}, S_{a,t+1}^{T_a} = s_{a,t+1}^{T_a} | S_{a,t} = j, \tilde{\xi}_a\right) \alpha'_j(t) \right\}, \end{aligned}$$

avec

$$\alpha'_j(t) = \max_{s_{a,1},...,s_{a,t-1}} P\left(S_{a,t} = j, S_{a,1}^{t-1} = s_{a,1}^{t-1}, Y_{a,1}^{t} = y_{a,1}^{t}, \tilde{\xi}_a\right).$$

L'algorithme est initialisé pour $t = 1$ et $j = 1, ..., J$ par

$$\begin{aligned}\alpha'_j(1) &= P\left(S_{a,1}=j, Y_{a,1}=y_{a,1}, \tilde{\xi}_a\right)\\ &= P\left(Y_{a,1}=y_{a,1}|S_{a,1}=j, \tilde{\xi}_a\right) f\left(\tilde{\xi}_a\right) P\left(S_{a,1}=j\right)\\ &= f\left(y_{a,1}|S_{a,1}=j, \tilde{\xi}_a\right) f\left(\tilde{\xi}_a\right) \pi_j.\end{aligned}$$

L'équation de programmation dynamique s'écrit pour $t = 2, ..., T_a$ et $j = 1, ..., J$:

$$\begin{aligned}\alpha'_j(t) &= \max_{s_{a,1},\ldots,s_{a,t-1}} P\left(S_{a,t}=j, S_{a,1}^{t-1}=s_{a,1}^{t-1}, Y_{a,1}^{t}=y_{a,1}^{t}, \tilde{\xi}_a\right)\\ &= P\left(Y_{a,t}=y_{a,t}|S_{a,t}=j, \tilde{\xi}_a\right) \max_i \left\{P\left(S_{a,t}=j|S_{a,t-1}=i\right)\right.\\ &\quad \left.\times \max_{s_{a,1},\ldots,s_{a,t-2}} P\left(S_{a,t-1}=i, S_{a,1}^{t-2}=s_{a,1}^{t-2}, Y_{a,1}^{t-1}=y_{a,1}^{t-1}, \tilde{\xi}_a\right)\right\}\\ &= f\left(y_{a,t}|S_{a,t}=j, \tilde{\xi}_a\right) \max_i \left\{p_{ij}\alpha'_i(t-1)\right\}.\end{aligned}$$

Cet algorithme de Viterbi avec effet aléatoire est semblable à l'algorithme de Viterbi "classique", si ce n'est qu'il y a, en plus, la présence de l'effet aléatoire dans le conditionnement des probabilités d'observation et que l'on injecte $f\left(\tilde{\xi}_a\right)$ lors de l'initialisation. Cet algorithme est similaire à la récurrence "avant" avec effet aléatoire décrite à la section 3.3.3 pour le modèle 1, à condition d'oublier l'étape de normalisation et de remplacer les sommations par des maximisations.

Remarquons que la vraisemblance de la séquence d'états globalement optimale associée à la séquence observée et à l'effet aléatoire prédit est donnée par

$$\begin{aligned}&\max_{s_{a,1},\ldots,s_{a,T_a}} P\left(S_{a,1}^{T_a}=s_{a,1}^{T_a}, Y_{a,1}^{T_a}=y_{a,1}^{T_a}, \tilde{\xi}_a\right)\\ =\ &\max_j \left\{\max_{s_{a,1},\ldots,s_{a,T_a-1}} P\left(S_{a,T_a}=j, S_{a,1}^{T_a-1}=s_{a,1}^{T_a-1}, Y_{a,1}^{T_a}=y_{a,1}^{T_a}, \tilde{\xi}_a\right)\right\}\\ =\ &\max_j \left\{\alpha'_j(T_a)\right\}.\end{aligned}$$

Modèle 2

On a la décomposition suivante :

$$\begin{aligned} & \max_{s_{a,1},\ldots,s_{a,T_a}} P\left(S_{a,1}^{T_a}=s_{a,1}^{T_a}, Y_{a,1}^{T_a}=y_{a,1}^{T_a}, \tilde{\xi}_{a,1}^{J}\right) \\ = & \max_{j}\left\{\max_{s_{a,t+1},\ldots,s_{a,T_a}} P\left(Y_{a,t+1}^{T_a}=y_{a,t+1}^{T_a}, S_{a,t+1}^{T_a}=s_{a,t+1}^{T_a}|S_{a,t}=j,\tilde{\xi}_{a,1}^{J}\right)\alpha_j''(t)\right\}, \end{aligned}$$

avec

$$\alpha_j''(t)=\max_{s_{a,1},\ldots,s_{a,t-1}} P\left(S_{a,t}=j, S_{a,1}^{t-1}=s_{a,1}^{t-1}, Y_{a,1}^{t}=Y_{a,1}^{t}, \tilde{\xi}_{a,1}^{J}\right).$$

L'algorithme est initialisé pour $t=1$ et $j=1,...,J$ par

$$\begin{aligned} \alpha_j''(1) &= P\left(S_{a,1}=j, Y_{a,1}=y_{a,1}, \tilde{\xi}_{a,1}^{J}\right) \\ &= P\left(Y_{a,1}=y_{a,1}|S_{a,1}=j,\tilde{\xi}_{a,1}^{J}\right)P\left(\tilde{\xi}_{a,1}^{J}\right)P\left(S_{a,1}=j\right) \\ &= f\left(y_{a,1}|S_{a,1}=j,\tilde{\xi}_{a,j}\right)\left\{\prod_{j=1}^{J} f\left(\tilde{\xi}_{a,j}\right)\right\}\pi_j. \end{aligned}$$

L'équation de programmation dynamique s'écrit pour $t=2,...,T_a$ et $j=1,...,J$:

$$\begin{aligned} \alpha_j''(t) &= \max_{s_{a,1},\ldots,s_{a,t-1}} P\left(S_{a,t}=j, S_{a,1}^{t-1}=s_{a,1}^{t-1}, Y_{a,1}^{t}=y_{a,1}^{t}, \tilde{\xi}_{a,1}^{J}\right) \\ &= P\left(Y_{a,t}=y_{a,t}|S_{a,t}=j,\tilde{\xi}_{a,j}\right)\max_i\left\{P\left(S_{a,t}=j|S_{a,t-1}=i\right)\right. \\ &\quad \left.\times\max_{s_{a,1},\ldots,s_{a,t-2}} P\left(S_{a,t-1}=i, S_{a,1}^{t-2}=s_{a,1}^{t-2}, Y_{a,1}^{t-1}=y_{a,1}^{t-1}, \tilde{\xi}_{a,1}^{J}\right)\right\} \\ &= f\left(y_{a,t}|S_{a,t}=j,\tilde{\xi}_{a,j}\right)\max_i\left\{p_{ij}\alpha_i''(t-1)\right\}. \end{aligned}$$

Cet algorithme est similaire à celui du modèle 1, si ce n'est qu'à l'initialisation, la densité de l'effet aléatoire ξ_a est remplacée par le produit des J densités des effets aléatoires $\xi_{a,j}$, et que, dans le conditionnement des probabilités d'observation, l'unique effet aléatoire prédit $\tilde{\xi}_a$ est remplacé par $\tilde{\xi}_{a,j}$, l'effet aléatoire prédit relatif à l'état j.

Sur le même principe que pour le modèle 1, la vraisemblance de la séquence d'états globalement optimale, associée à la séquence observée et au vecteur des effets aléatoires prédits, est donnée par

$$\begin{aligned}
&\max_{s_{a,1},\ldots,s_{a,T_a}} P\left(S_{a,1}^{T_a}=s_{a,1}^{T_a}, Y_{a,1}^{T_a}=y_{a,1}^{T_a}, \tilde{\xi}_{a,1}^{J}\right)\\
=\;& \max_{j}\left\{\max_{s_{a,1},\ldots,s_{a,T_a-1}} P\left(S_{a,T_a}=j, S_{a,1}^{T_a-1}=s_{a,1}^{T_a-1}, Y_{a,1}^{T_a}=y_{a,1}^{T_a}, \tilde{\xi}_{a,1}^{J}\right)\right\}\\
=\;& \max_{j}\left\{\alpha_j''(T_a)\right\}.
\end{aligned}$$

Seconde étape : maximisation

Pour effectuer une maximisation au sens de l'algorithme de Baum-Viterbi, il faudrait maximiser par rapport à chacun des paramètres à estimer la quantité $P\left(S_{a,1}^{T_a}=\widetilde{s}_{a,1}^{T_a}, Y_{a,1}^{T_a}=y_{a,1}^{T_a};\theta\right)$ où $\widetilde{s}_{a,1}^{T_a}$ désigne la séquence d'états optimale restaurée à l'étape précédente. Cette quantité ne pouvant pas se calculer dans le cadre des modèles 1 et 2, on maximise $P\left(S_{a,1}^{T_a}=\widetilde{s}_{a,1}^{T_a}, Y_{a,1}^{T_a}=y_{a,1}^{T_a}, \tilde{\xi}_a;\theta\right)$ pour le modèle 1 et $P\left(S_{a,1}^{T_a}=\widetilde{s}_{a,1}^{T_a}, Y_{a,1}^{T_a}=y_{a,1}^{T_a}, \tilde{\xi}_{a,1}^{J};\theta\right)$ pour le modèle 2.

Modèle 1

La vraisemblance de la séquence observée, associée à la séquence d'états restaurée et à l'effet aléatoire prédit, s'écrit sous la forme :

$$f\left(y_{a,1}^{T_a}, \widetilde{s}_{a,1}^{T_a}, \tilde{\xi}_a;\theta\right) = f\left(y_{a,1}^{T_a}|\tilde{\xi}_a, \widetilde{s}_{a,1}^{T_a}\right) f\left(\tilde{\xi}_a\right) P\left(S_{a,1}^{T_a}=\widetilde{s}_{a,1}^{T_a}\right),$$

ou encore

$$\begin{aligned}
f(y_{a,1}^{T_a}, \widetilde{s}_{a,1}^{T_a}, \tilde{\xi}_a;\theta) \;=\;& \left\{\prod_{t=1}^{T_a}\frac{1}{\sqrt{2\pi}\sigma_{\widetilde{s}_{a,t}}}\exp\left(-\frac{\left(y_{a,t}-\beta_{\widetilde{s}_{a,t}}-\tau_{\widetilde{s}_{a,t}}\tilde{\xi}_a\right)^2}{2\sigma^2_{\widetilde{s}_{a,t}}}\right)\right\}\\
&\times\left\{\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{\tilde{\xi}_a^2}{2}\right)\right\}\pi_{\widetilde{s}_{a,1}}\prod_{t=2}^{T_a}p_{\widetilde{s}_{a,t-1}\widetilde{s}_{a,t}}.
\end{aligned}$$

Il s'ensuit que la log-vraisemblance de la séquence observée, associée à la séquence d'états restaurée et à l'effet aléatoire prédit, est donnée par :

$$\begin{aligned}
\log f(y_{a,1}^{T_a}, \widetilde{s}_{a,1}^{T_a}, \tilde{\xi}_a;\theta) \;=\;& \log\pi_{\widetilde{s}_{a,1}}+\sum_{t=2}^{T_a}\log p_{\widetilde{s}_{a,t-1}\widetilde{s}_{a,t}}-\frac{(T_a+1)}{2}\log 2\pi-\frac{\tilde{\xi}_a^2}{2}\\
&+\sum_{t=1}^{T_a}\left(-\log\sigma_{\widetilde{s}_{a,t}}-\frac{\left(y_{a,t}-\beta_{\widetilde{s}_{a,t}}-\tau_{\widetilde{s}_{a,t}}\tilde{\xi}_a\right)^2}{2\sigma^2_{\widetilde{s}_{a,t}}}\right),
\end{aligned}$$

ce qui peut se réécrire :

$$\log f(y_{a,1}^{T_a}, \widetilde{s}_{a,1}^{T_a}, \tilde{\xi}_a; \theta) = \sum_{j=1}^{J} I(\widetilde{s}_{a,1} = j) \log \pi_j + \sum_{i,j=1}^{J} \sum_{t=2}^{T_a} I\left(\widetilde{s}_{a,t-1} = i, \widetilde{s}_{a,t} = j\right) \log p_{ij}$$
$$- \frac{(T_a + 1)}{2} \log 2\pi - \frac{\tilde{\xi}_a^2}{2}$$
$$+ \sum_{j=1}^{J} \sum_{t=1}^{T_a} I\left(\widetilde{s}_{a,t} = j\right) \left(- \log \sigma_j - \frac{\left(y_{a,t} - \beta_j - \tau_j \tilde{\xi}_a\right)^2}{2\sigma_j^2} \right). \quad (3.21)$$

Selon les mêmes principes de maximisation que pour l'étape M de l'algorithme EM, les formules de réestimation pour les paramètres markoviens sont données à l'itération k par :

$$\pi_j^{(k+1)} = I(\widetilde{s}_{a,1}^{(k)} = j) \qquad j = 1, ..., J,$$

$$p_{ij}^{(k+1)} = \frac{\sum_{t=1}^{T_a - 1} I\left(\widetilde{s}_{a,t}^{(k)} = i, \widetilde{s}_{a,t+1}^{(k)} = j\right)}{\sum_{t=1}^{T_a - 1} I\left(\widetilde{s}_{a,t}^{(k)} = i\right)} \qquad i, j = 1, ..., J.$$

De plus, l'annulation successive de la dérivée de (3.21) par rapport à β_j, τ_j et σ_j^2 conduit aux formules de réestimation suivantes :

$$\beta_j^{(k+1)} = \frac{\sum_{t=1}^{T_a} I\left(\widetilde{s}_{a,t}^{(k)} = j\right) \left(y_{a,t} - \tau_j^{(k)} \widetilde{\xi}_a^{(k)}\right)}{\sum_{t=1}^{T_a} I\left(\widetilde{s}_{a,t}^{(k)} = j\right)} \qquad j = 1, ..., J,$$

$$\tau_j^{(k+1)} = \frac{\sum_{t=1}^{T_a} I\left(\widetilde{s}_{a,t}^{(k)} = j\right) \left(y_{a,t} - \beta_j^{(k)}\right) \widetilde{\xi}_a^{(k)}}{\sum_{t=1}^{T_a} I\left(\widetilde{s}_{a,t}^{(k)} = j\right) \left(\widetilde{\xi}_a^{(k)}\right)^2} \qquad j = 1, ..., J,$$

$$\sigma_j^{2(k+1)} = \frac{\sum_{t=1}^{T_a} I\left(\widetilde{s}_{a,t}^{(k)} = j\right) \left(y_{a,t} - \beta_j^{(k)} - \tau_j^{(k)} \widetilde{\xi}_a^{(k)}\right)^2}{\sum_{t=1}^{T_a} I\left(\widetilde{s}_{a,t}^{(k)} = j\right)} \qquad j = 1, ..., J.$$

Ces formules sont similaires à celles obtenues à l'étape de maximisation de l'algorithme itératif avec restauration probabiliste et sachant l'effet aléatoire, si ce n'est que les probabilités $L_{a,j}^{\prime(k)}(t)$ sont remplacées par les indicatrices $I\left(\widetilde{s}_{a,t}^{(k)} = j\right)$.

Modèle 2

La vraisemblance de la séquence observée, associée à la séquence d'états restaurée et au vecteur des effets aléatoires prédits, s'écrit sous la forme :

$$f\left(y_{a,1}^{T_a}, \widetilde{s}_{a,1}^{T_a}, \tilde{\xi}_{a,1}^{J}; \theta\right) = f\left(y_{a,1}^{T_a} | \tilde{\xi}_{a,1}^{J}, \widetilde{s}_{a,1}^{T_a}\right) f\left(\tilde{\xi}_{a,1}^{J}\right) P\left(S_{a,1}^{T_a} = \widetilde{s}_{a,1}^{T_a}\right),$$

ou encore

$$\begin{aligned} f\left(y_{a,1}^{T_a}, \widetilde{s}_{a,1}^{T_a}, \tilde{\xi}_{a,1}^{J}; \theta\right) &= \left\{ \prod_{t=1}^{T_a} \frac{1}{\sqrt{2\pi}\sigma_{\widetilde{s}_{a,t}}} \exp\left(-\frac{(y_{a,t} - \beta_{\widetilde{s}_{a,t}} - \tilde{\xi}_{a,\widetilde{s}_{a,t}})^2}{2\sigma^2_{\widetilde{s}_{a,t}}} \right) \right\} \\ &\quad \times \left\{ \prod_{j=1}^{J} \frac{1}{\sqrt{2\pi}\tau_j} \exp\left(-\frac{\tilde{\xi}^2_{a,j}}{2\tau_j^2} \right) \right\} \pi_{\widetilde{s}_{a,1}} \prod_{t=2}^{T_a} p_{\widetilde{s}_{a,t-1}\widetilde{s}_{a,t}}. \end{aligned}$$

Il s'ensuit que la log-vraisemblance de la séquence observée associée à la séquence d'états restaurée et au vecteur des effets aléatoires prédits est donnée par :

$$\begin{aligned} \log f\left(y_{a,1}^{T_a}, \widetilde{s}_{a,1}^{T_a}, \tilde{\xi}_{a,1}^{J}; \theta\right) &= \log \pi_{\widetilde{s}_{a,1}} + \sum_{t=2}^{T_a} \log p_{\widetilde{s}_{a,t-1}\widetilde{s}_{a,t}} - \frac{(T_a + J)}{2} \log 2\pi \\ &\quad + \sum_{t=1}^{T_a} \left(-\log \sigma_{\widetilde{s}_{a,t}} - \frac{\left(y_{a,t} - \beta_{\widetilde{s}_{a,t}} - \xi_{\widetilde{s}_{a,t}}\right)^2}{2\sigma^2_{\widetilde{s}_{a,t}}} \right) + \sum_{j=1}^{J} \left(-\log \tau_j - \frac{\tilde{\xi}^2_{a,j}}{2\tau_j^2} \right), \end{aligned}$$

ce qui peut se réécrire :

$$\begin{aligned} \log f\left(y_{a,1}^{T_a}, \widetilde{s}_{a,1}^{T_a}, \tilde{\xi}_{a,1}^{J}; \theta\right) &= \sum_{j=1}^{J} I\left(\widetilde{s}_{a,1} = j\right) \log \pi_j + \sum_{i,j=1}^{J} \sum_{t=2}^{T_a} I\left(\widetilde{s}_{a,t-1} = i, \widetilde{s}_{a,t} = j\right) \log p_{ij} \\ &\quad - \frac{(T_a + J)}{2} \log 2\pi + \sum_{j=1}^{J} \sum_{t=1}^{T_a} I\left(\widetilde{s}_{a,t} = j\right) \left(-\log \sigma_j - \frac{\left(y_{a,t} - \beta_j - \tilde{\xi}_{a,j}\right)^2}{2\sigma_j^2} \right) \\ &\quad + \sum_{j=1}^{J} \left(-\log \tau_j - \frac{\tilde{\xi}^2_{a,j}}{2\tau_j^2} \right). \end{aligned} \tag{3.22}$$

À l'itération k, les formules de réestimation pour les paramètres markoviens sont identiques à celles données dans le cadre du modèle 1, puisque la structure de chaîne de Markov sous-jacente est identique pour les deux modèles ; seule la modélisation de l'effet aléatoire change.

L'annulation successive de la dérivée de (3.22) par rapport à β_j, τ_j^2 et σ_j^2 conduit aux

formules de réestimation suivantes :

$$\beta_j^{(k+1)} = \frac{\sum_{t=1}^{T_a} I\left(\widetilde{s}_{a,t}^{(k)} = j\right)\left(y_{a,t} - \widetilde{\xi}_{a,j}^{(k)}\right)}{\sum_{t=1}^{T_a} I\left(\widetilde{s}_{a,t}^{(k)} = j\right)} \qquad j = 1, ..., J,$$

$$\tau_j^{2(k+1)} = \left(\widetilde{\xi}_{a,j}^{(k)}\right)^2 \qquad j = 1, ..., J,$$

$$\sigma_j^{2(k+1)} = \frac{\sum_{t=1}^{T_a} I\left(\widetilde{s}_{a,t}^{(k)} = j\right)\left(y_{a,t} - \beta_j^{(k)} - \widetilde{\xi}_{a,j}^{(k)}\right)^2}{\sum_{t=1}^{T_a} I\left(\widetilde{s}_{a,t}^{(k)} = j\right)} \qquad j = 1, ..., J.$$

Ces formules sont similaires à celles obtenues à l'étape de maximisation de l'algorithme itératif avec restauration probabiliste et sachant le vecteur des effets aléatoires, si ce n'est que les probabilités $L_{a,j}^{\prime\prime(k)}(t)$ sont remplacées par les indicatrices $I\left(\widetilde{s}_{a,t}^{(k)} = j\right)$. Les formules pour le paramètre τ_j^2, relatif aux effets aléatoires, sont identiques.

Troisième étape : prédiction

À l'itération k, à partir de $\tilde{s}_{a,1}^{T_a(k)}$, la séquence d'états restaurée et à partir de $\theta^{(k+1)}$, la valeur du paramètre calculée à l'étape de maximisation, on calcule la valeur prédite $\widetilde{\xi}_a^{(k+1)}$ pour le modèle 1 et les J valeurs prédites $\widetilde{\xi}_{a,j}^{(k+1)}, j = 1, ..., J$ pour le modèle 2.

Modèle 1

En toute rigueur, $\widetilde{\xi}_a = E\left(\xi_a | Y_{a,1}^{T_a} = y_{a,1}^{T_a}\right)$. Cependant, nous avons mis en évidence à la section 3.3.3 que ce calcul n'était pas possible. Comme nous connaissons la loi de $Y_{a,1}^{T_a} | S_{a,1}^{T_a} = \tilde{s}_{a,1}^{T_a}$, nous choisissons $E\left(\xi_a | Y_{a,1}^{T_a} = y_{a,1}^{T_a}, S_{a,1}^{T_a} = \tilde{s}_{a,1}^{T_a}\right)$ comme valeur de $\widetilde{\xi}_a$.

Ainsi

$$\widetilde{\xi}_a^{(k+1)} = E\left(\xi_a | Y_{a,1}^{T_a} = y_{a,1}^{T_a}, S_{a,1}^{T_a} = \tilde{s}_{a,1}^{T_a(k)}; \theta^{(k+1)}\right).$$

D'après (3.10), nous avons :

$$\widetilde{\xi}_a^{(k+1)} = \tau'^{(k+1)} U_a' \Gamma_a^{-1(k+1)} \left(Y_a - X_a \beta^{(k+1)}\right),$$

avec $\Gamma_a = U_a \tau \tau' U_a{}' + V_a$.

Modèle 2

Sur le même principe que pour le modèle 1, nous calculons

$$\widetilde{\xi}_{a,j}^{(k+1)} = E\left(\xi_{a,j} | Y_{a,1}^{T_a} = y_{a,1}^{T_a}, S_{a,1}^{T_a} = \tilde{s}_{a,1}^{T_a(k)}; \theta^{(k+1)}\right).$$

D'après (3.14), nous avons :

$$\widetilde{\xi}_{a,j}^{(k+1)} = C_j'^{(k+1)} U_a' \Gamma_a^{-1(k+1)} \left(Y_a - X_a \beta^{(k+1)} \right).$$

avec

· $\Gamma_a = U_a D U_a' + V_a$,

· $C_j = \begin{pmatrix} 0 \\ \vdots \\ \tau_j^2 \\ 0 \\ \vdots \end{pmatrix}$ est le vecteur de dimension $J \times 1$, composé de $J-1$ zéros et ayant la valeur τ_j^2 sur sa $j^{ème}$ ligne.

Algorithme itératif avec restauration par simulation

Nous envisageons à présent une étape de restauration par simulation pour l'algorithme que nous proposons, algorithme itératif en trois étapes. L'algorithme itératif avec restauration par simulation est de type SEM. Rappelons que l'algorithme SEM a été décrit à la section 2.4.3. Dans une première étape, des séquences d'états sont simulées, et dans une seconde étape, on maximise par rapport aux paramètres la log-vraisemblance des données complètes (la séquence observée complétée par la séquence d'états simulée).

Si $\theta^{(k)}$ est la valeur courante du paramètre et $\widetilde{\xi}_a^{(k)}$ la valeur prédite courante de l'effet aléatoire, l'itération k de l'algorithme proposé est décrite par les trois étapes suivantes :

Modèle 1

Étape de restauration : on simule la séquence d'états $\tilde{s}_{a,1}^{T_a(k)} = \left(\tilde{s}_{a,1}^{(k)}, ..., \tilde{s}_{a,T_a}^{(k)} \right)$ selon la loi conditionnelle $P\left(S_{a,1}^{T_a} = s_{a,1}^{T_a} | Y_{a,1}^{T_a} = y_{a,1}^{T_a}, \tilde{\xi}_a^{(k)}; \theta^{(k)} \right)$.

Étape de maximisation : on maximise par rapport à θ la log-vraisemblance des données complètes (la séquence observée complétée par la séquence d'états restaurée et l'effet aléatoire prédit) pour réactualiser la valeur du paramètre. Autrement dit, on choisit $\theta^{(k+1)}$ de sorte que

$$\theta^{(k+1)} = \arg\max_{\theta \in \Theta} \log f \left(\tilde{s}_{a,1}^{T_a(k)}, y_{a,1}^{T_a}, \tilde{\xi}_a^{(k)}; \theta \right).$$

Étape de prédiction : on calcule la valeur prédite $\widetilde{\xi}_a^{(k+1)}$ à partir de $\tilde{s}_{a,1}^{T_a(k)}$, la séquence d'états simulée et à partir de $\theta^{(k+1)}$, la valeur du paramètre calculée à l'étape de maximisation :

$$\widetilde{\xi}_a^{(k+1)} = E\left(\xi_a | Y_{a,1}^{T_a} = y_{a,1}^{T_a}, S_{a,1}^{T_a} = \tilde{s}_{a,1}^{T_a(k)}; \theta^{(k+1)} \right).$$

Modèle 2

La description de l'algorithme est similaire à celle du modèle 1 si ce n'est que la valeur $\tilde{\xi}_a$ est remplacée par les valeurs $\tilde{\xi}_{a,j}, j = 1, ..., J$.

Nous allons à présent décrire en détail chacune des étapes de l'algorithme proposé pour les modèles 1 et 2.

Première étape : restauration par simulation

La première étape est basée sur le principe de l'étape S de l'algorithme SEM dans laquelle, à l'itération k, on simule la séquence des états cachés $\tilde{s}_{a,1}^{T_a(k)} = \left(\tilde{s}_{a,1}^{(k)}, ..., \tilde{s}_{a,T_a}^{(k)}\right)$ selon la loi conditionnelle $P\left(S_{a,1}^{T_a} = s_{a,1}^{T_a} | Y_{a,1}^{T_a} = y_{a,1}^{T_a}; \theta^{(k)}\right)$. Or, la présentation de l'algorithme SEM faite au chapitre 2, dans le cadre de l'estimation des paramètres d'une chaîne de Markov cachée, doit être adaptée pour l'estimation des paramètres des modèles 1 et 2. Comme les observations ne sont plus conditionnellement indépendantes sachant seulement les états, mais sachant en plus l'effet aléatoire ou le vecteur des effets aléatoires, nous prenons en compte les effets aléatoires dans la décomposition de la loi $P\left(S_{a,t} = j | S_{a,t+1}^{T_a} = s_{a,t+1}^{T_a}, Y_{a,1}^{T_a} = y_{a,1}^{T_a}; \theta\right)$. Nous souhaitons donc décomposer $P\left(S_{a,t} = j | S_{a,t+1}^{T_a} = s_{a,t+1}^{T_a}, Y_{a,1}^{T_a} = y_{a,1}^{T_a}, \tilde{\xi}_a; \theta\right)$ pour le modèle 1 et $P\left(S_{a,t} = j | S_{a,t+1}^{T_a} = s_{a,t+1}^{T_a}, Y_{a,1}^{T_a} = y_{a,1}^{T_a}, \tilde{\xi}_{a,1}^{J}; \theta\right)$ pour le modèle 2.

Modèle 1

La probabilité conditionnelle $P\left(S_{a,t} = j | S_{a,t+1}^{T_a} = s_{a,t+1}^{T_a}, Y_{a,1}^{T_a} = y_{a,1}^{T_a}, \tilde{\xi}_a\right)$ peut s'écrire sous la forme :

$$
\begin{aligned}
& P\left(S_{a,t} = j | S_{a,t+1}^{T_a} = s_{a,t+1}^{T_a}, Y_{a,1}^{T_a} = y_{a,1}^{T_a}, \tilde{\xi}_a\right) \\
= & \frac{P\left(Y_{a,t+1}^{T_a} = y_{a,t+1}^{T_a}, S_{a,t+2}^{T_a} = s_{a,t+2}^{T_a} | S_{a,t+1} = s_{a,t+1}\right) P\left(S_{a,t+1} = s_{a,t+1} | S_{a,t} = j\right)}{P\left(Y_{a,t+1}^{T_a} = y_{a,1}^{T_a}, S_{a,t+2}^{T_a} = s_{a,t+2}^{T_a} | S_{a,t+1} = s_{a,t+1}\right) P\left(S_{a,t+1} = s_{a,t+1}, Y_{a,1}^{t} = y_{a,1}^{t}, \tilde{\xi}_a\right)} \\
& \times P\left(S_{a,t} = j, Y_{a,1}^{t} = y_{a,1}^{t}, \tilde{\xi}_a\right) \\
= & \frac{P\left(S_{a,t+1} = s_{a,t+1} | S_{a,t} = j\right) P\left(S_{a,t} = j | Y_{a,1}^{t} = y_{a,1}^{t}, \tilde{\xi}_a\right)}{P\left(S_{a,t+1} = s_{a,t+1} | Y_{a,1}^{t} = y_{a,1}^{t}, \tilde{\xi}_a\right)} \\
= & \frac{p_{j s_{a,t+1}} F'_{a,j}(t)}{\sum_i p_{i s_{a,t+1}} F'_{a,i}(t)},
\end{aligned}
$$

où $F'_{a,j}(t) = P\left(S_{a,t} = j | Y_{a,1}^{t} = y_{a,1}^{t}, \tilde{\xi}_a; \theta\right)$, la probabilité filtrée et $P\left(S_{a,t+1} = s_{a,t+1} | Y_{a,1}^{t} = y_{a,1}^{t}, \tilde{\xi}_a; \theta\right) = \sum_i p_{i s_{a,t+1}} F'_{a,i}(t)$, la probabilité prédite sont toutes deux issues de la récurrence "avant" décrite dans le cadre du modèle 1 à la section 3.3.3.

Selon le même principe que pour l'étape S de l'algorithme SEM, à l'itération k, on procède de la manière suivante :

Pour $t = T_a$ et $j = 1, ..., J$, on calcule grâce à l'algorithme "avant" décrit section 3.3.3

$$P\left(S_{a,T_a} = j | Y_{a,1}^{T_a} = y_{a,1}^{T_a}, \tilde{\xi}_a^{(k)}; \theta^{(k)}\right) = F_{a,j}^{\prime(k)}(T_a).$$

On génère ensuite un nombre aléatoire entre 0 et 1 et on regarde à quel état il correspond sur la fonction de répartition de la loi de probabilité

$$\left\{P\left(S_{a,T_a} = j | Y_{a,1}^{T_a} = y_{a,1}^{T_a}, \tilde{\xi}_a^{(k)}; \theta^{(k)}\right); j = 1, ...J\right\}.$$

Pour $t = T_a - 1, ..., 1$ et $j = 1, ..., J$, on calcule

$$P\left(S_{a,t} = j | S_{a,t+1}^{T_a} = s_{a,t+1}^{T_a^{(k)}}, Y_{a,1}^{T_a} = y_{a,1}^{T_a}, \tilde{\xi}_a^{(k)}; \theta^{(k)}\right) = \frac{P\left(S_{a,t+1} = s_{a,t+1}^{(k)} | S_{a,t} = j\right) F_{a,j}^{\prime(k)}(t)}{\sum_i P\left(S_{a,t+1} = s_{a,t+1}^{(k)} | S_{a,t} = i\right) F_{a,i}^{\prime(k)}(t)}.$$

On génère ensuite un nombre aléatoire entre 0 et 1 et on regarde à quel état il correspond sur la fonction de répartition de la loi de probabilité

$$\left\{P\left(S_{a,t} = j | S_{a,t+1}^{T_a} = s_{a,t+1}^{T_a^{(k)}}, Y_{a,1}^{T_a} = y_{a,1}^{T_a}, \tilde{\xi}_a^{(k)}; \theta^{(k)}\right); j = 1, ...J\right\}.$$

Cette étape de restauration par simulation est similaire à l'étape S de l'algorithme SEM, si ce n'est que les probabilités filtrées $P\left(S_{a,t} = j | Y_{a,1}^t = y_{a,1}^t\right)$, utilisées pour SEM, sont remplacées par les probabilités filtrées tenant compte de l'effet aléatoire $P\left(S_{a,t} = j | Y_{a,1}^t = y_{a,1}^t, \tilde{\xi}_a\right)$.

Modèle 2

Selon le même principe que pour le modèle 1, la probabilité conditionnelle $P\left(S_{a,t} = j | S_{a,t+1}^{T_a} = s_{a,t+1}^{T_a}, Y_{a,1}^{T_a} = y_{a,1}^{T_a}, \tilde{\xi}_{a,1}^J\right)$ peut s'écrire sous la forme :

$$P\left(S_{a,t} = j | S_{a,t+1}^{T_a} = s_{a,t+1}^{T_a}, Y_{a,1}^{T_a} = y_{a,1}^{T_a}, \tilde{\xi}_{a,1}^J\right) = \frac{p_{j s_{a,t+1}} F_{a,j}^{\prime\prime}(t)}{\sum_i p_{i s_{a,t+1}} F_{a,i}^{\prime\prime}(t)},$$

où $F_{a,j}^{\prime\prime}(t) = P\left(S_{a,t} = j | Y_{a,1}^t = y_{a,1}^t, \tilde{\xi}_{a,1}^J; \theta\right)$, la probabilité filtrée et $P\left(S_{a,t+1} = s_{a,t+1} | Y_{a,1}^t = y_{a,1}^t, \tilde{\xi}_{a,1}^J; \theta\right) = \sum_i p_{i s_{a,t+1}} F_{a,i}^{\prime\prime}(t)$, la probabilité prédite sont toutes deux issues de la récurrence "avant" décrite dans le cadre du modèle 2 à la section 3.3.3.

Le procédé de simulation est identique à celui décrit pour le modèle 1. L'étape de restauration pour le modèle 2 est semblable à celle du modèle 1, à la différence près que les probabilités $F_{a,j}^{\prime}(t)$ sont remplacées par les probabilités $F_{a,j}^{\prime\prime}(t)$.

Seconde étape : maximisation

Pour effectuer une maximisation au sens de l'algorithme de SEM, il faudrait maximiser par rapport à chacun des paramètres à estimer la quantité $P\left(S_{a,1}^{T_a} = \tilde{s}_{a,1}^{T_a(k)}, Y_{a,1}^{T_a} = y_{a,1}^{T_a}; \theta\right)$ où $\tilde{s}_{a,1}^{T_a(k)}$ désigne la séquence d'états simulée à l'étape précédente. Cette quantité ne pouvant pas se calculer dans le cadre des modèles 1, et 2, on maximise $P\left(S_{a,1}^{T_a} = \tilde{s}_{a,1}^{T_a(k)}, Y_{a,1}^{T_a} = y_{a,1}^{T_a}, \tilde{\xi}_a^{(k)}; \theta\right)$ pour le modèle 1 et $P\left(S_{a,1}^{T_a} = \tilde{s}_{a,1}^{T_a(k)}, Y_{a,1}^{T_a} = y_{a,1}^{T_a}, \tilde{\xi}_{a,1}^{J(k)}; \theta\right)$ pour le modèle 2.

Les formules de réestimation pour les paramètres des modèles 1 et 2, obtenues dans le cadre d'une maximisation au sens de SEM, sont similaires à celles obtenues dans le cadre d'une maximisation au sens de Baum-Viterbi, si ce n'est que $\tilde{s}_{a,t}^{(k)}$ désigne l'état simulé à l'instant t au lieu de l'état globalement optimal à l'instant t.

Troisième étape : prédiction

Le principe de l'étape de prédiction est similaire à celui décrit dans le cadre de l'algortihme de type Baum-Viterbi, à la différence près que $\tilde{s}_{a,1}^{T_a}$ désigne la séquence d'états simulée au lieu de la séquence d'états globalement optimale.

3.4 Conclusion

La double structure cachée du modèle linéaire mixte multiphasique implique que les observations ne sont pas seulement conditionnellement indépendantes sachant les états comme pour une chaîne de Markov cachée. Les observations sont conditionnellement indépendantes sachant la séquence d'états et l'effet aléatoire pour le modèle 1, et elles sont conditionnellement indépendantes sachant la séquence d'états et le vecteur des effets aléatoires pour le modèle 2. Du fait de ces relations d'indépendance conditionnelles spécifiques, l'estimation des paramètres de ces deux familles de modèles n'est pas aisée. En particulier l'étape E de l'algorithme EM ne s'écrit pas de manière analytique et par conséquent l'algorithme EM est écarté comme méthode d'estimation des paramètres pour la famille des modèles linéaires mixtes multiphasiques. Nous avons également envisagé un algorithme de type EM sachant les valeurs prédites des effets aléatoires. Cet algorithme itératif est composé de trois étapes : restauration, maximisation et prédiction. L'étape de restauration est probabiliste, et s'implémente par un algorithme "avant-arrière" sachant les effets aléatoires. L'étape de maximisation s'écrit sans difficulté. Toutefois il est nécessaire de savoir prédire à chaque itération une nouvelle valeur des effets aléatoires. Cette étape de prédiction pose problème car l'étape de restauration probabiliste détermine l'ensemble des séquences d'états possibles.

Nous proposons comme alternative à l'algorithme EM, un algorithme itératif, en trois étapes : restauration, maximisation et prédiction. L'étape de restauration déterministe ou effectuée par simulation nécessite l'hypothèse supplémentaire d'intégrer les effets aléatoires au calcul de la séquence d'états restaurée. L'étape de maximisation nécessite cette même hypothèse pour la maximisation de la probabilité jointe de la séquence observée et de la séquence d'états restaurée. Par conséquent, une étape de prédiction est nécessaire

pour calculer les valeurs prédites des effets aléatoires, valeurs utilisées pour les étapes de restauration et de maximisation.

Le fait que l'étape de restauration soit déterministe ou effectuée par simulation permet d'effectuer une étape de prédiction sans difficulté. Par conséquent, nous pourrions envisager, pour l'algorithme avec restauration probabiliste et sachant les effets aléatoires, une étape de prédiction pour laquelle une étape de restauration déterministe ou par simulation serait effectuée, afin de restaurer une seule séquence d'états utilisée pour le calcul de la prédiction.

Chapitre 4

Application : analyse de la croissance en longueur d'arbres forestiers

4.1 Introduction

Ce chapitre est consacré à l'application de la modélisation linéaire mixte multiphasique présentée au chapitre 3 pour l'analyse de la croissance en longueur d'arbres forestiers. L'objectif est tout d'abord de déterminer le nombre de phases de croissance qui composent la période de croissance globale de l'échantillon, et de les identifier précisément par le biais de la restauration. Ensuite, l'objectif est de sélectionner un modèle linéaire (mixte) sur chacune des phases restaurées pour mettre en évidence l'influence éventuelle de facteurs climatiques (stress hydriques) ainsi que la présence ou l'absence d'hétérogénéité inter-individuelle à l'intérieur de la phase de croissance considérée. Nous restreignons la présentation à la modélisation de deux des jeux de données présentés au chapitre 1 – les chênes sessiles âgés de 15 ans (Fig 1-4) et les pins laricios âgés de 18 ans (Fig 1-8) – avec le modèle 2. Dans un premier temps, la démarche pour l'étude des deux jeux de données sera décrite. Par la suite, les résultats de l'application à chacun des jeux de données seront exposés et interprétés.

4.2 Démarche

Chaque jeu de données est modélisé avec le modèle linéaire mixte multiphasique ayant un effet aléatoire différent pour chaque état. Comme cela a été mis en évidence au chapitre 1, pour chacune des phases, la tendance reflétant l'évolution de la croissance sur la phase et les covariables climatiques sont modélisées par des effets fixes et un effet aléatoire "individu" modélise l'hétérogénéité entre les individus. La tendance sera représentée par un simple intercept reflétant le niveau moyen de croissance sur la phase de croissance ou par un polynôme d'ordre 1. La covariable climatique modélisant les stress hydriques détectés pendant la période d'allongement (resp. la période d'organogenèse) peut être prise en compte seule dans la modélisation ou bien les deux covariables climatiques, chacune relative à l'une des deux périodes, peuvent être prises en compte. À titre d'exemple,

nous donnons l'écriture du modèle 2 pour lequel les effets fixes sont représentés par une tendance polynomiale d'ordre 1 et par les deux covariables climatiques, chacune relative à l'une des deux périodes. La modélisation de l'observation $y_{a,t}$, relative à l'individu a se trouvant dans l'état j à l'instant t, s'écrit sous la forme suivante :

$$y_{a,t}|_{S_{a,t}=j} = \alpha_j+\delta_j t+\gamma_{1j}x_{1,t}+\gamma_{2j}x_{2,t-1}+\xi_{a,j}+\varepsilon_{a,t}, \qquad a = 1, ..., N; t = 1, ..., T_a; j = 1, ..., J,$$

où
· α_j et δ_j sont l'intercept et la pente de la tendance polynomiale d'ordre 1,
· γ_{1j} (resp. γ_{2j}) est le coefficient de la covariable climatique $x_{1,t}$ (resp. $x_{2,t-1}$) modélisant les stress hydriques détectés sur la période d'allongement (resp. sur la période d'organogenèse). La covariable climatique modélisant les stress hydriques détectés sur la période d'organogenèse est indicée par $t-1$ et non par t car un stress hydrique détecté au cours de la période d'organogenèse de l'année $n-1$ est susceptible d'influencer la longueur de la pousse annuelle de l'année n. Remarquons que pour $t = 1$, $x_{2,0}$ est connue car les données climatiques sont connues pour l'année précédant le début de la période d'étude. Les covariables climatiques $x_{1,t}$ et $x_{2,t-1}$ peuvent représenter les différents points de vue de la covariable stress hydrique définis au chapitre 1 : indicatrice, quantification, nombre de stress.
· $\xi_{a,j} \sim N(0, \tau_j^2)$ est l'effet aléatoire relatif à l'individu a se trouvant dans l'état j,
· $\varepsilon_{a,t} \sim N(0, \sigma_j^2)$ est le terme d'erreur aléatoire relatif à l'individu a se trouvant dans l'état j à l'instant t,
· N est le nombre d'individus, T_a est la longueur de la séquence observée relative à l'individu a et J est le nombre d'états.

Remarquons que l'expression $\alpha_j + \delta_j t + \gamma_{1j}x_{1,t} + \gamma_{2j}x_{2,t-1}$ représente le terme "β_j" utilisé dans les chapitres 2 et 3 comme terme générique pour désigner les effets fixes. Selon les notations adoptées dans la section 3.3.2, le vecteur β des effets fixes est dans ce cas précis un vecteur de dimension $4J \times 1$. La matrice X_a, matrice d'incidence de β est alors de dimension $T_a \times 4J$ et la matrice U_a, matrice d'incidence du vecteur des effets aléatoires $\xi_{a,1}^J$ est de dimension $T_a \times J$. Les matrices X_a et U_a ne sont identiques que lorsque le terme des effets fixes se résume à un simple intercept, sans covariable climatique.

Les paramètres du modèle 2 sont estimés avec l'algorithme itératif de type Baum-Viterbi, avec restauration déterministe, décrit à la section 3.3.4. L'implémentation de cet algorithme a été effectuée sous le logiciel statistique libre R (Ihaka et Gentleman, 1996). Dans le cas précis de la modélisation écrite ci-dessus, pour l'ensemble des N séquences, les formules de réestimation pour les paramètres des J modèles linéaires mixtes, à l'itération k, sont les suivantes :

$$\alpha_j^{(k+1)} = \frac{\sum_{a=1}^{N}\sum_{t=1}^{T_a} I\left(\widetilde{s}_{a,t}^{(k)} = j\right)\left(y_{a,t} - \delta_j^{(k)}t - \gamma_{1j}^{(k)}x_{1,t} - \gamma_{2j}^{(k)}x_{2,t-1} - \widetilde{\xi}_{a,j}^{(k)}\right)}{\sum_{a=1}^{N}\sum_{t=1}^{T_a} I\left(\widetilde{s}_{a,t}^{(k)} = j\right)} \qquad j = 1, ..., J,$$

$$\delta_j^{(k+1)} = \frac{\sum_{a=1}^{N}\sum_{t=1}^{T_a} I\left(\widetilde{s}_{a,t}^{(k)} = j\right)\left(y_{a,t} - \alpha_j^{(k)} - \gamma_{1j}^{(k)} x_{1,t} - \gamma_{2j}^{(k)} x_{2,t-1} - \widetilde{\xi}_{a,j}^{(k)}\right) t}{\sum_{a=1}^{N}\sum_{t=1}^{T_a} I\left(\widetilde{s}_{a,t}^{(k)} = j\right) t^2} \qquad j = 1, ..., J,$$

$$\gamma_{1j}^{(k+1)} = \frac{\sum_{a=1}^{N}\sum_{t=1}^{T_a} I\left(\widetilde{s}_{a,t}^{(k)} = j\right)\left(y_{a,t} - \alpha_j^{(k)} - \delta_j^{(k)} t - \gamma_{2j}^{(k)} x_{2,t-1} - \widetilde{\xi}_{a,j}^{(k)}\right) x_{1,t}}{\sum_{a=1}^{N}\sum_{t=1}^{T_a} I\left(\widetilde{s}_{a,t}^{(k)} = j\right) x_{1,t}^2} \qquad j = 1, ..., J,$$

$$\gamma_{2j}^{(k+1)} = \frac{\sum_{a=1}^{N}\sum_{t=1}^{T_a} I\left(\widetilde{s}_{a,t}^{(k)} = j\right)\left(y_{a,t} - \alpha_j^{(k)} - \delta_j^{(k)} t - \gamma_{1j}^{(k)} x_{1,t} - \widetilde{\xi}_{a,j}^{(k)}\right) x_{2,t-1}}{\sum_{a=1}^{N}\sum_{t=1}^{T_a} I\left(\widetilde{s}_{a,t}^{(k)} = j\right) x_{2,t-1}^2} \qquad j = 1, ..., J,$$

$$\tau_j^{2(k+1)} = \frac{\sum_{a=1}^{N}\left(\widetilde{\xi}_{a,j}^{(k)}\right)^2}{N} \qquad j = 1, ..., J,$$

$$\sigma_j^{2(k+1)} = \frac{\sum_{a=1}^{N}\sum_{t=1}^{T_a} I\left(\widetilde{s}_{a,t}^{(k)} = j\right)\left(y_{a,t} - \alpha_j^{(k)} - \delta_j^{(k)} t - \gamma_{1j}^{(k)} x_{1,t} - \gamma_{2j}^{(k)} x_{2,t-1} - \widetilde{\xi}_{a,j}^{(k)}\right)^2}{\sum_{a=1}^{N}\sum_{t=1}^{T_a} I\left(\widetilde{s}_{a,t}^{(k)} = j\right)} \quad j = 1, ..., J.$$

Initialisation de l'algorithme

Le premier travail consiste à définir les J phases de croissance successives qui composent la période de croissance globale des arbres. Cette étape est assez subjective, mais il s'agit de découper des phases bien distinctes, délimitées par des ruptures assez nettes.

Ensuite, sur chacune de ces phases, les valeurs des paramètres du modèle linéaire mixte qui ajuste au mieux les données, et les valeurs prédites des J effets aléatoires pour chacun des individus sont obtenues avec des procédures de R. Ces valeurs sont alors utilisées pour construire les J modèles linéaires mixtes et la matrice des $N \times J$ effets aléatoires prédits, nécessaires à l'initialisation.

L'initialisation requiert également des paramètres initiaux relatifs à la chaîne de Markov sous-jacente :

· Les J probabilités initiales sont choisies de sorte que leur somme vaut 1.

· La matrice des probabilités de transition est une matrice dont la triangulaire inférieure est nulle, traduisant la structure gauche-droite de la chaîne de Markov. Autrement dit, la croissance de l'arbre est traduite en supposant que lorsque l'arbre se trouve dans un état, il peut soit y rester soit passer dans un des états suivants, mais il ne peut pas revenir dans l'état précédent. Les probabilités de bouclage p_{jj}, $j = 1, ..., J$ sont déterminées en égalant la longueur moyenne de la phase j (obtenue à partir des données) à la moyenne

$1/(1-p_{jj})$ d'une loi géométrique de paramètre $1-p_{jj}$, qui est la loi d'occupation de l'état j d'une chaîne de Markov.

Principe de l'algorithme

Le principe de l'algorithme consiste à itérer les trois étapes de restauration, maximisation et prédiction jusqu'à stabilisation des séquences d'états optimales restaurées et jusqu'à convergence des paramètres et des effets aléatoires prédits.

La démarche pour l'étude de chacun des jeux de données est la suivante :
Tout d'abord, le nombre de phases de croissance composant la période de croissance de l'échantillon d'arbres est sélectionné sur la base de critères de sélection décrits ci-dessous. Le nombre de phases de croissance déterminé, chaque période de croissance est restaurée. Remarquons que la restauration est spécifique à chaque individu : comme nous le verrons dans l'exemple des chênes sessiles, pour une année donnée, tous les individus ne sont pas forcément dans la même phase de croissance. Toutefois, il y a une forte homogénéité dans la segmentation parmi l'ensemble des individus. Un critère de validation consiste à vérifier que la segmentation met bien en évidence des phases séparées par des ruptures nettes. Il est également intéressant de déterminer, de manière globale, lequel des trois éléments intervenant dans la modélisation (le nombre de phases de croissance, les covariables climatiques, l'effet aléatoire) est prépondérant.
Une fois le nombre de phases de croissance sélectionné et les phases de croissance restaurées, un modèle linéaire (mixte) est sélectionné sur chacune des phases pour déterminer et interpréter l'influence des différentes sources de variation : covariables climatiques, effet aléatoire.

Explicitons à présent les critères de sélection utilisés pour ce travail.

Sélection de modèles

La sélection des modèles s'effectue sur la base de critères de type AIC (Akaike Information Criterion) et BIC (Bayesian Information Criterion). Le principe des critères AIC et BIC (Burnham et Anderson, 2002) consiste à compenser la vraisemblance des paramètres estimés associée aux données, traduisant la qualité de l'ajustement, par une fonction judicieusement choisie du nombre de paramètres indépendants du modèle et éventuellement de la taille de l'échantillon. Autrement dit, ce sont des critères de vraisemblance pénalisée qui sélectionnent le modèle réalisant le meilleur compromis entre ajustement aux données et parcimonie du modèle. Ces deux critères se définissent comme suit :

$$\begin{aligned} \text{AIC} &= -2L + 2K, \\ \text{BIC} &= -2L + K\log(n), \end{aligned}$$

où
· L désigne la log-vraisemblance des paramètres associée aux données,
· K désigne le nombre de paramètres indépendants du modèle,

· n est la taille de l'échantillon.

Le meilleur modèle au sens du critère AIC ou BIC est celui pour lequel la valeur du critère est la plus faible. Rappelons que d'après les abaques de Jeffreys (1961), une différence de BIC est très significative et permet de prendre une décision si elle est strictement supérieure à 10.

L'utilisation de ces deux critères dans le cadre de la sélection de modèles markoviens n'est justifiée que si la chaîne de Markov est ergodique (formée d'une seule classe récurrente apériodique). Or dans notre cas, la chaîne de Markov sous-jacente n'est pas ergodique puisqu'elle est composée d'une succession d'états transitoires et d'un état final absorbant (structure "gauche-droite"). Toutefois, sur la base de ces deux critères, nous prendrons pour L la log-vraisemblance des données observées sachant le vecteur des effets aléatoires prédits. Pour la séquence d'observations $y_{a,1}^{T_a}$ relative à l'individu a, la log-vraisemblance des données observées sachant le vecteur des effets aléatoires prédits est donnée par $P\left(y_{a,1}^{T_a}|\xi_{a,1}^{J}\right) = \prod_{t=1}^{T_a} N''_{a,t}$ où les $N''_{a,t}$ sont les facteurs de normalisation définis lors de la description de la récurrence "avant" à la section 3.3.3. Le nombre K est obtenu en ajoutant au nombre de paramètres markoviens indépendants le nombre de paramètres relatifs aux J modèles linéaires mixtes. Par la suite, nous garderons la terminologie AIC et BIC pour les deux critères utilisés, qui ne sont pas de véritables AIC et BIC, mais qui en sont inspirés.

4.3 Les chênes sessiles âgés de 15 ans

4.3.1 Sélection du nombre de phases de croissance

Le premier jeu de données étudié est celui des 46 chênes sessiles âgés de 15 ans (Fig 1-4). L'étude débute par la sélection du nombre de phases de croissance. Le tableau 4-1 récapitule les différentes valeurs de L, K, et des critères AIC et BIC pour divers modèles qui diffèrent par le nombre d'états (c'est-à-dire de phases de croissance), par la présence ou l'absence de covariable climatique (notée covar clim), et par la présence ou l'absence d'effet aléatoire individuel (modèle linéaire mixte, noté L2M ou modèle linéaire, noté LM). Les résultats sont donnés pour une tendance polynomiale d'ordre 1 et pour la covariable climatique représentée par la quantification des stress détectés durant la période d'allongement. La figure 4-1 présente le cas où l'on envisage deux phases de croissance avec un changement de phases entre les années 1989 et 1990. La figure 4-2 présente le cas où l'on envisage trois phases de croissance avec un premier changement entre 1986 et 1987 et un second changement entre 1989 et 1990. Remarquons que les changements de phases sont envisagés de manière globale, pour la majorité des individus, mais ils pourraient se situer une ou deux années avant pour certains individus. Envisager plus de trois phases de croissance semble peu réaliste au vu des données.

Nb états	LM : tendance	LM : tendance + covar clim	L2M : tendance	L2M : tendance + covar clim
1	$L = -2792.85$ $K = 3$ AIC = 5591.69 BIC = 5605.03	$L = -2773.76$ $K = 4$ **AIC = 5555.51** **BIC = 5573.30**	$L = -2792.85$ $K = 4$ AIC = 5593.71 BIC = 5611.49	$L = -2773.76$ $K = 5$ AIC = 5557.53 BIC = 5579.76
2	$L = -2600.80$ $K = 7$ AIC = 5215.61 BIC = 5246.73	$L = -2573.46$ $K = 10$ **AIC = 5166.93** **BIC = 5211.39**	$L = -2598.42$ $K = 10$ AIC = 5216.84 BIC = 5261.29	$L = -2572.96$ $K = 12$ AIC = 5169.93 BIC = 5223.28
3	$L = -2573.72$ $K = 13$ AIC = 5172.91 BIC = 5230.70	$L = -2557.72$ $K = 16$ **AIC = 5146.97** **BIC = 5218.10**	$L = -2573.72$ $K = 16$ AIC = 5179.43 BIC = 5250.56	$L = -2556.96$ $K = 19$ AIC = 5151.92 BIC = 5236.39

TAB. 4-1 – Sélection du nombre de phases de croissance chez les chênes sessiles âgés de 15 ans.

À la lecture du tableau, il apparaît que l'effet aléatoire n'apporte rien dans la modélisation, et ce quel que soit le nombre d'états. Entre le modèle linéaire avec covariable climatique à 1 état et celui à 2 états, la différence d'AIC est de près de 400, et la différence de BIC est de plus de 300. Cela prouve que le passage de 1 à 2 états améliore très nettement la modélisation. Enfin, la prise en compte d'une covariable climatique améliore le modèle de façon significative. En effet, la différence d'AIC (resp. de BIC) entre le modèle linéaire à 2 états sans covariable climatique et le modèle linéaire à 2 états avec covariable climatique est de près de 50 (resp. de plus de 30). Ainsi, pour les chênes sessiles âgés de 15 ans, dans la modélisation, le nombre d'états est l'élément nettement prépondérant devant les covariables climatiques qui sont elles-mêmes prépondérantes devant l'effet aléatoire.

Le critère AIC sélectionne le modèle linéaire multiphasique à 3 états avec covariable climatique alors que le critère BIC sélectionne le modèle linéaire multiphasique à 2 états avec covariable climatique. Cela confirme le fait que le critère AIC a tendance à donner un modèle sur-paramétré alors que le critère BIC a tendance à donner un modèle sous-paramétré. Le fait que le modèle à 2 états ait moins de paramètres que le modèle à 3 états pourrait être un argument favorable au modèle 2. La comparaison entre les séquences d'états restaurées pour 2 états et les séquences d'états restaurées pour 3 états peut également être un argument de décision. Nous choisissons de présenter les résultats liés à la restauration des séquences d'états pour le modèle à 2 états, et les résultats pour la comparaison et la sélection de modèles phase par phase sont donnés pour le modèle à 3 états.

Comme nous l'avons déjà mentionné à la section 4.2, la restauration des séquences d'états est spécifique à chaque individu. À titre d'illustration, la figure 4-3 représente trois séquences d'états globalement optimales restaurées parmi les 46. Il s'agit de la restauration effectuée dans le cadre du modèle à 2 états. Pour l'individu représenté en trait plein noir,

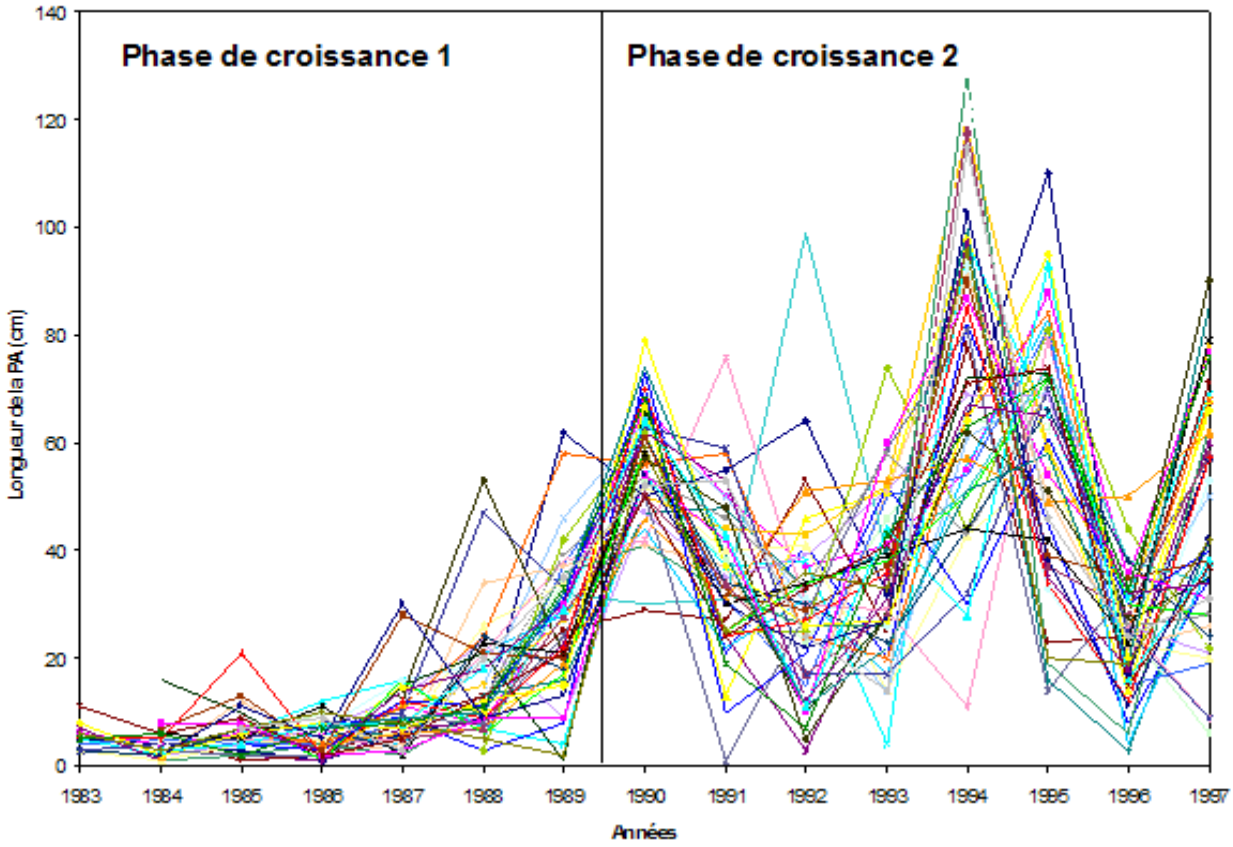

FIG. 4-1 – Deux phases de croissance envisagées pour la période de croissance des chênes sessiles âgés de 15 ans.

deux phases de croissance ont été restaurées avec un changement entre 1987 et 1988. Pour les deux autres individus représentés par des pointillés rouge et vert, le changement entre les deux phases de croissance restaurées a lieu entre 1989 et 1990. La restauration de la séquence d'états globalement optimale est spécifique à chaque individu dans le sens où, pour une année donnée, tous les individus ne sont pas dans la même phase de croissance. Toutefois, après restauration déterministe de l'ensemble des séquences d'états globalement optimales, une forte homogénéité dans la segmentation en phases est mise en évidence parmi les 46 individus.

De même, un travail de validation consiste à vérifier que les changements de phases correspondent bien à des sauts au niveau de la longueur de la pousse annuelle. Ainsi, pour les individus représentés en rouge et en vert, le changement de phase entre 1989 et 1990 est caractérisé par une nette variation de la longueur de la pousse annuelle. Cette

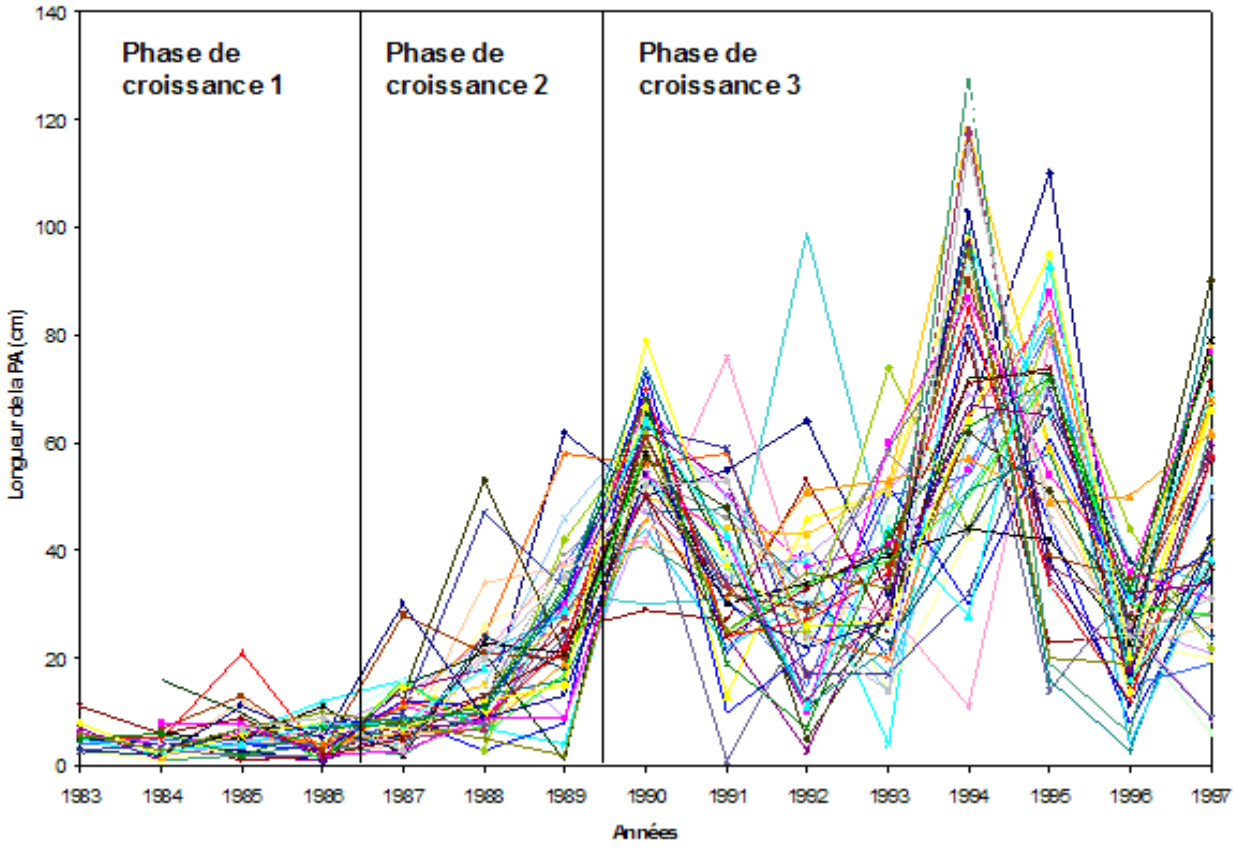

FIG. 4-2 – Trois phases de croissance envisagées pour la période de croissance des chênes sessiles âgés de 15 ans.

longueur passe de 20cm en 1989 à près de 70cm en 1990 pour l'individu en rouge, et de 25cm en 1989 à 70cm en 1990 pour l'individu en vert.

4.3.2 Comparaison et sélection de modèles phase par phase

Le nombre de phases de croissance déterminé, nous souhaitons sélectionner un modèle sur chacune des phases de croissance afin d'interpréter la croissance des arbres sur les différentes phases, et ce en comparant les deux sources de variation : les covariables climatiques et l'effet aléatoire. Sur la base des mêmes critères de sélection, les trois tableaux 4-2, 4-3 et 4-4 comparent pour chacune des trois phases de croissance divers modèles qui diffèrent par la présence ou l'absence d'effet aléatoire (L2M/LM) et par la présence

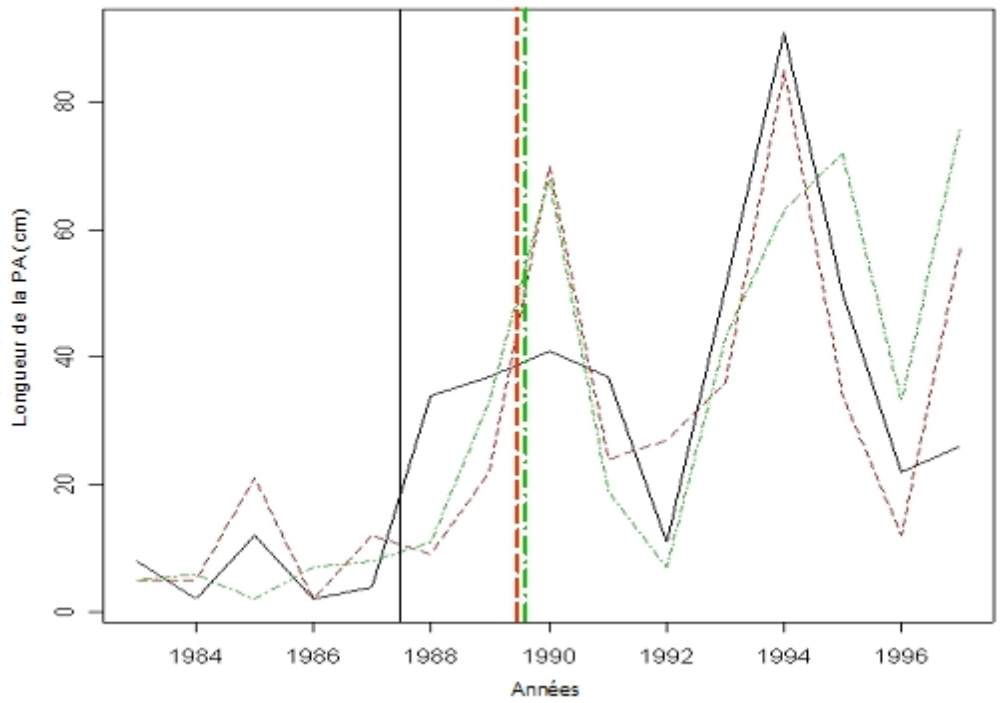

FIG. 4-3 – Restauration des séquences d'états globalement optimales pour 3 des chênes âgés de 15 ans.

ou l'absence de covariable climatique. Lorsqu'une covariable climatique est présente, on compare les différents points de vue – indicatrice, quantification et nombre de stress – de la covariable stress hydrique.

Phase de croissance 1

	Pas de covar clim	Indicatrice	Quantification	Nb stress
LM	AIC = 719.48 BIC = 725.57	AIC = 713.10 BIC = 722.23	AIC = 710.29 BIC = 719.42	**AIC = 706.24** **BIC = 715.37**
L2M	AIC = 723.48 BIC = 735.66	AIC = 717.11 BIC = 732.32	AIC = 714.30 BIC = 729.51	AIC = 710.25 BIC = 725.46

TAB. 4-2 – Comparaison et sélection de modèles pour la phase de croissance 1.

Le modèle linéaire simple, sans effet aléatoire et avec la covariable climatique caractérisée par le nombre de stress, est le meilleur modèle au sens des critères AIC et BIC. L'ajout d'un effet aléatoire n'améliore pas le modèle. L'ajout de la covariable climatique caractérisée par l'indicatrice ou par la quantification améliore le modèle, mais de façon

non significative. Les différences d'AIC (plus de 13) et de BIC (plus de 10) sont seulement significatives pour l'ajout de la covariable climatique caractérisée par le nombre de stress.

Phase de croissance 2

	Pas de covar clim	Indicatrice	Quantification	Nb stress
LM	AIC = 878.60 BIC = 884.03	AIC = 858.19 BIC = 866.35	AIC = 847.13 BIC = 855.29	**AIC = 725.52** **BIC = 733.67**
L2M	AIC = 882.60 BIC = 893.47	AIC = 862.20 BIC = 875.79	AIC = 851.14 BIC = 864.73	AIC = 729.52 BIC = 743.11

TAB. 4-3 – Comparaison et sélection de modèles pour la phase de croissance 2.

C'est encore le modèle linéaire simple avec la covariable climatique caractérisée par le nombre de stress qui est le meilleur modèle au sens des deux critères. Cette fois, l'ajout de l'une des trois caractérisations de la covariable climatique améliore de façon très significative la modélisation. La différence d'AIC (plus de 150) et de BIC (plus de 110) entre le modèle linéaire sans covariable climatique et le modèle linéaire avec la covariable climatique caractérisée par le nombre de stress est la plus importante.

Phase de croissance 3

	Pas de covar clim	Indicatrice	Quantification	Nb stress
LM	AIC = 3340.19 BIC = 3347.98	AIC = 3342.17 BIC = 3353.85	**AIC = 3310.09** **BIC = 3321.77**	AIC = 3327.07 BIC = 3338.76
L2M	AIC = 3344.20 BIC = 3359.77	AIC = 3346.17 BIC = 3365.64	AIC = 3314.09 BIC = 3333.56	AIC = 3331.08 BIC = 3350.54

TAB. 4-4 – Comparaison et sélection de modèles pour la phase de croissance 3.

Le modèle linéaire avec la covariable climatique caractérisée par la quantification est le meilleur modèle. L'ajout de la covariable climatique caractérisée par l'indicatrice n'améliore pas le modèle, alors que l'ajout de la covariable climatique caractérisée par la quantification améliore la modélisation de façon très significatrice. La croissance est donc soumise au climat sur cette période.

Ainsi, les chênes sessiles âgés de 15 ans ne sont soumis à l'effet aléatoire "individu" sur aucune des périodes. Il n'y a pas d'hétérogénéité entre les individus au cours des différentes phases de croissance. Cela signifie que les arbres ont une croissance homogène sur toute leur période de croissance. Le climat influence la croissance de manière importante sur les seconde et troisième périodes, alors qu'il l'influence de façon moindre sur la première période.

En plus de confirmer les résultats obtenus pour la modélisation de la période globale lors de la sélection du nombre de phases (l'effet aléatoire n'apporte rien dans la modélisation, la prise en compte d'une covariable climatique améliore la modélisation), la sélection de modèles phase par phase permet de préciser l'importance de l'influence du climat au cours des différents âges.

4.4 Les pins laricios âgés de 18 ans

4.4.1 Sélection du nombre de phases de croissance

Le second jeu de données étudié est celui des 30 pins laricios âgés de 18 ans (Fig 1-8). Le principe de l'étude est le même que pour les chênes. Le tableau 4-5 récapitule les différentes valeurs de L, K, et des critères AIC et BIC pour divers modèles qui diffèrent par le nombre d'états, par la présence ou l'absence de covariable climatique et par la présence ou l'absence d'effet aléatoire. Comme pour les chênes, les résultats sont donnés pour une tendance polynomiale d'ordre 1 et pour la covariable climatique représentée par la quantification des stress détectés durant la période d'allongement. La figure 4-4 présente le cas où l'on envisage deux phases de croissance avec un changement de phases pour la majorité des individus entre les années 1981 et 1982. La figure 4-5 présente le cas où l'on envisage trois phases de croissance avec un premier changement entre 1981 et 1982 et un second changement pour la plupart des individus entre 1987 et 1988. Comme pour les chênes, envisager plus de trois phases de croissance paraît peu réaliste au vu des données.

Même si l'hypothèse où la période de croissance est composée de trois phases de croissance paraît tout à fait plausible, la restauration déterministe des séquences d'états globalement optimales n'a restauré que deux phases de croissance pour la totalité des 30 individus. Par conséquent, nous ne présentons que les résultats comparatifs entre les modèles à 1 état et ceux à 2 états.

Nb états	LM : tendance	LM : tendance + covar clim	L2M : tendance	L2M : tendance + covar clim
1	$L = -1983.72$ $K = 3$ AIC = 3973.44 BIC = 3986.07	$L = -1982.55$ $K = 4$ AIC = 3973.11 BIC = 3989.94	$L = -1881.83$ $K = 4$ AIC = 3771.66 BIC = 3788.46	$L = -1879.82$ $K = 5$ **AIC = 3769.63** **BIC = 3787.67**
2	$L = -1863.23$ $K = 9$ AIC = 3744.47 BIC = 3782.34	$L = -1858.89$ $K = 11$ AIC = 3739.78 BIC = 3786.07	$L = -1861.82$ $K = 9$ AIC = 3741.64 BIC = 3779.51	$L = -1849.76$ $K = 11$ **AIC = 3721.53** **BIC = 3767.82**

TAB. 4-5 – Sélection du nombre de phases de croissance chez les pins laricios âgés de 18 ans.

Contrairement aux chênes, l'effet aléatoire améliore fortement la modélisation à un seul état. En effet, la différence d'AIC (resp. de BIC) entre le modèle linéaire sans covariable climatique et le modèle linéaire mixte sans covariable climatique est de plus de 200 (resp. près de 200). De même, le passage de 1 à 2 états engendre une différence d'AIC et de BIC de plus de 200 entre le modèle linéaire sans covariable climatique à 1 état et le modèle linéaire sans covariable climatique à 2 états. Ces mêmes différences se retrouvent entre le

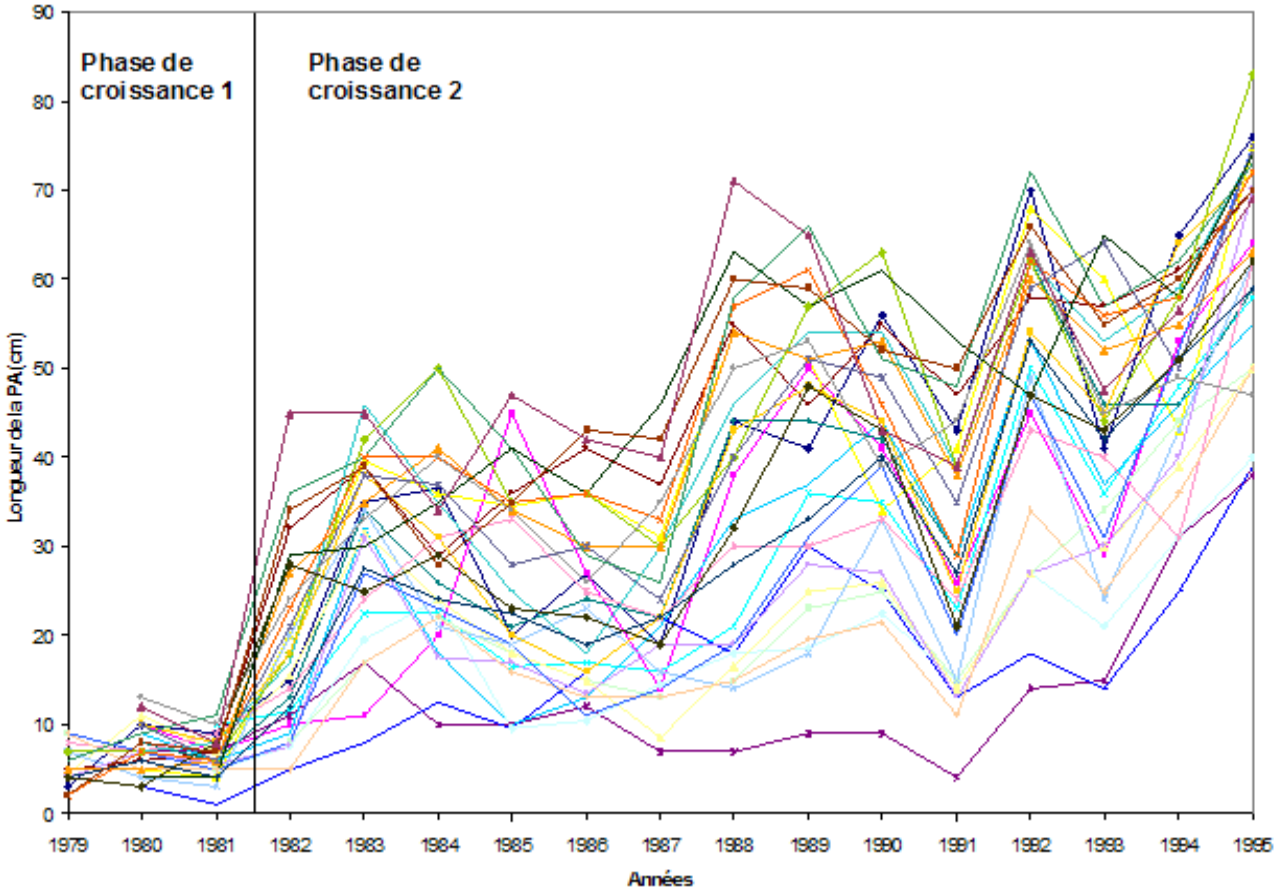

FIG. 4-4 – Deux phases de croissance envisagées pour la période de croissance des pins laricios âgés de 18 ans.

modèle linéaire avec covariable climatique à 1 état et celui à 2 états. Enfin, l'effet de la covariable climatique n'est pas significatif avec un seul état alors qu'il l'est pour le modèle linéaire mixte multiphasique à 2 états (différences d'AIC et de BIC supérieures à 20). Par conséquent, dans la modélisation des pins laricios de 18 ans, l'effet aléatoire et le nombre de phases de croissance sont prépondérants devant les covariables climatiques.

Les critères AIC et BIC sélectionnent tous deux le modèle linéaire mixte multiphasique à 2 états avec covariable climatique.

4.4.2 Comparaison et sélection de modèles phase par phase

Comme pour les chênes, les deux tableaux 4-6 et 4-7 comparent, pour chacune des deux phases de croissance, divers modèles qui diffèrent par la présence ou l'absence d'effet aléatoire et par la présence ou l'absence de covariable climatique. Lorsqu'une covariable

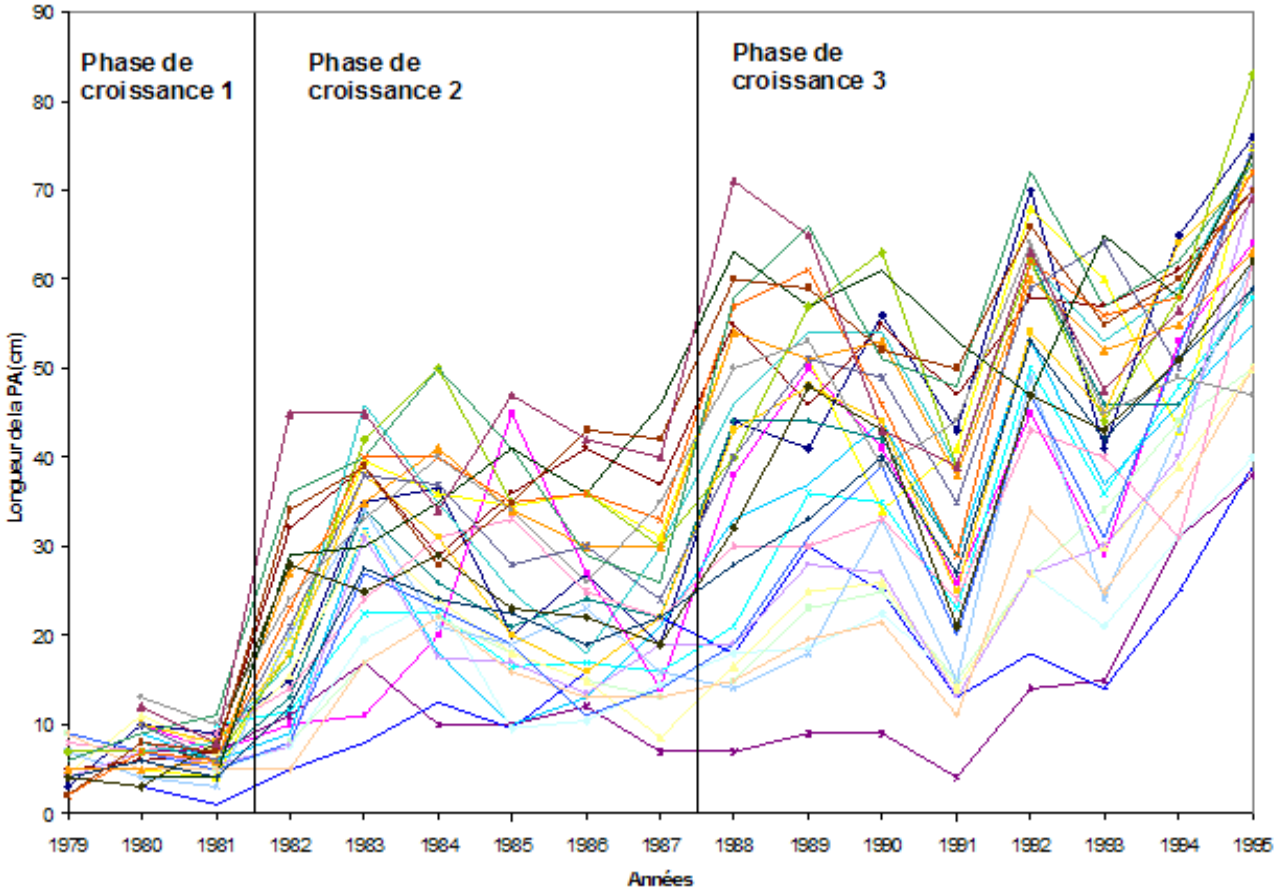

FIG. 4-5 – Trois phases de croissance envisagées pour la période de croissance des pins laricios âgés de 18 ans.

climatique est présente, on compare les différents points de vue – indicatrice, quantification et nombre de stress – de la covariable stress hydrique.

Phase de croissance 1

	Pas de covar clim	Indicatrice	Quantification	Nb stress
LM	AIC = 577.88 BIC = 583.32	AIC = 576.29 BIC = 584.45	**AIC = 569.38** **BIC = 577.53**	AIC = 569.95 BIC = 578.11
L2M	AIC = 581.85 BIC = 592.73	AIC = 580.26 BIC = 593.85	AIC = 573.26 BIC = 586.86	AIC = 573.89 BIC = 587.48

TAB. 4-6 – Comparaison et sélection de modèles pour la phase de croissance 1.

L'ajout de l'effet aléatoire n'améliore pas le modèle. Le modèle sélectionné est le modèle linéaire avec la covariable climatique caractérisée par la quantification. Au vu des

différences d'AIC et de BIC non significatives entre le modèle linéaire sans covariable climatique et un modèle linéaire avec covariable climatique, il est difficile de conclure que les stress hydriques influencent la croissance sur cette première phase.

Phase de croissance 2

	Pas de covar clim	Indicatrice	Quantification	Nb stress
LM	AIC = 3123.45 BIC = 3131.36	AIC = 3116.50 BIC = 3128.36	AIC = 3089.49 BIC = 3101.35	AIC = 3103.25 BIC = 3115.11
L2M	AIC = 2942.10 BIC = 2957.91	AIC = 2927.59 BIC = 2947.36	**AIC = 2871.61** **BIC = 2891.37**	AIC = 2894.65 BIC = 2914.42

TAB. 4-7 – Comparaison et sélection de modèles pour la phase de croissance 2.

Le modèle linéaire mixte avec la covariable climatique caractérisée par la quantification est sélectionné. L'ajout de l'effet aléatoire améliore nettement la modélisation. De même, les différences importantes d'AIC et de BIC entre le modèle linéaire mixte sans covariable climatique et le modèle linéaire mixte avec la covariable climatique, caractérisée par la quantification ou le nombre de stress, mettent en évidence une forte influence des stress hydriques durant cette seconde phase de croissance. Ainsi, lorsqu'ils sont âgés de 5 à 18 ans, les pins laricios sont fortement soumis à l'influence du climat et il existe une forte hétérogénéité de la croissance entre les individus. Les individus ne poussent pas tous de la même manière au cours de cette phase là.

Contrairement aux chênes, les pins sont très sensibles à l'effet aléatoire "individu" au cours d'une de leurs phases de croissance. Le découpage en phases permet donc de préciser et de comparer l'influence des deux sources de variation au cours des deux phases.

4.5 Commentaires

La modélisation linéaire mixte multiphasique facilite l'interprétation du botaniste. En effet, des modèles linéaires mixtes (non multiphasiques) ajustant au mieux les données ont été estimés sur toute la période de croissance. Les résultats ne sont pas présentés ici, mais par comparaison avec les résultats présentés dans ce chapitre sur la modélisation linéaire mixte multiphasique, il en ressort que le modèle linéaire mixte associé à une phase de croissance possède une structure beaucoup plus simple (tendance composée d'un simple intercept ou d'un polynôme d'ordre 1, intercept aléatoire, structure de variance-covariance simple) que celle du modèle linéaire mixte multiphasique estimé sur toute la période de croissance (tendance polynomiale d'ordre élevé, intercept et pente aléatoires, structure de variance-covariance complexe). L'interprétation du modèle linéaire mixte sur chacune des phases est donc plus simple et la modélisation linéaire mixte multiphasique permet à la fois de préciser et de différencier les comportements des arbres au cours des différentes phases de croissance.

Conclusion générale et perspectives

Bilan

L'objectif de cette thèse était de développer une famille de modèles pour l'analyse de données longitudinales structurées en phases successives, soumises à l'influence de covariables pouvant varier dans le temps, et présentant une hétérogénéité inter-individuelle. Nous avons proposé une nouvelle famille de modèles statistiques, appelés modèles linéaires mixtes multiphasiques. Un modèle linéaire mixte multiphasique est un modèle de type Markov caché, combinant une chaîne de Markov pour modéliser la succession de phases et des modèles linéaires mixtes associés aux états de la chaîne de Markov sous-jacente. Pour chacune des phases, la tendance et les covariables sont modélisées par des effets fixes, et un effet aléatoire modélise l'hétérogénéité entre les individus.

Des applications de ces modèles sont envisageables dans plusieurs domaines (biomédical, agronomique,...), mais la construction de cette famille de modèles a principalement été motivée par des applications dans le domaine de la botanique telle que l'analyse de la croissance d'arbres forestiers en fonction de facteurs climatiques. Or, prendre en compte des covariables climatiques dans la modélisation soulève un problème de changement d'échelle entre le pas de temps annuel des données botaniques constituant la variable réponse du modèle et le pas de temps journalier des covariables climatiques. Pour résoudre ce problème, nous avons proposé un modèle bioclimatologique relativement rudimentaire, résultant de considérations à la fois biologiques et climatologiques, afin d'obtenir des covariables climatiques ayant le même pas de temps annuel que les données botaniques.

Dans ce travail, ont été proposées deux familles de modèles linéaires mixtes multiphasiques qui diffèrent par le choix de la modélisation de l'effet aléatoire : la séquence observée peut être modélisée par un unique effet aléatoire ou bien chaque phase de la séquence observée peut être modélisée par un effet aléatoire différent.

Le modèle linéaire mixte multiphasique étant un modèle à structure cachée, l'algorithme EM pouvait être envisagé comme méthode d'estimation pour les paramètres de ce modèle. Toutefois, la double structure cachée de ce modèle entraîne des relations d'indépendance conditionnelles spécifiques qui font que l'étape E ne s'écrit pas de manière analytique. L'algorithme EM a donc été écarté comme méthode d'estimation des paramètres pour cette famille de modèles. Nous avons proposé comme alternative à l'algorithme EM un algorithme itératif en trois étapes : restauration, maximisation et prédiction. L'étape de restauration, probabiliste, déterministe, ou effectuée par simulation, nécessite d'intégrer les effets aléatoires pour le calcul des séquences d'états restaurées et nécessite donc

d'adapter les algorithmes habituels pour la restauration des états des chaînes de Markov cachées : algorithme "avant-arrière" pour une restauration probabiliste, algorithme de Viterbi pour une restauration déterministe et algorithme "avant-arrière" de simulation pour une restauration par simulation. De même, l'étape de maximisation nécessite de prendre en compte les effets aléatoires pour le calcul de la probabilité jointe de la séquence d'états restaurée et de la séquence observée. Cette étape est basée sur de simple comptages d'états et de transition d'états à partir des séquences d'états optimales restaurées pour une restauration déterministe ou à partir des séquences d'états simulées pour une restauration par simulation. Dans le cadre d'une restauration probabiliste, l'étape de maximisation utilise les probabilités lissées calculées par l'algorithme "avant-arrière". Par conséquent, une étape de prédiction est nécessaire pour calculer les valeurs prédites des effets aléatoires qui seront utilisées pour les étapes de restauration et de maximisation. Cette prédiction nécessite la connaissance d'une séquence d'états pour une séquence observée ainsi que la connaissance des paramètres estimés à l'itération courante. Elle s'effectue à partir de la séquence d'états optimale restaurée (resp. séquence d'états simulée) dans le cas d'une restauration déterministe (resp. restauration par simulation). Elle pose problème lorsque l'ensemble des séquences d'états possibles est déterminé par restauration probabiliste.

L'intérêt de cette nouvelle famille de modèles a été illustré par des applications à l'analyse de la croissance en longueur d'arbres forestiers. Trois phases de croissance ont été identifiées pour les chênes sessiles âgés de 15 ans. Il n'y a pas d'hétérogénéité entre les individus au cours des différentes phases de croissance. Le climat influence la croissance de manière importante au cours des seconde et troisième phases, alors qu'il l'influence de façon moindre sur la première phase. Deux phases de croissance ont été identifiées pour les pins laricios âgés de 18 ans. Contrairement aux chênes, les pins sont très sensibles à l'effet aléatoire "individu" au cours de leur seconde phase de croissance. Entre 5 ans et 18 ans, les pins laricios sont fortement soumis à l'influence du climat et il existe une forte hétérogénéité de la croissance entre les individus.

Perspectives

Comme perspectives biologiques immédiates au chapitre 4, il serait intéressant d'estimer les paramètres avec l'algorithme itératif avec restauration par simulation, et de comparer les résultats avec ceux obtenus dans le cadre de l'estimation avec une restauration déterministe. Il est aussi important d'effectuer ce même travail avec la modélisation linéaire mixte multiphasique ayant un effet aléatoire unique pour toute la séquence d'observations, afin de juger quelle modélisation est la plus appropriée selon les jeux de données botaniques. L'hypothèse biologique sous-jacente au modèle 1 traduit le fait qu'un arbre avec une croissance lente (resp. rapide) par rapport à la population, sur une des phases de croissance, a une croissance également lente (resp. rapide) durant les phases de croissance suivantes. Cette hypothèse paraît biologiquement plus réaliste que celle sous jacente au modèle 2 qui envisage un comportement d'un individu totalement différent par rapport à la population d'une phase de croissance à l'autre. De même, un travail de simulation de données est indispensable pour valider les deux modélisations proposées.

Jusqu'à présent, seules des séquences relatives à la croissance en longueur d'arbres forestiers ont été modélisées et analysées avec les modèles linéaires mixtes multiphasiques. Il est tout à fait envisageable d'utiliser cette famille de modèles pour la modélisation statistique de séquences relatives à la croissance en épaisseur des arbres (Schweingruber, 1988), en relation avec des facteurs climatiques. Les longueurs des pousses annuelles sont étudiées pour la croissance en longueur alors que les surfaces des cernes seront étudiées pour la croissance en épaisseur.

Un stress hydrique couplé aux températures maximales est susceptible d'influencer la croissance d'arbres forestiers. Toutefois, un gel printanier peut également affecter le début de l'allongement de la pousse annuelle. Par conséquent, il serait intéressant de prendre en compte ces températures dans la modélisation afin d'étudier leur effet sur la croissance.

Enfin, pour mieux comprendre l'effet année (dû aux facteurs climatiques) et l'effet ontogénique, une solution possible serait d'analyser plusieurs échantillons de la même espèce correspondant à des âges différents, mais ayant grandi dans des conditions similaires et durant une période commune.

D'un point de vue statistique, il est essentiel d'étudier les propriétés de l'algorithme itératif en trois étapes proposé. Selon le type de restauration, déterministe ou par simulation, les propriétés de convergence ainsi que le comportement des paramètres estimés et des effets aléatoires prédits doivent être examinés. Il serait intéressant de vérifier si la propriété de croissance monotone de la log-vraisemblance des séquences d'états optimales associées aux séquences observées, vérifiée par l'algorithme de Baum Viterbi, se transpose à l'algorithme proposé où l'on intégre en plus les effets aléatoires.

Plusieurs suggestions concernant la nature de l'étape de restauration sont rapidement envisageables. L'algorithme itératif avec restauration par simulation est de type SEM car une seule séquence d'états est simulée pour un individu donné. Il serait intéressant de faire une comparaison avec un algorithme de type MCEM (Wei et Tanner, 1990) pour lequel plusieurs séquences d'états sont simulées pour un individu. Cela permettrait de contourner l'inconvénient lié à une restauration de type SEM pour laquelle il peut arriver que, pour un individu, la seule séquence simulée soit peu probable.

Une autre possibilité serait de combiner restauration déterministe et restauration par simulation selon un principe de recuit simulé. Autrement dit, à chaque itération, la séquence d'états globalement optimale est restaurée de manière déterministe et plusieurs séquences d'états sont simulées. Selon le principe du recuit simulé, un poids est attribué à la séquence d'états optimale et un autre aux séquences d'états simulées. Le poids est donné sous forme de probabilité de sorte que la somme des deux poids vaut 1. Le poids des perturbations aléatoires joue en fait le même rôle que la température dans les algorithmes de recuit simulé. Au départ, un fort poids est donné aux séquences simulées afin que l'introduction de perturbations aléatoires permette d'éviter de converger vers un minimum local ou un point selle. Ensuite, on réduit le poids des perturbations aléatoires au fil des itérations. Autrement dit, on réduit le poids attribué aux séquences d'états simulées, et de ce fait, on augmente celui attribué à la séquence d'états optimale, afin d'atteindre un meilleur maximum local, voire un maximum global. Après l'étape de restauration, l'étape

de maximisation peut s'effectuer sur la base de simples comptages à partir des séquences d'états restaurées. Mais se pose le problème de la prédiction des effets aléatoires. Dans l'algorithme itératif que nous proposons, la prédiction de l'effet aléatoire ou des effets aléatoires – selon le modèle considéré – s'effectue à partir de la séquence d'états restaurée et des paramètres estimés à l'itération courante. Ici, faut-il prédire le ou les effet(s) aléatoire(s) comme une combinaison linéaire d'effets aléatoires prédits pour chacune des séquences d'états restaurées, en attribuant une pondération à chacun d'eux ? Ou bien tout simplement, vaut-il mieux les prédire à partir de la séquence d'états optimale restaurée avec les paramètres estimés à l'itération courante ?

Comme nous l'avons rappelé dans le bilan, l'étape de prédiction, qui nécessite la connaissance d'une seule séquence d'états pour une séquence observée, pose problème lorsque l'ensemble des séquences d'états possibles est déterminé par restauration probabiliste. La solution la plus simple serait peut être d'utiliser la séquence d'états globalement optimale restaurée avec les paramètres estimés à l'itération courante, et ce quelle que soit la nature probabiliste, déterministe, ou par simulation de l'étape de restauration.

Pour la modélisation linéaire mixte multiphasique, la variable réponse – par exemple la longueur de la pousse annuelle – est supposée suivre approximativement une loi normale. Toutefois, comme nous l'avons mentionné au chapitre 1, certaines informations concernant le nombre de branches par pousse ou le nombre de cycles de croissance sont disponibles pour certains jeux de données. Par conséquent, il serait intéressant de développer une modélisation linéaire généralisée mixte multiphasique pour prendre en compte l'hypothèse de non normalité de ces variables réponses. Autrement dit, on souhaiterait combiner une chaîne de Markov et des modèles linéaires généralisés mixtes (McCulloch et Searle, 2001). Les méthodes d'estimation dans les modèles linéaires généralisés mixtes (Trottier, 1998 ; Trottier et Lavergne, 2000) pourraient-elles s'adapter à l'ajout d'une structure cachée supplémentaire (les états) ? L'algorithme itératif en trois étapes proposé pourrait-il s'adapter aux cas des modèles linéaires généralisés mixtes ?

Bibliographie

Baum, L. E. & Petrie, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains. *Annals of Mathematical Statistics* 37, 1554-1563.

Baum, L. E., Petrie, T., Soules, G. & Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics* 41, 164-171.

Brockwell, P. J. & Davis, R. A. (2002). *Introduction to Time Series and Forecasting*, 2nd edition. New York : Springer.

Burnham, K. P. & Anderson, D. R. (2002). *Model Selection and Multimodel Inference. A Practical Information-Theoretic Approach*, 2nd edn. New York : Springer.

Caraglio, Y. & Barthélémy, D. (1997). Revue critique des termes relatifs à la croissance et à la ramification des tiges des végétaux vasculaires. In : J. Bouchon, P. de Reffye, D. Barthélémy (Eds.), *Modélisation et simulation de l'architecture des végétaux. Institut national de la recherche agronomique*. Paris : INRA Editions, Sciences Update ; p. 11-87.

Celeux, G. & Clairambault, J. (1991). Analyse discriminante appliquée à l'étude du rythme cardiaque : développements méthodologiques. *La Revue de Modulad* 8, 73-80.

Celeux, G. & Diebolt, J. (1985). The SEM algorithm : a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Computational Statistics Quaterly* 2, 73-82.

Celeux, G. & Govaert, G. (1992). A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics and Data Analysis* 14(3), 315-332.

Celeux, G. & Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Classification Journal* 13, 195-212.

Chatfield, C. (2003). *The Analysis of Time Series : An Introduction*, 6th edition. Boca Raton : Chapman & Hall/CRC.

Chib, S. (1996). Calculating posterior distributions and modal estimates in Markov mixture models. *Journal of Econometrics* 75, 79-97.

Churchill, G. A. (1989). Stochastic models for heterogeneous DNA sequences. *Bulletin of Mathematical Biology* 51, 79-94.

Churchill, G. A. & Lazareva, B. (1999). Bayesian Restoration of a Hidden Markov Chain with Applications to DNA Sequencing. *Journal of Computational Biology* 6(2), 261-277.

Crouse, M. S., Nowak, R. D. & Baraniuk, R. G. (1998). Wavelet-based statistical signal processing using hidden Markov models. *IEEE Transactions on Signal Processing* 46(4), 886-902.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series* B 39, 1-38.

Devijver, P. A. (1985). Baum's forward-backward algorithm revisited. *Pattern Recognition Letters* 3, 369-373.

Diggle, P. J. (1990). *Times Series : A Biostatistical Introduction.* Oxford : Oxford University Press.

Diggle, P. J., Heagerty, P., Liang, K.-Y. & Zeger, S. L. (2002). *Analysis of Longitudinal Data*, 2nd edn. Oxford : Oxford University Press.

Durbin, R., Eddy, S. R., Krogh, A. & Mitchison, G. J. (1998). *Biological Sequence Analysis : Probabilistic Models of Proteins and Nucleic Acids*, Cambridge : Cambridge University Press.

Eisenhart, C. (1947). The assumptions underlying the analysis of variance. *Biometrics* 3, 1-21.

Ephraim, Y. & Merhav, N. (2002). Hidden Markov processes. *IEEE Transactions on Information Theory* 48(6), 1518-1569.

Forney Jr., G. D. (1973). The Viterbi Algorithm. *In : Proceedings of the IEEE*, 268-278.

Foulley, J. L, Jaffrezic, F. & Robert-Granié, C. (2000). EM-REML estimation of covariance parameters in Gaussian mixed models for longitudinal analysis. *Genetics Selection Evolution* 32, 129-141.

Foulley, J. L. (2002). Algorithme EM : Théorie et application au modèle mixte. *Journal de la Société Française de Statistique* 143(3-4), 57-109.

Foulley, J. L. (2002). *Bases théoriques du modèle linéaire mixte.* Journées Modèles Mixtes et Biométrie, Paris. *Société Française de Biométrie.*

Fritts, H. C. & Swetnam, T. W. (1989). Dendroecology : A Tool for Evaluating Variations in Past and Present Forest Environments. *Advances in Ecological Research* 19, 111-188.

Geman, S. & Geman, D. (1984). Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images. *IEEE Trans. Pattn. Anal. Mach. Intel.* 6, 721-741.

Godin, C., Guédon, Y., Costes, E. & Caraglio, Y. (1997). Measuring and analysing plants with the AMAPmod software. In *Plants to Ecosystems - Advances in Computational Life Sciences* (Michalewicz, M. T., ed.), Vol. 1, Collingwood, Victoria : CSIRO Publishing, 53-84.

Godin, C., Guédon, Y. & Costes, E. (1999). Exploration of a plant architecture database with the AMAPmod software illustrated on an apple tree hybrid family. *Agronomie* 19, 163-184.

Green, P. J. (1990). Bayesian reconstructions from emission tomography data using a modified EM algorithm. *IEEE Transactions on Medical Imaging* 9, 84-93.

Grossman, M., Koops, W. J. & Den Daas, J. H. G. (1995). Multiphasic Analysis of Reproductive Efficiency of Dairy Bulls. *J. Dairy Sci.* 78, 2871-2876.

Guiot, J. (1986). ARMA techniques for modelling tree-ring response to climate and for reconstructing variations of Paleoclimates. *Ecological Modelling* 33, 149-171.

Hartley, H. O. & Rao, J. N. K. (1967). Maximum likelihood estimation for the mixed analysis of variance model. *Biometrika* 54, 93-108.

Harville, D. A. (1974). Bayesian inference for variance components using only error contrasts. *Biometrika* 61, 383-385.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association* 72, 320-340.

Henderson, C. R., Kempthorne, O., Searle, S. R. & Von Krosigh, C. N. (1959). Estimation of environmental and genetic trends from records subject to culling. *Biometrics* 15, 192-218.

Heuret, P. (2002). *Analyse et modélisation de séquences d'événements botaniques : applications à la compréhension de la régularité d'expression des processus de croissance, de ramification et de floraison.* Thèse de doctorat, Université Henri Poincaré, Nancy-I.

Heuret, P., Barthélémy, D., Nicolini, E. & Atger, C. (2000). Analyse des composantes de la croissance en hauteur et de la formation du tronc chez le chêne sessile (*Quercus petraea* (Matt.) *Liebl.*, *Fagaceae*) en sylviculture dynamique. *Canadian Journal of Botany* 78, 361-373.

Ihaka R, Gentleman R. (1996). R : a language for data analysis and graphics. *Journal of Computational and Graphical Statistics* 5, 299-314.

Jeffreys, H. (1961). *Theory of Probability*, 3rd Edn. Oxford : Oxford University Press.

Jelinek, F., Bahl, L. R. & Mercer, R. L. (1975). Design of a linguistic statistical decoder for the recognition of continuous speech. *IEEE Transactions on Information Theory* 21, 250-256.

Jelinek, F. (1976). Continuous speech recognition by statistical methods. *Proceedings of the IEEE* 64(4), 532-556.

Juang, B.-H. & Rabiner, L. R. (1990). The segmental K-means algorithm for estimating parameters of hidden Markov models. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 38(9), 1639-1641.

Krogh, A., Brown, M., Mian, I., Sjolander, K. & Haussler, D. (1994). Hidden Markov models in computational biology : Applications to protein modeling. *Journal of Molecular Biology* 235, 1501-1531.

Laird, N. M & Ware, J. H. (1982). Random effects models for longitudinal data. *Biometrics* 38, 963-974.

Lauritzen, S. L. (1996). *Graphical Models*. Oxford : Oxford University Press.

Levinson, S. E., Rabiner, L. R. & Sondhi, M. M. (1983). An Introduction to the Application of the Theory of Probabilistic Functions of a Markov Process in Automatic Speech Recognition. *Bell System Technical Journal* 62, 1035-1074.

Lindley, D. V. & Smith, A. F. M. (1972). Bayes Estimates for the Linear Model. *J. R. Statist Soc* B, 34, 1-41.

Martin, O. (2002). *Approches statistiques pour l'analyse de données de puces à ADN*. Thèse de doctorat, Université Joseph Fourier, Grenoble.

McCulloch, C. E. & Searle, S. R. (2001). *Generalized, Linear and Mixed Models*. New York : Wiley.

McLachlan, G. J. & Krishnan, T. (1997). *The EM Algorithm and Extensions*. New York : Wiley.

Méredieu, C. (1998). *Croissance et branchaison du Pin Laricio (Pinus nigra Arn. ssp. laricio (Poiret) Maire) : élaboration et évaluation d'un système de modèles pour la prévision de caractéristiques des arbres et du bois*. Thèse de doctorat, Université Claude Bernard Lyon-I.

Merhav, N. & Ephraim, Y. (1991). Hidden Markov modeling using a dominant state sequence with application to speech recognition. *Computer, Speech, and Language* 5, 327-339.

Monserud, R. A. (1986). Time-series analyses of tree-ring chronologies. *Forest Science* 32(2), 349-372.

Muri, F. (1997). *Comparaison d'algorithmes d'identification de chaînes de Markov cachées et application à la détection de régions homogènes dans les séquences d'ADN*. Thèse de doctorat, Université Paris-V.

Nougarède, A. (2001). Le méristème caulinaire des Angiospermes : nouveaux outils, nouvelles interprétations. *Acta Botanica Gallica* 148(1), 3-77.

Oldeman, R. A. A. (1974). *L'architecture de la forêt guyanaise*. Paris : ORSTOM, Mémoire n°73.

Patterson, H. D. & Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika* 58, 545-554.

Potthoff, R. F. & Rao, S. N. (1964). A generalized multivariate analysis of variance model usefull especially for growth curve problems. *Biometrika* 51, 313-326.

Qian, W. & Titterington, D. (1990). Parameter estimation for hidden Gibbs chains. *Statist. Probab. Lett.* 10, 49-58.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE* 77, 257-286.

Rao, C. R. & Kleffe, J. (1988). *Estimation of variance components and applications*. North Holland series in statistics and probability. Elsevier, Amsterdam.

Robert, C., Celeux, G. & Diebolt, J. (1993). Bayesian estimation of hidden Markov chains : A stochastic implementation. *Statist. Probab. Lett.* 16, 77-83.

Schweingruber, F. H. (1988). *Tree Rings. Basics and Applications of Dendrochronology*. Kluwer Academic Publishers.

Searle, S. R., Casella, G. & McCulloch, C. E. (1992). *Variance components*. New York : Wiley.

Sorensen, K., Grossman, M. & Koops, W. J. (2003). Multiphasic Growth Curves in Mink (Mustela vison) Selected for Feed Efficiency. *Acta Agric. Scand.* (A) : 53, 41-50.

Smyth, P., Heckerman, D. & Jordan, M. I. (1997). Probabilistic independence networks for hidden Markov probability models. *Neural Computation* 9, 227-269.

Trottier, C. (1998). *Estimation dans les modèles linéaires généralisés à effets aléatoires*. Thèse de doctorat, Institut National Polytechnique de Grenoble.

Trottier, C. & Lavergne, C. (2000). Sur l'estimation dans les modèles linéaires généralisés mixtes, *Revue de Statistiques Appliquées*, 48(1), 45-63.

Verbeke, G. & Molenberghs, G. (1997). *Linear Mixed Models in Practice : A SAS-Oriented Approach*. Lecture Notes in Statistics 126. New-York : Springer-Verlag.

Verbeke, G. & Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. New York : Springer.

Viterbi, A. J. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory* 13, 260-269.

Wei, G. C. G. & Tanner, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. J*ournal of the American Statistical Association* 85, 699-704.

Zangwill, W. I. (1969). *Nonlinear Programming : A Unified Approach*. Englewood Cliffs, NJ : Prentice-Hall.

Modèles linéaires mixtes multiphasiques pour l'analyse de données longitudinales - Application à la croissance des plantes

Nous proposons une nouvelle famille de modèles statistiques, appelés modèles linéaires mixtes multiphasiques, pour analyser des données longitudinales structurées en phases successives, soumises à l'influence de covariables pouvant varier dans le temps et présentant une hétérogénéité inter-individuelle. La construction de cette famille de modèles a été motivée par des applications dans le domaine de la botanique, mais d'autres applications dans le domaine biomédical ou agronomique sont possibles.

Le modèle linéaire mixte multiphasique est un modèle de type Markov caché, combinant une chaîne de Markov pour modéliser la succession de phases et des modèles linéaires mixtes associés aux états de la chaîne de Markov sous-jacente. Sont présentées deux familles de modèles linéaires mixtes multiphasiques qui différent par le choix de la modélisation de l'effet aléatoire.

Nous étudions le problème de l'estimation des paramètres de ces deux modèles. L'algorithme EM est écarté comme méthode d'estimation. Nous proposons comme alternative à l'algorithme EM un algorithme itératif en trois étapes : restauration, maximisation et prédiction. L'étape de restauration peut être probabiliste, déterministe ou effectuée par simulation.

L'intérêt de cette nouvelle famille de modèles est illustré par des applications à l'analyse de la croissance en longueur d'arbres forestiers en fonction de facteurs climatiques.

Mots-Clés : Données longitudinales, modèle de Markov caché, modèle linéaire mixte, effet aléatoire, algorithme de restauration maximisation prédiction, croissance en longueur, arbres forestiers.

Multiphasic lineard mixed models for the analysis of longitudinal data - Application to plant growth

We propose a new family of statistical models, called Markov switching linear mixed models or multiphasic linear mixed models, for analysing longitudinal data structured in successive phases, influenced by time-varying covariates and presenting heterogeneity between individuals. The construction of this family of models was justified by applications in the field of botany, but others applications in the biomedical or agronomic field are possible.

A Markov switching linear mixed model is a hidden Markovian model that combines a Markov chain to model the succession of phases and linear mixed models associated with the states of the underlying Markov chain. Two families of Markov switching linear mixed models, which differ by the choice of the random effect modelling, are presented.

The problem of estimating the parameters of these two models is addressed. The EM algorithm cannot be applied. As an alternative to the EM algorithm, we propose an iterative algorithm which decomposes in three steps : restoration, maximization and prediction. The restoration step can be probabilistic, deterministic or achieved by simulation.

The relevance of this new family of models is illustrated by applications to the analysis of forest trees growth according to climatic factors.

Keywords : Longitudinal data, hidden Markov model, linear mixed model, random effect, restoration maximization prediction algorithm, plant growth, forest trees.

Discipline : Statistique
Laboratoire : UMR **AMAP** bot**A**nique et bioinfor**M**atique de l'**A**rchitecture des **P**lantes
TA 40 / PS2 - boulevard de la lironde - 34398 Montpellier - France